Computational Methods for the Study of Plant-associated Microbial Communities

Kumulative Dissertation

zur
Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Heinrich-Heine-Universität Düsseldorf

vorgelegt von

Ruben Garrido Oter
aus Madrid

Düsseldorf, Februar 2017

aus dem Institut für Informatik
der Heinrich-Heine-Universität Düsseldorf

Gedruckt mit der Genehmigung der
Mathemathisch-Naturwissenschaftlichen Fakultät der
Heinrich-Heine-Universität Düsseldorf

Referent:	Prof. Dr. Paul Schulze-Lefert
Koreferent:	Prof. Dr. Alice C. McHardy
Koreferent:	Prof. Dr. Laura Rose
Datum der mündlichen Prüfung:	06.06.2017

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Dissertation eigenständig und ohne fremde Hilfe angefertig habe. Arbeiten Dritter wurden entsprechend zitiert. Diese Dissertation wurde bisher in dieser oder ähnlicher Form noch bei keiner anderen Institution eingereicht. Ich habe bisher keine erfolglosen Promotionsversuche unternommen.

Düsseldorf, Februar 2017

....

(Ruben Garrido Oter)

Statement of authorship

I hereby certify that this dissertation is the result of my own work. No other person's work has been used without due acknowledgement. This dissertation has not been submitted in the same or similar form to other institutions. I have not previously failed a doctoral examination procedure.

Summary

Higher organisms such as animals and land plants host diverse communities of microorganisms which are collectively designated as microbiota. There is a growing body of evidence that establishes links between these microbial assemblages and the fitness of their host, for example *via* indirect protection against pathogens or enhanced nutrient acquisition. Additionally, host-microbiota systems can be used as models to investigate the principles underlying microbial community structure and assembly and the dynamics of co-adaptation between multiple organisms.

The aim of this work was to develop and apply computational methods for the analysis of sequence data obtained from environmental samples of plant-associated microbes as well as genomic sequences of cultured isolates in a comparative framework. Employing culture-independent community profiling techniques (e.g. 16S rRNA gene amplicon surveys or shotgun-metagenomics) we were able to describe and characterize the taxonomic structure and functional potential of the plant microbiota across multiple hosts, including the model *Arabidopsis thaliana* and relatives, the crop barley (*Hordeum vulgare*) or the legume *Lotus japonicus*. In addition, we have developed a number of culture-dependent methods to study the plant microbiome, including the characterization of a large collection of cultured bacterial microbiota members using a sequence-indexed library of more than 5,000 colony-forming units. Whole-genome sequencing of a taxonomically representative subset of 400 isolates from this collection revealed a large overlap of functional capabilities between leaf- and root-derived bacteria as well as few significant differences at the level of individual functional categories. A targeted, large-scale isolation and sequencing effort focused on the Rhizobiales order, a taxonomic group of particular interest which includes members that are capable of engaging in highly adapted and beneficial symbiotic interactions with legumes resulted in the generation of a dataset of more than 900 draft genomes, which includes a large number from previously uncharted branches of the species tree of rhizobia. Phylogenomic analysis of these sequences provided evidence of an ancestral form of association between rhizobia and flowering plants that predates the capacity for nodulation and nitrogen fixation, which ancestral reconstruction of relevant genomic features suggests was acquired in multiple subsequent events, most likely *via* horizontal gene transfer, in an example of convergent evolution.

Finally, we developed a novel phylogenetic approach for determining clusters of co-evolving genes and their network organization by modeling gene gain and loss as a continuous process along the branches of the species tree. This method accounts for

uncertainty in the reconstruction of the ancestral states as well as in the inference of the species tree and robustly identifies clusters of co-evolving genes that significantly enrich for functional categories and pathways and which are relevant for adaptation to diverse environments. We demonstrate the potential of this approach to detect biologically meaningful gene family interactions and predict genotype-phenotype relationships by analyzing a total of 2,737 bacterial genomes from diverse environments, including plant commensals and symbionts as well as human pathogens.

In summary, we have generated and analyzed large quantities of sequencing data that provide a taxonomic and functional characterization of the plant microbiota, constituting a large dataset and a valuable resource for future research. Additionally, novel methodology for the analysis of collections of microbial genomes provides tools for the identification of sets of genes involved in relevant biological processes as well as links to corresponding phenotypes. Extending these datasets and the available number of sequenced genomes, together with further development of phylogenenomic methods has the potential to greatly improve our understanding of the processes that drive the adaptation of microbes in the context of the complex communities which they form.

Zusammenfassung

Höhere Organismen wie Tiere oder Landpflanzen beherbergen vielfältige Gemeinschaften von Mikroorganismen, die in ihrer Gesamtheit als Mikrobiota bezeichnet werden. Es gibt immer mehr Hinweise darauf, dass diese mikrobiellen Zusammenschlüsse mit der Fitness ihrer Wirte in Verbindung stehen, beispielsweise durch indirekten Schutz gegen Pathogene oder durch verbesserte Nährstoffaufnahme. Des Weiteren dienen Wirt-Mikrobiota Systeme als Modelle für die Untersuchung der Prinzipien, die der mikrobiellen Gemeinschaftsstruktur und des Gemeinschaftsaufbaus zugrundeliegenden, sowie der Dynamik der Co-Adaption zwischen zahlreichen Organismen.

Das Ziel dieser Arbeit war die Entwicklung und Anwendung bioinformatischer Methoden für die Analyse von Sequenzdaten, die aus Umweltproben pflanzenassoziierter Mikroben gewonnen wurden, und von genomischen Sequenzen im Labor kultivierter Isolate, in einem vergleichbaren Rahmen/Struktur. Kultur-unabhängige Techniken zur Profilbildung/Profiling von Gemeinschaften (z.B. 16S rRNA Amplikon-Untersuchung und Shotgun-Metagenomics) ermöglichten uns die Beschreibung und Charakterisierung der taxonomischen Struktur und des funktionellen Potentials der Pflanzenmikrobiotia mehrerer verschiedener Wirte, darunter die Modellpflanze Arabidopsis thaliana und verwandte Arten, die Nutzpflanze Gerste (Hordeum vulgare) und die Leguminose Lotus japonicus. Zusätzlich haben wir einige Kultur-abhängige Methoden entwickelt, um das Pflanzenmikrobiom zu untersuchen, wie zum Beispiel die Charakterisierung einer großen Kollektion von kultivierbaren bakteriellen Mikrobiotamitgliedern, wofür wir eine Barcode-basierte Sequenzierungsbibliothek mit mehr als 5,000 kolonieformenden Einheiten verwendet haben. Whole Genome Sequencing eines taxonomisch repräsentativen Anteils von 400 Isolaten aus dieser Kollektion hat eine große überlappung funktioneller Fähigkeiten zwischen Blatt- und Spross-stämmigen Bakterien ergeben, sowie wenige signifikante Unterschiede bezüglich einzelner funktioneller Kategorien. Eine gezielte großangelegte Isolierungs- und Sequenzierungsaktion konzentrierte sich auf die Ordnung Rhizobiales. Zu dieser taxonomische Gruppe von speziellem Interesse gehören Arten, die hoch angepasste und nutzbringende symbiotische Interaktionen mit Leguminosen eingehen können. Ein Datenset von mehr als 900 Genomen wurde generiert, unter diesen viele von bisher nicht kartierten ästen des Rhizobien-Artenbaums. Die phylogenetische Analyse dieser Sequenzen lieferte den Beweis für eine Urform des Verbands zwischen Rhizobien und Blütenpflanzen, die noch vor der Fähigkeit zur Nodulation und zur Stickstofffixierung datiert ist. Letztere wurde der Abstammungsrekonstruktion von relevanten Genomeigenschaften zufolge in mehreren aufeinanderfolgenden Vorgängen erworben, höchst wahrscheinlich durch horizontalen Gentransfer, als Beispiel konvergenter Evolution.

Schließlich haben wir einen neuen phylogenetischen Ansatz entwickelt, um die Cluster von co-entwickelnden Genen und deren Netzwerkorganisation zu bestimmen, indem wir Gengewinn und Genverlust als kontinuierlichen Prozess entlang der äste des Artenbaumes modelliert haben. Diese Methode bezieht die Unsicherheit der Rekonstruktion der Urzustandes mit ein, wie auch der Interferenz des Artenbaumes, und identifiziert robust Cluster von co-entwickelnden Genen, die signifikant mit funktionellen Kategorien und Signalwegen angereichert sind, welche relevant für die Anpassung an unterschiedliche Umgebungen sind. Wir demonstrieren das Potential dieses Ansatzes, indem wir Interaktionen von biologisch bedeutsamen Genfamilien ermitteln und Genotyp-Phänotyp Beziehungen anhand der Analyse von 2,737 Bakteriengenomen aus verschiedenen Umgebungen vorhersagen, darunter sowohl Pflanzensymbionten als auch Menschenpathogene.

Zusammengefasst haben wir große Mengen an Sequenzierungsdaten generiert und analysiert, die eine taxonomische und funktionelle Charakterisierung der Pflanzenmikrobiota bereitstellen, was eine wertvolle Ressource für die zukünftige Forschung darstellt. Weiterhin liefert solch eine neuartige Methodik für die Analyse von mikrobiellen Genomsammlungen Werkzeuge für die Identifizierung von Gensets, die in relevanten biologischen Prozessen eine Rolle spielen, sowie die Verbindungen zu den korrespondierenden Phänotypen. Die Erweiterung dieser Datensets und der Anzahl verfügbarer sequenzierter Genome hat, zusammen mit der Weiterentwicklung der phylogenetischen Methoden, das Potenzial, unser Verständnis der Prozesse stark zu verbessern, welche die Anpassung von Mikroben im Kontext der von ihnen gebildeten komplexen Gemeinschaften vorantreiben.

Acknowledgements

First, I would like to thank my advisors Paul Schulze-Lefert and Alice C. McHardy. It has been an honor and a privilege to be their student.

It was Alice who first gave me the opportunity to conduct the work presented in this thesis and who trained me with generosity and patience during the earlier stages of my doctorate; for that I will always be in debt. I have the deepest admiration for her keen intellect and her ability to get to the core of any problem while quickly discarding the unimportant details. I am very grateful for all the time and effort she has invested in me despite of ever-changing personal and professional circumstances.

I find myself lacking the words to thank my mentor, Paul, for his unwavering support. From the very first day he has done nothing other than encourage me and help me grow as a scientist. The enthusiasm and joy with which he approaches research have been a constant source of motivation during my years as a doctoral student. I am perhaps most of all grateful to him for keeping the door of his office always open –sometimes at the risk of getting caught in hours-long discussions– no matter how late in the evening or how busy the day. It is still hard for me to come to terms with all the time and effort and all the trust that he has deposited in me. I genuinely believe that it is not well deserved. I should very much hope to one day have the opportunity to prove him right.

Most of the projects that constitute the chapters of this thesis are the result of close collaborations and joint work with various people, sometimes shoulder-to-shoulder and over many months. I would like to give special thanks to these fantastic colleagues and co-first authors: Davide Bulgarelli, Yang Bai, Rafal Zgadzaj, Nina Dombrowski and Thomas Nakano. I was very lucky to be able to benefit so much from their hard work and expertise.

I would also like to thank the members of the former Algorithmic Bioinformatics group at the HHUD and of the Innate Immunity and the Plant Microbiota group at the MPIPZ for creating such great working environments. I have spent several years in constant awe at the amazingly high scientific and personal caliber of my colleagues, many of which I also have the pleasure to call my friends.

Without the help and love of my family I would not have come all this way. Everything that I have accomplished, including this thesis, I owe to them. I also want to aknowledge my little niece, Valeria, and my little nephew, Ernesto, for being patient in their own way with my long absences. I hope some day I can make it up to them.

Finally, I want to thank my partner Julian for all of his patience and support. I am truly fortunate to have such a kind and loving person by my side.

Contents

List of Figures xix

List of Tables xxiii

1 Introduction 1

1.1 Motivation and research aim 1

1.2 Outline 2

1.3 The plant microbiota 3

1.3.1 Compartment-specific microbial assemblages 3

1.3.2 The robust structure of the root microbiota 5

1.3.3 Factors driving community diversity 6

1.3.4 Host-microbiota co-evolution 8

1.4 Computational methods in microbiome studies 10

1.4.1 Marker gene amplicon data 10

1.4.2 Diversity assessment and statistical analyses 11

1.4.3 Shotgun metagenomics 11

1.4.4 Phylogenomic analyses of microbiota members 12

1.5 Outlook 13

2 Personal bibliography 17

I The structure of the plant root microbiota 19

3 First publication — Microbiota and host nutrition across plant and animal kingdoms 21
3.1 Abstract 22
3.2 Physiological functions of the vertebrate gut and plant roots 22
3.3 Compartmentalization of the gut and root microbiota 25
3.4 Community structure of the vertebrate gut and plant root microbiota . 26
3.4.1 Where do they come from? 26
3.4.2 Who are they? 27
3.4.3 Are there structural similarities across diverse host-associated microbial communities? 28
3.4.4 Do they fluctuate over time? 30
3.5 Major factors driving community establishment and composition 31
3.5.1 Environmental factors 31
3.5.2 Nutritional drivers 34
3.5.3 Microbe-microbe interactions 37
3.5.4 Host genotype 37
3.6 Host immune systems and microbiota homeostasis 38
3.7 Co-diversification of host-microbe communities 40
3.8 Metagenome analysis-inferred functions of the gut and the plant microbiota 41
3.9 Concluding remarks and perspectives 42
3.10 Author contributions 43
3.11 Acknowledgements 43
3.12 Supporting material 43

4 Second publication — Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives 45
4.1 Significance 46
4.2 Abstract 46
4.3 Introduction 47
4.4 Results 49
4.4.1 Defining abundant community members 50
4.4.2 Community composition is defined more strongly by environmental parameters than by host species 51

4.4.3 Naturally grown *A. thaliana* and *C. hirsuta* host a taxonomically narrow root microbiota 52
4.4.4 Phylogenetic distance of host species contributes to microbiota diversification 54
4.4.5 Members of the Actinomycetales, Burkholderiales, and Flavobacteriales are stable across host species and environments 57
4.5 Discussion 59
4.5.1 A conserved core root microbiota? 59
4.5.2 Root microbiota interactions with the environment 61
4.5.3 Host microbiota co-evolution 62
4.6 Materials and methods 63
4.7 Author contributions 63
4.8 Acknowledgements 64
4.9 Supporting material 64

5 Third publication — Structure and functions of the bacterial root microbiota in wild and domesticated barley 65

5.1 Summary 66
5.2 Introduction 66
5.3 Results 67
5.3.1 The structure of the barley bacterial microbiota 67
5.3.2 The barley rhizosphere microbiome 74
5.3.3 Comparison of SSU rRNA genes and metagenome taxonomic abundance estimates 74
5.3.4 Enrichment of biological functions in root- and rhizosphere-associated bacterial taxa 76
5.3.5 Positive selection in the barley rhizosphere 78
5.3.6 Microbial elicitors and effectors of plant immunity under positive selection 79
5.3.7 Positive selection acting on phages and CRISPR systems 81
5.4 Discussion 83
5.5 Experimental procedures 87
5.5.1 Experimental design 87
5.5.2 16S data analysis 87
5.5.3 Metagenome data analysis 88
5.6 Author contributions 88

5.7 Acknowledgements . . . 89
5.8 Accession numbers . . . 89
5.9 Supporting material . . . 89

6 Fourth publication — Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities 91
6.1 Abstract . . . 93
6.2 Significance . . . 93
6.3 Introduction . . . 93
6.4 Results . . . 96
6.4.1 Characterization of the *Lotus japonicus* root, nodule and rhizosphere microbiota . . . 96
6.4.2 Parallel selection of nodule- and root-specific bacteria from the rhizosphere compartment . . . 99
6.4.3 Impairment of nitrogen-fixing symbiosis dramatically alters bacterial community structure in the *Lotus* root and rhizosphere compartments . . . 99
6.4.4 The symbiosis pathway drives root and rhizosphere differentiation across multiple bacterial orders . . . 105
6.4.5 Comparable immune- and symbiosis-related metabolic responses in soil-grown WT and symbiotic mutant roots . . . 107
6.4.6 *Lotus japonicus* and various Brassicaceae species assemble highly diverged root-inhabiting bacterial communities . . . 107
6.4.7 Symbiosis-impaired mutants maintain an altered community structure in nitrogen-supplemented soil . . . 110
6.5 Discussion . . . 110
6.6 Materials and methods . . . 114
6.6.1 Soil and plant material . . . 114
6.6.2 Sample and 16S rRNA library preparations . . . 115
6.6.3 Quantitative RT-PCR . . . 115
6.6.4 Computational analyses . . . 115
6.6.5 Metabolite analyses . . . 117
6.7 Author contributions . . . 117
6.8 Acknowledgements . . . 117
6.9 Supporting material . . . 117

II Functions of root- and leaf-associated microbes 129

7 Fifth publication — Functional overlap of the *Arabidopsis* leaf and root microbiota 131

7.1 Abstract 133
7.2 Introduction 133
7.3 Results 134
7.3.1 Bacterial culture collections from roots and leaves 134
7.3.2 *At*-RSPHERE and *At*-LSPHERE culture collections 135
7.3.3 Comparative genome analysis of the culture collections 137
7.3.4 Synthetic community colonization of germ-free plants 141
7.3.5 Niche-specific microbiota establishment with SynComs 144
7.4 Conclusions 145
7.5 Methods 146
7.5.1 Sampling of *A. thaliana* plants and isolation of root-, leaf- and soil-derived bacteria 146
7.5.2 Culture-independent bacterial 16S rRNA gene profiling of *A. thaliana* leaf, root and corresponding soil samples 147
7.5.3 High-throughput identification of leaf-, root- and soil-derived bacterial isolates by 454 pyrosequencing 148
7.5.4 Preparation of *A. thaliana* leaf, root and soil bacterial culture collections 149
7.5.5 Preparation of bacterial genomic DNA for whole-genome sequencing 149
7.5.6 Genome assembly and annotation 150
7.5.7 Analyses of phylogenetic diversity within sequenced isolates . . . 151
7.5.8 Analyses of functional diversity between sequenced isolates . . . 151
7.5.9 Recolonization experiments of leaf-, root- and soil-derived bacteria on *Arabidopsis* 151
7.5.10 Accession numbers 154
7.5.11 Code availability 154
7.6 Author contributions 154
7.7 Acknowledgements 154
7.8 Supporting material 155

III Comparative genomics of Rhizobia 159

8 Sixth publication — Assessment of functional diversification and adaptation in Rhizobia by comparative genomics 161

8.1 Abstract 162
8.2 Introduction 162
8.3 Results 164
8.3.1 Rhizobia are important components of the core root microbiota across a wide taxonomic variety of plant hosts 164
8.3.2 Comparative analysis of sequenced isolates reveals a large pangenome with high functional diversity 169
8.3.3 Convergent evolution of nitrogen-fixing symbiosis in Rhizobia . . 173
8.3.4 The majority of rhizobial species engage in root-growth promoting interactions with non-legume hosts 175
8.4 Discussion 177
8.5 Materials and methods 179
8.5.1 Isolation and sequencing of strains 179
8.5.2 Genome assembly 181
8.5.3 Annotation and orthology inference 181
8.5.4 Natural community amplicon sequencing meta-analysis 181
8.5.5 Comparative genomics and ancestral character reconstruction . . 182
8.5.6 Binary association experiments with germ-free *Arabidopsis plants* 183
8.6 Author contributions 184
8.7 Acknowledgements 184
8.8 Supporting material 184

IV Evolutionary profiles 185

9 Seventh publication — Clustering of functionally related genes reveals novel symbiosis-relevant genes in Rhizobia 187

9.1 Abstract 188
9.2 Introduction 188
9.3 Results 190
9.3.1 Network of co-evolving genes and functional modules in Rhizobia 190
9.3.2 Modules enrich for functionally related genes 197
9.3.3 Identification of novel symbiosis-related genes in Rhizobia 198

9.3.4 Functional modules in the *Arabidopsis* root and leaf microbiota . 199
9.3.5 Predicting the genetic basis of antibiotic resistance in *S. aureus* . 202
9.4 Discussion 205
9.5 Methods 208
9.5.1 Genome assembly 208
9.5.2 Inference of homology relationships between genes 209
9.5.3 Maximum likelihood inference of ancestral characters 209
9.5.4 Modeling gene gains and losses 210
9.5.5 Clustering of evolutionary profiles into functional modules 211
9.5.6 Statistical analyses of module functional enrichment 214
9.5.7 Code availability 214
9.6 Author contributions 214
9.7 Acknowledgements 215
9.8 Supporting material 215

References **217**

A Journal versions of the published articles **253**

List of Figures

1.1 Plant microbiota members visualized by fluorescence in situ hybridization 5

3.1 Physiological functions of the plant roots and human gut 23
3.2 Alpha- and beta-diversity analyses 28
3.3 3D PCoA plots 30
3.4 Cumulative abundance plots (Caption) 32
3.4 Cumulative abundance plots (Figure) 33
3.5 Phylogenetic analysis of OTU abundances (Caption) 35
3.5 Phylogenetic analysis of OTU abundances (Figure) 36

4.1 Phylogeny of *Arabidopsis thaliana* and relative species 48
4.2 Hierarchical clustering of samples 52
4.3 Root microbiota comparisons of *A. thaliana* and *C. hirsuta* at the natural sites Cologne and Eifel. 54
4.4 Root microbiota comparisons of *A. halleri*, *A. lyrata*, *A. thaliana*, and *C. hirsuta* 56
4.5 Identification of the *Arabidopsis thaliana* and relative species core microbiota 57

5.1 The barley rhizosphere and root microbiota are gated communities . . 68
5.2 Constrained PCoA on the soil and barley bacterial microbiota 70

5.3 OTU enrichment at the barley root/soil interface (Figure) 71
5.3 Barley OTU enrichment analysis (Caption) 72
5.4 Taxonomic representation of the barley and *Arabidopsis* root-enriched bacterial taxa 73
5.5 Comparison of 16S rRNA amplicon and metagenome abundances 75
5.7 Proteins under selection in the barley rhizosphere microbiome (Cont.) . 80
5.6 Proteins under selection in the barley rhizosphere microbiome 82

6.1 Images depicting *L. japonicus* WT and mutant plants. 96
6.2 Constrained PCoA analyses of beta-diversity 98
6.3 Root- rhizosphere- and nodule-enriched OTUs in wild-type *L. japonicus* 100
6.4 Compartment-enriched OTUs in WT and mutant *Lotus japonicus* . . . 102
6.5 Manhattan plots showing root- and rhizosphere-enriched OTUs in WT and mutant *Lotus* (Caption) 103
6.5 Manhattan plots showing root- and rhizosphere-enriched OTUs in WT and mutant *Lotus* (Figure) 104
6.6 Differentially enriched OTUs between WT and mutant *Lotus* 106
6.7 Comparison between *Arabidopsis* and *Lotus* root bacterial communities. 108
6.8 Comparison between *Arabidopsis* and *Lotus* root bacterial communities (Cont.). 109
6.9 Harvesting procedures applied to *Lotus* plants 119
6.10 Analysis of alpha-diversity 120
6.11 PCoA analysis of beta-diversity by soil batch 121
6.12 CPCoA analysis after *in silico* depletion of nodule-enriched OTUs. . . . 122
6.13 Manhattan plots showing differentially abundant OTUs in WT and mutant *Lotus* (Figure) 123
6.13 Manhattan plots showing differentially abundant OTUs in WT and mutant *Lotus* (Caption) 124
6.14 Phylum-level relative abundances in WT and mutant *Lotus* roots 124
6.15 Expression of immune- and symbiosis-related marker genes in WT and mutant 125
6.16 Concentration of soluble proteins and nitrate in *Lotus* roots 126
6.17 Macroscopic phenotypes of *L. japonicus* WT and mutant plants grown under nitrogen-supplemented conditions 127
6.18 Beta-diversity analyses of nitrogen-suplemented *L. japonicus* WT and mutant plants 128

7.1 Taxonomic distribution of *At*-RSPHERE and *At*-LSPHERE (Caption) . 135
7.1 Taxonomic distribution of *At*-RSPHERE and *At*-LSPHERE (Figure) . . 136
7.2 Analysis of functional diversity between sequenced isolates 138
7.3 Functional analysis of sequenced isolates (Figure) 140
7.3 Functional analysis of sequenced isolates (Caption) 141
7.4 SynCom colonization of germ-free *A. thaliana* plants 142
7.5 SynCom competition supports host-organ-specific community assemblies 143
7.6 Culture-dependent coverage of *A. thaliana* root- and leaf-associated OTUs identified in several cultivation-independent studies 156
7.7 *At*-RSPHERE, *At*-LSPHERE and soil bacterial culture collections . . . 157

8.1 Rhizobia are important components of the core root microbiota across a wide taxonomic variety of hosts (Figure) 166
8.1 Rhizobia are important components of the core root microbiota across a wide taxonomic variety of hosts (Caption) 167
8.2 Non-metric Multidimensional Scaling (NMDS) of Bray-Curtis distances between rhizobial abundances 168
8.3 Analysis of functional diversity between sequenced genomes of rhizobia . 170
8.4 Accumulation curves displaying the estimated pan-genome size of the Rhizobia and *At*-SPERE collections 172
8.5 Whole-genome phylogeny of Rhizobia and Maximum Likelihood reconstruction of ancestral symbiotic genotypes 174
8.6 Binary interaction experiments between rhizobia and germ-free *Arabidopsis* plants grown on agarose media 176

9.1 Network of co-evolving genes in Rhizobia 191
9.2 Relationship between network proximity and functional similarity 192
9.3 Adjacency matrix showing the highly clustered structure of the co-evolution network of Rhizobia 193
9.4 Co-evolution of genes in the symbiosis module in Rhizobia 194
9.5 Co-evolution of genes in the T6SS module in Rhizobia (Figure) 195
9.5 Co-evolution of genes in the T6SS module in Rhizobia (Caption) 196
9.6 Co-evolution of genes in the exopolisaccaride biosynthesis module in Rhizobia 197
9.7 Co-evolution of genes in the flaggelar apparatus module in the *At*-SPHERE genomes 200

9.8 Co-evolution of genes in the methane metabolism module in the *At*-SPHERE genomes . . . 201

9.9 Significantly correlated genes with patterns of gain and loss of antibiotic resistance in *S. aureus* (Figure) . . . 202

9.9 Significantly correlated genes with gain and loss of antibiotic resistance phenotypes in *S. aureus* (Caption) . . . 203

9.10 Species tree and phylogenetic distribution of antibiotic resistance in *S. aureus* . . . 204

List of Tables

3.1 Percentage of shotgun metagenome reads assigned to each kingdom of life across metagenome studies 25

4.1 Numerical overview of the experimental setup 50

5.1 Functional categories significantly enriched in taxonomic bins corresponding to RR_OTUs found in the barley rhizosphere metagenome 77

9.1 Significant functional enrichment of large clusters 198

CHAPTER 1

Introduction

1.1 Motivation and research aim

Healthy higher organisms, such as animals and land plants are colonized by vast numbers of diverse microbes, which are collectively designated as microbiota. These microbes assemble into complex communities that are established and persist in close contact with their host according to organizing principles that only in recent years we are beginning to understand. The link between these microbial assemblages and the health and ecological performance of their host are the focus of a rapidly growing interdisciplinary field of research that lies at the intersection of molecular biology, ecology and computational biology.

The aim of this work was to develop and apply computational approaches to the analysis of sequence data obtained from environmental samples of plant-associated microbes as well as genomic sequences of cultured isolates in a comparative framework. In particular, the computational analyses performed should aid in answering questions such as what is the composition of the plant microbiota in terms of taxonomic affiliation and functions and what are the underlying principles governing microbiota establishment. Furthermore, by generating large collections of sequenced and annotated genomes of cultured community members (in total >1,400 newly assembled and annotated genomes), I aimed at providing a valuable resource to be used for future research. Finally, a novel phylogenomic approach based on the reconstruction of the evolutionary histories of gene families (evolutionary profiles) was designed and implemented with the goal of inferring genotype-phenotype relationships and identifying modules of genes that are involved in the same biological processes, with a focus on mechanisms of interaction between plant-associated microbes and their host.

1.2 Outline

This work is a cumulative dissertation that contains five peer-reviewed articles published in international journals as well as two articles which are presented here before submission for scientific review (see Personal Bibliography). The author of this thesis has made a significant contribution to all of the articles included (either as a joint or sole first author), the nature of which is detailed in the title page leading each chapter. The different chapters are sorted not by chronological order but rather by theme, with descriptive studies taking precedence over culture-dependent analyses and methodological approaches, which are presented at the end. The first two chapters consist of a general introduction (Chapter 1) followed by a list of all authored scientific articles published during the length of this doctorate (Chapter 2). The first part of the thesis (Chapter 3, Chapter 4, Chapter 5 and Chapter 6) consists of studies that describe and characterize the taxonomic structure or functional potential of the plant microbiota across various hosts using culture-independent community profiling techniques (i.e. marker-gene amplicon or shotgun-metagenomics). The second part (Chapter 7) focuses on culture-dependent approaches to study of the plant microbiome, including the characterization of a large collection of cultured bacterial microbiota members and a comparative analysis of their sequenced genomes. The third part of this thesis (Chapter 8) analyzes the data obtained from a targeted, large-scale isolation and sequencing effort focused on the Rhizobiales order, a taxonomic group of particular interest which includes members that are capable of engaging in beneficial symbiotic interactions and improve host plant growth and ecological performance. Lastly, the forth part (Chapter 9) presents a novel computational method designed for the inference of functional interactions between genes and the analyses of functional modules that might be relevant for phenotypic traits of interest such as microbe-host symbiotic interactions or pathogen antibiotic resistance.

Each article has been adapted to fit the formatting requirements for publication in this thesis but the content (main text, figures and tables) has not been significantly altered.

Furthermore, the published articles extracted from the respective scientific journals (when applicable) are provided in chronological order of publication as an Appendix. Supplementary materials such as raw and intermediate data, supplementary figures and tables as well as the scripts necessary to reproduce each of the figures and statistical tests presented in the published articles can be accessed from links provided at the end of each chapter and are also provided along with the original journal publications.

1.3 The plant microbiota

The surface and interior of the plant leaf and root organs are colonized by bacterial communities, collectively referred to as the plant microbiota, which are characterized by a remarkably robust taxonomic structure and consist chiefly of members of the Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes phyla. These microbial communities (whose composition recent evidence suggests also extends to other kingdoms of life, such as fungi and protists) are believed to provide a number of functions to the host, including defense against pathogens, nutrient acquisition and enhanced tolerance to abiotic stresses. Thus, the plant microbiota can be considered as an additional plant trait that affects its ecological performance, raising the possibility of host-microbial community co-adaptation. In this section, a brief introduction to key findings concerning the plant leaf and root microbiota is presented.

1.3.1 Compartment-specific microbial assemblages

The microbial communities that inhabit soil are among the most complex and diverse found in the biosphere (Schloss and Handelsman, 2006). They constitute the start inoculum of the root microbiota (Lundberg *et al.*, 2012; Bulgarelli *et al.*, 2012) and, to a lesser extent, of a portion of the leaf microbial community (Bai *et al.*, 2015; Zarraonaindia *et al.*, 2015; Wagner *et al.*, 2016). Despite of a partially shared source of microbial diversity, there are clear differences between the communities associated to these two plant organs, with distinguishable contrasting taxonomic profiles (Bodenhausen *et al.*, 2013; Bai *et al.*, 2015) that account for the most significant source of variation within the plant microbiota. In the case of roots it is possible to define four additional and distinct niches along a gradient of decreasing complexity. From soil to the root interior, these are the soil, rhizosphere, rhizoplane, and endosphere compartments (Hacquard *et al.*, 2015). The rhizosphere corresponds to the portion of soil that is influenced by the action of root exudates, the rhizoplane is constituted by microbiota members that colonize the surface of the organ, and the endospheric compartment consists of microbes that inhabit the root interior. (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Edwards *et al.*, 2015). In the case of legumes, which are generally capable of engaging in symbiotic relationships with nitrogen-fixing bacteria this distinction can be extended to the nodule compartment, which corresponds to microbes capable of colonizing the interior of these specialized organs and are largely dominated by compatible symbionts (Zgadzaj *et al.*, 2016). These niches can be clearly separated in terms of their bacterial taxonomic profiles. The extent of the variation, however, changes with respect to the soil and host species. For example, whereas some hosts such as barley have a rhizosphere which is clearly separated from soil (Bulgarelli *et al.*, 2015), in others, like *Arabidopsis*, these two environments are barely distinguishable. The decreasing gradient in microbial diversity from soil to the root interior strongly suggests a sequential differentiation process, by virtue of which a subset of the microbial species present in one compartment is selected for and colonizes the subsequent niche (Bulgarelli *et al.*, 2013). However, recent evidence obtained from culture-independent community profiling of the model legume *Lotus japonicus* indicates that selection of microbial taxa that colonize the highly restrictive environment of the root nodule occurs in parallel to selection of the rhizoplane and endosphere compartments from the rhizosphere (see Chapter 6). This finding suggests that a consecutive selection process in which each compartment is assembled from a subset of the taxa present in the previous one is likely to explain only part of the process of root microbiota stablishment, and that parallel as well as sequential assembly of compartment-specific communities takes place in a concerted manner. Of note, despite of the abundant data that allows us to characterize these distinct root-associated communities, it is currently not known to what extent each compartment represents a heterogeneous rather than homogeneous environment, and how spatial distribution of community members further subdivides each niche into multiple micro-habitats.

A synthesis of recent literature indicates that factors driving the differentiation from the soil community into the rhizosphere, rhizoplane and endosphere compartments include the following four broad categories: i) soil chemical and structural properties, ii) soil biome start innoculum characteristics, which determines a complex network of microbe-microbe interactions, iii) root exudates and cell wall features and iv) host-genotype mediated control, e.g. by means of the host innate immunity or the legume-rhizobia symbiosis pathway. Although the underlying principles governing how these variables determine community stablishment are poorly understood, recent studies that employ plant genetics have begun to dissect the effect of the host in community differentiation (Lebeis *et al.*, 2015; Zgadzaj *et al.*, 2016).

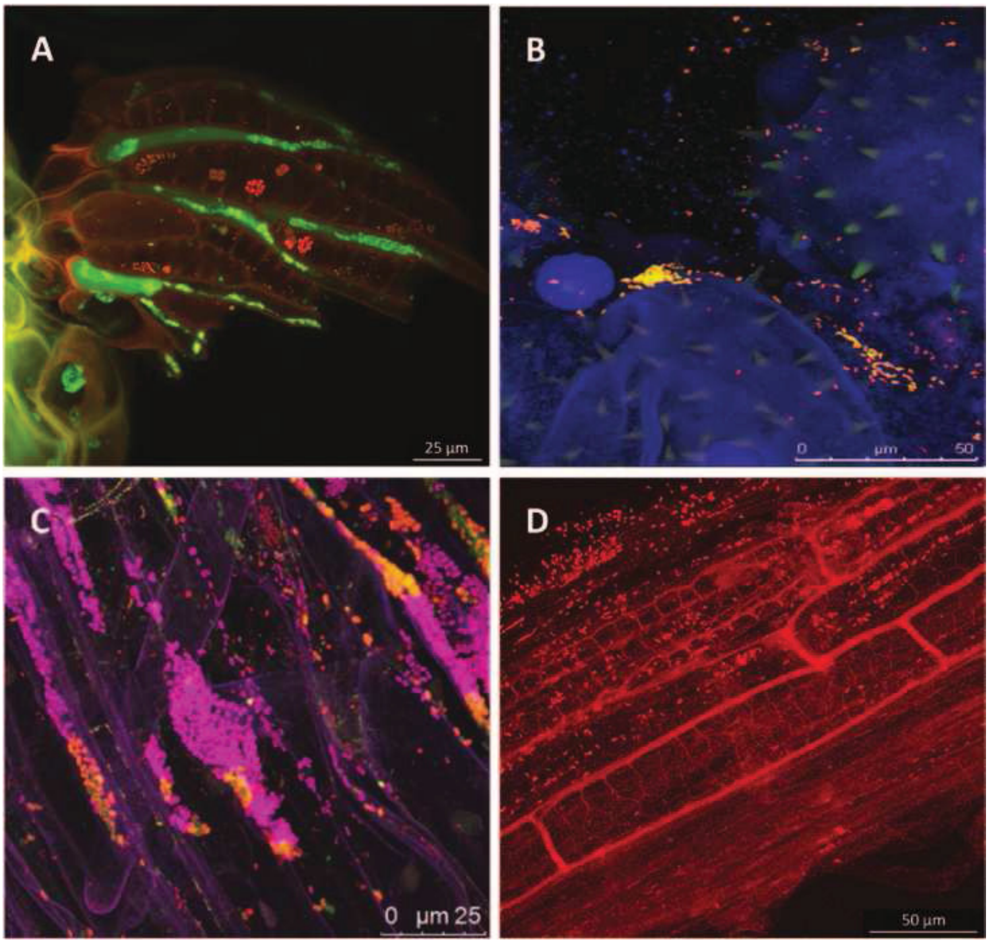


Figure 1.1: Plant microbiota members visualized by fluorescence in situ hybridization. (A-C) and confocal laser scanning microscopy. (A) Phyllosphere of a *Sphagnum* leave, (B) bacteria on pumpkin pollen, (C) bacteria in the rhizosphere of lettuce, and (D) root of an oilseed rape inoculated with the DsRed-labelled biocontrol agent *Pseudomonas trivialis* 3Re2-7. Adapted from Berg *et al.* (2015)

1.3.2 The robust structure of the root microbiota

The process of microbiota acquisition takes place at very early stages of host development. Time course profiling of rice root-associated compartments indicates that establishment begins 24h after germination and that within 2 weeks the process is already completed (Edwards *et al.*, 2015). Successive changes occurring throughout the plant life-cycle, e.g. during flowering, appear to have no influence in the taxonomic composition of the root community (Lundberg *et al.*, 2012; Dombrowski *et al.*, 2016). However, time course data obtained from long-lived perennials that captures seasonal dynamics over a period of years has not so far been obtained and our knowledge of long term variation remains limited. Comparative analyses of diversity across various host species demonstrate the presence of shared features that can be robustly identified in the root microbiota irrespective of the soil and host genotype (Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Hacquard *et al.*, 2015; Zgadzaj *et al.*, 2016). At a broad taxonomic rank, these communities are defined by members of a relatively small number of phyla, which includes Proteobacteria, Actinobacteria, Bacteroidetes and Firmicutes. Within these phyla there are several lower taxonomic groups that are identified as 'core' members of the root microbiota due to their consistent enrichment with respect to unplanted soil controls. These are the bacterial orders of Actinomycetales, Burkholderiales, Flavobacteriales and Rhizobiales (Schlaeppi *et al.*, 2014; Hacquard *et al.*, 2015). An extensive record in the literature concerning isolates from these bacterial clades that present plant growth promoting capabilities (Manter *et al.*, 2010; Johansen *et al.*, 2009; Kolton *et al.*, 2012) antagonistic activity toward soil-borne fungal and oomycete pathogens (Benitez and Gardener, 2009) or nutrient mobilization (Schmalenberger *et al.*, 2008; Yoshimoto *et al.*, 2002) indicates that the consistent and robust enrichment of these core community members might provide the host with beneficial functions. Fine-tuned recruitment of particular members within these bacterial orders shows considerable variation between plant species and even genotypes, e.g. in the case of Rhizobiales (see Chapter 8), and thus constitutes a host trait of ecological relevance whose genetic basis is the focus of ongoing research.

In spite of this robust and conserved taxonomic structure, multiple questions concerning microbiota stablishment and persistence remain unanswered. In particular, little is known about the necessary features that are required for pioneer microbes to stablish mature communities and to what extent succession of these early colonizers plays a significant role. Furthermore, the lack of time course data covering seasonal dynamics of perennial plants precludes us from drawing any conclusions concerning long-term dynamics of the plant root microbiota. Similarly, how plants influence the microbial composition of soil as well as successive generations of plants grown in the same area (often from multiple distinct species), a process which might have important practical repercussions, e.g. for the practice of crop rotation or for the development of efficient bio-fertilizers, is currently not known.

1.3.3 Factors driving community diversity

Modeling community structure (based on diversity estimates such as Bray-Curtis dissimilarity or UniFrac distances) as a function of environmental variables allows us to determine the proportion of the overall variance of the data that can be attributed to each factor (Peiffer *et al.*, 2013; Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Wagner *et al.*, 2016; Dombrowski *et al.*, 2016; Zgadzaj *et al.*, 2016). A synthesis of the available data suggests that the major sources of variation in community structure are, in order of decreasing importance: host organ, followed by compartment, host species, soil / geographic location and host genotype. Whereas the former factors typically explain between 10-30% of the variance and their effects can be robustly measured, the effect of the host genotype is much smaller and ranges between 5-10% (see Chapter 4, Chapter 5 and Chapter 6). These observed differences in community composition are related to various modulating factors, which include environmental features, such as soil physical and chemical characteristics or climate, microbe-microbe interactions (e.g. soil biota composition) as well as host-microbe interactions (e.g. the host innate immune system or the legume symbiosis pathway).

The environmental factors contributing to microbiota variation are estimated to explain approximately 20-30% of the variance of the data when correcting for technical factors (Peiffer *et al.*, 2013; Schlaeppi *et al.*, 2014; Dombrowski *et al.*, 2016; Wagner *et al.*, 2016). However, these variables are generally not fully independent and it is exceedingly difficult to disentangle their effects and interactions. For instance, sampling of plants grown in natural sites confounds the effect of the climate with the soil characteristics, which are often modeled together as the same variable ('site') (Wagner *et al.*, 2016). Perhaps more importantly, profiling of natural communities either in the wild or in the greenhouse under controlled conditions makes it impossible to distinguish between the soil physical and chemical characteristics (such as pH, nutrient availability, etc.) and the features of its endemic microbiota. These caveats impose a fundamental limit to our ability to extract causal relationships between individual environmental factors using culture-dependent approaches to microbial community analysis.

Microbes that successfully colonize plant roots and leaves and assemble into communities form complex webs of interactions. These interactions are not limited to associations of individual microbial species with their plant host but also include interactions between microbes, likely even across different microbial kingdoms of life (Agler *et al.*, 2016). Analysis of co-occurrence and co-exclusion patterns in microbial communities, including but not limited to those associated to eukaryotic hosts, reveal the importance of microbe-microbe in structuring and maintaining stability (Edwards *et al.*, 2015; Zhang *et al.*, 2014; Heijden and Hartmann, 2016; Faust *et al.*, 2012; Coyte *et al.*, 2015; Lima-Mendez *et al.*, 2015). Furthermore, analysis of (meta)genomic sequences obtained from the human gut (Qin *et al.*, 2010) and the plant root and rhizosphere (Ofek-Lalzar *et al.*, 2014; Bulgarelli *et al.*, 2015) suggest that genes relevant for competition and interaction with other microbes (such as secretion systems) are important for adaptation to these host-associated environments. Additionally, the identification of 'hub' microbes that occupy a central position in association networks derived from community data (Agler *et al.*, 2016) suggest that a few members may play a pivotal role in the overall structure of the community by directly interacting with other microbes. Estimates of diversity between genomes of the same host species indicates that host genotype has a much smaller influence in community variation compared to host organ (root or leave), root fraction (soil, rhizosphere, rhizoplane or endosphere) and smaller than soil type (or natural site). Variance deconvolution of beta-diversity estimates between maize genotypes (Peiffer *et al.*, 2013), 6 rice cultivars (Edwards *et al.*, 2015), 3 barley accession representing different stages along the domestication process of this crop (Bulgarelli *et al.*, 2015), and several *Arabidopsis thaliana* accessions and ecotypes (Schlaeppi *et al.*, 2014) reveals that only ∼5% of the variation in community structure can be attributed to host genotype. Further exploration of this genotype contribution to microbiota diversity requires an experimental setup with higher replicability and lower technical noise. At the same time, the large functional variation observed between exemplars within the same species (even between strains with identical 16S rRNA sequences; see Chapter 7 and Chapter 8), indicates that even small differences in community taxonomic structure may correspond to large variation in terms of function. An alternative approach designed to address this question consists on deep shotgun metagenome sequencing of the plant microbiome, which has the potential to provide high-resolution taxonomic profiles as well as functional data (Bulgarelli *et al.*, 2015; Mendes *et al.*, 2014; Ofek-Lalzar *et al.*, 2014).

1.3.4 Host-microbiota co-evolution

The concept of heritability is related to the amount of phenotypic variation of a particular trait can be explained by naturally occurring genetic variation within a population. In the case of the plant microbiota, variations can be measured using the taxonomic profiles of communities associated to genotypes of the same host species. By asking the question of whether beta-diversity estimates among individuals within a genotype are significantly lower than between individuals of different genotypes we are able to assess the level of heritability of the microbiota. Comparison of rhizosphere community structure of 27 maize inbred lines points to a very low but statistically significant heritability (Peiffer *et al.*, 2013). A similar study conducted on three barley accessions, including a wild-type genotype, a variety used by subsistence farmers and a modern cultivar, which represent three stages along a domestication spectrum identified a slightly larger heritable component in both the root (rhizoplane and endophitic compartment) and rhizosphere communities (see Chapter 5). This lack of a strong correlation between the genetic resemblance of the genotypes (kinship matrix) and the microbiota diversity profiles might indicate that the effect of the host genes are indirect and interact with other (environmental) factors. A recent study of the diversification of the root and leaf microbiota associated to a wild perennial *Arabidopsis* relative within the Brassicaceae family (*Boehera stricta*) that sampled individuals from different ecotypes grown in natural sites revealed that the correlation between the kinship matrix and community variation is site-dependent (Wagner *et al.*, 2016). This finding supports the hypothesis that heritability of the microbiota can be mediated by other factors, such as features of the soil endemic community and that host-genotype fine-tuning is restricted to certain environments. Large-scale field studies, where ecotypes of the same host species are reciprocally transplanted across natural habitats, or grown under controlled laboratory conditions, e.g. using collected soils and climate chambers, have the potential to uncover some of these interactions but remain challenging due to the potentially large number of genotypes and samples required to attain enough statistical power.

If heritability is sustained over sufficiently long periods of time, it becomes possible to observe a significant correlation between patterns of inter-species community structure variation and the host phylogeny. In the presence of interactions between the host and its community members, changes in each party imposes selective pressures on the other, and adaptation occurs in a concerted manner, in a process known as co-evolution. This pattern has for example been observed in primates and their gut microbiota, where the structure of the host species tree was found to be concurrent with patterns of microbiota diversification (Ochman *et al.*, 2010). A network analysis of amplicon data comparing 60 different species of mammals indicated that similar patterns can be also observed for taxonomically diverse groups of hosts (Ley *et al.*, 2008a). A previous attempt to contrast inter-species host phylogeny for a small subset of Bassicaceae plant species, including *Arabidopsis thaliana*, that share a common ancestor approximately 35 My ago revealed incongruences in the patterns of microbiota diversity that do not match the species tree, perhaps caused by the existence of ecological adaptations or due to the small number of samples species (see Chapter 4). To date, there is no direct and conclusive evidence of co-adaptation between the plant host and its microbiota at the whole community level. Surprisingly, a systematic meta-analysis of co-evolution that combines the large amounts of profiling data generated over the previous 5 years has not been conducted. It is important to note that the lack of heritability at the whole community level does not exclude the possibility of co-diversification between the plant and individual community members or microbial consortia. This is might be the case, e.g., for legumes and their nitrogen-fixing rhizobial symbionts, whose interactions are highly specific and restricted only to certain compatible species. This illustrates the importance of not restricting the analysis of co-evolution to the whole community level but also explore the possibility of heritability and co-adaptation between the host and individual microbiota members.

1.4 Computational methods in microbiome studies

Advances in large-scale sequencing of environmental samples have enabled researchers to ask questions regarding the taxonomic composition (SSU rRNA genes and ITS amplicon data), the functional potential (shotgun metagenomics) and the activity (metatranscriptopmics) of microbial communities, including those associated with the roots and leaves of healthy plants. The task of integrating these potentially very large datasets and extracting meaningful biological insights from them has seen in recent years the development years of numerous computational tools and pipelines and remains a vibrant area of research.

1.4.1 Marker gene amplicon data

Owing to the widespread use of bacterial 16S rRNA and fungal ITS amplicon sequencing, bioinformatic pipelines designed for analyzing marker gene data have been critical to advance our understanding of the diversity of microbes associated with plants grown in natural environments. Commonly used toolkits for amplicon data include Mothur (Schloss *et al.*, 2009) and QIIME (Caporaso *et al.*, 2010) and allow the pre-processing of sequencing data (de-noising and error correction, merging of paired-end reads, demultiplexing, etc.) as well as calculating diversity estimates. Analysis of marker gene data often requires the use of large collections of reference data in order to perform taxonomic classification of representative sequences. Among them, the most widely used databases are Greengenes (DeSantis *et al.*, 2006) the Ribosomal Database Project (Cole *et al.*, 2009), Silva, which includes 16S and also eukaryotic 18S sequences (Pruesse *et al.*, 2007) and the fungal and oomycete ITS UNITE database (Koljalg *et al.*, 2005; Abarenkov *et al.*, 2010). Despite of their fast growth and regular updates, these databases contain 'blind spots' that make them inadequate for the analysis of certain taxonomic groups of microbes, such as viruses, protists or fungi, and taxonomic classification of marker gene fragments at low ranks (genus or species) remains highly unreliable.

An important step during the processing of amplicon data is the grouping of sequences estimated to originate from the same microbial species into Operational Taxonomic Units (OTUs), generally using a fixed threshold of sequence similarity (e.g. 97% for bacteria and archaea). There are three main strategies for OTU inference: those that do not depend on a reference database of sequences (*de novo* OTU clustering), such as UCLUST (Edgar, 2010) or UPARSE (Edgar, 2013), reference-based methods (Edgar, 2010), and hybrid approaches, which typically perform a first pass reference iteration, followed by de-novo clustering of left-out sequences, such as SortMeRNA (Kopylova *et al.*, 2012). The OTU clustering step is necessary to reduce the amount of sequences to be processed to a manageable quantity by picking representatives of each OTUs in order to allow downstream analyses of diversity and taxonomic classification. However, this approach imposes severe limitations to amplicon-based studies, most importantly a resolution limit (typically 3% of sequence identity), the inclusion of data artifacts (OTUs that consist exclusively of PCR amplification or sequencing errors) and the need to base further diversity analyses on the assumption that OTUs are functionally and ecologically homogeneous units, despite of abundant contrary evidence, e.g. in the case of plant-associated bacteria (Bai *et al.*, 2015). Unlike culture-independent profiling of natural communities, where it is exceedingly difficult to differentiate between real and erroneous sequences, experiments with synthetic communities have the advantage of a simplified setup that can vastly reduce the complexity associated with processing of individual raw reads without the need to cluster them into taxonomic units. Unfortunately, current toolkits and computational pipelines are not designed to take advantage of this simplified experimental setup.

1.4.2 Diversity assessment and statistical analyses

Downstream analyses of abundance data typically include the inference of ecological networks, generally based on co-occurrence of OTUs across samples (Faust *et al.*, 2012), calculation of alpha-diversity (within sample) estimates such as the Shannon, Chao or Phylogenetic Diversity (PD) indices and beta-diversity (between samples), e.g. Bray-

Curtis or UniFrac (Lozupone *et al.*, 2011) distances. A common step to explore diversity estimates consists on performing a dimensionality reduction step in order to compare groups of samples from diverse environments (Analysis of Principal Coordinates, Non-metric Multidimensional Scaling, perMANOVA, ANOSIM, etc.) or to assess the contribution of individual environmental factors by variance deconvolution (e.g., Linear Mixed Models, Canonical Correspondence Analyses). Another important step in the analysis involves testing of differentially abundant OTUs between varying conditions, which usually require statistical tests designed for count data. A variety of toolkits and libraries have been developed for this purpose, such as vegan (Oksanen *et al.*, 2015), phyloseq (McMurdie and Holmes, 2013) or DESeq (Anders *et al.*, 2013) and their applicability for hypothesis testing in different experimental setups has been explored using simulated data (McMurdie and Holmes, 2014).

1.4.3 Shotgun metagenomics

An alternative to sequencing marker genes to explore the taxonomic diversity of a microbial community is to sequence all genomes contained in an environment, an approach known as shotgun metagenomics. Unlike marker gene metagenomics, shotgun metagenomics has the advantage of providing insights into the functions as well as the taxonomy of a microbial assemblage, thus moving beyond a mere catalogue of species present in an environment and into a mechanistic view of a community. However, the complexity of the computational analyses required for shotgun metagenome data far exceeds that of amplicon-based studies, especially for diverse communities such as those associated to plants for which *de novo* metagenome assembly difficult and where only a small percentage of protein-coding genes can be reliably annotated [∼41% of predicted reading frames; (Delmotte, 2009; Bulgarelli *et al.*, 2015; Zarraonaindia *et al.*, 2015; Ofek-Lalzar *et al.*, 2014)].

There are three important and challenging steps required prior to data interpretation, which are tackled by a variety of computational tools: 1) shotgun metagenome assembly, performed by standard assemblers such as SOAP (Li, 2010; Luo *et al.*, 2012) or specially designed for environmental data, like MetaVelvet (Namiki *et al.*, 2012), Meta-IDBA (Peng *et al.*, 2011), Ray Meta (Boisvert *et al.*, 2012) or Snowball (Gregor *et al.*, 2016), 2) binning of sequence fragments [e.g. MEGAN (Huson *et al.*, 2007, 2016), PhyloPytiaS+ (Gregor *et al.*, 2016), taxator-tk (Droege *et al.*, 2015), Kraken (Wood and Salzberg, 2014)], and 3) annotation and classification of metagenomes, which typically consists of prediction of open-reading frames [MetaGeneMark (Zhu *et al.*, 2010), PRODIGAL (Hyatt *et al.*, 2010)] followed by homology searches against annotation databases, such as KEGG (Kanehisa *et al.*, 2016), SEED (Overbeek, 2005), PFAM (Finn *et al.*, 2016) or eggNOG (Powell *et al.*, 2014). Alternative methods seek to estimate taxonomic abundances from shotgun metagenome data by generating clade-specific sets of marker genes (Segata *et al.*, 2012), an approach that can potentially be adapted to be used on synthetic communities to allow intra-species and strain level resolution, provided that a sufficiently high ratio of microbial to plant reads can be obtained.

1.4.4 Phylogenomic analyses of microbiota members

Phylogenomics refers to the reconstruction of relationships between organisms based on their genomic sequences and of the study within this context of gene function and genome evolution. Most commonly, the inference of a phylogeny or species tree is required as the basis of a wide variety of analysis, generally with the goal of understanding a biological process by reconstructing its evolutionary history. This approach constitutes one of the most versatile and commonly used computational methods for the analysis of sequence data, but suffers from several caveats that limit their applicability. First, some of the steps typically involve pairwise sequence comparisons between all genomes, for example during the process of determining homology relationships between gene-coding genes (inference of orthologous groups or gene families) (Li *et al.*, 2003; Sonnhammer and Ostlund, 2015; Emms and Kelly, 2015), which scale very poorly (generally quadratic cost) with the number of organisms included in the study. This imposes a limit to the size of the dataset with which high-quality *de novo* inference of homology can be taken advantage (see Chapter 8 and Chapter 9) that is effectively on the order of the hundreds of genomes. Second, sampling biases (e.g. non-culturable organisms, experimental preference in favor of model organisms or clinical sources, etc.) often lead to unbalanced dataset that contain uncharted gaps and do not capture the true diversity present in a community or a population and may lead to incorrect phylogenetic inferences (Kumar *et al.*, 2012). Finally, the predominantly clonal method of microbial reproduction prevents mutations affected by selection to be found in multiple genetic backgrounds, as it is the case e.g. in plants or vertebrates, and makes it difficult to distinguish between the effect of a mutation and its homogeneous genetic background (Falush and Bowden, 2006; Chen and Shapiro, 2015). Phylogenomic analysis of traits of interest in datasets with strong population structure are at risk of detecting a large number of false positives and fail to pinpoint true causal genetic determinants (Earle *et al.*, 2016). One alternative that is robust with respect to the effects of relatedness consists on focusing not on the genomic features present in the extant genomes, which is affected by clonal structure, but rather in their changes during their evolutionary histories, (see Chapter 9 and Chapter 8).

In the case of microbiome studies, it is of additional interest to study the evolutionary history of a microbial community (or their integrating members) in the context of the host phylogeny. The goal of this analysis is to identify if and how adaptation occurred in a concerted manner between the host and its associated microbes. Testing this hypothesis is not possible by assessing the extant microbial taxonomic or genomic diversity between a set of host species, as observing significant differences does not allow distinguishing between niche adaptation and true co-evolution Furthermore, although significant correlation between community member phylogenies and the host species tree can be taken as an evidence of co-diversification, it does not provide insight into the mechanistic or genetic basis of the interactions. Systematic isolation and sequencing of community members from taxonomically unrelated hosts could allow us to perform whole-genome comparative analyses of adapted microbiota members and has the potential to provide further insights into the process of co-evolution.

1.5 Outlook

Despite of substantial progress, much is still unknown about the processes involved in plant-microbiota interactions. Moving past a purely descriptive analysis consisting of correlations and observational data alone towards an experimental framework where specific hypothesis can be tested and models generated from sequence data falsified or validated, remains a great challenge for microbiome studies. In particular, the work here presented (see Chapter 7) constitutes some of the earlier steps taken in this direction. The generation of culture collections of isolates, together with their associated libraries of sequenced genomes and their use in synthetic community reconstitution experiments provide the possibility to engineer communities guided by models inferred from culture-dependent data and test outcomes in terms of output community structure or plant macroscopic phenotypes, such as growth promotion. Efforts to extend this experimental approach to other plant hosts (e.g. *Lotus japonicus*) and soil types (e.g. nutrient-poor calcareous soils) are currently ongoing. These culture collections will constitute an unprecedented resource to perform comparative experiments and analyses in order to extract basic guiding principles behind microbiota stablishment conserved across hosts and conditions.

One of the main limitations of the use amplicon data to microbial ecology lays on the fact that sequencing errors and PCR artifacts impose a limit in the resolution that can be achieved, typically up to 97% sequence identity over several hundred basepairs of conserved markers such as 16S rRNA. Given the large genomic and phenotypic variation observed within the same taxonomic unit (see e.g. Chapter 8 and Chapter 7), this lack of resolution constitutes an important caveat in interpreting amplicon data. Whereas de-noising and chimera removal are steps generally performed *prior* community data analyses in most microbiome studies, there is the possibility for further improvement, particularly in the case of reduced-complexity synthetic community experiments with germ-free hosts for which accurate reference sequences are available. Furthermore, applying phylogenomics to error correction within the 3% sequence identity threshold typically used, i.e. by placing new sequences in a reference tree and discarding errors that deviate substantially from what a model of DNA evolution would predict, has the potential to improve accuracy and resolution.

Another promising avenue of research consists of taking advantage of the large quantities of available data by integrating it into large-scale meta-analyses (see Chapter 3 for a proof of principle). In particular, inference of networks of ecologically meaningful interactions (e.g. from co-occurrence matrices) in a host phylogenetic framework could allow researchers to analyze specific microbiota features (such as the presence of hub organisms and their impact) by applying the comparative method across plant hosts. This work would require a fundamentally new approach to traditional network-enabled microbial ecology analyses and constitutes an exciting challenge for computational and experimental plant microbiota studies.

In Chapter 9, a novel approach designed to extract biologically meaningful links between protein families was presented. These results, however, underline the finding that a large proportion of the predicted coding sequences in bacterial genomes are poorly annotated (see e.g. Chapter 8). When the number of genomes included in a comparative analysis or association study increases over a certain number (in the order of hundreds of genomes), inference of orthology relationships based on sequence alone is no longer feasible and homology-based approaches using models from databases of annotated sequences is required. This greatly limits our ability to discover the true genetic basis of traits of interest or meaningful gene family clusters to previously annotated sequences while ignoring previously uncharacterized genes. A possible extension of the computational framework which is presented in the last chapter of this thesis (clustering of functionally related genes using evolutionary profiles) consists on moving beyond presence or absence of inferred orthologous groups and using more generic genomic features instead, such as k-mers of arbitrary size, that do not rely on annotation and can capture other sequence variants. The major obstacle facing this methodological update is the need to handle potentially vast matrices of features, a task that may prove to be computationally intractable despite of being easily parallelized.

CHAPTER 2

Personal bibliography

As first author

- Bulgarelli, D., R. Garrido-Oter, P. Muench, A. Weiman, J. Droege *et al.* (2015). Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley. *Cell Host & Microbe*, **17** (3): 392–403.

  Included in this thesis (Chapter 5).

- Hacquard, S., R. Garrido-Oter, A. Gonzlez, S. Spaepen, G. Ackermann *et al.* (2015). Microbiota and Host Nutrition across Plant and Animal Kingdoms. *Cell Host & Microbe*, **17** (5): 603–616.

  Included in this thesis (Chapter 3).

- Bai, Y., D. B. Mueller, G. Srinivas, R. Garrido-Oter, E. Potthoff *et al.* (2015). Functional overlap of the Arabidopsis leaf and root microbiota. *Nature*, **528** (7582): 364–369.

  Included in this thesis (Chapter 7).

- Zgadzaj, R., R. Garrido-Oter, D. B. Jensen, A. Koprivova, P. Schulze-Lefert *et al.* (2016). Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. *Proceedings of the National Academy of Sciences*, **113** (49): E7996–E8005.

  Included in this thesis (Chapter 6).

- Garrido-Oter, R., T. Nakano, N. Dombrowski, A. C. McHardy and P. Schulze-Lefert (2016). Assessment of functional diversification and adaptation in rhizobia by comparative genomics. (In preparation).

  Included in this thesis (Chapter 8).

- Garrido-Oter, R. and A. C. McHardy (2016). Clustering of functionally related genes reveals novel symbiosis-relevant genes in rhizobia. (In preparation).

  Included in this thesis (Chapter 9).

As contributing author

- Schlaeppi, K., N. Dombrowski, R. Garrido-Oter, E. V. L. v. Themaat and P. Schulze-Lefert (2014). Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. *Proceedings of the National Academy of Sciences*, **111** (2): 585–592.

  Included in this thesis (Chapter 4).

- Hacquard, S., B. Kracher, K. Hiruma, P. C. Muench, R. Garrido-Oter *et al.* (2016). Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nature Communications*, **7**: 11362.

- Dombrowski, N., K. Schlaeppi, M. T. Agler, S. Hacquard, E. Kemen *et al.* (2016). Root microbiota dynamics of perennial Arabis alpina are dependent on soil residence time but independent of flowering time. *The ISME Journal*.

Part I

The structure of the plant root microbiota

CHAPTER 3

Microbiota and host nutrition across plant and animal kingdoms

Status	**Published**
Journal	*Cell Host & Microbe* (Impact factor 12.328)
Citation	Hacquard, S. *, Garrido-Oter, R. *, González, A. *, Spaepen, S. *, Ackermann, G., Lebeis, S., McHardy, A.C. †, Dangl, J.L. †, Knight, R. †, Ley, R. †, Schulze-Lefert, P. † (2015). Microbiota and Host Nutrition across Plant and Animal Kingdoms. *Cell Host & Microbe*, **17**, 603-616. * joint first authors; † joint corresponding authors
URL	http://dx.doi.org/10.1016/j.chom.2015.04.009
Own contribution	Designed research (with co-authors) Performed the experiments (with co-authors) Analyzed the data (with co-authors) Interpreted the data (with co-authors) Wrote the manuscript (with co-authors)

3.1 Abstract

Plants and animals each have evolved specialized organs dedicated to nutrient acquisition, and these harbor specific bacterial communities that extend the host's metabolic repertoire. Similar forces driving microbial community establishment in the gut and plant roots include diet/soil-type, host genotype, and immune system as well as microbe-microbe interactions. Here we show that there is no overlap of abundant bacterial taxa between the microbiotas of the mammalian gut and plant roots, whereas taxa overlap does exist between fish gut and plant root communities. A comparison of root and gut microbiota composition in multiple host species belonging to the same evolutionary lineage reveals host phylogenetic signals in both eukaryotic kingdoms. The reasons underlying striking differences in microbiota composition in independently evolved, yet functionally related, organs in plants and animals remain unclear but might include differences in start inoculum and niche-specific factors such as oxygen levels, temperature, pH, and organic carbon availability.

3.2 Physiological functions of the vertebrate gut and plant roots

The vertebrate gut and plant roots evolved independently in animal and plant kingdoms but serve a similar primary physiological function in nutrient uptake (Figure 3.1). One major difference between plant and animal nutritional modes is their distinct energy production strategy. Plants are autotrophs, producing their own energy through photosynthesis (carbohydrate photo-assimilates), while animals rely entirely on the energy originally captured by other living organisms (heterotrophs). Long-distance transport mechanisms ensure the distribution of carbohydrate photo-assimilates from chloroplasts in leaves to all other body parts, including roots. Nutrient acquisition by roots to support plant growth is therefore almost exclusively limited to uptake of mineral ions and water from soil. In contrast, the mammalian gut has evolved to facilitate the uptake of simple sugars, amino acids, lipids, and vitamins in addition to ions. It is typically compartmentalized into sections with low microbial biomass in which the products of host enzymatic activity are absorbed (i.e., the human small intestine, SI) and a section for the uptake of microbe-derived fermentation products (human large intestine or hindgut, LI).

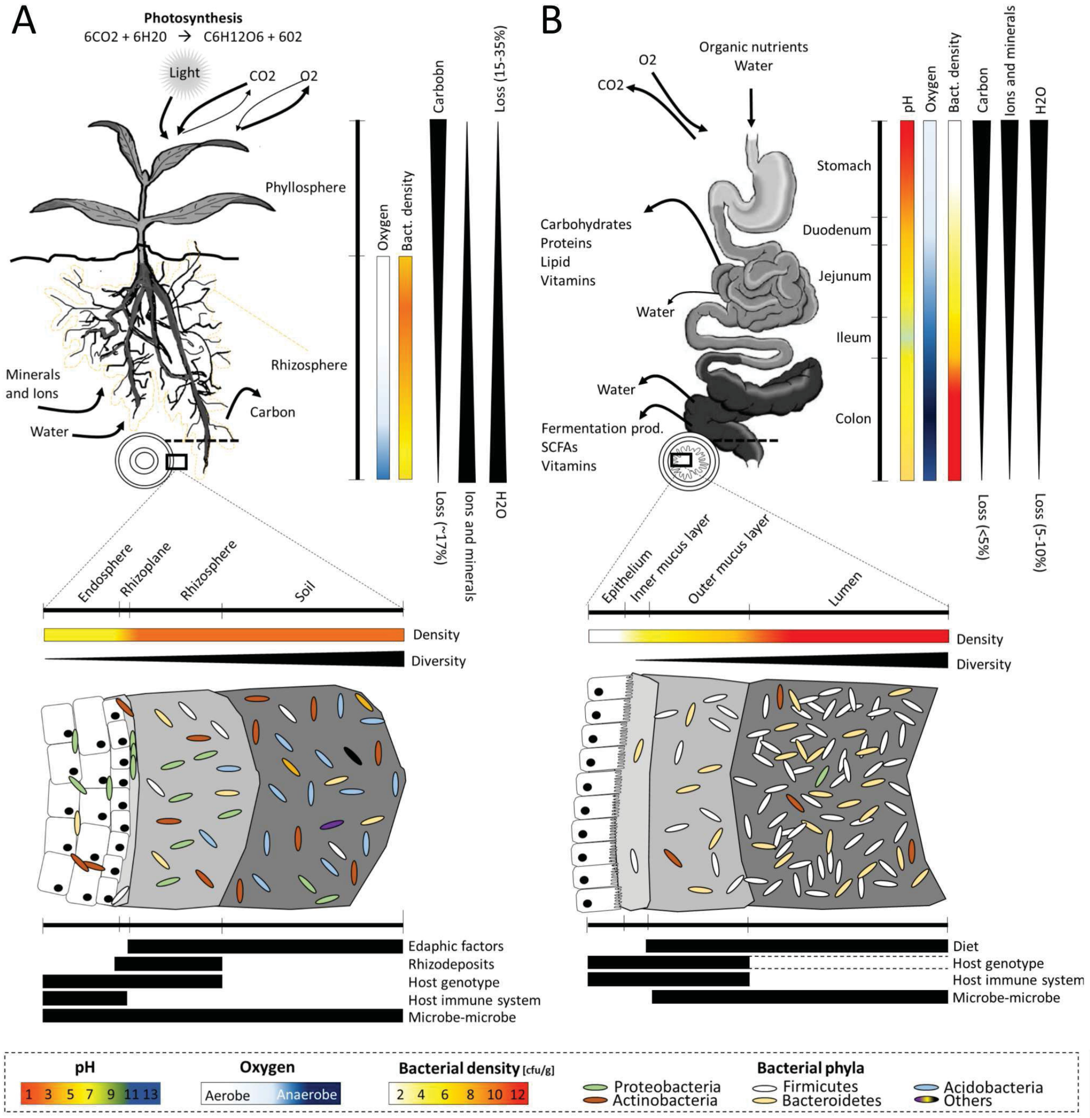


Figure 3.1: Physiological functions of the plant roots and human gut in nutrient uptake, spatial aspects of microbiota composition, and factors driving community establishment. (A and B) Spatial compartmentalization of the plant root microbiota (A) and the human gut microbiota (B). Upper panels: the major nutrient fluxes are indicated, as well as pH and oxygen gradients in relation with the bacterial density. Lower panels: compartmentalization of the microbiota along the lumen-epithelium continuum in the gut or along the soil-endosphere continuum in the root. For each compartment, the bacterial density, the bacterial diversity, and the major represented phyla are represented for both the gut and the root organs. The main factors driving community establishment in these distinct compartments are depicted with black bars. The gut drawing is adapted from Tsabouri *et al.* (2014) with permission from the publisher.

A significant fraction of the soil nutritive complement and of the dietary intake remains unavailable for plants and animals, respectively, and this defines their dietary constraints. Critical nutrients for plant growth and productivity in soil are nitrogen and phosphorus. However, plant roots can absorb only inorganic nitrogen and orthophosphate (Pi), although phosphorus is abundant in soil both in inorganic and organic pools. Pi can be assimilated via low-Pi-inducible (high-affinity) and constitutive Pi uptake systems (low-affinity) (Lambers *et al.*, 2008; Lopez-Arredondo *et al.*, 2014). Plant species adapted to neutral or higher soil pH, and more aerobic soils have a preference for nitrate and deploy two nitrate uptake and transport systems that act in coordination. By contrast, plants adapted to low pH (reducing soil) as found in forests or the arctic tundra appear to assimilate ammonium or amino acids (Maathuis, 2009). Similarly, a fraction of normal human dietary intake remains undigested and therefore non-bioavailable (fiber). These non-digestible components include plant cell wall constituents such as cellulose, hemicellulose, xylan, and pectin, and certain polysaccharides such as β-glucan, inulin, and oligosaccharides that contain bonds that cannot be cleaved by mammalian hydrolytic enzymes (Tungland and Meyer, 2002).

Plant roots and animal guts are colonized by diverse microbial classes, including bacteria and archaea, fungi, oomycetes, as well as viruses (Table 3.1). These communities can be regarded as the host's extended genome, providing a huge range of potential functional capacities (Berendsen *et al.*, 2012; Gill *et al.*, 2006; Qin *et al.*, 2010; Turner *et al.*, 2013). Here we focus on bacterial microbiotas because these were shown to form reproducible taxonomic assemblies in animal and plant individuals with well-defined functions.

In plant roots, the microbiota mobilizes and provides nutrients by increasing nutrient bioavailability from soil (Bulgarelli *et al.*, 2013). Non-nutritional functions include increased host tolerance to biotic stresses, e.g., against soil-borne pathogens (Mendes *et al.*, 2011), and likely abiotic stresses. In addition, the root microbiota can also affect plant fitness by impacting flowering plasticity (Panke-Buisse *et al.*, 2015; Wagner *et al.*, 2014).

Similarly, the gut microbiota has a major role in host nutrition. It contributes nutrients and energy to the host via fermentation of indigestible polysaccharides into short-chain fatty acids (SCFAs) in the colon (Martins dos Santos *et al.*, 2010; Tremaroli and Baeckhed, 2012). The human LI has incomplete peristalsis and a longer retention time, allowing fermentative microbiota to break down complex glycan bonds and liberate additional energy from the diet (Stevens and Hume, 1998). Additionally, gut microbiota provide essential vitamins to the host and modulate the absorptive capacity of the intestinal epithelium. An additional common feature of the gut and root microbiota is their protective role by competitive exclusion against invasion by opportunistic pathogens (Kamada *et al.*, 2013).

	Cucumber (a)	Wheat (a)	Soybean (b)	Wheat (c)	Oat (c)	Pea (c)	Barley (d)	Gut (e)
Bacteria	99.36	99.45	96	88.5	77.3	73.7	94.04	99.1
Archaea	0.02	0.02	< 1	< 0.5	< 0.5	< 0.5	0.054	
Eukaryotes	0.54	0.48	3	3.3	16.6	20.7	5.90	< 0.1

Table 3.1: Percentage of shotgun metagenome reads assigned to each kingdom of life across metagenome studies.

a: Ofek-Lalzar *et al.* (2014)
b: Mendes *et al.* (2014)
c: Turner *et al.* (2013)
d: Bulgarelli *et al.* (2015)
e: Qin *et al.* (2010)

Homeostatic balance between both microbe-microbe and host-microbe interactions is critical for a healthy host-microbiota relationship. Alteration of this balance via perturbation of the gut or the plant microbiota composition (microbial dysbiosis) may represent an important mechanism of disease (Martins dos Santos *et al.*, 2010; Kemen, 2014; Sekirov *et al.*, 2010). In plants, a healthy status is the norm, and soil-resident microbes contribute to plant health. This is illustrated by a higher disease severity following pathogen inoculation when plants are grown in pasteurized compared to non-pasteurized soils (Weller *et al.*, 2002). In addition, so-called disease-suppressive soils protect plants against particular soil-borne pathogens. For example, specific bacterial genera belonging to gamma-Proteobacteria were associated with a high level of soil disease suppressiveness. The underlying mechanisms comprise competition between soil-borne microbes for plant-derived nutrients and antimicrobial compound production (Berendsen *et al.*, 2012; Mendes *et al.*, 2011). In the gut, commensal microbes can also suppress pathogen invasion through secretion of antimicrobial compounds, alteration of local pH, or stimulation of host immunity (Kamada *et al.*, 2013).

3.3 Compartmentalization of the gut and root microbiota

Relevant biotic and abiotic gradients exist in both the gut and root, leading to microbial compartmentalization (Figure 3.1). Along the soil-root continuum, four compartments can be distinguished: soil, rhizosphere, rhizoplane, and endosphere (Figure 3.1A). Bacterial diversity in soil is high, with estimates suggesting that >2,000 species populate 0.5 g of soil (Schloss and Handelsman, 2006). The rhizosphere corresponds to the zone of soil directly influenced by root exudation, while the root compartment can be separated in two distinct niches, rhizoplane and endosphere. The rhizoplane harbors a suite of microbes that tightly adhere to the root surface, while the endosphere is composed of microbes inhabiting the interior of roots. Microbial density is high in the rhizosphere, and species richness gradually decreases along the soil-endosphere continuum (Bulgarelli *et al.*, 2012, 2015; Edwards *et al.*, 2015; Lundberg *et al.*, 2012) (Figure 3.1A). Therefore, the bacterial community shifts from a dense and diverse soil-borne community to a host-adapted community with reduced diversity.

A spatial heterogeneity of microbial density exists along the digestive track (Stearns *et al.*, 2011). Densities are lowest in the stomach and duodenum (proximal SI) (101-103 bacteria per gram of content) and increase along the length of the SI with a higher density in the distal ileum (104-107 bacteria per gram). Cell densities in the LI can reach 1012-1013 bacteria per gram of content, representing the highest density recorded so far in any environment and exceeding the density detected in the rhizosphere by 2-3 orders of magnitude. Although the density is high, the diversity is relatively low (Stearns *et al.*, 2011; Walter and Ley, 2011). Using low-error 16S rRNA gene sequencing (LEA-seq) of the human fecal gut microbiota (low depth coverage), the number of bacterial species is estimated at 101 ± 27, which is in alignment with estimates of culture-based techniques (Faith *et al.*, 2013; Mitsuoka, 1992) Compartmentalization exists also from the inside to the outside of the intestinal tube, defined by the intestinal lumen, mucus, and epithelial surface. Similar to the compartmentalization in the root, a decrease in bacterial density is observed from the lumen to the epithelial surface (Swidsinski *et al.*, 2005; Abbeele *et al.*, 2011; Zhang *et al.*, 2014) (Figure 3.1B). In the LI, the mucus is subdivided into an inner firmly adherent layer largely devoid of bacteria and an outer layer that is looser and non-adherent and allows some microbial colonization (Johansson *et al.*, 2008).

3.4 Community structure of the vertebrate gut and plant root microbiota

3.4.1 Where do they come from?

A relevant difference for experimentation on the plant root and vertebrate gut microbiota is the ease with which the start inoculum of the root microbiota can be defined.

This is due to a predominant horizontal acquisition of root endophytes from the surrounding soil biome, although in some plant species there is evidence for additional vertical transmission of seed-borne endophytes Barret *et al.* (2015). These endophytes mainly belong to Proteobacteria and can colonize seeds via different colonization routes, including flowers, fruits as well as roots, leaves, and stems (Truyens *et al.*, 2015). Even though vertical transmission in mammals is not as explicit as in plants (none are transferred with the germline), vertical transmission nevertheless occurs. The transmission from parent to offspring results from the birth process itself, from milk, and from the close contact that comes from parental care (Unger *et al.*, 2015). In humans, vaginal birth inoculates the newborn with a set of strains that can be matched to the mother, whereas caesarean section results in colonization with skin microbes originating from various caregivers (Dominguez-Bello *et al.*, 2010). Breast milk is also an important source of microbiota and antibodies that shape the gut microbiome (Newburg and Morelli, 2015), and introduction of solid foods brings rapid shifts in the bacterial community composition toward an adult-like microbiome (Koenig *et al.*, 2011). Vertical transmission from mother to infant gut microbiota is sometimes behaviorally increased in mammals by feeding mother's fecal matter to their infants. In koalas, for instance, this transmission is believed to participate in the digestion of eucalyptus (Osawa *et al.*, 1993). Additionally, group living is known to aid the transmission of commensal microbes between members of family groups (humans), troupes (primates), and most likely herds as well. Co-habitation in humans leads to sharing of microbiota, which is enhanced when dogs also co-habit in the same house (Song *et al.*, 2013). Ironically, hygiene measures aimed at reducing pathogen transmission may have had broad negative impacts on the transmission of commensals and may underlie the loss of diversity observed in the West (Blaser and Falkow, 2009).

3.4.2 Who are they?

Despite the vast prokaryotic biodiversity found in the biosphere (currently >80 bacterial phyla are described), the host-associated microbiota is dominated numerically by a few phyla. The rhizosphere and the root endophytic compartment of unrelated plant species is often enriched for bacteria belonging to three main phyla (Proteobacteria, Actinobacteria, and Bacteroidetes). In contrast, abundant soil bacteria belonging to the phylum Acidobacteria are excluded from the endophytic compartment (Bulgarelli *et al.*, 2013). Compared with the surrounding soil, microbiota members belonging to the phylum Proteobacteria are consistently enriched in the rhizosphere/endosphere compartments of monocotyledonous and dicotyledonous plants, including perennial and annual plants (Bulgarelli *et al.*, 2012, 2015; Edwards *et al.*, 2015; Lundberg *et al.*, 2012; Ofek-Lalzar *et al.*, 2014; Peiffer *et al.*, 2013; Schlaeppi *et al.*, 2014; Shakya *et al.*, 2013; Zarraonaindia *et al.*, 2015). This likely reflects niche adaptation (nutrient availability, oxygen levels) and the ability to efficiently invade and persist inside or outside the roots of divergent plant species. Firmicutes and Bacteroidetes are by far the two most-abundant phyla detected in adult human and mouse feces. Other phyla represented include the Actinobacteria, Verrucomicrobia, and a number of less-abundant phyla such as the Proteobacteria, Fusobacteria, and Cyanobacteria (Eckburg *et al.*, 2005). Similar to the rhizosphere compartment, the mucus layer of the gut represents a particular niche favoring the proliferation of specialized inhabitants. It has been estimated that at least 1% of the gut microbiota can degrade mucins as a source for carbon and nitrogen (Hoskins and Boulding, 1981). Select types of bacteria can also attach to mucins, such as Bifidobacterium bifidum, which has the ability to stimulate mucin production via butyrate-induced expression of MUC2, while others can degrade the nine-carbon sugar sialic acid found in host glycoconjugates (Almagro-Moreno and Boyd, 2009; Gaudier *et al.*, 2004; Leitch *et al.*, 2007).

3.4.3 Are there structural similarities across diverse host-associated microbial communities?

Striking physiological (dis-)similarities exist between organs dedicated to nutrient acquisition in hosts belonging to different taxonomic lineages. However, the extent to which microbial communities living in association with phylogenetically divergent hosts overlap with each other is largely unknown. In an attempt to unravel host-specific and conserved signatures in the microbiota, we retrieved and re-analyzed the raw sequencing data contributed by 14 previous large-scale 16S rRNA gene survey studies (Table S1). These comprise >3,200 samples from more than 40 different host species, including human, other mammals, and fish gut, as well from the root and rhizosphere of the flowering plant Arabidopsis thaliana and relative species, maize, rice, barley, and grapevine. In addition, we included samples from several species of cnidarian hydra, a freshwater basal animal featuring a gut forming a hollow cavity within the body with one opening, the mouth.

To analyze the data, we followed the QIIME (Caporaso *et al.*, 2010) closed-reference protocol and used SortMeRNA (Kopylova *et al.*, 2012) to cluster the sequences into operational taxonomic units (OTUs) at 97% sequence similarity (see Supporting Ma-

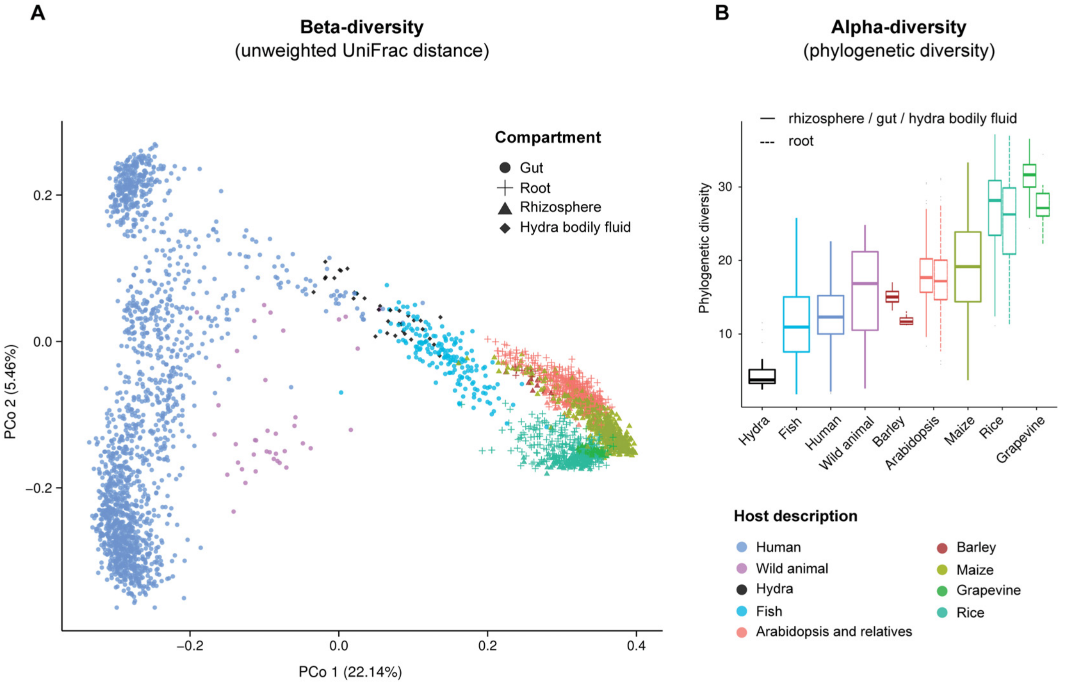


Figure 3.2: alpha- and beta-diversity Analyses. (A) Principal coordinate analysis (PCoA) of pairwise unweighted UniFrac distances between samples. The color and shape of each point represent the host and compartment, respectively. (B) Comparison of alpha-diversity between hosts based on the whole tree phylogenetic diversity index (PD), sorted by ascending order of complexity. See Table S1 for more information about the individual host species included in each study.

terial). Analyses of beta-diversity using principal coordinate analysis (PCoA) revealed a clear clustering of samples according to their respective host species (Figure 3.2A). Although all samples are derived from organs with a dedicated function in nutrient uptake, we found striking qualitative differences between their associated microbial communities. This disparity can be explained by the increased abundance of members of the Bacteroidetes phylum in the mammalian stool samples (particularly those belonging to the orders Bacteroidales and Clostridiales) and the enrichment of members of the families Pseudomonadaceae, Streptomycetaceae, and Comamonadaceae in the rhizosphere and plant root compartments (Figure 3.3). Intriguingly, the bacterial communities in the fish gut are more closely related to those in the root and rhizosphere samples than to the mammalian gut, partially due to an increased abundance in Proteobacteria (45.08% and 54.44% in root-associated samples and fish gut, respectively, compared with 4.20% in the case of the human gut; Figure 3.4). In addition, the microbial communities from infant gut (from Koenig et al., 2011) are more closely related to those of plant roots (and therefore soil microbiota) than those associated to adults Figure S1). Together, this suggests that shared environmental and physiological features, rather than phylogenetic relatedness of the hosts, are decisive for community establishment.

Analysis of alpha-diversity (Figure 3.2B and Figure 3.3B). show that the bacterial richness is low in the gut of aquatic organisms and higher in the root and in the rhizosphere of different plant species, consistent with the bacterial diversity detected in their respective surrounding environments (aquatic versus soil environments) Curtis *et al.* (2002). For all plant species surveyed, the bacterial diversity is lower in the endosphere compartment (root) compared to the rhizosphere compartment (Figure 3.2), in concordance with previous studies Bulgarelli *et al.* (2012); Edwards *et al.* (2015); Lundberg *et al.* (2012). The extent of this gradient in diversity, as well as the differentiation between the two compartments, appears to be dependent on the plant species, indicating a strong host-dependent effect on community establishment.

A phylogenetic comparison of the abundant community members across hosts (OTUs, with a relative abundance higher than 0.1% on average) reveals clear qualitative structural differences between mammalian gut and plant root and rhizosphere samples (Figure 5). These distinct sets of bacterial communities show virtually no overlap even at high taxonomic levels. Samples obtained from human and mammalian guts are dominated by OTUs belonging to the orders Bacteroidales and Clostridiales (34.55% and 51.26% relative abundances, respectively), while these are almost completely absent in the root and rhizosphere samples (0.70% and 0.80%, respectively). This striking difference in community composition in independently evolved, yet functionally related, gut and root organs might be explained by adaptations to specific host and environmental needs, including niche-specific factors such as oxygen levels, pH, and organic carbon availability. Our findings also make a direct transfer and persistence of microbiota members from numerous root-derived dietary plant products in the human gut unlikely.

3.4.4 Do they fluctuate over time?

Despite the fact that infancy or the seedling stage for plants are critical windows for microbiota assembly, very little is known about the earliest steps driving host colo-

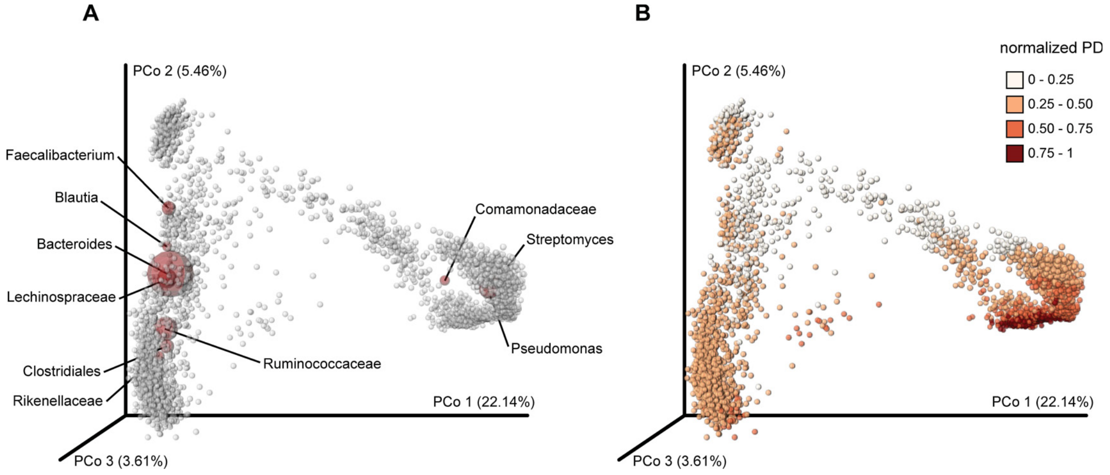


Figure 3.3: 3D PCoA plots. (A) Biplots depicting the taxa with the largest contribution to the ordination space (order Clostridiales; families Ruminococcaceae, Rikenellaceae, Lechinospraceae, Comamonadaceae; genera Streptomyces, Pseudomonas, Bacteroides, Blautia, Faecalibacterium). (B) PCoA plot showing the alpha-diversity variation as measured by the PD index across all samples included in the study.

nization by pioneer bacteria. Assembly of the infant gut microbiome begins at birth (early reports described it as chaotic), and diversity levels slowly increase until ~2-3 years of age (Koenig *et al.*, 2011; Palmer *et al.*, 2007; Yatsunenko *et al.*, 2012). Sampling from birth to 2.5 years of age revealed the following: (i) community richness increased gradually over time, (ii) the use of antibiotics, changes in diet, and infections led to jumps from one stable consortium of species to another, and (iii) members of the Bacteroidetes phylum were co-dominant with members of the Firmicutes phylum after the introduction of solid foods (Koenig *et al.*, 2011). The adult-like microbiota is characterized by a greater stability (David *et al.*, 2014a; Spor *et al.*, 2011). About 60% of the bacterial strains in the intestine are detected over a 5-year time frame, and Bacteroidetes and Actinobacteria were identified as the most stable phyla (Faith *et al.*, 2013). In contrast to the chaotic microbial succession described for the infant gut, the structure of the root microbiota during the plant life cycle appears rather stable. Despite a higher variability observed during the seedling stage (Chaparro *et al.*, 2014), microbiota acquisition from soil appears to occur relatively rapidly, initiating within 24 hr after sowing and approaching a steady state within 2 weeks (Edwards *et al.*, 2015). Once established, there is little evidence for dramatic changes even late in the life cycle of annual *A. thaliana* plants, when organic carbon and nitrogen are spatially re-allocated during the transition from vegetative to reproductive growth for seed formation (Lundberg *et al.*, 2012). This surprising stability might be explained by the sessile nature of plants, together with a rather stable soil-borne inoculum source, which prevents extreme fluctuations in input communities throughout a rapid annual plant's life cycle. Whether this also applies to longer-lived perennials and to repeated croppings of the same species at the same location remains to be further substantiated (Donn *et al.*, 2015).

3.5 Major factors driving community establishment and composition

Inter-individual differences in the gut and the plant microbiota are likely to be dictated by many modulating factors, including environmental parameters but also diet/soil-type, microbe-microbe interactions, host genotype, and host immune system (Figure 3.1).

3.5.1 Environmental factors

pH

Bacterial community composition is strongly correlated with differences in soil pH, with soils at near-neutral pH showing the highest microbial diversity (Fierer and Jackson, 2006). Roots can acidify the rhizosphere up to two pH units compared to the surrounding soil through release of protons, bicarbonate, organic acids, and CO2 (Hinsinger *et al.*, 2003). Along the digestive tract, the increase in bacterial titer can be attributed to several factors, such as pH and bile acids. The pH is very low in the stomach (pH 1.5-5), restricting bacterial growth, increases in the SI (duodenum pH 5-7, jejunum 7-9, ileum 7-8) and drops in the colon (pH 5-7) (Walter and Ley, 2011) (Figure 3.1B). Many types of bacteria, in both the gut and the soil, are sensitive to pH, and this is thought to structure communities to a large degree (Duncan *et al.*, 2009), although it is difficult to disentangle the exact contribution of pH on the overall community structure due to likely interaction with many other factors.

Figure 3.4: Cumulative abundance plots. (Figure on next page).
Relative abundances grouped at the phylum or class taxonomic level for each sample included in the meta-analysis. The bar plots have been arranged along the x axis separating different host groups as well as different species and compartments.

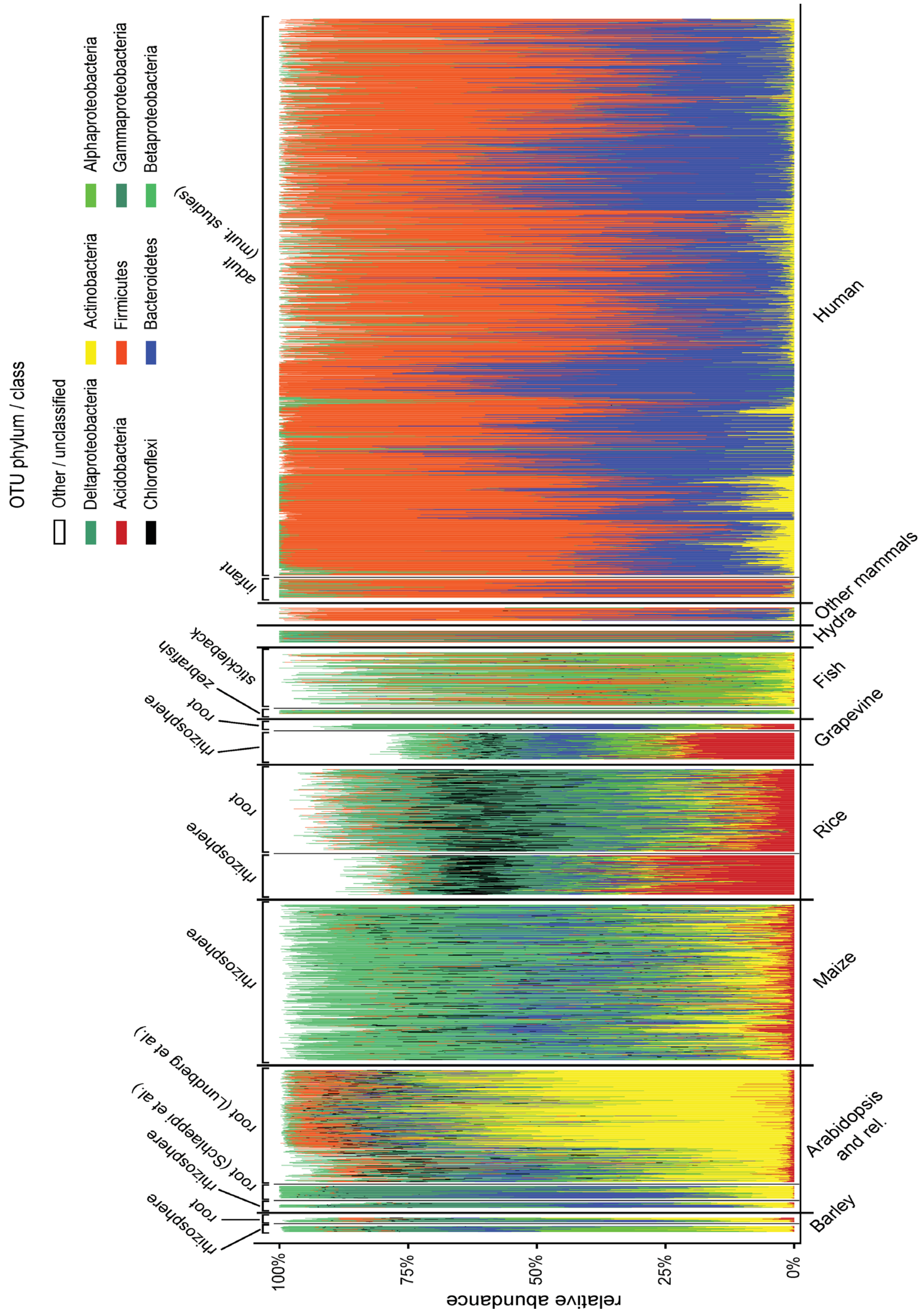


Figure 3.4: Cumulative abundance plots. (Caption on previous page).

Oxygen

Although both gut and root systems are dedicated for nutrient uptake, O2 levels are controlled in opposing directions. In the vertebrate gut, luminal microbes generally face anaerobic conditions favoring fermentative metabolism, while in soil and along the root (micro-)aerobic conditions are found (Figure 3.1). This might be a major factor explaining structural and functional differences between the microbiota of the vertebrate gut and plant roots (Figure 5). The gut microbiota of healthy individuals is dominated by anaerobic bacteria, which outnumber aerobic and facultative anaerobic bacteria by a factor of 100-1,000:1 (Quigley and Quera, 2006), while the root microbiota is enriched for Proteobacteria, a phylum dominated by aerobic species. Consistent with this, genes encoding high-affinity oxidases that use O_2 as a terminal electron acceptor are overrepresented in gut metagenomes, whereas those encoding low-affinity oxidases are enriched in soil metagenomes (Morris and Schmidt, 2013). It is arguably in the host's interest to limit respiration, because (i) limiting respiration will control bacterial growth and (ii) promoting fermentation will result in SCFA availability. Nonetheless, there is a biologically relevant gradient of oxygen levels in both the soil and the gut that is likely to influence microbial community structure at the micro-levels. Despite the fact that plant roots generally face (micro-)aerobic conditions, soil O_2 levels can also fluctuate as a function of soil wetting/drying (Noll *et al.*, 2005), with anoxic niches in the center of soil aggregates. Similarly, a higher O_2 concentration is found at the surface of the epithelium compared with the lumen. Some facultative aerobes can grow along this oxygen gradient by respiring O2 close to the epithelium using flavins and thiols as electron shuttles to respire at 'long distance' (Khan *et al.*, 2012).

Temperature

While thermal stability exists in the gut of mammals (endotherm), higher temperature fluctuation is observed for plants or ectothermic animals that rely on the external temperature to regulate their internal body temperature. It has been reported that the bacterial community in soil is modulated by temperature (Barcenas-Moreno *et al.*, 2009), although plant microbiota functions must remain stable under a wide range of temperatures.

3.5.2 Nutritional drivers

For both plant roots and vertebrate guts, diet (for plants, soil type defines the diet) is a major driver for microbial community structure (Bulgarelli *et al.*, 2012; Cotillard *et al.*, 2013; Carmody *et al.*, 2015; David *et al.*, 2014b; Edwards *et al.*, 2015; Ley *et al.*, 2008a; Lundberg *et al.*, 2012; Muegge *et al.*, 2011; Schlaeppi *et al.*, 2014; Peiffer *et al.*, 2013; Turnbaugh *et al.*, 2009).

Organic carbon is widely considered to be the most important factor limiting bacterial growth in different soils (Demoling *et al.*, 2007). Isotope probing experiments using different plant species revealed that an average of 17% of all photosynthetically fixed carbon is transferred to the rhizosphere through root exudates (Nguyen, 2003), highlighting a considerable organic carbon deposition in soil. Low molecular weight carbon substrates such as dicarboxylic acids, exuded by roots in large quantities to acidify the rhizosphere, also enhance the availability of Pi and micronutrients such as manganese, iron, and zinc. These dicarboxylic acids are an important driver mediating soil community shifts, leading to an increase in the relative abundance of beta-Proteobacteria, gamma-Proteobacteria, and Actinobacteria (Eilers *et al.*, 2010).

The evolution of the mammalian gut microbiota has been greatly influenced by host diet. Mammals, their gut microbiota, and their diet types are part of a dynamic tripartite coevolution (Ley *et al.*, 2008b). The majority (80%) of extant mammals are herbivorous, which stands in contrast to the early mammals that were most likely carnivorous based on their tooth morphology. The rise in herbivory could only have been accomplished with the necessary changes in gut microbes, since mammalian genomes lack the necessary genes encoding plant cell wall degrading enzymes. Comparisons of microbiomes between host species highlight the specific adaptations of the microbiota to the host diet, such as an increased abundance of genes encoding the necessary enzymes and their respective pathways (Eilam *et al.*, 2014), as exemplified in a comparison between the termite hindgut and the bovine rumen metagenome (Brulc *et al.*, 2009). The latter is enriched for genes encoding glycoside hydrolases, cellulosome enzymes, and nitrogen-related uptake proteins. In contrast, the termite hindgut microbiome showed an enrichment for genes involved in the degradation of the cellulose backbone and nitrogen fixation. This clearly reflects the differences in diet of the hosts (forages and legumes versus nitrogen-poor wood).

Figure 3.5: Phylogenetic analysis of OTU abundances. (Figure on next page). (A) Phylogeny inferred from the representative sequences of all OTUs that had at least 0.1% relative abundance on average for all samples of a host species (1,133 in total). The color of each leaf depicts the taxonomic classification of its corresponding OTU. (B) Average relative abundances of abundant OTUs across all samples of each host (log-transformed).

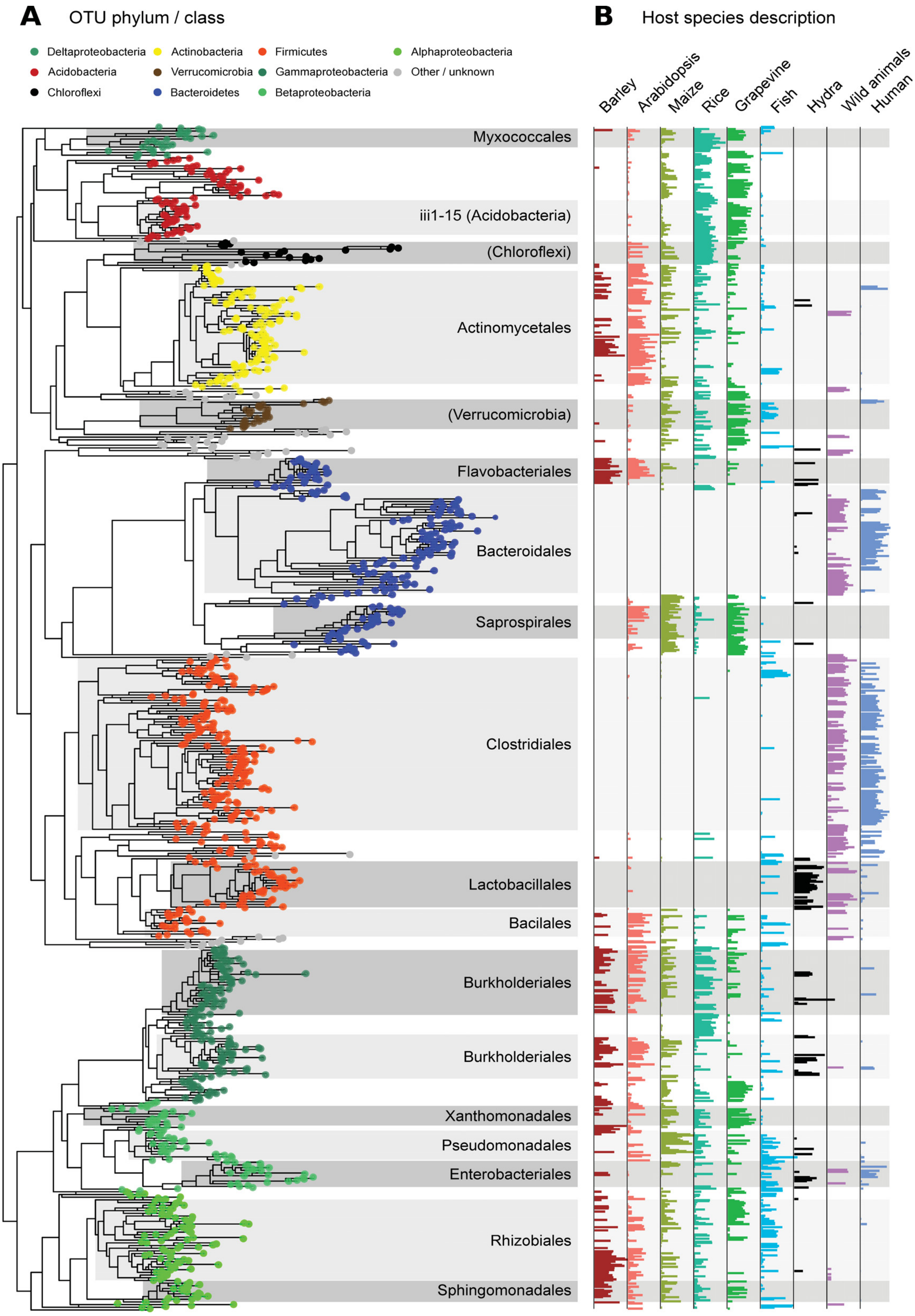


Figure 3.5: Phylogenetic analysis of OTU abundances. (Caption in prev. page)

3.5.3 Microbe-microbe interactions

The role of microbe-microbe interactions is also critical for shaping microbiota structure in both plant and animal systems (Bulgarelli *et al.*, 2015; Fraune *et al.*, 2015; Hacquard and Schadt, 2015; Trosvik *et al.*, 2009). The combination of synergistic, beneficial, and antagonistic interactions among microbiota members colonizing the gut and plants is likely to have a major impact on overall community structure. Therefore, individual members of a community may contribute to the overall stability of the system, and consequently, each community member must be viewed as a potential internal driver of microbial community assemblage. Microbial co-occurrence and co-exclusion patterns are now emerging as important concepts for understanding the rules guiding microbial community assembly (Cardinale *et al.*, 2015; Faust *et al.*, 2012; Zhang *et al.*, 2014).

3.5.4 Host genotype

Intra-species plant genetic diversity explains less variation in community structure than soil type and root fraction (soil, rhizosphere, and endosphere). Surveys of the bacterial community structure of 27 maize inbred lines, 6 cultivated rice varieties, 3 barley accessions, and several A. thaliana accessions each point to a small (∼5%-6% of variation) but significant role of the host genotype on community composition (Bulgarelli *et al.*, 2012, 2015; Edwards *et al.*, 2015; Lundberg *et al.*, 2012; Peiffer *et al.*, 2013; Schlaeppi *et al.*, 2014). This suggests a link between host diversification and microbial community establishment (see below).

In humans, family members are often observed to have more similar microbiotas than unrelated individuals (Tims *et al.*, 2013; Turnbaugh *et al.*, 2009; Yatsunenko *et al.*, 2012). Familial similarities are usually attributed to shared environmental influences, such as dietary preference, a powerful shaper of microbiome composition (Cotillard *et al.*, 2013; David *et al.*, 2014b; Wu *et al.*, 2011). However, host genetics also play a small but statistically significant role in shaping the composition and structure of the gut microbiome. Studies comparing microbiota between human subjects differing at specific genetic loci have shown gene-microbiota interactions (Khachatryan *et al.*, 2008; Rehman *et al.*, 2011). A more general approach to this question has linked genetic loci with abundances of gut bacteria in mice (Benson *et al.*, 2010; McKnite *et al.*, 2012), although diet effects outweigh the host genotype effects (Parks *et al.*, 2013). In humans, earlier twin studies failed to reveal significant genotype effects on microbiome diversity (Turnbaugh *et al.*, 2009; Yatsunenko *et al.*, 2012). However, a recent report by Goodrich et al. (2014) comparing monozygotic (MZ) with dizygotic (DZ) twin pairs identified specific taxa as heritable (i.e., the variability in the relative abundances of these taxa across the population was partially driven by host genotype variation). These taxa include health-associated Faecalibacterium and Bifidobacterium and lean phenotype-mediating Christensenella (Goodrich *et al.*, 2014).

3.6 Host immune systems and microbiota homeostasis

Plants and animals each engage structurally related pattern recognition receptors (PRRs) for recognition of evolutionarily conserved non-self microbial structures (i.e., lipopolysaccharides [LPS], lipopeptides, flagellin, chitin) at the cell surface, and activation of these is typically sufficient to halt microbial proliferation. However, successful plant and animal pathogens have evolved mechanisms to dampen or escape PRR-mediated host responses to foster virulence. In response, members of the NLR (nucleotide-binding domain leucine-rich repeat containing) family of intracellular immune receptors in plants and animals are activated by the action of pathogen virulence factors or by direct binding of the virulence factors themselves (Boller and Felix, 2009; Jones and Dangl, 2006; Maekawa *et al.*, 2011). Active animal PRRs and NLR inflammasomes each can instruct the mammalian adaptive immune system and cause spatially dispersed response in plants, as detailed below.

Detection of microbial patterns via PRRs constitutes the first layer of immunity in plants and animals and triggers a variety of output responses. In animals, these include instruction and either activation or suppression of the adaptive immune system via cytokine signaling and cell migration to and from infection sites and lymphoid organs. Because there are no circulating cells in plants, PRR- and NLR-dependent signaling can lead to differential local and systemic signals that result in adequate defense outputs at and directly surrounding the site of infection and a poised defense in distal organs. Analogous to cytokines, plants deploy a handful of defense phytohormones that have variable domains of signaling and instruct cells neighboring an infection site, and even systemically to distal organs, to be ready to respond to infection (Pieterse *et al.*, 2012).

The lack of circulating immunocytes also demands that each plant cell in an organ be capable of recognizing all pathogens adapted to that organ. This drives a complicated requirement for coordination of normal cellular functions, mediated by growth-regulating hormones, and immune output mediated by the defense phytohormones. This coordination is manifested as trade-offs between growth and immunity (Belkhadir and Jaillais, 2015). Thus, systemic acquired resistance in above-ground organs is triggered by biotrophic pathogens and mediated by salicylic acid (SA), while induced systemic resistance, also active in leaves, is triggered in roots by rhizobacteria and is mediated by jasmonic acid (JA) and ethylene (Spoel and Dong, 2008; van Loon *et al.*, 1998).

Because plant defense phytohormones are key signaling molecules between microbial perception and immune system outputs, their production and perception are common pathways targeted by both potential pathogens and beneficial microbes. Hence, there is evidence that during the early stages of colonization both arbuscular mycorrhizal (AM) and Rhizobium species locally suppress SA signaling (Garcia-Garrido and Ocampo, 2002; Stacey *et al.*, 2006), suggesting that defense phytohormones normally act to inhibit microbial survival in the root. Indeed, culture-dependent studies in A. thaliana have demonstrated a significantly lower load of culturable bacteria in rhizospheres of plants with either defective JA signaling or, conversely, constitutive SA production (Doornbos *et al.*, 2010). Beyond defense phytohormones, other immune outputs have also been implicated by recent studies. In particular, metagenomic studies in rice uncovered genes present in root endophytic bacteria, notably detoxification of reactive oxygen species (Sessitsch *et al.*, 2012).

The overall structure of the *Arabidopsis* root microbiota remains largely robust to host mutations leading to hypo- or hyper-immunity. However, sets of mutants with altered defense phytohormone biosynthesis and/or perception had specifically altered root microbiome taxonomic compositions compared to wild-type. These alterations were congruent with the known effects of the mutants on immune system outputs in leaves. Experiments using both wild soil and its natural community or synthetic soil microcosms in the presence of a synthetic bacterial community demonstrated that SA and/or SA-dependent processes are major contributors to root microbiome composition (S.L., unpublished data). Together, these studies represent some of the insights into mechanisms used by the plant immune system to shape its microbiota.

In the animal gut, a first line of defense consists of the secretion of antimicrobial peptides that are produced deep within the crevices of the epithelial layer, in the crypts between the villi. While some antimicrobial agents are continuously secreted, others are secreted in response to bacterial triggering of specific PRRs (Toll-like receptors, TLRs) on the epithelial cell surfaces. The mucus layer is crucial to prevent systematic activation of these immune responses. When the inner mucus layer is removed chemically (i.e., with dextran sodium sulfate [DSS]) or through gene mutation (MUC2 mutants), bacteria come into contact with epithelial cells and cause an inflammatory response (Johansson *et al.*, 2010; Van der Sluis *et al.*, 2006). In contrast to plants, the adaptive immune system also plays a role for sequestering symbiotic bacteria in the lumen through the secretion of immunoglobin A (IgA) that target epitopes of intestinal bacteria. Like the antimicrobial activity of the innate immune system, the adaptive immune system can be regulated in parts by TLR signaling (Iwasaki and Medzhitov, 2010). Together, the adaptive and innate immune systems have mechanisms for detecting surface-associated bacteria and work together to reduce inflammation. Because the adaptive immune system is (largely) unique to vertebrates, and based on the observation that vertebrates, notably mammals, harbor microbial communities with much greater complexity than do invertebrates, McFall-Ngai *et al.* (2013) have proposed that the adaptive immune system itself is important in the shaping and maintenance of high microbial diversity.

3.7 Co-diversification of host-microbe communities

By comparing the bacterial communities associated with maize genotypes or other grasses, a significant correlation between rhizobacterial communities and the host phylogenetic distance has been detected, suggesting that the host's evolutionary history can be a good predictor of root microbiota structure (Bouffaud *et al.*, 2014). A comparison of inter-species host phylogeny and microbiota diversification in four Brassicaceae plant species, including *A. thaliana*, which diverged ∼35 Ma revealed only quantitative differences. This diversification cannot be explained solely by the phylogenetic distance of these hosts but likely includes plant species-specific ecological adaptations (Schlaeppi *et al.*, 2014). However, qualitative differences can be observed when comparing more distantly related plant species such as *A. thaliana* and barley (dicotyledonous versus monocotyledonous plants), which diverged ∼150 Ma (Bulgarelli *et al.*, 2015). Marked differences in microbiota composition were also reported for Hydra vulgaris and Hydra oligactis, cnidarian animal groups that diverged approximately 100 Ma and have been cultivated under identical laboratory conditions for decades (Franzenburg *et al.*, 2013). In mammals, similarities in microbial community composition between members of the same species raise the question of whether the bacterial communities track mammalian phylogeny. This would be expected if the bacteria are passed vertically from parent to offspring, which some mammal species encourage behaviorally. Patterns of relatedness of the bacterial communities were compared to the mammalian phylogeny (Ley *et al.*, 2008a). For subsets of the mammalian phylogeny, the trees matched at a rate that is greater than expected by chance. For instance, this pattern was observed in the case of bears, which are an animal group candidate for mother-offspring transmission due to prolonged contact between the cub and the mother, implying that an ancestral microbial population diversified at the same time that bears speciated. A comparison of the microbial communities associated to great ape species, including Homo sapiens, also revealed that the host species phylogeny was congruent to the pattern of relatedness of their gut microbial communities, which diverged in a manner consistent with vertical inheritance (Ochman *et al.*, 2010). However, a comparative analysis of the gut microbiota of humans with the ape species indicates an accelerated change in the microbiota composition of humans that cannot be explained by evolutionary distance (Moeller *et al.*, 2014). A recent study of one isolated Amazonian tribe revealed the highly diverse gut microbiota, in both composition and functions, including a broad range of antibiotic resistance genes, suggesting that the Western lifestyle has dramatically reduced bacterial diversity (Clemente *et al.*, 2015).

Taken together, these data indicate generally that a correlation between microbiota and host phylogeny can be explained by co-diversification from common ancestors. Nonetheless, the hugely different generation times of bacteria compared to their associated eukaryotic hosts together with the high density of microbes in the gut or surrounding the root system suggest that the evolution of host-microbe communities is mainly determined by other selective forces, including microbe-microbe and host-microbe-environment interactions.

3.8 Metagenome analysis-inferred functions of the gut and the plant microbiota

The gut microbiota is dominated by a few bacterial phyla, but more variation is observed when focusing on lower taxonomic levels. The relative abundance of individual species can vary over a 10-fold range among individual humans Spor *et al.* (2011). In contrast, at the level of gene functions, less variability is observed among individuals, pointing to functional redundancy within the bacterial microbiota and the existence of a conserved functional core (Huttenhower *et al.*, 2012; Turnbaugh *et al.*, 2009).

Given the critical function in nutrient acquisition, it is not surprising that gene functions found in the gut microbial community are influenced by both long- and short-term changes in diet (David *et al.*, 2014b; Muegge *et al.*, 2011; Suez *et al.*, 2014; Wu *et al.*, 2011) Pathways found over all human body parts ('core' pathways) include translational machinery, nucleotide charging, ATP synthesis, and glycolysis (Huttenhower et al., 2012). The functional categories found specifically enriched in the gut microbiota are related to metabolism categories (genes involved in starch, sucrose, and monosaccharide metabolism, including many glycoside hydrolase families). More specifically,

functions related to fermentation of complex sugars and glycans to SCFAs, methanogenesis, synthesis of essential amino acids and vitamins, and hydrolysis of phenolic glycosidic conjugates are enriched (Gill *et al.*, 2006; Huttenhower *et al.*, 2012; Qin *et al.*, 2010; Turnbaugh *et al.*, 2009). Some of these functions, such as fermentation and carbohydrate metabolism and vitamin biosynthesis, are also highly expressed in the gut microbiome, as assessed by metatranscriptome analysis (Turnbaugh *et al.*, 2010).
For plant studies, experimental design is more standardized across individuals, which often allows for direct or indirect tests of functional enrichment (Bulgarelli *et al.*, 2015; Mendes *et al.*, 2014; Ofek-Lalzar *et al.*, 2014), in contrast to the human gut microbiome. Shared functional categories found across at least two plant rhizosphere studies relate to iron transport and metabolism, nitrogen metabolism, transport and secretion systems, as well as chemotaxis and motility (Mendes *et al.*, 2014; Ofek-Lalzar *et al.*, 2014; Sessitsch *et al.*, 2012). Similar functions were also found in a metaproteogenomics study of the rice rhizosphere, although in addition, a major role for one-carbon compound recycling could be identified (Knief *et al.*, 2012). However, considerable differences were found in these studies, and additionally no specific function can be assigned for a large proportion of annotated genes in metagenomic studies (42%-86% in the gut; 59% in the plant rhizosphere) (Gill *et al.*, 2006; Ofek-Lalzar *et al.*, 2014; Qin *et al.*, 2010). A striking commonality between the gut and root metagenome studies is the significant enrichment/high abundance of phage-related functions (Bulgarelli *et al.*, 2015; Qin *et al.*, 2010), but the exact role of these functions is not known.
To gain further insight into the evolutionary forces acting on genes in relation to their functional roles, natural selection was assessed using dN/dS ratios for gene families in the barley rhizosphere and human gut microbiomes (Bulgarelli *et al.*, 2015; Schloissnig *et al.*, 2013). Positive selection is a hallmark of protein families implicated in molecular arms races between two competing organisms. In the rhizosphere, proteins involved in host-pathogen interactions showed significant signs of positive selection, such as the type III secretion system and its associated effectors, phage elements, and microbial CRISPR proteins (Bulgarelli *et al.*, 2015). Similarly, CRISPR-related families, as well as transposases and families related to antibiotic resistance, showed signatures of positive selection in the human gut microbiome (Schloissnig *et al.*, 2013).

3.9 Concluding remarks and perspectives

To complement large-scale community profile and metagenome studies, reference collections of several hundred isolates from different human body sites and their corresponding genome sequences have been generated (Goodman *et al.*, 2011). For plant-associated microbial communities, similar projects aiming to maximize phylogenetic diversity of cultured bacteria through cross-referencing with culture-independent community profiling experiments are about to be concluded (P.S.-L. and J.L.D., unpublished data). In the future, these genome collections may allow determination of multi-locus reference gene collections for the identification of individual strains within a community, as an alternative to lower-resolution 16S rRNA-based taxon identification, as well as comparative analyses of thousands of genomes for association-based analyses, to link genes and genetic variants to particular phenotypes. The construction of defined (synthetic) communities and their assessment under controlled environments with germ-free eukaryotic hosts allows studies of community resilience and responses to perturbation at the level of individual members and simplifies testing of specific hypotheses relating to individual attributes of other community members and the host (Faith *et al.*, 2014; Guttman *et al.*, 2014). Controlled experimental systems will reduce the noise inherent to any natural environmental sample and will drive the next phase of plant and gut microbiota research in which scientific conclusions are based on causation rather than correlations. For a detailed description of the meta-analysis, see Supporting Material. The OTU count matrices and taxonomic information as well as the scripts used to analyze the data and generate the figures of this study are available at `http://www.mpipz.mpg.de/R_scripts`.

3.10 Author contributions

S.H., R.G.-O., S.S., and P.S.-L. designed research. R.G.-O., A.G., and R.K. designed and R.G.-O., A.G., and G.A. performed the computational analysis. S.H., R.G.-O., S.S., S.L., A.C.M., J.L.D., R.K., R.L., and P.S.-L. wrote the paper.

3.11 Acknowledgements

Work in the different research groups was supported by the Max Planck Society (to P.S.-L.), a European Research Council advanced grant (ROOTMICROBIOTA to P.S.-L.), the Cluster of Excellence on Plant Sciences program (funded by the Deutsche Forschungsgemeinschaft to P.S.-L. and A.C.M.), NSF Microbial Systems Biology (grant IOS-0958245 to J.L.D.), NSF INSPIRE (grant IOS-1343020 to J.L.D.), and DOE Feedstocks (grant SC0010423 to J.L.D). S.S. was supported by a postdoctoral grant from the Research Foundation - Flanders (FWO-Vlaanderen). J.L.D. is an Investigator of the Howard Hughes Medical Institute, supported by the HHMI and the Gordon and Betty Moore Foundation (GBMF3030).

3.12 Supporting material

The supporting material corresponding to this section, including all supplementary tables and figures can be accessed via the online version of the published article and have not been included in this thesis due to space limitations. All intermediate data as well as the scripts used to analyze the data and generate the figures of this study are available at `http://www.mpipz.mpg.de/R_scripts`.

CHAPTER 4

Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives

Status	**Published**
Journal	*PNAS* (Impact factor 9.423)
Citation	Schlaeppi, K., Dombrowski, N., Garrido-Oter, R., Themaat, E.v.L.v. and Schulze-Lefert, P. (2014). Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. *Proceedings of the National Academy of Sciences*, **111** (2): 585-592.
URL	http://dx.doi.org/10.1073/pnas.1321597111
Own contribution	Analyzed the data (with co-authors) Interpreted the data (with co-authors)

4.1 Significance

All plants carry distinctive bacterial communities on and inside organs such as roots and leaves, collectively called the plant microbiota. How this microbiota diversifies in related plant species is unknown. We investigated the diversity of the bacterial root microbiota in the Brassicaceae family, including three *Arabidopsis thaliana* ecotypes, its sister species *Arabidopsis halleri* and *Arabidopsis lyrata*, and *Cardamine hirsuta*. We show that differences in root microbiota profiles between these hosts are largely quantitative and that host phylogenetic distance alone cannot explain the observed microbiota diversification. Our work also reveals a largely conserved and taxonomically narrow root microbiota, which comprises stable community members belonging to the Actinomycetales, Burkholderiales, and Flavobacteriales.

4.2 Abstract

Plants host at the contact zone with soil a distinctive root-associated bacterial microbiota believed to function in plant nutrition and health. We investigated the diversity of the root microbiota within a phylogenetic framework of hosts: three *Arabidopsis thaliana* ecotypes along with its sister species *Arabidopsis halleri* and *Arabidopsis lyrata*, as well as *Cardamine hirsuta*, which diverged from the former ∼35 Mya. We surveyed their microbiota under controlled environmental conditions and of *A. thaliana* and *C. hirsuta* in two natural habitats. Deep 16S rRNA gene profiling of root and corresponding soil samples identified a total of 237 quantifiable bacterial ribotypes, of which an average of 73 community members were enriched in roots. The composition of this root microbiota depends more on interactions with the environment than with host species. Interhost species microbiota diversity is largely quantitative and is greater between the three *Arabidopsis* species than the three *A. thaliana* ecotypes. Host species-specific microbiota were identified at the levels of individual community members, taxonomic groups, and whole root communities. Most of these signatures were observed in the phylogenetically distant *C. hirsuta*. However, the branching order of host phylogeny is incongruent with interspecies root microbiota diversity, indicating that host phylogenetic distance alone cannot explain root microbiota diversification. Our work reveals within 35 My of host divergence a largely conserved and taxonomically narrow root microbiota, which comprises stable community members belonging to the Actinomycetales, Burkholderiales, and Flavobacteriales.

4.3 Introduction

Plants host distinct bacterial communities associated with roots and leaves (Vorholt, 2012; Bulgarelli *et al.*, 2013). Both the leaf and root microbiota contain bacteria that provide indirect pathogen protection, but root microbiota members appear to serve additional host functions through the acquisition of nutrients from soil supporting plant growth (Bulgarelli *et al.*, 2013). The plant-root microbiota emerges as a fundamental trait that includes mutualism enabled through diverse biochemical mechanisms, as exemplified by previous studies on numerous plant growth and plant health-promoting bacteria (Bulgarelli *et al.*, 2013).

Recent deep profiling of the root microbiota of *Arabidopsis thaliana* ecotypes, grown under controlled environments, confirmed soil type as major source of variation in root microbiota membership and provided evidence for limited host genotype-dependent variation (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012). Using four soil types on two continents and based on two 16S rRNA gene PCR primer sets, these replicated experiments revealed a similar phylogenetic structure of the root-associated microbiota at high taxonomic rank, including the phyla Actinobacteria, Bacteroidetes, and Proteobacteria. In addition, these studies revealed a minor 'rhizosphere effect' in *A. thaliana*, i.e., a weak differentiation of the bacterial communities in the rhizosphere (soil that is firmly attached to roots) compared with the corresponding unplanted bulk soil.

The genus *Arabidopsis* consists of the four major lineages *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Arabidopsis halleri* and *Arabidopsis arenosa*. The former is the sole self-fertile species and diverged from the rest of the genus ∼13 Mya whereas the other three species radiated approximately ∼8 Mya ((Beilstein *et al.*, 2010); (Figure 4.1). *Cardamine hirsuta* diverged from the *Arabidopsis* species ∼35 Mya and often shares the same habitat with *A. thaliana*. *A. thaliana* has a cosmopolitan distribution whereas the other species occur in spatially restricted populations or developed even endemic subspecies, indicative of their adaptation to specific ecological niches (Hoffmann, 2005). The two diploid species, *A.halleri* and *A. lyrata*, co-occur in Eurasia, but, in contrast to *A. lyrata* (Northern rock-cress), the geographical distribution of *A. halleri* rarely extends into northern latitudes. *A. lyrata* primarily colonizes, similar to *A. thaliana*, low-competition habitats as, for example, tundra, stream banks, lakeshores, or rocky slopes, whereas *A. halleri* (Meadow rock-cress) is tolerant of shading and competition, growing in habitats such as mesic meadow sites (Clauss and Koch, 2006). In contrast to its sister species, *A. halleri* can grow on heavy metal-contaminated soils and serves as a model species for metal hyperaccumulation and associated metal hypertolerance and for extremophile adaptation (Kraemer, 2010).

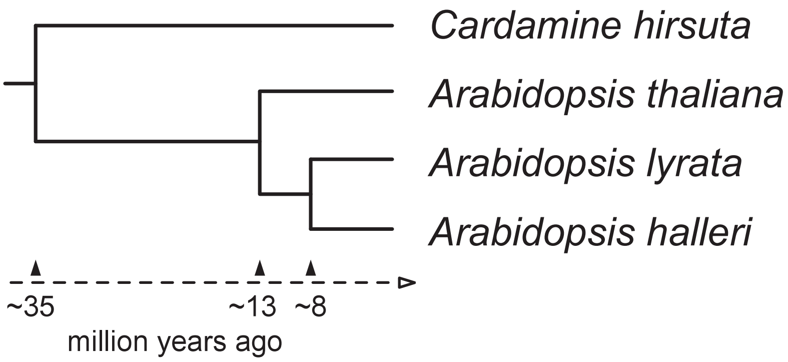


Figure 4.1: Phylogeny of *Arabidopsis thaliana* and relative species. Phylogenetic placement of the *Arabidopsis* species *A. halleri*, *A. lyrata*, and *A. thaliana* and relative species *Cardamine hirsuta*. The relationships and divergence time estimates are based on molecular systematics using combined data of NADH dehydrogenase subunit F and phytochrome A sequences anchored by four fossil age constraints (Beilstein *et al.*, 2010).

The bacterial root microbiota of plants –'plants wear their guts on the outside' (Janzen, 1985)– is conceptually analogous to the gut microbiota of animals owing to a shared primary physiological function of root and gut organs for nutrient uptake. The idea of a core microbiota within a species has been initially explored in humans by revealing an extensive array of shared microbial genes among sampled individuals, comprising an identifiable gut core microbiome at the gene, rather than at the level of organismal lineages (Turnbaugh *et al.*, 2009; Qin *et al.*, 2010). However, using a phylogroup- and tree-independent approach, two prevalent core phylogroups belonging to the clostridial family Lachnospiraceae were identified in the human colon among a total of 210 human beings with widespread geographic origin, ethnic background, and diet (Sekelja *et al.*, 2011). These phylogroups were also detected in a wide range of other mammals and are thought to play a conserved role in gut homeostasis and health. The findings of a core set of species in the human gut microbiota remain contentious as a wider set of samples including developing countries and a broader age range becomes available (Lozupone *et al.*, 2012). However, spatial stratification of the gut microbiota, which is normally missing in fecal samples, led to the definition of a crypt-specific core microbiota in the mouse colon, dominated by aerobic Acinetobacter, regardless of the mouse line used or breeding origin of these mice (Pedron *et al.*, 2012). Finally, evidence for a shared core gut microbiota was found in domesticated and recently caught zebrafish, dominated by Proteobacteria, some Fusobacteria, and Firmicutes (Roeselers *et al.*, 2011). This shared core is believed to reflect common selective pressures governing microbial community assembly within this intestinal habitat. Although root microbiota profiles of numerous plant species, including crops, have been examined (Peiffer *et al.*, 2013; Inceoglu *et al.*, 2011; Hardoim *et al.*, 2011; Sharma *et al.*, 2005), different sampling protocols and low-resolution profiling methods make it difficult to reexamine and compare these for the existence of a conserved core microbiota between plant species.

Here, we present a systematic investigation of host-microbiota diversification within a phylogenetically defined plant species framework, combined with replicated experiments under controlled conditions and sampling in natural habitats. Using deep 16S rRNA gene profiling of root and corresponding soil samples of four host species of the Brassicaceae family, together with rigorous statistical analysis, we show that interhost species microbiota diversity is largely quantitative, and we discuss a possible microbiota coevolution with these hosts. We also compared bacterial community structure variation within and between the tested host species. We provide evidence for the existence of a largely conserved and taxonomically narrow root microbiota between the tested host species, which remains stable in natural and controlled environments. This identified core comprises Actinomycetales, Burkholderiales, and Flavobacteriales. Members of each of these bacterial families are known to promote plant growth and plant health. It is possible that the conserved microbiota represents a standing reservoir of retrievable host services independent of environmental parameters and host species-specific niche adaptations.

4.4 Results

We collected side-by-side growing *A. thaliana* and *C. hirsuta* plants at two natural sites, designated 'Cologne' and 'Eifel', and prepared quadruplicate root and rhizosphere samples for bacterial 16S rRNA gene community profiling (Table 4.1; Supporting Material). In parallel, we conducted two replicate greenhouse experiments using two seasonal batches of natural experimental Cologne soil (SI Appendix, Table S1) on which we cultivated *A. halleri* (Auby), *A. lyrata* (Mn47), and *C. hirsuta* (Oxford), together with the three *A. thaliana* accessions Shakdara (Sha), Landsberg (Ler) and Columbia (Col), and prepared triplicate root samples for microbiota analysis (Table 4.1; Supporting Material). Rhizosphere samples are defined as firmly root-adhering soil particles removed by a washing step and collected by centrifugation. Root samples were washed a second time and treated with ultrasound to deplete root surface-associated bacteria and to enrich for endophytic bacteria (Bulgarelli *et al.*, 2012) (Supporting Material). To quantify the start inoculum for the root-associated bacterial communities, we prepared triplicate samples from unplanted pots of each greenhouse experiment, as well as four samples from bulk soil collected at each natural site (Table 4.1; Supporting Material). Barcoded pyrosequencing of bacterial 16S rRNA gene amplicon libraries generated with the PCR primers 799F (Chelius and Triplett, 2001) and 1193R (Bodenhausen *et al.*, 2013) was used to display bacterial communities (Supporting Material).

We generated 2,110,506 raw reads from 77 samples of the replicated natural-site and greenhouse experiments (Dataset S1). For subsequent analysis, we included 1,567,657 quality sequences (Supporting Material), resulting in a median of 15,603 quality sequences per sample (range 6,339-58,150 sequences per sample). Quality sequences were binned at >97% sequence identity using QIIME (Caporaso *et al.*, 2010) to define operational taxonomic units (OTUs), corrected for differences in sequencing depth between samples by rarefaction to 6,000 sequences per sample. OTU representative sequences were taxonomically classified based on the Greengenes database (McDonald *et al.*, 2012) (Supporting Material) and we identified a total of 88,731 unique bacterial OTUs and a single archea OTU across all samples.

Sample	Species	Cologne	Eifel	GH 1	GH 2
Soil		4	4	3	3
Rhizosphere	*A. thaliana*	4	4	-	-
Root	*A. thaliana*	4	4	8	9
Rhizosphere	*C. hirsuta*	4	3	-	-
Root	*C. hirsuta*	4	3	3	3
Root	*A. halleri*	-	-	3	2
Root	*A. lyrata*	-	-	2	3

Table 4.1: Numerical overview of the experimental setup. Numerical overview of biological replicate samples per sample type, plant species, and experiments. See Supporting Material for the detailed experimental design, including the sequencing effort.

4.4.1 Defining abundant community members

Technical reproducibility of community profiles was determined by repeated library sequencing, and we defined a minimum of 20 sequences per OTU for reproducible quantification of OTU abundance (Figure S1). This reproducibility threshold is similar to previous studies Bulgarelli *et al.* (2012); Pedron *et al.* (2012); Roeselers *et al.* (2011). We noted a low reproducibility for soil microbiota profiles and found that OTU richness does not reach a plateau even at a sequencing depth of 50,000 quality sequences per sample (Figure S2A), These observations, together with the exclusion of low-abundant (<20 sequences) and nonreproducible OTUs for rarefaction analysis (Figure S2B), suggested that OTU richness in soil is the result of a vast number of low-count OTUs. In the root samples, the richness of the abundant community members (OTUs with >20 sequences) was sufficiently captured at a sequencing depth of 6,000 sequences per sample. Consequently, we focus our analyses on the abundant community members (ACMs) of the dataset, which we define to comprise OTUs reaching the threshold of 20 quality sequences in at least one sample. Without application of this abundance threshold, we refer to community profiles rarefied to 6,000 sequences as threshold-independent communities (TICs; Supporting Material).

The ACM, including soil, rhizosphere, and root samples, was represented by 237 bacterial OTUs comprising 55.3% of rarefied quality sequences. Soil and rhizosphere samples contained fewer sequences after thresholding compared with root samples, likely due to increased richness by low-count OTUs in the former two compartments. We normalized the counts of the ACM OTUs per sample, expressed their relative abundance as per mille, and used log2-transformed values for statistical comparisons.

4.4.2 Community composition is defined more strongly by environmental parameters than by host species

We first examined taxonomic composition and ecological diversity parameters in the whole dataset consisting of samples from two natural sites and two greenhouse experiments. (Figure S4). All OTUs of the ACM belonged to the domain of bacteria. In root samples, the majority of OTUs belonged to Proteobacteria (4.2%, 33.6%, 6.4%, and 1.8% in the Alpha-, Beta-, Gamma-, and Deltaproteobacteria subphyla, respectively), Bacteroidetes (27.5%), Actinobacteria (22.1%), and Chloroflexi (2.2%). Soil samples also contain Proteobacteria (52.2%) and Actinobacteria (26.8%), but few Bacteroidetes (3.5%) and, characteristic for this compartment, Firmicutes (10.1%) and Nitrospirae (2.4%). Similar taxonomic characteristics of soil and root samples were also found for TICs. We noted the dominance of a single *Flavobacterium* (OTU162362) in root communities of natural-site and greenhouse experiments, representing, in some of the samples, more than half of the total community. A high OTU diversity in family-rich phyla, such as the Proteobacteria (128 OTUs in 24 families) or Actinobacteria (67 OTUs in 17 families), contrasts with a low taxonomic diversity within the root-specific Bacteroidetes (20 OTUs), all belonging to the family of the Flavobacteriaceae.

To compare community diversity between samples, we used the weighted UniFrac metric (Lozupone and Knight, 2005). Consistent with previous studies (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Peiffer *et al.*, 2013), the hierarchical clustering of UniFrac distances revealed that compartments and environmental conditions (soil types/soil batches, controlled/noncontrolled climates) are the major sources of variation both in the ACMs (Figure 4.2) and TICs (Figure S5). Due to independent library preparation and sequencing, we validated that the variation in the replicate greenhouse experiments reflects biological rather than technical variation (Figure S6). For both natural-site and controlled environment samples, we did not detect a consistent clustering by host species, evidencing that the present sample-to-sample variation obscures a possible host species effect on beta diversity.

We estimated OTU diversity within samples based on the number of OTUs detected (richness) and Faith's Phylogenetic Diversity (PD) metric (Faith, 1992). Root TICs are of lower richness and diversity compared with the soil and rhizosphere microbiota (Figure S7). Of note, roots of plants grown under natural conditions host bacterial communities of increased richness and Faith's PD compared with greenhouse-grown plants. Root TICs and root ACMs did not differ in richness and Faith's PD among the tested host species in natural and greenhouse experiments (Figure S7). This finding further supports the existence of qualitatively similar root-associated bacterial assemblies among *A. thaliana* relatives.

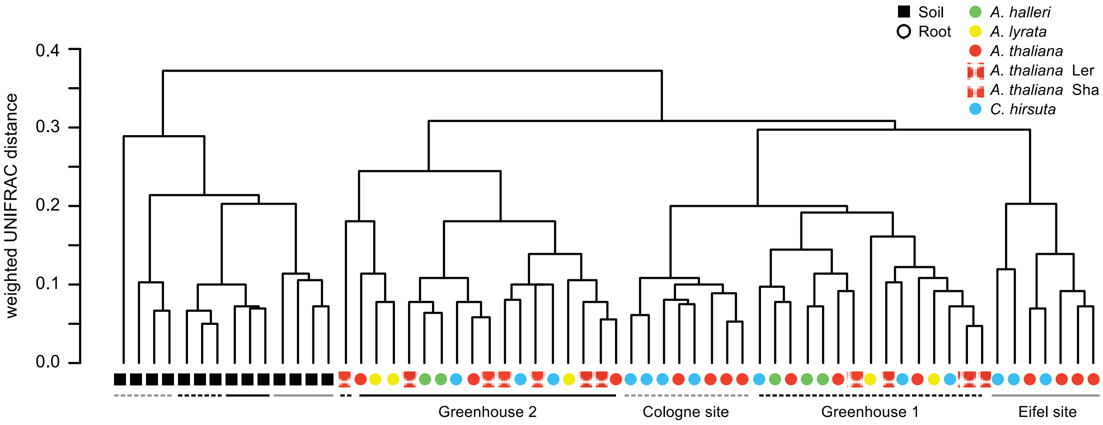


Figure 4.2: Hierarchical clustering of samples. Beta diversity of the ACM. Between-sample similarities were estimated on 1,400 sequences per sample using the phylogeny-based UniFrac distance metric. Weighted UniFrac is sensitive to the sequence abundances. The *A. thaliana* ecotype Col (nonshaded red) was used in the greenhouse experiments.

4.4.3 Naturally grown *A. thaliana* and *C. hirsuta* host a taxonomically narrow root microbiota

In a second step, we investigated the variation in root microbiota composition between the plant species *A. thaliana* and *C. hirsuta* from both natural-site experiments. We compared the root bacterial communities using ANOVA-based statistics to detect taxonomic groups of OTUs ('community modules') and individual OTUs ('community members') that differ quantitatively between sites and/or host species (Supporting Material). The community member analysis was performed on the ACM, and, for the community module analyses, we prepared abundance matrices at phylum and family rank of all ACM OTUs representing 9 and 51 taxa, respectively.

We searched the abundance matrices at phylum and family rank for modules that differ between the root communities as a function of the variables site and host species (Supporting Material). The taxonomic structure of the root communities varies mainly by site (six phyla, 35 families) and less by host species (two phyla, two families; ANOVA, $P < 0.1$) At both sites, *A. thaliana* and *C. hirsuta* root communities displayed similar relative distributions of bacterial phyla, except for an increased abundance of Bacteroidetes in *C. hirsuta* at the Eifel site (Figure S8; Tukey, $P < 0.1$). The single dominant Flavobacterium OTU mentioned earlier (OTU162362) was more abundant in *C.hirsuta* compared with *A. thaliana* root communities. We conclude that *A. thaliana* and *C. hirsuta* root microbiota consist of similar community modules and that the root communities differ quantitatively by their biogeography.

Next, we identified individual community members that differ quantitatively between the two host species. For this analysis, we initially defined for both natural sites the 'RootOTUs', which represent 70 OTUs that are enriched in *A. thaliana* or *C. hirsuta* roots compared with the corresponding soil communities (Tukey, $P < 0.1$; Figure 4.3; Figure S9). Spearman rank correlation coefficients of the RootOTU communities between these two hosts are 0.89 and 0.74 for the Cologne and Eifel sites, respectively, indicating an overall similar RootOTU composition. The RootOTUs in root communities vary mainly by the variable site (50 of the 70 RootOTUs) followed by host species (18 RootOTUs; ANOVA, $P < 0.1$) The comparison of RootOTUs between sites revealed a taxonomically narrow and shared set of 14 RootOTUs, consisting of seven Actinomycetales, three Burkholderiales, three Flavobacteriales, and a Myxococcales OTU Figure S10). These shared RootOTUs were validated by parametric Tukey and nonparametric Mann-Whitney and Bayesian statistics (Supporting Material). and represent in their abundance half of the community. At the Eifel site, quantitative differences between the two plant species were found in 9 of the 70 RootOTU members, where 7 RootOTUs were more abundant in A. thaliana compared with *C. hirsuta* and 2 RootOTUs were less abundant (Tukey, $P < 0.1$; Figure S11). This finding and the few aforementioned host species-differentiating community modules point to the existence of largely shared bacterial root communities with similar relative abundances in *A. thaliana* and *C. hirsuta*.

Previous studies revealed a weak rhizosphere effect for A. thaliana (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012). To quantify the rhizosphere effect, we determined for both sites the OTUs that are enriched in the rhizosphere of *A. thaliana* or *C. hirsuta* compared with the corresponding soil (termed 'RhizoOTUs'). Similar to these studies, we detected only few RhizoOTUs at the Cologne site (Tukey, $P < 0.1$; Figure S12). The occurrence of a rhizosphere effect was found to be site-dependent: 6 (*A. thaliana*) and 11 RhizoOTUs (*C. hirsuta*) discriminated the rhizosphere from soil communities at the Cologne site whereas no RhizoOTUs (both host species) were found at the Eifel site. We conclude that the magnitude of the rhizosphere effect is site-dependent but independent of the tested host species.

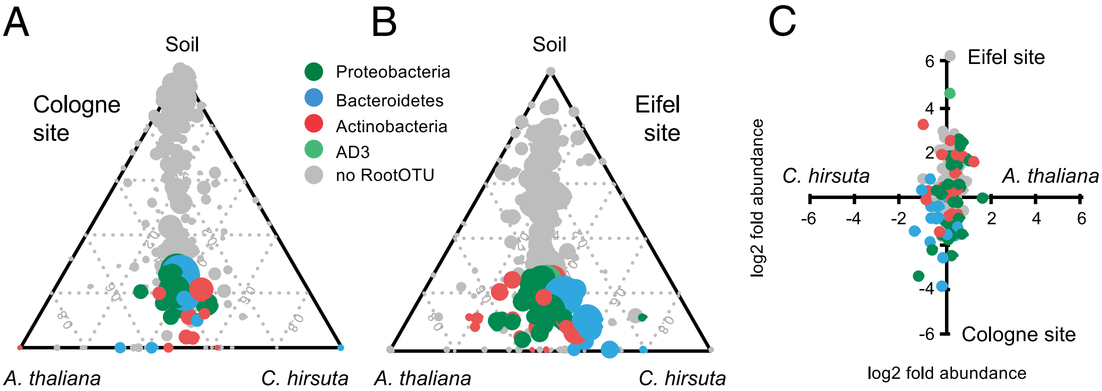
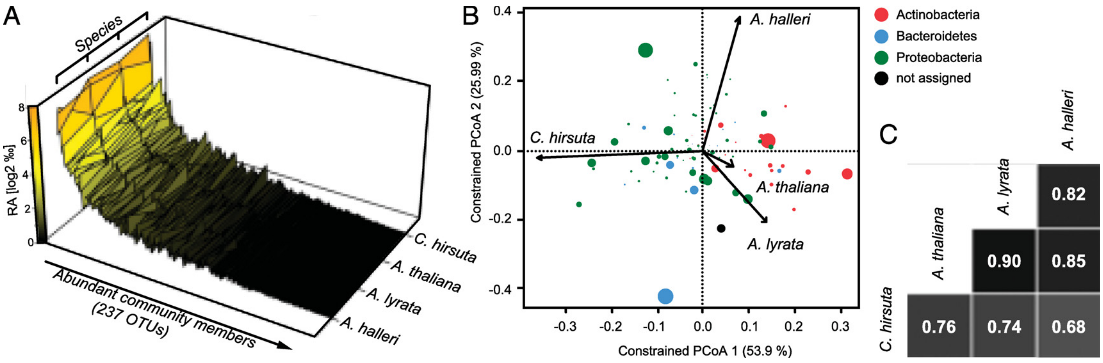


Figure 4.3: Root microbiota comparisons of *A. thaliana* and *C. hirsuta* at the natural sites Cologne and Eifel. The ternary plots depict the relative occurrence of individual OTUs (circles) in root samples of *A. thaliana* and *C. hirsuta* compared with the respective soil samples for the Cologne (A) and the Eifel site (B). RootOTUs (OTUs enriched in roots compared with soil; Tukey, $P < 0.1$) are colored by taxonomy, and OTUs, which are not enriched in root communities, are plotted in gray. The size of the circles is proportional to the mean abundance in the community. (C) Variation in mean relative abundance (RA) of individual OTUs (circles) across species and sites, where axes depict logtwofold variation.

4.4.4 Phylogenetic distance of host species contributes to microbiota diversification

Next, we examined the bacterial root communities retrieved from the *A. thaliana* and the relative species *A. halleri*, *A. lyrata*, and *C. hirsuta* grown under controlled environmental conditions in replicated greenhouse experiments. A similar overall rank abundance profile of the ACMs in root communities between these four hosts reveals qualitatively similar community structures, indicating that variation in root microbiota is largely quantitative (Figure 4.4A). We used ANOVA-based statistics to detect community modules and members that vary in abundance between the tested host species (Supporting Material). A few taxonomic modules differed in relative abundance between the root microbiota of the tested plant species (Figure S13), exemplified by significantly lower Bacteroidetes levels in *A. halleri* (Tukey, $P < 0.1$) This phenotype was again due to the differential abundance of the dominant Flavobacterium mentioned above (OTU162362). At family rank, *A. halleri* and *A. lyrata* display a species-specific quantitative reduction of Flavobacteriaceae and Oxalobacteriaceae, respectively (Figure S13).

Analogous to the community member analysis of the natural-site experiments, we identified 76 RootOTUs enriched in roots of at least one plant species compared with soil (Tukey, $P < 0.1$) Analogous to the community member analysis of the natural-site experiments, we (Figure S14). We then examined the between-sample (beta diversity) variation in the composition of RootOTUs among the host species using canonical analysis of principal coordinates (CAP) (26). CAP analysis constrained for the variable species revealed that 17% of the variation in beta diversity, as measured by Bray-Curtis distance metric, was explained by the host species (Figure S15; $P < 0.005$; 95% confidence interval = 12%, 25%). The samples clustered by host species and distances between host species revealed that the root communities of *A. thaliana* were more similar to *A. lyrata* than to *A. halleri* and that the root microbiota of the three *Arabidopsis* species are more similar to each other than to the root microbiota of *C. hirsuta*. Thus, within the genus *Arabidopsis* (*A. thaliana*, *A. halleri*, and *A. lyrata*), microbiota diversification is incongruent with the phylogenetic distance of these hosts (compare Figure 4.1 and Figure S15) Further exploration of the CAP analysis revealed a correspondence between the taxonomy of the RootOTUs and their contribution to the microbial diversity between host species: RootOTUs of the phylum Actinobacteria showed the strongest influence on the variation between the Arabidopsis species and *C. hirsuta* root communities (Figure 4.4B). Similarly, the abundance of Bacteroidetes largely explains the differentiation between *A. halleri* and the other host species.



	A. thaliana	A. lyrata	A. halleri
A. lyrata			0.82
A. thaliana		0.90	0.85
C. hirsuta	0.76	0.74	0.68

Figure 4.4: Root microbiota comparisons of *A. halleri*, *A. lyrata*, *A. thaliana*, and *C. hirsuta*. (A) The mean abundance of individual OTUs (both replicate experiments) was calculated for the indicated species and plotted ranked by average OTU abundance across all species. (B) OTU scores of principal coordinate analysis of the RootOTU community, constrained by host species and based on Bray-Curtis distances among root samples. The arrows point to the centroid of the constrained factor. Circle sizes correspond to relative abundances of RootOTUs, and colors are assigned to different phyla. The percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by host species. (C) Pairwise Spearman rank correlation analysis of the RootOTU communities between the indicated species.

Community similarities were confirmed by pairwise correlation analysis of the RootOTU communities between the four host species, revealing Spearman rank coefficients ranging from 0.68 to 0.90 (Figure 4.4C). The RootOTU composition of *A. thaliana* correlated best with each of its sister species *A. halleri* and *A. lyrata*, and all three pair-wise comparisons of *C. hirsuta* with the *Arabidopsis* species showed low correlation coefficients, suggesting that the evolutionarily most ancient plant species hosts a RootOTU community, which is quantitatively most diversified. Thus, inclusion of the more distant *C. hirsuta* suggests that phylogenetic distance of host species contributes to microbiota diversification across all four tested hosts. These observations were supported by the highest number of species-specific RootOTU accumulation patterns for *C. hirsuta* (Tukey, $P < 0.1$; Figure S16) In total, we identified 14 species-specific RootOTUs that consisted of 1, 2, 4, and 7 RootOTUs for *A. thaliana*, *A. halleri*, *A. lyrata*, and *C. hirsuta*, respectively The lower accumulation of the *A. halleri*-specific Flavobacterium (OTU162362) and the Oxalobacteriaceae member (OTU91279) in *A. lyrata* contributed to the species-specific accumulation of the corresponding community modules Similarly, the Actinocorallia RootOTU (OTU97580) contributed to the trend of lower accumulation of the Thermomonosporaceae, a distinctive feature of *C. hirsuta* We independently validated the lower accumulation of Thermononosporaceae in *C. hirsuta* using quantitative PCR with taxon-specific PCR primers Figure S17).

We assessed variation in microbiota composition between and within host species by a direct comparison of the three *A. thaliana* ecotypes with the three Arabidopsis sister species. Ternary plots revealed a larger spread of abundant OTUs between the sister species than between the *A. thaliana* ecotypes (Figure S18). This observation is supported by the identification of 13 host species-dependent OTUs and one host genotype-dependent OTU (ANOVA, P < 0.1). This direct comparison demonstrates a greater inter- compared with intraspecies variation in microbiota composition.

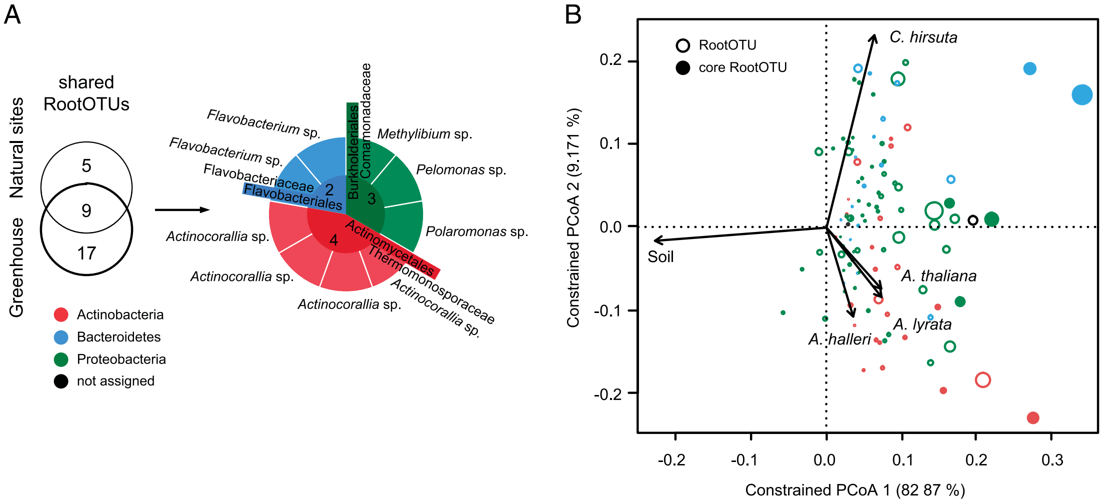


Figure 4.5: Identification of the *Arabidopsis thaliana* and relative species core microbiota. (A) Core members result from the intersection of the shared RootOTUs found at the natural sites and the shared RootOTUs detected in the greenhouse experiments. The pie chart segments are colored by the bacterial phyla of the corresponding taxa. The taxonomic assignments of the core RootOTUs are reported at order and family rank in the center and the first ring of the pie chart, respectively. The genera of the core members are noted at the periphery of the pie chart. (B) OTU scores of principal coordinate analysis of the RootOTU community, constrained by sample groups and based on Bray-Curtis distances among soil and root samples. Sample groups include root samples by species and soil samples. The arrows point to the centroid of the constrained factor. Circle sizes correspond to relative abundances, colors are assigned to different phyla, and core members are marked as solid circles. The percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by host species.

4.4.5 Members of the Actinomycetales, Burkholderiales, and Flavobacteriales are stable across host species and environments

We identified in the controlled environment experiments 26 RootOTUs shared among the four tested host species (Figure S19). These shared RootOTUs belonged to the orders Burkholderiales (11 RootOTUs), Actinomycetales (7), Rhizobiales (3), Flavobacteriales (2), Myxococcales (1), Xanthomonadales (1), and Herpetosiphonales (1). These OTUs were validated by parametric Tukey and nonparametric Mann-Whitney and Bayesian statistics and constituted by their relative abundance the bulk of the root community (~75%). Remarkably, the most abundant orders were also recovered in the shared RootOTUs in the natural-site experiments. The intersection of RootOTUs shared between plant species found at natural sites and in greenhouse experiments determined the core microbiota (Figure 4.5A and Figure S20). This core consisted of nine RootOTUs assigned to the orders Actinomycetales (four RootOTUs, genus Actinocorallia), Burkholderiales (three, family Comamonadaceae), and Flavobacteriales (two, genus Flavobacterium; Figure 4.5A) This core represented a taxonomically extremely reduced subcommunity of the microbiota in all tested host species, and together these RootOTUs constituted by their abundance up to half of the root microbiota in all samples tested (Figure S20A). The enrichment of these core microbiota members relative to soil across plant species and sites was identified by three statistical methods and confirmed by a subsampling technique (i.e., bootstrapping) (Figure S20B). In addition, bootstrapping predicted OTUs of the orders Rhizobiales and Myxococcales to be part of the core microbiota. However, the abundance of the latter two orders was less stable between environments, and, therefore, they did not pass the stringent identification of significant RootOTUs in the original data set using three different statistical methods (Figure 4.5). Taken together, the core RootOTUs found across all host species and sites belonged to only three bacterial orders: the Actinomycetales, Burkholderiales, and Flavobacteriales. We compared the composition of this core microbiota to *A. thaliana* root endophyte communities from previous studies (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Bodenhausen *et al.*, 2013), which were based on different soil types, environments, and PCR primer combinations (Figure S21). The raw 16S rRNA gene sequences of these studies were coclustered with the sequences of this study, and the common OTU table was examined for the core microbiota in each data subset using the statistical procedure of this study A common core at OTU level cannot be confirmed in other *A. thaliana* root microbiome studies (Figure S21C). whereas, at order rank, the presence of Actinomycetales presents the common denominator across all studies.

Burkholderiales and Flavobacteriales were detected in three and two of the four studies, respectively. Additionally, Rhizobiales and Sphingomonadales were each detected once as core members.

Using CAP analysis, we finally investigated the contribution of the core RootOTUs to the overall variation in root – compared with soil samples in all experimental conditions. Therefore, we constrained the analysis for all sample groups, i.e., root samples by species and the soil samples as additional group (Figure S22 and Figure 4.5). Consistent with the unconstrained beta diversity analysis (Figure 4.2), the compartment constituted the major source of variation We observed a clear differentiation between soil and root samples along the first principal coordinate, which explained the largest fraction of the variation (82.87%). We noted that the core RootOTUs (filled circles in Figure 4.5B) –having the largest species descriptors– contributed most to the formation of the ordination space. This observation was consistent with their definition (enriched in root samples) and identification in all experimental conditions. Importantly, we confirmed the correspondence between the taxonomy of the RootOTUs and root microbiota diversity across host species over all experimental conditions: the root microbiota of *Arabidopsis* species were distinguished by RootOTUs of the phylum Actinobacteria (open and closed red circles in Figure 4.5B) whereas root bacterial communities of *C. hirsuta* were differentiated by Bacteroidetes (open and closed blue circles in Figure 4.5B). We interpreted these correspondences as evidence of a host impact on the root microbiota at a high taxonomic rank.

4.5 Discussion

4.5.1 A conserved core root microbiota?

Here, we have examined the bacterial root microbiota of *A. thaliana* along with its sister species *A. lyrata* and *A. halleri* and of *C. hirsuta*. This study revealed the existence of a core root microbiota comprising members from the three bacterial orders Actinomycetales, Burkholderiales, and Flavobacteriales (Figure 4.5A). Previous studies (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Bodenhausen *et al.*, 2013), using different 16S rRNA PCR primer combinations, reported that *A. thaliana* roots host mainly Actinobacteria, Bacteroidetes, and Proteobacteria. This taxonomic structure at phylum level is congruent with the core composition described here because the order Actinomycetales belongs to the phylum Actinobacteria, the Burkholderiales belongs to the subphylum Betaproteobacteria, and the Flavobacteriales to the phylum Bacteroidetes. Bootstrapping also identified the core members and revealed additional RootOTUs, expanding the core composition These Rhizobiales and Myxococcales members became apparent in approximately half of random subsets of the original data. Enhanced variation in their abundance between replicate samples and tested environments could explain their absence from the core microbiota.

The significance of our definition of the core microbiota is potentially constrained by the PCR primer used and a low number of tested environments (Cologne and Eifel natural sites and controlled environment). It remains to be seen whether additional samples, also from extreme environments, modify the composition of the core. Corroborating evidence for its stability in additional environments comes from a recent *A. thaliana* field study comprising four disturbed sites in the United States using the same PCR primer combination (Bodenhausen *et al.*, 2013). In these root samples, Actinomycetales, Burkholderiales, and Flavobacteriales were found by 16S rRNA gene pyrosequencing among other prevalent taxa, and the taxonomic composition of the core at order level is similar to our study (Figure S22C). Differences in the selectivity of different 16S PCR primers and variation in 16S rRNA gene copy number likely distort the composition of the core root microbiota. The comparison across *A. thaliana* root microbiome studies Previous studies (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Bodenhausen *et al.*, 2013) did not reveal a common core at the (Figure S22C). However, we cannot discriminate whether PCR primer bias, the soil type/start inoculum, or combinations thereof account for the lack of a common OTU core. Despite this lack of clarity, we noted that, at higher taxonomic rank, the enrichment of members of the Actinomycetales in roots was a common feature of all *A. thaliana* root microbiome studies. Actinobacteria, including the Actinomycetales, appear to be enriched from soil by cues from living A. thaliana roots (Bulgarelli *et al.*, 2012). Our study suggests that such host-derived assembly signal(s) are evolutionarily conserved, at least in the Brassicaceae. Future examination of root microbiomes from additional plant species in the same environments and using the same PCR primer combination will test whether this core is a lineage-specific innovation of the Brassicaceae family.

The core root microbiota members accounted for up to half of the total community size. The core consisted of both abundant and low-abundant community members, suggesting that assembly and physiological function(s) depend on selective membership and regulation of their relative abundance. In addition, we found a correspondence between the taxonomy of the bacterial communities and the diversity pattern of the root microbiota across host species and three different environments (Figure 4.5B). The core members largely supported this correspondence. These observations, together with the reduced taxonomic complexity of the core community, point to the existence of a common organizing principle for their establishment. We speculate about two potential and mutually not exclusive mechanisms that take part in the establishment process: (i) each bacterial lineage autonomously responds to host-derived cues and (ii) microbe-microbe interactions enable a selective advantage for cocolonization by core members. For example, the commensal relationship between Bacillus cereus and bacteria of the Cytophaga-Flavobacterium (CF) group in the soybean rhizosphere is mediated by peptidoglycan, which is produced by B. cereus and stimulates the growth of CF bacteria (Peterson *et al.*, 2006). From future root-microbiota metagenome and -transcriptome analyses, we expect insights into the connectivity among microbes (Shade and Handelsman, 2012). Such approaches will define the core microbiota at the level of genes rather than taxonomic lineages and will provide a deeper understanding of host services of the core as pioneered by human gut microbiota research (Turnbaugh *et al.*, 2009; Qin *et al.*, 2010).

Members of each taxon of the core are potentially beneficial for their hosts. Grassland rhizosphere-derived cultured strains belonging to the Burkholderiales were shown to have antagonistic activities toward multiple soilborne oomycete and fungal pathogens (Benitez and Gardener, 2009). Wheat rhizosphere-derived Comamonadaceae strains appear to be functional specialists in soil sulfonate transformation as part of the biogeochemical sulfur cycle, likely enabling mineralization of organic sulfur for sulfate acquisition by high-affinity sulfate transporters at the root surface (Schmalenberger *et al.*, 2008; Yoshimoto *et al.*, 2002). *Flavobacterium*, a common soil and water bacterium, was positively correlated with potato biomass and frequently isolated from barley and bell pepper rhizospheres (Manter *et al.*, 2010; Johansen *et al.*, 2009; Kolton *et al.*, 2012). A reference *Flavobacterium* isolated from the rhizosphere of the latter plant tested positive for several biochemical assays associated with plant growth promotion and pathogen protection, and accessible genomes of soil/rhizosphere-derived Flavobacteria indicate that these bacteria define a distinct clade compared with Flavobacteria from aquatic environments (Kolton *et al.*, 2012). Finally, root-derived isolates of the Thermomonosporaceae and other core members such as Polaromonas isolates are capable of fixing atmospheric nitrogen (Valdes *et al.*, 2005; Hanson *et al.*, 2012), potentially increasing the amount of bioavailable nitrogen for plant growth.

4.5.2 Root microbiota interactions with the environment

If the aforementioned physiological function(s) of the core hold true for isolates present in Brassicaceae roots, then these functions could present a standing reservoir of retrievable host services independently of environmental parameters and host species-specific niche adaptations (e.g., metal tolerance of *A. halleri*) or life history traits (perennialism of *A. lyrata* and *A. halleri*). Although the conserved core is established in both controlled and natural environments, the composition of the entire root microbiota depends most on interactions with the environment (Figure 4.2). This dependence is consistent with prior findings that soil type is the key determinant of root community structure (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Bodenhausen *et al.*, 2013). Thus, the whole root microbiota consists of two parts, the conserved core and an environment-responsive subcommunity. For example, members of the order Rhizobiales represent a root community module found only as shared RootOTUs in the controlled environment using experimental Cologne soil In addition, the relative abundance of the family Oxalobacteriaceae is environment-responsive resulting in a change in rank abundance from rank 2 in the greenhouse experiments to rank 3 at the natural sites. Likewise, dominant Streptomycetaceae in the controlled environment (rank 3) are a low-abundant taxon at the natural sites (rank 14). Thus, the relative abundance of these taxa is tunable by the environment. We speculate that these environment-responsive community modules provide services to all tested host species in an environment-dependent manner.

The strong responsiveness of the root microbiota to the environment might be explained by the fact that soil does not only define the start inoculum but also the 'diet' for plants, e.g., bioavailability of macro- and micronutrients. Diet as a major driver for community structure was previously reported in microbiota hosted by other eukaryotes. For example, the mammalian gut microbiota follows primarily the dietary habits of the animals, where communities from herbivores, carnivores, and omnivores clustered clearly apart from each other (Ley *et al.*, 2008a). Dietary patterns in humans appear to determine gut microbiota enterotypes (Arumugam *et al.*, 2011; Wu *et al.*, 2011). However, it remains to be examined whether dietary effects independent of the soil start inoculum are sufficiently strong to provoke consistent shifts in root microbiota community composition. A further exploration of this question would require the modulation of individual nutrients or their composition in the same soil type/start inoculum and subsequent determination of possible effects on community structure.

4.5.3 Host microbiota co-evolution

A systematic investigation of host-microbiota diversification within a phylogenetically defined plant species framework, combined with replicated experiments under controlled conditions, has not been reported before. A major finding of our work is that the diversification of the root microbiota of the tested host species is largely quantitative. This conclusion is based on abundant community members (ACM, >20 sequences per OTU in at least one sample of the dataset), and, therefore, qualitative differences might exist in the rare biosphere, which is currently not quantifiable. Despite an overall inter-host species microbiota similarity, we found host species-specific community modules and members. The host species-specific community modules are clearly part of the environment-responsive subcommunity, as illustrated by the observation that they are site-dependent (Figure 4.3). Both the most divergent root microbiota (Figure 4.4 and Figure 4.5) and the highest number of species-specific community members were found in the phylogenetically most distant *C. hirsuta* (Figure 4.1), suggesting that phylogenetic distance of the hosts could contribute to microbiota diversification. Future studies using additional plant lineages are required to conclude whether phylogenetic distance time correlates with microbiota diversification. For example, in primates, the branching order of host-species phylogeny was found to be congruent with gut community composition (Ochman *et al.*, 2010). The greatest similarities in root microbiota were found between *A. thaliana* and *A. lyrata* whereas the root microbiota of *A. halleri* was more dissimilar (Figure 4.4), demonstrating that, within the genus Arabidopsis (*A. thaliana*, *A. halleri*, and *A. lyrata*), microbiota diversification is incongruent with the phylogenetic distances of these hosts (Figure 4.1). The two species *A. thaliana* and *A. lyrata* occur in similar habitats whereas *A. halleri* has evolved a distinctive lifestyle enabling growth in mesic sites and tolerance to high-competition habitats. This particularity could imply that the recent speciation event of *A. halleri*, coupled to an adaptation to a distinctive habitat, resulted in the selection of a distinctive microbiota with habitat-specific services. Taken together, both host species-specific ecological adaptation and phylogenetic distance might have driven microbiota diversification among the tested hosts. Whether the proportion of species-specific community members/modules increases when the host species are exposed to stressful conditions where a plant species has an adaptive advantage (e.g., tolerance of *A. halleri* to metalliferous soils) (Kraemer, 2010) remains to be tested. Similarly, it will be interesting to examine whether the proportion of species-specific community members/modules increases when the perennials *A. lyrata* and *A. halleri* are grown according to their lifestyle for longer than 1 year. Finally, future experimentation using synthetic bacterial communities with isolates of the core microbiota members and gnotobiotic *Arabidopsis* plants will directly test whether their presumed beneficial roles in plant growth and health can be reproduced under laboratory conditions and are retrievable by the host under normal and stressful conditions.

4.6 Materials and methods

We collected roots of naturally occurring *A. thaliana* and *C. hirsuta* growing side by side at the two replicate sites Cologne and Eifel. Additionally, we sampled in two replicate experiments roots of *A. thaliana* and the relative species *A. lyrata*, *A. halleri*, and *C. hirsuta*, which were grown under controlled conditions in the greenhouse in pots containing natural microbe-rich soil. We used a root-sampling protocol similar to Bulgarelli et al. (Bulgarelli *et al.*, 2012) to examine the root-inhabiting bacterial microbiota. For comparison, we also sampled bulk soil and rhizosphere compartments. Bacterial communities were characterized by pyrosequencing 16S rRNA gene amplicons derived from the PCR primers 799F (Chelius and Triplett, 2001) and 1193R (Bodenhausen *et al.*, 2013). The pyrosequencing reads were processed and analyzed with the software QIIME (Caporaso *et al.*, 2010), and custom R scripts were used for statistical analyses. For details, see (Supporting Material).

4.7 Author contributions

K.S. and P.S.-L. designed research; K.S., N.D., R.G.O., and E.V.L.v.T. performed research; K.S., N.D., R.G.O., E.V.L.v.T., and P.S.-L. analyzed data; and K.S. and P.S.-L. wrote the paper.

4.8 Acknowledgements

We thank Dr. Bruno HÃ¼ttel and Dr. Kurt StÃ¼ber (Max Planck Genome Centre Cologne) for excellent support in amplicon library preparation and pyrosequencing. We thank Dr. Yang Bai and Dr. Davide Bulgarelli (Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research) for critical comments on the manuscript. Dr. Pierre Saumitou-Laprade (Laboratoire de Giénétique et Évolution des Populations Végétales, Formations de Recherche en Évolution, Centre National de la Recherche Scientifique 3268, UniversitÃ© de Lille) kindly provided seeds of *A. halleri*.

K.S. was supported by Swiss National Science Foundation Grants PBFRP3-133544 and PA00P3-136473. This work was supported by funds to P.S.-L. from the Max Planck Society (M.IF.A. ZUCH8048), a European Research Council advanced grant (ROOT-MICROBIOTA), and the 'Cluster of Excellence on Plant Sciences' program funded by the Deutsche Forschungsgemeinschaft.

4.9 Supporting material

The supporting material corresponding to this section, including all supplementary tables and figures can be accessed via the online version of the published article and have not been included in this thesis due to space limitations. All intermediate data as well as the scripts used to analyze the data and generate the figures of this study are available at `http://www.mpipz.mpg.de/R_scripts`.

CHAPTER 5

Structure and functions of the bacterial root microbiota in wild and domesticated barley

Status	**Published**
Journal	*Cell Host & Microbe* (Impact factor 12.328)
Citation	Bulgarelli, D. *****, Garrido-Oter, R. *****, Muench, P.C., Weimann, A., Droege, J., Pan, Y., McHardy, A.C. †, Schulze-Lefert, P. † (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host & Microbe*, **17**, 392-403. ***** joint first authors; † joint corresponding authors
URL	http://dx.doi.org/10.1016/j.chom.2015.01.011
Own contribution	Analyzed the data (with co-authors) Interpreted the data (with co-authors) Wrote the manuscript (with co-authors)

5.1 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.

5.2 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 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.

5.3 Results

5.3.1 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 (5.5). 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 (5.5).

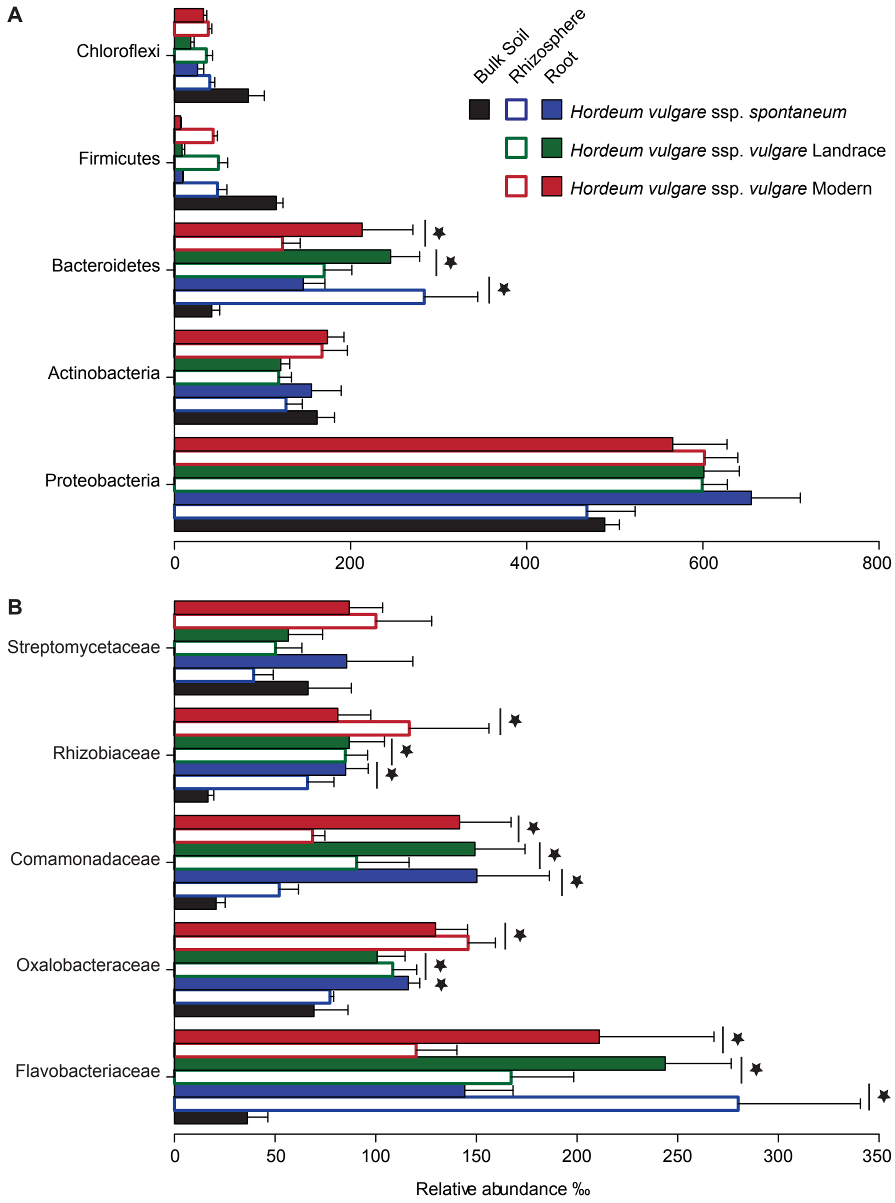


Figure 5.1: The barley rhizosphere and root microbiota are gated communities. Average relative abundance (RA ± SEM) of the five most abundant (A) phyla and (B) families in soil, rhizosphere, and root samples as revealed by the 16S rRNA gene ribotyping. For each sample type, the number of replicates is 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.

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 5.1). 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, p value < 0.05; Figure 5.1). 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 5.1). 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 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 5.2A).

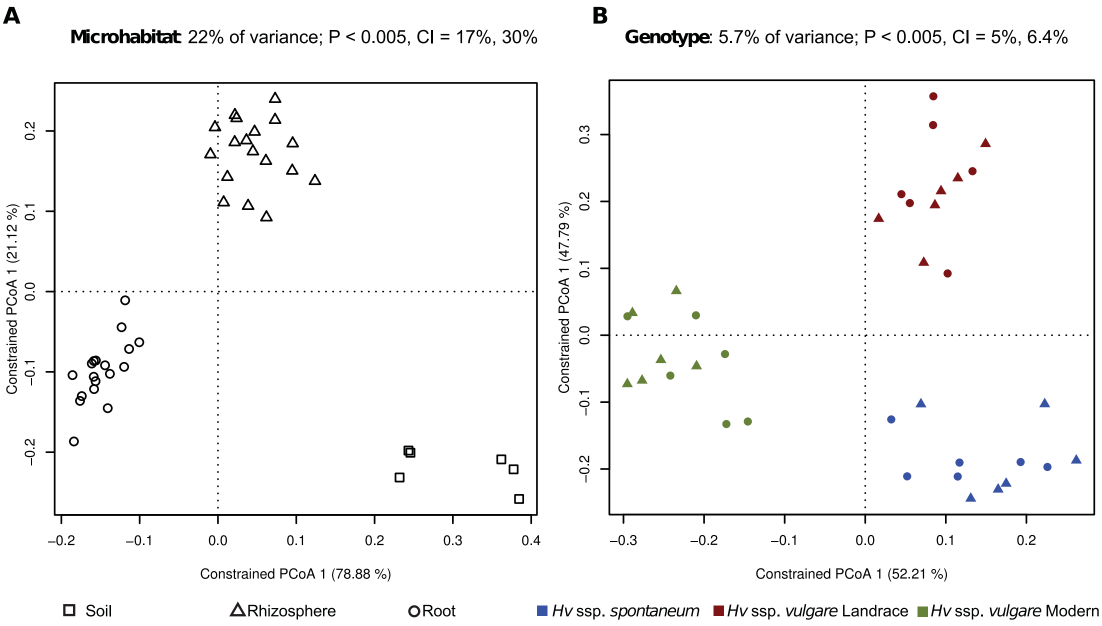


Figure 5.2: Constrained PCoA on the soil and barley bacterial microbiota. (A) Variation between samples in Bray-Curtis distances constrained by microhabitat (22% of the overall variance; p < 5.00E-2) and (B) by accession (5.7% of the overall variance; p < 5.00E-2). In both panels, triangles correspond 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.

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 5.2B). This proportion of the variation, albeit small, was found significant by permutation-based ANOVA (p < 0.005; Figure 5.2). 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 S1A). Bootstrapping analysis of the constrained ordination (5.5) 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 S1C). 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 S1B). 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 S1B). 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.

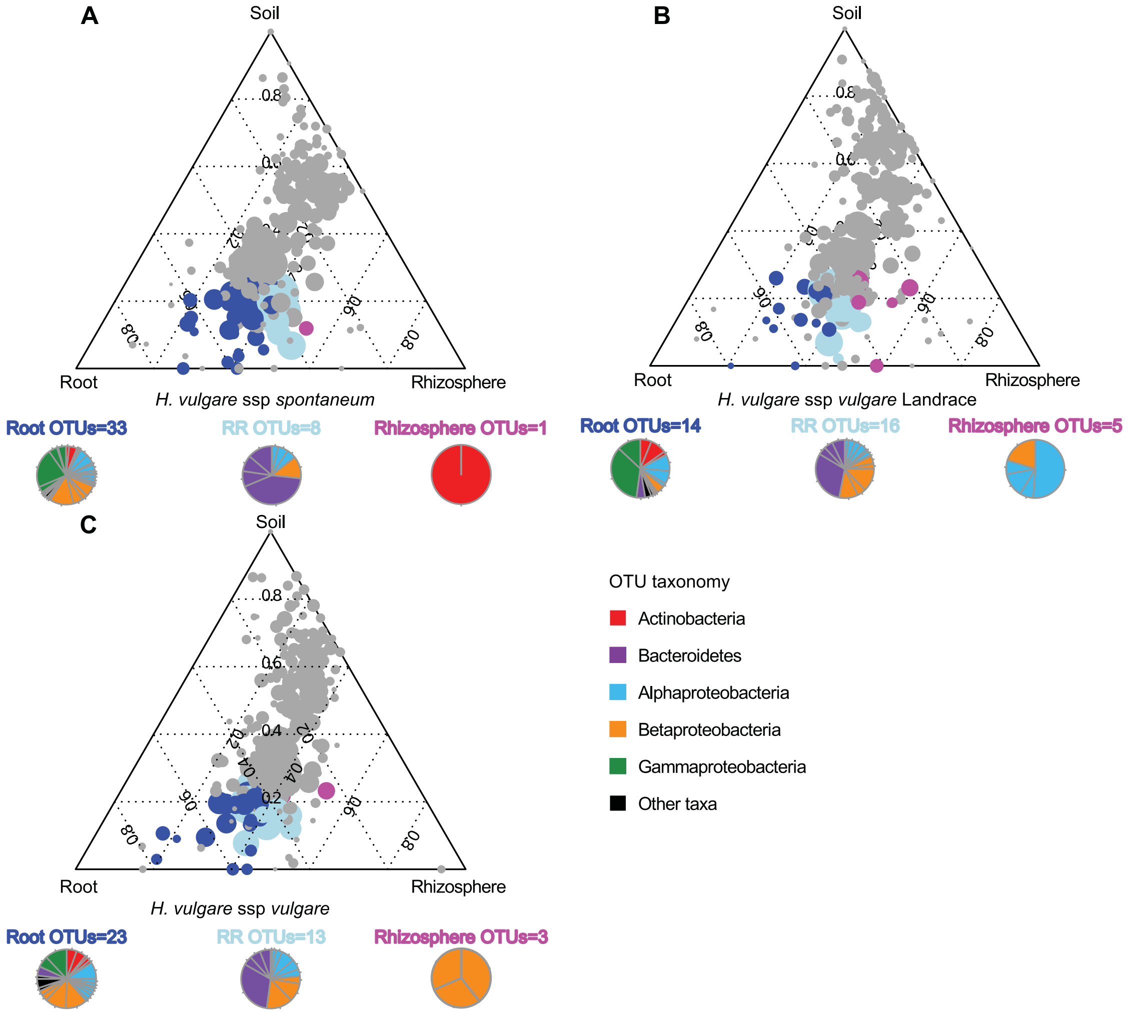


Figure 5.3: Barley OTU enrichment analysis. (Caption on next page).

Figure 5.3: Barley OTU enrichment analysis. (Figure on previous page). 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, $p < 0.05$).

To identify bacteria responsible for the diversification between the two root-associated microhabitats we employed a linear model analysis (Supporting Material). 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 at the root-soil interface Figure 5.3). 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. 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. 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. 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. 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 5.4) 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.

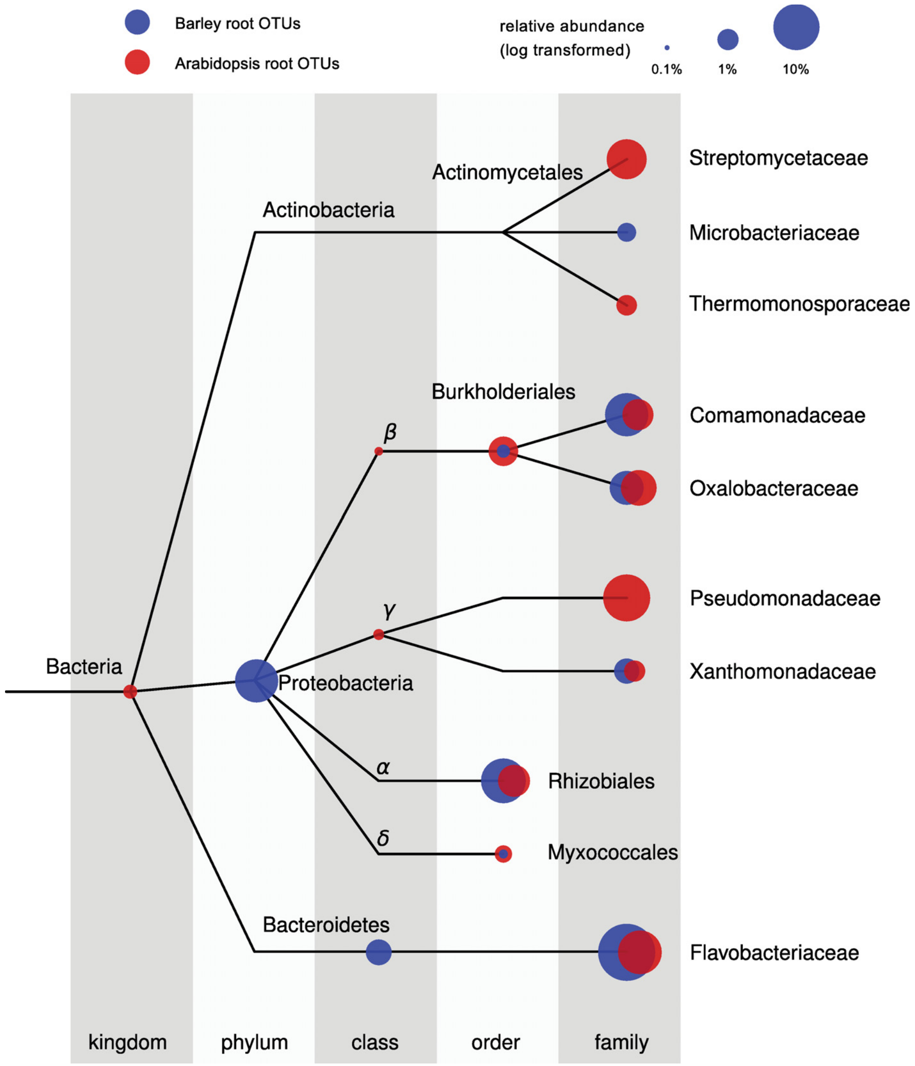


Figure 5.4: Taxonomic representation of the barley and *Arabidopsis* root-enriched bacterial taxa. The tree represents a subset of the NCBI taxonomy containing all OTUs found to be enriched in the barley and Arabidopsis root samples with respect to soil. The branches of the tree do not reflect evolutionary 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%).

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).

5.3.2 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 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); 5.5. 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 (Droege *et al.*, 2015), a tool for the taxonomic assignment of shotgun metagenomes (5.5). 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.

5.3.3 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 5.5).

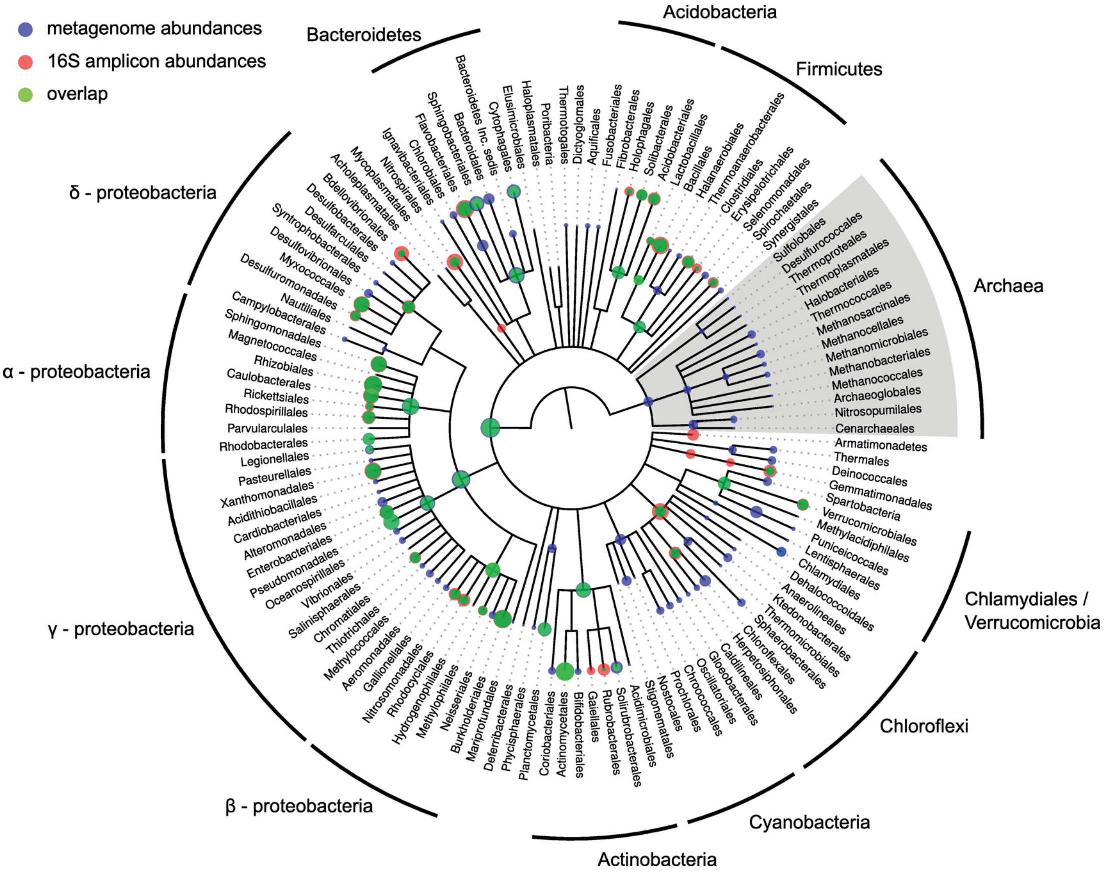


Figure 5.5: Comparison of 16S rRNA amplicon and metagenome abundances. The tree represents the NCBI taxonomy for all taxonomically classified OTUs from the rhizosphere samples of the 16S rRNA survey as well as all metagenome bins, resolved down to the order rank. The branches of the tree do not reflect 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.

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 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 5.5; 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).

Functional Category	p Value
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

Table 5.1: Functional categories significantly enriched in taxonomic bins corresponding to RR_OTUs found in the barley rhizosphere metagenome. Differentially abundances were determined using a Mann-Whitney test, controlling for false discovery rate (FDR).

5.3.4 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 (5.5) 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 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 5.1; 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 (5.5; Supporting Material) 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 ($p = 0.044$).

5.3.5 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 (Dn) and synonymous (Ds) changes, a proxy for evolutionary pressure. Our analyses showed that 9% of the gene families had on average significantly higher Dn values and lower Ds 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 5.1; Furthermore, we determined that 10.66% (115) of protein families encoded by the barley metagenome displayed a Dn/Ds ratio significantly greater than the metagenome mean Dn/Ds 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). 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 (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).

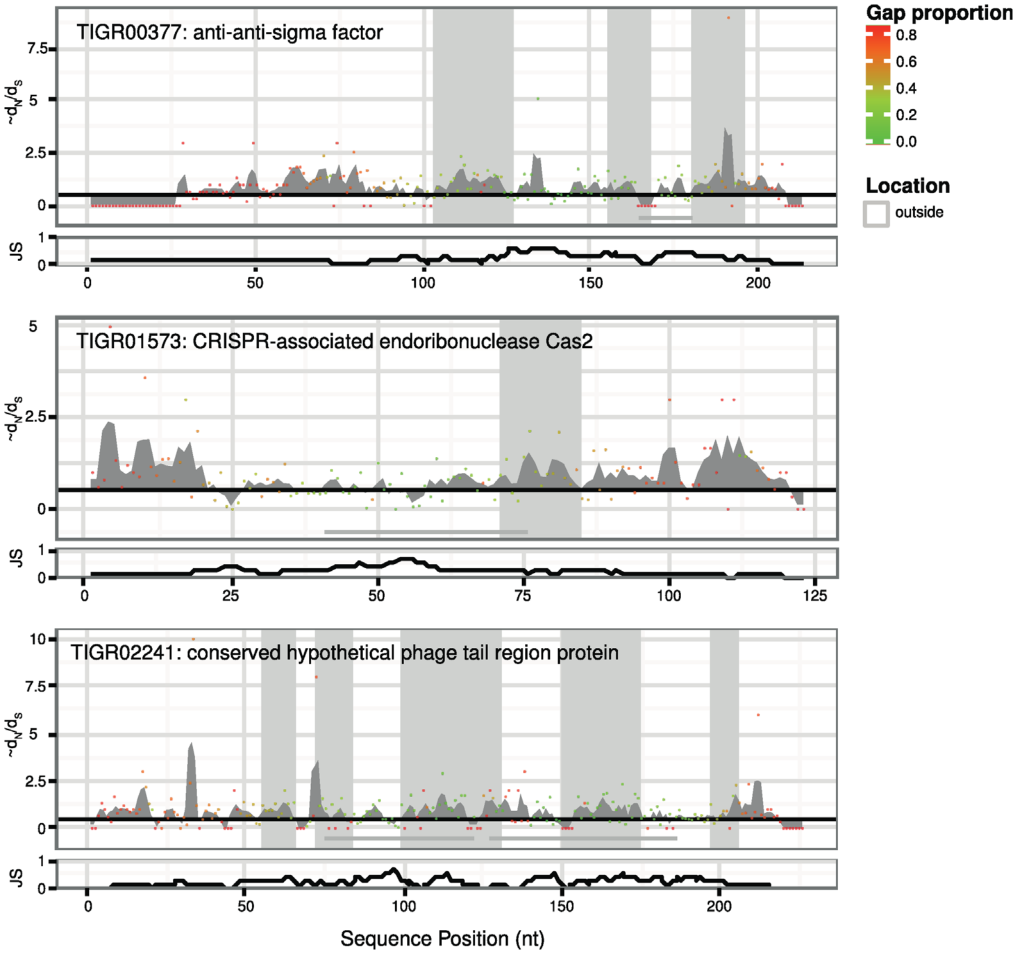


Figure 5.6: Proteins under selection in the barley rhizosphere microbiome. Sequence clusters of residues under positive selection in selected protein families. Top: dots indicate ∼Dn/Ds for a given position in the protein sequence, and their color corresponds to the proportion of gaps in the multiple 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.

5.3.6 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 5.6; 5.5). 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 under selection in the barley rhizosphere microbiome. 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 Dn/Ds 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. 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).

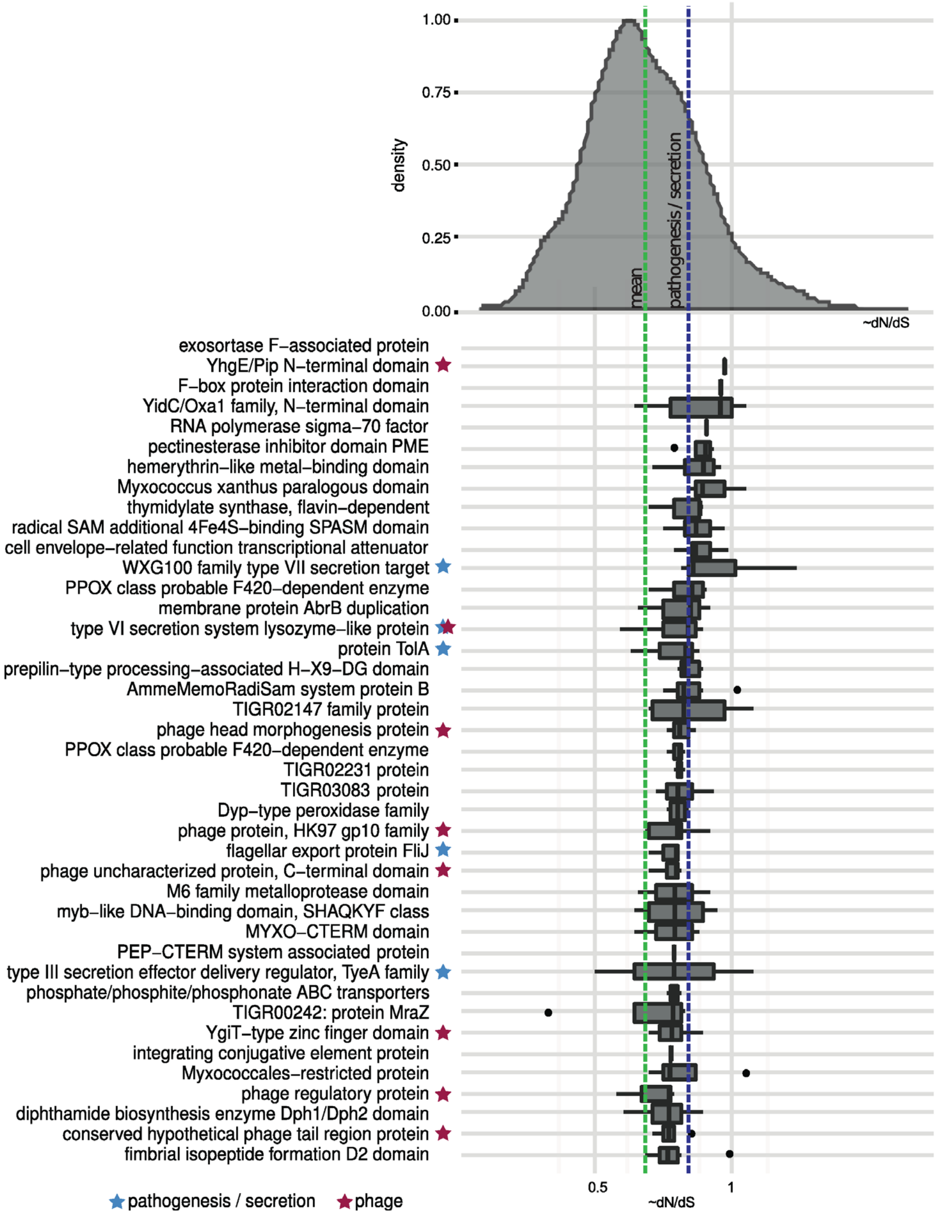


Figure 5.7: Proteins under selection in the barley rhizosphere microbiome (Cont.). Top-ranking protein families under positive selection with significantly increased Dn/Ds statistic. The distribution at the top shows the density function over all protein families smoothed with a Gaussian kernel function. The green bar indicates the average ∼Dn/Ds over all the samples, the blue bar the average ∼Dn/Ds for all TIGFRAMS annotated with the term 'patho' and/or 'secretion'. The boxplot shows the distribution of the ∼Dn/Ds across all samples for the top 50 ranked TIGRFAM families under positive selection, with families sorted by their median ∼Dn/Ds in descending order. TIGRFAMs annotated with 'repeat' or with a mean repetitive value of more than 50% were discarded.

5.3.7 Positive selection acting on phages and CRISPR systems

Interestingly, in our Dn/Ds 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 5.7). The enrichment of functional categories related to interactions with bacterial phages in RR_OTUs (Table 5.1) 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.

5.4 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 5.3), 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 5.4). 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 5.2B). 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 5.2B) 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; $p < 1.75E\text{-}12$) 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 5.1) 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 5.1) 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 5.6 and Figure 5.7). 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 5.4) 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.

5.5 Experimental procedures

5.5.1 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. 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.

5.5.2 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 (Supporting Material).

5.5.3 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 (Supporting Material). We used taxator-tk (Droege *et al.*, 2015) 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 Ds and Dn 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.

5.6 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.

5.7 Acknowledgements

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. StÃ©phane 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.

5.8 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`.

5.9 Supporting material

The supporting material corresponding to this section, including all supplementary tables and figures can be accessed via the online version of the published article and have not been included in this thesis due to space limitations. All intermediate data as well as the scripts used to analyze the data and generate the figures of this study are available at `http://www.mpipz.mpg.de/R_scripts`.

CHAPTER 6

Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities

Status	**Published**
Journal	*PNAS* (Impact factor 9.423)
Citation	Zgadzaj, R. *, Garrido-Oter, R. *, Jensen, D.B., Koprivova, A., Schulze-Lefert, P. † and Radutoiu, S. † (2016). Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities *Proceedings of the National Academy of Sciences*, In press. * joint first authors; † joint corresponding authors
Own contribution	Performed the computational experiments Analyzed the data Interpreted the data (with co-authors) Wrote the manuscript (with co-authors)

6.1 Abstract

Lotus japonicus has been employed for decades as a model legume to study the establishment of binary symbiotic relationships with nitrogen-fixing rhizobia that trigger root nodule organogenesis for bacterial accommodation. Using community profiling of 16S rRNA gene amplicons we reveal that in *Lotus* distinctive nodule- and root-inhabiting communities are established by parallel rather than consecutive selection of bacteria from the rhizosphere and root compartments. Comparative analyses of WT and symbiotic mutants in Nod factor receptor5 *nfr5*, Nodule inception *nin* and Lotus histidine kinase1 *lhk1*, identified a previously unsuspected role of the nodulation pathway in the establishment of different bacterial assemblages in the root and rhizosphere. We found that the loss of nitrogen-fixing symbiosis dramatically alters community structure in the latter two compartments, affecting at least 14 bacterial orders. Our findings imply a role of the legume host in selecting a broad taxonomic range of root-associated bacteria that, in addition to rhizobia, likely contribute to plant growth and ecological performance.

6.2 Significance

Legumes are known as pioneer plants colonizing marginal soils, and as enhancers of the nutritional status in cultivated soils. This beneficial activity has been explained by their capacity to engage in symbiotic relationship with nitrogen-fixing rhizobia. We performed a community profiling analysis of *Lotus japonicus* wild type and mutants to investigate the role of nitrogen fixing symbiosis on the structure of the root-associated bacterial microbiota. We found that several bacterial orders were almost entirely depleted from the mutant roots, and that an intact symbiosis is needed for the establishment of taxonomically diverse and distinctive bacterial communities in root and rhizosphere. Our findings imply that a symbiosis-linked bacterial community rather than dinitrogen-fixing rhizobia alone contribute to legume growth and ecological performance.

6.3 Introduction

The transition from aquatic to terrestrial lifestyle during plant evolution required the formation of roots as organ for water, macro- and micro-nutrient retrieval from soil. Nutrient-uptake systems of roots are usually specific for bioavailable forms of nutrients, e.g. inorganic nitrogen such as nitrate (NO_3-) or inorganic orthophosphate (Pi) (Maathuis, 2009). However, phosphorus per se is abundant in soil in plant-inaccessible pools and, likewise, atmospheric dinitrogen (N_2) is abundant in aerobic soil (78% v/v), but cannot be accessed by plants. Soil-resident microbes play important roles in the solubilization and conversion of mineral nutrients into bioavailable forms, and a subset of these microbes have acquired the capacity to engage in mutualistic interactions with plant roots to trade soil-derived bioavailable macronutrients for plant-derived photoassimilates (Fierer and Jackson, 2006; Bulgarelli *et al.*, 2013; Parniske, 2008).

It is now widely accepted that in nature all healthy, asymptomatic plants live in association with diverse microbes including bacteria, fungi, viruses, and protists, collectively called the plant microbiota (Bulgarelli *et al.*, 2013; Vorholt, 2012). The bacterial root microbiota is taxonomically structured and characterized by the co-occurrence of three main phyla comprising Actinobacteria, Bacteroidetes, and Proteobacteria across different soil types and divergent plant hosts (Hacquard *et al.*, 2015; Guttman *et al.*, 2014). This root-associated bacterial assemblage is mostly derived from the highly diverse bacterial soil biome surrounding roots and is established rapidly within few days after seed germination (Hacquard *et al.*, 2015; Edwards *et al.*, 2015). Soil type is the main driver of diversification of the bacterial root microbiota at low taxonomic ranks, e.i. at genus and species level (Edwards *et al.*, 2015; Lundberg *et al.*, 2012; Bulgarelli *et al.*, 2012; Dombrowski *et al.*, 2016). However, root exudates are thought to play an important role as cues to initiate a substrate-driven competition between and differential proliferation of soil-resident microbes for root colonization (Bulgarelli *et al.*, 2013; Eilers *et al.*, 2010). On average, 17% of photosynthetically fixed carbon is transferred to the rhizosphere, the thin layer of soil surrounding the root, through root exudation (Nguyen, 2003). These carbon substrates likely contribute to bacterial community shifts that are often detected in the rhizosphere. A fraction of the bacterial taxa present in the rhizosphere colonize roots either as epiphytes on the root surface (rhizoplane) or as bacterial endophytes inside roots (Bulgarelli *et al.*, 2013; Edwards *et al.*, 2015)). In particular, members of Proteobacteria are consistently found enriched in root and rhizosphere compartments, and diazotrophs in this phylum have evolved the capacity to establish a sophisticated form of mutualistic interaction with plant roots, designated root nodule symbiosis. Unlike the taxonomically diverse root- and rhizosphere-associated bacterial communities that comprise a network of microbe-microbe and plant-microbe associations, the root nodule symbiosis defines a highly specific binary plant-microbe interaction where the compatible nitrogen-fixing soil bacterium is selected by the host for intracellular infection often via plant-derived infection threads and subsequent accommodation and amplification inside nodule cells.

Decades of bacterial and legume genetics allowed a detailed dissection of the regulatory networks behind the stepwise symbiotic association with diazotroph alpha-Proteobacteria. A two-way signal recognition initiates the interaction. Root-secreted (iso)flavonoids are perceived by the compatible soil bacteria, which start the production and secretion of the rhizobial symbiotic signal, the Nod factor. On the host side, LysM receptor kinases, like NFR1 and NFR5 in *Lotus japonicus*, specifically recognize and bind the compatible Nod factors (Radutoiu *et al.*, 2007; Broghammer *et al.*, 2012) and initiate the symbiotic signalling cascade. *Nin* was identified as an early key regulator of both nodule organogenesis and infection thread formation (Schauser *et al.*, 1999), while cytokinin signalling (Held *et al.*, 2014) *LHK1* in particular, controls progression of the signalling events from root epidermis into the cortex (Murray *et al.*, 2007). Inside nodules, a low-oxygen, carbon-rich environment is established by the host allowing bacteria, upon endocytosis, to start the nitrogen fixation (Udvardi and Poole, 2013). Symbiotic nitrogen fixation reprograms the whole root transcriptional and metabolic landscape (El Yahyaoui *et al.*, 2004; Colebatch *et al.*, 2004; Hogslund *et al.*, 2009; Nakagawa *et al.*, 2011). Moreover, the process is reiterative and highly asynchronous, as rhizobia from the rhizosphere recapitulate the infection on newly formed, competent root hairs. Nevertheless, the legume host controls the number of infection events and nodule primordia via shoot-derived signal(s) (Krusell *et al.*, 2002; Searle *et al.*, 2003).

Symbiotic nitrogen fixation allows legumes to thrive in habitats with limited nitrogen availability (Peoples *et al.*, 2009; Batterman *et al.*, 2013; Adams *et al.*, 2016). However, the beneficial effect of this symbiosis is not limited to legume hosts, but extends to subsequent or concurrent plantings with non-legumes as exemplified by ancient agricultural practices with legume cropping sequences or intercropping systems. This likely involves a beneficial activity of legume roots and their associated microbes on the nutritional status of the soil as well as the soil biome. However, the mechanisms underpinning these symbiotic interactions in a community context and their impact on the complex microbial assemblages associated with roots remain largely unknown. Integrating these highly specific binary interactions into an ecological community context is critical for understanding the evolution of symbiosis and efficient use of rhizobia inoculum in agricultural systems.

Here, we investigated the role of symbiotic nitrogen fixation on the structure of the root-associated bacterial microbiota of the model legume *Lotus japonicus*. We performed bacterial 16S rRNA gene-based community profiling experiments of WT plants, grown in natural soil, and symbiotic mutants impaired at different stages of the symbiotic process. We have found that an intact nitrogen-fixing symbiosis in WT *Lotus* plants is needed outside of nodules for the establishment of taxonomically diverse and distinctive bacterial communities in root and rhizosphere compartments, raising the possibility that the influence of legumes on soil performance in agricultural and ecological contexts is mediated by the enrichment of a symbiosis-linked bacterial community rather than dinitrogen-fixing rhizobia alone.

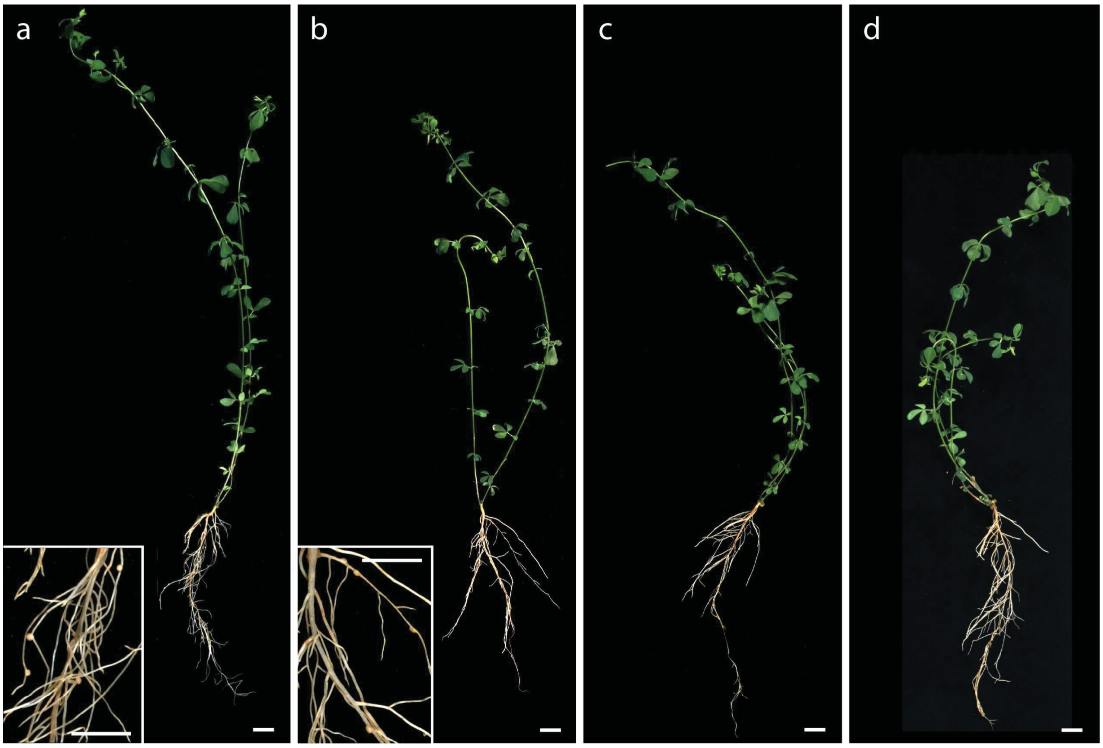


Figure 6.1: Images depicting *L. japonicus* WT and mutant plants. Images depicting *Lotus japonicus* wild-type (a) and nodule symbiosis-deficient mutant plants: *lhk1*-1 (b), *nfr5*-3 (c), *nin*-2 (d) following harvest. For nodulating genotypes (a and b), insets present close-up view of nodules. Scale bars correspond to 1cm.

6.4 Results

6.4.1 Characterization of the *Lotus japonicus* root, nodule and rhizosphere microbiota

We established a root fractionation protocol for 10 week-old *Lotus japonicus* plants (accession Gifu, designated wild-type, or WT), grown in three batches of natural Cologne soil to account for batch-to-batch and seasonal variation at the soil sampling site (Figure 6.1A; 6.6). This fractionation enabled us to compare the structure of bacterial communities present in nodules, roots without nodules (denoted hereafter 'root compartment'), rhizosphere and unplanted soil (Supporting Figure 6.9; 6.6). Briefly, the 'rhizosphere compartment' defines soil particles tightly adhering to *Lotus* roots that were collected after the first of two successive washing steps. Macroscopically visible nodules and nodule initials were excised from roots with a scalpel and designated 'nodule compartment'. Pooled nodules and washed roots without nodules were separately subjected to a sonication treatment to deplete epiphytes and enrich for endophytic bacteria. Abundant nodulation (approx. 20 nodules per plant) of healthy WT plants demonstrates that this soil is conducive for nodule formation and contains *Lotus* compatible rhizobia (inset, Figure 6.1A). We subjected a total of 27 unplanted soil, 73 rhizosphere, 75 root and 27 nodule samples to amplification of the 16S rRNA gene with PCR primers targeting the V5-V7 hypervariable regions (Schlaeppi *et al.*, 2014) (6.6). and generated ∼1M high-quality sequencing reads (4,670 reads per sample, in average; Supplementary Data 1). After removal of low-quality reads, chimeras or sequences assigned to plant mitochondria, we clustered the data into 1,834 Operational Taxonomic Units (OTUs) at 97% sequence similarity (6.6).

To assess the effect of the different compartments on the assembly of bacterial communities, we compared the beta-diversity (between samples diversity) using Bray-Curtis distances and performed a Canonical Analysis of Principal oordinates (Anderson and Willis, 2003) (CAP; 6.6). This revealed a clear differentiation of samples belonging to the root, rhizosphere, nodule and soil compartments that explains as much as 19.97% of the overall variance of the data (Supporting Figure 6.10A; P<0.001), while the effect attributable to the soil batch was comparatively small (8.01% of the variance; P<0.001). Analysis of alpha-diversity (within samples diversity) using the Shannon index indicated a decreasing gradient of complexity from the soil bacterial communities (highest richness) to the rhizosphere, root and finally the nodule microbiota (Supporting Figure 6.10).

Our finding of a bacterial community shift in the *Lotus* rhizosphere compared to the bulk soil reservoir are consistent with previous reports from WT pea (Turner *et al.*, 2013), soybean (Mendes *et al.*, 2014) and peanut (Chen *et al.*, 2014), in which a similar enrichment of members of Burkholderiales, Flavobacteriales, and Rhizobiales has been shown, whereas information on the community structure of the root microbiota is unavailable for other legumes.

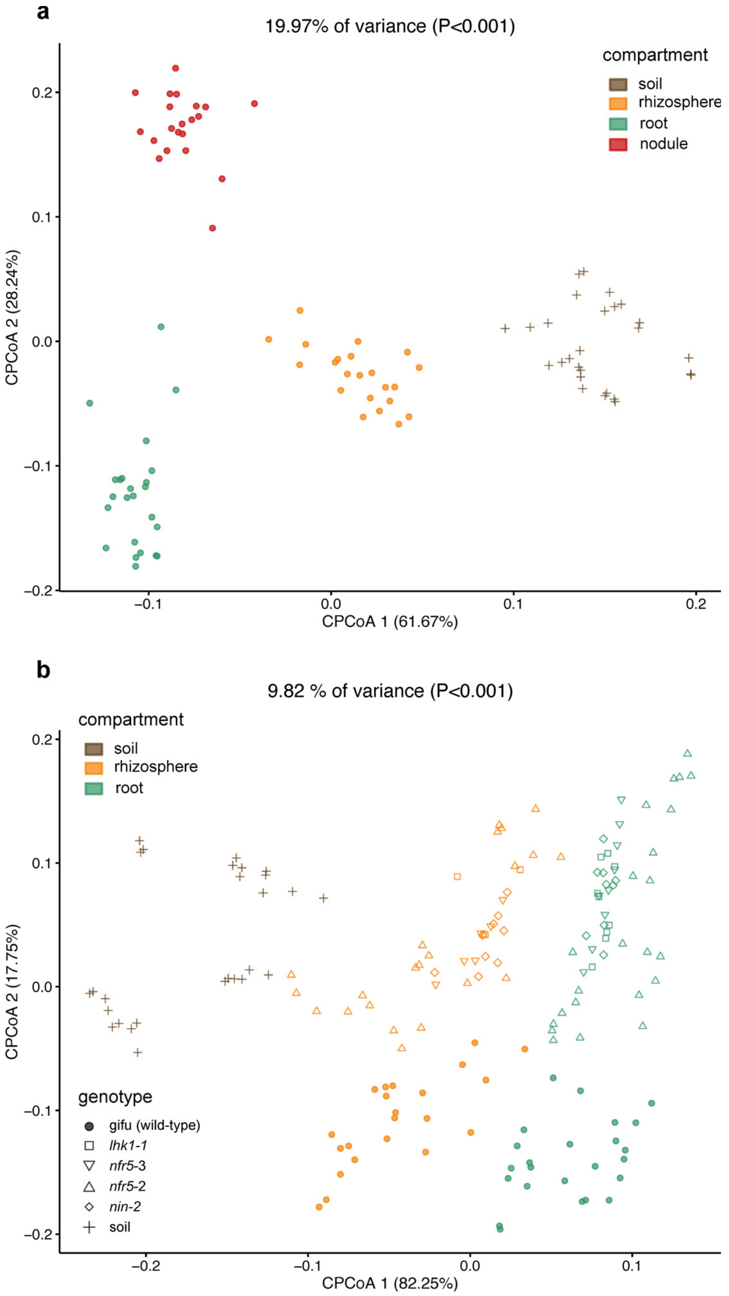


Figure 6.2: Constrained PCoA analyses of beta-diversity. (a) Constrained PCoA plot of Bray-Curtis distances between samples including only the wild-type gifu constrained by compartment (19,97% of variance, P>0.001; n=94). (b) Contrained PCoA plot of Bray-Curtis distances constrained by genotype (9.82% of variance explained, P<0.001; n=164). Each point corresponds to a different sample, colored by compartment and each host genotype is represented by a different shape. The percentage of variation indicated in each axis corresponds to the fraction of the total variance explained by the projection. Corresponding unconstrained PCoA plots for each soil batch are shown in Supporting Figure 6.11.

6.4.2 Parallel selection of nodule- and root-specific bacteria from the rhizosphere compartment

Legume nodules represent a unique environmental niche derived from differentiated cortical root cells where both symbiotic and nonsymbiotic bacteria are allowed accommodation and proliferation. Laboratory studies with single WT or mutant symbiotic strains demonstrated a step-wise, host-controlled colonization process ensuring symbiont selection (Suzaki *et al.*, 2015). By contrast, little is known about the extent or the diversity of nodule and root colonization by nonsymbionts (Zgadzaj *et al.*, 2015). We took advantage of the compatible symbiotic association between *Lotus* and rhizobia present in Cologne soil and performed an analysis of the bacterial community of epiphyte-depleted, functional nodules of WT plants grown in this soil (6.6). We found that nodules were inhabited by a distinctive bacterial community compared to those present in the root and rhizosphere (Figure 6.2A). Only a small number of the 1834 OTUs were identified to be nodule-enriched (12 red circles in Figure 6.3A) with one dominant member classified as belonging to the Mesorhizobium genus, substantiating that nodules represent a highly selective bacterial niche also for soil-grown *Lotus* plants. Taxonomic assignments at the order level for all OTUs with RA >0.5% in the nodule samples revealed a clear preferential enrichment for bacteria belonging to the order Rhizobiales (23.79% average RA; Figure 6.3B), which is mainly due to a selective enrichment of Mesorhizobium members (Figure 6.3C). Rank abundance plots revealed Burkholderiales and Actinobacteriales as abundant orders in nodules (12.11% and 10%, respectively) besides several other low-abundance bacterial orders (Figure 6.3B), showing that nodules of soil-grown *Lotus* are preferentially, but not exclusively, colonized by symbiotic rhizobia. Importantly, the nodule-enriched OTUs were found in similar abundances in the root and rhizosphere samples (Figure 6.3A). This finding strongly suggests a parallel rather than consecutive selection of these taxa from the rhizosphere assemblage for enrichment in the two endocompartments, most likely via host-induced infection threads.

6.4.3 Impairment of nitrogen-fixing symbiosis dramatically alters bacterial community structure in the *Lotus* root and rhizosphere compartments

Next, we applied the same growth conditions, fractionation protocol and bacterial community analysis to four symbiotic *Lotus* mutants (*nfr5*-2, *nfr5*-3, *nin*-2, *lhk1*-1) to identify the role of the nitrogen-fixing signaling pathway on bacterial assemblages.

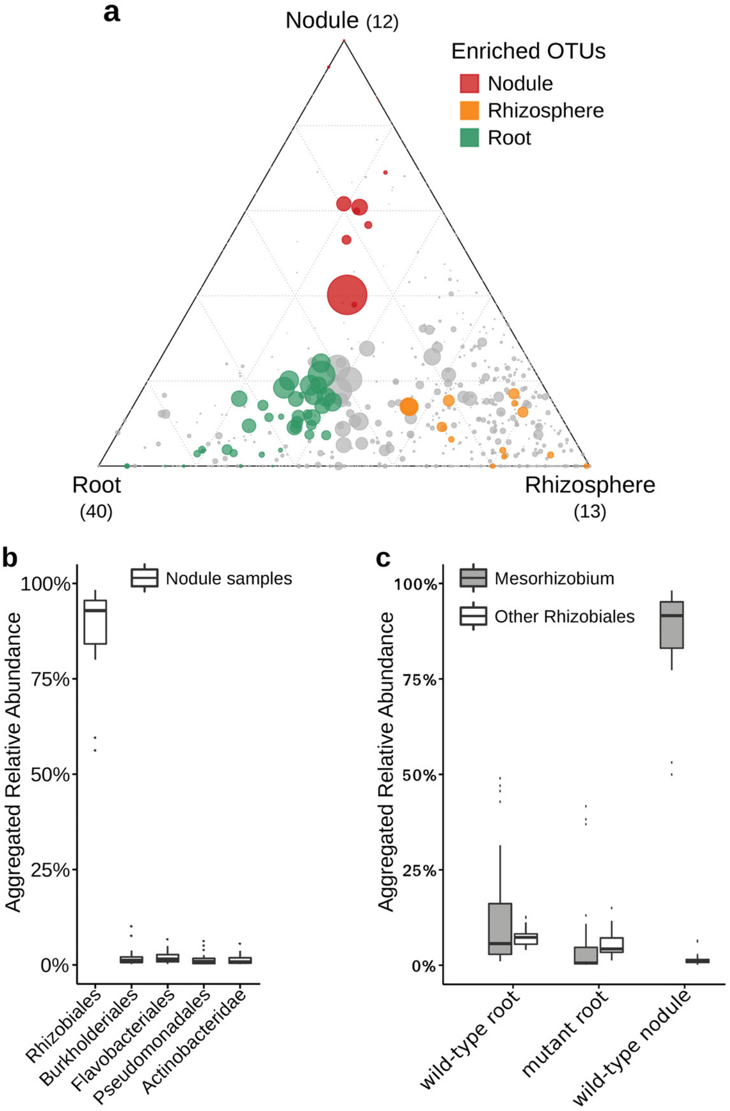


Figure 6.3: Root- rhizosphere- and nodule-enriched OTUs in wild-type *L. japonicus*. Ternary plot depicting compartment relative abundance of all OTUs (> 0.5%) for wild-type root, rhizosphere and nodule samples (a; n=67) across three soil batches (CAS8-10). Each point corresponds to an OTU. Its position represents its relative abundance with respect to each compartment and its size the average cross all three compartments. Colored circles represent OTUs enriched in one compartment compared to the others (green in root, orange in rhizosphere and red in nodule samples). (b) Rank abundance plot depicting relative abundances aggregated to the order taxonomic level for the top abundant taxa found in the wild-type nodule samples (n=21). (c) Comparison of abundances between Mesorhizobium and other Rhizobiales genera in wild-type roots (n=48), mutant roots (n=100) and wild-type nodule samples (n=21).

Similarly to WT, symbiotic mutant plants appeared healthy, but were smaller and had slightly pale green leaves (Figure 6.1B-D). With the exception of occasional nodules on *lhk1* roots, no nodules were found on *nfr5* or *nin* root systems (Figure 6.1B-D). Remarkably, we found that communities associated to the root and rhizosphere of each of the four symbiosis mutants were similar to each other, but significantly different from those of WT plants (Figure 6.2B; Supporting Figure 6.11). This separation between the mutant and WT samples was found to be robust, as indicated by unconstrained Principal Coordinate Analysis (PCoA) performed independently for each soil batch (Supporting Figure 6.11). Furthermore, CAP analysis performed on the entire dataset revealed a prominent effect of the host genotype on bacterial communities, explai*ning* 9.82% of the variance (Figure 6.2B).

Nodules are root-derived and -anchored structures, and yet the two organs host distinctive bacterial assemblages (Figure 6.2A). As a consequence, despite rigorous preparation of root compartments, WT root segments might contain incipient root-concealed nodule primordia and, vice versa, the nodule samples might be contaminated with surrounding root tissue. To clarify whether these potential limitations of our sampling protocol confound the observed host genotype-dependent community differentiation, we performed an in silico depletion of all nodule-enriched OTUs from the WT root dataset and repeated the PCoA and CAP analyses (Supporting Figure 6.12). This experiment revealed only a negligible reduction in the portion of the community variance explained by the host genotype (9.82% versus 9.72%), indicating that the differences in the root-associated assemblages caused by the impairment of nitrogen-fixing symbiosis are largely robust against residual cross-contamination between the two compartments.

To better understand how the *Lotus* nodulation pathway influences bacterial community composition we identified OTUs that are specifically enriched in the root and rhizosphere of WT or mutants compared to unplanted soil (6.6). Due to the fact that the bacterial assemblages of the tested symbiosis mutants do not significantly differ among each other (Figure 6.2B), we performed our analyses using the combined samples from all mutant genotypes across all soil batches (Figure 6.4). The *Lotus* WT root microbiota is characterized by a large number of root-enriched OTUs, mostly belonging to Proteobacteria, Actinobacteria and Bacteroidetes (105 green circles in Figure 6.4A; (Supporting Figure 6.14). By contrast, only a small number of OTUs were found specifically enriched in the WT rhizosphere samples (8 orange circles in Figure 6.4A). Compared to WT, roots of the symbiotic mutants are dramatically depleted of root-enriched OTUs (28 green circles in Figure 6.4B), whilst the number of rhizosphere-enriched OTUs increased by a factor of eight (68 orange circles in Figure

6.4B). This pattern was reproducible when we performed the same analysis for each soil batch and mutant genotype independently (data not shown).

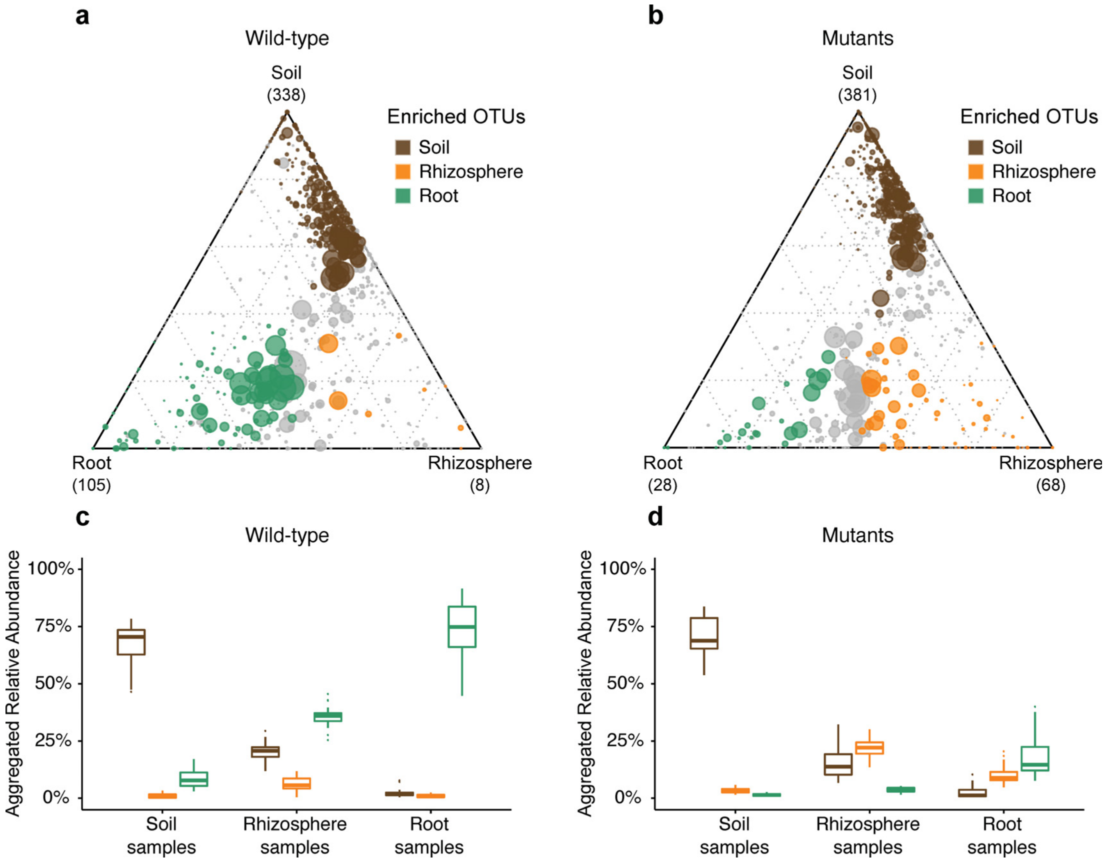


Figure 6.4: Compartment-enriched OTUs in WT and mutant *Lotus japonicus*. Ternary plots depicting compartment relative abundance of all OTUs ($> 0.5\%$) for wildtype samples (a, wild-type; n=73) and mutant samples (b, *nfr5*-2, *nfr5*-3, *nin*-2, *lhk1*-1; n=118) across three soil batches (CAS8-10). Each point corresponds to an OTU. Its position represents its relative abundance with respect to each compartment and its size the average cross all three compartments. Colored circles represent OTUs enriched in one compartment compared to the others (green in root, orange in rhizosphere and brown in root samples). Aggregated relative abundances of each group of enriched OTUs (root-, rhizosphere- and soil-enriched OTUs) in each compartment for the wildype samples (c, wild-type; n=73) and mutant samples (d, *nfr5*-2, *nfr5*-3, *nin*-2, *lhk1*-1; n=118).

To further characterize the bacterial community shifts, we calculated separately for WT and symbiotic mutant plants aggregated relative abundances (RAs) of OTUs that are specifically enriched in one compartment. As expected, this revealed a decreasing contribution of the soil-enriched OTUs in soil, rhizosphere and root samples (69.40%, 17.03% and 2.40% mean aggregated RA, respectively; dark brown boxplots in Figure 6.4C-D). in both, WT and mutant samples. This suggests that impairment of the symbiosis pathway does not affect the capacity of *Lotus* to exclude colonization by the majority of the detectable bacterial soil biome and to form characteristic root-associated microbiota, fully differentiated from those present in bulk soil. We observed an inverse pattern for the root-enriched OTUs across the three WT compartments (green boxplots in Figure 6.4C). The steep increase in the aggregated RAs, from 8.76% in the soil, 35.72% in rhizosphere and 72.34% in roots for WT samples was almost completely abolished in the mutants (1.49%, 3.63% and 17.74%, respectively; green boxplots in Figure 6.4D). Conversely, the aggregated RAs of rhizosphere-specific OTUs is similarly low in the rhizosphere samples of WT plants compared to roots and soil (orange boxplots in Figure 6.4C). but they are significantly higher in the mutant rhizosphere samples with respect to the other compartments (3.29% average RA in soil, 22.09% in rhizosphere and 9.94% in root samples; orange boxplots in Figure 6.4D). Taken together, these data support the hypothesis that the *Lotus* symbiosis pathway is a key component for the progressive enrichment/selection of specific soil-derived OTUs and the establishment of fully differentiated microbiota in rhizosphere and root compartments.

Figure 6.5: Manhattan plots showing root- and rhizosphere-enriched OTUs in WT and mutant *Lotus*. (Figure on next page).
Manhattan plots showing root-enriched OTUs in wild-type (a) or in the mutants (b) with respect to the rhizosphere and rhizosphere-enriched OTUs in wild-type (c) or in the mutants (d) with respect to root. OTUs that are significantly enriched (also with respect to soil) are depicted as full circles. The dashed line corresponds to the FDR corrected P-value threshold of significance (α=0.05). The color of each dot represents the different taxonomic affiliation of the OTUs (order level) and their sizes to their relative abundances in their respective samples (a, gifu root samples; b, mutant root samples; c, wild-type rhizosphere samples; d, mutant rhizosphere samples). Grey boxes are used to denote the different taxonomic groups (order level).

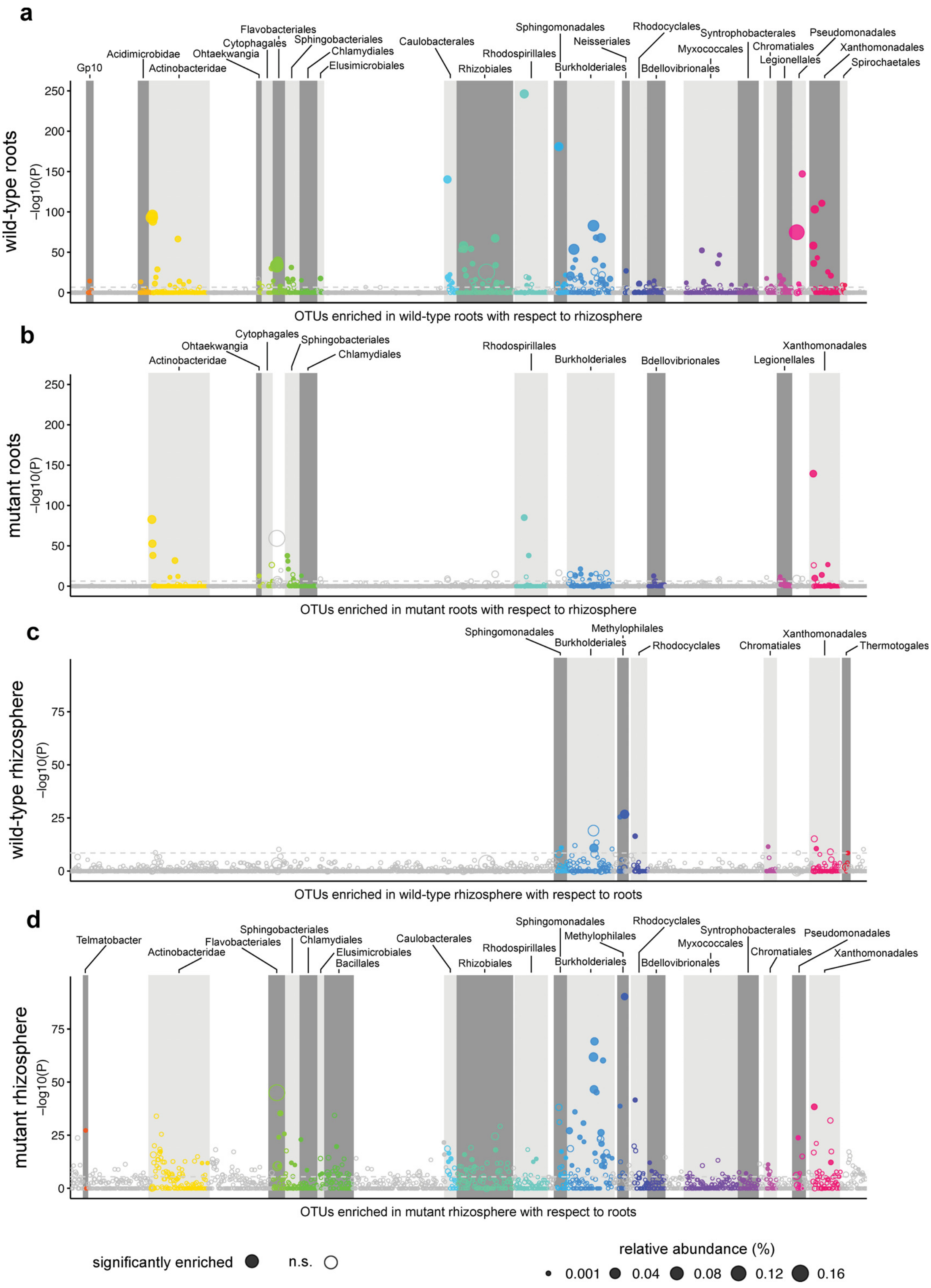


Figure 6.5: Manhattan plots showing root- and rhizosphere-enriched OTUs in WT and mutant *Lotus*. (Caption on previous page).

6.4.4 The symbiosis pathway drives root and rhizosphere differentiation across multiple bacterial orders

We dissected the observed bacterial community shifts by arranging OTUs according to their taxonomy and displaying their enrichment in root or rhizosphere of WT and symbiotic mutants in a set of Manhattan plots (6.6). The results revealed unexpectedly nuanced taxonomic alterations underlying the community shifts in the plant-associated compartments (Figure 6.5A-B). Whereas WT plants host root-enriched OTUs belonging to a wide range of bacterial orders, mutants roots fail to enrich any member of the order Flavobacteriales, Myxococcales, Pseudomonadales, Rhizobiales, and Sphingomonadales above a threshold of significance (FDR corrected P-values; α=0.05). In addition, a striking enrichment of more than 15 Burkholderiales OTUs in WT roots contrasts with a marginal enrichment of this order in the symbiotic mutants. However, the mutant roots retain the capacity to enrich OTUs belonging to the orders Actinobacteridae, Rhodospirilalles, Sphingobacteriales and Xanthomonadales (Figure 6.5A-B). Strikingly, we found an almost inverse pattern when we consider the rhizosphere-enriched OTUs in WT and mutant plants: both the number and the taxonomic diversity of significantly enriched OTUs increased dramatically in the mutants compared to WT (Figure 6.5C-D).

Next, we compared directly the WT and mutants to identify OTUs differentially abundant in root or rhizosphere (Figure 6.6A and Supporting Figure 6.13). We found that the community shift that separates host genotypes (Figure 6.2B). is largely caused by numerous OTUs that are specifically enriched (n=45) or depleted (n=15) in WT roots with respect to mutant root samples (Figure 6.6A) belonging to at least 14 bacterial orders (Supporting Figure 6.13A-B). We observed a parallel effect on OTUs of a similar taxonomic profile when comparing rhizosphere samples across genotypes, and identified numerous OTUs enriched (n=27) or depleted (n=6) in the WT rhizosphere (Figure 6.6B and Supporting Figure 6.13C-D).

Next, we examined whether the complex community shifts observed at the order taxonomic rank were also detectable at the higher phylum rank. Interestingly, we found only minor differences between the roots of WT and mutant *Lotus* plants for Actinobacteria (15.07% and 11.91% average RA, respectively) and Bacteroidetes (7.56% and 11.06% RA, respectively) in the root samples, whereas the aggregated RAs of OTUs assigned to Proteobacteria (~70%) and Firmicutes (~2%) were almost indistinguishable (Supporting Figure 6.14A). Similarly, we found no significant differences between WT and mutant rhizosphere analyzed for these four dominant phyla (Supporting Figure 6.14B).

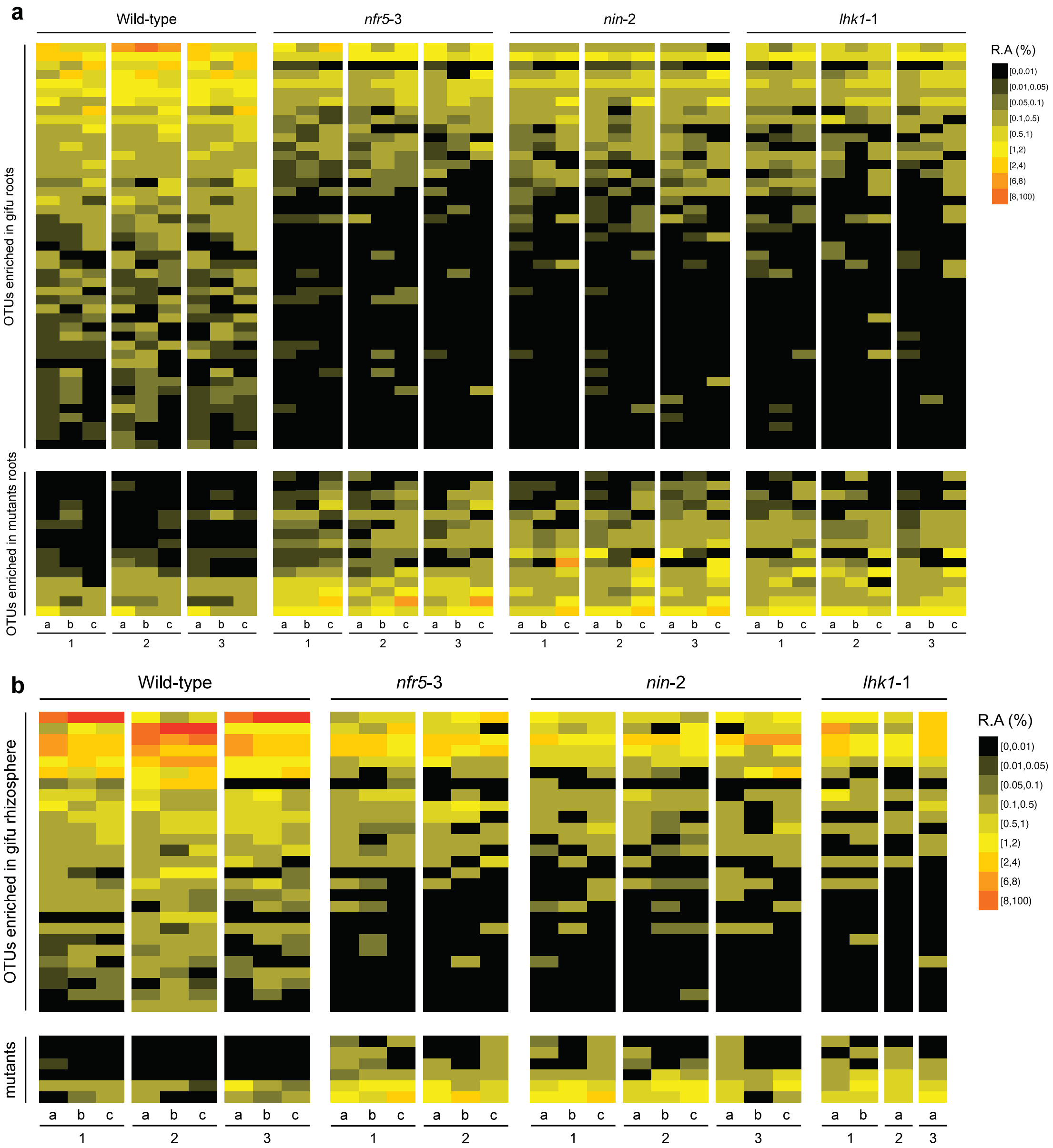


Figure 6.6: Differentially enriched OTUs between WT and mutant *Lotus*. (a) Heatmaps showing relative abundances of all OTUs found to be significantly enriched in the root samples of wild-type plants grown in CAS10 soil with respect to the symbiosis-impaired mutants (top panel) or significantly enriched in the root samples of the mutants compared to gifu (bottom panel). (b) Heatmaps showing relative abundances of all OTUs found to be significantly enriched in the rhizosphere samples of wild-type plants grown in CAS10 soil with respect to the symbiosis- impaired mutants (top panel) or significantly enriched in the rhizosphere samples of the mutants compared to gifu (bottom panel). Labels at the bottom of each column correspond to biological (numbers) and technical replicates (letters). Enrichment tests were performed using the negative binomial generalized log-linear model and P-values were FDR adjusted.

6.4.5 Comparable immune- and symbiosis-related metabolic responses in soil-grown WT and symbiotic mutant roots

The extensive changes of root microbiota structure across multiple bacterial orders in the symbiosis mutants prompted us to investigate whether mutant roots display altered immune- or symbiosis-related metabolic responses that, indirectly, perturb an orderly microbiota establishment. We quantified relative transcript levels for a panel of defense and symbiotic marker genes using WT and mutant root tissue samples that were processed as for the 16S rRNA gene community profiling. Analysis of eight genes induced during pathogen defense in Lotus or likely representing *Lotus* orthologs of *Arabidopsis* defense marker genes revealed that WT and mutant roots accumulate similar transcript levels, indicative of a comparable immune status (Supporting Figure 6.15A). We also tested whether WT and mutant roots differed in expression levels of genes that have been reported to contribute to the metabolic state established between host and nitrogen-fixing symbiont (Szczyglowski *et al.*, 1998; Hwang *et al.*, 2010; Flemetakis *et al.*, 2003; Welham *et al.*, 2009). We found comparable transcript levels of Nodulin26, Nodulin70, Sucrose transporter4, and Invertase1 in the tested genotypes, suggesting similar metabolic responses in the WT and mutants roots (Supporting Figure 6.15B). On the other hand, early symbiotic genes like *Nin*, Peroxidase and Thaumatin were induced in WT or *lhk1* and *nin*-2 mutants, but not in *nfr5*-2 roots, indicating that soil-grown symbiotic mutants maintain their previously described, gradually impaired, root response to nitrogen-fixing rhizobia (Schauser *et al.*, 1999; Murray *et al.*, 2007; Hogslund *et al.*, 2009; Madsen *et al.*, 2003) (Supporting Figure 6.15C). Direct measurements of total protein content revealed comparable levels in WT and symbiotic mutants (Supporting Figure 6.16). while quantification of nitrate levels revealed significant differences between *nfr5*, *nin* and *lhk1* or WT, indicating that regulation of nitrate uptake, which has a known inhibitory effect on nodulation (Carroll and Gresshoff, 1983; Reid *et al.*, 2016), operates downstream of *Nin* (Supporting Figure 6.16). Together, these results suggest that a nitrogen sufficient status is reached in all tested genotypes, but that the nitrogen source, N2 or nitrate, might differ among them.

6.4.6 *Lotus japonicus* and various Brassicaceae species assemble highly diverged root-inhabiting bacterial communities

We have previously shown that *Arabidopsis thaliana* and three other Brassicaceae species (*Cardamine hirsuta*, *A. halleri* and *A. lyrata*), grown in Cologne soil, assemble a highly similar root microbiota, characterized only by small quantitative differences of community profiles (Schlaeppi *et al.*, 2014). We retrieved the corresponding raw 16S sequence reads and performed de novo OTU clustering together with the amplicon data of WT and symbiotic *Lotus* mutants (Figure 6.7). PCoA of Bray-Curtis distances revealed a clear separation of root and soil samples, but also a marked distinction between all *Lotus* and Brassicaceae samples (Figure 6.7), indicating that both WT and symbiosis-impaired *Lotus* plants harbor strikingly distinctive root microbiota compared to those of the four tested Brassicaceae species.

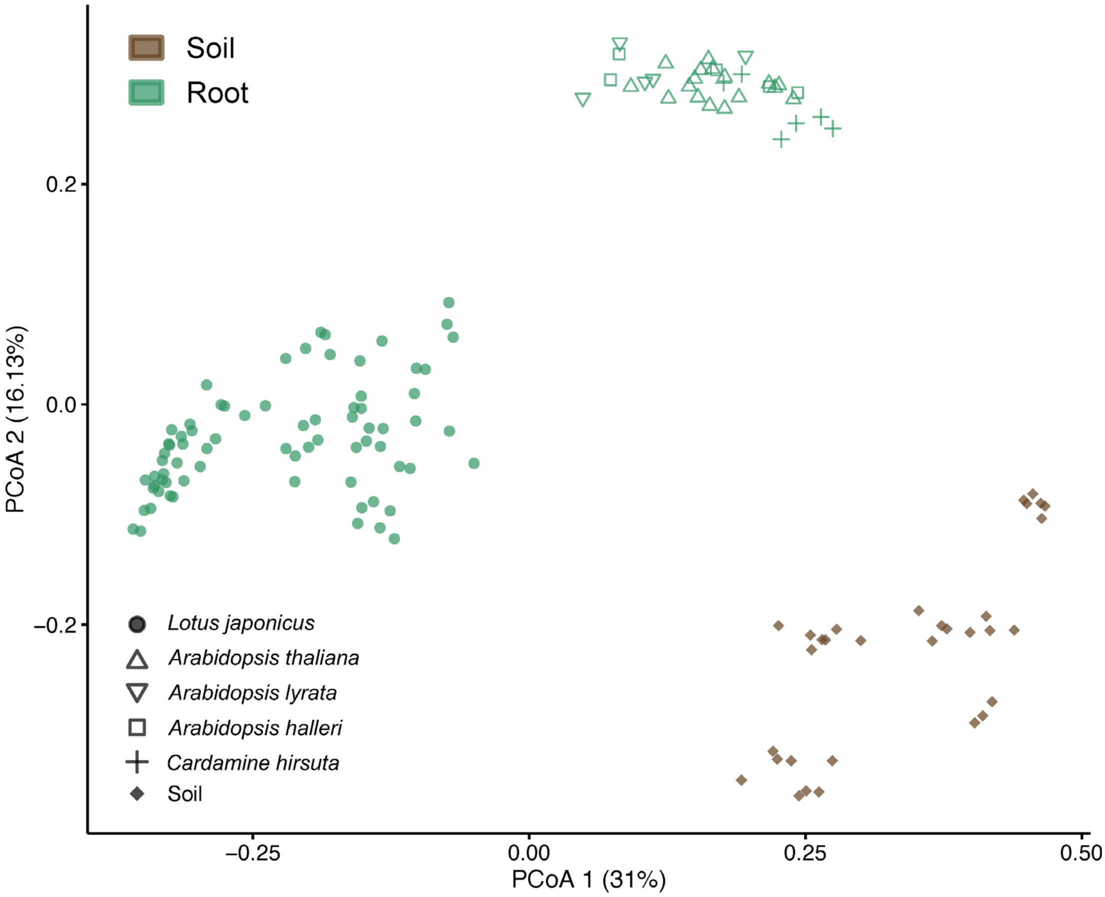


Figure 6.7: Comparison between *Arabidopsis* and *Lotus* root bacterial communities. PCoA plot of Bray-Curtis distances for root an soil samples showing a clear separation between the roots of all *Lotus* genotypes (circles) compared to those of *Arabidopsis* and relative species (hollow shapes) grown in Cologne soil and sequenced using the same primer set.

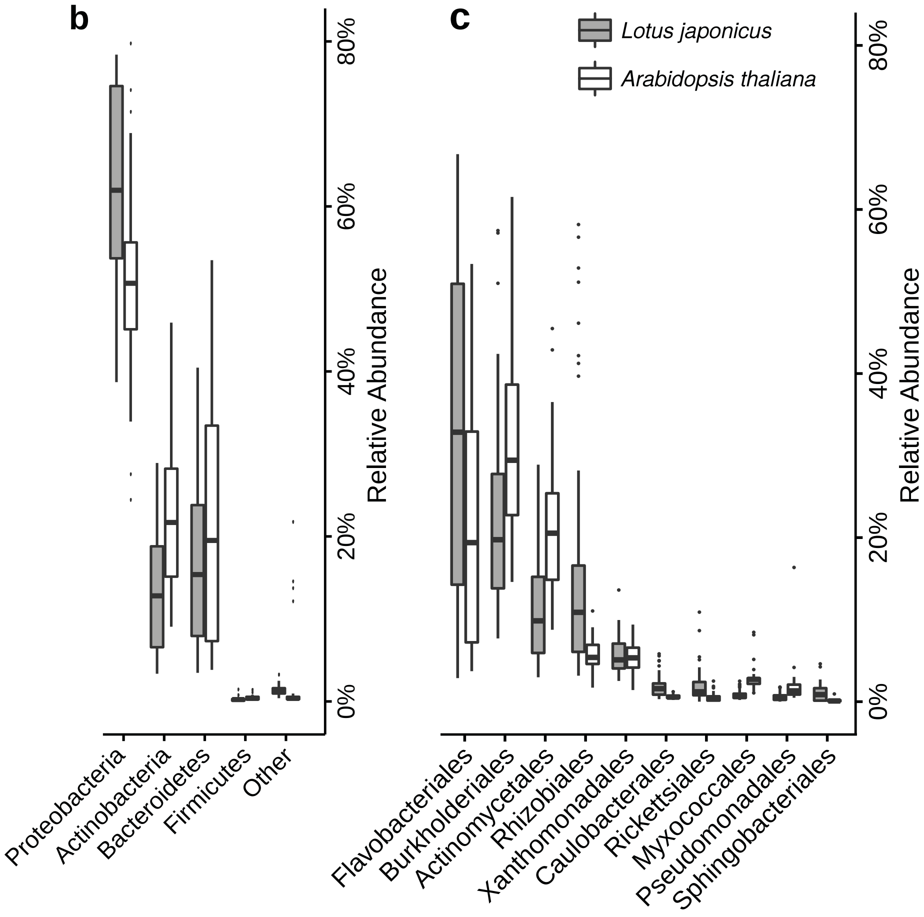


Figure 6.8: Comparison between *Arabidopsis* and *Lotus* root bacterial communities. Relative abundances aggregated to the phylum (b) and order taxonomic levels (b; 10 most abundant orders) showing a comparison between *Arabidopsis thaliana* (n=26) and *Lotus japonicus* root samples (n=74).

Quantitative analysis of WT *Lotus japonicus* and *Arabidopsis thaliana* root-enriched OTUs revealed significant and contrasting differences already at the phylum level primarily reflected in the abundances of Proteobacteria and Actinobacteria (Figure 6.8A). Similar rank abundance analysis performed at the order level identified particular taxonomic lineages contributing to the differences between these two plant species; Burkholderiales and Caulobacterales are more abundant in *Lotus* roots, while Actinomycetales, Myxococcales and Sphingomonadales have higher abundances in *Arabidopsis* roots (Figure 6.8B).

6.4.7 Symbiosis-impaired mutants maintain an altered community structure in nitrogen-supplemented soil

Nitrogen-fixing symbiosis is nitrogen-sensitive, and already from 2 mM KNO3 concentration, reduced nodulation and infection were observed in Lotus (Reid *et al.*, 2016). To determine if the community shifts observed in Lotus mutant roots and the rhizosphere were caused by a potentially differential nitrogen status, we performed a similar community analysis using plants grown in Cologne soil (different soil batch) supplemented with 10 mM KNO3 . In these conditions, the symbiotic mutants no long had a pale leaf phenotype as observed in non-supplemented soil (Supporting Figure 6.17) and the WT plants developed no functional nodules, based on their low number and small and white appearance at the time of harvest. Despite similar nitrogen content in WT and mutant roots (Supporting Figure 6.17) we found that the differential phenotypes (stature and fresh weight) seen in non-supplemented soil were retained under nitrogen-supplemented conditions (Supporting Figure 6.17) indicating that an intact symbiosis pathway promotes plant growth irrespective of the presence of functional nodules. Based on the similar macroscopic phenotypes of the symbiotic mutants in response to the nitrogen supplementation, we then analyzed the composition of bacterial communities in WT and two Nod factor receptor mutants, *nfr5*-2 and *nfr5*-3. Remarkably, the PCoA revealed a similar shift in the root and rhizosphere communities of the mutants relative to the corresponding WT compartments for plants grown in nitrogen-supplemented soil (21.20% of the variance, $P < 0.001$; Supporting Figure 6.18A) as in the nonsupplemented Cologne soil (21.80% of the variance, $P < 0.001$; (Supporting Figure 6.18B). Finally, no detectable differences in soil biome composition were seen between nitrogen-supplemented and non-supplemented unplanted soil samples (Dataset S4). Together, these results provide evidence for a direct impact of the disabled symbiosis pathway on the root-associated community structure rather than an indirect effect resulting from abolished symbiotic nitrogen fixation.

6.5 Discussion

Here, we have characterized the root microbiota of the model legume *L. japonicus* using a 16S rRNA amplicon survey. By employing a panel of nitrogen-fixing symbiosis-impaired mutants we have investigated the role of host genes with known functions in the establishment of a highly specific and binary symbiotic plant-microbe association in the context of the root-associated bacterial community. Our study reveals that key symbiotic genes play a major role in the establishment of taxonomically structured bacterial communities in the root and rhizosphere of *L. japonicus* (Figure 6.4). which extends their role beyond the perception and selection of nitrogen-fixing rhizobia for intracellular accommodation in nodules. The observed impact of disenabling symbiosis in *Lotus* seemingly differs from analogous perturbations in insects and mammals, where impairment of symbiosis in the gut leads to the accumulation of pathogens and dramatic consequences for the health of the host (Brandl *et al.*, 2008; Dessein *et al.*, 2009; Round and Mazmanian, 2009). Cologne soil-grown *Lotus* symbiotic mutants showed no signs of disease or altered immune response in comparison to WT (Figure 6.1 and Supporting Figure 6.15). However, it remains possible that the observed reduced plant growth (Figure 6.1) leads to lower reproductive fitness of the mutant plants in a natural ecosystem.

Genetic disruption of nitrogen-fixing symbiosis resulted in depletion of six bacterial orders from the root compartment, including the most abundant order identified in WT, Burkholderiales (Figure 6.5A-B and Figure 6.7B). Rhizobium acts therefore as a bacterial 'hub' for *Lotus* roots, in analogy to the Albugo oomycete pathogen that largely affected the phyllosphere microbiome of Arabidopsis after infection (Agler *et al.*, 2016). Two mutually non-exclusive scenarios could account for the observed dramatic community shift inside roots: i) cooperative microbe-microbe interactions between symbiotic nitrogen-fixing bacteria and a subset of root microbiota members or ii) direct use of the nodulation pathway by soil-borne bacteria other than nodulating rhizobia for entry and proliferation inside legume roots. Quantification of selected transcripts in WT and mutant roots provided evidence for a similar immune status and symbiosis-specific metabolic responses, but as expected, differential expression of the symbiotic genes (Supporting Figure 6.15). These results support the hypothesis of a direct rather than indirect engagement of the nitrogen-fixing symbiosis pathway in the selection of specific bacterial taxa other than nodulating rhizobia. The beneficial association with Burkholderiales is habitual for legumes. For example, Mimosa genera within South America and legumes from the Papilionoideae subfamily in South Africa are known to form an ancient and stable symbiosis with nitrogen-fixing Burkholderia lineages inside nodules (Bontemps *et al.*, 2010; Garau *et al.*, 2009; Angus and Hirsch, 2010). Members of the order Burkholderiales, which are dramatically depleted in the *Lotus* mutants, are also known for potent plant growth promotion activities in non-leguminous plants (Touceda-Gonzalez *et al.*, 2015). The inclusion of legumes in a cropping rotation sequence in agriculture generally enriches the soil, but the symbiotic nitrogen fixation alone cannot explain the productivity increase in the subsequent crop (Peoples *et al.*, 2009). Thus, our study raises the intriguing possibility that, besides the activity of nitrogen-fixing bacteria, the selective enrichment of other symbiosis-linked root microbiota members influences the soil biome and, consequently, soil bio-fertilization by legumes might involve a much wider taxonomic range than currently thought.

Our statistical analyses revealed a major impact of the host genotype on the Lotus root microbiota, which explains 9.82% of the variance of the data (Figure 6.2B). This genotype effect is almost two-fold larger than that identified when comparing the root-associated communities of WT and severely immunocompromised *A. thaliana mutants* (Lebeis *et al.*, 2015) or between *A. thaliana* and three Brassicaceae species that diverged up to 35 Mya (Schlaeppi *et al.*, 2014). These findings demonstrate that the early nodulation signaling pathway, besides its established role for the onset of nitrogen-fixing symbiosis, exerts a major influence on the composition of *Lotus* root- and rhizosphere-associated bacterial communities. Interestingly, the community shifts detected in *nfr5*, *nin* and *lhk1* mutant roots are very similar (Figure 6.2B), despite differences in nitrate content (Supporting Figure 6.16B). and colonization phenotypes with nitrogen-fixing rhizobia (Figure 6.1). In *nfr5* and *nin* mutants the infection process is either not initiated, or terminated at the microcolony stage, respectively, whereas *lhk1* plants develop a large number of root hair infection threads that subsequently fail to infect cortical cells (Schauser *et al.*, 1999; Murray *et al.*, 2007; Madsen *et al.*, 2010). The similarity in community shifts between all three tested mutants is likely related to our stringent root fractionation protocol, which is based on a combination of washing and sonication treatments and is known to damage the integrity of the root epidermis (Bulgarelli *et al.*, 2012; Dombrowski *et al.*, 2016), depleting both epiphytes and epidermal endophytes. Since the three mutants had comparable shoot appearance and protein levels, while differing in nitrate content, this suggests that the similar community shifts are unlikely to be caused by nitrogen stress.

Recent studies have shown that, upon microbiota acquisition, the root-associated bacterial assemblage remains robust against major changes in plant stature and source-sink relationships, i.e. in non-flowering and flowering mutants of the *Arabidopsis thaliana* relative *Arabis alpina* (Dombrowski *et al.*, 2016). This microbiota stability is also observed here, where a clear separation between the root and rhizosphere of WT and symbiosis-impaired *Lotus* plants is retained when plants are grown under different nitrate conditions (Figure 6.18). The latter observation is also strong evidence that the root-associated community shift in the symbiotic mutants is a direct consequence of the disabled symbiosis pathway rather than an indirect effect resulting from abolished symbiotic nitrogen fixation.

Previous controlled co-inoculation experiments with *Mesorhizobium loti* and various root endophytes have shown that *Lotus* can selectively guide various endophytic bacteria towards nodule primordia via symbiont-induced infection threads, and that endophyte and symbiont can promote each other's infection of the host (Zgadzaj *et al.*, 2015). These experiments focused on nodule colonization and were conducted using a gnotobiotic plant system with a limited number of endophytes. Our bacterial community profiling data obtained with soil-grown symbiotic mutants allowed testing the contribution of infection thread-dependent root colonization by natural populations of compatible endophytes present in soil. Indeed, the large number of bacterial taxa found depleted in *nfr5*, *nin* and *lhk1* genotypes (Supporting Figure 6.13) suggests that infection threads arrested in WT root epidermis or cortex may facilitate root colonization by endophytes. This finding implies additional functions of these host genes, besides rhizobial infection, for efficient root cortex colonization by a subset of the root microbiota. Our community profiling data also identified some bacterial orders that are enriched in the roots of the soil-grown symbiotic mutants, thus their root colonization takes place independently of nitrogen-fixing symbiosis. This suggests that these endophytic taxa employ alternative entry routes for root infection. One of these mechanisms is crack entry, which occurs at the base of emerging lateral roots and is likely also an important entry portal for root endophytes in non-leguminous plants (Patriquin *et al.*, 1983; Egener *et al.*, 1998; Chi *et al.*, 2005).

Our study revealed that *Lotus* growth in Cologne soil did not interfere with the host-symbiont compatibility described previously in mono-associations with plants grown in artificial media (Handberg and Stougaard, 1992). This high selectivity is evidenced by only 12 nodule-enriched OTUs among a total of 1,834 OTUs in our dataset and by Mesorhizobium members representing the most abundant OTUs inside nodules (Figure 6.3A). Remarkably, nodule-enriched OTUs had a similar relative abundance in root and rhizosphere compartments, suggesting linked selection process(es) in all three compartments. Furthermore, nodule-enriched OTUs were depleted from the mutant root samples, which corroborates our hypothesis that the intact symbiotic pathway is selecting rhizobia during infection in both root cortex and nodules. Interestingly, the remaining portion of the nodule community contained representatives from more than 20 bacterial orders (OTUs $> 0.05\%$ RA; Figure 6.3B), showing that nodules define a selective, but not exclusive, niche for rhizobia. The overlap between root- and nodule-inhabiting OTUs could be explained by the root-derived origin of nodules or by a specific capacity of these taxa to infect both organs via crack entry (Madsen *et al.*, 2010).

Interestingly, nodule-enriched Mesorhizobium OTUs were not only depleted from symbiotic mutant roots but also from their rhizosphere, indicating that root-derived diffusible compounds, produced by WT plants with the intact symbiotic pathway, exert a role in enriching symbionts in soil that adheres to the legume root surface. Thus, our findings support previous observations that maintenance of highly symbiotic isolates in the soil is not only a function of rhizobia release from decaying nodules, but is also dependent on persistent host selective pressure (Triplett *et al.*, 1993; Thies *et al.*, 1995). A likely scenario for this enrichment is a positive feedback mechanism in which symbiont-initiated signaling leads to further enrichment of symbionts in the rhizosphere. Legume root-derived flavonoids are candidate diffusible signaling molecules in such a feedback mechanism as their profile was shown to change during root-nodule symbiosis (Zuanazzi *et al.*, 1998; Rispail *et al.*, 2010) and their broad impact on soil bacterial communities (White *et al.*, 2015; Szoboszlay *et al.*, 2016), especially on symbiotic rhizobia, has been documented (Hartwig *et al.*, 1991; Dakora and Phillips, 1996).

Our comparative analyses of the root microbiota from *L. japonicus* and four Brassicaceae species grown in Cologne soil (Schlaeppi *et al.*, 2014) revealed a highly distinctive community composition irrespective of an intact or dysfunctional symbiosis pathway (Figure 6.7). Lotus and *Arabidopsis* ancestors diverged ∼118 MYA (Magallon *et al.*, 2015) and two major evolutionary events took place soon after their separation: loss of arbuscular mycorrhiza (AM) symbiosis in the Brassicaceae (Delaux *et al.*, 2014) and gain of nitrogen-fixation predisposition in the legume predecessor (Werner *et al.*, 2014)). Thus, our findings suggest that the marked distinctiveness of the *Lotus*-specific root microbiota is not governed by the evolved functions of *Nfr5*, *Nin* or *Lhk1*, despite their major effect on root microbiota composition, but is possibly linked to the loss of AM symbiosis in the Brassicaceae lineage. This can be tested by future experiments with mutants affecting *Lotus* 'common symbiosis genes' that fail to establish both symbiotic relationships with AM fungi and nodulating rhizobia (Stracke *et al.*, 2002; Kanamori *et al.*, 2006; Charpentier *et al.*, 2008).

6.6 Materials and methods

6.6.1 Soil and plant material

Seeds of *L. japonicus* WT, ecotype Gifu B-129 and the corresponding symbiosis-deficient mutants *nfr5*-2, *nfr5*-3, *lhk1*-1 and *nin*-2 were grown in soil batches collected from an agricultural field located at the Max Planck Institute for Plant Breeding Research in Cologne, Germany (50.958N, 6.865E) in the following seasons: CAS8-spring 2013, CAS9-fall 2013, CAS10-spring 2014. For each genotype and soil batch, 3 biological replicas were obtained, and the samples were harvested from 10 weeks-old plants grown in the greenhouse (day/night cycle 16/8h, light intensity 6000 LUX, temperature: 20 degrees C, relative humidity: 60%).

6.6.2 Sample and 16S rRNA library preparations

Fragments of the root systems (4 cm-long starting 1 cm below hypocotyl) were washed and ultrasound treated to separate the rhizosphere, root and nodule compartments. First wash containing the root-adhering soil layer defined the rhizosphere compartment. Nodules and visible primordia, were separated from washed root fragments of nodulating genotypes (WT and *lhk1*-1) with a scalpel. Root samples were exposed to 10 cycles of 30' ultrasound treatment, and the nodules to 3 cycles in order to avoid tissue out-burst. Homogenized samples were transferred to lysis buffer and DNA extraction was performed following the manufacturer's protocol (MP Bioproducts). DNA concentrations were measured fluorometrically (Quant-iT PicoGreen dsDNA assay kit, Life Technologies, Darmstadt, Germany), and finally adjusted to 3,5 ng/μl. Primers targeting the variable V5-V7 regions of bacterial 16S rRNA genes (799F and 1193R) (Bulgarelli *et al.*, 2012; Chelius and Triplett, 2001) were used for amplification. For each sample, triplicate amplifications were performed using three independent PCR mixtures (9 replicates per sample in total, along with no template controls). Amplification products were combined, purified and subjected to 454 sequencing.

6.6.3 Quantitative RT-PCR

WT and mutant root samples were used for mRNA isolation using oligodT DYNA beads following the protocol of the manufacturer (Invitrogen). The mRNA was subsequently used as template for cDNA synthesis using an oligodT primer. The same cDNA pool was used for amplification of all tested transcripts in each sample. Quantitative PCR was performed on the LightCycler (Roche Molecular Biochemicals) using FastStart DNA master SYBR greenI kit (Roche). The relative quantification software (Roche) was used to determine the efficiency-corrected relative transcript concentration, normalized to a calibrator sample. ATP, UBC, PP2A were used as housekeeping genes. For each genotype, normalized relative ratios of the target genes and the three independent housekeeping genes have been calculated using the Relative Quantification

Software (Quant) from Roche. The geometric mean of the relative expression ratios for the three biological and three technical replicates and the corresponding 95% intervals of confidence have been calculated.

6.6.4 Computational analyses

The 16S rRNA gene sequences were processed using a combination of custom scripts as well as tools from the QIIME (Caporaso *et al.*, 2010) and USEARCH (Edgar, 2010) pipelines. First, reads were truncated to an even length (290 bp) using the *truncate_fasta_qual_files.py* QIIME script. Libraries were demultiplexed (*split_libraries.py*) and only reads with a quality score $Q > 25$ were retained for subsequent analysis. After dereplication and removal of singletons we conducted *de novo* clustering of sequences into OTUs using the UPARSE algorithm (Edgar, 2013) at 97% identity. Next, chimeric sequences were filtered using the UCHIME algorithm (Edgar *et al.*, 2011) implemented in the USEARCH pipeline and the 'gold' database (`http://drive5.com/uchime/gold.fa`) as a reference. An additional filtering step was performed by first aligning all OTU representative sequences to the greengenes database (DeSantis *et al.*, 2006) using PyNAST (Caporaso *et al.*, 2010) (*align_seqs.py* script in QIIME) and subsequently discarding sequences not aligned to the database at least at 75% identity. OTU representative sequences were classified taxonomically using the RDP classifier (Wang *et al.*, 2007) and the uclust algorithm (Edgar, 2010) trained on the greengenes database. The resulting OTU table was used in all subsequent statistical analyses of differentially abundant taxa as well analyses of alpha- and beta-diversity.

Indices of alpha-diversity indices (Shannon, chao1 and number of observed OTUs) were calculated after subsampling to an even depth of 1,000 reads (QIIME script *alpha_diversity.py*). Measures of beta-diversity (Bray-Curtis and weighted and unweighted UniFrac) (Lozupone *et al.*, 2011) were calculated on a normalized OTU table (CSS method with QIIME script function *beta_diversity.py*). Principle Coordinate Analysis (PCoA) was done by classical multidimensional scaling of beta-diversity distance matrices using the cmdscale function in R. Canonical Analysis of Principle Components Coordinates (Anderson and Willis, 2003) was computed using the capscale function implemented in the 'vegan' R library, by constraining for the variable of interest and conditio*ning* for the remaining factors. Statistical significance was determined using a permutation-based ANOVA test implemented in anova.cca function using 5,000 permutations. Statistical analyses of differentially abundant OTUs were performed using the edgeR library (Robinson *et al.*, 2010). Briefly, we first obtained normalization factors using the function calcNormFactors and subsequently estimated common and tag-wise dispersions for a Negative Binomial Generalized Linear Model (GLM) using the estimateGLMCommonDisp and estimateGLMTagwiseDisp functions, respectively. In order to test for differential OTU abundances, we then fit a negative binomial generalized log-linear model to the read counts for OTU using the *glmFit* function. P values were corrected for multiple tests using the approach of (Benjamini and Hochberg, 1995) with α=0.05.

6.6.5 Metabolite analyses

Root nitrate contents were determined by ion chromatography method as previously described (Koprivova *et al.*, 2014). Approximately 20 mg plant tissue was homogenized and extracted in 1 mL of sterile water for 1 hour at 4 degrees C. The extracts were heated for 15 min at 95 degrees C and centrifuged for 15 min at 13,000 rpm. Ten μL of the extracts were analysed on Dionex IonPac AS22 RFIC 4x250 mm analytical column (Thermo Scientific) using a carbonate buffer (4.5 mM Na_2CO_3, 1.4 mM $NaHCO_3$) as eluent at 1 mL/min in an isocratic mode and a Dionex ICS-1100 ion chromatography system. Proteins were extracted in 10 mM Tris/HCl buffer, pH 8 and determined with a Bio-Rad Protein Assay Kit using bovine serum albumin as standard.

6.7 Author contributions

S. R. and P. S.-L. conceived this study. R. Z. collected plant material and performed culture-independent community profiling, R. G.-O. conducted the computational analysis, D. B. performed the qRT-PCR experiments, A. K. measured the metabolite concentrations. R. Z., R. G.-O., A.K., P. S.-L. and S. R. interpreted the data. R. Z., R. G.-O., P. S.-L. and S. R. wrote the manuscript.

6.8 Acknowledgements

We thank Diana Kuehn and Bruno Huettel (Max Planck Genome Centre Cologne) for the preparation and sequencing of the 454 libraries and Irene Klinkhammer (Botanical Institute Cologne) for assistance with nitrate analysis. We thank Jens Stougaard, Lene H. Madsen, Stig U. Andersen, Ryohei Nakano, Stephane Hacquard and Thorsten Thiergart for valuable comments on the manuscript. This work was supported by funds to S.R. from the Danish Council for Independent Research (Grant no. 09-070023) and the Danish National Research Foundation (Grant no. DNRF79), 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.

6.9 Supporting material

Supporting figures related to this section can be found bellow. All raw and ntermediate data as well as the scripts used to analyze the data and generate the figures of this study are available at `http://www.mpipz.mpg.de/R_scripts`.

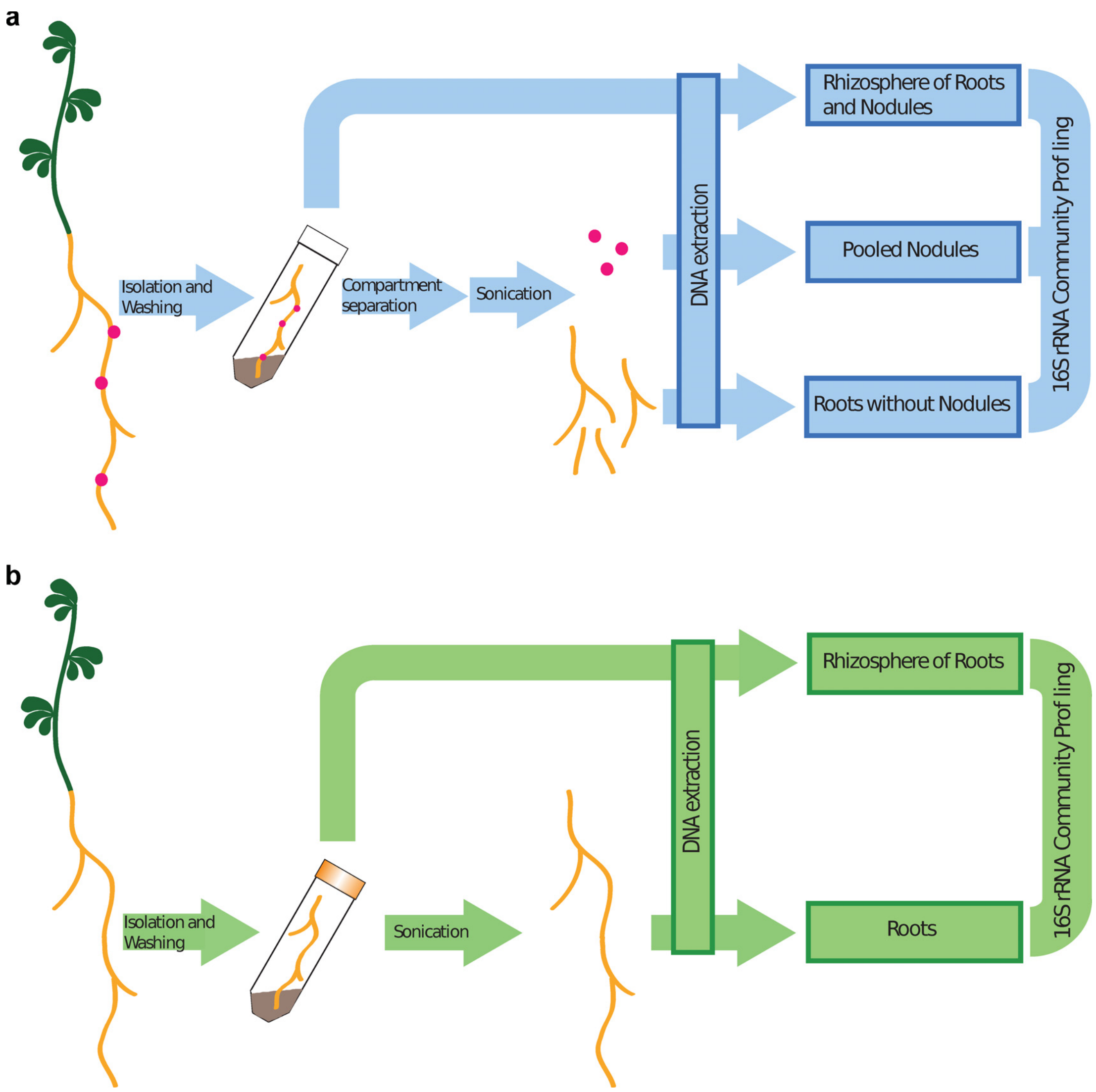


Figure 6.9: Harvesting procedures applied for compartment separation of nodulating (a) and non-nodulating (b) genotypes of *Lotus japonicus*.

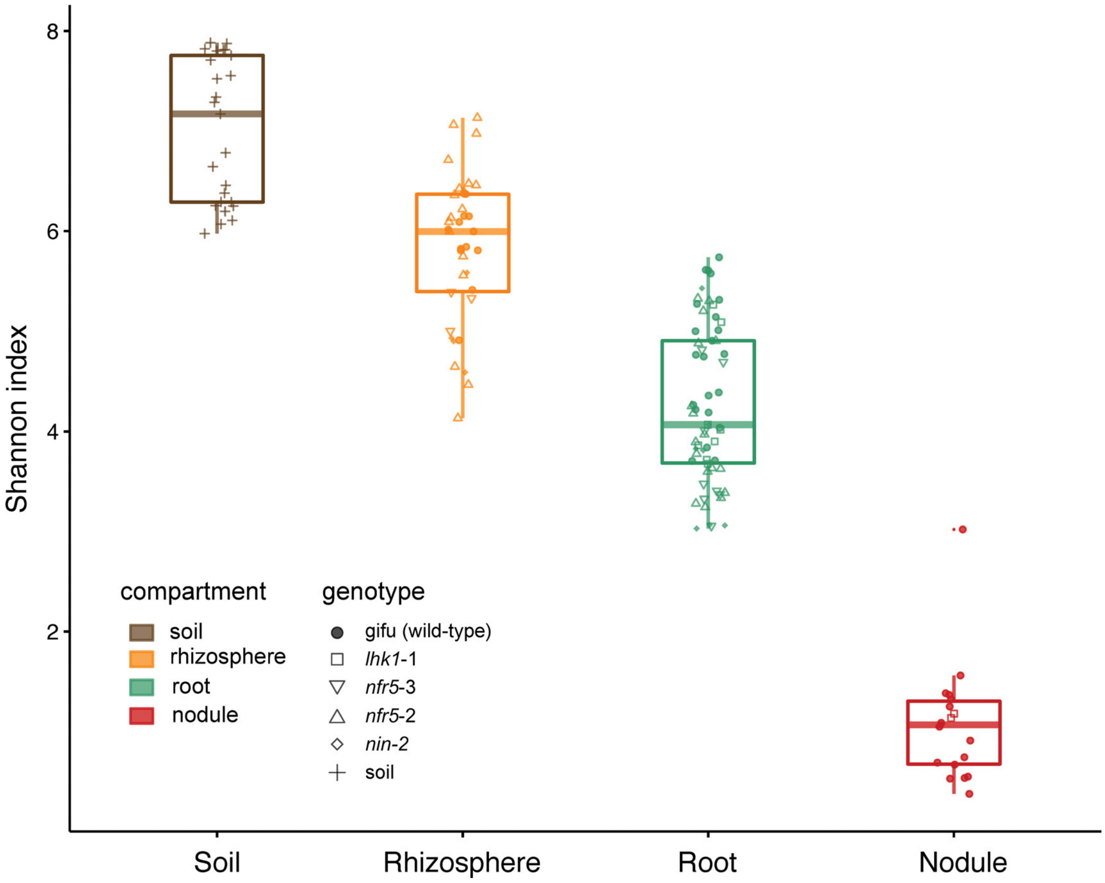


Figure 6.10: Analysis of alpha-diversity. Analisys of alpha-diversity (within sample diversity) based on the Shannon index in the soil (n=27), rhizosphere (n=38), root (n=62) and nodule (n=18) compartments. Each dot corresponds to an individual sample, colored by compartment and each host genotype is represented by a different shape. Each sample was rarefied to an depth of 1,000 reads before the analysis.

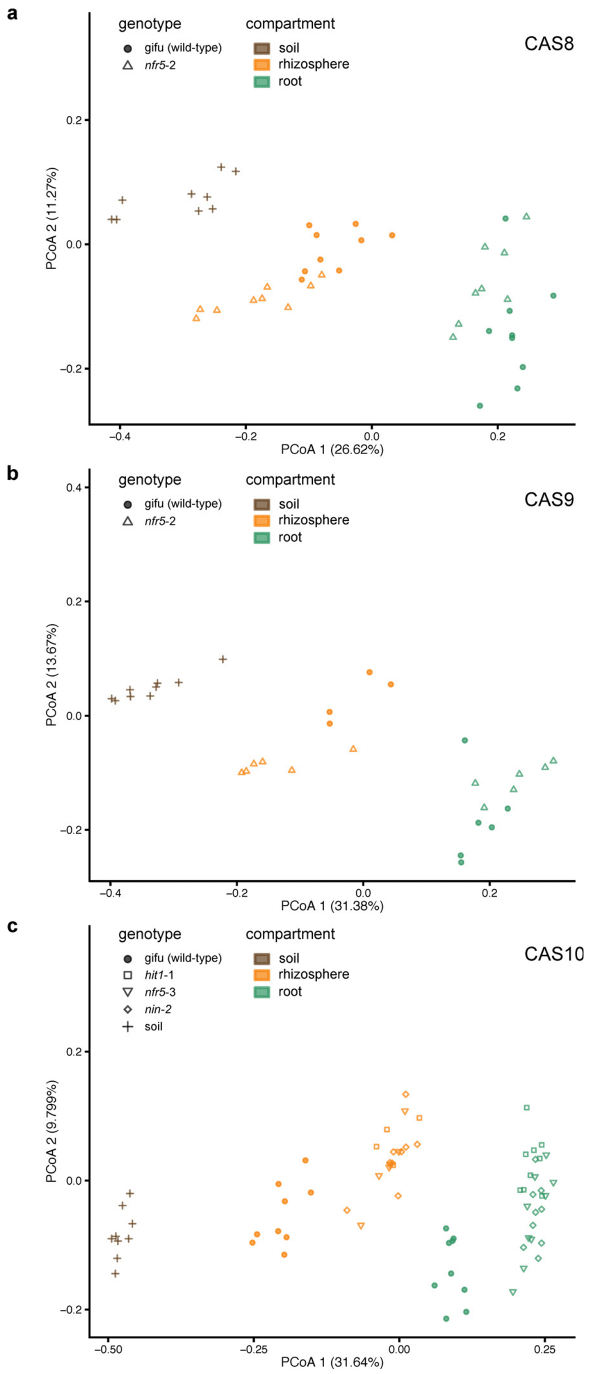


Figure 6.11: PCoA analysis of beta-diversity by soil batch. PCoA plots of Bray-Curtis distances for samples from each soil batch (a, CAS8, n=63; b, CAS9, n=31 and c, CAS10, n=72). Each point corresponds to a different sample, colored by compartment. Host genotype is represented by different shapes.

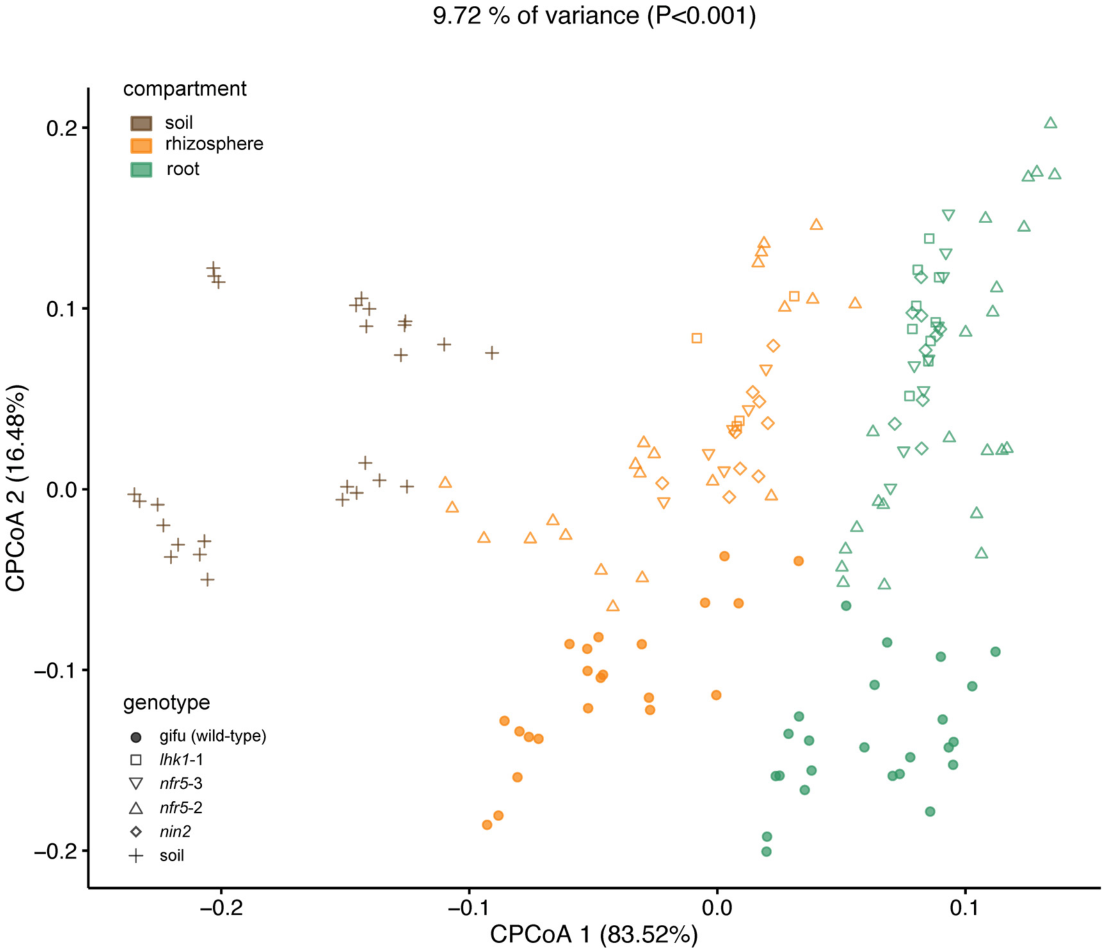


Figure 6.12: CPCoA analysis after *in silico* depletion of nodule-enriched OTUs. Constrained PCoA plot of Bray-Curtis distances constrained by genotype after in silico depletion of nodule-enriched OTUs (9.72% of variance explained, P<0.001; n=164). Each point corresponds to a different sample, colored by compartment and each host genotype is represented by a different shape. The percentage of variation indicated in each axis corresponds to the fraction of the total variance explained by the projection.

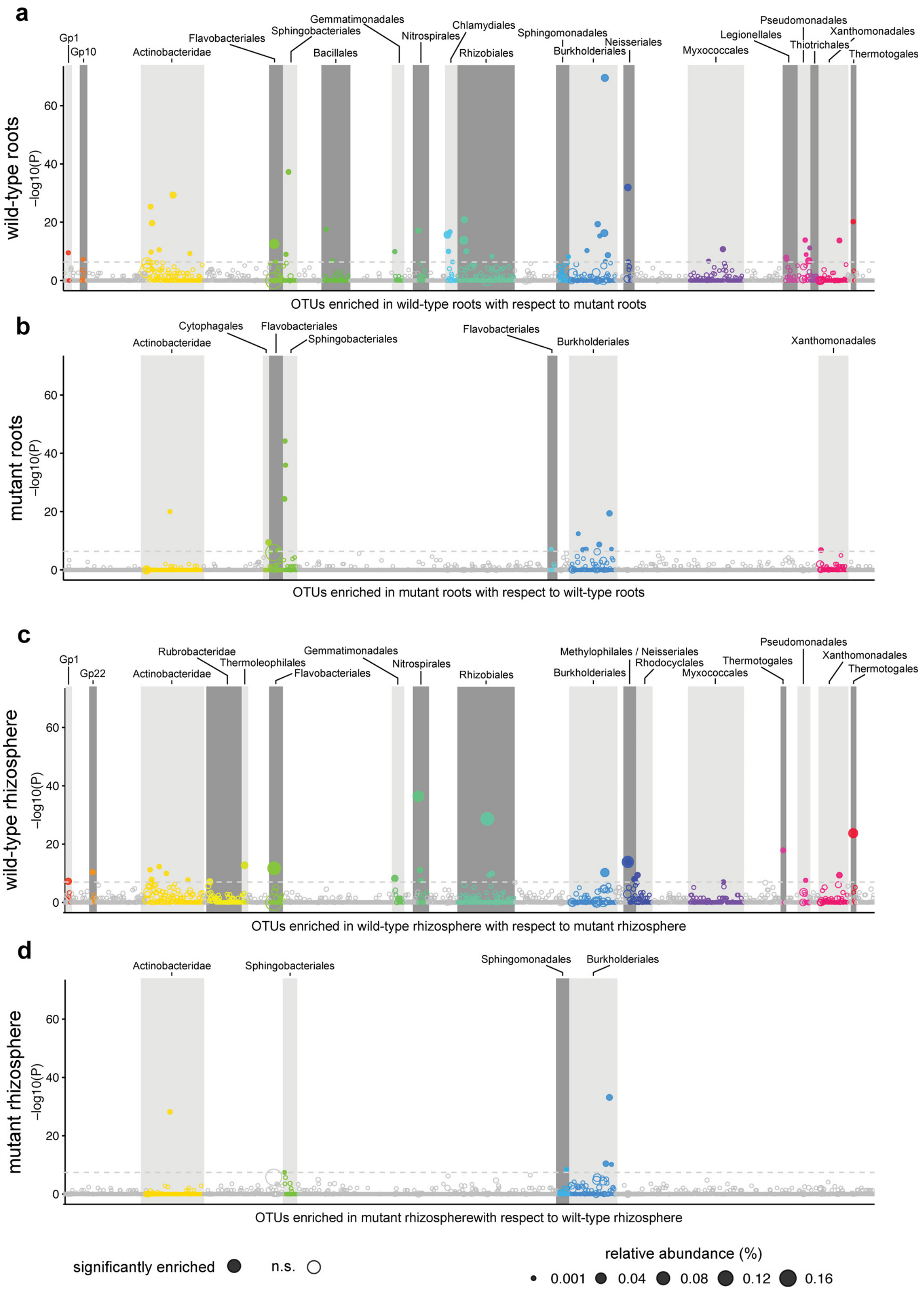


Figure 6.13: Manhattan plots showing differentially abundant OTUs in WT and mutant *Lotus*. (Caption on next page).

Figure 6.13: Manhattan plots showing differentially abundant OTUs in WT and mutant *Lotus*. (Figure on previous page).
Manhattan plots showing OTUs enriched in wild-type root compared to mutant root samples (a), depleted in the wild-type with respect to the mutants (b), enriched in the wild-type rhizosphere compared with respect to the mutants (c) or depleted in gifu rhizosphere with respect to the mutant rhizosphere (d). Significantly enriched OTUs are depicted as full circles. The dashed line corresponds to the FDR corrected P-value threshold of significance (α=0.05). The color of each dot represents the different taxonomic affiliation of the OTUs (order level) and their sizes to their relative abundances in their respective samples (a, wild-type root samples; b, mutant root samples). Grey boxes are used to denote the different taxonomic groups (order level).

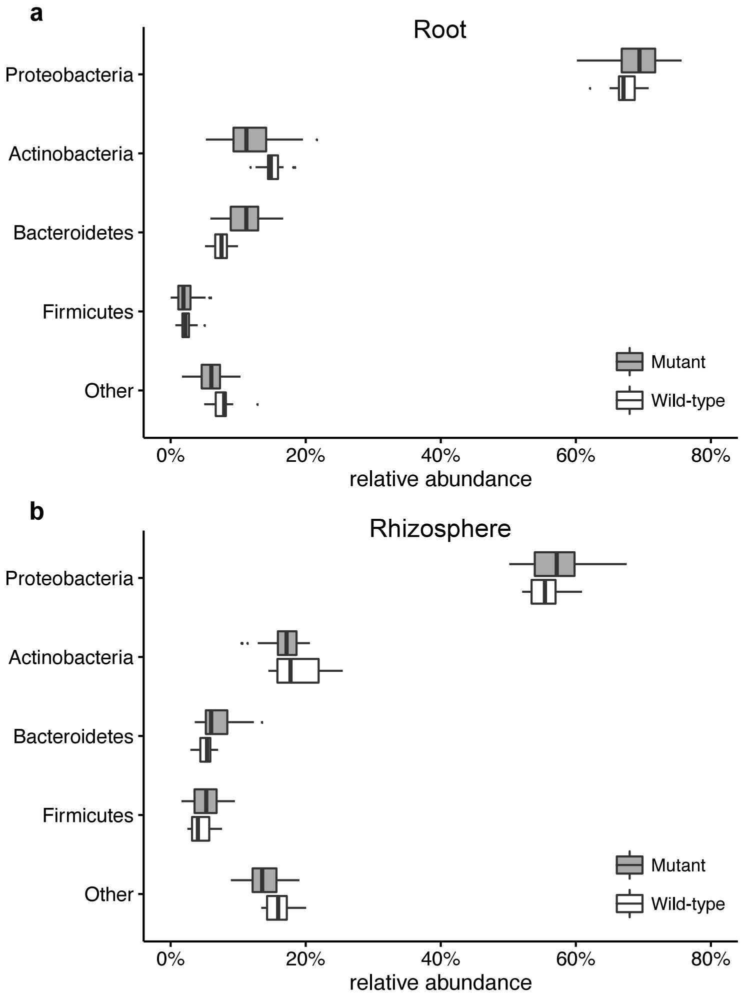


Figure 6.14: Phylum-level relative abundances in WT and mutant *Lotus* roots. Relative abundances aggregated to the phylum taxonomic level showing a comparison between wild-type and the mutant root (a; n=74) and rhizosphere (b; n=63) samples.

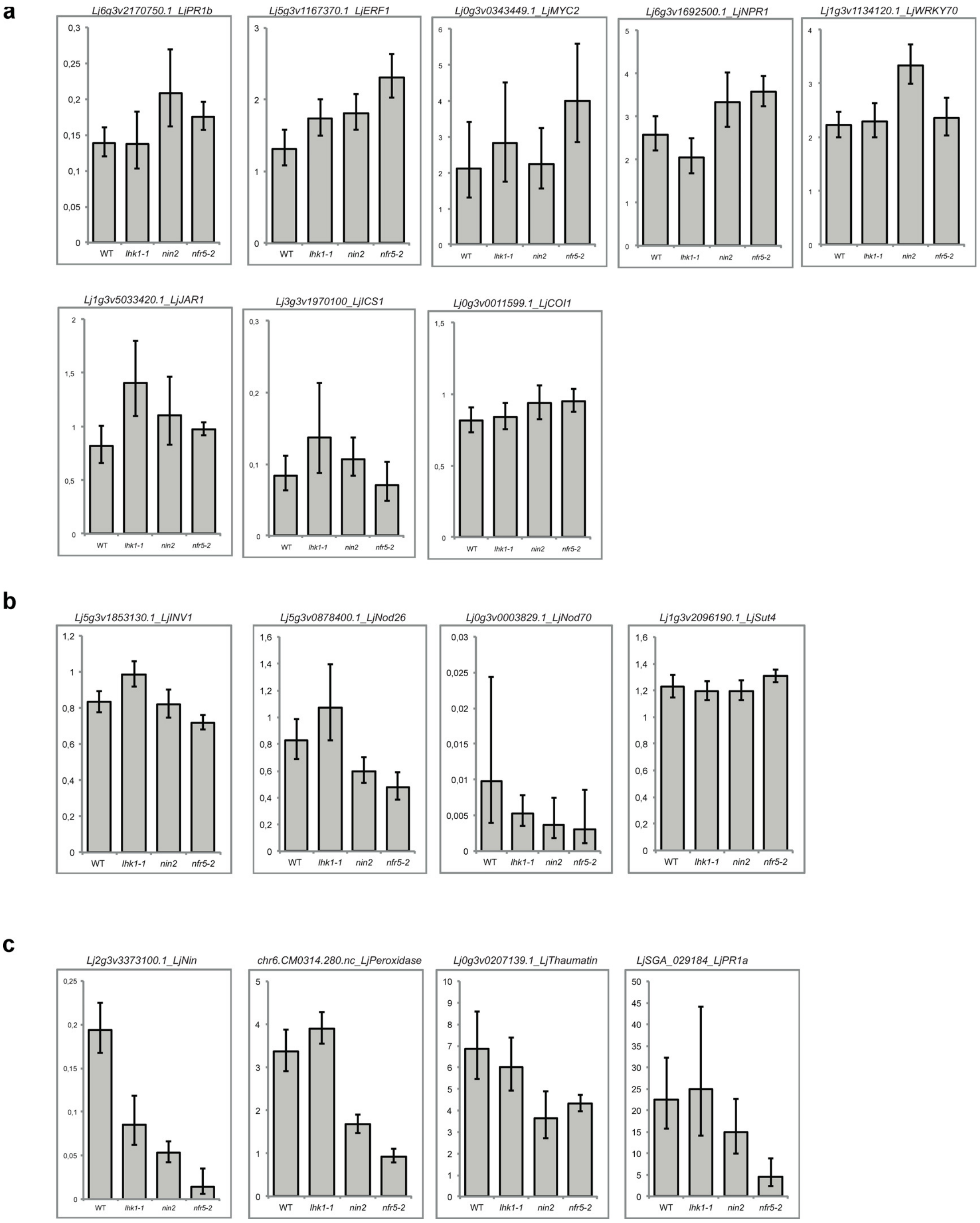


Figure 6.15: Expression of immune- and symbiosis-related marker genes in WT and mutant. Soil-grown wild-type and symbiotic mutants differ in the expression of early symbiotic genes, but show comparable immune- and symbiosis-related metabolic responses. Relative transcript levels of (a) immune-response genes (LjPR1b, LjErf1, LjMyc2 LjNpr1, Ljwrky70, LjJar1, LjIcs1, LjCoi1), (b) symbiosis-related metabolic responses (LjInv1, LjNod26, LjNod70, LjSut4) and (c) early symbiotic genes (Lj*Nin*, LjPeroxidase, LjThaumatin, LjPR1a).

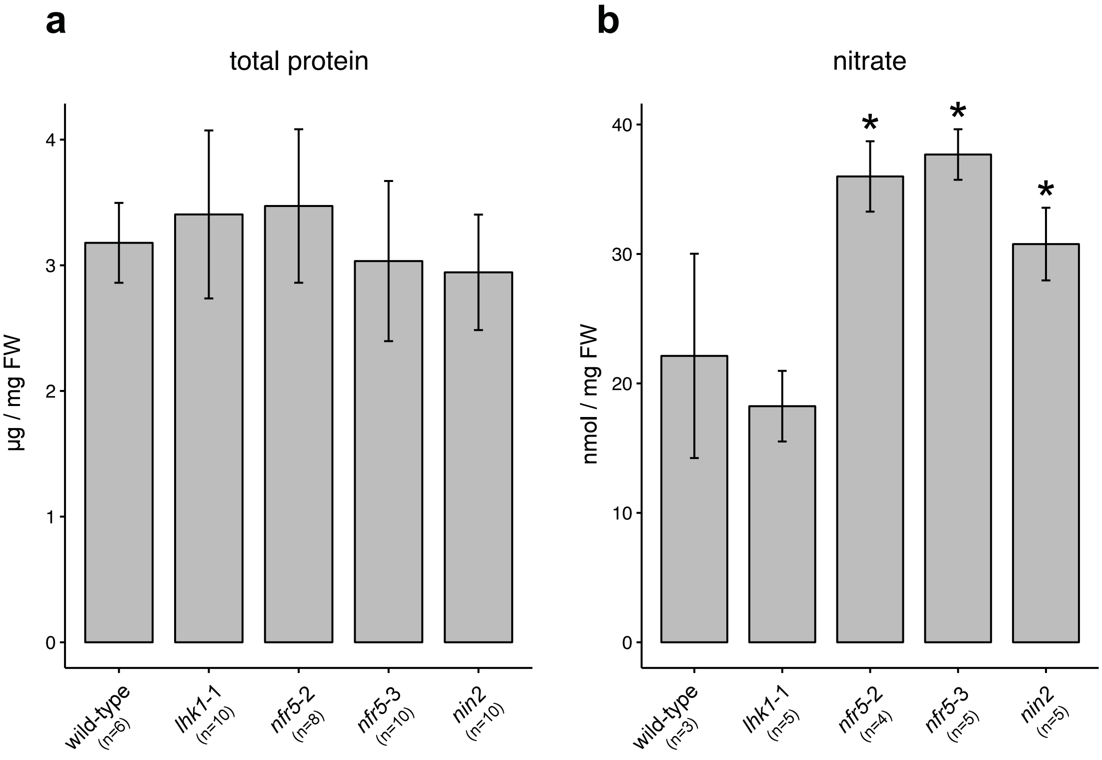


Figure 6.16: Concentration of soluble proteins and nitrate in *Lotus* roots. Roots of soil-grown wild-type and symbiotic mutants have the same protein concentration but differ in nitrate pool. Concentration of (a) soluble proteins and (b) nitrate in roots across *Lotus* genotypes. Bars correspond to the standard deviation between samples.

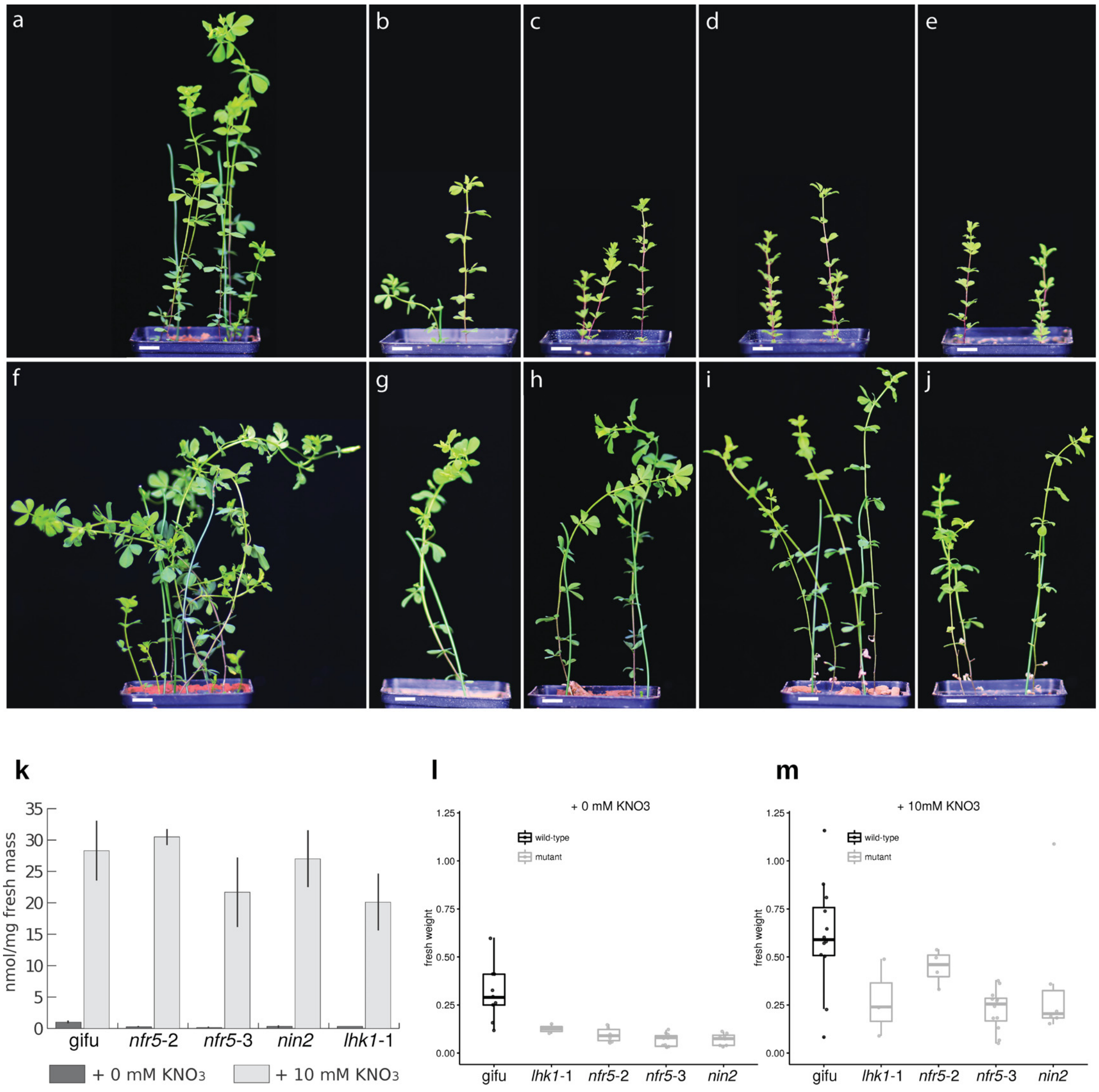


Figure 6.17: Macroscopic phenotypes of *L. japonicus* WT and mutant plants grown under nitrogen-supplemented conditions. (a) Images depicting *L. japonicus* wild-type (a) and root nodule symbiosis-deficient mutant plants *lhk1*-1 (b) *nfr5*-2 (c) and *nfr5*-3 (d) and *nin2* (e) grown in natural soil. The lower panels display an identical set of genotypes, wild-type (f), *hit1*-1 (g), *nfr5*-2 (h), *nfr5*-3 (i) and *nin2* (j), grown in natural soil supplemented with 10 mM potassium nitrate. Scale bars correspond to 1cm. (k) differences in nitrate concentrations in the roots of wild-type and Lotus mutants under both conditions. Differences in fresh weight of wild-type Lotus and mutant plants grown in natural soil (l) and in natural soil supplemented with potassium nitrate (m).

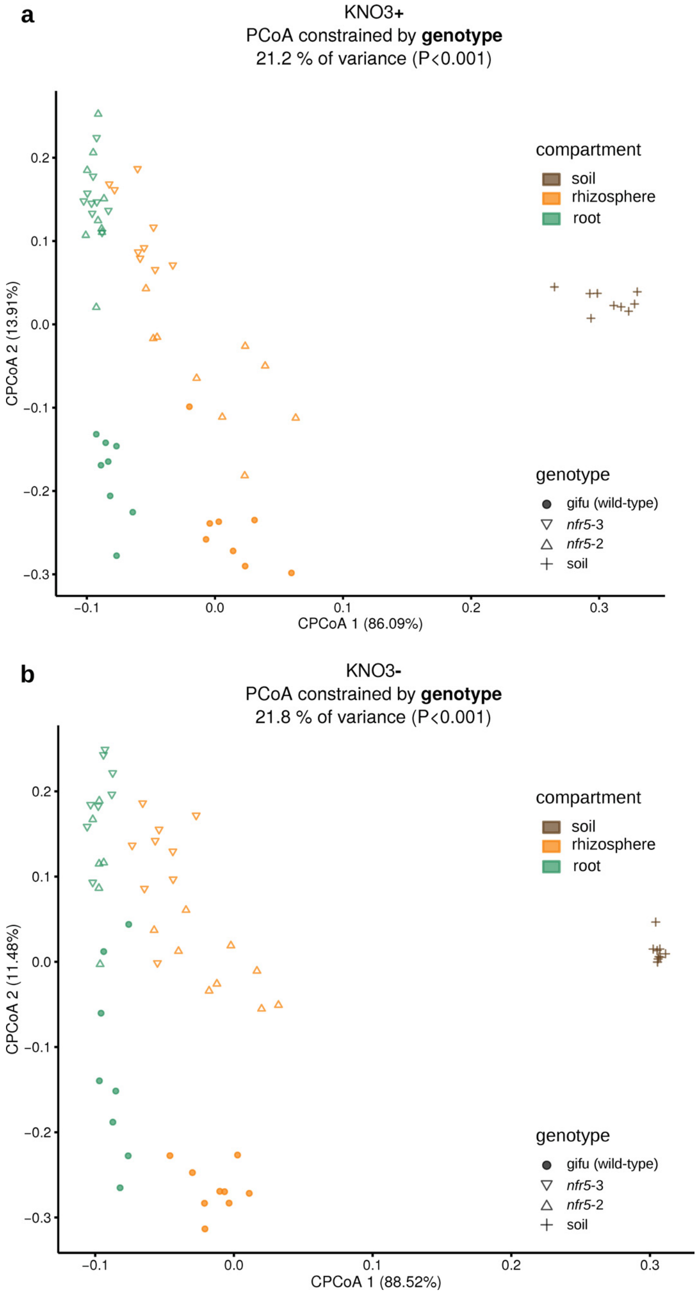


Figure 6.18: Beta-diversity analyses of nitrogen-suplemented *L. japonicus* WT and mutant plants. (a) Constrained PCoA plot of Bray-Curtis distances between samples from KNO3 treated CAS11 soil constrained by genotype (21.2% of variance, P>0.001; n=61). (b) Contrained PCoA plot of Bray-Curtis distances between CAS11 untreated samples constrained by genotype (21.8% of variance explained, P<0.001; n=58). Each point corresponds to a different sample, colored by compartment and each host genotype is represented by a different shape. The percentage of variation indicated in each axis corresponds to the fraction of the total variance explained by the projection.

Part II

Functions of root- and leaf-associated microbes

CHAPTER 7

Functional overlap of the *Arabidopsis* leaf and root microbiota

Status	**Published**
Journal	*Nature* (Impact factor 38.138)
Citation	Bai, Y.*, Mueller, D.B.,*, Srinivas, G.*, Garrido-Oter, R.*, Potthoff, E., Rott, M., Dombroski, N., Muench, P.C., Spaepen, S., Remus-Emsermann, M., Huettel, B., McHardy, A.C., Vorholt, J.A. †, Schulze-Lefert, P. † (2015). Functional overlap of the *Arabidopsis* leaf and root microbiota *Nature*, **528** (7582):364-369. * joint first authors; † joint corresponding authors
URL	http://dx.doi.org/doi:10.1038/nature16192
Own contribution	Performed the computational experiments (with co-authors) Analyzed the natural community 16S data (with co-authors) Assembled, annotated and curated library of bacterial genomes Analyzed the whole-genome sequencing data Analyzed the synthetic community 16S data (with co-authors) Interpreted the data (with co-authors) Wrote the manuscript (with co-authors)

7.1 Abstract

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

7.2 Introduction

Plants and animals harbour abundant and diverse bacterial microbiota (Rosenberg and Xilber-Rosenberg, 2013). These taxonomically structured bacterial communities have important functions for the health of their multicellular eukaryotic hosts (Spor *et al.*, 2011; Berendsen *et al.*, 2012; Subramanian, 2015). The leaf and root microbiota of flowering plants have been extensively studied by culture-independent analyses, which have consistently revealed the co-occurrence of four main bacterial phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria (Delmotte, 2009; Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Vorholt, 2012; Bodenhausen *et al.*, 2013; Guttman *et al.*, 2014; Horton, 2014; Schlaeppi *et al.*, 2014; Edwards *et al.*, 2015; Hacquard *et al.*, 2015; Bulgarelli *et al.*, 2015). Determinants of microbiota composition at lower taxonomic ranks, that is, at genus and species level, are host compartment, environmental factors and host genotype (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Schlaeppi *et al.*, 2014; Lebeis *et al.*, 2015).

Soil harbours an extraordinary rich diversity of bacteria and these define the start inoculum of the *Arabidopsis thaliana* root microbiota (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012). The inoculum source of the leaf microbiota is thought to be more variable owing to the inherently open nature of the leaf ecosystem, probably involving bacteria transmitted by aerosols, insects, or soil (Vorholt, 2012; Bodenhausen *et al.*, 2013; Maignien *et al.*, 2014). A recent study of the grapevine (*Vitis vinifera)* microbiota showed that the root-associated bacterial assemblies differed significantly from aboveground communities, but that microbiota of leaves, flowers, and grapes shared a greater proportion of taxa with soil communities than with each other, suggesting that soil may serve as a common bacterial reservoir for belowground and aboveground plant microbiota (Zarraonaindia *et al.*, 2015).

A major limitation of current plant microbiota research is the lack of systematic microbiota culture collections that can be employed in microbiota reconstitution experiments with germ-free plants to address principles underlying community assembly and proposed microbiota functions for plant health under laboratory conditions (Lebeis *et al.*, 2012).

7.3 Results

7.3.1 Bacterial culture collections from roots and leaves

We employed three bacterial isolation procedures to establish taxonomically diverse culture collections of the *A. thaliana* root and leaf microbiota. Bacterial isolates were recovered from pooled or individual roots or leaves of healthy plants using colony picking from agar plates, limiting dilution in liquid media in 96-well microtitre plates, or microbial cell sorting (see 7.5). We adopted a two-step bar-coded pyrosequencing protocol (Goodman *et al.*, 2011) for taxonomic classification of the cultured bacteria by determining ≥550 base pairs (bp) 16S ribosomal RNA (rRNA) gene sequences (Supporting Figure 7.6; 7.5). In parallel, parts of the root and leaf material was used for cultivation-independent 16S rRNA gene community sequencing to cross-reference Operational Taxonomic Unit (OTU)-defined taxa from the microbiota with individual colony forming units (CFUs) in the culture collections.

A total of 5,812 CFUs were recovered from 59 independently pooled *A. thaliana* root samples of plants mainly grown in Cologne soil, Germany, whereas 2,131 CFUs were retrieved from leaf washes of individual leaves collected from *A. thaliana* populations at six locations near Tübingen, Germany, or Zurich, Switzerland (Supporting Material). Recovery estimates for root-associated OTUs were calculated using the culture-independent community profiles of the present and two earlier studies (Bulgarelli *et al.*, 2012; Schlaeppi *et al.*, 2014) and varied for the top 100 OTUs (70% of sequencing reads) between 54-65% and at ≥0.1% relative abundance (RA) between 52-64% (7.5;

Supporting Material). For leaf samples, the culture-independent 16S rRNA gene analyses from individual and pooled leaves (60 samples from six sites) revealed similar community profiles at all tested geographic sites and high leaf-to-leaf consistency (Supporting Material). Recovery estimates of the top 100 leaf-associated bacterial OTUs (86% of all sequencing reads) were 54% and at ≥0.1% RA 47% (Supporting Material). The root-derived CFUs correspond to 23 of 38 and the leaf-derived CFUs belong to 28 of 45 detectable bacterial families. Root- and leaf-derived CFUs each represent all four bacterial phyla typically associated with *A. thaliana* roots and leaves. Thus, most bacterial families that are reproducibly associated with *A. thaliana* roots and leaves have culturable members.

7.3.2 *At*-RSPHERE and *At*-LSPHERE culture collections

We selected from the aforementioned culture collections a taxonomi- cally representative core set of bacterial strains after Sanger sequencing of a ≥550 bp fragment of the 16S rRNA gene and additional strain purification (7.5). To increase the intra-species genetic diversity of the culture collections, and because the quantitative contribution of a single isolate to its corresponding OTU cannot be estimated, we included bacterial strains sharing ≥97% 16S rRNA gene sequence identity (widely used for bacterial species definition), but representing independent host colonization events, that is, recovered from different plant roots or leaves. In total we selected 206 root-derived isolates that comprise 28 bacterial families belonging to four phyla (designated *At*-RSPHERE) and 224 leaf-derived isolates that comprise 29 bacterial families belonging to five phyla (designated *At*-LSPHERE) (Supporting Material; 7.5). Additionally, to represent abundant soil OTUs (≥0.1% RA) we selected 33 bacterial isolates encompassing eight bacterial families belonging to three phyla from unplanted Cologne soil (Supporting Material).

Figure 7.1: Taxonomic distribution of *At*-RSPHERE and *At*-LSPHERE. (Figure on next page). Phylogenetic trees of *At*-RSPHERE (a; n=206 isolates) and *At*-LSPHERE (b; n=224 isolates) bacteria. Their taxonomic overlap is shown in the outermost ring (green or brown triangles). a, Representation of *At*-RSPHERE bacteria in each of four indicated culture-independent profiling studies of the *A. thaliana* root microbiota; root-associated OTUs with RAs ≥0.1% (dark orange) or ≥0.1% (light orange). b, Representation of *At*-LSPHERE bacteria in the two indicated culture-independent phyllosphere profiling studies; leaf-associated OTUs with RAs ≥0.1% (dark green) or <0.1% (light green). Taxonomic assignment and phylogenetic tree inference were based on partial 16S rRNA gene Sanger sequences.

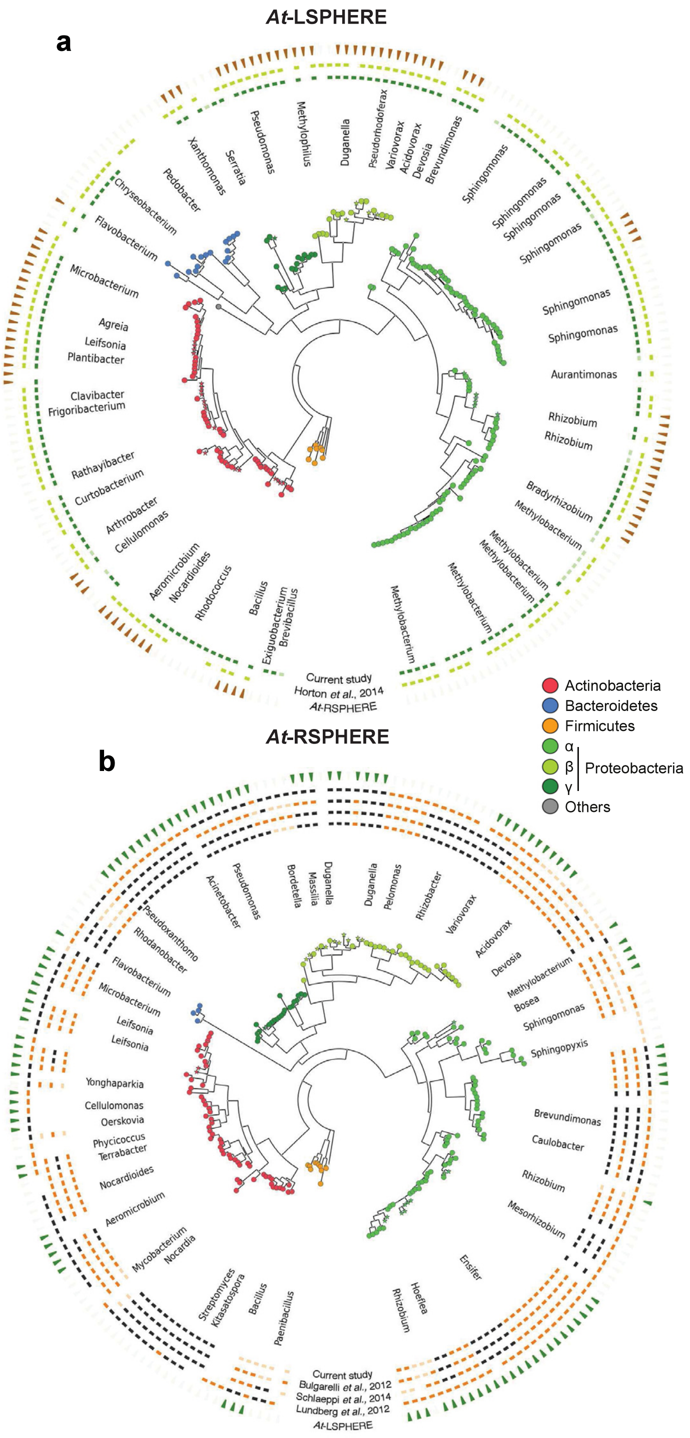


Figure 7.1: Taxonomic distribution of *At*-RSPHERE and *At*-LSPHERE. (Caption on previous page).

Notably, the majority of the At-RSPHERE isolates share ≥97% 16S rRNA gene sequence identity matches with root-associated OTUs reported in four independent studies in which A. thaliana plants had been grown in Cologne soil or other European (Bulgarelli *et al.*, 2012; Schlaeppi *et al.*, 2014) or US soils (Lundberg *et al.*, 2012) (inner four circles in Figure 7.1A; 7.5). Similarly, the bulk of *At*-LSPHERE isolates match leaf-derived OTUs detected in *A. thaliana* populations at the Tübingen/Zurich locations or US-grown plants (innermost two circles in Figure 7.1B). This indicates that representatives of the majority of *At*-RSPHERE and *At*-LSPHERE members co-populate the corresponding *A. thaliana* organs in multiple tested environments, including two continents, Europe and North America.

Phylogenetic analysis based on 16S rRNA gene Sanger sequences revealed that 119 out of 206 At-RSPHERE isolates (58%) share ≥97% sequence identity matches with corresponding 16S rRNA gene fragments of *At*-LSPHERE members (outermost circle in Figure 7.1A). Similarly, 108 out of 224 *At*-LSPHERE isolates (48%) share ≥97% sequence identity matches with At-RSPHERE members (outermost circle in Figure 7.1B). This extensive overlap both at the rank of bacterial genera and bacterial families (20 out of 38 detectable families) between leaf- and root-derived bacteria is notable because we collected leaf and root specimen from environments that are geographically widely separated (>500 km) and is consistent with a previous report on leaf and root microbiota overlap in *V. vinifera* (Zarraonaindia *et al.*, 2015). This overlap is corroborated by the corresponding culture-independent leaf and root community profiles (Supporting Material). As essentially all *A. thaliana* root-associated bacteria are recruited from the surrounding soil biome (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Schlaeppi *et al.*, 2014), this raises the possibility that unplanted soil also defines the start inoculum for a substantial proportion of the leaf microbiota with subsequent selection for niche-adapted organisms.

7.3.3 Comparative genome analysis of the culture collections

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

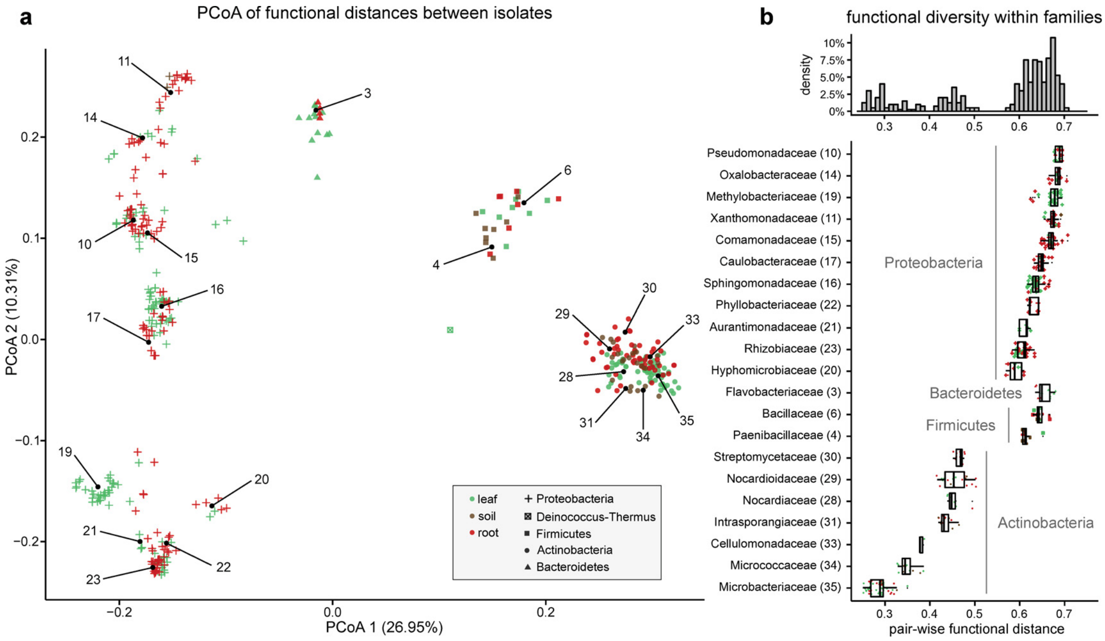


Figure 7.2: Analysis of functional diversity between sequenced isolates. a, Principal coordinate analysis (PCoA) plot depicting functional distances between sequenced genomes (n=432) based on the KEGG Orthology (KO) database annotation. Each point represents a genome. Colours represent the organ of isolation and shapes correspond to their taxonomy. Numbers inside the plot refer to bacterial families listed in b. b, Analysis of functional diversity within bacterial families as measured by pair-wise functional distances between genomes (bottom panel; n=432). Higher pairwise distances between members of a family indicate a larger degree of functional diversity. Only families with at least five members are shown.

Based on the whole-genome taxonomic information, we grouped the isolates into family-level clusters. We found that clusters of genomes are characterized by a relatively large core-genome, with an average of 33.6% of the annotated proteins present in each member and a smaller fraction of singleton genes identified in only one genome per cluster (14.0%). Detailed analysis of phylogenetic diversity of each cluster revealed a substantial overlap between leaf, root and soil isolates Many clusters showed no clear separation of isolates based on their ecological niche, suggesting shared core functions. However, other clusters contained isolates of one organ or showed clear separation among them, suggesting niche specialization within some clusters (Supporting Material). We then examined the functional diversity between the sequenced isolates in order to determine whether the observed phylogenetic overlap corresponded with functional similarities between leaf and root isolates. Principal coordinates analysis (PCoA) of functional distances (Figure 7.2A; 7.5) revealed a clear clustering of genomes on the basis of their taxonomy, but only limited separation of genomes on the basis of their ecological compartment. Taken together, both phylogenetic and functional diversification of the genomes is strongly driven by their taxonomic affiliation and weakly by the ecological niche.

We examined the functional diversity within each bacterial family (Figure 7.2B) in order to identify bacterial taxa with varying degrees of functional versatility. Families belonging to Actinobacteria show a lower functional diversity (average distance 0.37) compared to those belonging to Bacteroidetes, Firmicutes and especially Proteobacteria (0.65 average pair-wise distance), which exhibit a higher degree of within-family functional diversification, even though all family-level groups have a comparable degree of phylogenetic relatedness. Among these groups, Pseudomonadaceae, Oxalobacteraceae and Methylobacteriaceae members show the highest functional heteroge- neity, compared to Microbacteriaceae strains, which we identified as the least functionally diverse family (Figure 7.2B).

We searched for signatures of niche specialization at individual functional categories using enrichment analysis to identify functional categories over-represented in genomes from root and leaf or soil isolates (Figure 7.3; 7.5). Specifically, we found the category 'carbohydrate metabolism' to be enriched in the leaf and soil genomes compared to those isolated from roots (Mann-Whitney test, P = 1.29E10-7; Figure 7.3B). We speculate that this differential enrichment could reflect the availability of simple carbon sources in roots through the process of root exudation (sugars, amino acids, aliphatic acids) (Faure *et al.*, 2009; Bais *et al.*, 2006), whereas bacteria associated with leaves or unplanted soil might rely on a more diverse repertoire of carbohydrate metabolism genes to access scarce and complex organic carbon, for example, polysaccharides and leaf cuticular waxes. The category 'xenobiotics biodegradation and catabolism' is enriched in the root genomes with respect to those isolated from leaves (P = 2.60E10-11; Figure 7.3B), which is consistent with previous evidence that genes for aromatic compound utilization are expressed in the rhizosphere (Ramachandran *et al.*, 2011). No single taxon is responsible for these significant differences, but this seems to be a general feature across the sequenced bacterial genomes of the respective ecological niche (Supporting Material). Interestingly, we observed the same trends of differential abundance of functional categories in *V. vinifera* root metagenome samples (Zarraonaindia *et al.*, 2015) compared to their respective unplanted soil controls (Supporting Material).

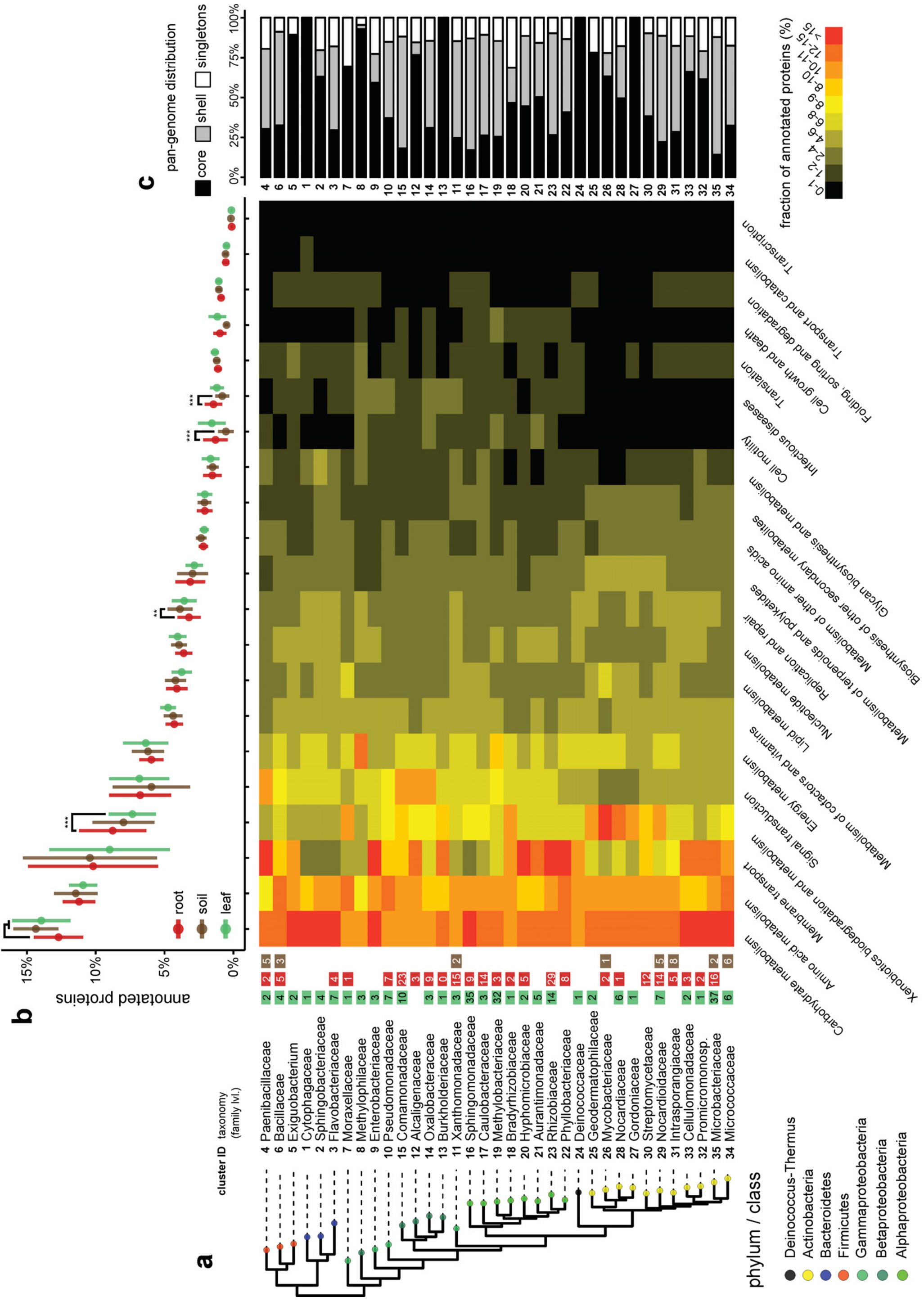


Figure 7.3: Functional analysis of sequenced isolates. (Caption on next page).

Figure 7.3: Functional analysis of sequenced isolates. (Figure on previous page). a, Phylogeny of family-level clusters of bacterial isolates. The tips of the tree are annotated, from left to right, with the cluster ID, taxonomic classification, followed by the number of sequenced isolates from leaf, root or soil that constitute each cluster. The heat map depicts the average percentage of annotated proteins of each cluster belonging to each functional category. b, Functional enrichment analysis between leaf (n=206), root (n=194) and soil (n=32) genomes. Points and bars correspond to the mean abundance and standard deviation of each functional category. P values were obtained using the non-parametric Mann-Whitney test corrected by the Bonferroni approach. c, Analysis of pan-genome distribution for each cluster of genomes, indicating the percentage of annotated proteins found in only one isolate (singletons), in more than one but not all (shell) or in all genomes within the cluster (core).

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

7.3.4 Synthetic community colonization of germ-free plants

We colonized germ-free *A. thaliana* plants with synthetic communities (SynComs) consisting of bacterial isolates from our culture collections to assess their potential for host colonization in a gnotobiotic system containing calcined clay as inert soil substitute (7.5). To mimic the taxonomic diversity of leaf and root microbiota in natural environments we employed mainly two SynComs: 'L' comprising 218 leaf-derived bacteria and 'R+S' consisting of 188 members of which 158 are root-derived and 30 are soil-derived bacteria (Supporting Material). Input SynComs were either inoculated directly before sowing of surface-sterilized seeds in calcined clay and/or spray-inoculated on leaves of three-week-old germ-free plants. For all defined communities we examined three independent SynCom preparations, each tested in three closed containers containing four plants. We employed 16S rRNA gene community profiling with a method validated for defined communities (Edgar, 2013) to detect potential community shifts between input and output SynComs in samples of seven week-old roots, leaves, or unplanted clay. In this community analysis, 'indicator OTUs' either represent a single strain or a known group of isolates.

Upon application of the input R+S SynCom to clay ('R+S in clay') and co-cultivation with A. thaliana plants for seven weeks we retrieved reproducible R+ S output com-

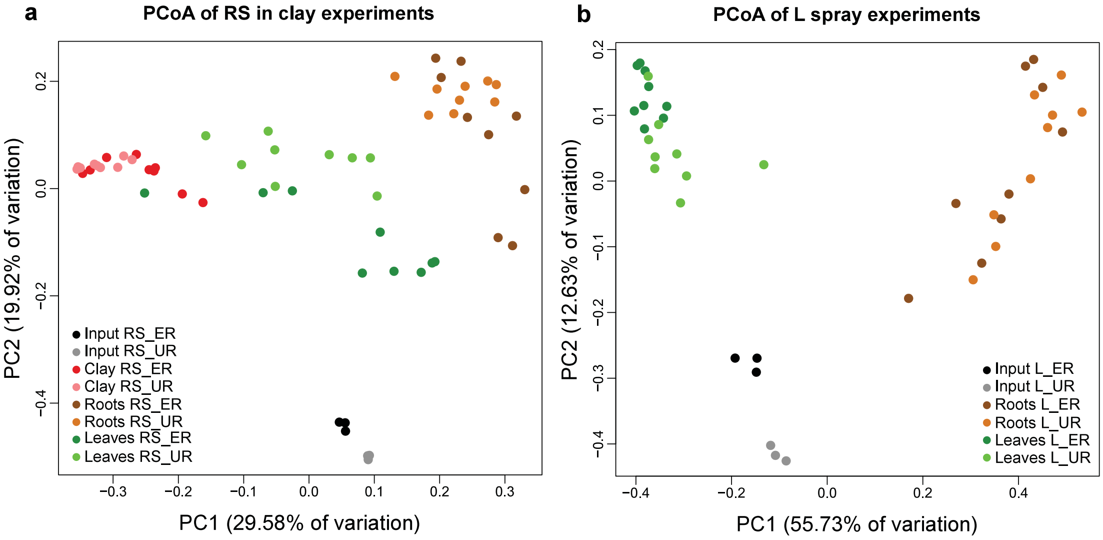


Figure 7.4: SynCom colonization of germ-free *A. thaliana* plants. Principal coordinate analysis (PCoA) of Bray-Curtis distances of input and output SynCom profiles of RS in clay (a; n=60) and L spray (b; n=42) experiments. Each condition was tested with 6 independently prepared SynComs; each preparation was used for 3 independent inoculations. L, leaf-derived strains; RS, root- and soil-derived strains; ER, equal strain ratio; UR, unequal strain ratio.

munities from clay (without host), root, and leaf compartments These output SynCom profiles were robust against a 75% reduction in RA of Proteobacteria compared to Actinobacteria, Bacteroidetes and Firmicutes in the input R+S SynCom (input ratios 1:1:1:1 or 1:1:1:0.25, respectively), which was confirmed by PCoA (Figure 7.4A). PCoA also revealed distinct output communities in each of the three tested compartments (Figure 7.4A; P < 0.001 Supporting Material). This indicates that a marked host-independent community change occurred in clay (without host) as well as host-dependent community shifts that are specific for leaves and roots. Next, we tested the 'L' SynCom of leaf-derived bacteria by spray inoculation on leaves of three week-old plants. After four weeks of L SynCom co-incubation with plants, output communities were detected in leaves and roots.

PCoA revealed that these two output communities were different between each other, but robust against a 75% reduction in RA of input Proteobacteria (Figure 7.4B; P < 0.001; Supporting Material). The converging output communities despite varying RAs of input SynComs suggest that the communities have reached a steady state. These experiments also reveal that both R+S and L SynCom members not only colonize cognate host organs, but are capable of ectopic colonization of leaves and roots, which might be linked to the extensive species overlap of *A. thaliana* leaf and root microbiota in natural environments (Figure 7.1A-B). Additionally, this provides experimental support for the hypothesis that a subset of leaf-colonizing bacteria originates from unplanted soil and raises the possibility for reciprocal bacterial colonization events between roots and leaves during and/or after the establishment of the respective microbiota, for example, by ascending migration of rhizobacteria from roots to leaves (Chi *et al.*, 2005). Upon leaf spray application of SynComs, a small amount of leaf bacteria is likely to land on the clay surface and thereafter colonize roots, which is not fundamentally different from processes occurring in natural environments, for example, during rain showers and/or leaf dehiscence.

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

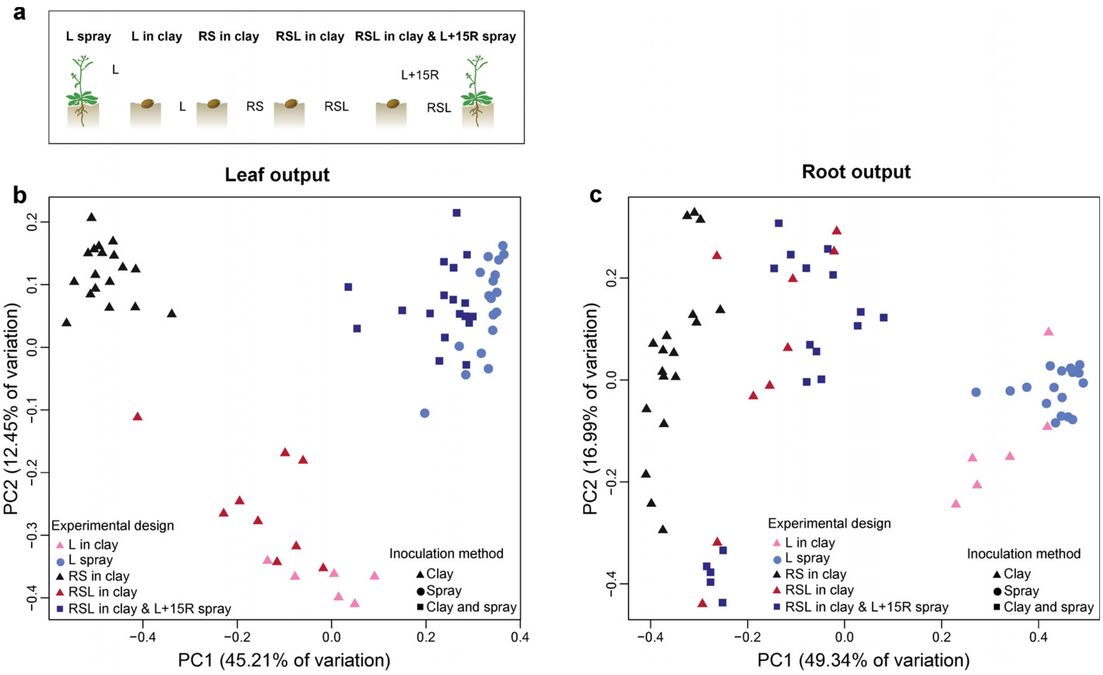


Figure 7.5: SynCom competition supports host-organ-specific community assemblies. a, Pictograms illustrating 'L spray', 'L in clay', 'RS in clay', 'RSL in cla', and 'RSL in clay & L+15R spray' SynCom experiments. b, c, PCoA of Bray-Curtis distances of leaf (b; n=69) and root (c; n=69) outputs of the five experiments illustrated in a. R, root-derived isolates; S, soil-derived isolates; L, leaf-derived isolates.

7.3.5 Niche-specific microbiota establishment with SynComs

The species overlap between root and leaf microbiota and their corresponding culture collections (Figure 7.1A-B; Supporting Material). prompted us to test whether R+S and L SynComs equally contribute to root and leaf microbiota establishment. Both SynComs were pooled and inoculated in clay together with surface-sterilized *A. thaliana* seeds (designated 'RSL in clay', (Figure 7.5A). We also tested whether a preformed root-associated community can interfere with leaf-associated community establishment. After three weeks of co-cultivation, half of the plants grown with the 'RSL in clay' SynCom were treated by leaf-spray inoculation with the L SynCom supplemented with 15 root-derived strains (designated 'RSL in clay & L+15R spray'). Plant organ-specific output communities were determined after a further four weeks of co-incubation. We also inoculated the L SynCom alone in clay and determined output SynComs (designated 'L in clay', Figure 7.5A).

We found significant differences between leaf-associated output communities of the 'RSL in clay' and 'RS in clay' experiments (Figure 7.5B; $P < 0.001$; Supporting Material). and that the output community on leaves after 'L in clay' inoculation is similar to the leaf outputs of 'RSL in clay' inoculation (Figure 7.5B; $P < 0.001$; Supporting Material), indicating that in this comparison the leaf-derived SynCom has a stronger influence on leaf microbiota structure than root- and soil-derived bacteria. However, both 'RSL in clay' and 'L in clay' leaf outputs are significantly different from the leaf output of the 'L spray' experiment (Figure 7.5B; $P < 0.001$, Supporting Material), showing that many leaf-derived isolates do not successfully colonize leaves when only inoculated in the clay environment. For example, of the top 16 genera a total of three are grossly underrepresented in leaf outputs of the 'RSL in clay' compared to the 'RSL in clay & L+15R spray' experiment (Chryseobacterium, Sphingomonas and Variovorax). and these three genera are abundant in the natural leaf microbiota (Supporting Material). Finally, leaf outputs were strikingly similar between 'RSL in clay & L+15R spray' and 'L spray' only experiments (Figure 7.5B; Supporting Figure 7.7), indicating that the L+15R SynCom, leaf spray-inoculated three weeks after RSL application to clay, can displace the RSL leaf output. Collectively, these results support the hypothesis that leaf microbiota establishment benefits from air- and soil-borne inoculations (Vorholt, 2012; Maignien *et al.*, 2014), although we note that our single application of bacteria to leaves does not mimic the continuous exposure of plant leaves to airborne microorganisms in nature.

A comparison of the root-associated community outputs of the experiments described above revealed that the 'RSL in clay' experiment is more similar to root outputs of the 'RS in clay' than 'L in clay' experiments (Figure 7.5C; $P < 0.001$; Supporting Material), suggesting that the root- and soil-derived SynCom has a stronger influence on root microbiota structure than the leaf-derived SynCom. In this experiment the fractional contribution of root-specific indicator OTUs increases in the output, but decreases for leaf-specific indicator OTUs, relative to their input, pointing to a potential adaptation of root-derived bac- teria for root colonization (Supporting Material; Mann-Whitney; $P < 0.05$). This is further supported by the observation that in the 'RSL in clay' experiment root colonization rates for root-specific indicator OTUs are higher compared to those specific for leaves when applying a 0.1% relative abundance threshold in at least one biological replicate (69% and 33%, respectively). Taken together, this suggests that root-derived bacteria are better adapted to colonize their cognate host niche than leaf-derived bacteria. Further comparisons of the root-associated output communities of the 'L in clay' and 'L spray' experiments (Figure 7.5C; Supporting Figure 7.7) revealed similar community composition, indicating convergence of ectopic root-associated community outputs despite different inoculation time points or sites of application. Additional reciprocal transplantation experiments using a 'R' (root strains only) SynCom either applied to clay ('R in clay') or by spray inoculation ('R spray') confirmed the convergence of ectopic community outputs also for root-derived bacteria on leaves (Supporting Material). Convergence of ectopic SynCom outputs is consistent with the hypothesis that a subset of leaf and root colonizing bacteria has the potential to relocate between leaves and roots.

7.4 Conclusions

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

7.5 Methods

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

A. thaliana plants were either harvested from natural populations or grown in different natural soils and used for bacterial isolations by colony picking, limiting dilution or bacterial cell sorting as well as 16S rRNA gene-based community profiling. To obtain a library of representative root colonizing bacteria, *A. thaliana* plants were grown in different soils (50.958 N, 6.856 E, Cologne, Germany; 52.416 N, 12.968 E, Golm, Germany; 50.982 N, 6.827 E, Widdersdorf, Germany; 47.941 N, 04.012 W, Saint-Evarzec, France; 48.725 N, 3.989 W, Roscoff, France) and harvested before bolting. Briefly, *Arabidopsis* roots were washed twice in washing buffers (10 mM MgCl2 for limiting dilution and PBS for colony picking (Bulgarelli *et al.*, 2012) on a shaking platform for 20 min at 180 rpm and then homogenized twice by Precellys (Edgar, 2013) tissue lyser (Bertin Technologies) using 3 mM metal beads at 5,600 rpm for 30 s. Homogenates were diluted and used for isolation approaches on several bacterial growth media (Supporting Material). For isolations based on colony picking, diluted cell suspensions were plated on solidified media and incubated, before isolates of plates containing less than 20 colony-forming units (CFUs) were picked after a maximum of two weeks of incubation. For limiting dilution, homogenized roots from each root pool were sedimented for 15 min and the supernatant was empirically diluted, distributed and cultivated in 96-well microtitre plates (Goodman *et al.*, 2011). In parallel to the isolation of root-derived bacteria, roots of plants grown in Cologne soil were harvested and used to assess bacterial diversity by culture-independent 16S rRNA gene sequencing. Additionally, soil-derived bacteria were extracted from unplanted Cologne soil by washing soil with PBS buffer, supplemented with 0.02% Silwet L-77 and subjected to bacterial isolation as well as 16S rRNA gene community profiling. For the isolation of representative phyllosphere strains, naturally grown *Arabidopsis* plants were collected at eight different sites in southern Germany and Switzerland (six main sampling sites used for bacterial isolations and community profiling: 47.4090306 N, 8.470169444 E, Hoengg, Switzerland; 47.474825 N, 8.305008333 E, Baden, Switzerland; 47.4816806 N, 8.217547222 E, Brugg, Switzerland; 48.5560194 N, 9.134944444 E, Farm, Tuebingen, Germany; 48.5989861 N, 9.201655556 E, Haeslach, Germany; 48.602682 N, 9.213247258 E, Haeslach, Germany; and two additional sites only used for bacterial isolation: 47.4074722 N, 8.50825 E, Zurich, Switzerland; 47.4227222 N, 8.548666667 E, Seebach, Switzerland) during spring and autumn of 2013 and used for bacterial isolations as well as 16S rRNA gene profiling. Leaf-colonizing bacteria of individual leaves were washed off by alternating steps of intense mixing and sonication. The suspension was subsequently filtered (CellTrics filters, 10 μM, Partec GmbH, Görlitz, Germany) in order to remove remaining plant or debris particles as well as cell aggregates and applied to cell sorting on a BD FACS Aria III (BD Biosciences) as well as to plating on different media (Supporting Material). All isolates were subsequently stored in 30% or 40% glycerol at -80 degrees C.

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

Parts of *A. thaliana* leaves, roots and corresponding unplanted soil samples used for bacterial isolation were also processed for bacterial 16S rRNA gene community profiling using 454 pyrosequencing. Frozen root and corresponding soil samples were homogenized, DNA was extracted with Lysing Matrix E (MP Biomedicals) at 5,600 rpm for 30 s, and DNA was extracted from all samples using the FastDNA SPIN Kit for soil (MP Biomedicals) according to the manufacturer's instructions. Lyophilized leaf samples were transferred into 2 ml microcentrifuge tubes containing one metal bead and subsequently homogenized twice for 2 min at 25 Hz using a Retsch tissue lyser (Retsch, Haan, Germany). Homogenized leaf material was resuspended in lysis buffer of the MO BIO PowerSoil DNA isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA, USA), transferred into lysis tubes, provided by the supplier, and DNA extraction was performed following the manufacturer's protocol. DNA concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), and subsequently diluted to 3.5 ng μ / l. Bacterial 16S rRNA genes were subsequently amplified (Bulgarelli *et al.*, 2012) using primers targeting the variable regions V5-V7 (799F (Chelius and Triplett, 2001) and 1193R (Bulgarelli *et al.*, 2012), Supporting Material). Each sample was amplified in triplicate by two independent PCR mixtures (a total of 6 replicates per sample plus respective no template controls). PCR products of triplicate were subsequently combined, purified and subjected to 454 sequencing. Obtained sequences were demultiplexed as well as quality and length filtered (average quality score $\geq$25, minimum length 319 bp with no ambiguous bases and no errors in the barcode sequences allowed) (Caporaso *et al.*, 2010). High-quality sequences were subsequently processed using the UPARSE24 pipeline and OTUs were taxonomically classified using the Greengenes database28 and the PyNAST (Caporaso *et al.*, 2010) method.

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

We adopted a two-step barcoded PCR protocol (Goodman *et al.*, 2011) in combination with 454 pyrosequencing to define V5-V8 sequences of bacterial 16S rRNA genes of all leaf, root- and soil-derived bacterial (Supporting Figure 7.6). DNA of isolates was extracted by lysis of 6 μ l of bacterial cultures in 10 μ l of buffer I containing 25 mM NaOH, 0.2 mM EDTA, pH 12 at 95 degrees C for 30 min, before the pH value was lowered by addition of 10 μ l of buffer II containing 40 mM Tris-HCl at pH 7.5. Position and taxonomy of isolates in 96-well microtitre plates were indexed by a two-step PCR protocol using the degenerate primers 799F and 1392R containing well- and plate-specific barcodes (Supporting Material) to amplify the variable regions V5 to V8. During the first step of PCR amplification, DNA from 1.5 μ l of lysed cells was amplified using 2 U DSF-Taq DNA polymerase, 1x complete buffer (both Bioron GmbH), 0.2 mM dNTPs (Life technologies), 0.2 μ M of 1 of 96 barcoded forward primer with a 18-bp linker sequence (for example, A1_454_799F1_PCR1_wells). and 0.2 μ M reverse primer (454B_1392R) in a 25 μ l reaction. PCR amplification was performed under the following conditions: DNA was initially denaturised at 95 degrees C for 2 min, followed by 40 cycles of 95 degrees C for 30 s, 50 degrees C for 30 s and 72 degrees C for 45 s, and a final elongation step at 72 degrees C for 10 min. PCR products of each 96-well microtitre plate were combined and subsequently purified in a two-step procedure using the Agencourt AMPure XP Kit (Beckman Coulter GmbH, Krefeld, Germany) first, then DNA fragments were excised from a 1% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). DNA concentration was measured by Nanodrop and diluted to 1 ng μ / l.

During the second PCR step, 1 ng of pooled DNA (each pool represents one 96-well microtitre plate) was amplified by 1.25 U PrimeSTAR HS DNA Polymerase, 1x PrimeSTAR Buffer (both TaKaRa Bio S.A.S, Saint-Germain-en-Laye, France), 0.2 mM dNTPs (Thermo Fisher Scientific Inc.), 0.2 μ M of 1 of 96 barcoded forward primer targeting the 18-bp linker sequence (for example, P1_454_PCR2). and 0.2 μ M reverse primer (454B_1392R) in a 50 μ l reaction. The PCR cycling conditions were as follows. First, denaturation at 98 degrees C for 30 s, followed by 25 cycles of 98 degrees C for 10 s, 58 degrees C for 15 s and 72 degrees C for 30 s, and a final elongation at 72 degrees C for 5 min. PCR products were purified using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and QIAquick Gel Extraction Kit (Qiagen) as described for the purification of first step PCR amplicons. DNA concentration was determined by PicoGreen dsDNA Assay Kit (Life technologies) and samples were pooled in equal amounts. The final PCR product libraries were sequenced on the Roche 454 Genome Sequencer GS FLX+. Each sequence contained a plate-barcode, a well-barcode and V5-V8 sequences. The sequences were quality filtered, demultiplexed according to well and plate identifiersi (Caporaso *et al.*, 2010). OTUs were clustered at 97% similarity by UPARSE algorithm (Edgar, 2013). Nucleotide-based blast (v. 2.2.29) was used to align representative sequences of isolated OTUs to culture-independent OTUs and only hits ≥97% sequence identity covering at least 99% of the length of the sequences were considered.

7.5.4 Preparation of *A. thaliana* leaf, root and soil bacterial culture collections

Based on representative sequences of OTUs from this as well as previously published culture-independent community analysis, bac- terial CFUs in the culture collections with ≥97% 16S rRNA gene identity to root-, leaf- and soil-derived OTUs were purified by three consecutive platings on the respective solidified media before an individual colony was used to inoculate liquid cultures. These liquid cultures were used for validation by Sanger sequencing with both 799F and 1392R primers as well as for the preparation of glycerol stocks for the culture collections and for the extraction of genomic DNA for whole-genome sequencing. A total of 21 leaf-derived strains, previously described as phyllosphere bacteria (Vorholt, 2012; Bodenhausen *et al.*, 2013), were added to the *At*-LSPHERE collection although these were undetectable in the present culture-independent leaf community profiling.

7.5.5 Preparation of bacterial genomic DNA for whole-genome sequencing

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

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

7.5.6 Genome assembly and annotation

Paired-end Illumina reads were subjected to quality and length trimming using Trimmomatic v. 0.33 (Bolger *et al.*, 2014) and assembled using two independent methods: A5 (Tritt *et al.*, 2012) and SOAPdenovo v. 20.1 (Li, 2010). In each case, the assembly with the smaller number of scaffolds was selected. Detailed assembly statistics for each sequenced isolate can be found in the Supporting Material. Identification of putative protein-encoding genes and annotation of the genomes were performed using GLIMMER v. 3.02 (Delcher *et al.*, 1999). Functional annotation of genes was conducted using Prokka v. 1.11 (Seemann, 2014) and the SEED subsystems approach using the RAST server API (Overbeek, 2005). Additionally, annotation of KEGG Orthologue (KO) groups was performed by first generating HMM models for each KO in the database

(Kanehisa and Goto, 2000; Kanehisa, 2014) the HMMER toolkit (v. 3.1b2) (Eddy, 2011). Next, we employed the HMM models to search all predicted ORFs using the hmmsearch tool, with an E value threshold of 10E10-5. Only hits covering at least 70% of the protein sequence were retained and for each gene and the match with the lowest E value was selected.

7.5.7 Analyses of phylogenetic diversity within sequenced isolates

Each proteome was searched for the presence of the 31 well-conserved, single-copy, bacterial AMPHORA genes (Wu and Eisen, 2008), designed for the purpose of high-resolution phylogeny reconstruction of genomes. Subsequently, a concatenated alignment of these marker genes was performed using Clustal Omega (Sievers, 2011) v. 1.2.1. Based on this multiple sequence alignment, a species tree was inferred using FastTree (Price *et al.*, 2010) v. 2.1, a maximum likelihood tool for phylogeny inference. Whole-genome taxonomic classification of sequenced isolates was conducting using taxator-tk (Droege *et al.*, 2015), an homology/based tool for accurate classification of sequences. Analyses of phylogenetic diversity were performed independently for each cluster based on pairwise tree distances between all isolates (Supporting Material).

7.5.8 Analyses of functional diversity between sequenced isolates

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

7.5.9 Recolonization experiments of leaf-, root- and soil-derived bacteria on *Arabidopsis*

Calcined clay (Lebeis *et al.*, 2015), an inert soil substitute, was washed with water, sterilized twice by autoclaving and heat-incubated until being completely dehydrated. *A. thaliana* Col-0 seeds were surface-sterilized with ethanol and stratified overnight at 4 degrees C. Leaf-, root- and soil-derived bacteria of the culture collections were cultivated in 96-deep-well plates and subsequently pooled (in equal or unequal ratios) in order to prepare synthetic bacterial communities (SynComs) for inoculations below the carrying capacity of leaves and roots (Whitman *et al.*, 1998; Bodenhausen *et al.*, 2014). To inoculate SynComs into the calcined clay matrix, OD600 was adjusted to 0.5 and 1 ml (∼2.75 x 108 cells) was added to 70 ml 0.5i x MS media (pH 7; including vitamins, without sucrose), and mixed with 100 g calcined clay in Magenta boxes (∼2.75 x 106 cells per gr calcined clay), directly before sowing of surface-sterilized seeds. Plants were grown at 22 degrees C, 11 h light, and 54% humidity. Alive cell counts (CFUs) of root-associated bacteria by serial dilutions of root homogenates after seven weeks of co-incubation were 1.4 x 108 ± 8.4 x 107 cells per gram root tissue. For leaf spray-inoculation of *A. thaliana* plants, bacterial SynComs were prepared as described above and adjusted to OD600 0.2, before the solution was diluted tenfold and 170 μ l (∼1.87 x 106 cells) were sprayed into each magenta box containing four three-week-old plants using a TLC chromatographic reagent sprayer (BS124.000, Biostep GmbH, Jahnsdorf, Germany). The average volume per spraying event was determined by spraying repeatedly into 50 ml tubes and weighing before and after. All plants and corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. All plants and corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. During harvest, leaves and roots of individual plants were carefully separated using sterilized tweezers and scissors to avoid cross-contamination and processed separately thereafter. All leaves being obviously contaminated with clay particles or touching the ground were carefully removed and omitted from further processing. Remaining aerial parts of four plants collected from one magenta box were combined and transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at -80 degrees C until used for DNA extraction. Roots from one Magenta box were pooled, washed twice in 5 ml PBS at 180 rpm for 20 min, dried on sterilized Whatman glass microfibre filters (GE Healthcare Life Sciences), transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at -80 degrees C until further processing.

The corresponding unplanted clay samples were washed in 100 ml PBS supplemented with 0.02% Silwet L-77 at 180 rpm for 10 min, before particles were allowed to settle down for 5 min. The supernatant was collected by centrifugation at 3,220g for 15 min. The pellet was subsequently resuspended in 1 ml water, transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at -80 degrees C.

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

Forward and reverse reads were joined, demultiplexed and subjected to quality controls using scripts from the QIIME toolkit (Caporaso *et al.*, 2010), v. 180 (Phred $\geq$ 20). The resulting high quality sequences were further clustered at 97% sequence identity together with Sanger sequences of leaf, root and soil isolates using the UPARSE (Edgar, 2013) pipeline as described above. Taxonomic assignments of representative sequences were performed as explained in the previous sections. OTUs only corresponding to one or more Sanger 16S rRNA gene sequence(s) of purified strains in the *At*-RSPHERE, *At*-LSPHERE or soil collection were selected and designated 'indicator OTUs'. The heat maps were generated using the ggplot2 R package.

7.5.10 Accession numbers

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

7.5.11 Code availability

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

7.6 Author contributions

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

7.7 Acknowledgements

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

7.8 Supporting material

A subset of the supporting figures related to this section can be found bellow. The reminder supporting material corresponding to this section, including all supplementary tables and figures can be accessed via the online version of the published article and have not been included in this thesis due to space limitations. The sequenced bacterial genomes as well as any future updates are available at `http://www.at-sphere.com`. All intermediate data as well as the scripts used to analyze the data and generate the figures of this study are available at `http://www.mpipz.mpg.de/R_scripts`.

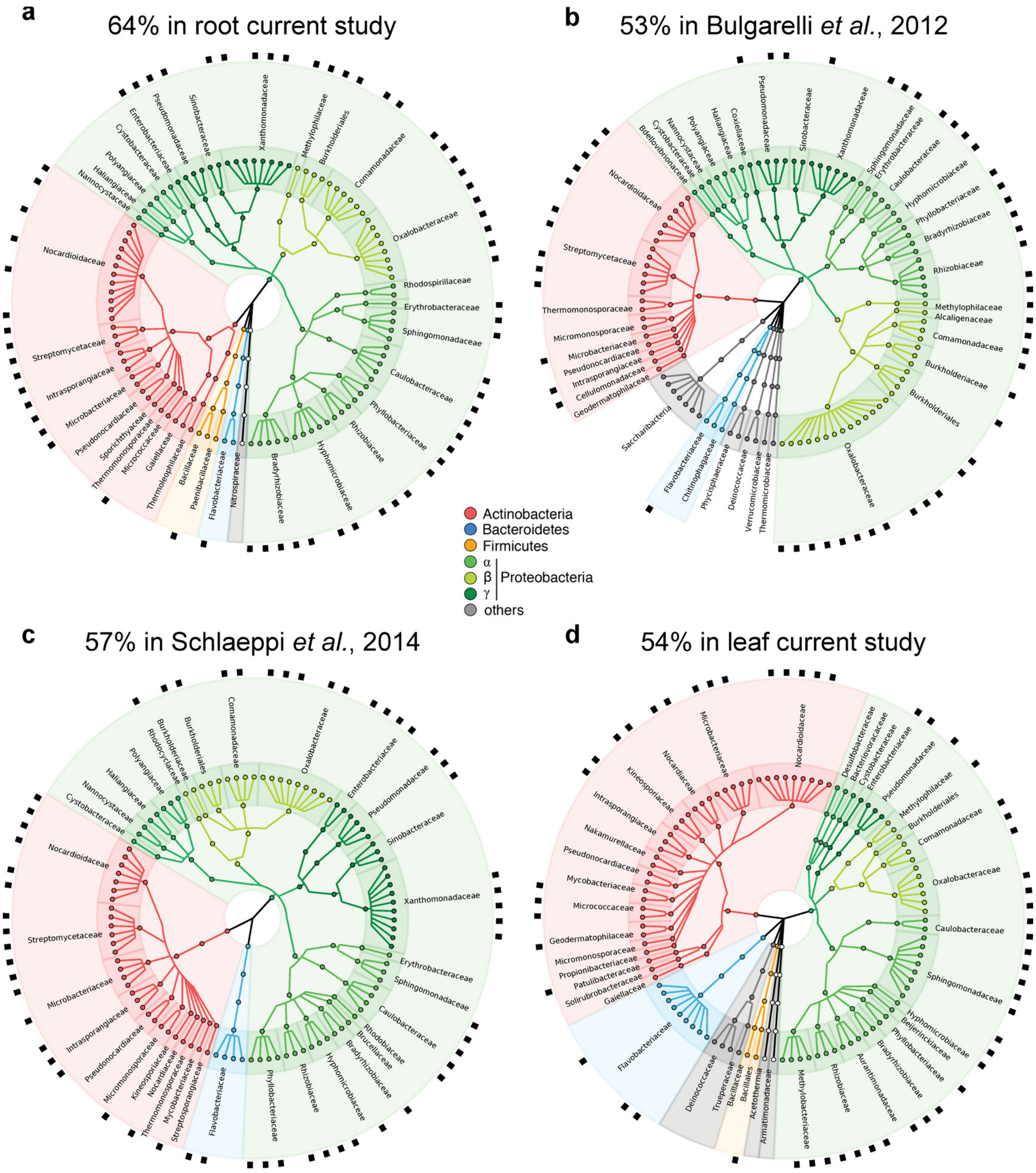


Figure 7.6: Culture-dependent coverage of *A. thaliana* root- and leaf-associated OTUs identified in several cultivation-independent studies. The inner circle depicts taxonomic assignments of top 100 root-associated OTUs (filled dots) for the indicated phyla and families that were identified in the current (a), (Bulgarelli *et al.*, 2012) (b) and (Schlaeppi *et al.*, 2014) (c) studies with Cologne-soil-grown plants, and current leaf (d) study at locations around Tübingen and Zurich. Black squares of the outer ring highlight OTUs sharing ≥97% 16S rRNA gene similarity to *Arabidopsis* root or leaf bacterial culture collection.

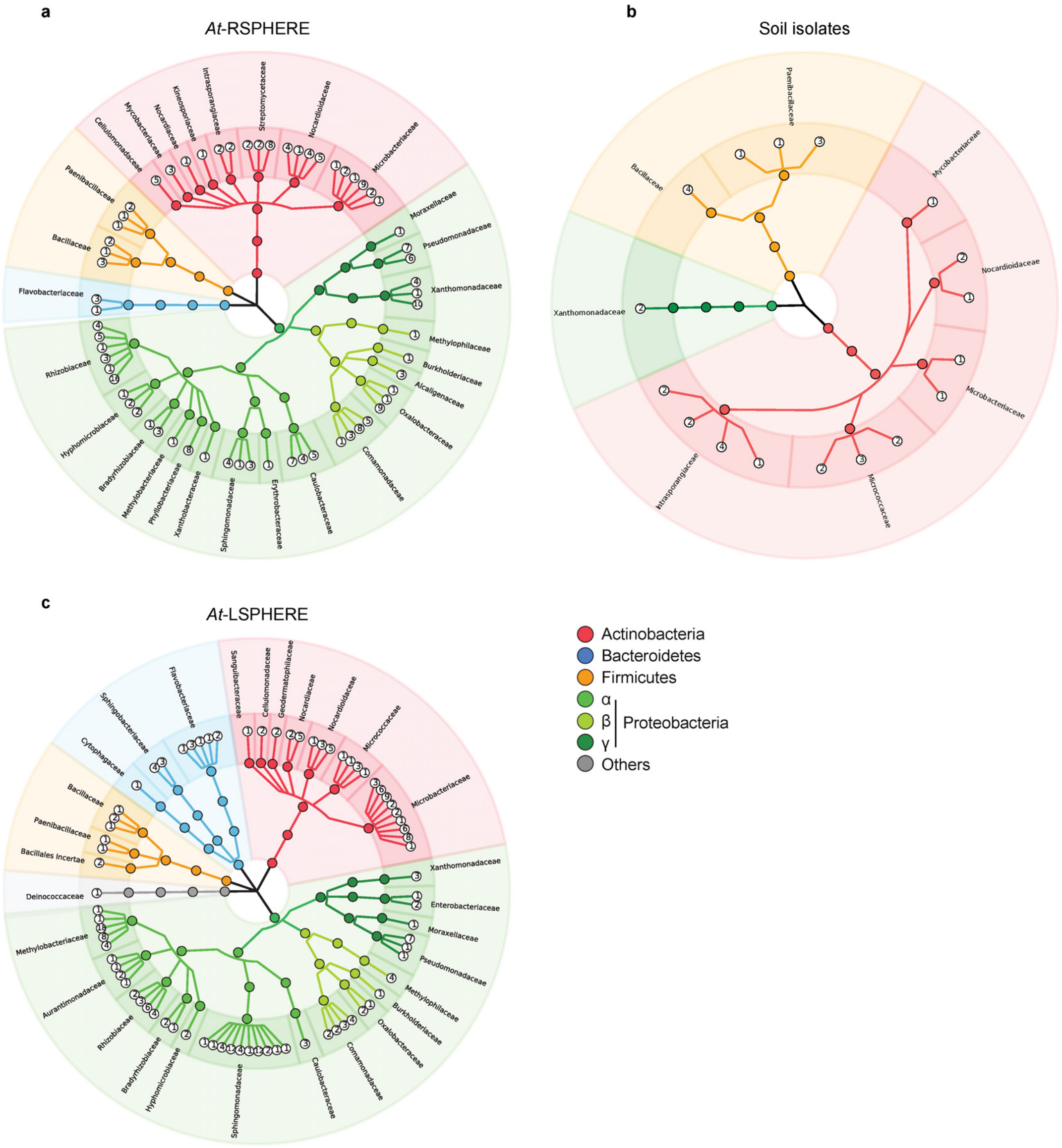


Figure 7.7: *At*-RSPHERE, *At*-LSPHERE and soil bacterial culture collections. a, *At*-RSPHERE (n=206 isolates), a culture collection of the *A. thaliana* root microbiota. b, *At*-LSPHERE (n=224 isolates), a culture collection of the *A. thaliana* leaf microbiota. c, Bacteria isolated from Cologne soil (n=33 isolates). Numbers inside white circles indicate the number of bacterial isolates sharing ≥97% sequence identity, but isolated from independent roots, leaves and soil batches.

Part III

Comparative genomics of Rhizobia

CHAPTER 8

Assessment of functional diversification and adaptation in Rhizobia by comparative genomics

Status	**In preparation**
Citation	Garrido-Oter, R. *, Nakano, R.T. *, Dombrowski, N. *, McHardy, A.C. and Shulze-Lefert, P. (2016). Assessment of functional diversification and adaptation in Rhizobia by comparative genomics * joint first authors
Own contribution	Conceived research (with co-authors) Analyzed the data (with co-authors) Interpreted the data (with co-authors) Performed the computational experiments Wrote the manuscript

8.1 Abstract

Rhizobia are a paraphyletic group of soil-borne bacteria defined by their ability to induce nodule development on legume roots and fix atmospheric nitrogen to deliver bioavailable ammonium for plant growth under nitrogen-limiting conditions. Although this form of symbiosis only occurs between certain combinations of rhizobia and leguminous plants, recent studies have identified species within the Rhizobiales order as core components of the plant microbiota. The robust membership of rhizobial species in the root microbiota across taxonomically distant plant species (ranging from monocots to dicots) suggests the presence of conserved function(s) of rhizobia besides nodulation and nitrogen fixation. To address this possibility, we have performed a large-scale comparative genomic study utilizing more than 1,300 whole-genome sequences of isolates within the order Rhizobiales, including previously characterized nodulating species isolated from legumes as well as more than 940 newly sequenced non-nodulating and non-fixing exemplars isolated from a variety of non-leguminous plants and related environmental sources. By focusing on the set of genes required for nodulation and nitrogen fixation, we found that these traits have been acquired multiple independent times within each sublineage, suggesting the existence of an ancestral form of association with plant hosts, independent of nodule formation and nitrogen fixation. Next, we utilized a subset from this collection of cultured isolates in experiments with germ-free *Arabidopsis thaliana* plants with the aim of identifying potential physiological functions of rhizobia in association with a non-leguminous plant. Our results illustrate that the majority of tested root-associated rhizobia are able to colonize and promote root growth in *Arabidopsis* without nodulation or fixing nitrogen. We propose that rhizobial root colonization and root growth promotion in associations with flowering plants are ancestral traits and that the capacity for nodulation and nitrogen fixation was acquired in multiple subsequent events, most likely *via* horizontal gene transfer, thereby constituting an example of convergent evolution.

8.2 Introduction

Land plants must acquire nitrogen from the surrounding soil predominantly in the form of nitrate or ammonia to sustain growth. Legumes have developed a strategy for survival in nitrogen-poor soils that consists of engaging in beneficial interactions with members of several bacterial genera within the order Rhizobiales, collectively known as rhizobia, that are capable of converting atmospheric nitrogen (N_2) into ammonia.

Understanding the evolutionary history of this innovation, which allows legumes to thrive in habitats with limited biologically active forms of nitrogen (Peoples *et al.*, 2009; Batterman *et al.*, 2013; Adams *et al.*, 2016), is a crucial component in the development of new bio-fertilizers and of sustainable agriculture. Unlike the association with the taxonomically diverse bacterial communities that constitute the plant microbiota, (Bulgarelli *et al.*, 2013; Vorholt, 2012; Hacquard *et al.*, 2015) these binary relationships with nitrogen-fixing rhizobia are highly-specific and require the exchange of multiple signaling molecules that coordinate recognition of compatible symbionts (Radutoiu *et al.*, 2007; Broghammer *et al.*, 2012; Oldroyd, 2013) that initiate the process of nodule organogenesis and the subsequent reprogramming of the root transcriptional and metabolic status (Oldroyd *et al.*, 2011; El Yahyaoui *et al.*, 2004; Colebatch *et al.*, 2004; Hogslund *et al.*, 2009; Nakagawa *et al.*, 2011). Although this form of symbiosis only occurs between certain combinations of rhizobia and leguminous plants, recent efforts in the characterization of the microbial communities associated with various plant hosts have identified species within the Rhizobiales order as core components of the root plant microbiota (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Edwards *et al.*, 2015; Dombrowski *et al.*, 2016). The presence of members of the Rhizobiales lineage in the root microbiota across taxonomically distant plant species (ranging from monocots to dicots) suggests the existence of conserved function(s) besides nodulation and nitrogen fixation and a shared capability to colonize the root and rhizosphere environments. However, the low taxonomic resolution of marker gene amplicon data and the lack of genomic and functional information provide only a limited understanding of the taxonomy and the functions of these root- and rhizosphere-inhabiting rhizobia. Furthermore, efforts in the isolation and whole-genome sequencing of rhizobia have been largely confined to exemplars derived from legume nodules, introducing a strong bias in the representation of these bacterial lineages in the databases. This bias and the limited number of sequenced representatives of certain genera have made the reconstruction of the evolutionary history of these commensal and mutualistic interactions exceedingly difficult (Tian *et al.*, 2012).

Here, we introduce a large-scale isolation and sequencing effort resulting in high-quality whole-genome assemblies of 904 exemplars of rhizobia originating from a panel of taxonomically distant plant hosts, including non-legumes as well as associated environmental sources. These newly-sequenced isolates cover the majority of the known sublineages of rhizobia as well as potentially uncharacterized phylogenetic groups, providing novel insights into the taxonomic and genomic diversity of this ecologically relevant bacterial clade. Analyses of these sequences, together with a set of 370 high-quality genome assemblies of nitrogen-fixing legume symbionts retrieved from public databases, and 16S rRNA gene amplicon data from five previous studies, revealed that the ability to colonize plant roots is an evolutionarily conserved and ancestral feature of rhizobia, which is not limited to nitrogen-fixing nodule symbionts. Experiments employing a subset of strains from this collection with germ-free *Arabidopsis thaliana* plants indicate that the majority of rhizobia are able to colonize the roots of and promote root growth in binary associations with a non-legume host. Finally, by reconstructing the evolutionary history of the genes required for nodule symbiosis and nitrogen-fixation (*nod*, *fix* and *exo* genes) we found evidence indicating that these phenotypes have been acquired multiple independent times within each sublineage of rhizobia, providing an example of convergent evolution in bacteria. Our results suggest the existence of an ancestral form of association of rhizobia with plants that predates the acquisition of the genetic toolkit necessary for nodule formation and nitrogen fixation, raising the possibility that root- and rhizosphere- competence are ecologically relevant traits also for legume symbionts.

8.3 Results

8.3.1 Rhizobia are important components of the core root microbiota across a wide taxonomic variety of plant hosts

In this study we present annotated whole-genome assemblies of 943 newly sequenced strains of rhizobia with representatives covering all major plant-associated taxonomic groups isolated from a variety of hosts, including roots and leaves of more than 10 plant species as well as other associated environmental sources such as soil (4.77%), nematodes (0.11%) or insects (6.47%) (Supporting Information). In addition, we have retrieved and annotated publicly available genome assemblies of 370 strains of nitrogen-fixing rhizobia from various species, the majority of which were isolated from functional legume nodules. In order to explore the evolutionary origin of the symbiotic or commensal relationship of these bacterial species with their plant hosts, we performed a comparative genomic analysis of this dataset, which includes multiple representative strains from previously uncharted taxa as well as close relatives of well-studied legume nodule symbionts isolated from soil and from the roots and leaves of non-legume hosts such as *Arabidopsis thaliana*, rice or corn.

First, we sought to assess the distribution and ecological prevalence of the sequenced taxonomic groups across various soil types and plant host species. In an attempt to link the genome sequences with environmental data, we retrieved and re-analyzed the raw sequences of five previous large-scale 16S rRNA gene surveys (Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Bai *et al.*, 2015; Zgadzaj *et al.*, 2016) covering a taxonomically broad set of hosts, including *Arabidopsis thaliana* and relative species (*Cardamine hirsuta*, *Arabidopsis halleri* and *Arabidopsis lyrata*), barley (*Hordeum vulgare*), and the model legume *Lotus japonicus*. This dataset includes more than 450 samples that use the same set of sequencing primers targeting the V5-V7 hypervariable regions of the 16S rRNA gene and are distributed across five compartments (soil, root, rhizosphere, leaf and legume nodule) of plants grown in 6 different soil types (Supporting Information). First, we extracted full-length 16S sequences from the whole-genome assemblies of the sequenced rhizobial strains (Supporting Information). Subsequently, we performed a multiple sequence alignment, extracted the regions targeted by the set of primers employed in the amplicon taxonomic surveys and collapsed identical sequences into 216 high-resolution (≥ 1 SNP) Operational Taxonomic Units (OTUs; tree in Figure 8.1). We then employed these sequences as a reference template to cluster the environmental 16S reads using a stringent ($\geq 99\%$ sequence identity) threshold and assessed the abundance of each OTU across different compartments and hosts (Figure 8.1). By calculating the ratio of assigned to unassigned reads, we were able to assess the total abundance of rhizobia in each condition (boxplots in Figure 8.1) and to calculate the normalized relative abundance of each OTU in the dataset (barplots in Figure 8.1). This analysis confirmed that rhizobia are a component of the core plant root and leaf microbiota, in line with previous reports (Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Hacquard *et al.*, 2015), and showed that a number of OTUs, consisting of members from a broad taxonomic range and diverse functional capability are consistently present in large relative abundances in the root and rhizosphere of all studied hosts.

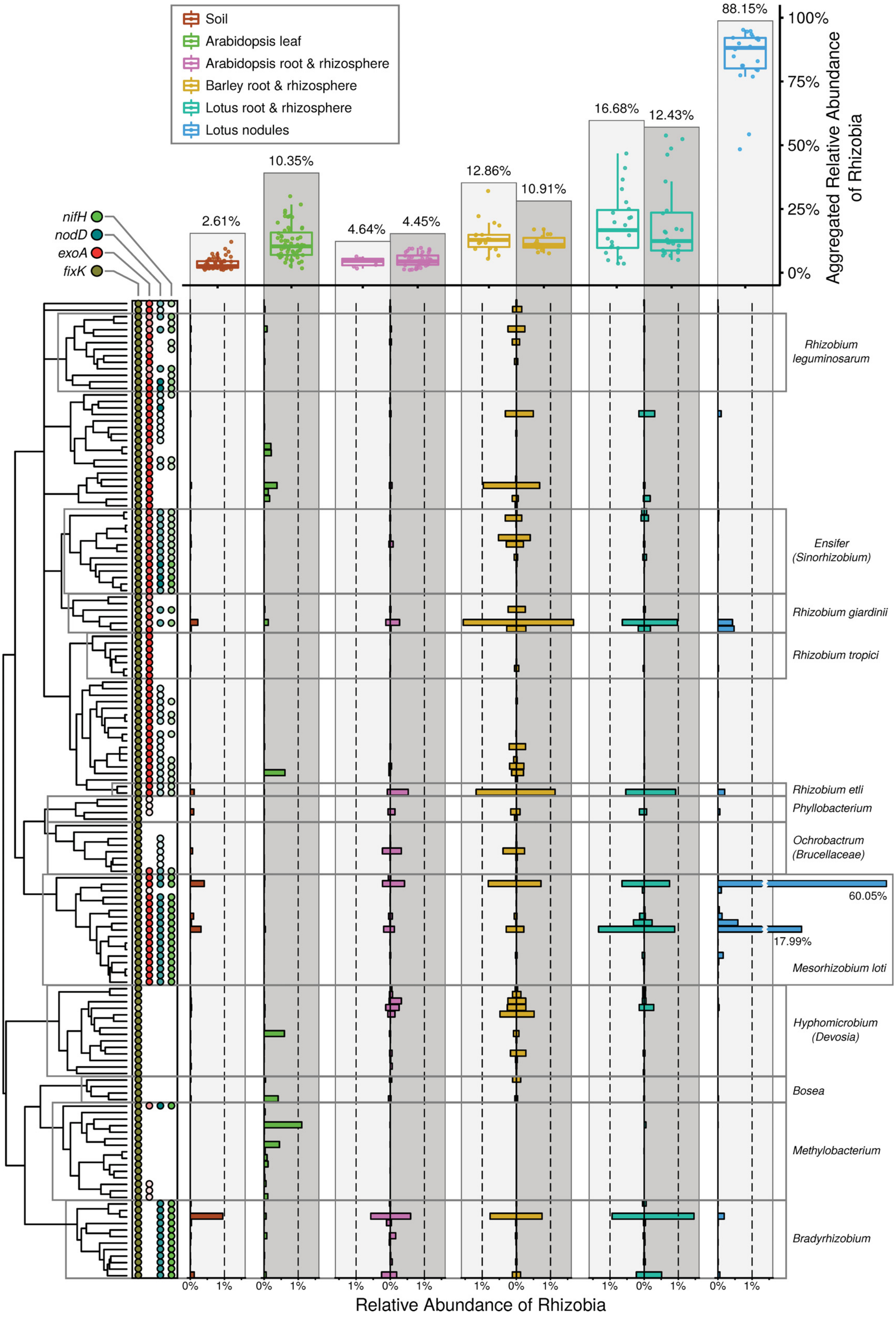


Figure 8.1: Rhizobia are important components of the core root microbiota across a wide taxonomic variety of hosts. (Caption on next page).

Figure 8.1: Rhizobia are important components of the core root microbiota across a wide taxonomic variety of hosts. (Figure on previous page).
Analysis of high-resolution (≥ 99% sequence identity) OTUs using data from 5 previous 16S rRNA gene amplicon surveys covering root, rhizosphere and nodule samples of a taxonomically diverse panel of plant hosts grown in a variety of natural and agricultural soils. Boxplots (top) illustrate accumulated relative abundances of rhizobia in each host / compartment (n=453). Phylogenetic tree (bottom left pannel) of 216 representative sequences corresponding to all non-identical V5-V7 16S rRNA gene sequences found in the whole-genome dataset. Colored circles (bottom middle pannel) displaying to the prevalence of genes relevant for symbiosis in each rhizobial OTU. Transparency of each circle corresponds to the percentage of genomes within each OTU where each gene is present. Barplots (bottom right pannel) show relative abundances of each OTU across host and compartment. Note that the scale of the two most abundant OTUs found in *Lotus* nodules is not to scale.

We observed that rhizobia from a wide taxonomic range, including those found in the root-associated compartments, are also present in soil samples (∼2.5% aggregated RA) although at lower abundances compared to the root-associated compartments. The largest contribution to the root and rhizosphere bacterial community was found in *Lotus japonicus* samples (∼12-17% RA). Interestingly, the presence of rhizobia in *Lotus* roots was not limited to its compatible symbiont *Mesorhizobium loti* but extended to other taxa, including OTUs belonging to other Rhizobiaceae and Bradyrhizobiaceae species (Figure 8.1). As expected, the largest contribution was observed in the *Lotus* nodule samples, where rhizobia account for more than 88% of the overall bacterial community and which is dominated by two distinct *Mesorhizobium* high-resolution OTUs. This results are in line with previously reported findings obtained by analyzing the entire bacterial community associated to *Lotus japonicus* roots (Zgadzaj *et al.*, 2016). Of all hosts, barley harbors the most complex rhizobial community, which includes members of all major taxonomic groups and accounts for ∼11-13% of all bacteria, followed by the *Arabidopsis* root rhizobial community at approximately 5% aggregated RA, which was found to be only marginally above that of the soil samples. Comparison of rhizobial OTUs across compartments revealed that strains which are undetectable in soil are nevertheless able to colonize the root-associated compartments. These OTUs become in some cases abundant community members (Figure 8.1), indicating a high level of specialization in colonizing their respective host-associated environmental niches. Similarly, a narrower subset of rhizobia is able to colonize *Arabidopsis* leaves, which are mainly, but not limited to Methylobacteria (in the Bradyrhizobiaceae family). Other rhizobial OTUs are also present in the leaf compartment (up to ∼5% aggregated RA) but unlike Methylobacteria are also present the root compartment, indicating that they might originate from soil.

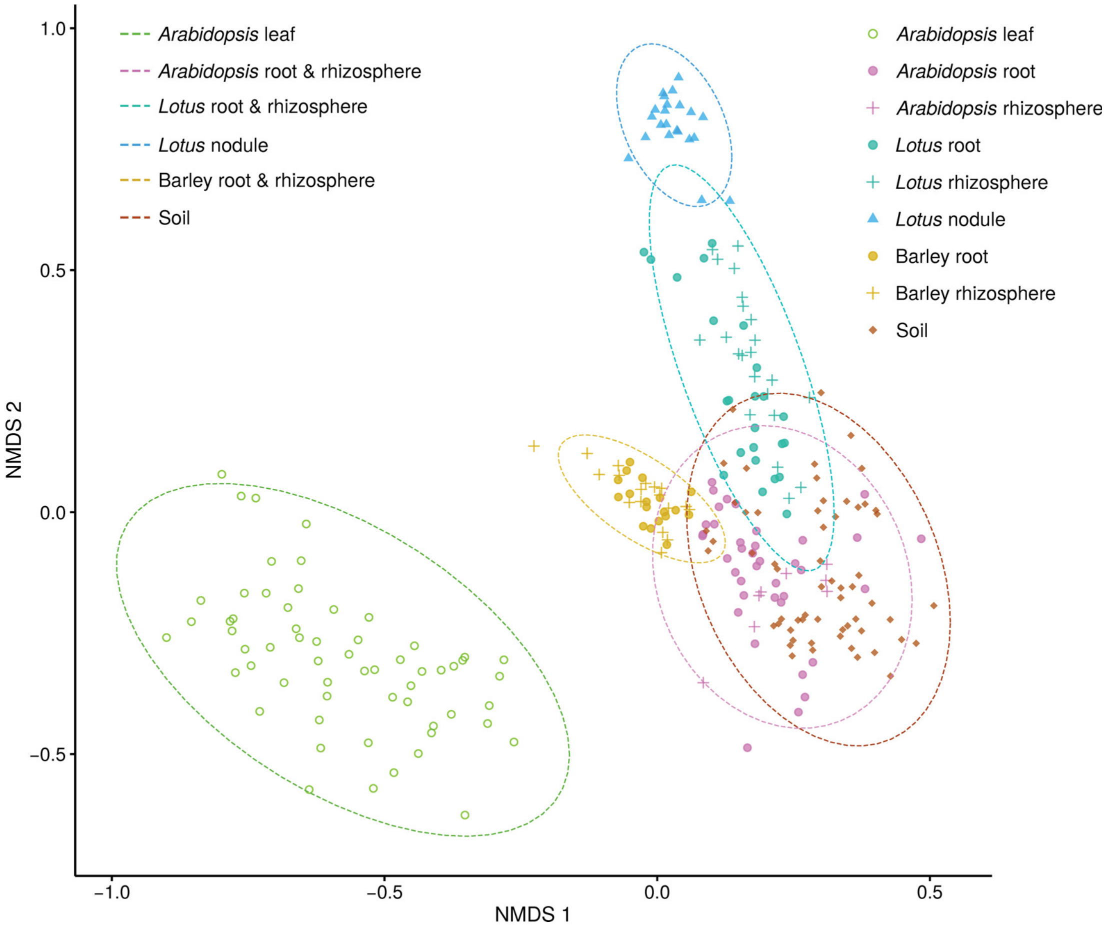


Figure 8.2: Non-metric Multidimensional Scaling (NMDS) of Bray-Curtis distances between rhizobial abundances. Analysis of beta-diversity of rhizobia between samples (n=453) across hosts (in different colors) and compartments (in different shapes). Dashed lines correpond to a Gaussian distributtion fitted to each cluster (95% confidence interval).

Next, we analyzed the beta-diversity (between sample diversity) of rhizobia present in the culture-dependent taxonomic surveys in order to determine differences at the whole community level between hosts and compartments (Methods). Non-metric Multidimensional Scaling (NMDS) analysis of Bray-Curtis distances between samples (Figure 8.2) revealed a clear separation of the leaf- and root-associated compartments, owing to the differential presence of Methylobacteria which dominated the leaf samples and which were absent from root and rhizosphere. Within the root-associated compartments, *Lotus* nodules were separated from the root, rhizosphere and soil clusters along the *y*-

axis, which is largely explained by a gradient of decreasing *Mesorhizobium* abundance. We observed a substantial overlap between the soil and the *Arabidopsis* root and rhizosphere communities which, despite of varying total abundances (nearly a two-fold increase in *Arabidopsis*), are markedly similar. In contrast, the barley root-associated community constitutes a separate cluster, showing a higher alpha-diversity (within sample diversity) than all of the other communities, including those found in soil, due to the presence of undetectable very low abundant taxa in the soil samples (Figure 8.2). Together, these results indicate that rhizobia are consistently found in the root and leaf communities across a range of plant hosts, including but not limited to nodulating legumes. Meta-analysis of 5 different amplicon marker gene studies indicates that the sequenced isolates include representatives of all major taxonomic groups of rhizobia found in natural communities, demonstrating their ecological relevance. The wide taxonomic spread of isolates found in large relative abundances in root and rhizosphere samples, together with the conserved community composition observed in hosts that diverged more than 200 My ago, indicates that the ability to colonize plant roots is an evolutionarily conserved and ancestral feature of rhizobia, which is not limited to nitrogen-fixing nodule symbionts.

8.3.2 Comparative analysis of sequenced isolates reveals a large pan-genome with high functional diversity

The genomes of rhizobia, not unlike those of other soil-inhabiting bacterial taxa, display a highly mosaic and multipartite structure, consisting of a stable and conserved core as well as accessory components that show high variability between species and even between strains within the same species (Galibert, 2001; Young *et al.*, 2006; Harrison *et al.*, 2010). In the case of rhizobia, these accessory components are typically located on large plasmids or secondary replicons that can encode more than 30% of all proteins present in the genome (Masson-Boivin *et al.*, 2009; Galardini *et al.*, 2013). These 'megaplasmids' exhibit strong signatures of horizontal gene transfer and can be specific of a larger taxonomic group (e.g. the symbiotic plasmid that contains the genes necessary for the establishment of functional nodules) or variable even between strains of the same species (Galardini *et al.*, 2011; Tian *et al.*, 2012). In an attempt to accurately reconstruct the evolutionary origin of the various components of the rhizobial pan-genome, including the symbiosis genomic toolkit, we have conducted a comparative genomics analysis of a dataset containing 943 newly sequenced genomes, a subset of which originates from previously uncharted branches of the Rhizobiales order, together with 370 high-quality and publicly available genome sequences of nitrogen-fixing rhizobia isolated from legume nodules.

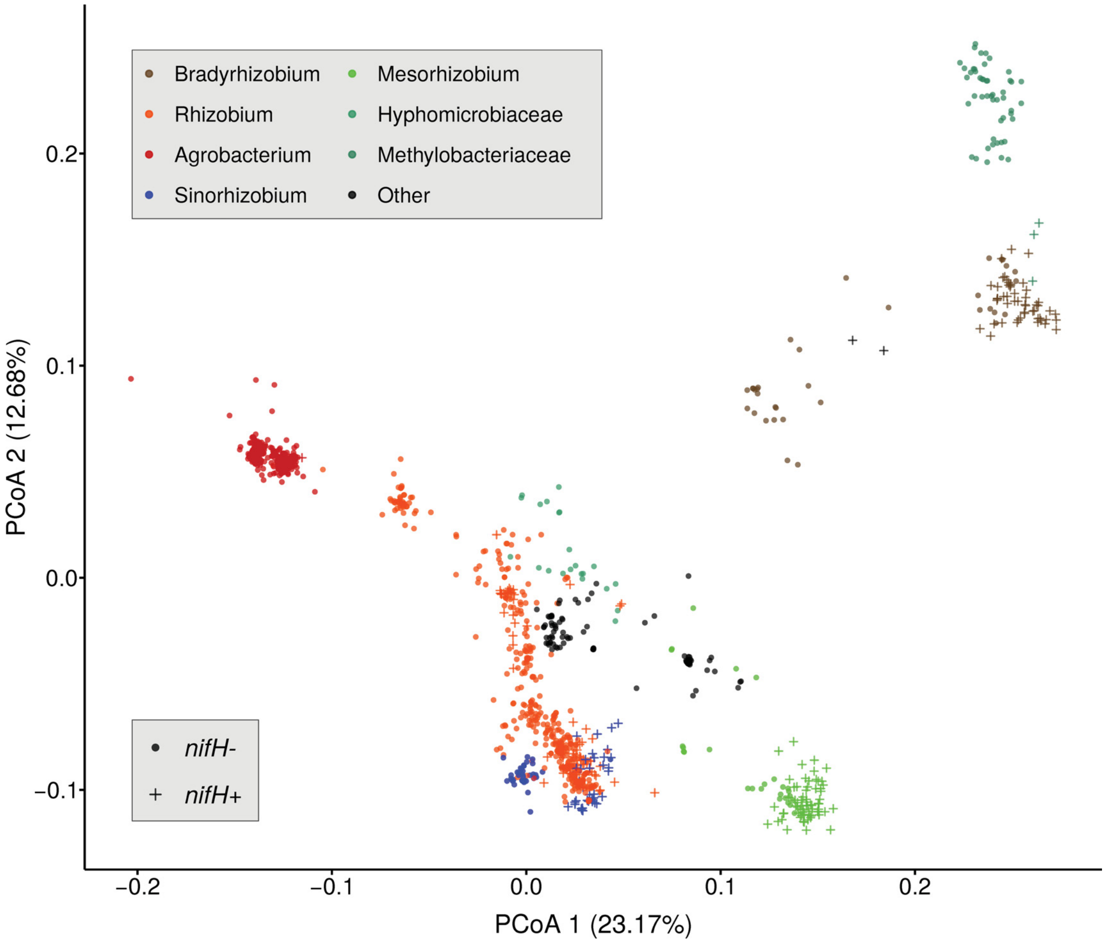


Figure 8.3: Analysis of functional diversity between sequenced genomes of rhizobia. Principal coordinate analysis (PCoA) plot depicting functional distances between sequenced genomes (n=1,313) based on the KEGG Orthology (KO) database annotation. Each point represents a genome, colors correspond to their taxonomic affiliation and shapes depict presence (crosses) or absence (circles) of symbiosis genes using the marker *nifH* gene as a proxy.

First, we annotated the proteomes using the KEGG database or Orthologous groups (KOs; Methods). Genes annotated with the same KO-term were grouped into the same gene family, resulting in a set of 4,935 annotated Clusters of Orthologous Genes (COGs). This number is comparable to the 5,738 annotated gene families found in the *At*-SPHERE dataset (Bai *et al.*, 2015), despite of the much larger phylogenetic diversity of the former, which contains exemplars from 5 different phyla and more than 35 bacterial families (see Chapter 6). This large number of gene families is indicative of an exceedingly high functional diversity. Principal Coordinates Analysis (PCoA) of functional distances between all genomes (Methods) revealed a strong phylogenetic signal, and isolates from the same taxonomic groups tended to cluster together (Figure 8.3). The largest separation was observed between members of the Bradyrhizobiaceae family (mostly consisting of *Methylobacterium* and *Bradyrhizobium* species) and the remaining isolates. Within the larger Rhizobiaceae cluster, *Agrobacterium* and *Mesorhizobium* formed separated functional groups, clearly distinct from the remaining Rhizobiales, which include *Sinorhizobium* genomes and those belonging to various species of *Rhizobium* (e.g. *R. leguminosarum*, *R. etli* or *R. tropicii*). Other taxa such as *Devosia* (Hyphomicrobiaceae), *Ochrobactrum* and *Phyllobacterium* were found to be functionally related to other Rhizobiaceae species, despite of their phylogenetic diversity (black points in Figure 8.3). Interestingly, a subset of *Ochrobactrum* and *Phyllobacterium* strains isolated from insects (Supporting Information) were indistinguishable from other members of the Rhizobiaceae group at the whole genome level, indicating that insects might constitute a vector of transmission of rhizobia. Finally, this analysis revealed that nitrogen-fixing symbionts (crosses in Figure 8.3) do not form a distinct cluster and are not functionally differentiated at the whole genome level from the remaining non-nodulating rhizobia. The genomes of rhizobia isolated from legume nodules are, by contrast, functionally diverse and more similar to non-nodulating rhizobia within the same taxonomic group than to N-fixing symbionts from other taxa. These data, together with the widespread capability to colonize taxonomically distant hosts (Figure 8.1), are indicative of an evolutionarily conserved functional adaptation to the plant-associated niches that is independent from nodule symbiosis. The observed functional diversity further indicated the presence of accessory genome components with high variability across taxa, which was observable even at the level of annotated gene families. Next, we calculated gene accumulation curves using a permutation-based analysis in order to estimate the approximate pan-genome size and to determine to what extent the sequenced genomes are able to cover the gene space of rhizobia (Methods).

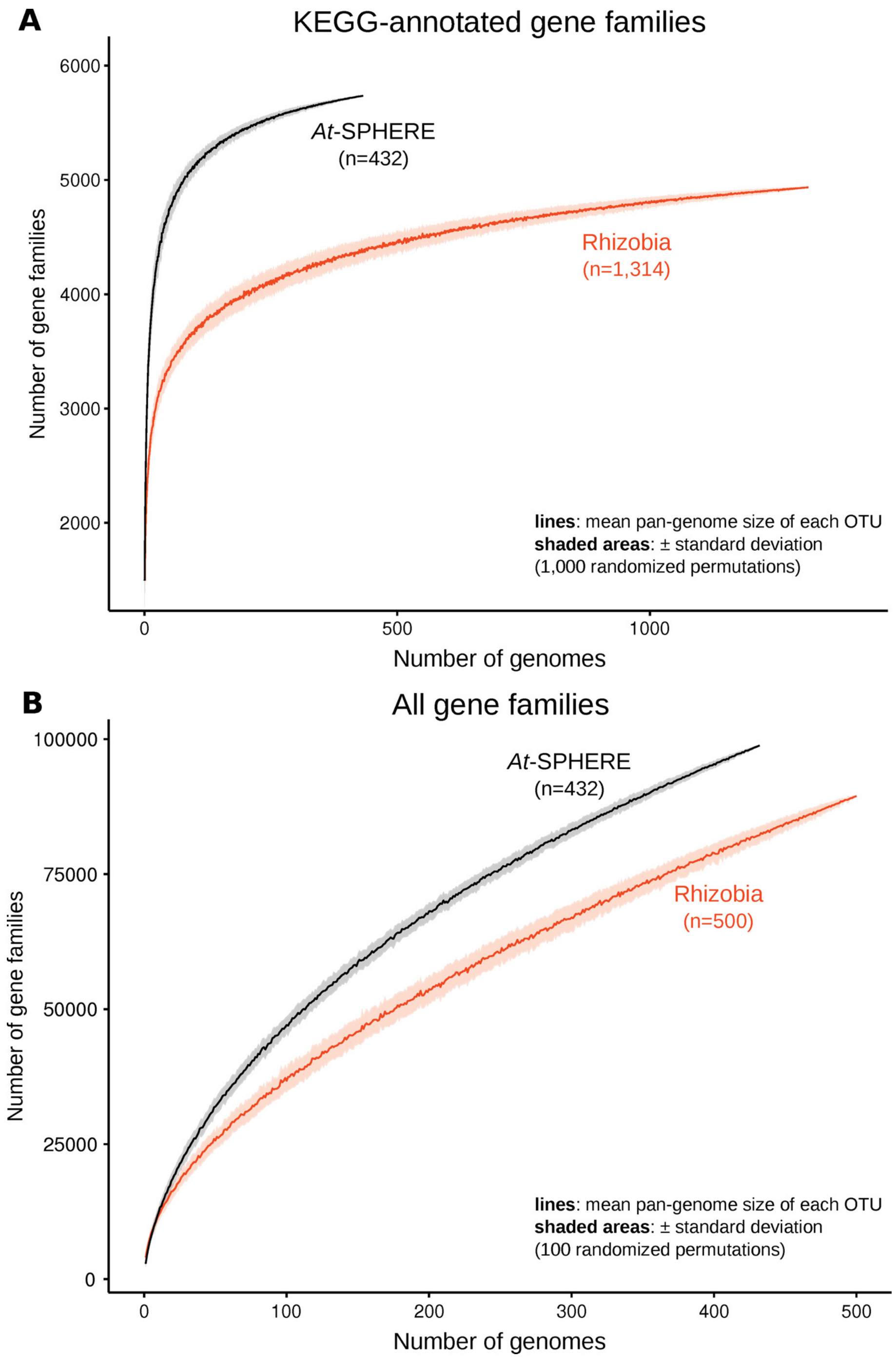


Figure 8.4: Accumulation curves displaying the estimated pan-genome size of the Rhizobia and *At*-SPERE collections. Permutation-based analysis of pan-genome size displaying accumulation curves of KEGG-annotated gene families (A) or *de novo* inferred orthologues (B) of the Rhizobia and *At*-SPHERE collection of sequenced isolates. Solid lines correpond to the mean pan-genome size across permutations (A, n=1,000; B, n=100) and shaded areas to ± 1 standard deviation.

This analysis revealed gene rarefaction curves close to the saturation regime (Figure 8.4A) and an approximate pan-genome size of ~5,000 COGs. However, the size of the database used for annotation imposes a limit to the number of gene families found in the dataset, and disregards all coding sequenced without an homologous annotated entry in the database (in average, only 54.33% of the genes found in a genome are classified into KO groups). We then extended the analysis of homology relationships between genes to sequences not represented in the database by performing *de novo* inference of orthologues. This operation is computationally demanding (its cost grows quadratically with the number of genomes), and its calculation was unfeasible for the entire dataset. Therefore, we limited the analysis to a tractable subset of 500 genomes covering representatives of all taxonomic groups and sublineages (Methods). This analysis clustered all protein sequences into 89,451 COGs or gene families (compared to the 98,813 COGs found in the *At*-SPHERE dataset using the same approach) and revealed a relatively small core-genome consisting of only 802 gene families (5.01% of the proteome, in average) and a very large, open pan-genome (88,649 COGs). This was further evidenced by a permutation-based analysis of gene content, which resulted in non-saturated rarefaction curves (Figure 8.4B) that indicate a substantial and yet uncharacterized functional and genomic diversity.

8.3.3 Convergent evolution of nitrogen-fixing symbiosis in Rhizobia

In order to explore the evolutionary history of the symbiosis in rhizobia, we performed a reconstruction of ancestral genotypes of all sequenced genomes using a phylogenomic approach. First, a rooted species tree was generated from a multiple sequence alignment of a set of conserved, single-copy and vertically inherited genes (Methods) using a Bayesian approach with a strict molecular clock and General Time Reversiveble model (GTR+G+I; Supporting Information). Next, we reconstructed the presence or absence of the nitrogenase-encoding *nif* and nodulation factor *nod* genes in each of the ancestral genomes along the branches of the species tree using a Maximum Likelihood (ML) approach (Methods). Due to the high correlation found between the gene families within this set of genes (Pearson correlation coefficient ≤ 0.94; Supporting Information), we decided to employ the conserved *nifH* marker gene as a proxy to indicate the presence of symbiosis genes in each ancestral genome. This analysis revealed that the most recent common ancestor to all rhizobia lacked the nitrogenase and nodulation factors (Figure 8.5; and that the capacity to form nodules and fix atmospheric nitrogen was acquired after the speciation events leading to the formation of the major rhizobial

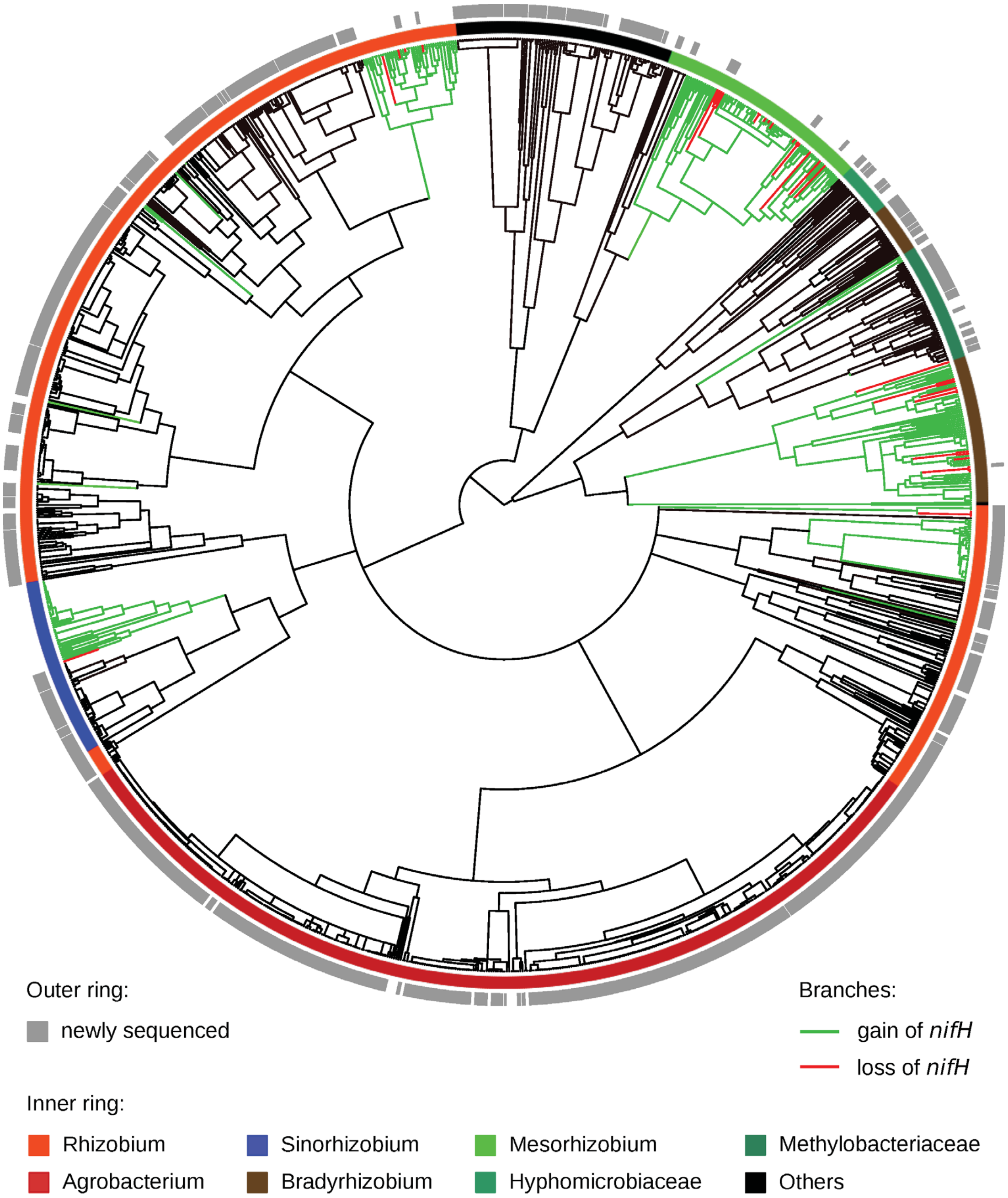


Figure 8.5: Whole-genome phylogeny of Rhizobia and Maximum Likelihood reconstruction of ancestral symbiotic genotypes. Phylogenetic tree of sequenced isolated inferred from aligned single-copy marker genes using a Bayesian approach and a strict molecular clock. Different taxonomic groups are shown in various colors in the first ring (n=1,313). The second ring depicts newly sequenced isolates in gray (n=943). Branches leading to clades colored in green correspond to likely gains of the symbiosis genes whereas those in red correspond to probable loss events.

taxonomic groups. We found strong evidence indicating multiple (≤ 16) instances of independent gains of the symbiosis genes along the branches of the species tree, all of them occurring in branches below the emergence of the sublineages corresponding to the extant families of rhizobia (Figure 8.5). The highly correlated pattern of gene gain and loss found for all considered symbiosis-relevant genes (see Chapter 9) indicates that the most likely origin of these sequences was a taxonomic group outside of the rhizobiales. However, joint phylogenetic analysis of *nifH* sequences found in the current dataset (n=296), together with orthologous sequences found in genomes outside of rhizobia (n=585) showed that the rhizobial genes constitute a separate clade Supporting Information), indicating that a first aquisition even from an outside donor was subsequently followed by HGT of the symbiosis genes within the Rhizobiales order.
In summary, analysis of reconstructed ancestral states of rhizobia indicates that the most recent ancestor of all rhizobia did not have the capacity to fix atmospheric nitrogen or induce nodule formation (Figure 8.5) but was able to successfully colonize the roots of a wide range of hosts, as indicated by the presence of exemplars from all major taxonomic groups in 16S amplicon surveys of plants grown in natural soils (Figure 8.1). These results, together with the finding that the vast majority of the exemplars isolated from roots of non-legumes were lacking the symbiosis genes (gray bars in Figure 8.5) indicates that, while the capability to colonize roots is shared among all major taxonomic groups of rhizobia, strains lacking the symbiosis genes are able to out-compete their nitrogen-fixing counterparts not only in soil, but also in the root and rhizosphere environments in the absence of nodulation. Furthermore, evidence of multiple independent instances of acquisition of the symbiosis genes indicates that development of the capacity to engage in symbiotic relationships with legumes is an example of convergent evolution in bacteria.

8.3.4 The majority of rhizobial species engage in root-growth promoting interactions with non-legume hosts

Based on the observation that 16S rRNA sequences from members of all major taxonomic groups were found in samples from plants grown in natural soils (Figure 8.1), we speculated that most rhizobia have developed a mechanism for interacting with their host that results in successful colonization of the root and rhizosphere environments. We decided to test this hypothesis by performing binary interaction experiments with rhizobia and germ-free plants grown under controlled laboratory conditions. First, we selected a subset of phylogenetically and functionally diverse strains ($n = 19$) belonging to all major rhizobial clades found in roots in the culture-independent datasets which include exemplars isolated from *Arabidopsis* roots as well as well characterized, nitrogen-fixing nodule symbionts (Supporting Information).

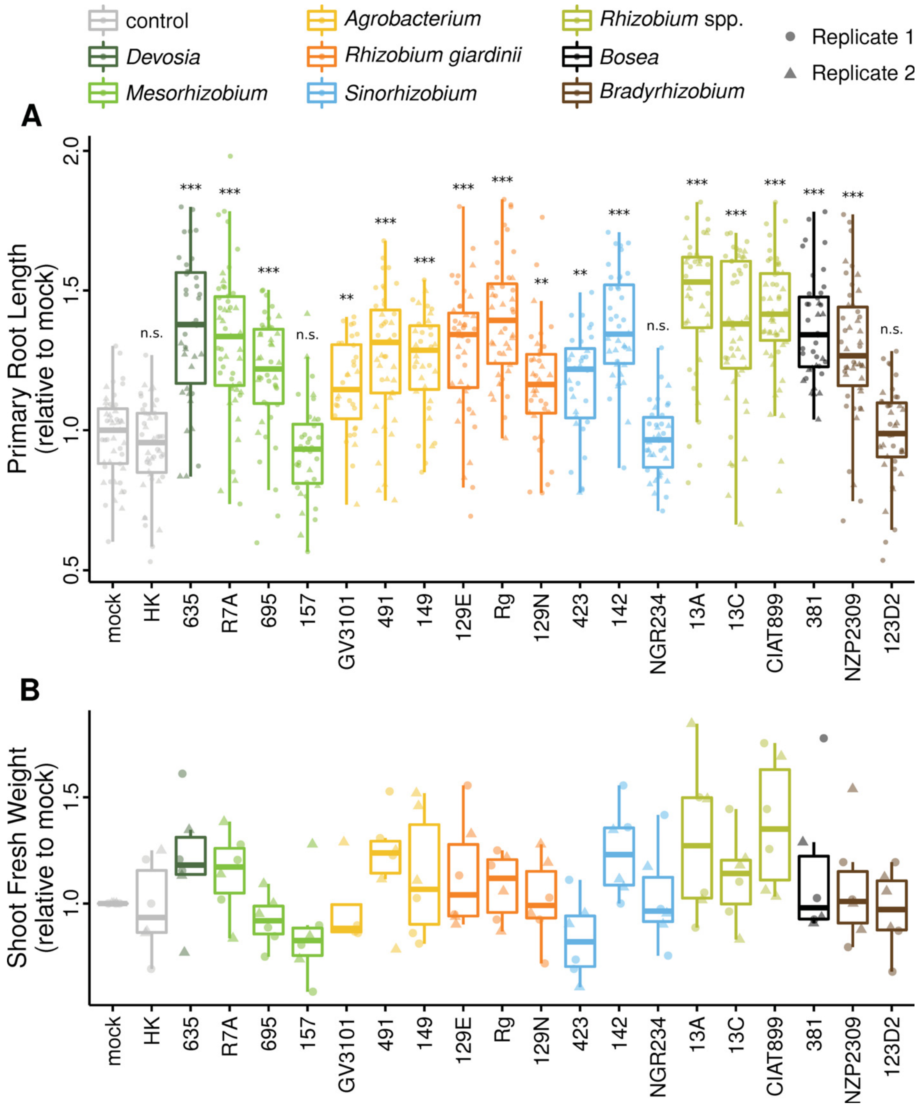


Figure 8.6: Binary interaction experiments between rhizobia and germ-free *Arabidopsis* plants grown on agarose media. Macroscopic plant phenotypes after inoculation of germ-free *Arabidopsis* wild-type plants grown on agarose media containing individual rhizobial isolates selected from all major taxonomic groups (different colors). (A) Primary root length relative to mock control for each treatment as well as an additional heat-killed (HK) bacteria control (n=874). (B) Shoot fresh weight relative to mock control for each treatment and HK bacteria control (n=126). Different shappes depict datapoints from two separate full-factorial replicates. Statistical significance corresponds to a Dunnett's test with FDR correcting ($\alpha = 0..5$).

We grew *Arabidopsis thaliana* wild type plants for three weeks on agarose media that contained individual rhizobial strains and measured their root length ($n = 874$) and shoot fresh weight ($n = 126$) relative to the germ-free controls (Methods). These experiments revealed that the majority of the tested exemplars displayed a strong root growth promotion phenotype, which was statistically significant (Dunnett's test with FDR correction; $\alpha = 0.05$; Methods) and conserved across multiple biological replicates (Figure 8.6A). Importantly, we observed that this phenotype was present in all major rhizobial taxonomic groups (different colors in Figure 8.6) and that both, *Arabidopsis* isolates as well as nitrogen-fixing legume symbionts had a comparable effect on the host. However, it remains unclear whether this root-growth promotion phenotype results in an increase in the fitness of the plant, as we observed a tendency to increased shoot biomass but no statistically significant difference in shoot fresh weight compared to the controls (Figure 8.6B). The presence of the same strain or very close relatives in the roots and rhizospheres of healthy plants collected from natural sites or grown in the greenhouse using agricultural soil suggest that the physiological functions of rhizobia in associations with flowering plants are not harmful to the host. Taken together, these data suggest the presence of an evolutionarily conserved mechanism that allows rhizobia to colonize the root environment and to interact with plant hosts from a broad taxonomic range that predates the acquisition of the ability to fix atmospheric nitrogen and induce nodule formation in rhizobia.

8.4 Discussion

Here, we introduce a large-scale isolation and sequencing effort resulting in high-quality whole-genome assemblies of 904 exemplars of rhizobia originating from a panel of taxonomically distant plant hosts, including non-legumes, as well associated environmental sources, such as soil, nematodes or insects (Supporting Information). These newly-sequenced isolates cover the majority of the known sublineages of rhizobia as well as potentially uncharacterized phylogenetic groups, providing novel insights into the taxonomic and genomic diversity of this ecologically relevant bacterial clade. These extensive collection of cultured isolates, together with their annotated genomes and a comprehensive and detailed metadata exploring their origin, biogeography and their ecological significance by cross-referencing with previous culture-independent community profiling studies using a taxonomically diverse panel of hosts (Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Bai *et al.*, 2015; Zgadzaj *et al.*, 2016) constitutes a valuable resource in for the study of symbiotic and commensal bacterial evolution by sequence analysis as well as by experimentation in the laboratory under controlled conditions. In this study we have also retrieved and jointly analyzed a set of high-quality genome assemblies of 370 strains of rhizobia mostly isolated from nodules of legumes belonging to all major taxonomic groups containing exemplars capable of developing symbiotic relationships with their plant host (Supporting Information). The resulting dataset of 1,313 genomes forms the basis of a large-scale comparative genomics analysis that sheds light in the evolutionary history of symbiosis and commensalism in rhizobia.

First, we assessed the ecological relevance of the various taxonomic groups present in the dataset by retrieving and jointly processing 5 previous 16S rRNA gene amplicon surveys of the microbial communities of plants grown in natural soils (Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Bai *et al.*, 2015; Zgadzaj *et al.*, 2016). We then used marker gene sequences extracted from the genome sequences to anchor all raw 16S reads and perform a high-resolution ($\geq$ 99% sequence identity) analysis of diversity. The results of this computational experiment showed that a taxonomically and functionally wide set of rhizobia are consistently found in the root and leaf microbial communities of a wide taxonomic range of plant hosts including the monocot barley and the dicots *Arabidopsis thaliana* and *Lotus japonicus*, the latter of which is able to form symbiotic interactions with compatible strains. (Figure 8.1). The broad taxonomic range of the OTUs found in significant levels in the root and rhizosphere samples, together with the conserved community composition observed in hosts that diverged more than 200 My ago, indicates that the ability to colonize plant roots is an evolutionarily conserved and ancestral feature of rhizobia which is not limited to nitrogen-fixing nodule symbionts. This hypothesis is further supported by the observation that the vast majority of the exemplars that were isolated from non-legume hosts and other environmental sources are lacking the genes necessary for nodule formation and nitrogen fixation, despite of containing very close relatives to known legume symbionts (Figure 8.5).

Analysis of functional diversity revealed a strong phylogenetic signal separating the genomes of the different taxonomic groups, without any clear clustering of isolates with respect to their origin. Furthermore, symbionts isolated from legume nodules showed no functional differentiation compared to non-nodulating rhizobia, implying that the acquisition of the symbiosis genes and subsequent process of nodule and host specialization was not accompanied by any significant changes at the whole genome level (Figure 8.3). *De novo* prediction of orthologous genes independent of a reference database for a taxonomically diverse subset of genomes ($n = 500$) revealed a very small core-genome consisting on an average of 27.30% of the proteome (802 gene families) and an extremely large pan-genome (88,649 COGs) (Figure 8.4). These results reveal a pre-

viously uncharacterized genomic heterogeneity which likely corresponds to adaptation to multiple soil and plant-associated habitats. Despite of the deep targeted sequencing effort conducted in this study, permutation analyses resulted in non-saturated accumulation curves that correspond to a yet to be explored functional diversity (Figure 8.4B). Ancestral state reconstruction of symbiosis genes (*nif* and *nod*) using a Maximum Likelihood approach indicates that the most recent common ancestor of all rhizobia was not capable to induce nodule formation and fix atmospheric nitrogen and reveals strong evidence of multiple independent gains of the symbiosis genes occurring after the speciation events leading to the formation of the clades that constitute the modern rhizobial families (Figure 8.5). These instances of independent acquisition of the symbiosis phenotype, likely via horizontal gene transfer from an out-group taxon, constitutes an example of convergent evolution in rhizobia.
Binary association experiments between a selected panel of rhizobial strains and germ-free *Arabidopsis thaliana* revealed a phylogenetically conserved form of interaction with the plant host consisting on a significant plant growth promotion phenotype (Figure 8.6). Interestingly, this phenotype was not limited to exemplars isolated from *Arabidopsis* roots but it was likewise present in plants inoculated with strains isolated from functional legume nodules and even with the originally pathogenic Agrobacterium tumefaciens (GV3101). The results from these experiments suggest the presence of a conserved form of interaction with the plant which is independent from the canonical symbiosis pathway and that might be a relevant feature for ecological fitness and survival in the plant root environment. Our data also supports the hypothesis that persistence of rhizobia in soils, including symbiotic strains, depends on their capability to interact with plant hosts and colonize the root and rhizosphere environments (Triplett *et al.*, 1993; Thies *et al.*, 1995; Zgadzaj *et al.*, 2016). Understanding the genetic basis and the evolutionary history of this capability to interact with a broad range of hosts might be a crucial component in improving the performance of highly compatible symbionts that can be employed as biofertilizers with increased persistence in nutrient-poor agricultural soils where the common practice of legume crop rotation is employed.

8.5 Materials and methods

8.5.1 Isolation and sequencing of strains

Surface sterilized seeds of Arabidopis thaliana were sown at a density of four plants per pot (7x7 cm) and stratified for three to four days. Plants were grown in the greenhouse for 6 weeks under short day conditi ons (8/16 hours day/night with a temperature of 22°C/18°C and a relative humidity of 70%). After harvest, roots were mechanically separated from the adhering soil particles and a defined root segment of 3cm starting 0.5cm distant from the hypocotyl was sampled. Soil particles still attached to the roots were removed by gentle tapping. Roots were collected in 15 ml falcons containing 5 ml phosphorus buffered saline (PBS)-S buffer (130 mM NaCl, 7 mM Na 2 HPO 4 , 3 mM NaH 2 PO 4 , pH 7.0, 0.02% Silwet L-77) and washed for 15 minutes at 180 rpm on a shaking platform. Roots were transferred to a new falcon, the remaining soil particles were centrifuged for 20 minutes at 15,000 x g. After washing for a second time, roots were transferred to a new falcon tube and sonicated 10 times at 160 W with 30 second brakes (Bioruptor Next Gen UCD-300, Diagenode, LiÃ¨ge, Belgium). Roots were transferred to a 1.5 ml tube containing 10 mM MgCl2 and mechanically disrupted using the Precellys24 tissue lyzer at 5,000 rpm for 3x at 30 seconds. Afterwards, the root tissue was centrifuged for 5 minutes at 1,000 x g and the supernatant transferred to a new tube. Serial dilutions were prepared from the supernatant and plated onto flour and TWYE media (flour: 6 g/L flour, 0.3 g/L yeast extract, 0.3 g/L sucrose, 0.3 g/L CaCO3 , 1.8%. TWYE: 0.25 g/L yeast extract, 0.5 g/L K2HPO4 , 1.8% agar, respectively), including 50 μg/ml Benzimidazole to inhibit fungal growth. Plates were incubated for three to four days at 28°C. Afterwards, single colonies were transferred with a sterile pipette tip to 400 μl liquid media in a 96-well format and incubated for up to 7 days at 28°C at 200 rpm. Then, 100 μl of the culture was transferred to a 96-well PCR plate and bacterial cells disrupted for 10 minutes at 100Â°C. Cell fragments were centrifuged for 10 minutes at 3,000 rpm and the supernatant transferred into a new plate. In a PCR reaction, used to obtain sequence information for taxonomic assignment, the primers 799F-noadaptor (5'-AACMGGATTAGATACCCKG-3') and 1392R (5'-ACGGGCGGTGTGTRC-3') were used. PCRs were performed using 3 μl of isolated bacterial DNA in a total volume of 25 μl. PCR components include 1.25 U DFS-Taq DNA Polymerase (Bioron, Ludwigshafen, Germany), 1x complete reaction buffer, 0.3% BSA, 200 μM of dNTPs and 400 nM of each primer. The PCR reaction was pipetted in a laminar flow and amplified (94°C/ 5 minutes, 94°C/30 seconds, 50°C/ 30 seconds, 72°C/30 seconds, 72°C/5 minutes for 35 PCR cycles). PCR quality was assessed by loading 5 μl of the amplicon on a 1% agarose gel. DNA concentration was measured using a Nanodrop photometer (PeqLab, Erlangen, Germany), adjusted to 20 ng/μl and subjected to Sanger sequencing (performed by the Max Planck Genome Center, Cologne, Germany). Obtained sequences were analyzed with BioEdit and blasted against an internal 454 se quence database (as described in Bulgarelli et al., 2012) to correlate the isolated bacteria to OTUs identified with in previous culture-independent studies.

8.5.2 Genome assembly

Genomes sequenced using the Illumina platform were assembled by first passing their short reads through a quality and length trimming filtering with Trimmomatic (Bolger *et al.*, 2014) and subsequently assembled using an ensemble of two approaches: A5 (Tritt *et al.*, 2012) and SOAPdenovo (Li, 2010). In each case, the assembly with the smaller number of scaffolds was selected. Subsequently, contigs smaller than 1,000 bases were removed from the assemblies. Genomes sequenced using the Pacific Biosciences platform and their long reads were assembled into complete genomes using the HGAP assembler (Chin *et al.*, 2013). Detailed statistics concerning the assemblies of all genomes analyzed in this study can be found in the Supporting Material. Accession numbers corresponding to the raw reads of all genomes, including also those of previously published organisms can likewise be found in the Supporting Material.

8.5.3 Annotation and orthology inference

To determine homology relationships between gene sequences we first predicted putative protein-coding sequences using Prodigal (Hyatt *et al.*, 2010). Annotation of candidate Open Reading Frames (ORFs) was then conduced using the KEGG Orthologue (KO) database (Kanehisa *et al.*, 2016) as previously described (Bai *et al.*, 2015). Briefly: sequences from each KO were aligned using Clustal Omega (Sievers, 2011) and Hidden Markov Models (HMMs) generated based on each multiple sequence alignment using the HMMER suite (Eddy, 2011). Subsequently, the HMMs were employed to search all putative protein-coding sequences. Sequences matching a KO with an E value lower than 10E10-5 with a coverage of at least 70% of the total length were assigned to said KO. All sequences assigned to the same KO were considered homologous and classified as belonging to the same Cluster of Orthologous Genes (COG). For the pan-genome analysis and permutation-based analysis of saturation curves we also conducted a more comprehensive prediction of COGs using OrthoFinder (Emms and Kelly, 2015), a sequence-based *de novo* method for orthogroup inference. Subsequently, for each Cluster of Orthologous Genes (COG) we generated a phyletic pattern consisting on a binary vector of presence / absence of each KO group or COG in each genome of the dataset. The total of all phyletic patterns, or phylettic matrices, were used as input for all downstream analyses (Supporting Material).

8.5.4 Natural community amplicon sequencing meta-analysis

First, we retrieved from the public databases all raw, unprocessed sequences corresponding to the analyzed 16S rRNA gene surveys (Schlaeppi *et al.*, 2014; Bulgarelli *et al.*, 2015; Bai *et al.*, 2015; Zgadzaj *et al.*, 2016) and compiled a master meta-data table containing information of all samples across studies (Supporting Material). Next, a database of high-quality reference sequences was built by extracting 16S gene sequences from the rhizobial genome assemblies using RNAmmer (Lagesen *et al.*, 2007). Identical reference sequences were prior downstream analyses dereplicated (Supporting Material). Subsequently, 16S rRNA gene sequences were processed using a combination of custom scripts as well as tools from the QIIME (Caporaso *et al.*, 2010) and USEARCH (Edgar, 2010) pipelines. First, reads were truncated to an even length (290 bp) using the *truncate_fasta_qual_files.py* QIIME script. Libraries were demultiplexed (*split_libraries.py*) and only reads with a quality score $Q > 25$ were retained for subsequent analysis. After dereplication and removal of singletons we conducted a reference-based clustering of sequences into OTUs using the UPARSE algorithm (Edgar, 2013) at 99% identity using high-quality 16S dereplicated sequences extracted from the whole-genome assemblies. In parallel, we conducted *de novo* clustering of all sequences into OTUs using the standard UPARSE algorithm (Edgar, 2013) at 97% identity. After chimeric sequences were filtered using UCHIME (Edgar *et al.*, 2011) and the 'gold' database (`http://drive5.com/uchime/gold.fa`), we used the total number of non-artefactual OTUs per sample to normalize the rhizobial abundances as a percentage of the total sample size. The resulting OTU table was used in all subsequent statistical analyses of differentially abundant taxa as well analyses of alpha- and beta-diversity (Supporting Material). Prevalence of symbiosis genes in each reference OTU was calculated as the percentage of genomes within each OTU (genomes with identical 16S V5-V7 sequences) containing each marker gene with respect to the OTU size.

8.5.5 Comparative genomics and ancestral character reconstruction

Inference of an accurate species tree was conducted as follows: first, we extracted from each proteome a set of single-copy phylogenetic markers that are present in the majority of sequenced bacterial genomes called AMPHORA genes (Wu and Eisen, 2008) using a set of HMMs and the hmmsearch tool (Eddy, 2011). The resulting sequences were concatenated independently for each genome and then aligned using Clustal Omega (Sievers, 2011). For each dataset, rooted species trees were generated using the AMPHORA multiple sequence alignments using MrBayes (Huelsenbeck and

Ronquist, 2001; Ronquist and Huelsenbeck, 2003) with a strict molecular clock and a General Time Reversible model (GTR+G+I; Supporting Material). Similar results as reported above were obtained by reconstructing the species tree using an alternative Maximum Likelihood approach implemented in FastTree (Price *et al.*, 2010) with a GTR model of DNA evolution and without a molecular clock constrain (data not shown). Polytomies in the trees were resolved by inserting branches of zero length to avoid conflicts in the inference of ancestral characters.

Given the matrix of phyletic patterns and the species tree we used a Maximum Likelihood (ML) approach (Pagel, 1994) for the estimation of ancestral states using the implementation provided with the R package ape (Paradis *et al.*, 2004), which employs a 'two-pass' algorithm for the joint estimation of the likelihood of the ancestral states (Felsenstein and Felenstein, 2004; Yang, 2006). This approach gives similar results as stochastic mapping (Pupko *et al.*, 2000; Paradis *et al.*, 2004) while being much faster, thus enabling the estimation of ancestral characters for very large datasets (several thousands of genomes or more).

Analyses of functional diversity between sequenced isolates were as in (Bai *et al.*, 2015) First, we generated for each genome in the data set, a profile of presence/absence of each KO group (or phyletic pattern). Subsequently, a distance measure based on the Pearson correlation of each pair of phyletic patterns was calculated, which allowed us to embed each genome as a data point in a metric space.

8.5.6 Binary association experiments with germ-free *Arabidopsis plants*

The rhizobial strains were grown in liquid TY media (5 g/L tryptone, 3 g/L yeast extract, and 10 mM CaCl2) for 2-3 days and precultured for additional 2-3 hours following 5-fold dilution with new TY media. Bacterial cells were collected and resuspended in 10 mM MgCl2, followed by OD600 adjustment to 0.5, and mixed into 10,000 volume of half-strength MS media (750 μM MgSO4, 625 μ KH2P04, 10.3 mM NH4NO3, 9.4 mM KNO3, 1.5 mM CaCl2, 0.05 μM CoCl2, 005 μM CuCl2, 50 μM H3BO3, 2.5 μM KI, 50 μM MnCl2, 0.5 μM Na2MoO4, 15 μM ZnCl2, 75 μM Fe-EDTA, 0.5 mM MES-KHO pH 5.5) supplemented with 1% granulated agar (Difco). Aliquots of rhizobia strains after OD600 adjustment was mixed with equal volume ratio and incubated on 99 degrees blocks for 10 minutes. Heat treated rhizobia strains mixture was added to 526 volume of half-strength MS media and used as a heat-killed (HK) control. As a mock control, 10 mM MgCl2 was added to 10,000 volume of half-strength MS media. Surface sterilized seeds of Col-0 wild type (N60000) were sown on to the media and cultured for three weeks under short-day conditions (10 hours light under 21 degrees and 14 hours dark under 19 degrees) in a randomized design. To avoid heterogeneous light intensity, the location of plates in the chamber was mixed every week. Primary root length was measured by ImageJ after scanning the plates at 360 dpi resolution. Primary root length and shoot fresh weight were normalized to the median of respective mock control within each technical replicates.

8.6 Author contributions

R.G.-O., R.T.N., N.D., P.S.-L designed research. N.D. isolated rhizobial strains from *Arabidopsis* and prepared DNA for whole-genome sequencing. R.T.N. and N.D. conducted and performed statistical analysis of phenotypic data. R.G.-O. performed genome assembly and annotation, amplicon data meta-analysis, comparative genomic and statistical tests and interpreted the data. R.G.-O., N.D. and R.T.N. wrote the manuscript.

8.7 Acknowledgements

We thank tue Nottingham *Arabidopsis* Stock Centre (NASC) for providing Col-0 wild type seeds, Dr. Simona Radutoiu of Aarhus University for providing *Mesorhizobium loti* R7A, *Sinorhizobium* NGR234 and *Bradyrhizobium* NZP2309 strains, Dr. Rene Geurts of Wageningen University for providing *Rhizobium tropicii* CIAT899 strain. We would also like to thank Dr. Rafal Zgadzaj for his comments on this manuscript.

8.8 Supporting material

The supporting material corresponding to this section, including all supplementary tables, databases and figures can be accessed during revision at the following address http://www.at-sphere.com/downloads/rhizobia.tar.gz and have not been included in this thesis due to space limitations.

Part IV

Evolutionary profiles

CHAPTER 9

Seventh publication — Clustering of functionally related genes reveals novel symbiosis-relevant genes in Rhizobia

Status	**In preparation**
Citation	Garrido-Oter, R. and McHardy, A.C. (2016). Clustering of functionally related genes reveals novel symbiosis-relevant genes in Rhizobia
Own contribution	Conceived research (with co-authors) Implemented the method and performed the experiments Interpreted the data (with co-authors) Wrote the manuscript

9.1 Abstract

Determining functional and co-evolutionary relationships between genes based on sequence alone is a crucial step in integrating large genomic datasets and understanding how bacterial populations and communities adapt to diverse environments. Here, we propose a phylogenetic approach for determining clusters of co-evolving genes and their network organization by modeling gene gain and loss as a continuous process along the branches of the species tree. Our method accounts for uncertainty in the reconstruction of the ancestral states as well as in the inference of the species tree and robustly identifies clusters of co-evolving genes that significantly enrich for functional categories and pathways and which are relevant for adaptation to diverse environments. We demonstrate its ability to detect biologically meaningful gene family interactions by analyzing a total of 2,737 bacterial genomes, including diverse populations of multiple plant-associated and symbiotic Rhizobia, a phylogenetically wide collection of bacterial genomes representative of the *Arabidopsis thaliana* leaf and root microbiota as well as a population of clinically relevant antibiotic-resistant *Staphylococcus aureus* strains. We generate co-evolutionary networks that show a hierarchical and highly clustered structure and find that proximity in the network clearly correlates with functional relatedness in a curated database. Detailed analysis of a large and compact cluster containing gene families known to be crucial for the interaction between Rhizobia and their plant hosts, revealed potentially novel genes relevant for nodule nitrogen-fixing symbiosis. In addition, by analyzing a large set of *Staphylococcus aureus* genomes we were able to accurately predict the genetic basis of resistance to several antibiotics, further validating our approach to link genotypes and phenotypes. Finally, based on these results, we conclude that the predominant mechanisms driving bacterial genome evolution, based on gene gain and loss (via Horizontal Gene Transfer and genome reduction), act in a concerted manner at the level of gene modules rather than at the level of individual genes.

9.2 Introduction

Prokaryotic genome evolution is dominated by two main processes: gene gain (mostly *via* Horizontal Gene Transfer, or HGT) and gene loss, for instance when lifestyle changes in symbiotic or pathogenic bacteria derivate in reductive genome evolution (Soucy *et al.*, 2015; Kuo and Ochman, 2009). However, selective pressure acting on a given gene might change as a result of gains or losses of other genes relevant for the same or related processes, causing sets of genes to co-evolve. Furthermore, genes whose products form part of the same complex tend to be located in the same regions of the genome, causing them to be jointly lost or transferred (Ettema *et al.*, 2001). Studying these inter-relationships between individual gene evolutionary histories is a crucial step in understanding the process of prokaryotic genome evolution.

Previous attempts to measure signatures of co-evolution between pairs of genes can be devided broadly into two categories: methods that detect co-evolving sites or residues in gene familiy sequence alignments, and which are mostly applied for predicting protein-protein interactions (de Juan *et al.*, 2013), and phylogenetic methods that use patterns of gene presence or absence (phyletic patterns) and that take into account the species tree to distinguish joint adaptive evolution from the effects of population structure (Campillos *et al.*, 2006; Cordero *et al.*, 2008; Cohen *et al.*, 2012). These previous approaches have shown that correlating discrete events of gene gain and loss along branches of the species tree allows identification of co-evolving and functionally related pairs of genes (Cordero *et al.*, 2008). However, modeling gene evolutionary histories as a set of discrete events does not adequately account for uncertainty in the inference of these events and treats gains and losses as separate, independent features (Cohen *et al.*, 2012).

Here we introduce a novel phylogenetic approach that uses Maximum Likelihood (ML) reconstruction of ancestral states to model gene evolutionary histories as a continuous spectrum of gain / loss probabilities which we term evolutionary profiles. We employ this approach to cluster gene families into functionally relevant modules by comparing their evolutionary profiles while accounting for uncertainties in the inference of the ancestral states and the species phylogeny. The underlying hypothesis underpinning this approach is that gene families that have been jointly acquired and lost more often than expected by chance are likely to be involved in the same biological process. Comparing the evolutionary profiles between all pairs of genes allowed us to generate networks of co-evolving proteins for several large datasets of bacterial genomes. We were able to validate the biological significance of the inferred networks and clustering solutions using a curated annotation database as well identified a number of potentially novel gene families involved in the ecologically and economically relevant symbiotic interaction between nitrogen-fixing Rhizobia and legume plants. Additionally, we applied our approach to a large dataset of antibiotic-resistant *Staphylococcus aureus* genomes and were able to robustly and accurately determine the genetic basis of the resistance to the tested antibiotics, despite of the presence of a strong population structure.

9.3 Results

9.3.1 Network of co-evolving genes and functional modules in Rhizobia

We sought to determine modules or clusters of co-evolving genes in a dataset of 1,314 genomes of plant-associated bacteria belonging to symbiotic and commensal species within the order *Rhizobiales*, collectively known as Rhizobia. Members of these species have been profusely studied in the past for their capability to engage in various forms of symbiotic interactions with plants. The most studied form of symbiosis requires the formation of specialized structures in the roots of legumes, called nodules, which serve as a niche for nitrogen-fixing bacteria. Recently, we have shown the majority of rhizobial isolated strains are commensals found in great abundances in the roots and leaves of both, legume and non-legume plants, and that the specialized, nitrogen-fixing form of symbiosis, is one possible of several adaptive strategies for these bacterial species (Chapter 7). Our goal was to identify *in silico* modules of co-evolving genes relevant for these strategies of adaptation to the plant root and leaf environments.

First, we generated a matrix of annotated phyletic patterns for each of the 4,936 Clusters of Orthologous Genes (thereafter referred to as COGs or, interchangeably, 'gene families' or simply 'genes') found in the dataset. Each row in the matrix corresponds to the phyletic profile of a gene family, consisting in a binary vector containing information of presence / absence of the gene in every genome of the dataset. Taking this matrix as input of our method (see Methods for a detailed description), we were able to generate a network of genes whose evolutionary histories showed significant evidence of co-evolution (Figure 9.1). In total, we determined a network with 4,125 edges denoting high functional relatedness between COGs (Methods). In accordance with previous research (Cordero *et al.*, 2008; Cohen *et al.*, 2012), we found the resulting network to be scale free (Figure 9.1). A deeper analysis of the network structure revealed a high level of granularity (Watts-Strogatz clustering coefficient 0.47; Watts and Strogatz (1998)) and a large number of genes that showed evidence of co-evolution with at least another gene (1,805 COGs or 36.57% of all COGs).

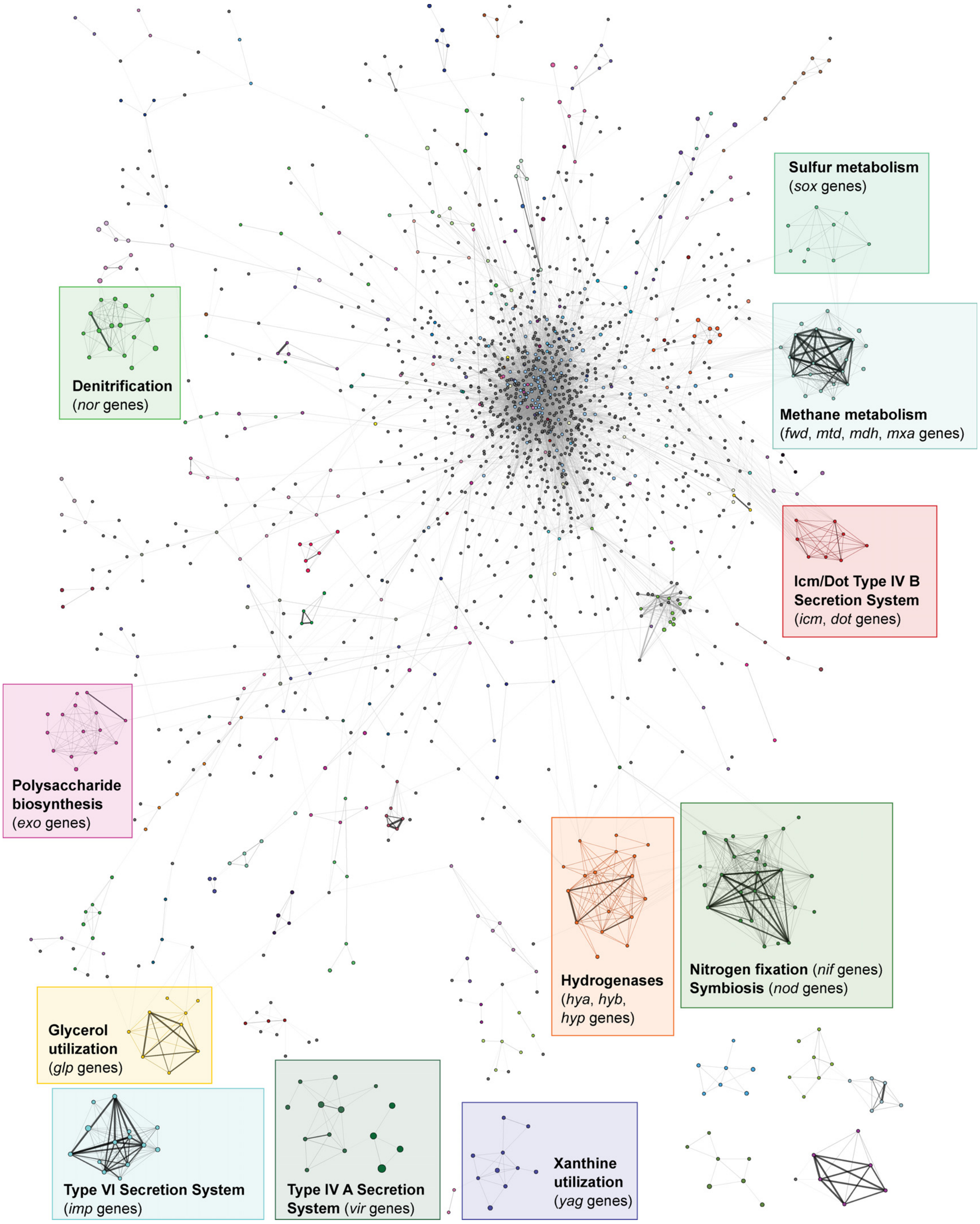


Figure 9.1: Network of co-evolving genes in Rhizobia. Graphical view of the co-evolution network of 1,314 genomes of Rhizobia, where each node corresponds to a gene family and edges correpond to similarity between the evolutionary profiles of each pair of genes. The network was produced after a filtering step and retaining edges with a weighted correlation coefficient above a threshold ($c_{\min} \geq 0.30$) and considering only informative nodes ($I_{\min} > 10$). Clusters with the highest connectivity are depicted in different colors and highlighted in boxes. Thickness and transparency of the connecting lines correspond weighted correlation coefficient between evolutionary profiles.

Next, we tried to determine the biological relevance of the evolutionary network by comparing the adjacency of pairs of gene families with their functional annotation in the Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway database (Kanehisa *et al.*, 2016), a well curated resource that groups orthologous sets of genes into pathways and modules representing their interactions in diverse biological processes. We found that pairs of genes which are highly adjacent in the co-evolutionary network show a large overlap in their annotated pathways (Figure 9.2) and observed a clear correlation between their functional relatedness (Jaccard similarity index) and their proximity in the network (Figure 9.2; Methods). These results indicate that the infered links that form the basis of the co-evolution network have a biological relevance and recapitulate experimentally validated functional interactions.

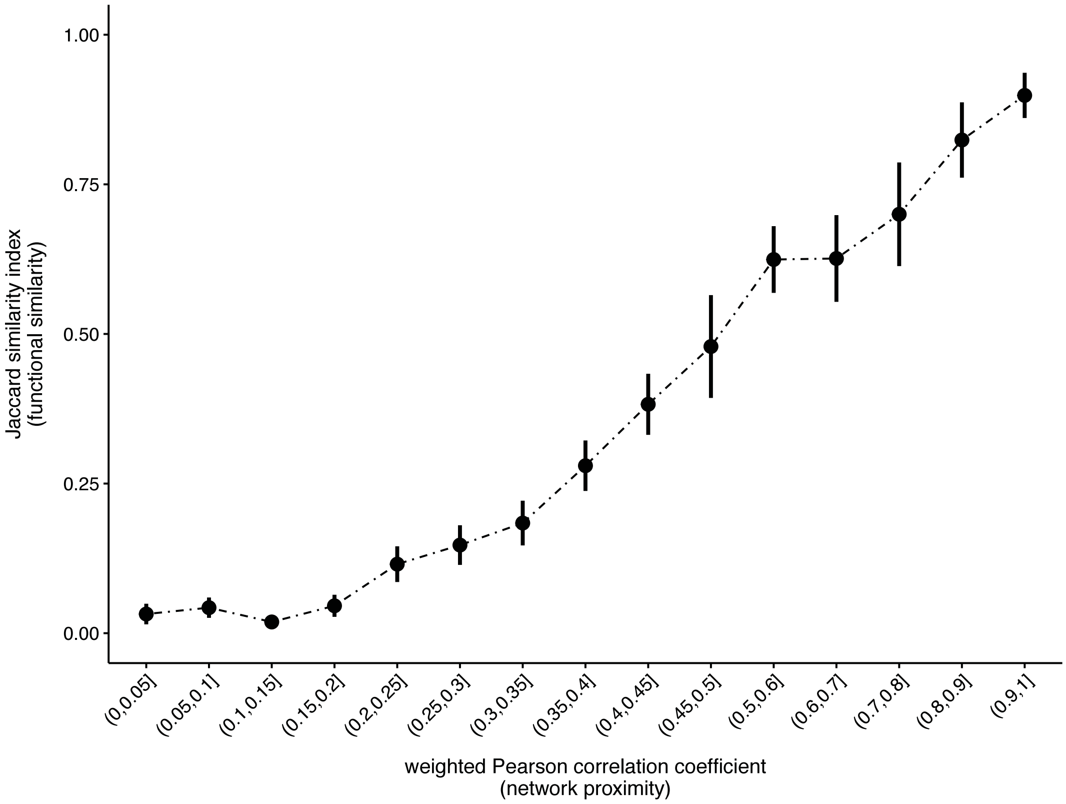


Figure 9.2: Relationship between network proximity and functional similarity. Plot showing the correlation between functional similarity (Jaccard similarity index) between the predicted KEGG pathway of each pair of gene families and their network proximity, determined by the weighted Pearson correlation between their evolutionary profiles. Posterior probabilities from the inferred species tree were used as weights. (Methods) Dots and bars correspond to the mean and standard deviation of each bin.

Based on the highly granular structure of the network, we decided to apply an unsupervised clustering method to robustly determine well-connected components or clusters potentially involved in the same biological processes. We applied the Markov Cluster algorithm (MCL) (Van Dongen, 2001) to our co-evolution network (Methods) and obtained a set of 233 clusters (Supporting Information). Among these, 44 clusters containing at least 5 genes showed a high degree of connectivity and the strongest signatures of co-evolution, clearly indicating the presence of biologically relevant modules of functionally related gene families.

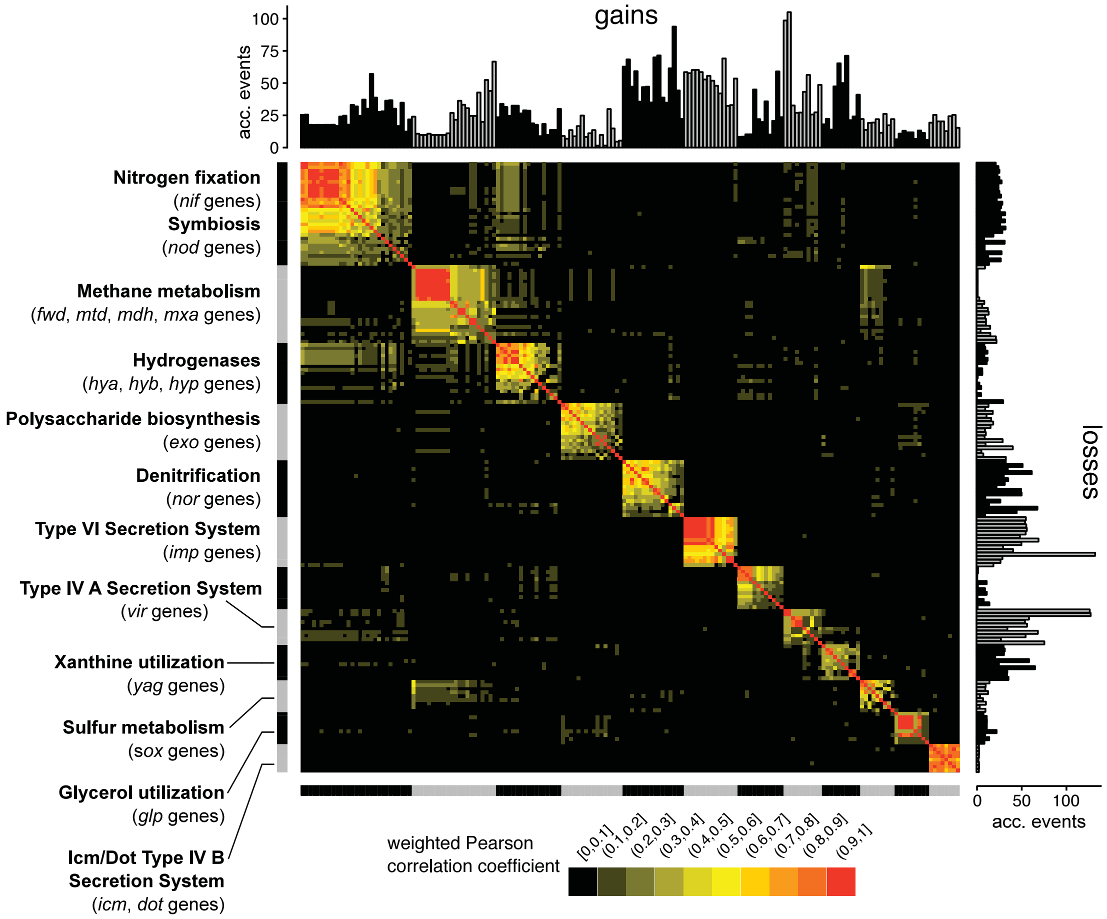
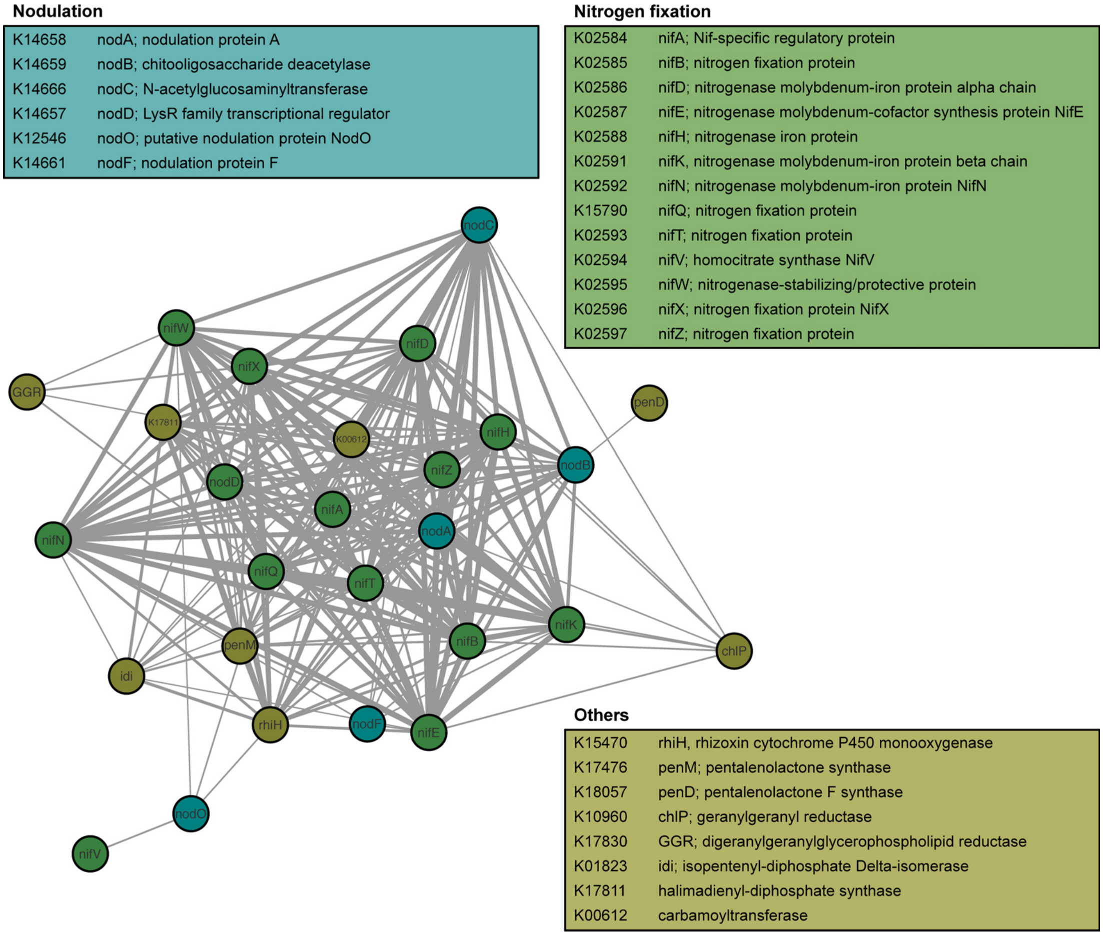


Figure 9.3: Adjacency matrix showing the highly clustered structure of the co-evolution network of Rhizobia. Heatmap depicting weigthed correlation coefficients between each pair of genes in the Rhizobia co-evolution network. Only the largest connected clusters were included. Genes were ordered according to their cluster membership and clusters where arranged by size. The annotation on the left-hand side is attributed by majority vote of all the cluster members.

Combining the clustering results with the underlying adjacency matrix provided a visual representation of the structure of these modules or components of co-evolving genes (Figure 9.3). This analysis showed that gene families belonging to a module are well connected to other members of the same cluster but seldom to other genes in the network, indicating that the driving forces underlying gene acquisition and loss act at the level of functional modules rather than at the level of individual genes.



Nodulation

K14658	nodA; nodulation protein A
K14659	nodB; chitooligosaccharide deacetylase
K14666	nodC; N-acetylglucosaminyltransferase
K14657	nodD; LysR family transcriptional regulator
K12546	nodO; putative nodulation protein NodO
K14661	nodF; nodulation protein F

Nitrogen fixation

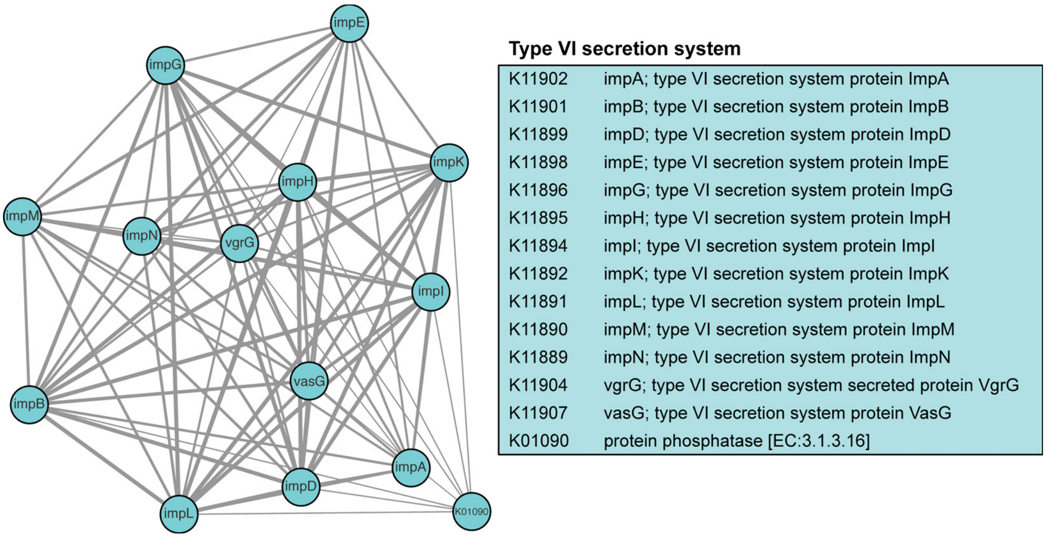
K02584	nifA; Nif-specific regulatory protein
K02585	nifB; nitrogen fixation protein
K02586	nifD; nitrogenase molybdenum-iron protein alpha chain
K02587	nifE; nitrogenase molybdenum-cofactor synthesis protein NifE
K02588	nifH; nitrogenase iron protein
K02591	nifK, nitrogenase molybdenum-iron protein beta chain
K02592	nifN; nitrogenase molybdenum-iron protein NifN
K15790	nifQ; nitrogen fixation protein
K02593	nifT; nitrogen fixation protein
K02594	nifV; homocitrate synthase NifV
K02595	nifW; nitrogenase-stabilizing/protective protein
K02596	nifX; nitrogen fixation protein NifX
K02597	nifZ; nitrogen fixation protein

Others

K15470	rhiH, rhizoxin cytochrome P450 monooxygenase
K17476	penM; pentalenolactone synthase
K18057	penD; pentalenolactone F synthase
K10960	chlP; geranylgeranyl reductase
K17830	GGR; digeranylgeranylglycerophospholipid reductase
K01823	idi; isopentenyl-diphosphate Delta-isomerase
K17811	halimadienyl-diphosphate synthase
K00612	carbamoyltransferase

Figure 9.4: Co-evolution of genes in the symbiosis module in Rhizobia. Graphical representation of the subnetwork containing gene families in the symbiosis cluster. Each node corresponds to a gene family, colored by attributed function. The thickness and transparency of the connecting edges corresponds to the weighted Pearson correlation coefficient between the evolutionary profiles of each pair of genes. Only genes with a correlation above a minimum threshold ($c_{\text{min}} \geq 0.30$) are depicted. Boxes contain the KEGG Orthologue (KO) identifier and description of each gene.

Surprisingly, all of the primary functions associated to the largest and better connected clusters belong to categories of potential ecological relevance for Rhizobia (Figure 9.1 and Figure 9.3) For example, one of the top modules is highly enriched with gene families which are crucial for the establishment of successful nitrogen-fixing symbiosis with legume plants (Figure 9.3 and Figure 9.4) and contains the genes encoding for the nitrogenase (*nif* genes) and their regulatory components (Dixon and Kahn, 2004) as well as the nodulation factors (*nod* genes), which encode for secreted proteins that act as molecular messengers between symbiotic Rhizobia and their host during the process of colonization and nodule formation (Lerouge *et al.*, 1990; Oldroyd, 2013). These two sets of genes are tightly clustered together and highly connected by edges in the network (Figure 9.1 and Figure 9.3) denoting strong signatures of co-evolution. This result is in accordance with the current underestanding of the genetic basis of the symbiotic process. However, despite of the clear involvement of these two groups of genes in the same biological process, the *nif* and *nod* genes are annotated in the database as belonging to different pathways ('Nitrogen fixation' and 'Symbiosis'). This implies that our positive assessment of the biological relevance of the co-evolutionary network (Figure 9.2) likely understimates the number of meaningful links between gene families which are nonetheless annotated as belonging to distinct funtional categories, thus validating our approach to uncover non-trivial relationships between co-evolving subsets of genes of diverse functions.



Type VI secretion system

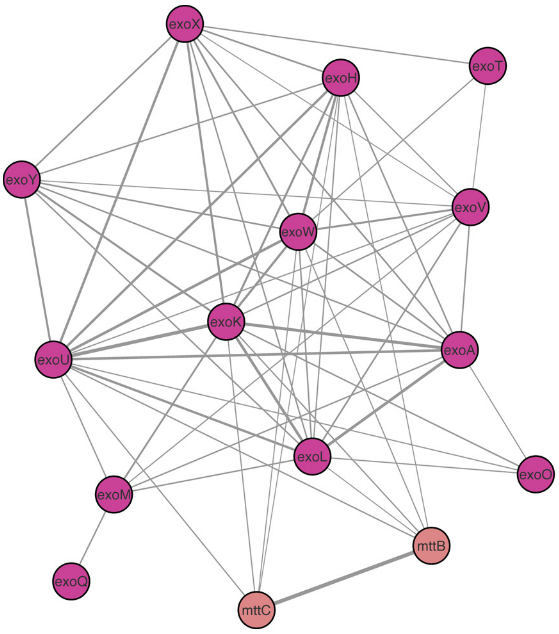
K11902	impA; type VI secretion system protein ImpA
K11901	impB; type VI secretion system protein ImpB
K11899	impD; type VI secretion system protein ImpD
K11898	impE; type VI secretion system protein ImpE
K11896	impG; type VI secretion system protein ImpG
K11895	impH; type VI secretion system protein ImpH
K11894	impI; type VI secretion system protein ImpI
K11892	impK; type VI secretion system protein ImpK
K11891	impL; type VI secretion system protein ImpL
K11890	impM; type VI secretion system protein ImpM
K11889	impN; type VI secretion system protein ImpN
K11904	vgrG; type VI secretion system secreted protein VgrG
K11907	vasG; type VI secretion system protein VasG
K01090	protein phosphatase [EC:3.1.3.16]

Figure 9.5: Co-evolution of genes in the T6SS module in Rhizobia. (Caption in following page).

Figure 9.5: Co-evolution of genes in the T6SS module in Rhizobia. (Figure on previous page). Graphical representation of the subnetwork containing gene families in the symbiosis cluster. Each node corresponds to a gene family, colored by attributed function. The thickness and transparency of the connecting edges corresponds to the weighted Pearson correlation coefficient between the evolutionary profiles of each pair of genes. Only genes with a correlation above a minimum threshold ($c_{\min} \geq 0.30$) are depicted. Boxes contain the KEGG Orthologue (KO) identifier and description of each gene.

Another example of strong signatures of co-evolution is found between genes encoding for molecular componets of different bacterial secretion systems (Figure 9.3), such as the canonical Type IV Secretion System (T4SS), which is known to be present in members of the Rhizobiales order, most notably in the *Agrobacterium* genus, and relevant for interactions between rhizobacteria and their plant host (Cascales and Christie, 2003; Lessl and Lanka, 1994). Surprisingly, in addition to the canonical T4SS encoded by the *vir* genes, we found a strongly co-evolving cluster formed by elements of an alternative T4SS system, also known as Icm/Dot complex or T4SS B, encoded by the *icm* and *dot* genes (Figure 9.3). This system is known to be present in pathogenic bacteria such as *Legionella pneumophila* and found to be crucial for virulence by means of effector protein delivery into the eukaryotic host cell (Vogel *et al.*, 1998; Zusman *et al.*, 2004; Christie *et al.*, 2014). However, among the three detected clusters containing genes related to secretion, the strongest signs of co-evolving gene families were found in the module corresponding to the Type VI Secretion System (T6SS), consisting mostly of the *imp* genes (Figure 9.5), which has been shown to be relevant in microbe-microbe interactions, e.g. by providing defense against single-celled eukaryotic predators such as Protists or by playing a role in competition against other bacteria. (Toft and Andersson, 2010; Russell *et al.*, 2014). Together, these results provide further evidence of the importance of bacterial secretion in the adaptation to the root and rhizosphere environment by plant-associated taxa.

An additional cluster found among the better connected components of the network is formed by the *exo* genes (Figure 9.6), which are necessary for the biosynthesis of a large array of exopolysaccharides known to play a crucial role in the invasion of the root interior (endophytic compartment) that leads to succesful formation of indeterminate nodules such as those found in *Medicago truncatula* (Jones *et al.*, 2009). Thus, together with the nitrogenase and *Nod* factor cluster (Figure 9.3 and Figure 9.4) our method successfuly identified, among the highest scoring clusters, several modules containing a large portion of the repertoire of genes necessary for nodule symbiosis formation.



Bacterial polysaccharide synthesis

K16557	exoA; succinoglycan biosynthesis protein ExoA
K16560	exoH; succinoglycan biosynthesis protein ExoH
K16559	exoK; endo-1,3-1,4-beta-glycanase ExoK
K16558	exoL; succinoglycan biosynthesis protein ExoL
K16556	exoM; succinoglycan biosynthesis protein ExoM
K16555	exoO; succinoglycan biosynthesis protein ExoO
K16567	exoQ; exopolysaccharide production protein ExoQ
K16553	exoT; succinoglycan exporter
K16564	exoU; succinoglycan biosynthesis protein ExoU
K16563	exoV; succinoglycan biosynthesis protein ExoV
K16562	exoW; succinoglycan biosynthesis protein ExoW
K16565	exoX; exopolysaccharide production repressor protein
K16566	exoY; exopolysaccharide production protein ExoY

Others

K14083	mttB; trimethylamine corrinoid protein Co-methyltransferase
K14084	mttC; trimethylamine corrinoid protein

Figure 9.6: Co-evolution of genes in the exopolisaccaride biosynthesis module in Rhizobia. Graphical representation of the subnetwork containing gene families in the symbiosis cluster. Each node corresponds to a gene family, colored by attributed function. The thickness and transparency of the connecting edges corresponds to the weighted Pearson correlation coefficient between the evolutionary profiles of each pair of genes. Only genes with a correlation above a minimum threshold ($c_{\min} \geq 0.30$) are depicted. Boxes contain the KEGG Orthologue (KO) identifier and description of each gene.

9.3.2 Modules enrich for functionally related genes

We have demonstrated that our method is able to generate a network of gene families and robustly identify modules within it consisting of genes encoding functions relevant for the adaptation of Rhizobia to its environment. Next, sought to test whether the modules inferred by our approach significantly enrich for similar biological functions. In order to do this, we used a hypergeometric test with the Benjamini-Hochberg correction for multiple hypothesis testing (Methods). Briefly, this statistical test allows us to determine if genes in the same cluster are annotated with the same functions more often than expected by chance and thus determine if the cluster is enriched with a particular functional category or pathway. We found that the majority of the well-connected clusters (Figure 9.1 and Figure 9.3) have a statistically significant overlap with one or more KEGG pathways. The results of this analysis, summarized in Table 9.1, provide evidence that the modules found in the co-evolutionary network are biologically meaningful.

Furthermore, among the remaining clusters with the strongest signatures of co-evolution we also found modules enriching for functions relevant for adaptation of Rhizobia to their environment (Figure 9.1 and Figure 9.3) e.g. uptake of metabolites from the host plant as carbon sources (Glycerol utilization cluster; *glp* genes) (Ding *et al.*, 2012) or nitrogen sources (Xanthene uptake cluster; *yag* genes) (Botou *et al.*, 2014). Of note, we found a cluster loosely connected to the symbiosis module (Figure 9.3, third from the top) mostly consisting of genes encoding for uptake hydrogenases, which allow Rhizobia to recycle the hydrogen generated during the atmospheric nitrogen fixation process occuring inside of functional nodules (Baginsky *et al.*, 2002). Interestingly, we also found the cluster containing the genes encoding the canonical T4SS Vir complex loosely linked to the symbiosis module (Figure 9.3), indicating that this bacterial secretion system might also play a role in the interactions between Rhizobia and their host. Together, these results illustrate that the placement of each component in the network in relation with other clusters has potential implications about their functional context and their shared evolutionary histories.

Attributed functions	Size	KEGG pathway / module	P value
Symbiosis	29	Nitrogen metabolism	1.49E-04
		Symbiosis	≈ 0
Methane metabolism	22	Methane metabolism	4.97E-14
Hydrogenases	17	Degradation of aromatic compounds	2.67E-02
Exopolisaccaride biosynthesis	16	Exopolisaccaride	≈ 0
Denitrification	16	Nitrogen metabolism	3.89E-10
Type VI Secretion System	14	Bacterial secretion system	≈ 0
Type IV A Secretion System	10	Bacterial secretion system	≈ 0
Sulfur metabolism	9	Sulfur metabolism	≈ 0
Glycerol utilization	9	Saccharide, polyol, and lipid transport system	6.58E-14
Type IV B Secretion System	8	Bacterial secretion system	≈ 0

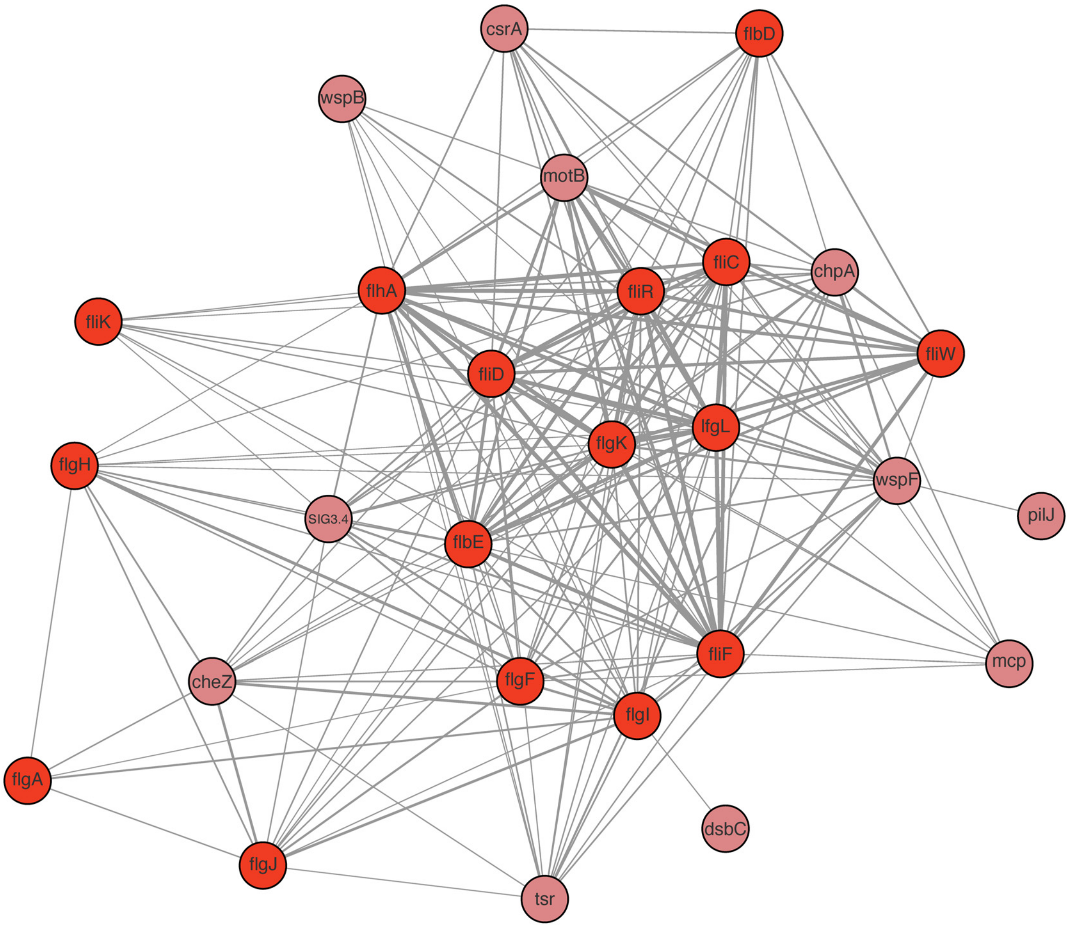
Table 9.1: Significant functional enrichment of large clusters. Table listing the results of a hypergeometric test of the overlap between genes in a cluster entries in the database classified as belonging to a given KEGG pathway or funcional module. Only the results of large clusters whose attributed function matched a pathway or module present in the database are shown. P values were adjusted for multiple hypothesis testing using the Benjamini-Hochberg correction.

9.3.3 Identification of novel symbiosis-related genes in Rhizobia

Next, we sought to determine if our approach would allow us to identify potentially novel gene families relevant for the symbiotic interaction between Rhizobia and their legume host. For instance, within the symbiotic cluster (Figure 9.4) we found two reductases that transform geranylgeranyl pyrophosphate, an intermediate metabolite thought to be a precursor of the gibberellin plant hormones (Hershey *et al.*, 2014). Interestingly, we also found the *rhiH* gene, which is a component of the operon encoding for the biosynthesis of an antimitotic rhizoxin complex (Partida-Martinez and Hertweck, 2005) to show signs of co-evolution with the known symbiosis genes (Figure 9.4). Notably, an ortholog of this gene was found to be present in the genome of the plant commensal *Pseudomonas fluorescens* and to play a role in the synthesis of and antifungal rhizoxin derivatives (Brendel *et al.*, 2007; Takeuchi *et al.*, 2015). We speculate that the strong co-evolutionary link between these genes and other genes known to be crucial for the development of functional nodules, such as the nodulation factors or the nitrogenase components, suggest that they too might play a role in the symbiotic interaction between Rhizobia and their hosts. Further experiments in the laboratory with knock-out mutants of model symbiotic Rhizobia in mono-associations with their legume host, which are currently underway, are necessary to determine the role that these potentially relevant gene families play in the establishment of functional nitrogen-fixing root nodules.

9.3.4 Functional modules in the *Arabidopsis* root and leaf microbiota

We next applied our method to a dataset with a broader phylogenetic range consisting of the sequenced genomes of the *Arabidopsis thaliana* leaf and root culture collections (Chapter 7). This dataset, which consists of 432 genomes belonging to 5 different phyla and 35 families, thus covering a wider portion of the bacterial tree of life, allowed us to further validate our approach. We generated a highly clustered (Watts-Strogatz coefficient 0.50), scale-free network consisting of 932 informative gene families and 1,937 edges depicting high correlation between their evolutionary profiles (Supporting Information). Clustering of the nodes of this network using the same approach as described above (see also Methods) revealed modules with clear signatures of co-evolution. Among these clusters, we found several examples of modules encoding genes potentially relevant for adaptation to the root and leaf environments, such as genes related to environmental information processing and motility (Bulgarelli *et al.*, 2015; Bai *et al.*, 2015). In particular, we could identify the cluster of genes related to the flagellar apparatus (Figure 9.7) showing strong signs of correlated evolutionary histories. Expectedly, due to the prevalence of these genes in the vast majority of the rhizobial genomes analyzed before, this cluster was not clearly identifiable in the Rhizobia network but nonetheless present among the top clusters in the *At*-SPHERE network.



Flagellar assembly

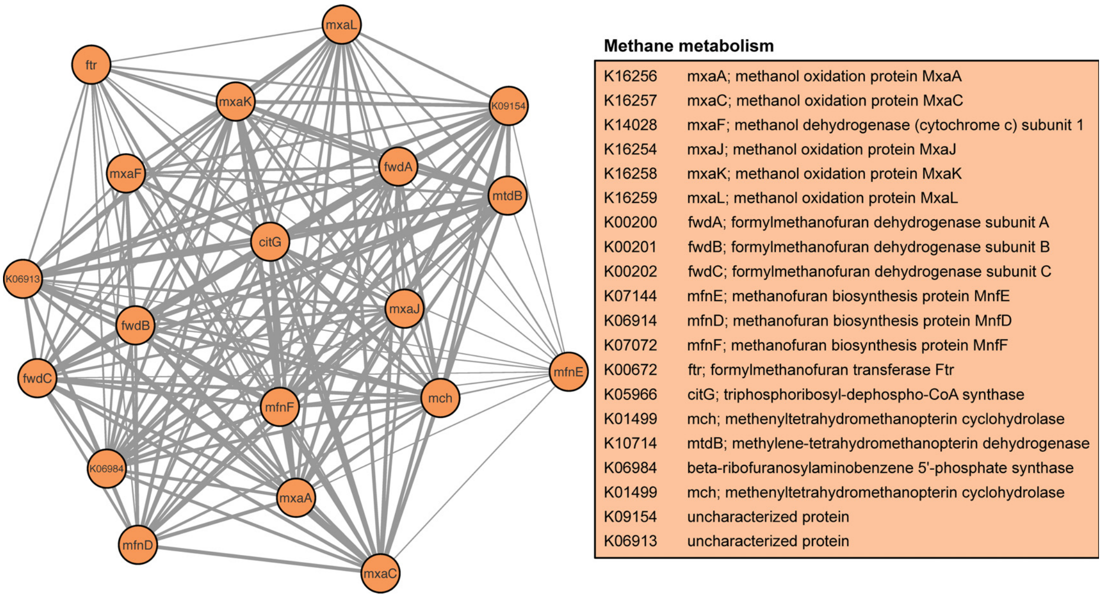
K02400	flhA; flagellar biosynthesis protein FlhA
K02406	fliC; flagellin
K02407	fliD; flagellar hook-associated protein 2 FliD
K02409	fliF; flagellar M-ring protein FliF
K02414	fliK; flagellar hook-length control protein FliK
K13820	fliR; flagellar biosynthetic protein FliR
K13626	fliW; flagellar assembly factor FliW
K02386	flgA; flagella basal body P-ring formation protein FlgA
K02385	flbD; flagellar protein FlbD
K02390	flbE; flagellar protein FlbE
K02391	flgF; flagellar basal-body rod protein FlgF
K02393	flgH; flagellar L-ring protein precursor FlgH
K02394	flgI; flagellar P-ring protein precursor FlgI
K02395	flgJ; flagellar protein FlgJ
K02396	flgK; flagellar hook-associated protein 1 FlgK
K02397	flgL; flagellar hook-associated protein 3 FlgL

Cell motility / others

K02660	pilJ; twitching motility protein PilJ
K06596	chpA; chemosensory pili system protein ChpA
K13488	wspB; chemotaxis-related protein WspB
K13491	wspF; chemotaxis response regulator WspF
K02557	motB; chemotaxis protein MotB
K03406	mcp; methyl-accepting chemotaxis protein
K03091	SIG3.4; RNA polymerase sporulation-specific sigma facto
K03563	csrA; carbon storage regulator
K03414	cheZ; chemotaxis protein CheZ
K05874	tsr; methyl-accepting chemotaxis protein I
K03981	dsbC; thiol:disulfide interchange protein DsbC

Figure 9.7: Co-evolution of genes in the flaggelar apparatus module in the *At*-SPHERE genomes. Graphical representation of the subnetwork containing gene families in the symbiosis cluster. Each node corresponds to a gene family, colored by attributed function. The thickness and transparency of the connecting edges corresponds to the weighted Pearson correlation coefficient between the evolutionary profiles of each pair of genes. Only genes with a correlation above a minimum threshold ($c_{\min} \geq 0.30$) are depicted. Boxes contain the KEGG Orthologue (KO) identifier and description of each gene.

Another gene module of environmental relevance was found to include a large number of COGs encoding proteins relevant for methane metabolism (Figure 9.8). Interestingly, gene families involved in methanogenesis and methanotrophy, which are present in a wide variety but not all of the plant-associated bacterial taxa, clustered together with methanol-based methylotrophy COGs present mostly in members of the genus *Methylobacterium*, predominantly isolated from the *Arabidopsis thaliana* phylosphere. The presence of this cluster of genes and their high degree of co-evolution is a finding consistent with previous analyses of metagenomic and metaproteomic data obtained from roots and leaves of rice (Knief *et al.*, 2012) and supports the hypothesis that methanogenesis and methanotrophy are traits of ecological importance under considerable selective pressure for plant-associated bacteria (Delmotte, 2009; Vorholt, 2012). Studying other gene families linked to this cluster, some of which remain unknown and poorly annotated, could provide novel insights into the evolution and the genetic basis of this relevant biogeochemical conversion process and might consitute the basis of further experimental work.



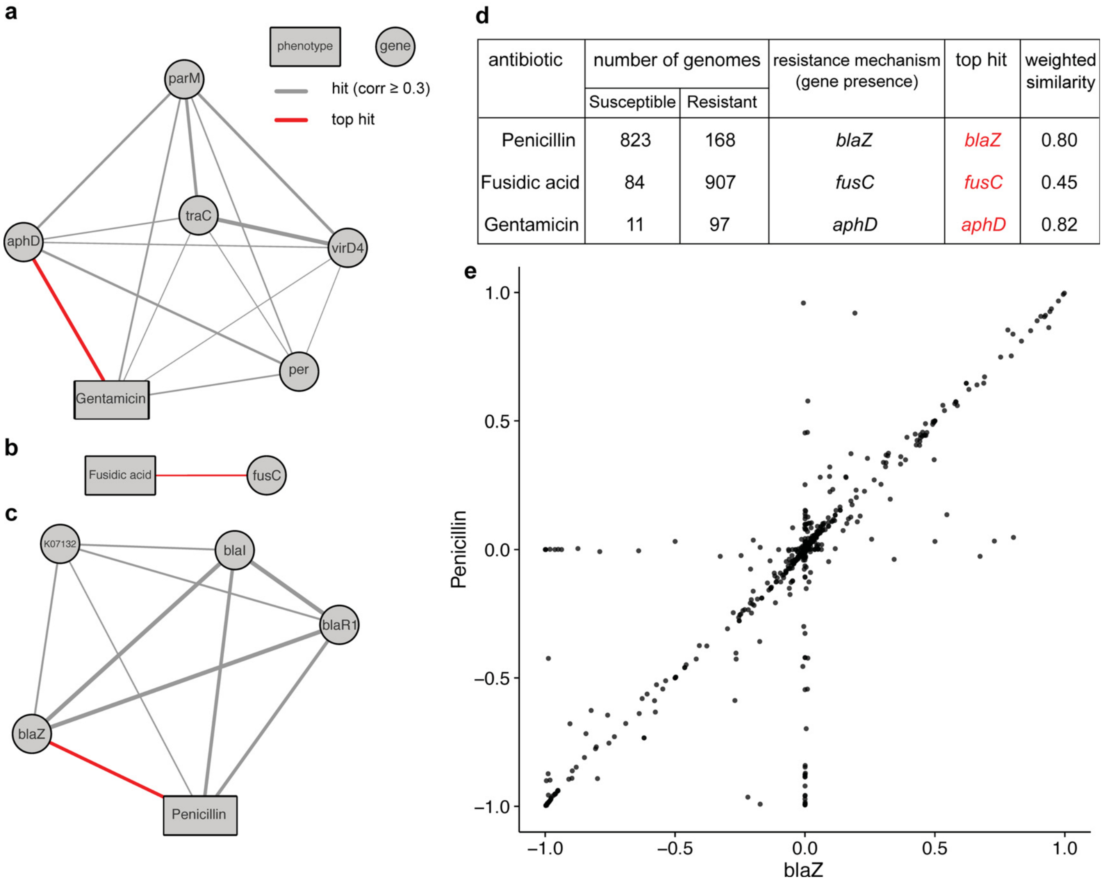
Methane metabolism

K16256	mxaA; methanol oxidation protein MxaA
K16257	mxaC; methanol oxidation protein MxaC
K14028	mxaF; methanol dehydrogenase (cytochrome c) subunit 1
K16254	mxaJ; methanol oxidation protein MxaJ
K16258	mxaK; methanol oxidation protein MxaK
K16259	mxaL; methanol oxidation protein MxaL
K00200	fwdA; formylmethanofuran dehydrogenase subunit A
K00201	fwdB; formylmethanofuran dehydrogenase subunit B
K00202	fwdC; formylmethanofuran dehydrogenase subunit C
K07144	mfnE; methanofuran biosynthesis protein MnfE
K06914	mfnD; methanofuran biosynthesis protein MnfD
K07072	mfnF; methanofuran biosynthesis protein MnfF
K00672	ftr; formylmethanofuran transferase Ftr
K05966	citG; triphosphoribosyl-dephospho-CoA synthase
K01499	mch; methenyltetrahydromethanopterin cyclohydrolase
K10714	mtdB; methylene-tetrahydromethanopterin dehydrogenase
K06984	beta-ribofuranosylaminobenzene 5'-phosphate synthase
K01499	mch; methenyltetrahydromethanopterin cyclohydrolase
K09154	uncharacterized protein
K06913	uncharacterized protein

Figure 9.8: Co-evolution of genes in the methane metabolism module in the *At*-SPHERE genomes. Graphical representation of the subnetwork containing gene families in the symbiosis cluster. Each node corresponds to a gene family, colored by attributed function. The thickness and transparency of the connecting edges corresponds to the weighted Pearson correlation coefficient between the evolutionary profiles of each pair of genes. Only genes with a correlation above a minimum threshold ($c_{\min} \geq 0.30$) are depicted.

9.3.5 Predicting the genetic basis of antibiotic resistance in *S. aureus*

Next we applied our approach to a set of closely related genomes from a different environment to investigate the interaction between co-evolving genes and a given phenotype. We retrieved the raw sequencing reads, assembled and annotated the genomes corresponding to 991 strains of *Staphylococcus aureus* for which their phenotype of resistance to a variety of antibiotics had been previously determined (Gordon *et al.*, 2014) and which have been used as an example in recent efforts to apply genome-wide associaton studies (or GWAS) to bacteria (Earle *et al.*, 2016). In particular, we focused on three antibiotics whose resistance or susceptibility is determined by the presence or absence, respectively, of a particular gene (Figure 9.9).



antibiotic	number of genomes		resistance mechanism (gene presence)	top hit	weighted similarity
	Susceptible	Resistant			
Penicillin	823	168	*blaZ*	*blaZ*	0.80
Fusidic acid	84	907	*fusC*	*fusC*	0.45
Gentamicin	11	97	*aphD*	*aphD*	0.82

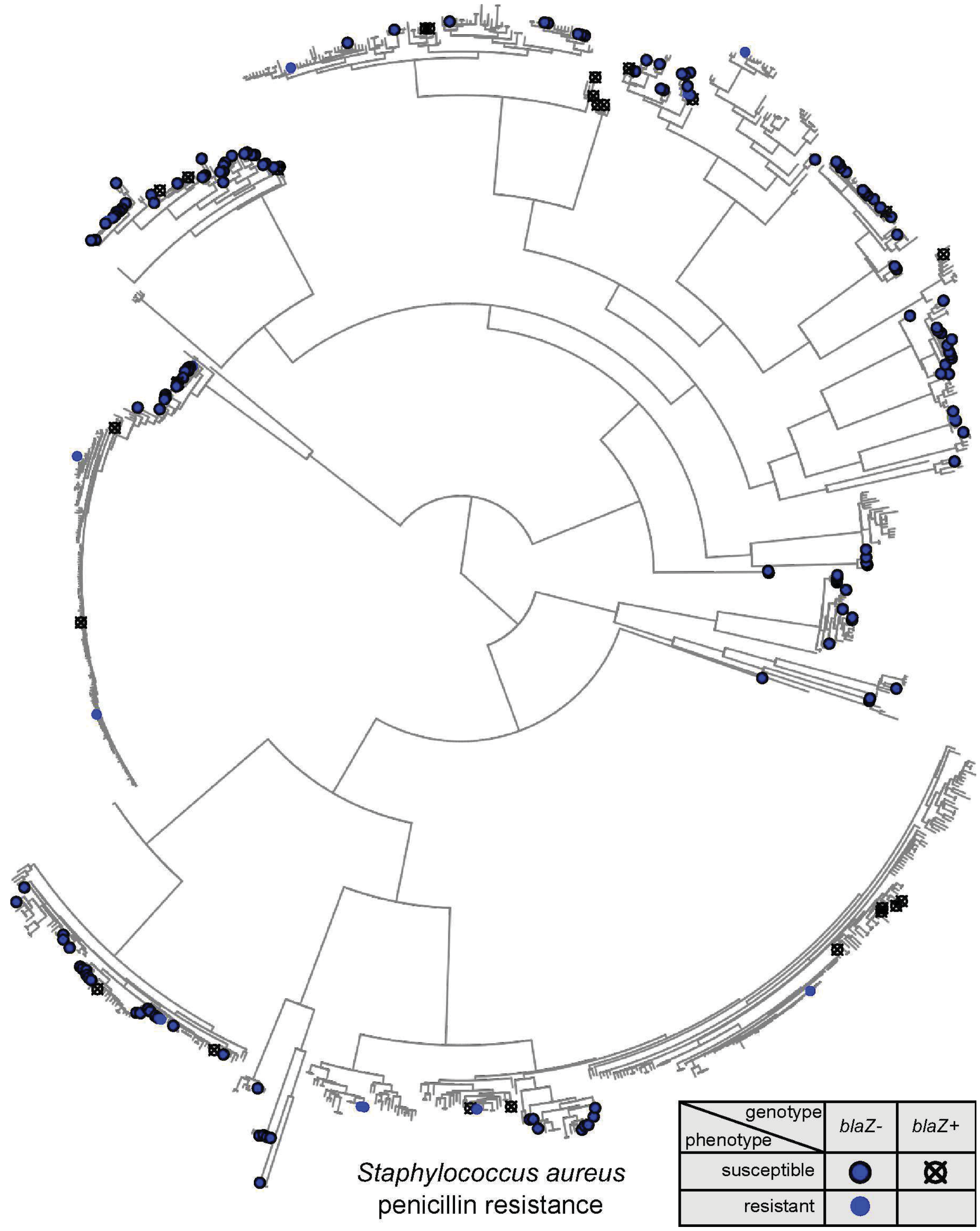
Figure 9.9: Significantly correlated genes with patterns gain and loss of antibiotic resistance in *S. aureus*. (Caption on next page)

Figure 9.9: Significantly correlated genes with gain and loss of antibiotic resistance phenotypes in *S. aureus*. (Figure on previous page) **a-c**, antibiotic resistance phenotypes and their significantly correlated gene families. Thickness and transparency of the connecting edges corresponds to the weighted correlation coefficients. Only genes with a correlation above a minimum threshold ($c_{\min} \geq 0.30$) are depicted. The top hit of each phenotype is highlighted with a red connecting line. **d**, table listing the known genetic basis of each phenotype, the spread of susceptibility in the dataset and the inferred causal gene. **e**, weighted Pearson correlation of gain / loss patterns inferred for the phenotype of penicillin resistance and its causal gene *blaZ*.

Here, we applied a variant of our method in which the phyletic patterns for each phenotype were included together with the profiles of presence or absence of each gene family with the goal of identifying genes or modules associated with the phenotypes. For all three of the tested antibiotics, the top hit accurately corresponded with the known gene responsible for the phenotype (Figure 9.9), and our method succeeded in correctly determining the genetic basis of the resistance. Our approach also captured other co-evolving genes as secondary hits, some of which are known to be relevant for the phenotype, as well as others of undetermined function. For instance, the phenotype of resistance to gentamicin was found to be strongly linked to the *aphD* gene (weighted correlation corr=0.82; Figure 9.9 A and D) whose presence is known to cause the phenotype, as well as a small number of hits with a much lower correlation.

Similarly, resistance to beta-lactam antibiotics is related to the presence of genes encoding the degrading beta-lactamases (*bla* genes; Zeng and Lin (2013)). In particular, resistance to penicillin in *Staphylococcus aureus* is mediated by the *blaZ* gene (Figure 9.10). Interestingly, our method indicates a very high correlation between the beta-lactamase genes, with *blaZ* as the top hit (corr=0.80; Figure 9.9 C and D). A closer inspection at the evolutionary profile of the *blaZ* gene revealed that the evolutionary profile of the gene correlates with gains and losses of the penicillin resistance phenotype, not only for branches with high likelihood of gain or loss but also for those with intermediate probabilities (Figure 9.9 E; correlated values in the range between -1 and 1). This result illustrates the advantage of considering continuous values to model the gene gain and loss process instead of using discrete events as branches with intermediate values are also highly informative to infer co-evolution between gene families and / or phenotypes.

We also explored resistance to fusidic acid, which is partially associated to the presence of *fusC*, a gene present in a transposable element whose product interferes with the interaction between fusidic acid and the prokaryotic elongation factor (EF-G).



genotype / phenotype	*blaZ-*	*blaZ+*
susceptible	●	⊠
resistant	●	

Figure 9.10: Species tree and phylogenetic distribution of antibiotic resistance in *S. aureus* The species tree was generated using a Maximum Likelihood approach based on concatenated biallelic sites. Tip shapes and colors represent the spread of the penicillin resistance phenotype as well as its causal gene.

Our method successfully determined the genetic basis for the resistance to fusidic acid, albeit in this instance with a substantially lower weighted correlation (corr=0.45; Figure 9.9 B and D). This is due to the fact that *fusC*-encoded resistance is not the only mechanism to prevent susceptibility to fusidic acid, which can be also caused by SNP substitutions in the *fusA* gene, which is part of the *Staphylococcus aureus* core genome (O'Neill *et al.*, 2007). Despite of the fact that resistance-conferring *fusA* polymorphisms explain around a third of the positive instances of the phenotype in the dataset, and of the strong population structure (most of the resistance strains mediated by the presence of *fusC* are present in only two *S. aureus* clades), our method correctly reported *fusC* as the most likely causal gene for the phenotype, with no false positives (Figure 9.9 B and D). Taken together, these results illustrate the potential of our method to link gene families and gene family clusters to phenotypes with a predictive power comparable to state-of-the-art GWAS approaches even for sets of genomes of very closely related strains.

9.4 Discussion

We present a phylogenetic approach for determining clusters of co-evolving genes and their network organization by modeling gene gain and loss as continuous instead of discrete events based on Maximum Likelihood reconstruction of ancestral states. Our method accounts for uncertainty in the reconstruction of the ancestral states as well as in the inference of the species tree. By applying our method to genomes of Rhizobia (n=1,314), the *Arabidopsis thaliana* root and leaf culure collections (n=432) and a population of antibiotic resistant *Staphylococcus aureus* (n=991) we have demonstrated that our approach can be applied to very large datasets of genomes and that it can robustly infer functionally-related modules of genes relevant for adaptation to diverse environments, as well as the genetic basis of tested phenotypes, even in the presence of strong population structure. Furthermore, by comparing the results obtained with a well-curated functional annotation database, we have determined that in plant associated bacteria in general and in rhizobia in particular, genes predicted to co-evolve tend to be involved in the same pathways much more frequently than expected by chance (Table 9.1) and that proximity in the co-evolutionary network clearly correlates with functional similarity (Figure 9.2).

We have reconstructed a network of co-evolving genes for for several species of Alphaproteobacteria, collectivelly known as Rhizobia, which are known to engage in symbiotic relationships with legumes and which are core components of the plant microbiota (see Chapter 6 and Chapter 8) (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Bulgarelli *et al.*, 2015; Schlaeppi *et al.*, 2014; Edwards *et al.*, 2015; Hacquard *et al.*, 2016). Analysis of this network revealed several gene modules with strong signatures of co-evolution and which are related to important adaptive processes in rhizobia, specifically for their ability to colonize and interact with their plant hosts (Figure 9.3). Among the largest and better connected clusters we found one containing gene families known to be crucial for nitrogen-fixing symbionts (Figure 9.4), such as the nodulation factor synthesis and regulation genes (*nod* genes) as well as the *nif* genes, which encode components of the nitrogenase complex. Surprisingly, we found several additional gene families that seem to co-evolve with these known key symbiosis genes, implying their potential and currently unknown role in the complex molecular dialog between nitrogen-fixing rhizobia and legumes. Further testing –currently underway– of these candidate genes in the laboratory in binary interactions with their plant hosts will help elucidate their biological relevance. In addition, we also found clusters containing a repertoire of exopolysaccaride biosynthesis genes (*exo* genes; Figure 9.3 and Figure 9.6) which are previously known to be relevant for Rhizobia-host interactions (Jones *et al.*, 2009) as well several gene clusters encoding diverse proteins related to the uptake of plant-derived compounds (Figure 9.1 and Figure 9.3). Interestingly, we also found three separate clusters of genes encoding the bacterial T6SS (Figure 9.3) and two distinct T4SS: the canonical system, encoded by the *vir* genes and which is used by *Agrobacterium* to transfer of T-DNA into the plant cell (Figure 9.5), and the T4SS B or Icm/Dot complex, encoded by the *icm* and *dot* genes which is known to be present almost exclusively in pathogenic bacteria such as *Legionella pneumophila* and found to be important for virulence by means of effector protein delivery into the eukaryotic host cell (Christie *et al.*, 2014). Together, these results underline the strength with which adaptation to the plant environment imposes selective pressures that coordinate genome evolution in Rhizobia beyond the well-studied nitrogen-fixing and nodule-inducing symbiosis with legumes. Our data also suggest that secretion systems, particularly T4SS, is an important genomic feature for the adaptation of Rhizobia to their environment. Laboratory experiments with T4SS A and Icm/Dot T4SS knockout mutants with axenic plants will allow us to test these hypothesis in the near future.

Furthermore, we have shown that this approach can also be applied to datasets containing genomes from distantly related bacterial species, such as the 432 isolates from 5 phyla and 35 different families that constitute the *Arabidopsis thaliana* root and leaf culture collections (Bai *et al.*, 2015). This *in silico* experiment allowed us to further validate our method by detecting clusters of genes that in the dataset of more closely related Rhizobia were part of the core genome, such as those involved in motility (Figure 9.7). Our results also confirmed previous studies indicating the importance of motility genes in the root microbiome (Bulgarelli *et al.*, 2015; Bai *et al.*, 2015) and methane metabolism for the root and the leaf microbiota (Delmotte, 2009; Bai *et al.*, 2015), where gene families involved in methanogenesis and methanotrophy (Figure 9.8) showed strong evidence of co-evolution and of being relevant genomic features for adaptation to the plant environment.

In addition, we have tested our approach in a dataset of 991 closely related *Staphylococcus aureus* genomes for which resistance to a panel of antibiotics had previously been determined (Gordon *et al.*, 2014; Earle *et al.*, 2016). We have shown that our method can also be used to determine links between discrete phenotypes and individual genes or larger gene functional modules by reconstructing the probability of resistance to several antibiotics along each branch of the *S. aureus* phylogeny. For each tested drug whose resistance phenotype was determined partially or totally by presence or absence of a gene, the top hit identified by our method corresponded to the known causal gene (Figure 9.9). These results are comparable with state-of-the-art approaches for bacterial association studies that account for lineage effects (Earle *et al.*, 2016). Of note, unlike most of these approaches, the use of our method is not limited to sets of closely related strains from the same population but is also applicable to genomes from unrelated bacterial species, even from different phyla, as indicated by the results discussed above. Conversely, our approach is only able to determine links between genes and phenotypes when the causal mechanism is the presence or absence of a gene, as it does not account for polymorphic variants within the same gene families. A possible way to overcome this limitation consists of reconstructing evolutionary profiles of all k-mers found in the dataset above a certain frequency threshold instead of gene families, thus allowing to infer relationships between co-evolving genomic variants and phenotypes. This approach would also remove another crucial caveat of this and other phylogenetic methods for the inference of co-evolution, which is their limitation to genes that are subject to gain and loss and therefore not part of the core-genome. However, despite of being a promising avenue for future research, this extension would involve the calculation of a very large number pair-wise weighted correlations, and additional optimizations are required to make these operations feasible for large datasets such as those analyzed in this study.

Previous studies have shown that gene gain (mostly via HGT) (Soucy *et al.*, 2015) and gene loss (Kuo and Ochman, 2009) are the predominant drivers of evolution in bacteria. Taken together, our results also indicate that these adaptive processes in plant-associated bacteria act at in a concerted manner at the level of gene modules rather than at the level of single genes, a finding which has important implications for the study of microbial genome evolution and adaptation to host-associated environments, not only of individual species but also of their entire microbial communitites. Phylogenetic methods such as the one presented in this study have the potential to detect these processes at the community level (e.g. in Rhizobia and in the *At*-SPHERE genomes) as well as at the populataion level (e.g. for *Staphylococcus aureus*). These potential applications might be specially relevant for microbiome studies, where microbe-microbe as well as microbe-host interactions determine complex co-evolutionary relationships between multiple parties, rather than e.g. adaptation of isolated species to a novel or changing environment.

In summary, we have developed a phylogenetic method to detect co-evolution between pairs of gene families which accounts for uncertainty in the reconstruction of ancestral events and the inference of the species tree. Applying our method to three distinct datasets of varying phylogenetic diversity (from community to population level data) we present three key results: i) clustering of the inferred co-evolution network reveals functional modules relevant for colonization and survival in different environments, including host-associated commensal and symbionts. ii) Clusters of co-evolving genes significantly enrich for related functional categories and pathways. iii) The hierarchical and highly clustered structure of the co-evolutionary network implies that the predominant mechanisms of bacterial genome evolution act at the level of gene modules instead of at the level of individual genes. Although these results are encouraging, the choice of taxa to be analyzed and their sampling depth play a crucial role in determining the predictive power of this and related methods (Campillos *et al.*, 2006; Cordero *et al.*, 2008; Cohen *et al.*, 2012). However, we have shown that, for traits and gene families showing significant variation within a dataset, our approach can robustly identify sets of genes involved in relevant biological processes as well as links to corresponding phenotypes, largely unaffected by population structure or sampling bias. As the available number of sequenced genomes increases, further development of phylogenetic methods (for example to be able to account for other gene variants besides presence or absence) has the potential to greatly improve our understanding of the processes that drive the adaptation of microbes in the context of the complex communities they form.

9.5 Methods

9.5.1 Genome assembly

Genomes sequenced using the Illumina platform were assembled by first passing their short reads through a quality and length trimming filtering with Trimmomatic (Bolger *et al.*, 2014) and subsequently assembled using an ensemble of two approaches: A5 (Tritt *et al.*, 2012) and SOAPdenovo (Li, 2010). In each case, the assembly with the smaller number of scaffolds was selected. Subsequently, contigs smaller than 1,000 basepairs were removed from the assemblies. Selected genomes of the Rhizobia dataset were sequenced using the Pacific Biosciences platform and their long reads assembled into complete genomes using the HGAP assembler (Chin *et al.*, 2013). Detailed statistics concerning the assemblies of all genomes analyzed in this study can be found in the Supporting Material. Accession numbers corresponding to the raw reads of all genomes, including also those of previously published organisms can likewise be found in the Supporting Material.

9.5.2 Inference of homology relationships between genes

To determine homology relationships between gene sequences we predicted putative protein-coding sequences using Prodigal (Hyatt *et al.*, 2010). Annotation of candidate Open Reading Frames (ORFs) was then conducted using the KEGG Orthologue (KO) database (Kanehisa and Goto, 2000; Kanehisa, 2014) as previously described (Bai *et al.*, 2015). Briefly: sequences from each KO were aligned using Clustal Omega (Sievers, 2011) and Hidden Markov Models (HMMs) generated based on each multiple sequence alignment using the HMMER suite (Eddy, 2011). Subsequently, the HMMs were employed to search all putative protein-coding sequences. Sequences matching a KO with an E value lower than 10^{-5} with a coverage of at least 70% of the total length were assigned to said KO. All sequences assigned to the same KO were considered homologous and classified as belonging to the same Cluster of Orthologous Genes (COG). For the *At*-SPHERE dataset, given its comparatively smaller size, we also conducted a more comprehensive prediction of COGs using OrthoFinder (Emms and Kelly, 2015), a sequence-based de-novo method for orthogroup inference. Subsequently, for each Cluster of Orthologous Genes (COG) we generated a phyletic pattern consisting of a binary vector of presence / absence in each genome of the dataset. The total of all phyletic patterns, or phyletic matrices, were used as input for all downstream analyses (Supporting Material).

9.5.3 Maximum likelihood inference of ancestral characters

In addition to the matrix of phyletic patters, our method requires a fully dychotomous tree that accurately reflects the phylogenetic relationship between the genomes. For this purpose, we extracted from each proteome a set of single-copy, vertically inherited phylogenetic markers that are present in the majority of sequenced bacterial genomes called AMPHORA genes (Wu and Eisen, 2008) using a set of HMMs and the hmmsearch tool (Eddy, 2011). The resulting sequences were concatenated independently for each genome and then aligned using Clustal Omega (Sievers, 2011). For each dataset, rooted species trees were generated using the AMPHORA multiple sequence alignments using MrBayes (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) with a strick molecular clock and a General Time Reversive model (GTR+G+I; Supporting Material). Similar results as reported above were obtained by reconstructing the species tree using an alternative Maximum Likelihood approach implemented in FastTree (Price *et al.*, 2010) with a GTR model of DNA evolution and without a molecular clock constrain (data not shown). For the closely related *Staphylococcus aureus* genomes, which do not have enough polymorphism in their AMPHORA marker genes to infer a highly resolved phylogeny, a species tree was constructed using the same approach as in (Earle *et al.*, 2016) by using RAxML version 7.7.6 (Stamatakis, 2014) on SNP data obtained from aligning raw reads to a reference genome. Polytomies in the trees were resolved by inserting branches of zero length to avoid conflicts in the inference of ancestral characters.

Given the matrix of phyletic patterns and the species tree we used a Maximum Likelihood (ML) approach (Pagel, 1994) for the estimation of ancestral states using the implementation provided with the R package ape (Paradis *et al.*, 2004), which employs a 'two-pass' algorithm for the joint estimation of the likelihood of the ancestral states (Felsenstein and Felenstein, 2004; Yang, 2006). This approach gives similar results as stochastic mapping (Pupko *et al.*, 2000; Paradis *et al.*, 2004) while being much faster, thus enabling the estimation of ancestral characters for very large datasets (several thousands of genomes or more).

For a dataset with N gene families or COGs and M extant genomes, the output of this analysis is a vector $\mathbf{x}_k$ of length $2M-1$, equal to the number of nodes in the rooted species tree, for each gene family $k = 1, 2, ..., N$ in the dataset:

$$\mathbf{x}_k = \{p(x_1), p(x_2), ..., p(x_{M-1}), x_1, x_2, ..., x_M\}$$

where $p(x_i)$ is the probability of the gene family of being present in each ancestor genome $i = 1, 2, ..., M - 1$, and $x_j = \{0, 1\}$ indicates the presence or absence of the gene family in each of the $j = 1, 2, ..., M$ extant genomes.

9.5.4 Modeling gene gains and losses

Based on the estimated probabilities of a gene family to be present in the ancestral genomes, $\mathbf{x}_k$, we model the gain and loss events for each branch $i = 1, 2, ..., 2N - 2$ of the species tree as

$$b_i = p(x_{\text{child}}) - p(x_{\text{parent}})$$

where $p(x_{\text{child}})$ and $p(x_{\text{parent}})$ are the probabilities of the gene family of being present at the ancestral genome corresponding to the parent or child node of the i-th branch of the species tree, respectively. This gives a vector of real values

$$\mathbf{b}_k = \{b_1, b_2, ..., b_{2N-2}\}$$

with $b_i \in (-1, 1)$ for each COG or gene family $k = 1, 2, ..., N$ in the dataset. We designate these vectors b_k evolutionary profiles. Negative values correspond to a likely loss event whereas positive values represent probably gain events along a given branch of the species tree. We reason that these vectors contain more information concerning the past evolutionary history of a gene family than a representation of gain / loss events as separate discrete values as it has been previously done for similar purposes (Cordero *et al.*, 2008; Cohen *et al.*, 2012), given the fact that low probability events are represented with values close to 0, and not only highly probable gains or losses, which here correspond to values closer to 1 or -1, respectively. Therefore, the absolute value of each entry in an evolutionary profile contains infromation about how likely an event is to have occurred at a given branch of the species tree.

9.5.5 Clustering of evolutionary profiles into functional modules

We speculate that gene families whose evolutionary profiles show a high correlation are more likely to be involved in the same biological process as gene families with a low correlation. In order to quantify and test this hypothesis, we next calculate a meassure of similarity between each pair of evolutionary profiles based on a weighted variant of the Pearson correlation coefficient. Briefly, if the Pearson correlation coefficient ρ between two vectors $\mathbf{x}$ and $\mathbf{y}$ is

$$\rho_{\mathbf{x},\mathbf{y}} = \frac{\operatorname{cov}(\mathbf{x},\mathbf{y})}{\sigma_{\mathbf{x}}\sigma_{\mathbf{y}}} = \frac{E[(\mathbf{x}-\mu_{\mathbf{x}})(\mathbf{y}-\mu_{\mathbf{y}})]}{\sigma_{\mathbf{X}}\sigma_{\mathbf{Y}}}$$

where $\mu_{\mathbf{x}}$ and $\rho_{\mathbf{x}}$ are the mean and standard deviation of $\mathbf{x}$, respectively, E is the expectation and $\operatorname{cov}(\mathbf{x},\mathbf{y})$ is the covariance between $\mathbf{x}$ and $\mathbf{y}$. Similarly, the weighted correlation coefficient corr is defined as follows:

$$\operatorname{corr}(\mathbf{x},\mathbf{y};\mathbf{w}) = \frac{\operatorname{cov}(\mathbf{x},\mathbf{y};\mathbf{w})}{\sqrt{\operatorname{cov}(\mathbf{x},\mathbf{x};\mathbf{w})\operatorname{cov}(\mathbf{y},\mathbf{y};\mathbf{w})}}$$

where

$$\operatorname{cov}(\mathbf{x},\mathbf{y};\mathbf{w}) = \frac{\sum_i w_i(x_i - \mathrm{m}(x;w_i))(y_i - \mathrm{m}(y;w_i))}{\sum_i w_i}$$

and $\mathrm{m}(\mathbf{x})$ is the weighted mean of vector $\mathbf{x}$

$$\mathrm{m}(\mathbf{x};\mathbf{w}) = \frac{\sum_i w_i x_i}{\sum_i w_i}$$

where each weight w_i is associated with a branch in the species tree. In this case, we take the posterior probabilities for each branch obtained during the Bayesian inference of the phylogeny. This weighted correlation coefficient considers that branches have varying degrees of importance, represented by the vector weights or posterior probabilities $\mathbf{w} = \{w_1, w_2, ..., w_i\}$, thus reducing the impact that observations from branches with low confidence have in the overal calculation of similarities between evolutionary profiles. Similarly, the vector of weights can be chosen arbitrarily to assign a higher importance to particular branches of interest in the phylogeny, e.g. those corresponding to particularly relevant speciation events.

For a set of N gene families, these calculations result in a symmetric adjacency matrix

$$A = (a_{i,j})$$

of dimensions $N \times N$ whose elements

$$a_{i,j} = \operatorname{corr}(\mathbf{b}_i, \mathbf{b}_j; \mathbf{w})$$

correspond to the weighted correlation coefficient between the profiles $\mathbf{b}_i$ and $\mathbf{b}_j$, corresponding to the i-th and j-th gene families. Subsequently, from this adjacency matrix we infer a network of evolutionary interactions by generating links connecting gene families or COGs whose weighted correlation coefficient exceeds a certain value $c_{\min}$. Additionally, gene families which are present in almost all extant organism (core genome) or, alternatively, in a very small subset of the genomes in the dataset, will show high correlation coefficients without necessarily implying true co-evolution. In order to remove nodes with non-informative profiles, that is, from genes with only few inferred events of gain or loss, we filter all gene families with evolutionary profiles having an total amount of information $I(\mathbf{b}_i) = |\mathbf{b}_i|$ above a certain minimum value $I_{\min}$, which should be related to the total accumulated branch length of the species tree. We provide a table of edges for each of the analyzed networks using thresholds $c_{\min} = 0.3$ and $I_{\min} = 10$ for the Rhizobia and *At*-SPHERE datasets and $c_{\min} = 0.3$ and $I_{\min} = 5$ for the *Staphylococcus aureus* genomes.

A large number of methods for clustering edges of a network representing interactions have been proposed (Eisen *et al.*, 1998; Tavazoie *et al.*, 1999; Raychaudhuri *et al.*, 2000; Alter *et al.*, 2000; Wittkop *et al.*, 2011), and applied profusely, especially on networks inferred from expression data. Here we use an unsupervised clustering algorithm for networks based of simulation of stochastic flow in graphs known as Markov Cluster Algorithm (MCL) (Van Dongen, 2001). The MCL algorithm finds clusters in a network by simulating random walks consisting on alternating two operations designated expansion and inflation. Briefly, these two alternating phases assign probabilities to each pair of nodes of being the point of departure and arrival of a random walk. Since nodes belonging to the same cluster will generally have more paths in common than those in different clusters, the probability of two nodes being classified together will relate to the number of paths between them. After a sufficiently large number of iterations, convergence is achieved by segmenting the network into separate components, which are then considered as separate clusters. The MCL approach has only one parameter, I, which relates to the 'granularity' of the clustering solution. For all calculations presented here, we have selected an intermediate value of $I = 1.5$. As a likely alternative to MCL, an alternative approach known as Affinity Propagation (AF) (Frey and Dueck, 2007) has been proposed as a fast and reliable method to finding cluster structure in large and complex networks. However, a comparison between these two approaches has shown that MCL is slightly more robust with respect to noise than AF in biological networks (Vlasblom and Wodak, 2009).

An alternative to derivating an adjacency matrix and a network and performing clustering of the nodes would be defining a distance meassure based on the previously described correlation coefficient. Simply, a well-defined distance between two evolutionary profiles $\mathbf{b}_i$ and $\mathbf{b}_j$ corresponding to two gene families i and j, is the weighted Pearson's distance

$$d_{\mathbf{b}_i,\mathbf{b}_j} = 1 - \mathrm{corr}(\mathbf{b}_i, \mathbf{b}_j; \mathbf{w})$$

which provides values in the interval $(0, 2)$ and satisfies the triangle inequality. By calculating all pairwise weighted correlation distances we can embed the evolutionary profiles in a matric space which will then allow to apply standard clustering methods such as hierarchical clustering, k-means, PAM, etc. Further computational experiments are necessary, however, to compare the performance of said approach with respect to the robust network-based clustering methodology which we have employed.

In order to analyze and visualize the resulting networks we employ Cytoscape (Shannon *et al.*, 2003) and the R packages qgraph (Epskamp *et al.*, 2012), igraph (Csardi and Nepusz, 2006) and ggplot2 (Wickham, 2009). Annotated network tables compatible with Cytoscape are provided for all analyzed datasets (Supporting Material).

9.5.6 Statistical analyses of module functional enrichment

In order to test whether the gene family clusters found using our methodology have a biological relevance, we perform a hypergeometric test on the gene family KEGG pathway and module annotations, similarly as previously done for microbial gene network analyses (Trevino III *et al.*, 2012). This statistical test allows us to determine if COG belonging to the same cluster in our network are also classified as pertaining to the same pathway in a curated database more often than expected by chance. The hypergeometric test calculates the probability that a cluster of size n of having k genes in common with a KEGG pathway of size M in the context of a network consisting of a total of N gene families. For a randomly drawn set of genes, which we take as the null hypothesis, the probability is

$$P = \frac{\binom{M}{k}\binom{N-M}{n-k}}{\binom{N}{n}}.$$

In order to reject the null hypothesis we sought to determine whether the overlap between the cluster and the KEGG pathway is unlikely to happen by chance, thus implying a meaningful biological relationship between the members of the cluster. The P value is then calculated as the probability of randomly drawing a number k or larger genes annotated as belonging to a specific pathway from the total set of N present in the network. Subsequently, P values are corrected for multiple testing using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995), with a false discovery rate α=0.05.

9.5.7 Code availability

All code is available upon request and will be uploaded to a public repository upon publication, in addition to all raw and intermediate data, which will be made available at the following site: http://www.mpipz.mpg.de/R_scripts.

9.6 Author contributions

R.G.-O. and A.C.M. conceived research. R.G.-O. implemented the method and performed the experiments. R.G.-O. and A.C.M. interpreted the data. R.G.-O. wrote the manuscript with comments from A.C.M.

9.7 Acknowledgements

This work was supported by funds from the 'Cluster of Excellence on Plant Sciences' program funded by the Deutsche Forschungsgemeinschaft.

9.8 Supporting material

The supporting material corresponding to this section, including all supplementary tables, databases and figures can be accessed during revision at the following address http://www.at-sphere.com/downloads/evol.tar.gz and have not been included in this thesis due to space limitations.

References

Abarenkov, K., R. Henrik Nilsson, K.-H. Larsson, I. J. Alexander, U. Eberhardt *et al.* (2010). The UNITE database for molecular identification of fungi–recent updates and future perspectives. *The New Phytologist*, **186** (2): 281–285.

Abbeele, P. V. d., T. V. d. Wiele, W. Verstraete and S. Possemiers (2011). The host selects mucosal and luminal associations of coevolved gut microorganisms: a novel concept. *FEMS Microbiology Reviews*, **35** (4): 681–704.

Abbo, S., R. Pinhasi van Oss, A. Gopher, Y. Saranga, I. Ofner *et al.* (2014). Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends in Plant Science*, **19** (6): 351–360.

Adams, M. A., T. L. Turnbull, J. I. Sprent and N. Buchmann (2016). Legumes are different: Leaf nitrogen, photosynthesis, and water use efficiency. *Proceedings of the National Academy of Sciences of the United States of America*, **113** (15): 4098–4103.

Agler, M. T., J. Ruhe, S. Kroll, C. Morhenn, S.-T. Kim *et al.* (2016). Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLoS biology*, **14** (1): e1002352.

Almagro-Moreno, S. and E. F. Boyd (2009). Sialic Acid Catabolism Confers a Competitive Advantage to Pathogenic Vibrio cholerae in the Mouse Intestine. *Infection and Immunity*, **77** (9): 3807–3816.

Alter, O., P. O. Brown and D. Botstein (2000). Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences of the United States of America*, **97** (18): 10101–10106.

Amaral-Zettler, L. A., E. A. McCliment, H. W. Ducklow and S. M. Huse (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS One*, **4** (7): e6372.

Anders, S., D. J. McCarthy, Y. Chen, M. Okoniewski, G. K. Smyth *et al.* (2013). Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nature Protocols*, **8** (9): 1765–1786.

Anderson, M. J. and T. J. Willis (2003). CANONICAL ANALYSIS OF PRINCIPAL COORDINATES: A USEFUL METHOD OF CONSTRAINED ORDINATION FOR ECOLOGY. *Ecology*, **84** (2): 511–525.

Angus, A. A. and A. M. Hirsch (2010). Insights into the history of the legume-betaproteobacterial symbiosis. *Molecular Ecology*, **19** (1): 28–30.

Arumugam, M., J. Raes, E. Pelletier, D. Le Paslier, T. Yamada *et al.* (2011). Enterotypes of the human gut microbiome. *Nature*, **473** (7346): 174–180.

Baginsky, C., B. Brito, J. Imperial, J.-M. Palacios and T. Ruiz-ArgÃ¼eso (2002). Diversity and evolution of hydrogenase systems in rhizobia. *Applied and Environmental Microbiology*, **68** (10): 4915–4924.

Bai, Y., D. B. Mueller, G. Srinivas, R. Garrido-Oter, E. Potthoff *et al.* (2015). Functional overlap of the Arabidopsis leaf and root microbiota. *Nature*, **528** (7582): 364–369.

Bais, H. P., T. L. Weir, L. G. Perry, S. Gilroy and J. M. Vivanco (2006). The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.*, **57**: 233–266.

Barcenas-Moreno, G., M. Gomez-Brandon, J. Rousk and E. Baath (2009). Adaptation of soil microbial communities to temperature: comparison of fungi and bacteria in a laboratory experiment. *Global Change Biology*, **15** (12): 2950–2957.

Barrangou, R., C. Fremaux, H. Deveau, M. Richards, P. Boyaval *et al.* (2007). CRISPR provides acquired resistance against viruses in prokaryotes. *Science (New York, N.Y.)*, **315** (5819): 1709–1712.

Barret, M., M. Briand, S. Bonneau, A. PrÃ©veaux, S. ValiÃ¨re *et al.* (2015). Emergence Shapes the Structure of the Seed Microbiota. *Applied and Environmental Microbiology*, **81** (4): 1257–1266.

Batterman, S. A., N. Wurzburger and L. O. Hedin (2013). Nitrogen and phosphorus interact to control tropical symbiotic N2 fixation: a test in Inga punctata. *Journal of Ecology*, **101** (6): 1400–1408.

Beilstein, M. A., N. S. Nagalingum, M. D. Clements, S. R. Manchester and S. Mathews (2010). Dated molecular phylogenies indicate a Miocene origin for Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America*, **107** (43): 18724–18728.

Belkhadir, Y. and Y. Jaillais (2015). The molecular circuitry of brassinosteroid signaling. *New Phytologist*, **206** (2): 522–540.

Benitez, M.-S. and B. B. M. Gardener (2009). Linking Sequence to Function in Soil Bacteria: Sequence-Directed Isolation of Novel Bacteria Contributing to Soilborne Plant Disease Suppression. *Applied and Environmental Microbiology*, **75** (4): 915–924.

Benjamini, Y. and Y. Hochberg (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57** (1): 289–300.

Benson, A. K., S. A. Kelly, R. Legge, F. Ma, S. J. Low *et al.* (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. *Proceedings of the National Academy of Sciences*, **107** (44): 18933–18938.

Berendsen, R. L., C. M. J. Pieterse and P. A. H. M. Bakker (2012). The rhizosphere microbiome and plant health. *Trends in Plant Science*, **17** (8): 478–486.

Berg, G., D. Rybakova, M. Grube and M. KÃ¶berl (2015). The plant microbiome explored: implications for experimental botany. *Journal of Experimental Botany*, page erv466.

Blaser, M. J. and S. Falkow (2009). What are the consequences of the disappearing human microbiota? *Nature Reviews Microbiology*, **7** (12): 887–894.

Bodenhausen, N., M. Bortfeld-Miller, M. Ackermann and J. A. Vorholt (2014). A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS Genet.*, **10**: e1004283.

Bodenhausen, N., M. W. Horton and J. Bergelson (2013). Bacterial Communities Associated with the Leaves and the Roots of Arabidopsis thaliana. *PLoS ONE*, **8** (2): e56329.

Boisvert, S., F. Raymond, Ã. Godzaridis, F. Laviolette and J. Corbeil (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, **13**: R122.

Bolger, A. M., M. Lohse and B. Usadel (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**: 2114–2120.

Boller, T. and G. Felix (2009). A Renaissance of Elicitors: Perception of Microbe-Associated Molecular Patterns and Danger Signals by Pattern-Recognition Receptors. *Annual Review of Plant Biology*, **60** (1): 379–406.

Bontemps, C., G. N. Elliott, M. F. Simon, F. B. Dos Reis JÃºnior, E. Gross *et al.* (2010). Burkholderia species are ancient symbionts of legumes. *Molecular Ecology*, **19** (1): 44–52.

Botou, M., K. K.i, K. G, K. C, P. K *et al.* (2014). Cellular uptake and utilization of xanthine by the nitrogen-fixing symbiotic rhizobium Sinorhizobium meliloti. In *ResearchGate*.

Bouffaud, M.-L., M. KyselkovÃ¡, B. Gouesnard, G. Grundmann, D. Muller *et al.* (2012). Is diversification history of maize influencing selection of soil bacteria by roots? *Molecular Ecology*, **21** (1): 195–206.

Bouffaud, M.-L., M.-A. Poirier, D. Muller and Y. MoÃ«nne-Loccoz (2014). Root microbiome relates to plant host evolution in maize and other Poaceae. *Environmental Microbiology*, **16** (9): 2804–2814.

Brandl, K., G. Plitas, C. N. Mihu, C. Ubeda, T. Jia *et al.* (2008). Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits. *Nature*, **455** (7214): 804–807.

Brendel, N., L. P. Partida-Martinez, K. Scherlach and C. Hertweck (2007). A cryptic PKS-NRPS gene locus in the plant commensal Pseudomonas fluorescens Pf-5 codes for the biosynthesis of an antimitotic rhizoxin complex. *Organic & Biomolecular Chemistry*, **5** (14): 2211–2213.

Broghammer, A., L. Krusell, M. Blaise, J. Sauer, J. T. Sullivan *et al.* (2012). Legume receptors perceive the rhizobial lipochitin oligosaccharide signal molecules by direct binding. *Proceedings of the National Academy of Sciences of the United States of America*, **109** (34): 13859–13864.

Brulc, J. M., D. A. Antonopoulos, M. E. B. Miller, M. K. Wilson, A. C. Yannarell *et al.* (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. *Proceedings of the National Academy of Sciences*, **106** (6): 1948–1953.

Bulgarelli, D., R. Garrido-Oter, P. Muench, A. Weiman, J. Droege *et al.* (2015). Structure and Function of the Bacterial Root Microbiota in Wild and Domesticated Barley. *Cell Host & Microbe*, **17** (3): 392–403.

Bulgarelli, D., M. Rott, K. Schlaeppi, E. Ver Loren van Themaat, N. Ahmadinejad *et al.* (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature*, **488** (7409): 91–95.

Bulgarelli, D., K. Schlaeppi, S. Spaepen, E. V. L. v. Themaat and P. Schulze-Lefert (2013). Structure and Functions of the Bacterial Microbiota of Plants. *Annual Review of Plant Biology*, **64** (1): 807–838.

Campillos, m., c. v. mering, l. j. jensen and p. bork (2006). identification and analysis of evolutionarily cohesive functional modules in protein networks. *genome research*, **16** (3): 374–382.

Caporaso, J. G., J. Kuczynski, J. Stombaugh, K. Bittinger, F. D. Bushman *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7** (5): 335–336.

Cardinale, M., M. Grube, A. Erlacher, J. Quehenberger and G. Berg (2015). Bacterial networks and co-occurrence relationships in the lettuce root microbiota. *Environmental Microbiology*, **17** (1): 239–252.

Carmody, R. N., G. K. Gerber, J. M. Luevano, D. M. Gatti, L. Somes *et al.* (2015). Diet dominates host genotype in shaping the murine gut microbiota. *Cell Host & Microbe*, **17** (1): 72–84.

Carroll, B. J. and P. M. Gresshoff (1983). Nitrate Inhibition of Nodulation and Nitrogen Fixation in White Clover. *Zeitschrift fÃ¼r Pflanzenphysiologie*, **110** (1): 77–88.

Cascales, E. and P. J. Christie (2003). The versatile bacterial type IV secretion systems. *Nature Reviews Microbiology*, **1** (2): 137–149.

Case, R. J., Y. Boucher, I. DahllÃ¶f, C. HolmstrÃ¶m, W. F. Doolittle *et al.* (2007). Use of 16s rRNA and rpoB genes as molecular markers for microbial ecology studies. *Applied and Environmental Microbiology*, **73** (1): 278–288.

Chaparro, J. M., D. V. Badri and J. M. Vivanco (2014). Rhizosphere microbiome assemblage is affected by plant development. *The ISME Journal*, **8** (4): 790–803.

Charpentier, M., R. Bredemeier, G. Wanner, N. Takeda, E. Schleiff *et al.* (2008). Lotus japonicus CASTOR and POLLUX are ion channels essential for perinuclear calcium spiking in legume root endosymbiosis. *The Plant Cell*, **20** (12): 3467–3479.

Chelius, M. K. and E. W. Triplett (2001). The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. *Microbial Ecology*, **41** (3): 252–263.

Chen, M., X. Li, Q. Yang, X. Chi, L. Pan *et al.* (2014). Dynamic succession of soil bacterial community during continuous cropping of peanut (Arachis hypogaea L.). *PloS One*, **9** (7): e101355.

Chen, P. E. and B. J. Shapiro (2015). The advent of genome-wide association studies for bacteria. *Current Opinion in Microbiology*, **25**: 17–24.

Chi, F., S.-H. Shen, H.-P. Cheng, Y.-X. Jing, Y. G. Yanni *et al.* (2005). Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. *Applied and Environmental Microbiology*, **71** (11): 7271–7278.

Chin, C.-S., D. H. Alexander, P. Marks, A. A. Klammer, J. Drake *et al.* (2013). Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods*, **10** (6): 563–569.

Christie, P. J., N. Whitaker and C. GonzÃ¡lez-Rivera (2014). Mechanism and structure of the bacterial type IV secretion systems. *Biochimica Et Biophysica Acta*, **1843** (8): 1578–1591.

Clauss, M. J. and M. A. Koch (2006). Poorly known relatives of Arabidopsis thaliana. *Trends in Plant Science*, **11** (9): 449–459.

Clemente, J. C., E. C. Pehrsson, M. J. Blaser, K. Sandhu, Z. Gao *et al.* (2015). The microbiome of uncontacted Amerindians. *Science Advances*, **1** (3).

Cohen, O., H. Ashkenazy, D. Burstein and T. Pupko (2012). Uncovering the co-evolutionary network among prokaryotic genes. *Bioinformatics*, **28** (18): i389–i394.

Cole, J. R., Q. Wang, E. Cardenas, J. Fish, B. Chai *et al.* (2009). The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Research*, **37** (Database issue): D141–145.

Colebatch, G., G. Desbrosses, T. Ott, L. Krusell, O. Montanari *et al.* (2004). Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in Lotus japonicus. *The Plant Journal: For Cell and Molecular Biology*, **39** (4): 487–512.

Comadran, J., B. Kilian, J. Russell, L. Ramsay, N. Stein *et al.* (2012). Natural variation in a homolog of Antirrhinum CENTRORADIALIS contributed to spring growth habit and environmental adaptation in cultivated barley. *Nature Genetics*, **44** (12): 1388–1392.

Cordero, O. X., B. Snel and P. Hogeweg (2008). Coevolution of gene families in prokaryotes. *Genome Research*, **18** (3): 462–468.

Cornelis, G. R. and F. Van Gijsegem (2000). Assembly and function of type III secretory systems. *Annual Review of Microbiology*, **54**: 735–774.

Cotillard, A., S. P. Kennedy, L. C. Kong, E. Prifti, N. Pons *et al.* (2013). Dietary intervention impact on gut microbial gene richness. *Nature*, **500** (7464): 585–588.

Coyte, K. Z., J. Schluter and K. R. Foster (2015). The ecology of the microbiome: Networks, competition, and stability. *Science (New York, N.Y.)*, **350** (6261): 663–666.

Csardi, G. and T. Nepusz (2006). The igraph software package for complex network research. *InterJournal*, **Complex Systems**: 1695.

Curtis, T. P., W. T. Sloan and J. W. Scannell (2002). Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences*, **99** (16): 10494–10499.

Dakora, F. and D. Phillips (1996). Diverse functions of isoflavonoids in legumes transcend anti-microbial definitions of phytoalexins. *Physiological and Molecular Plant Pathology*, **49** (1): 1–20.

Dakora, F. D. and D. A. Phillips (2002). Root exudates as mediators of mineral acquisition in low-nutrient environments. *Plant and Soil*, **245** (1): 35–47.

David, L. A., A. C. Materna, J. Friedman, M. I. Campos-Baptista, M. C. Blackburn *et al.* (2014a). Host lifestyle affects human microbiota on daily timescales. *Genome Biology*, **15**: R89.

David, L. A., C. F. Maurice, R. N. Carmody, D. B. Gootenberg, J. E. Button *et al.* (2014b). Diet rapidly and reproducibly alters the human gut microbiome. *Nature*, **505** (7484): 559–563.

de Juan, D., F. Pazos and A. Valencia (2013). Emerging methods in protein co-evolution. *Nature Reviews. Genetics*, **14** (4): 249–261.

Delaux, P.-M., K. Varala, P. P. Edger, G. M. Coruzzi, J. C. Pires *et al.* (2014). Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS genetics*, **10** (7): e1004487.

Delcher, A. L., D. Harmon, S. Kasif, O. White and S. L. Salzberg (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.*, **27**: 4636–4641.

Delmotte, N. (2009). Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl Acad. Sci. USA*, **106**: 16428–16433.

Demoling, F., D. Figueroa and E. Baath (2007). Comparison of factors limiting bacterial growth in different soils. *Soil Biology and Biochemistry*, **39** (10): 2485–2495.

DeSantis, T. Z., P. Hugenholtz, N. Larsen, M. Rojas, E. L. Brodie *et al.* (2006). Greengenes, a chimera-checked 16s rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72** (7): 5069–5072.

Dessein, R., M. Gironella, C. Vignal, L. Peyrin-Biroulet, H. Sokol *et al.* (2009). Toll-like receptor 2 is critical for induction of Reg3 beta expression and intestinal clearance of Yersinia pseudotuberculosis. *Gut*, **58** (6): 771–776.

Ding, H., C. B. Yip, B. A. Geddes, I. J. Oresnik and M. F. Hynes (2012). Glycerol utilization by Rhizobium leguminosarum requires an ABC transporter and affects competition for nodulation. *Microbiology (Reading, England)*, **158** (Pt 5): 1369–1378.

Dixon, R. and D. Kahn (2004). Genetic regulation of biological nitrogen fixation. *Nature Reviews Microbiology*, **2** (8): 621–631.

Dombrowski, N., K. Schlaeppi, M. T. Agler, S. Hacquard, E. Kemen *et al.* (2016). Root microbiota dynamics of perennial Arabis alpina are dependent on soil residence time but independent of flowering time. *The ISME Journal*.

Dominguez-Bello, M. G., E. K. Costello, M. Contreras, M. Magris, G. Hidalgo *et al.* (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. *Proceedings of the National Academy of Sciences*, **107** (26): 11971–11975.

Donn, S., J. A. Kirkegaard, G. Perera, A. E. Richardson and M. Watt (2015). Evolution of bacterial communities in the wheat crop rhizosphere. *Environmental Microbiology*, **17** (3): 610–621.

Doornbos, R. F., B. P. J. Geraats, E. E. Kuramae, L. C. Van Loon and P. A. H. M. Bakker (2010). Effects of Jasmonic Acid, Ethylene, and Salicylic Acid Signaling on the Rhizosphere Bacterial Community of Arabidopsis thaliana. *Molecular Plant-Microbe Interactions*, **24** (4): 395–407.

Droege, J., I. Gregor and A. C. McHardy (2015). Taxator-tk: precise taxonomic assignment of metagenomes by fast approximation of evolutionary neighborhoods. *Bioinformatics (Oxford, England)*, **31** (6): 817–824.

Duncan, S. H., P. Louis, J. M. Thomson and H. J. Flint (2009). The role of pH in determining the species composition of the human colonic microbiota. *Environmental Microbiology*, **11** (8): 2112–2122.

Durocher, D. and S. P. Jackson (2002). The FHA domain. *FEBS letters*, **513** (1): 58–66.

Earle, S. G., C.-H. Wu, J. Charlesworth, N. Stoesser, N. C. Gordon *et al.* (2016). Identifying lineage effects when controlling for population structure improves power in bacterial association studies. *Nature Microbiology*, **1**: 16041.

Eckburg, P. B., E. M. Bik, C. N. Bernstein, E. Purdom, L. Dethlefsen *et al.* (2005). Diversity of the Human Intestinal Microbial Flora. *Science*, **308** (5728): 1635–1638.

Eddy, S. R. (2011). Accelerated profile HMM searches. *PLOS Comput. Biol.*, **7**: e1002195.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics (Oxford, England)*, **26** (19): 2460–2461.

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

Edgar, R. C., B. J. Haas, J. C. Clemente, C. Quince and R. Knight (2011). UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics (Oxford, England)*, **27** (16): 2194–2200.

Edwards, J., C. Johnson, C. Santos-MedellÃn, E. Lurie, N. K. Podishetty *et al.* (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. *Proceedings of the National Academy of Sciences*, **112** (8): E911–E920.

Edwards, R. A., R. Olson, T. Disz, G. D. Pusch, V. Vonstein *et al.* (2012). Real time metagenomics: using k-mers to annotate metagenomes. *Bioinformatics (Oxford, England)*, **28** (24): 3316–3317.

Egener, T., T. Hurek and B. Reinhold-Hurek (1998). Use of green fluorescent protein to detect expression of nif genes of Azoarcus sp. BH72, a grass-associated diazotroph, on rice roots. *Molecular plant-microbe interactions: MPMI*, **11** (1): 71–75.

Eilam, O., R. Zarecki, M. Oberhardt, L. K. Ursell, M. Kupiec *et al.* (2014). Glycan degradation (GlyDeR) analysis predicts mammalian gut microbiota abundance and host diet-specific adaptations. *mBio*, **5** (4).

Eilers, K. G., C. L. Lauber, R. Knight and N. Fierer (2010). Shifts in bacterial community structure associated with inputs of low molecular weight carbon compounds to soil. *Soil Biology and Biochemistry*, **42** (6): 896–903.

Eisen, M. B., P. T. Spellman, P. O. Brown and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences of the United States of America*, **95** (25): 14863–14868.

El Yahyaoui, F., H. KÃ$\frac{1}{4}$ster, B. Ben Amor, N. Hohnjec, A. PÃ$\frac{1}{4}$hler *et al.* (2004). Expression profiling in Medicago truncatula identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. *Plant Physiology*, **136** (2): 3159–3176.

Emms, D. M. and S. Kelly (2015). OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, **16**: 157.

Epskamp, S., A. O. J. Cramer, L. J. Waldorp, V. D. Schmittmann and D. Borsboom (2012). qgraph: Network visualizations of relationships in psychometric data. *Journal of Statistical Software*, **48** (4): 1–18.

Ettema, T., J. v. d. Oost and M. Huynen (2001). Modularity in the gain and loss of genes: applications for function prediction. *Trends in Genetics*, **17** (9): 485–487.

Faith, D. P. (1992). Conservation evaluation and phylogenetic diversity. *Biological Conservation*, **61** (1): 1–10.

Faith, J. J., P. P. Ahern, V. K. Ridaura, J. Cheng and J. I. Gordon (2014). Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. *Science Translational Medicine*, **6** (220): 220ra11.

Faith, J. J., J. L. Guruge, M. Charbonneau, S. Subramanian, H. Seedorf *et al.* (2013). The Long-Term Stability of the Human Gut Microbiota. *Science*, **341** (6141): 1237439.

Falush, D. and R. Bowden (2006). Genome-wide association mapping in bacteria? *Trends in Microbiology*, **14** (8): 353–355.

Faure, D., D. Vereecke and J. J. Leveau (2009). Molecular communication in the rhizosphere. *Plant Soil*, **321**: 279–303.

Faust, K., J. F. Sathirapongsasuti, J. Izard, N. Segata, D. Gevers *et al.* (2012). Microbial Co-occurrence Relationships in the Human Microbiome. *PLOS Comput Biol*, **8** (7): e1002606.

Felsenstein, J. and J. Felenstein (2004). *Inferring phylogenies*, volume 2. Sinauer Associates Sunderland.

Fierer, N. and R. B. Jackson (2006). The diversity and biogeography of soil bacterial communities. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (3): 626–631.

Finn, R. D., P. Coggill, R. Y. Eberhardt, S. R. Eddy, J. Mistry *et al.* (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, **44** (D1): D279–285.

Flemetakis, E., M. Dimou, D. Cotzur, R. C. Efrose, G. Aivalakis *et al.* (2003). A sucrose transporter, LjSUT4, is up-regulated during Lotus japonicus nodule development. *Journal of Experimental Botany*, **54** (388): 1789–1791.

Franzenburg, S., J. Walter, S. KÃ$\frac{1}{4}$nzel, J. Wang, J. F. Baines *et al.* (2013). Distinct antimicrobial peptide expression determines host species-specific bacterial associations. *Proceedings of the National Academy of Sciences of the United States of America*, **110** (39): E3730–3738.

Fraune, S., F. Anton-Erxleben, R. Augustin, S. Franzenburg, M. Knop *et al.* (2015). Bacteria-bacteria interactions within the microbiota of the ancestral metazoan Hydra contribute to fungal resistance. *The ISME journal*, **9** (7): 1543–1556.

Frey, B. J. and D. Dueck (2007). Clustering by Passing Messages Between Data Points. *Science*, **315** (5814): 972–976.

Galardini, M., A. Mengoni, M. Brilli, F. Pini, A. Fioravanti *et al.* (2011). Exploring the symbiotic pangenome of the nitrogen-fixing bacterium Sinorhizobium meliloti. *BMC Genomics*, **12** (1).

Galardini, M., F. Pini, M. Bazzicalupo, E. G. Biondi and A. Mengoni (2013). Replicon-Dependent Bacterial Genome Evolution: The Case of Sinorhizobium meliloti. *Genome Biology and Evolution*, **5** (3): 542–558.

Galibert, F. (2001). The Composite Genome of the Legume Symbiont Sinorhizobium meliloti. *Science*, **293** (5530): 668–672.

Garau, G., R. J. Yates, P. Deiana and J. G. Howieson (2009). Novel strains of nodulating Burkholderia have a role in nitrogen fixation with papilionoid herbaceous legumes adapted to acid, infertile soils. *Soil Biology and Biochemistry*, **41** (1): 125–134.

Garcia-Garrido, J. M. and J. A. Ocampo (2002). Regulation of the plant defence response in arbuscular mycorrhizal symbiosis. *Journal of Experimental Botany*, **53** (373): 1377–1386.

Garrido-Oter, R. and A. C. McHardy (2016). Clustering of functionally related genes reveals novel symbiosis-relevant genes in rhizobia. (In preparation).

Garrido-Oter, R., T. Nakano, N. Dombrowski, A. C. McHardy and P. Schulze-Lefert (2016). Assessment of functional diversification and adaptation in rhizobia by comparative genomics. (In preparation).

Gaudier, E., A. Jarry, H. M. BlottiÃ¨re, P. d. Coppet, M. P. Buisine *et al.* (2004). Butyrate specifically modulates MUC gene expression in intestinal epithelial goblet cells deprived of glucose. *American Journal of Physiology - Gastrointestinal and Liver Physiology*, **287** (6): G1168–G1174.

Gill, S. R., M. Pop, R. T. DeBoy, P. B. Eckburg, P. J. Turnbaugh *et al.* (2006). Metagenomic Analysis of the Human Distal Gut Microbiome. *Science*, **312** (5778): 1355–1359.

Goodman, A. L., G. Kallstrom, J. J. Faith, A. Reyes, A. Moore *et al.* (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proceedings of the National Academy of Sciences of the United States of America*, **108** (15): 6252–6257.

Goodrich, J. K., J. L. Waters, A. C. Poole, J. L. Sutter, O. Koren *et al.* (2014). Human genetics shape the gut microbiome. *Cell*, **159** (4): 789–799.

Gordon, N. C., J. R. Price, K. Cole, R. Everitt, M. Morgan *et al.* (2014). Prediction of Staphylococcus aureus Antimicrobial Resistance by Whole-Genome Sequencing. *Journal of Clinical Microbiology*, **52** (4): 1182–1191.

Gregor, I., J. DrÃ¶ge, M. Schirmer, C. Quince and A. C. McHardy (2016). PhyloPythiaS+: a self-training method for the rapid reconstruction of low-ranking taxonomic bins from metagenomes. *PeerJ*, **4**.

Guttman, D. S., S. J. Gropp, R. L. Morgan and P. W. Wang (2006). Diversifying selection drives the evolution of the type III secretion system pilus of Pseudomonas syringae. *Molecular Biology and Evolution*, **23** (12): 2342–2354.

Guttman, D. S., A. C. McHardy and P. Schulze-Lefert (2014). Microbial genome-enabled insights into plant-microorganism interactions. *Nature Reviews Genetics*, **15** (12): 797–813.

Hacquard, S., R. Garrido-Oter, A. GonzÃ¡lez, S. Spaepen, G. Ackermann *et al.* (2015). Microbiota and Host Nutrition across Plant and Animal Kingdoms. *Cell Host & Microbe*, **17** (5): 603–616.

Hacquard, S., B. Kracher, K. Hiruma, P. C. Muench, R. Garrido-Oter *et al.* (2016). Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nature Communications*, **7**: 11362.

Hacquard, S. and C. W. Schadt (2015). Towards a holistic understanding of the beneficial interactions across the Populus microbiome. *New Phytologist*, **205** (4): 1424–1430.

Haft, D. H., J. D. Selengut, R. A. Richter, D. Harkins, M. K. Basu *et al.* (2013). TIGRFAMs and Genome Properties in 2013. *Nucleic Acids Research*, **41** (Database issue): D387–395.

Handberg, K. and J. Stougaard (1992). Lotus japonicus, an autogamous, diploid legume species for classical and molecular genetics. *The Plant Journal*, **2** (4): 487–496.

Hanson, B. T., J. M. Yagi, C. O. Jeon and E. M. Madsen (2012). Role of nitrogen fixation in the autecology of Polaromonas naphthalenivorans in contaminated sediments. *Environmental Microbiology*, **14** (6): 1544–1557.

Hardoim, P. R., F. D. Andreote, B. Reinhold-Hurek, A. Sessitsch, L. S. van Overbeek *et al.* (2011). Rice root-associated bacteria: insights into community structures across 10 cultivars: Insights into the bacterial community structure of rice cultivars. *FEMS Microbiology Ecology*, **77** (1): 154–164.

Harrison, P. W., R. P. Lower, N. K. Kim and J. P. W. Young (2010). Introducing the bacterial chromid: not a chromosome, not a plasmid. *Trends in Microbiology*, **18** (4): 141–148.

Hartwig, U. A., C. M. Joseph and D. A. Phillips (1991). Flavonoids Released Naturally from Alfalfa Seeds Enhance Growth Rate of Rhizobium meliloti. *Plant Physiology*, **95** (3): 797–803.

Heger, A. and L. Holm (2000). Rapid automatic detection and alignment of repeats in protein sequences. *Proteins*, **41** (2): 224–237.

Heijden, M. G. A. v. d. and M. Hartmann (2016). Networking in the Plant Microbiome. *PLOS Biol*, **14** (2): e1002378.

Held, M., H. Hou, M. Miri, C. Huynh, L. Ross *et al.* (2014). Lotus japonicus cytokinin receptors work partially redundantly to mediate nodule formation. *The Plant Cell*, **26** (2): 678–694.

Hershey, D. M., X. Lu, J. Zi and R. J. Peters (2014). Functional conservation of the capacity for ent-kaurene biosynthesis and an associated operon in certain rhizobia. *Journal of Bacteriology*, **196** (1): 100–106.

Hinsinger, P., C. Plassard, C. Tang and B. Jaillard (2003). Origins of root-mediated pH changes in the rhizosphere and their responses to environmental constraints: A review. *Plant and Soil*, **248** (1-2): 43–59.

Hoffmann, M. H. (2005). EVOLUTION OF THE REALIZED CLIMATIC NICHE IN THE GENUS ARABIDOPSIS (BRASSICACEAE). *Evolution*, **59** (7): 1425.

Hogslund, N., S. Radutoiu, L. Krusell, V. Voroshilova, M. A. Hannah *et al.* (2009). Dissection of symbiosis and organ development by integrated transcriptome analysis of lotus japonicus mutant and wild-type plants. *PloS One*, **4** (8): e6556.

Hong, S., J. Bunge, C. Leslin, S. Jeon and S. S. Epstein (2009). Polymerase chain reaction primers miss half of rRNA microbial diversity. *The ISME journal*, **3** (12): 1365–1373.

Horton, M. W. (2014). Genome-wide association study of Arabidopsis thaliana leaf microbial community. *Nat. Commun.*, **5**: 5320.

Hoskins, L. C. and E. T. Boulding (1981). Mucin Degradation in Human Colon Ecosystems. *Journal of Clinical Investigation*, **67** (1): 163–172.

Huelsenbeck, J. P. and F. Ronquist (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, **17** (8): 754–755.

Huson, D. H., A. F. Auch, J. Qi and S. C. Schuster (2007). MEGAN analysis of metagenomic data. *Genome Research*, **17** (3): 377–386.

Huson, D. H., S. Beier, I. Flade, A. Gorska, M. El-Hadidi *et al.* (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, **12** (6).

Huttenhower, C., D. Gevers, R. Knight and T. H. M. P. Consortium (2012). Structure, function and diversity of the healthy human microbiome. *Nature*, **486** (7402): 207–214.

Hwang, J. H., S. R. Ellingson and D. M. Roberts (2010). Ammonia permeability of the soybean nodulin 26 channel. *FEBS letters*, **584** (20): 4339–4343.

Hyatt, D., G.-L. Chen, P. F. LoCascio, M. L. Land, F. W. Larimer *et al.* (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**: 119.

Inceoglu, O., W. A. Al-Soud, J. F. Salles, A. V. Semenov and J. D. van Elsas (2011). Comparative Analysis of Bacterial Communities in a Potato Field as Determined by Pyrosequencing. *PLoS ONE*, **6** (8): e23321.

Iwasaki, A. and R. Medzhitov (2010). Regulation of Adaptive Immunity by the Innate Immune System. *Science*, **327** (5963): 291–295.

Janzen, D. (1985). The natural history of mutualisms. *The biology of mutualism: Ecology and evolution*, pages 40–99.

Jeong, J. and M. L. Guerinot (2009). Homing in on iron homeostasis in plants. *Trends in Plant Science*, **14** (5): 280–285.

Johansen, J. E., P. Nielsen and S. J. Binnerup (2009). Identification and potential enzyme capacity of flavobacteria isolated from the rhizosphere of barley (*Hordeum vulgare* L.). *Canadian Journal of Microbiology*, **55** (3): 234–241.

Johansson, M. E. V., J. K. Gustafsson, K. E. SjÃ¶berg, J. Petersson, L. Holm *et al.* (2010). Bacteria Penetrate the Inner Mucus Layer before Inflammation in the Dextran Sulfate Colitis Model. *PLOS ONE*, **5** (8): e12238.

Johansson, M. E. V., M. Phillipson, J. Petersson, A. Velcich, L. Holm *et al.* (2008). The inner of the two Muc2 mucin-dependent mucus layers in colon is devoid of bacteria. *Proceedings of the National Academy of Sciences*, **105** (39): 15064–15069.

Jones, D. L., C. Nguyen and R. D. Finlay (2009). Carbon flow in the rhizosphere: carbon trading at the soil-root interface. *Plant and Soil*, **321** (1-2): 5–33.

Jones, J. D. G. and J. L. Dangl (2006). The plant immune system. *Nature*, **444** (7117): 323–329.

Kamada, N., G. Y. Chen, N. Inohara and G. Nunez (2013). Control of pathogens and pathobionts by the gut microbiota. *Nature Immunology*, **14** (7): 685–690.

Kanamori, N., L. H. Madsen, S. Radutoiu, M. Frantescu, E. M. H. Quistgaard *et al.* (2006). A nucleoporin is required for induction of Ca2+ spiking in legume nodule development and essential for rhizobial and fungal symbiosis. *Proceedings of the National Academy of Sciences of the United States of America*, **103** (2): 359–364.

Kanehisa, M. (2014). Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**: D199–D205.

Kanehisa, M. and S. Goto (2000). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**: 27–30.

Kanehisa, M., Y. Sato, M. Kawashima, M. Furumichi and M. Tanabe (2016). KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Research*, **44** (Database issue): D457–D462.

Karasov, T. L., J. M. Kniskern, L. Gao, B. J. DeYoung, J. Ding *et al.* (2014). The long-term maintenance of a resistance polymorphism through diffuse interactions. *Nature*, **512** (7515): 436–440.

Kemen, E. (2014). Microbe-microbe interactions determine oomycete and fungal host colonization. *Current Opinion in Plant Biology*, **20**: 75–81.

Khachatryan, Z. A., Z. A. Ktsoyan, G. P. Manukyan, D. Kelly, K. A. Ghazaryan *et al.* (2008). Predominant Role of Host Genetics in Controlling the Composition of Gut Microbiota. *PLOS ONE*, **3** (8): e3064.

Khan, M. T., S. H. Duncan, A. J. M. Stams, J. M. van Dijl, H. J. Flint *et al.* (2012). The gut anaerobe Faecalibacterium prausnitzii uses an extracellular electron shuttle to grow at oxic-anoxic interphases. *The ISME Journal*, **6** (8): 1578–1585.

Kislev, M. E., D. Nadel and I. Carmi (1992). Festschrift For Professor Van Zeist Epipalaeolithic (19,000 BP) cereal and fruit diet at Ohalo II, Sea of Galilee, Israel. *Review of Palaeobotany and Palynology*, **73** (1): 161–166.

Klindworth, A., E. Pruesse, T. Schweer, J. Peplies, C. Quast *et al.* (2013). Evaluation of general 16s ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41** (1): e1.

Knief, C., N. Delmotte, S. Chaffron, M. Stark, G. Innerebner *et al.* (2012). Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *The ISME Journal*, **6** (7): 1378–1390.

Koenig, J. E., A. Spor, N. Scalfone, A. D. Fricker, J. Stombaugh *et al.* (2011). Succession of microbial consortia in the developing infant gut microbiome. *Proceedings of the National Academy of Sciences*, **108** (Supplement 1): 4578–4585.

Koljalg, U., K.-H. Larsson, K. Abarenkov, R. H. Nilsson, I. J. Alexander *et al.* (2005). UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi. *The New Phytologist*, **166** (3): 1063–1068.

Kolton, M., S. J. Green, Y. M. Harel, N. Sela, Y. Elad *et al.* (2012). Draft Genome Sequence of Flavobacterium sp. Strain F52, Isolated from the Rhizosphere of Bell Pepper (Capsicum annuum L. cv. Maccabi). *Journal of Bacteriology*, **194** (19): 5462–5463.

Koprivova, A., A. L. Harper, M. Trick, I. Bancroft and S. Kopriva (2014). Dissection of the control of anion homeostasis by associative transcriptomics in Brassica napus. *Plant Physiology*, **166** (1): 442–450.

Kopylova, E., L. NoÃ© and H. Touzet (2012). SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*, **28** (24): 3211–3217.

Kraemer, U. (2010). Metal Hyperaccumulation in Plants. *Annual Review of Plant Biology*, **61** (1): 517–534.

Krusell, L., L. H. Madsen, S. Sato, G. Aubert, A. Genua *et al.* (2002). Shoot control of root development and nodulation is mediated by a receptor-like kinase. *Nature*, **420** (6914): 422–426.

Kumar, S., A. J. Filipski, F. U. Battistuzzi, S. L. K. Pond and K. Tamura (2012). Statistics and Truth in Phylogenomics. *Molecular Biology and Evolution*, **29** (2): 457–472.

Kuo, C.-H. and H. Ochman (2009). Deletional bias across the three domains of life. *Genome Biology and Evolution*, **1**: 145–152.

Lagesen, K., P. Hallin, E. A. Rodland, H.-H. Staerfeldt, T. Rognes *et al.* (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, **35** (9): 3100–3108.

Lambers, H., J. A. Raven, G. R. Shaver and S. E. Smith (2008). Plant nutrient-acquisition strategies change with soil age. *Trends in Ecology & Evolution*, **23** (2): 95–103.

Lebeis, S. L., S. H. Paredes, D. S. Lundberg, N. Breakfield, J. Gehring *et al.* (2015). PLANT MICROBIOME. Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science (New York, N.Y.)*, **349** (6250): 860–864.

Lebeis, S. L., M. Rott, J. L. Dangl and P. Schulze-Lefert (2012). Culturing a plant microbiome community at the cross-Rhodes. *New Phytol.*, **196**: 341–344.

Leitch, E. C. M., A. W. Walker, S. H. Duncan, G. Holtrop and H. J. Flint (2007). Selective colonization of insoluble substrates by human faecal bacteria. *Environmental Microbiology*, **9** (3): 667–679.

Lerouge, P., P. Roche, C. Faucher, F. Maillet, G. Truchet *et al.* (1990). Symbiotic host-specificity of Rhizobium meliloti is determined by a sulphated and acylated glucosamine oligosaccharide signal. *Nature*, **344** (6268): 781–784.

Lessl, M. and E. Lanka (1994). Common mechanisms in bacterial conjugation and Ti-mediated T-DNA transfer to plant cells. *Cell*, **77** (3): 321–324.

Ley, R. E., M. Hamady, C. Lozupone, P. J. Turnbaugh, R. R. Ramey *et al.* (2008a). Evolution of Mammals and Their Gut Microbes. *Science*, **320** (5883): 1647–1651.

Ley, R. E., C. A. Lozupone, M. Hamady, R. Knight and J. I. Gordon (2008b). Worlds within worlds: evolution of the vertebrate gut microbiota. *Nature Reviews Microbiology*, **6** (10): 776–788.

Li, L., C. J. Stoeckert and D. S. Roos (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, **13** (9): 2178–2189.

Li, R. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**: 265–272.

Lima-Mendez, G., K. Faust, N. Henry, J. Decelle, S. Colin *et al.* (2015). Ocean plankton. Determinants of community structure in the global plankton interactome. *Science (New York, N.Y.)*, **348** (6237): 1262073.

Lopez-Arredondo, D. L., M. A. Leyva-Gonzalez, S. I. Gonzalez-Morales, J. Lopez-Bucio and L. Herrera-Estrella (2014). Phosphate nutrition: improving low-phosphate tolerance in crops. *Annual Review of Plant Biology*, **65**: 95–123.

Lozupone, C. and R. Knight (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71** (12): 8228–8235.

Lozupone, C., M. E. Lladser, D. Knights, J. Stombaugh and R. Knight (2011). UniFrac: an effective distance metric for microbial community comparison. *The ISME journal*, **5** (2): 169–172.

Lozupone, C. A., J. I. Stombaugh, J. I. Gordon, J. K. Jansson and R. Knight (2012). Diversity, stability and resilience of the human gut microbiota. *Nature*, **489** (7415): 220–230.

Lugtenberg, B. and F. Kamilova (2009). Plant-growth-promoting rhizobacteria. *Annual Review of Microbiology*, **63**: 541–556.

Lundberg, D. S., S. L. Lebeis, S. H. Paredes, S. Yourstone, J. Gehring *et al.* (2012). Defining the core Arabidopsis thaliana root microbiome. *Nature*, **488** (7409): 86–90.

Luo, R., B. Liu, Y. Xie, Z. Li, W. Huang *et al.* (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*, **1**: 18.

Maathuis, F. J. (2009). Physiological functions of mineral macronutrients. *Current Opinion in Plant Biology*, **12** (3): 250–258.

Madsen, E. B., L. H. Madsen, S. Radutoiu, M. Olbryt, M. Rakwalska *et al.* (2003). A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. *Nature*, **425** (6958): 637–640.

Madsen, L. H., L. Tirichine, A. Jurkiewicz, J. T. Sullivan, A. B. Heckmann *et al.* (2010). The molecular network governing nodule organogenesis and infection in the model legume Lotus japonicus. *Nature Communications*, **1**: 10.

Maekawa, T., T. A. Kufer and P. Schulze-Lefert (2011). NLR functions in plant and animal immune systems: so far and yet so close. *Nature Immunology*, **12** (9): 817–826.

Magallon, S., S. Gomez-Acevedo, L. L. Sanchez-Reyes and T. Hernandez-Hernandez (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *The New Phytologist*, **207** (2): 437–453.

Maignien, L., E. A. DeForce, M. E. Chafee, A. M. Eren and S. L. Simmons (2014). Ecological succession and stochastic variation in the assembly of Arabidopsis thaliana phyllosphere communities. *MBio*, **5**: e00682–e13.

Manter, D. K., J. A. Delgado, D. G. Holm and R. A. Stong (2010). Pyrosequencing Reveals a Highly Diverse and Cultivar-Specific Bacterial Endophyte Community in Potato Roots. *Microbial Ecology*, **60** (1): 157–166.

Martins dos Santos, V., M. MÃ$\frac{1}{4}$ller and W. M. de Vos (2010). Systems biology of the gut: the interplay of food, microbiota and host at the mucosal interface. *Current Opinion in Biotechnology*, **21** (4): 539–550.

Masson-Boivin, C., E. Giraud, X. Perret and J. Batut (2009). Establishing nitrogen-fixing symbiosis with legumes: how many rhizobium recipes? *Trends in Microbiology*, **17** (10): 458–466.

McCann, H. C., H. Nahal, S. Thakur and D. S. Guttman (2012). Identification of innate immunity elicitors using molecular signatures of natural selection. *Proceedings of the National Academy of Sciences of the United States of America*, **109** (11): 4215–4220.

McCully, M. E. (1999). ROOTS IN SOIL: Unearthing the Complexities of Roots and Their Rhizospheres. *Annual Review of Plant Physiology and Plant Molecular Biology*, **50**: 695–718.

McDonald, D., M. N. Price, J. Goodrich, E. P. Nawrocki, T. Z. DeSantis *et al.* (2012). An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *The ISME Journal*, **6** (3): 610–618.

McFall-Ngai, M., M. G. Hadfield, T. C. G. Bosch, H. V. Carey, T. Domazet-LoÅ¡o *et al.* (2013). Animals in a bacterial world, a new imperative for the life sciences. *Proceedings of the National Academy of Sciences of the United States of America*, **110** (9): 3229–3236.

McKnite, A. M., M. E. Perez-Munoz, L. Lu, E. G. Williams, S. Brewer *et al.* (2012). Murine Gut Microbiota Is Defined by Host Genetics and Modulates Variation of Metabolic Traits. *PLOS ONE*, **7** (6): e39191.

McMurdie, P. J. and S. Holmes (2013). phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS ONE*, **8** (4).

McMurdie, P. J. and S. Holmes (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS computational biology*, **10** (4): e1003531.

Mendes, L. W., E. E. Kuramae, A. A. Navarrete, J. A. van Veen and S. M. Tsai (2014). Taxonomical and functional microbial community selection in soybean rhizosphere. *The ISME Journal*, **8** (8): 1577–1587.

Mendes, R., M. Kruijt, I. d. Bruijn, E. Dekkers, M. v. d. Voort *et al.* (2011). Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. *Science*, **332** (6033): 1097–1100.

Meyer, R. S., A. E. DuVal and H. R. Jensen (2012). Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. *The New Phytologist*, **196** (1): 29–48.

Mitsuoka, T. (1992). Intestinal flora and aging. *Nutrition Reviews*, **50** (12): 438–446.

Moeller, A. H., Y. Li, E. M. Ngole, S. Ahuka-Mundeke, E. V. Lonsdorf *et al.* (2014). Rapid changes in the gut microbiome during human evolution. *Proceedings of the National Academy of Sciences*, **111** (46): 16431–16435.

Morris, R. L. and T. M. Schmidt (2013). Shallow breathing: bacterial life at low O2. *Nature Reviews Microbiology*, **11** (3): 205–212.

Muegge, B. D., J. Kuczynski, D. Knights, J. C. Clemente, A. GonzÃ¡lez *et al.* (2011). Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans. *Science*, **332** (6032): 970–974.

Murray, J. D., B. J. Karas, S. Sato, S. Tabata, L. Amyot *et al.* (2007). A cytokinin perception mutant colonized by Rhizobium in the absence of nodule organogenesis. *Science (New York, N.Y.)*, **315** (5808): 101–104.

Nakagawa, T., H. Kaku, Y. Shimoda, A. Sugiyama, M. Shimamura *et al.* (2011). From defense to symbiosis: limited alterations in the kinase domain of LysM receptor-like kinases are crucial for evolution of legume-Rhizobium symbiosis. *The Plant Journal: For Cell and Molecular Biology*, **65** (2): 169–180.

Namiki, T., T. Hachiya, H. Tanaka and Y. Sakakibara (2012). MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, **40** (20): e155.

Newburg, D. S. and L. Morelli (2015). Human milk and infant intestinal mucosal glycans guide succession of the neonatal intestinal microbiota. *Pediatric Research*, **77** (1-2): 115–120.

Newton, A. C., A. J. Flavell, T. S. George, P. Leat, B. Mullholland *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** (2): 141–178.

Nguyen, C. (2003). Rhizodeposition of organic C by plants: mechanisms and controls. *Agronomy for Sustainable Development*, **23** (5-6): 22.

Noll, M., D. Matthies, P. Frenzel, M. Derakshani and W. Liesack (2005). Succession of bacterial community structure and diversity in a paddy soil oxygen gradient. *Environmental Microbiology*, **7** (3): 382–395.

Ochman, H., M. Worobey, C.-H. Kuo, J.-B. N. Ndjango, M. Peeters *et al.* (2010). Evolutionary Relationships of Wild Hominids Recapitulated by Gut Microbial Communities. *PLOS Biol*, **8** (11): e1000546.

Ofek, M., M. Voronov-Goldman, Y. Hadar and D. Minz (2014). Host signature effect on plant root-associated microbiomes revealed through analyses of resident vs. active communities. *Environmental Microbiology*, **16** (7): 2157–2167.

Ofek-Lalzar, M., N. Sela, M. Goldman-Voronov, S. J. Green, Y. Hadar *et al.* (2014). Niche and host-associated functional signatures of the root surface microbiome. *Nature Communications*, **5**: 4950.

Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin *et al.* (2015). *vegan: Community Ecology Package*. R package version 2.3-2.

Oldroyd, G. E. D. (2013). Speak, friend, and enter: signalling systems that promote beneficial symbiotic associations in plants. *Nature Reviews. Microbiology*, **11** (4): 252–263.

Oldroyd, G. E. D., J. D. Murray, P. S. Poole and J. A. Downie (2011). The Rules of Engagement in the Legume-Rhizobial Symbiosis. *Annual Review of Genetics*, **45** (1): 119–144.

O'Neill, A. J., F. McLaws, G. Kahlmeter, A. S. Henriksen and I. Chopra (2007). Genetic basis of resistance to fusidic acid in staphylococci. *Antimicrobial Agents and Chemotherapy*, **51** (5): 1737–1740.

Osawa, R., W. Blanshard and P. Ocallaghan (1993). Microbiological Studies of the Intestinal Microflora of the Koala, Phascolarctos-Cinereus .2. Pap, a Special Maternal Feces Consumed by Juvenile Koalas. *Australian Journal of Zoology*, **41** (6): 611–620.

Overbeek, R. (2005). The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res.*, **33**: 5691–5702.

Pagel, M. (1994). Detecting Correlated Evolution on Phylogenies: A General Method for the Comparative Analysis of Discrete Characters. *Proceedings of the Royal Society of London B: Biological Sciences*, **255** (1342): 37–45.

Palmer, C., E. M. Bik, D. B. DiGiulio, D. A. Relman and P. O. Brown (2007). Development of the Human Infant Intestinal Microbiota. *PLOS Biol*, **5** (7): e177.

Panke-Buisse, K., A. C. Poole, J. K. Goodrich, R. E. Ley and J. Kao-Kniffin (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. *The ISME Journal*, **9** (4): 980–989.

Paradis, E., J. Claude and K. Strimmer (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics*, **20** (2): 289–290.

Parks, B. W., E. Nam, E. Org, E. Kostem, F. Norheim *et al.* (2013). Genetic control of obesity and gut microbiota composition in response to high-fat, high-sucrose diet in mice. *Cell Metabolism*, **17** (1): 141–152.

Parniske, M. (2008). Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nature Reviews. Microbiology*, **6** (10): 763–775.

Partida-Martinez, L. P. and C. Hertweck (2005). Pathogenic fungus harbours endosymbiotic bacteria for toxin production. *Nature*, **437** (7060): 884–888.

Patriquin, D. G., J. DÃ¶bereiner and D. K. Jain (1983). Sites and processes of association between diazotrophs and grasses. *Canadian Journal of Microbiology*, **29** (8): 900–915.

Pedron, T., C. Mulet, C. Dauga, L. Frangeul, C. Chervaux *et al.* (2012). A crypt-specific core microbiota resides in the mouse colon. *mBio*, **3** (3).

Peiffer, J. A., A. Spor, O. Koren, Z. Jin, S. G. Tringe *et al.* (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proceedings of the National Academy of Sciences of the United States of America*, **110** (16): 6548–6553.

Peng, Y., H. C. M. Leung, S. M. Yiu and F. Y. L. Chin (2011). Meta-IDBA: a de Novo assembler for metagenomic data. *Bioinformatics*, **27** (13): i94–i101.

Peoples, M. B., J. Brockwell, D. F. Herridge, I. J. Rochester, B. J. R. Alves *et al.* (2009). The contributions of nitrogen-fixing crop legumes to the productivity of agricultural systems. *Symbiosis*, **48** (1-3): 1–17.

Peterson, S. B., A. K. Dunn, A. K. Klimowicz and J. Handelsman (2006). Peptidoglycan from Bacillus cereus mediates commensalism with rhizosphere bacteria from the

Cytophaga-Flavobacterium group. *Applied and Environmental Microbiology*, **72** (8): 5421–5427.

Pieterse, C. M. J., D. V. d. Does, C. Zamioudis, A. Leon-Reyes and S. C. M. V. Wees (2012). Hormonal Modulation of Plant Immunity. *Annual Review of Cell and Developmental Biology*, **28** (1): 489–521.

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth *et al.* (2014). eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, **42** (Database issue): D231–239.

Price, M. N., P. S. Dehal and A. P. Arkin (2010). FastTree 2 - approximately maximum-likelihood trees for large alignments. *PLoS One*, **5**: e9490.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig *et al.* (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, **35** (21): 7188–7196.

Punta, M., P. C. Coggill, R. Y. Eberhardt, J. Mistry, J. Tate *et al.* (2012). The Pfam protein families database. *Nucleic Acids Research*, **40** (Database issue): D290–301.

Pupko, T., I. Pe'er, R. Shamir and D. Graur (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular Biology and Evolution*, **17** (6): 890–896.

Qin, J., R. Li, J. Raes, M. Arumugam, K. S. Burgdorf *et al.* (2010). A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464** (7285): 59–65.

Quigley, E. M. M. and R. Quera (2006). Small intestinal bacterial overgrowth: roles of antibiotics, prebiotics, and probiotics. *Gastroenterology*, **130** (2 Suppl 1): S78–90.

Radutoiu, S., L. H. Madsen, E. B. Madsen, A. Jurkiewicz, E. Fukai *et al.* (2007). LysM domains mediate lipochitin-oligosaccharide recognition and Nfr genes extend the symbiotic host range. *The EMBO journal*, **26** (17): 3923–3935.

Ramachandran, V. K., A. K. East, R. Karunakaran, J. A. Downie and P. S. Poole (2011). Adaptation of Rhizobium leguminosarum to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.*, **12**: R106.

Raychaudhuri, S., J. M. Stuart and R. B. Altman (2000). Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 455–466.

Rehman, A., C. Sina, O. Gavrilova, R. Haesler, S. Ott *et al.* (2011). Nod2 is essential for temporal development of intestinal microbial communities. *Gut*, **60** (10): 1354–1362.

Reid, D. E., A. B. Heckmann, O. NovÃ¡k, S. Kelly and J. Stougaard (2016). CYTOKININ OXIDASE/DEHYDROGENASE3 Maintains Cytokinin Homeostasis during Root and Nodule Development in Lotus japonicus. *Plant Physiology*, **170** (2): 1060–1074.

Rispail, N., B. Hauck, B. Bartholomew, A. A. Watson, R. J. Nash *et al.* (2010). Secondary metabolite profiling of the model legume Lotus japonicus during its symbiotic interaction with Mesorhizobium loti. *Symbiosis*, **50** (3): 119–128.

Robinson, M. D., D. J. McCarthy and G. K. Smyth (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics (Oxford, England)*, **26** (1): 139–140.

Roeselers, G., E. K. Mittge, W. Z. Stephens, D. M. Parichy, C. M. Cavanaugh *et al.* (2011). Evidence for a core gut microbiota in the zebrafish. *The ISME Journal*, **5** (10): 1595–1608.

Ronquist, F. and J. P. Huelsenbeck (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics (Oxford, England)*, **19** (12): 1572–1574.

Rosenberg, E. and I. Xilber-Rosenberg (2013). *The Hologenome Concept: Human, Animal and Plant Microbiota.*

Round, J. L. and S. K. Mazmanian (2009). The gut microbiota shapes intestinal immune responses during health and disease. *Nature Reviews. Immunology*, **9** (5): 313–323.

Russell, A. B., S. B. Peterson and J. D. Mougous (2014). Type VI secretion system effectors: poisons with a purpose. *Nature Reviews. Microbiology*, **12** (2): 137–148.

Sayers, E. W., T. Barrett, D. A. Benson, S. H. Bryant, K. Canese *et al.* (2009). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, **37** (Database issue): D5–15.

Schauser, L., A. Roussis, J. Stiller and J. Stougaard (1999). A plant regulator controlling development of symbiotic root nodules. *Nature*, **402** (6758): 191–195.

Schlaeppi, K., N. Dombrowski, R. Garrido-Oter, E. V. L. v. Themaat and P. Schulze-Lefert (2014). Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. *Proceedings of the National Academy of Sciences*, **111** (2): 585–592.

Schloissnig, S., M. Arumugam, S. Sunagawa, M. Mitreva, J. Tap *et al.* (2013). Genomic variation landscape of the human gut microbiome. *Nature*, **493** (7430): 45–50.

Schloss, P. D. and J. Handelsman (2006). Toward a Census of Bacteria in Soil. *PLOS Comput Biol*, **2** (7): e92.

Schloss, P. D., S. L. Westcott, T. Ryabin, J. R. Hall, M. Hartmann *et al.* (2009). Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75** (23): 7537–7541.

Schmalenberger, A., S. Hodge, A. Bryant, M. J. Hawkesford, B. K. Singh *et al.* (2008). The role of Variovorax and other Comamonadaceae in sulfur transformations by microbial wheat rhizosphere communities exposed to different sulfur fertilization regimes. *Environmental Microbiology*, **10** (6): 1486–1500.

Searle, I. R., A. E. Men, T. S. Laniya, D. M. Buzas, I. Iturbe-Ormaetxe *et al.* (2003). Long-distance signaling in nodulation directed by a CLAVATA1-like receptor kinase. *Science (New York, N.Y.)*, **299** (5603): 109–112.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**: 2068–2069.

Segata, N., L. Waldron, A. Ballarini, V. Narasimhan, O. Jousson *et al.* (2012). Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature methods*, **9** (8): 811–814.

Sekelja, M., I. Berget, T. Naes and K. Rudi (2011). Unveiling an abundant core microbiota in the human adult colon by a phylogroup-independent searching approach. *The ISME Journal*, **5** (3): 519–531.

Sekirov, I., S. L. Russell, L. C. M. Antunes and B. B. Finlay (2010). Gut Microbiota in Health and Disease. *Physiological Reviews*, **90** (3): 859–904.

Sessitsch, A., P. Hardoim, J. DÃ¶ring, A. Weilharter, A. Krause *et al.* (2012). Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Molecular plant-microbe interactions: MPMI*, **25** (1): 28–36.

Shade, A. and J. Handelsman (2012). Beyond the Venn diagram: the hunt for a core microbiome: The hunt for a core microbiome. *Environmental Microbiology*, **14** (1): 4–12.

Shakya, M., N. Gottel, H. Castro, Z. K. Yang, L. Gunter *et al.* (2013). A Multifactor Analysis of Fungal and Bacterial Community Structure in the Root Microbiome of Mature Populus deltoides Trees. *PLOS ONE*, **8** (10): e76382.

Shames, S. R. and B. B. Finlay (2012). Bacterial effector interplay: a new way to view effector function. *Trends in Microbiology*, **20** (5): 214–219.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang *et al.* (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13** (11): 2498–2504.

Sharma, S., M. Aneja, J. Mayer, J. Munch and M. Schloter (2005). Characterization of Bacterial Community Structure in Rhizosphere Soil of Grain Legumes. *Microbial Ecology*, **49** (3): 407–415.

Sievers, F. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**: 539–539.

Simm, R., M. Morr, A. Kader, M. Nimtz and U. RÃ¶mling (2004). GGDEF and EAL domains inversely regulate cyclic di-GMP levels and transition from sessility to motility. *Molecular Microbiology*, **53** (4): 1123–1134.

Song, S. J., C. Lauber, E. K. Costello, C. A. Lozupone, G. Humphrey *et al.* (2013). Cohabiting family members share microbiota with one another and with their dogs. *eLife*, **2**: e00458.

Sonnhammer, E. L. L. and G. Ostlund (2015). InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*, **43** (Database issue): D234–239.

Soucy, S. M., J. Huang and J. P. Gogarten (2015). Horizontal gene transfer: building the web of life. *Nature Reviews Genetics*, **16** (8): 472–482.

Spoel, S. H. and X. Dong (2008). Making sense of hormone crosstalk during plant immune responses. *Cell Host & Microbe*, **3** (6): 348–351.

Spor, A., O. Koren and R. Ley (2011). Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Reviews Microbiology*, **9** (4): 279–290.

Stacey, G., C. B. McAlvin, S.-Y. Kim, J. Olivares and M. J. Soto (2006). Effects of Endogenous Salicylic Acid on Nodulation in the Model Legumes Lotus japonicus and Medicago truncatula. *Plant Physiology*, **141** (4): 1473–1481.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30** (9): 1312–1313.

Stearns, J. C., M. D. J. Lynch, D. B. Senadheera, H. C. Tenenbaum, M. B. Goldberg *et al.* (2011). Bacterial biogeography of the human digestive tract. *Scientific Reports*, **1**: 170.

Stevens, C. E. and I. D. Hume (1998). Contributions of microbes in vertebrate gastrointestinal tract to production and conservation of nutrients. *Physiological Reviews*, **78** (2): 393–427.

Stracke, S., C. Kistner, S. Yoshida, L. Mulder, S. Sato *et al.* (2002). A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature*, **417** (6892): 959–962.

Subramanian, S. (2015). Cultivating healthy growth and nutrition through the gut microbiota. *Cell*, **161**: 36–48.

Suez, J., T. Korem, D. Zeevi, G. Zilberman-Schapira, C. A. Thaiss *et al.* (2014). Artificial sweeteners induce glucose intolerance by altering the gut microbiota. *Nature*, **514** (7521): 181–186.

Suzaki, T., E. Yoro and M. Kawaguchi (2015). Leguminous plants: inventors of root nodules to accommodate symbiotic bacteria. *International Review of Cell and Molecular Biology*, **316**: 111–158.

Swidsinski, A., J. Weber, V. Loening-Baucke, L. P. Hale and H. Lochs (2005). Spatial Organization and Composition of the Mucosal Flora in Patients with Inflammatory Bowel Disease. *Journal of Clinical Microbiology*, **43** (7): 3380–3389.

Szczyglowski, K., P. Kapranov, D. Hamburger and F. J. de Bruijn (1998). The Lotus japonicus LjNOD70 nodulin gene encodes a protein with similarities to transporters. *Plant Molecular Biology*, **37** (4): 651–661.

Szoboszlay, M., A. White-Monsant and L. A. Moe (2016). The Effect of Root Exudate 7,4'-Dihydroxyflavone and Naringenin on Soil Bacterial Community Structure. *PloS One*, **11** (1): e0146555.

Takeuchi, K., N. Noda, Y. Katayose, Y. Mukai, H. Numa *et al.* (2015). Rhizoxin analogs contribute to the biocontrol activity of a newly isolated pseudomonas strain. *Molecular plant-microbe interactions: MPMI*, **28** (3): 333–342.

Takeuchi, N., Y. I. Wolf, K. S. Makarova and E. V. Koonin (2012). Nature and intensity of selection pressure on CRISPR-associated genes. *Journal of Bacteriology*, **194** (5): 1216–1225.

Tavazoie, S., J. D. Hughes, M. J. Campbell, R. J. Cho and G. M. Church (1999). Systematic determination of genetic network architecture. *Nature Genetics*, **22** (3): 281–285.

Thies, J. E., P. L. Woomer and P. W. Singleton (1995). Enrichment of Bradyrhizobium spp populations in soil due to cropping of the homologous host legume. *Soil Biology and Biochemistry*, **27** (4): 633–636.

Tian, C. F., Y. J. Zhou, Y. M. Zhang, Q. Q. Li, Y. Z. Zhang *et al.* (2012). Comparative genomics of rhizobia nodulating soybean suggests extensive recruitment of lineage-specific genes in adaptations. *Proceedings of the National Academy of Sciences of the United States of America*, **109** (22): 8629–8634.

Tims, S., C. Derom, D. M. Jonkers, R. Vlietinck, W. H. Saris *et al.* (2013). Microbiota conservation and BMI signatures in adult monozygotic twins. *The ISME Journal*, **7** (4): 707–717.

Toft, C. and S. G. E. Andersson (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nature Reviews Genetics*, **11** (7): 465–475.

Touceda-Gonzalez, M., G. Brader, L. Antonielli, V. B. Ravindran, G. Waldner *et al.* (2015). Combined amendment of immobilizers and the plant growth-promoting strain Burkholderia phytofirmans PsJN favours plant growth and reduces heavy metal uptake. *Soil Biology and Biochemistry*, **91**: 140–150.

Tremaroli, V. and F. Baeckhed (2012). Functional interactions between the gut microbiota and host metabolism. *Nature*, **489** (7415): 242–249.

Trevino III, S., Y. Sun, T. F. Cooper and K. E. Bassler (2012). Robust Detection of Hierarchical Communities from Escherichia coli Gene Expression Data. *PLOS Comput Biol*, **8** (2): e1002391.

Triplett, E. W., K. A. Albrecht and E. S. Oplinger (1993). Crop rotation effects on populations of Bradyrhizobium japonicum and Rhizobium meliloti. *Soil Biology and Biochemistry*, **25** (6): 781–784.

Tritt, A., J. A. Eisen, M. T. Facciotti and A. E. Darling (2012). An integrated pipeline for de novo assembly of microbial genomes. *PLoS One*, **7**: e42304.

Trosvik, P., N. C. Stenseth and K. Rudi (2009). Convergent temporal dynamics of the human infant gut microbiota. *The ISME Journal*, **4** (2): 151–158.

Truyens, S., N. Weyens, A. Cuypers and J. Vangronsveld (2015). Bacterial seed endophytes: genera, vertical transmission and interaction with plants. *Environmental Microbiology Reports*, **7** (1): 40–50.

Tsabouri, S., K. N. Priftis, N. Chaliasos and A. Siamopoulou (2014). Modulation of gut microbiota downregulates the development of food allergy in infancy. *Allergologia Et Immunopathologia*, **42** (1): 69–77.

Tungland, B. and D. Meyer (2002). Nondigestible Oligo- and Polysaccharides (Dietary Fiber): Their Physiology and Role in Human Health and Food. *Comprehensive Reviews in Food Science and Food Safety*, **1** (3): 90–109.

Turnbaugh, P. J., M. Hamady, T. Yatsunenko, B. L. Cantarel, A. Duncan *et al.* (2009). A core gut microbiome in obese and lean twins. *Nature*, **457** (7228): 480–484.

Turnbaugh, P. J., C. Quince, J. J. Faith, A. C. McHardy, T. Yatsunenko *et al.* (2010). Organismal, genetic, and transcriptional variation in the deeply sequenced gut microbiomes of identical twins. *Proceedings of the National Academy of Sciences of the United States of America*, **107** (16): 7503–7508.

Turner, T. R., E. K. James and P. S. Poole (2013). The plant microbiome. *Genome Biology*, **14**: 209.

Udvardi, M. and P. S. Poole (2013). Transport and metabolism in legume-rhizobia symbioses. *Annual Review of Plant Biology*, **64**: 781–805.

Unger, S., A. Stintzi, P. Shah, D. Mack and D. L. O'Connor (2015). Gut microbiota of the very-low-birth-weight infant. *Pediatric Research*, **77** (1-2): 205–213.

Valdes, M., N.-O. Perez, P. E.-d. l. Santos, J. Caballero-Mellado, J. J. Pena-Cabriales *et al.* (2005). Non-Frankia Actinomycetes Isolated from Surface-Sterilized Roots of Casuarina equisetifolia Fix Nitrogen. *Applied and Environmental Microbiology*, **71** (1): 460–466.

Van der Sluis, M., B. A. E. De Koning, A. C. J. M. De Bruijn, A. Velcich, J. P. P. Meijerink *et al.* (2006). Muc2-deficient mice spontaneously develop colitis, indicating that MUC2 is critical for colonic protection. *Gastroenterology*, **131** (1): 117–129.

Van Dongen, S. M. (2001). Graph clustering by flow simulation.

van Loon, L. C., P. A. H. M. Bakker and a. C. M. J. Pieterse (1998). Systemic Resistance Induced by Rhizosphere Bacteria. *Annual Review of Phytopathology*, **36** (1): 453–483.

Vlasblom, J. and S. J. Wodak (2009). Markov clustering versus affinity propagation for the partitioning of protein interaction graphs. *BMC Bioinformatics*, **10**: 99.

Vogel, J. P., H. L. Andrews, S. K. Wong and R. R. Isberg (1998). Conjugative Transfer by the Virulence System of Legionella pneumophila. *Science*, **279** (5352): 873–876.

Vorholt, J. A. (2012). Microbial life in the phyllosphere. *Nature Reviews Microbiology*, **10** (12): 828–840.

Wagner, M. R., D. S. Lundberg, D. Coleman-Derr, S. G. Tringe, J. L. Dangl *et al.* (2014). Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild Arabidopsis relative. *Ecology Letters*, **17** (6): 717–726.

Wagner, M. R., D. S. Lundberg, T. G. d. Rio, S. G. Tringe, J. L. Dangl *et al.* (2016). Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications*, **7**: 12151.

Walter, J. and R. Ley (2011). The Human Gut Microbiome: Ecology and Recent Evolutionary Changes. *Annual Review of Microbiology*, **65** (1): 411–429.

Wang, Q., G. M. Garrity, J. M. Tiedje and J. R. Cole (2007). Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology*, **73** (16): 5261–5267.

Watts, D. J. and S. H. Strogatz (1998). Collective dynamics of small-world networks. *Nature*, **393** (6684): 440–442.

Weber, E. and R. Koebnik (2006). Positive selection of the Hrp pilin HrpE of the plant pathogen Xanthomonas. *Journal of Bacteriology*, **188** (4): 1405–1410.

Welham, T., J. Pike, I. Horst, E. Flemetakis, P. Katinakis *et al.* (2009). A cytosolic invertase is required for normal growth and cell development in the model legume, Lotus japonicus. *Journal of Experimental Botany*, **60** (12): 3353–3365.

Weller, D. M., J. M. Raaijmakers, B. B. M. Gardener and L. S. Thomashow (2002). Microbial Populations Responsible for Specific Soil Suppressiveness to Plant Pathogens. *Annual Review of Phytopathology*, **40** (1): 309–348.

Werner, G. D. A., W. K. Cornwell, J. I. Sprent, J. Kattge and E. T. Kiers (2014). A single evolutionary innovation drives the deep evolution of symbiotic N2-fixation in angiosperms. *Nature Communications*, **5**: 4087.

Westra, E. R., A. Buckling and P. C. Fineran (2014). CRISPR-Cas systems: beyond adaptive immunity. *Nature Reviews. Microbiology*, **12** (5): 317–326.

White, L. J., K. Jothibasu, R. N. Reese, V. S. BrÃ¶zel and S. Subramanian (2015). Spatio Temporal Influence of Isoflavonoids on Bacterial Diversity in the Soybean Rhizosphere. *Molecular plant-microbe interactions: MPMI*, **28** (1): 22–29.

Whitman, W. B., D. C. Coleman and W. J. Wiebe (1998). Prokaryotes: the unseen majority. *Proc. Natl Acad. Sci. USA*, **95**: 6578–6583.

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York.

Wittkop, T., D. Emig, A. Truss, M. Albrecht, S. BÃ¶cker *et al.* (2011). Comprehensive cluster analysis with Transitivity Clustering. *Nature Protocols*, **6** (3): 285–295.

Wood, D. E. and S. L. Salzberg (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, **15**: R46.

Wu, G. D., J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen *et al.* (2011). Linking Long-Term Dietary Patterns with Gut Microbial Enterotypes. *Science*, **334** (6052): 105–108.

Wu, M. and J. A. Eisen (2008). A simple, fast, and accurate method of phylogenomic inference. *Genome Biol.*, **9**: R151.

Yang, Z. (2006). *Computational molecular evolution*. Oxford University Press Oxford.

Yatsunenko, T., F. E. Rey, M. J. Manary, I. Trehan, M. G. Dominguez-Bello *et al.* (2012). Human gut microbiome viewed across age and geography. *Nature*, **486** (7402): 222–227.

Yoshimoto, N., H. Takahashi, F. W. Smith, T. Yamaya and K. Saito (2002). Two distinct high-affinity sulfate transporters with different inducibilities mediate uptake of sulfate in Arabidopsis roots. *The Plant Journal*, **29** (4): 465–473.

Young, J. P. W., L. C. Crossman, A. W. B. Johnston, N. R. Thomson, Z. F. Ghazoui *et al.* (2006). The genome of Rhizobium leguminosarum has recognizable core and accessory components. *Genome Biology*, **7** (4): R34.

Zarraonaindia, I., S. M. Owens, P. Weisenhorn, K. West, J. Hampton-Marcell *et al.* (2015). The soil microbiome influences grapevine-associated microbiota. *mBio*, **6** (2).

Zeng, X. and J. Lin (2013). Beta-lactamase induction and cell wall metabolism in Gram-negative bacteria. *Frontiers in Microbiology*, **4**.

Zgadzaj, R., R. Garrido-Oter, D. B. Jensen, A. Koprivova, P. Schulze-Lefert *et al.* (2016). Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. *Proceedings of the National Academy of Sciences*, **113** (49): E7996–E8005.

Zgadzaj, R., E. K. James, S. Kelly, Y. Kawaharada, N. de Jonge *et al.* (2015). A legume genetic framework controls infection of nodules by symbiotic and endophytic bacteria. *PLoS genetics*, **11** (6): e1005280.

Zhang, Z., J. Geng, X. Tang, H. Fan, J. Xu *et al.* (2014). Spatial heterogeneity and co-occurrence patterns of human mucosal-associated intestinal microbiota. *The ISME Journal*, **8** (4): 881–893.

Zhu, W., A. Lomsadze and M. Borodovsky (2010). Ab initio gene identification in metagenomic sequences. *Nucleic Acids Research*, **38** (12): e132.

Zuanazzi, J. A. S., P. H. Clergeot, J.-C. Quirion, H.-P. Husson, A. Kondorosi *et al.* (1998). Production of *Sinorhizobium meliloti nod* Gene Activator and Repressor Flavonoids from *Medicago sativa* Roots. *Molecular Plant-Microbe Interactions*, **11** (8): 784–794.

Zusman, T., M. Feldman, E. Halperin and G. Segal (2004). Characterization of the icmH and icmF Genes Required for Legionella pneumophila Intracellular Growth, Genes That Are Present in Many Bacteria Associated with Eukaryotic Cells. *Infection and Immunity*, **72** (6): 3398–3409.

APPENDIX A

Journal versions of the published articles

INAUGURAL ARTICLE

Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives

Klaus Schlaeppi[a,b], Nina Dombrowski[a], Ruben Garrido Oter[a,c,d], Emiel Ver Loren van Themaat[a], and Paul Schulze-Lefert[a,d,1]

[a]Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany; [b]Plant–Soil-Interactions, Institute for Sustainability Sciences, Agroscope, Reckenholzstrasse 191, 8046 Zurich, Switzerland; [c]Department of Algorithmic Bioinformatics, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany; and [d]Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2010.

Contributed by Paul Schulze-Lefert, November 27, 2013 (sent for review May 25, 2013)

Plants host at the contact zone with soil a distinctive root-associated bacterial microbiota believed to function in plant nutrition and health. We investigated the diversity of the root microbiota within a phylogenetic framework of hosts: three *Arabidopsis thaliana* ecotypes along with its sister species *Arabidopsis halleri* and *Arabidopsis lyrata*, as well as *Cardamine hirsuta*, which diverged from the former ∼35 Mya. We surveyed their microbiota under controlled environmental conditions and of *A. thaliana* and *C. hirsuta* in two natural habitats. Deep 16S rRNA gene profiling of root and corresponding soil samples identified a total of 237 quantifiable bacterial ribotypes, of which an average of 73 community members were enriched in roots. The composition of this root microbiota depends more on interactions with the environment than with host species. Interhost species microbiota diversity is largely quantitative and is greater between the three *Arabidopsis* species than the three *A. thaliana* ecotypes. Host species-specific microbiota were identified at the levels of individual community members, taxonomic groups, and whole root communities. Most of these signatures were observed in the phylogenetically distant *C. hirsuta*. However, the branching order of host phylogeny is incongruent with interspecies root microbiota diversity, indicating that host phylogenetic distance alone cannot explain root microbiota diversification. Our work reveals within 35 My of host divergence a largely conserved and taxonomically narrow root microbiota, which comprises stable community members belonging to the Actinomycetales, Burkholderiales, and Flavobacteriales.

Brassicaceae species | bacterial communities | 16S amplicon ribotyping

Plants host distinct bacterial communities associated with roots and leaves (1, 2). Both the leaf and root microbiota contain bacteria that provide indirect pathogen protection, but root microbiota members appear to serve additional host functions through the acquisition of nutrients from soil supporting plant growth (2). The plant-root microbiota emerges as a fundamental trait that includes mutualism enabled through diverse biochemical mechanisms, as exemplified by previous studies on numerous plant growth and plant health-promoting bacteria (2).

Recent deep profiling of the root microbiota of *Arabidopsis thaliana* ecotypes, grown under controlled environments, confirmed soil type as major source of variation in root microbiota membership and provided evidence for limited host genotype-dependent variation (3, 4). Using four soil types on two continents and based on two 16S rRNA gene PCR primer sets, these replicated experiments revealed a similar phylogenetic structure of the root-associated microbiota at high taxonomic rank, including the phyla Actinobacteria, Bacteroidetes, and Proteobacteria. In addition, these studies revealed a minor "rhizosphere effect" in *A. thaliana*, i.e., a weak differentiation of the bacterial communities in the rhizosphere (soil that is firmly attached to roots) compared with the corresponding unplanted bulk soil.

The genus *Arabidopsis* consists of the four major lineages *Arabidopsis thaliana*, *Arabidopsis lyrata*, *Arabidopsis halleri* and *Arabidopsis arenosa*. The former is the sole self-fertile species and diverged from the rest of the genus ∼13 Mya whereas the other three species radiated approximately ∼8 Mya (5; Fig. 1). *Cardamine hirsuta* diverged from the *Arabidopsis* species ∼35 Mya and often shares the same habitat with *A. thaliana*. *A. thaliana* has a cosmopolitan distribution whereas the other species occur in spatially restricted populations or developed even endemic subspecies, indicative of their adaptation to specific ecological niches (6). The two diploid species, *A. halleri* and *A. lyrata*, co-occur in Eurasia, but, in contrast to *A. lyrata* (Northern rock-cress), the geographical distribution of *A. halleri* rarely extends into northern latitudes. *A. lyrata* primarily colonizes, similar to *A. thaliana*, low-competition habitats as, for example, tundra, stream banks, lakeshores, or rocky slopes, whereas *A. halleri* (Meadow rock-cress) is tolerant of shading and competition, growing in habitats such as mesic meadow sites (7). In contrast to its sister species, *A. halleri* can grow on heavy metal-contaminated soils and serves as a model species for metal hyperaccumulation and associated metal hypertolerance and for extremophile adaptation (8).

The bacterial root microbiota of plants—"plants wear their guts on the outside" (9)—is conceptually analogous to the gut

Significance

All plants carry distinctive bacterial communities on and inside organs such as roots and leaves, collectively called the plant microbiota. How this microbiota diversifies in related plant species is unknown. We investigated the diversity of the bacterial root microbiota in the Brassicaceae family, including three *Arabidopsis thaliana* ecotypes, its sister species *Arabidopsis halleri* and *Arabidopsis lyrata*, and *Cardamine hirsuta*. We show that differences in root microbiota profiles between these hosts are largely quantitative and that host phylogenetic distance alone cannot explain the observed microbiota diversification. Our work also reveals a largely conserved and taxonomically narrow root microbiota, which comprises stable community members belonging to the Actinomycetales, Burkholderiales, and Flavobacteriales.

Author contributions: K.S. and P.S.-L. designed research; K.S., N.D., R.G.O., and E.V.L.v.T. performed research; K.S., N.D., R.G.O., E.V.L.v.T., and P.S.-L. analyzed data; and K.S. and P.S.-L. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The data reported in this paper have been deposited in the European Nucleotide Archive database (accession no. PRJEB5058).

See Profile on page 570.

[1]To whom correspondence should be addressed. E-mail: schlef@mpipz.mpg.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1321597111/-/DCSupplemental.

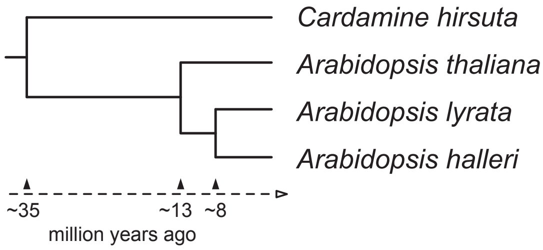


Fig. 1. Phylogenetic placement of the *Arabidopsis* species *A. halleri, A. lyrata*, and *A. thaliana* and relative species *Cardamine hirsuta*. The relationships and divergence time estimates are based on molecular systematics using combined data of NADH dehydrogenase subunit F and phytochrome A sequences anchored by four fossil age constraints (5).

microbiota of animals owing to a shared primary physiological function of root and gut organs for nutrient uptake. The idea of a core microbiota within a species has been initially explored in humans by revealing an extensive array of shared microbial genes among sampled individuals, comprising an identifiable gut core microbiome at the gene, rather than at the level of organismal lineages (10, 11). However, using a phylogroup- and tree-independent approach, two prevalent core phylogroups belonging to the clostridial family Lachnospiraceae were identified in the human colon among a total of 210 human beings with widespread geographic origin, ethnic background, and diet (12). These phylogroups were also detected in a wide range of other mammals and are thought to play a conserved role in gut homeostasis and health. The findings of a core set of species in the human gut microbiota remain contentious as a wider set of samples including developing countries and a broader age range becomes available (13). However, spatial stratification of the gut microbiota, which is normally missing in fecal samples, led to the definition of a crypt-specific core microbiota in the mouse colon, dominated by aerobic Acinetobacter, regardless of the mouse line used or breeding origin of these mice (14). Finally, evidence for a shared core gut microbiota was found in domesticated and recently caught zebrafish, dominated by Proteobacteria, some Fusobacteria, and Firmicutes (15). This shared core is believed to reflect common selective pressures governing microbial community assembly within this intestinal habitat. Although root microbiota profiles of numerous plant species, including crops, have been examined (16–19), different sampling protocols and low-resolution profiling methods make it difficult to reexamine and compare these for the existence of a conserved core microbiota between plant species.

Here, we present a systematic investigation of host–microbiota diversification within a phylogenetically defined plant species framework, combined with replicated experiments under controlled conditions and sampling in natural habitats. Using deep 16S rRNA gene profiling of root and corresponding soil samples of four host species of the Brassicaceae family, together with rigorous statistical analysis, we show that interhost species microbiota diversity is largely quantitative, and we discuss a possible microbiota coevolution with these hosts. We also compared bacterial community structure variation within and between the tested host species. We provide evidence for the existence of a largely conserved and taxonomically narrow root microbiota between the tested host species, which remains stable in natural and controlled environments. This identified core comprises Actinomycetales, Burkholderiales, and Flavobacteriales. Members of each of these bacterial families are known to promote plant growth and plant health. It is possible that the conserved microbiota represents a standing reservoir of retrievable host services independent of environmental parameters and host species-specific niche adaptations.

Results

We collected side-by-side growing *A. thaliana* and *C. hirsuta* plants at two natural sites, designated "Cologne" and "Eifel," and prepared quadruplicate root and rhizosphere samples for bacterial 16S rRNA gene community profiling (Table 1, Dataset S1, and *SI Appendix*). In parallel, we conducted two replicate greenhouse experiments using two seasonal batches of natural experimental Cologne soil (*SI Appendix*, Table S1) on which we cultivated *A. halleri* (Auby), *A. lyrata* (Mn47), and *C. hirsuta* (Oxford), together with the three *A. thaliana* accessions Shakdara (Sha), Landsberg (Ler) and Columbia (Col), and prepared triplicate root samples for microbiota analysis (Table 1, Dataset S1, and *SI Appendix*). Rhizosphere samples are defined as firmly root-adhering soil particles removed by a washing step and collected by centrifugation. Root samples were washed a second time and treated with ultrasound to deplete root surface-associated bacteria and to enrich for endophytic bacteria (3) (*SI Appendix*). To quantify the start inoculum for the root-associated bacterial communities, we prepared triplicate samples from unplanted pots of each greenhouse experiment, as well as four samples from bulk soil collected at each natural site (Table 1 and Dataset S1). Barcoded pyrosequencing of bacterial 16S rRNA gene amplicon libraries generated with the PCR primers 799F (20) and 1193R (21) was used to display bacterial communities (*SI Appendix*).

We generated 2,110,506 raw reads from 77 samples of the replicated natural-site and greenhouse experiments (Dataset S1). For subsequent analysis, we included 1,567,657 quality sequences (*SI Appendix*), resulting in a median of 15,603 quality sequences per sample (range 6,339–58,150 sequences per sample). Quality sequences were binned at >97% sequence identity using QIIME (22) to define operational taxonomic units (OTUs), corrected for differences in sequencing depth between samples by rarefaction to 6,000 sequences per sample. OTU representative sequences were taxonomically classified based on the Greengenes database (23) (*SI Appendix*) and we identified a total of 88,731 unique bacterial OTUs and a single archea OTU across all samples.

Defining Abundant Community Members. Technical reproducibility of community profiles was determined by repeated library sequencing, and we defined a minimum of 20 sequences per OTU for reproducible quantification of OTU abundance (*SI Appendix*, Fig. S1, and Dataset S2). This reproducibility threshold is similar to previous studies (3, 14, 15). We noted a low reproducibility for soil microbiota profiles and found that OTU richness does not reach a plateau even at a sequencing depth of 50,000 quality sequences per sample (*SI Appendix*, Fig. S2*A*). These observations, together with the exclusion of low-abundant (<20 sequences) and nonreproducible OTUs for rarefaction analysis (*SI Appendix*, Fig. S2*B*), suggested that OTU richness in soil is the result of a vast

Table 1. Experimental design

		Natural sites		Greenhouse	
Sample	Species	Cologne	Eifel	Exp. 1	Exp. 2
Soil		4	4	3	3
Rhizosphere	*A. thaliana*	4	4	—	—
Root	*A. thaliana*	4	4	8*	9†
Rhizosphere	*C. hirsuta*	4	3	—	—
Root	*C. hirsuta*	4	3	3	3
Root	*A. halleri*	—	—	3	2
Root	*A. lyrata*	—	—	2	3

Numerical overview of biological replicate samples per sample type, plant species, and experiments. —, sample types which were not prepared. See Dataset S1 for the detailed experimental design, including the sequencing effort.

*Three samples of genotype Col, 3x Ler and 2x Sha.

†Three samples of genotype Col, 3x Ler and 3x Sha.

number of low-count OTUs. In the root samples, the richness of the abundant community members (OTUs with >20 sequences) was sufficiently captured at a sequencing depth of 6,000 sequences per sample. Consequently, we focus our analyses on the abundant community members (ACMs) of the dataset, which we define to comprise OTUs reaching the threshold of 20 quality sequences in at least one sample (*SI Appendix* and Dataset S3). Without application of this abundance threshold, we refer to community profiles rarefied to 6,000 sequences as threshold-independent communities (TICs) (Dataset S4).

The ACM, including soil, rhizosphere, and root samples, was represented by 237 bacterial OTUs comprising 55.3% of rarefied quality sequences (Datasets S1 and S3). Soil and rhizosphere samples contained fewer sequences after thresholding compared with root samples, likely due to increased richness by low-count OTUs in the former two compartments (*SI Appendix*, Figs. S2 and S3). We normalized the counts of the ACM OTUs per sample, expressed their relative abundance as per mille, and used log2-transformed values for statistical comparisons (*SI Appendix*).

Community Composition Is Defined More Strongly by Environmental Parameters than by Host Species. We first examined taxonomic composition and ecological diversity parameters in the whole dataset consisting of samples from two natural sites and two greenhouse experiments. All OTUs of the ACM belonged to the domain of bacteria. In root samples, the majority of OTUs belonged to Proteobacteria (4.2%, 33.6%, 6.4%, and 1.8% in the Alpha-, Beta-, Gamma-, and Deltaproteobacteria subphyla, respectively), Bacteroidetes (27.5%), Actinobacteria (22.1%), and Chloroflexi (2.2%) (*SI Appendix*, Fig. S4*A*). Soil samples also contain Proteobacteria (52.2%) and Actinobacteria (26.8%), but few Bacteroidetes (3.5%) and, characteristic for this compartment, Firmicutes (10.1%) and Nitrospirae (2.4%). Similar taxonomic characteristics of soil and root samples were also found for TICs (*SI Appendix*, Fig. S4*B*). We noted the dominance of a single *Flavobacterium* (OTU162362) in root communities of natural-site and greenhouse experiments, representing, in some of the samples, more than half of the total community. A high OTU diversity in family-rich phyla, such as the Proteobacteria (128 OTUs in 24 families) or Actinobacteria (67 OTUs in 17 families), contrasts with a low taxonomic diversity within the root-specific Bacteroidetes (20 OTUs), all belonging to the family of the Flavobacteriaceae (Dataset S3).

To compare community diversity between samples, we used the weighted UniFrac metric (24) (*SI Appendix*). Consistent with previous studies (3, 4, 16), the hierarchical clustering of UniFrac distances revealed that compartments and environmental conditions (soil types/soil batches, controlled/noncontrolled climates) are the major sources of variation both in the ACMs (Fig. 2) and TICs (*SI Appendix*, Fig. S5). Due to independent library preparation and sequencing, we validated that the variation in the replicate greenhouse experiments reflects biological rather than technical variation (*SI Appendix*, Fig. S6, Dataset S5, and *SI Appendix*). For both natural-site and controlled environment samples, we did not detect a consistent clustering by host species, evidencing that the present sample-to-sample variation obscures a possible host species effect on beta diversity.

We estimated OTU diversity within samples based on the number of OTUs detected (richness) and Faith's Phylogenetic Diversity (PD) metric (25). Root TICs are of lower richness and diversity compared with the soil and rhizosphere microbiota (*SI Appendix*, Fig. S7). Of note, roots of plants grown under natural conditions host bacterial communities of increased richness and Faith's PD compared with greenhouse-grown plants. Root TICs and root ACMs did not differ in richness and Faith's PD among the tested host species in natural and greenhouse experiments (*SI Appendix*, Fig. S7). This finding further supports the existence of qualitatively similar root-associated bacterial assemblies among *A. thaliana* relatives.

Naturally Grown *A. thaliana* and *C. hirsuta* Host a Taxonomically Narrow Root Microbiota. In a second step, we investigated the variation in root microbiota composition between the plant species *A. thaliana* and *C. hirsuta* from both natural-site experiments. We compared the root bacterial communities using ANOVA-based statistics to detect taxonomic groups of OTUs ("community modules") and individual OTUs ("community members") that differ quantitatively between sites and/or host species (*SI Appendix*). The community member analysis was performed on the ACM, and, for the community module analyses, we prepared abundance matrices at phylum and family rank of all ACM OTUs representing 9 and 51 taxa, respectively.

We searched the abundance matrices at phylum and family rank for modules that differ between the root communities as a function of the variables *site* and *host species* (*SI Appendix*). The taxonomic structure of the root communities varies mainly by *site*

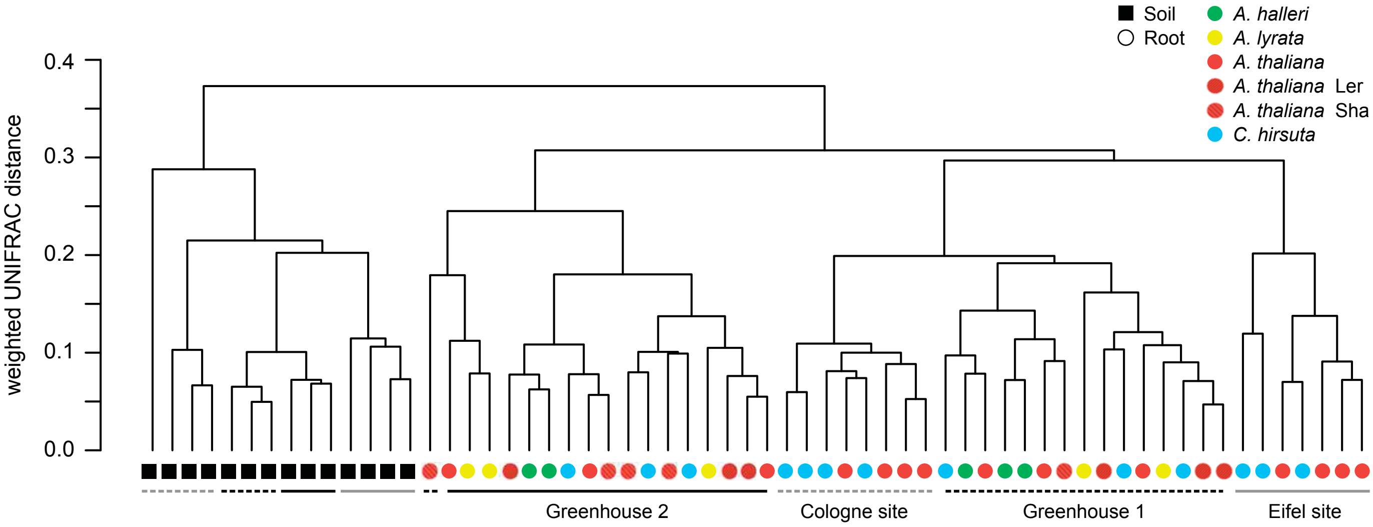


Fig. 2. Beta diversity of the ACM. Between-sample similarities were estimated on 1,400 sequences per sample using the phylogeny-based UniFrac distance metric. Weighted UniFrac is sensitive to the sequence abundances. The *A. thaliana* ecotype Col (nonshaded red) was used in the greenhouse experiments.

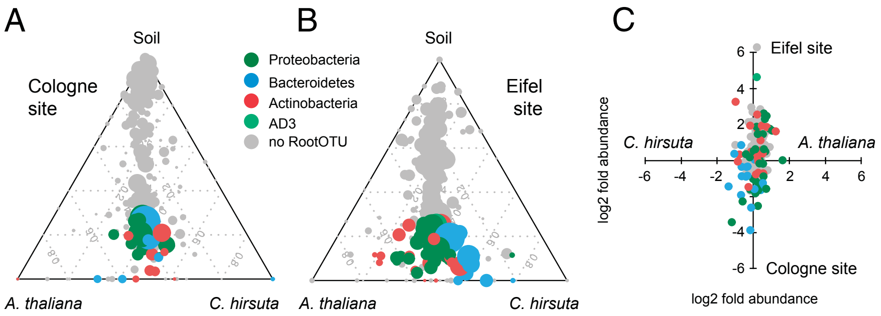


Fig. 3. Root microbiota comparisons of *A. thaliana* and *C. hirsuta* at the natural sites Cologne and Eifel. The ternary plots depict the relative occurrence of individual OTUs (circles) in root samples of *A. thaliana* and *C. hirsuta* compared with the respective soil samples for the Cologne (*A*) and the Eifel site (*B*). RootOTUs [OTUs enriched in roots compared with soil; Tukey, $P < 0.1$ (FDR)] are colored by taxonomy, and OTUs, which are not enriched in root communities, are plotted in gray. The size of the circles is proportional to the mean abundance in the community. (*C*) Variation in mean relative abundance (RA) of individual OTUs (circles) across species and sites, where axes depict logtwofold variation [*x* axis is log2(*A. thaliana*/*C. hirsuta*) and *y* axis is log2(Eifel/Cologne)]. Color coding as in *A* and *B*.

(six phyla, 35 families) and less by *host species* (two phyla, two families; ANOVA, $P < 0.1$ [false discovery rate (FDR) corrected]) (Dataset S6, worksheet A). At both sites, *A. thaliana* and *C. hirsuta* root communities displayed similar relative distributions of bacterial phyla, except for an increased abundance of Bacteroidetes in *C. hirsuta* at the Eifel site [*SI Appendix*, Fig. S8, Tukey, $P < 0.1$ (FDR) and Dataset S6*B*]. The single dominant *Flavobacterium* OTU mentioned earlier (OTU162362) was more abundant in *C. hirsuta* compared with *A. thaliana* root communities. We conclude that *A. thaliana* and *C. hirsuta* root microbiota consist of similar community modules and that the root communities differ quantitatively by their biogeography.

Next, we identified individual community members that differ quantitatively between the two host species. For this analysis, we initially defined for both natural sites the "RootOTUs" (*SI Appendix*), which represent 70 OTUs that are enriched in *A. thaliana* or *C. hirsuta* roots compared with the corresponding soil communities [Tukey, $P < 0.1$ (FDR); Fig. 3 *A* and *B*, *SI Appendix*, Fig. S9, and Dataset S6, worksheets C and D]. Spearman rank correlation coefficients of the RootOTU communities between these two hosts are 0.89 and 0.74 for the Cologne and Eifel sites, respectively, indicating an overall similar RootOTU composition. The RootOTUs in root communities vary mainly by the variable *site* (50 of the 70 RootOTUs) followed by *host species* [18 RootOTUs; ANOVA, $P < 0.1$ (FDR)] (Dataset S6, worksheet E). The comparison of RootOTUs between sites revealed a taxonomically narrow and shared set of 14 RootOTUs, consisting of seven Actinomycetales, three Burkholderiales, three Flavobacteriales, and a Myxococcales OTU (*SI Appendix*, Fig. S10). These shared RootOTUs were validated by parametric Tukey and nonparametric Mann–Whitney and Bayesian statistics (*SI Appendix*) and represent in their abundance half of the community. At the Eifel site, quantitative differences between the two plant species were found in 9 of the 70 RootOTU members, where 7 RootOTUs were more abundant in *A. thaliana* compared with *C. hirsuta* and 2 RootOTUs were less abundant [Tukey, $P < 0.1$ (FDR)] (*SI Appendix*, Fig. S11 and Dataset S6, worksheet F). This finding and the few aforementioned host species-differentiating community modules point to the existence of largely shared bacterial root communities with similar relative abundances in *A. thaliana* and *C. hirsuta*.

Previous studies revealed a weak rhizosphere effect for *A. thaliana* (3, 4). To quantify the rhizosphere effect, we determined for both sites the OTUs that are enriched in the rhizosphere of

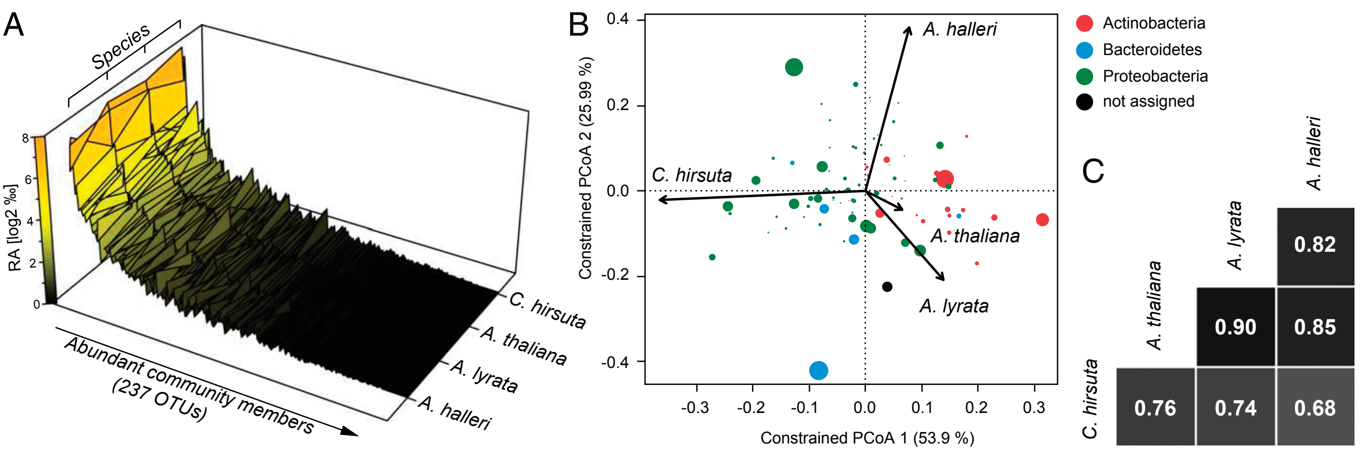


Fig. 4. Root microbiota comparisons of *A. halleri, A. lyrata, A. thaliana*, and *C. hirsuta*. (*A*) The mean abundance of individual OTUs (both replicate experiments) was calculated for the indicated species and plotted ranked by average OTU abundance across all species. (*B*) OTU scores of principal coordinate analysis of the RootOTU community, constrained by *host species* and based on Bray–Curtis distances among root samples (see corresponding *SI Appendix*, Fig. S15). The arrows point to the centroid of the constrained factor. Circle sizes correspond to relative abundances of RootOTUs, and colors are assigned to different phyla. The percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by *host species*. (*C*) Pairwise Spearman rank correlation analysis of the RootOTU communities between the indicated species.

A. thaliana or *C. hirsuta* compared with the corresponding soil (termed "RhizoOTUs") (*SI Appendix*). Similar to these studies, we detected only few RhizoOTUs at the Cologne site [Tukey, $P < 0.1$ (FDR)] (*SI Appendix*, Fig. S12 and Dataset S6, worksheets C and D). The occurrence of a rhizosphere effect was found to be site-dependent: 6 (*A. thaliana*) and 11 RhizoOTUs (*C. hirsuta*) discriminated the rhizosphere from soil communities at the Cologne site whereas no RhizoOTUs (both host species) were found at the Eifel site. We conclude that the magnitude of the rhizosphere effect is site-dependent but independent of the tested host species.

Phylogenetic Distance of Host Species Contributes to Microbiota Diversification. Next, we examined the bacterial root communities retrieved from the *A. thaliana* and the relative species *A. halleri, A. lyrata*, and *C. hirsuta* grown under controlled environmental conditions in replicated greenhouse experiments. A similar overall rank abundance profile of the ACMs in root communities between these four hosts reveals qualitatively similar community structures, indicating that variation in root microbiota is largely quantitative (Fig. 4*A*). We used ANOVA-based statistics to detect community modules and members that vary in abundance between the tested host species (*SI Appendix*). A few taxonomic modules differed in relative abundance between the root microbiota of the tested plant species (*SI Appendix*, Fig. S13), exemplified by significantly lower Bacteroidetes levels in *A. halleri* [Tukey, $P < 0.1$ (FDR)] (Dataset S6, worksheet G). This phenotype was again due to the differential abundance of the dominant *Flavobacterium* mentioned above (OTU162362). At family rank, *A. halleri* and *A. lyrata* display a species-specific quantitative reduction of Flavobacteriaceae and Oxalobacteriaceae, respectively (*SI Appendix*, Fig. S13*B* and Dataset S6, worksheet G).

Analogous to the community member analysis of the natural-site experiments, we identified 76 RootOTUs enriched in roots of at least one plant species compared with soil [Tukey, $P < 0.1$ (FDR)] (*SI Appendix*, Fig. S14 and Dataset S6, worksheets H and I). We then examined the between-sample (beta diversity) variation in the composition of RootOTUs among the host species using canonical analysis of principal coordinates (CAP) (26). CAP analysis constrained for the variable *species* revealed that 17% of the variation in beta diversity, as measured by Bray–Curtis distance metric, was explained by the host species (*SI Appendix*, Fig. S15) ($P < 0.005$; 95% confidence interval = 12%, 25%). The samples clustered by host species and distances between host species revealed that the root communities of *A. thaliana* were more similar to *A. lyrata* than to *A. halleri* and that the root microbiota of the three *Arabidopsis* species are more similar to each other than to the root microbiota of *C. hirsuta*. Thus, within the genus *Arabidopsis* (*A. thaliana*, *A. halleri*, and *A. lyrata*), microbiota diversification is incongruent with the phylogenetic distance of these hosts (compare Fig. 1 and *SI Appendix*, Fig. S15). Further exploration of the CAP analysis revealed a correspondence between the taxonomy of the RootOTUs and their contribution to the microbial diversity between host species: RootOTUs of the phylum Actinobacteria showed the strongest influence on the variation between the *Arabidopsis* species and *C. hirsuta* root communities (Fig. 4*B*). Similarly, the abundance of Bacteroidetes largely explains the differentiation between *A. halleri* and the other host species.

Community similarities were confirmed by pairwise correlation analysis of the RootOTU communities between the four host species, revealing Spearman rank coefficients ranging from 0.68 to 0.90 (Fig. 4*C*). The RootOTU composition of *A. thaliana* correlated best with each of its sister species *A. halleri* and *A. lyrata*, and all three pair-wise comparisons of *C. hirsuta* with the *Arabidopsis* species showed low correlation coefficients, suggesting that the evolutionarily most ancient plant species hosts a RootOTU community, which is quantitatively most diversified. Thus, inclusion of the more distant *C. hirsuta* suggests that phylogenetic distance of host species contributes to microbiota diversification across all four tested hosts. These observations were supported by the highest number of species-specific RootOTU accumulation patterns for *C. hirsuta* [Tukey, $P < 0.1$ (FDR)] (*SI Appendix*, Fig. S16 and Dataset S6, worksheet J). In total, we identified 14 species-specific RootOTUs that consisted of 1, 2, 4, and 7 RootOTUs for *A. thaliana, A. halleri*, *A. lyrata*, and *C. hirsuta*, respectively (*SI Appendix*, Fig. S16). The lower accumulation of the *A. halleri*-specific *Flavobacterium* (OTU162362) and the Oxalobacteriaceae member (OTU91279) in *A. lyrata*

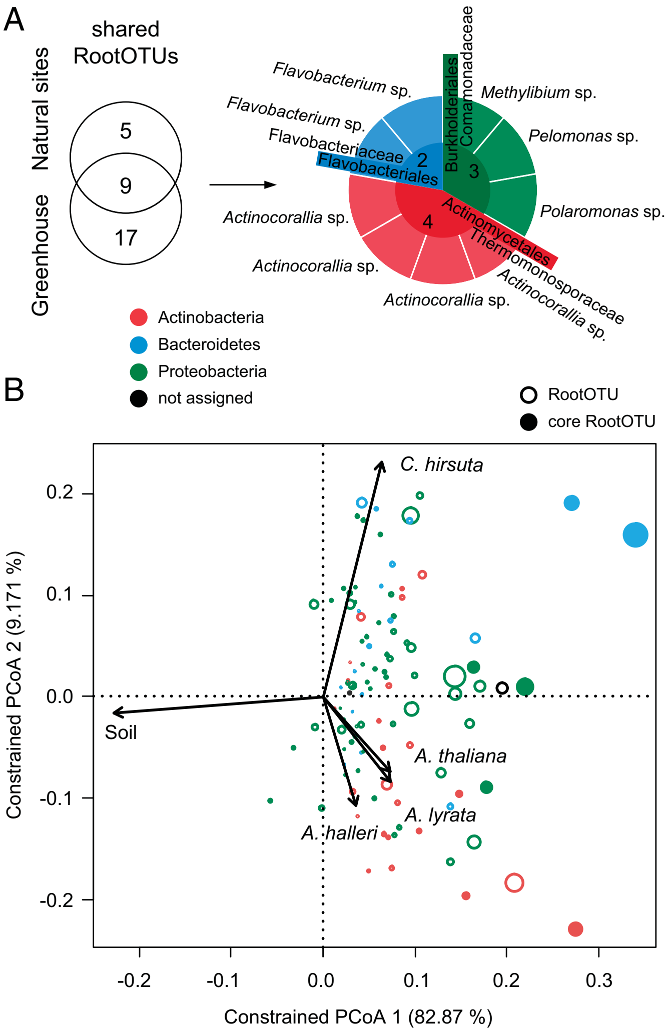


Fig. 5. Identification of the core microbiota. (*A*) Core members result from the intersection of the shared RootOTUs found at the natural sites (*SI Appendix*, Fig. S10) and the shared RootOTUs detected in the greenhouse experiments (*SI Appendix*, Fig. S19). The pie chart segments are colored by the bacterial phyla of the corresponding taxa. The taxonomic assignments of the core RootOTUs are reported at order and family rank in the center and the first ring of the pie chart, respectively. The genera of the core members are noted at the periphery of the pie chart. (*B*) OTU scores of principal coordinate analysis of the RootOTU community, constrained by *sample groups* and based on Bray–Curtis distances among soil and root samples (see corresponding *SI Appendix*, Fig. S22). Sample groups include root samples by species and soil samples. The arrows point to the centroid of the constrained factor. Circle sizes correspond to relative abundances, colors are assigned to different phyla, and core members are marked as solid circles. The percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by host species.

(*SI Appendix*, Fig. S16) contributed to the species-specific accumulation of the corresponding community modules (*SI Appendix*, Fig. S13). Similarly, the Actinocorallia RootOTU (OTU97580) contributed to the trend of lower accumulation of the Thermomonosporaceae, a distinctive feature of *C. hirsuta* (*SI Appendix*, Fig. S13). We independently validated the lower accumulation of Thermononosporaceae in *C. hirsuta* using quantitative PCR with taxon-specific PCR primers (*SI Appendix*, Fig. S17 and *SI Appendix*).

We assessed variation in microbiota composition between and within host species by a direct comparison of the three *A. thaliana* ecotypes with the three *Arabidopsis* sister species (*SI Appendix*). Ternary plots revealed a larger spread of abundant OTUs between the sister species than between the *A. thaliana* ecotypes (*SI Appendix*, Fig. S18). This observation is supported by the identification of 13 host species-dependent OTUs and one host genotype-dependent OTU [ANOVA, $P < 0.1$ (FDR)] (Dataset S6, worksheet K). This direct comparison demonstrates a greater inter- compared with intraspecies variation in microbiota composition.

Members of the Actinomycetales, Burkholderiales, and Flavobacteriales Are Stable Across Host Species and Environments. We identified in the controlled environment experiments 26 RootOTUs shared among the four tested host species (*SI Appendix*, Fig. S19*B*). These shared RootOTUs belonged to the orders Burkholderiales (11 RootOTUs), Actinomycetales (7), Rhizobiales (3), Flavobacteriales (2), Myxococcales (1), Xanthomonadales (1), and Herpetosiphonales (1). These OTUs were validated by parametric Tukey and nonparametric Mann–Whitney and Bayesian statistics and constituted by their relative abundance the bulk of the root community (~75%). Remarkably, the most abundant orders were also recovered in the shared RootOTUs in the natural-site experiments (*SI Appendix*, Fig. S10). The intersection of RootOTUs shared between plant species found at natural sites and in greenhouse experiments determined the core microbiota (Fig. 5*A* and *SI Appendix*, Fig. S20). This core consisted of nine RootOTUs assigned to the orders Actinomycetales (four RootOTUs, genus *Actinocorallia*), Burkholderiales (three, family Comamonadaceae), and Flavobacteriales (two, genus *Flavobacterium*) (Fig. 5*A*). This core represented a taxonomically extremely reduced subcommunity of the microbiota in all tested host species, and together these RootOTUs constituted by their abundance up to half of the root microbiota in all samples tested (*SI Appendix*, Fig. S20*A*). The enrichment of these core microbiota members relative to soil across plant species and sites was identified by three statistical methods (*SI Appendix*, Figs. S10*B* and S19*B*) and confirmed by a subsampling technique (i.e., bootstrapping) (*SI Appendix*, Fig. S20*B*, and *SI Appendix*). In addition, bootstrapping predicted OTUs of the orders Rhizobiales and Myxococcales to be part of the core microbiota (*SI Appendix*, Fig. S20*B*). However, the abundance of the latter two orders was less stable between environments, and, therefore, they did not pass the stringent identification of significant RootOTUs in the original data set using three different statistical methods (Fig. 5). Taken together, the core RootOTUs found across all host species and sites belonged to only three bacterial orders: the Actinomycetales, Burkholderiales, and Flavobacteriales. We compared the composition of this core microbiota to *A. thaliana* root endophyte communities from previous studies (3, 4, 21), which were based on different soil types, environments, and PCR primer combinations (*SI Appendix*, Fig. S21). The raw 16S rRNA gene sequences of these studies were coclustered with the sequences of this study, and the common OTU table was examined for the core microbiota in each data subset using the statistical procedure of this study (*SI Appendix*, Fig. S21 *A* and *B*, Dataset S7, and *SI Appendix*). A common core at OTU level cannot be confirmed in other *A. thaliana* root microbiome studies (*SI Appendix*, Fig. S21*C*) whereas, at order rank, the presence of Actinomycetales presents the common denominator across all studies. Burkholderiales and Flavobacteriales were detected in three and two of the four studies, respectively. Additionally, Rhizobiales and Sphingomonadales were each detected once as core members.

Using CAP analysis, we finally investigated the contribution of the core RootOTUs to the overall variation in root—compared with soil samples in all experimental conditions. Therefore, we constrained the analysis for all *sample groups*, i.e., root samples by species and the soil samples as additional group (*SI Appendix*, Fig. S22 and Fig. 5*B*). Consistent with the unconstrained beta diversity analysis (Fig. 2), the compartment constituted the major source of variation (*SI Appendix*, Fig. S22). We observed a clear differentiation between soil and root samples along the first principal coordinate, which explained the largest fraction of the variation (82.87%). We noted that the core RootOTUs (filled circles in Fig. 5*B*)—having the largest species descriptors—contributed most to the formation of the ordination space. This observation was consistent with their definition (enriched in root samples) and identification in all experimental conditions. Importantly, we confirmed the correspondence between the taxonomy of the RootOTUs and root microbiota diversity across host species over all experimental conditions: the root microbiota of *Arabidopsis* species were distinguished by RootOTUs of the phylum Actinobacteria (open and closed red circles in Fig. 5*B*) whereas root bacterial communities of *C. hirsuta* were differentiated by Bacteroidetes (open and closed blue circles in Fig. 5*B*). We interpreted these correspondences as evidence of a host impact on the root microbiota at a high taxonomic rank.

Discussion

A Conserved Core Root Microbiota? Here, we have examined the bacterial root microbiota of *A. thaliana* along with its sister species *A. lyrata* and *A. halleri* and of *C. hirsuta*. This study revealed the existence of a core root microbiota comprising members from the three bacterial orders Actinomycetales, Burkholderiales, and Flavobacteriales (Fig. 5*A*). Previous studies (3, 4, 21), using different 16S rRNA PCR primer combinations, reported that *A. thaliana* roots host mainly Actinobacteria, Bacteroidetes, and Proteobacteria. This taxonomic structure at phylum level is congruent with the core composition described here because the order Actinomycetales belongs to the phylum Actinobacteria, the Burkholderiales belongs to the subphylum Betaproteobacteria, and the Flavobacteriales to the phylum Bacteroidetes. Bootstrapping also identified the core members and revealed additional RootOTUs, expanding the core composition (*SI Appendix*, Fig. S20*B*). These Rhizobiales and Myxococcales members became apparent in approximately half of random subsets of the original data. Enhanced variation in their abundance between replicate samples and tested environments could explain their absence from the core microbiota.

The significance of our definition of the core microbiota is potentially constrained by the PCR primer used and a low number of tested environments (Cologne and Eifel natural sites and controlled environment). It remains to be seen whether additional samples, also from extreme environments, modify the composition of the core. Corroborating evidence for its stability in additional environments comes from a recent *A. thaliana* field study comprising four disturbed sites in the United States using the same PCR primer combination (21). In these root samples, Actinomycetales, Burkholderiales, and Flavobacteriales were found by 16S rRNA gene pyrosequencing among other prevalent taxa, and the taxonomic composition of the core at order level is similar to our study (*SI Appendix*, Fig. S21*C*). Differences in the selectivity of different 16S PCR primers and variation in 16S rRNA gene copy number likely distort the composition of the core root microbiota. The comparison across *A. thaliana* root microbiome studies (3, 4, 21) did not reveal a common core at the OTU level (*SI Appendix*, Fig S21*C*). However, we cannot discriminate whether PCR primer bias, the soil type/start inoculum, or combinations thereof account for the lack of a common OTU core. Despite this lack of clarity, we noted that, at higher taxonomic rank, the enrichment of members of the Actinomycetales in roots was a common feature of all *A. thaliana* root microbiome studies (*SI Appendix*, Fig. S21*B*). Actinobacteria, including the Actinomycetales, appear to be enriched from soil by cues from living *A. thaliana* roots (3). Our study suggests that such host-derived assembly signal(s) are evolutionarily conserved, at least in the Brassicaceae. Future examination of root microbiomes from additional plant species in the same environments and using the same PCR primer combination will test whether this core is a lineage-specific innovation of the Brassicaceae family.

The core root microbiota members accounted for up to half of the total community size (*SI Appendix*, Fig. S20*A*). The core consisted of both abundant and low-abundant community members, suggesting that assembly and physiological function(s) depend on selective membership and regulation of their relative abundance. In addition, we found a correspondence between the taxonomy of the bacterial communities and the diversity pattern of the root microbiota across host species and three different environments (Fig. 5B). The core members largely supported this correspondence. These observations, together with the reduced taxonomic complexity of the core community, point to the existence of a common organizing principle for their establishment. We speculate about two potential and mutually not exclusive mechanisms that take part in the establishment process: (*i*) each bacterial lineage autonomously responds to host-derived cues and (*ii*) microbe–microbe interactions enable a selective advantage for cocolonization by core members. For example, the commensal relationship between *Bacillus cereus* and bacteria of the *Cytophaga-Flavobacterium* (CF) group in the soybean rhizosphere is mediated by peptidoglycan, which is produced by *B. cereus* and stimulates the growth of CF bacteria (27). From future root-microbiota metagenome and -transcriptome analyses, we expect insights into the connectivity among microbes (28). Such approaches will define the core microbiota at the level of genes rather than taxonomic lineages and will provide a deeper understanding of host services of the core as pioneered by human gut microbiota research (10, 11).

Members of each taxon of the core are potentially beneficial for their hosts. Grassland rhizosphere-derived cultured strains belonging to the Burkholderiales were shown to have antagonistic activities toward multiple soilborne oomycete and fungal pathogens (29). Wheat rhizosphere-derived Comamonadaceae strains appear to be functional specialists in soil sulfonate transformation as part of the biogeochemical sulfur cycle, likely enabling mineralization of organic sulfur for sulfate acquisition by high-affinity sulfate transporters at the root surface (30, 31). *Flavobacterium*, a common soil and water bacterium, was positively correlated with potato biomass and frequently isolated from barley and bell pepper rhizospheres (32–34). A reference *Flavobacterium* isolated from the rhizosphere of the latter plant tested positive for several biochemical assays associated with plant growth promotion and pathogen protection, and accessible genomes of soil/rhizosphere-derived Flavobacteria indicate that these bacteria define a distinct clade compared with Flavobacteria from aquatic environments (34). Finally, root-derived isolates of the Thermomonosporaceae and other core members such as Polaromonas isolates are capable of fixing atmospheric nitrogen (35, 36), potentially increasing the amount of bioavailable nitrogen for plant growth.

Root Microbiota Interactions with the Environment. If the aforementioned physiological function(s) of the core hold true for isolates present in Brassicaceae roots, then these functions could present a standing reservoir of retrievable host services independently of environmental parameters and host species-specific niche adaptations (e.g., metal tolerance of *A. halleri*) or life history traits (perennialism of *A. lyrata* and *A. halleri*). Although the conserved core is established in both controlled and natural environments, the composition of the entire root microbiota depends most on interactions with the environment (Fig. 2). This dependence is consistent with prior findings that soil type is the key determinant of root community structure (3, 4, 16). Thus, the whole root microbiota consists of two parts, the conserved core and an environment-responsive subcommunity. For example, members of the order Rhizobiales represent a root community module found only as shared RootOTUs in the controlled environment using experimental Cologne soil (*SI Appendix*, Fig. S19). In addition, the relative abundance of the family Oxalobacteriaceae is environment-responsive (*SI Appendix*, Figs. S8 and S13), resulting in a change in rank abundance from rank 2 in the greenhouse experiments to rank 3 at the natural sites. Likewise, dominant Streptomycetaceae in the controlled environment (rank 3) are a low-abundant taxon at the natural sites (rank 14). Thus, the relative abundance of these taxa is tunable by the environment. We speculate that these environment-responsive community modules provide services to all tested host species in an environment-dependent manner.

The strong responsiveness of the root microbiota to the environment might be explained by the fact that soil does not only define the start inoculum but also the "diet" for plants, e.g., bioavailability of macro- and micronutrients. Diet as a major driver for community structure was previously reported in microbiota hosted by other eukaryotes. For example, the mammalian gut microbiota follows primarily the dietary habits of the animals, where communities from herbivores, carnivores, and omnivores clustered clearly apart from each other (37). Dietary patterns in humans appear to determine gut microbiota enterotypes (38, 39). However, it remains to be examined whether dietary effects independent of the soil start inoculum are sufficiently strong to provoke consistent shifts in root microbiota community composition. A further exploration of this question would require the modulation of individual nutrients or their composition in the same soil type/start inoculum and subsequent determination of possible effects on community structure.

Host Microbiota Coevolution. A systematic investigation of host-microbiota diversification within a phylogenetically defined plant species framework, combined with replicated experiments under controlled conditions, has not been reported before. A major finding of our work is that the diversification of the root microbiota of the tested host species is largely quantitative. This conclusion is based on abundant community members (ACM, >20 sequences per OTU in at least one sample of the dataset), and, therefore, qualitative differences might exist in the rare biosphere, which is currently not quantifiable. Despite an overall interhost species microbiota similarity, we found host species-specific community modules (*SI Appendix*, Figs. S8 and S13) and members (*SI Appendix*, Figs. S11 and S16). The host species-specific community modules are clearly part of the environment-responsive subcommunity, as illustrated by the observation that they are site-dependent (Fig. 3). Both the most divergent root microbiota (Figs. 4 *B* and *C* and 5*B*) and the highest number of species-specific community members (*SI Appendix*, Fig. S16) were found in the phylogenetically most distant *C. hirsuta* (Fig. 1), suggesting that phylogenetic distance of the hosts could contribute to microbiota diversification. Future studies using additional plant lineages are required to conclude whether phylogenetic distance time correlates with microbiota diversification. For example, in primates, the branching order of host-species phylogeny was found to be congruent with gut community composition (40). The greatest similarities in root microbiota were found between *A. thaliana* and *A. lyrata* whereas the root microbiota of *A. halleri* was more dissimilar (Fig. 4 *B* and *C*), demonstrating that, within the genus *Arabidopsis* (*A. thaliana*, *A. halleri*, and *A. lyrata*), microbiota diversification is incongruent with the phylogenetic distances of these hosts (Fig. 1). The two species *A. thaliana* and *A. lyrata* occur in similar habitats whereas *A. halleri* has evolved a distinctive lifestyle enabling growth in mesic sites and tolerance to high-competition habitats. This particularity could imply that the recent speciation event of *A. halleri*, coupled to an adaptation to a distinctive habitat, resulted in the selection of a distinctive microbiota with habitat-specific services. Taken together, both host species-specific ecological adaptation and phylogenetic distance might have driven microbiota diversification among the tested hosts. Whether the proportion of species-specific community members/modules increases when the host species are exposed to stressful conditions where a plant species has an adaptive advantage (e.g., tolerance of *A. halleri* to metalliferous soils) (8) remains to be tested. Similarly, it will be interesting to examine whether the proportion of species-specific community members/modules increases when the perennials *A. lyrata* and *A. halleri* are grown according to their lifestyle for longer than 1 y. Finally, future experimentation using synthetic bacterial communities with isolates of the core microbiota members and gnotobiotic *Arabidopsis* plants will directly test whether their presumed beneficial roles in plant growth and health can be reproduced under laboratory conditions and are retrievable by the host under normal and stressful conditions.

Materials and Methods

We collected roots of naturally occurring *A. thaliana* and *C. hirsuta* growing side by side at the two replicate sites Cologne and Eifel. Additionally, we sampled in two replicate experiments roots of *A. thaliana* and the relative species *A. lyrata*, *A. halleri*, and *C. hirsuta*, which were grown under controlled conditions in the greenhouse in pots containing natural microbe-rich soil. We used a root-sampling protocol similar to Bulgarelli et al. (3) to examine the root-inhabiting bacterial microbiota. For comparison, we also sampled bulk soil and rhizosphere compartments. Bacterial communities were characterized by pyrosequencing 16S rRNA gene amplicons derived from the PCR primers 799F (20) and 1193R (21). The pyrosequencing reads were processed and analyzed with the software QIIME (22), and custom R scripts were used for statistical analyses. For details, see *SI Appendix*.

ACKNOWLEDGMENTS. We thank Dr. Bruno Hüttel and Dr. Kurt Stüber (Max Planck Genome Centre Cologne) for excellent support in amplicon library preparation and pyrosequencing. We thank Dr. Yang Bai and Dr. Davide Bulgarelli (Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research) for critical comments on the manuscript. Dr. Pierre Saumitou-Laprade (Laboratoire de Génétique et Évolution des Populations Végétales, Formations de Recherche en Évolution, Centre National de la Recherche Scientifique 3268, Université de Lille) kindly provided seeds of *A. halleri*. K.S. was supported by Swiss National Science Foundation Grants PBFRP3-133544 and PA00P3-136473. This work was supported by funds to P.S.-L. from the Max Planck Society (M.IF.A. ZUCH8048), a European Research Council advanced grant (ROOTMICROBIOTA), and the "Cluster of Excellence on Plant Sciences" program funded by the Deutsche Forschungsgemeinschaft.

1. Vorholt JA (2012) Microbial life in the phyllosphere. *Nat Rev Microbiol* 10(12):828–840.
2. Bulgarelli D, Schlaeppi K, Spaepen S, Ver Loren van Themaat E, Schulze-Lefert P (2013) Structure and functions of the bacterial microbiota of plants. *Annu Rev Plant Biol* 64:807–838.
3. Bulgarelli D, et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* 488(7409):91–95.
4. Lundberg DS, et al. (2012) Defining the core *Arabidopsis thaliana* root microbiome. *Nature* 488(7409):86–90.
5. Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S (2010) Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* 107(43):18724–18728.
6. Hoffmann MH (2005) Evolution of the realized climatic niche in the genus Arabidopsis (Brassicaceae). *Evolution* 59(7):1425–1436.
7. Clauss MJ, Koch MA (2006) Poorly known relatives of *Arabidopsis thaliana*. *Trends Plant Sci* 11(9):449–459.
8. Krämer U (2010) Metal hyperaccumulation in plants. *Annu Rev Plant Biol* 61:517–534.
9. Janzen DH (1985) The natural history of mutualisms. *The Biology of Mutualism: Ecology and Evolution*, ed Boucher DH (Oxford Univ Press, New York).
10. Turnbaugh PJ, et al. (2009) A core gut microbiome in obese and lean twins. *Nature* 457(7228):480–484.
11. Qin JJ, et al.; MetaHIT Consortium (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* 464(7285):59–65.
12. Sekelja M, Berget I, Næs T, Rudi K (2011) Unveiling an abundant core microbiota in the human adult colon by a phylogroup-independent searching approach. *ISME J* 5(3):519–531.
13. Lozupone CA, Stombaugh JI, Gordon JI, Jansson JK, Knight R (2012) Diversity, stability and resilience of the human gut microbiota. *Nature* 489(7415):220–230.
14. Pedron T, et al. (2012) A crypt-specific core microbiota resides in the mouse colon. *Mbio* 3(3):e00116-12.
15. Roeselers G, et al. (2011) Evidence for a core gut microbiota in the zebrafish. *ISME J* 5(10):1595–1608.
16. Peiffer JA, et al. (2013) Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc Natl Acad Sci USA* 110(16):6548–6553.
17. Inceoğlu O, Al-Soud WA, Salles JF, Semenov AV, van Elsas JD (2011) Comparative analysis of bacterial communities in a potato field as determined by pyrosequencing. *PLoS ONE* 6(8):e23321.
18. Hardoim PR, et al. (2011) Rice root-associated bacteria: Insights into community structures across 10 cultivars. *FEMS Microbiol Ecol* 77(1):154–164.
19. Sharma S, Aneja MK, Mayer J, Munch JC, Schloter M (2005) Characterization of bacterial community structure in rhizosphere soil of grain legumes. *Microb Ecol* 49(3):407–415.
20. Chelius MK, Triplett EW (2001) The diversity of Archaea and Bacteria in association with the roots of *Zea mays* L. *Microb Ecol* 41(3):252–263.
21. Bodenhausen N, Horton MW, Bergelson J (2013) Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PLoS ONE* 8(2):e56329.
22. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336.
23. McDonald D, et al. (2012) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* 6(3):610–618.
24. Lozupone C, Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* 71(12):8228–8235.
25. Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biol Conserv* 61(1):1–10.
26. Anderson MJ, Willis TJ (2003) Canonical Analysis of Principal Coordinates: A useful method of constrained ordination for ecology. *Ecology* 84:511–525.
27. Peterson SB, Dunn AK, Klimowicz AK, Handelsman J (2006) Peptidoglycan from *Bacillus cereus* mediates commensalism with rhizosphere bacteria from the Cytophaga-Flavobacterium group. *Appl Environ Microbiol* 72(8):5421–5427.
28. Shade A, Handelsman J (2012) Beyond the Venn diagram: The hunt for a core microbiome. *Environ Microbiol* 14(1):4–12.
29. Benítez MS, Gardener BBM (2009) Linking sequence to function in soil bacteria: Sequence-directed isolation of novel bacteria contributing to soilborne plant disease suppression. *Appl Environ Microbiol* 75(4):915–924.
30. Schmalenberger A, et al. (2008) The role of Variovorax and other Comamonadaceae in sulfur transformations by microbial wheat rhizosphere communities exposed to different sulfur fertilization regimes. *Environ Microbiol* 10(6):1486–1500.
31. Yoshimoto N, Takahashi H, Smith FW, Yamaya T, Saito K (2002) Two distinct high-affinity sulfate transporters with different inducibilities mediate uptake of sulfate in Arabidopsis roots. *Plant J* 29(4):465–473.
32. Manter DK, Delgado JA, Holm DG, Stong RA (2010) Pyrosequencing reveals a highly diverse and cultivar-specific bacterial endophyte community in potato roots. *Microb Ecol* 60(1):157–166.
33. Johansen JE, Nielsen P, Binnerup SJ (2009) Identification and potential enzyme capacity of flavobacteria isolated from the rhizosphere of barley (*Hordeum vulgare* L.). *Can J Microbiol* 55(3):234–241.
34. Kolton M, et al. (2012) Draft genome sequence of *Flavobacterium* sp. strain F52, isolated from the rhizosphere of bell pepper (Capsicum annuum L. cv. Maccabi). *J Bacteriol* 194(19):5462–5463.
35. Valdés M, et al. (2005) Non-Frankia actinomycetes isolated from surface-sterilized roots of Casuarina equisetifolia fix nitrogen. *Appl Environ Microbiol* 71(1):460–466.
36. Hanson BT, Yagi JM, Jeon CO, Madsen EM (2012) Role of nitrogen fixation in the autecology of *Polaromonas naphthalenivorans* in contaminated sediments. *Environ Microbiol* 14(6):1544–1557.
37. Ley RE, et al. (2008) Evolution of mammals and their gut microbes. *Science* 320(5883):1647–1651.
38. Arumugam M, et al.; MetaHIT Consortium (2011) Enterotypes of the human gut microbiome. *Nature* 473(7346):174–180.
39. Wu GD, et al. (2011) Linking long-term dietary patterns with gut microbial enterotypes. *Science* 334(6052):105–108.
40. Ochman H, et al. (2010) Evolutionary relationships of wild hominids recapitulated by gut microbial communities. *PLoS Biol* 8(11):e1000546.

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

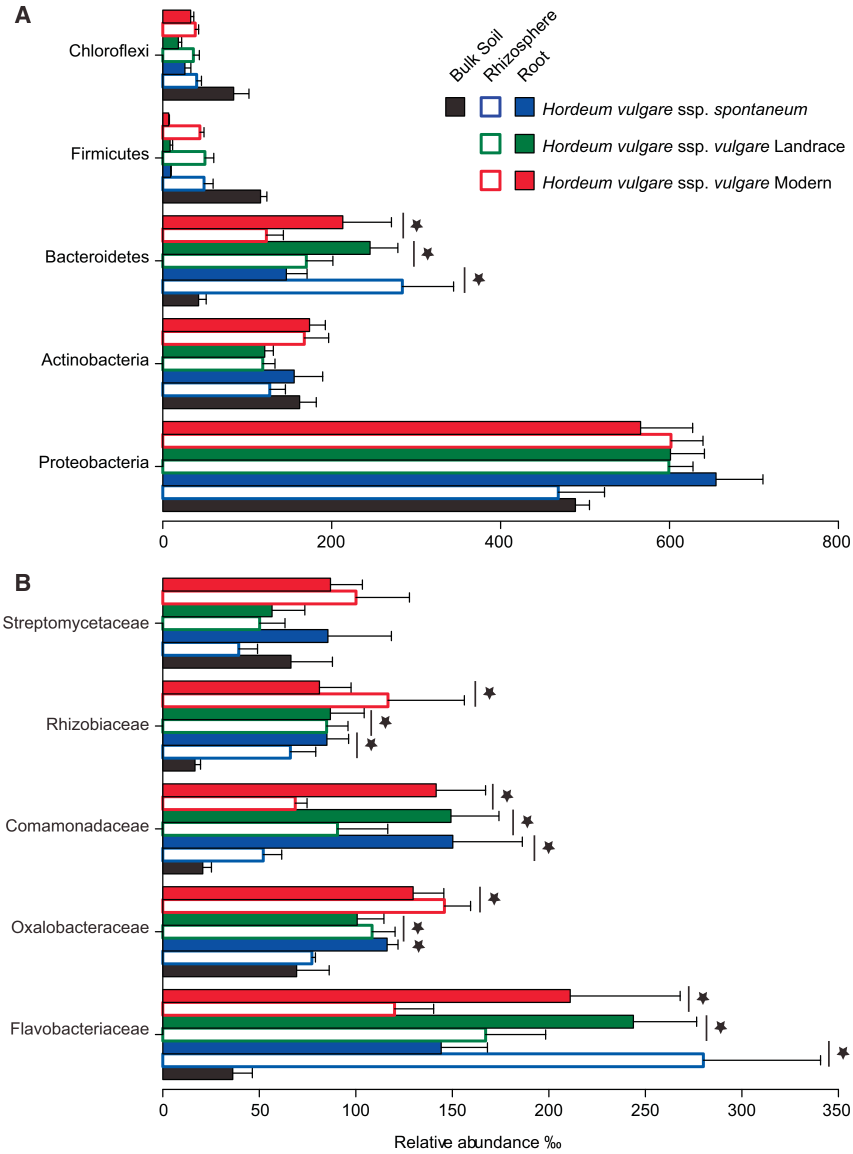


Figure 1. The Barley Rhizosphere and Root Microbiota Are Gated Communities

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

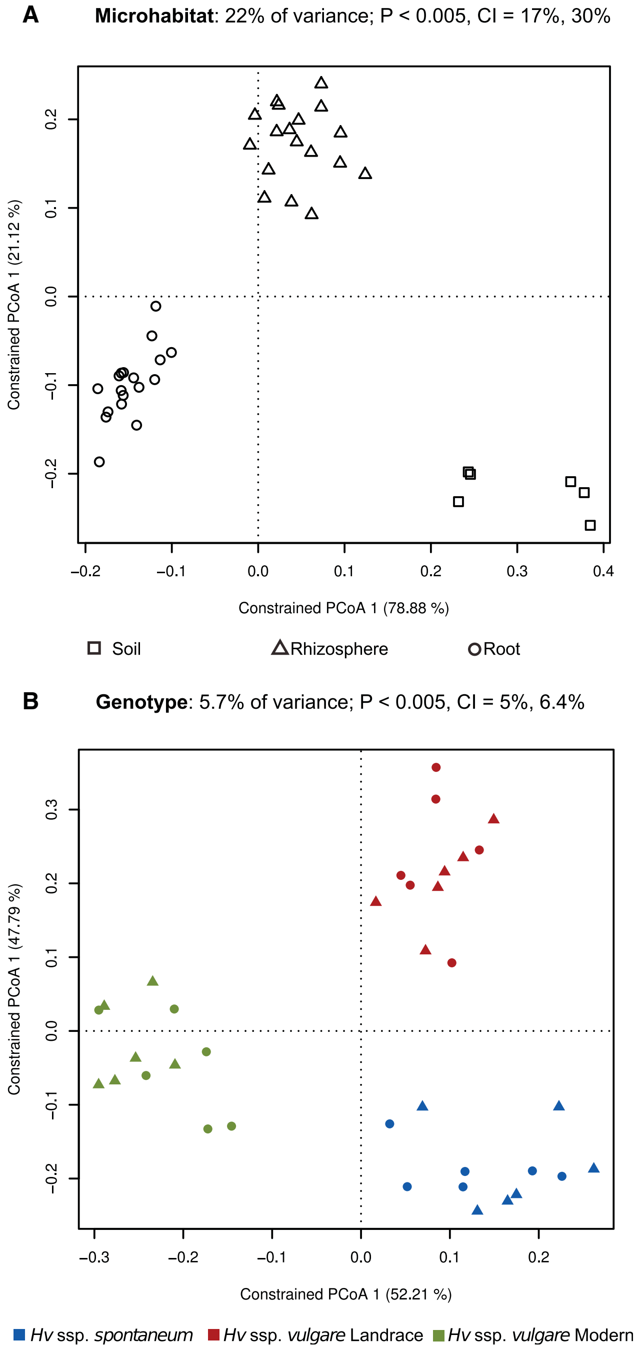


Figure 2. Constrained Principal Coordinate Analysis on the Soil and Barley Bacterial Microbiota

(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 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

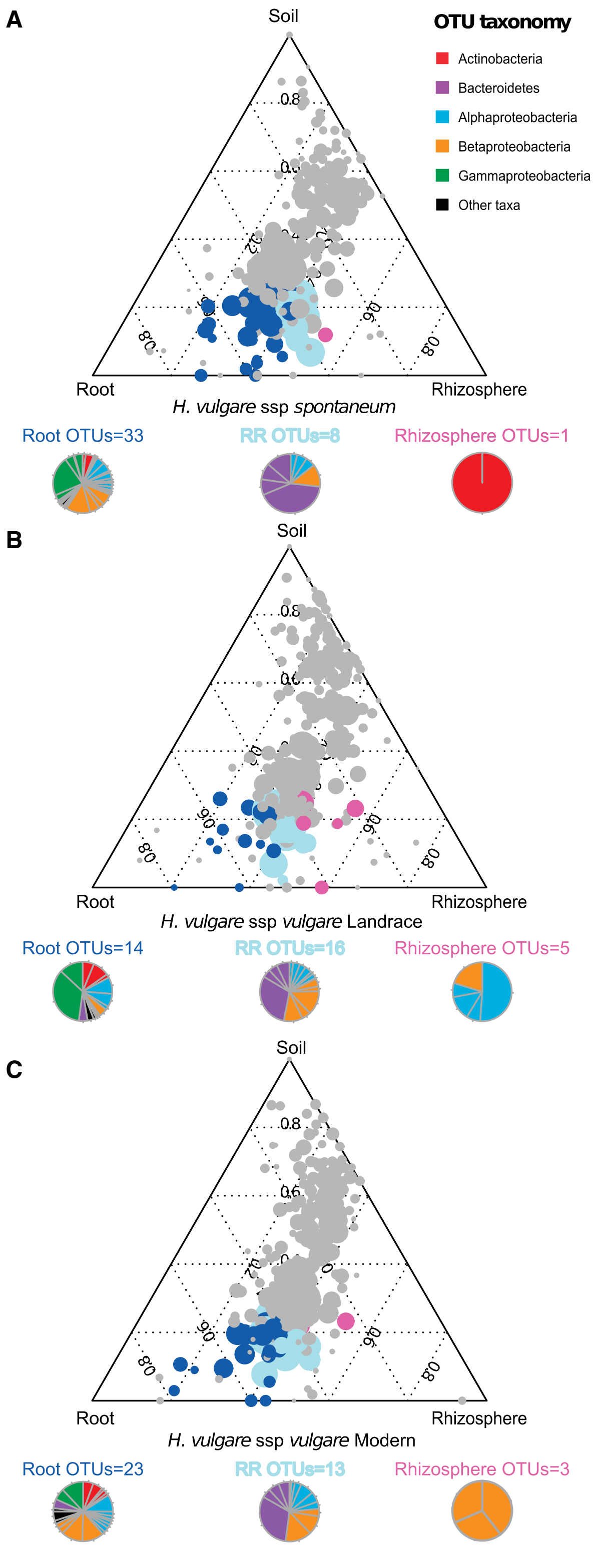


Figure 3. OTU Enrichment at the Barley Root/Soil Interface

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).

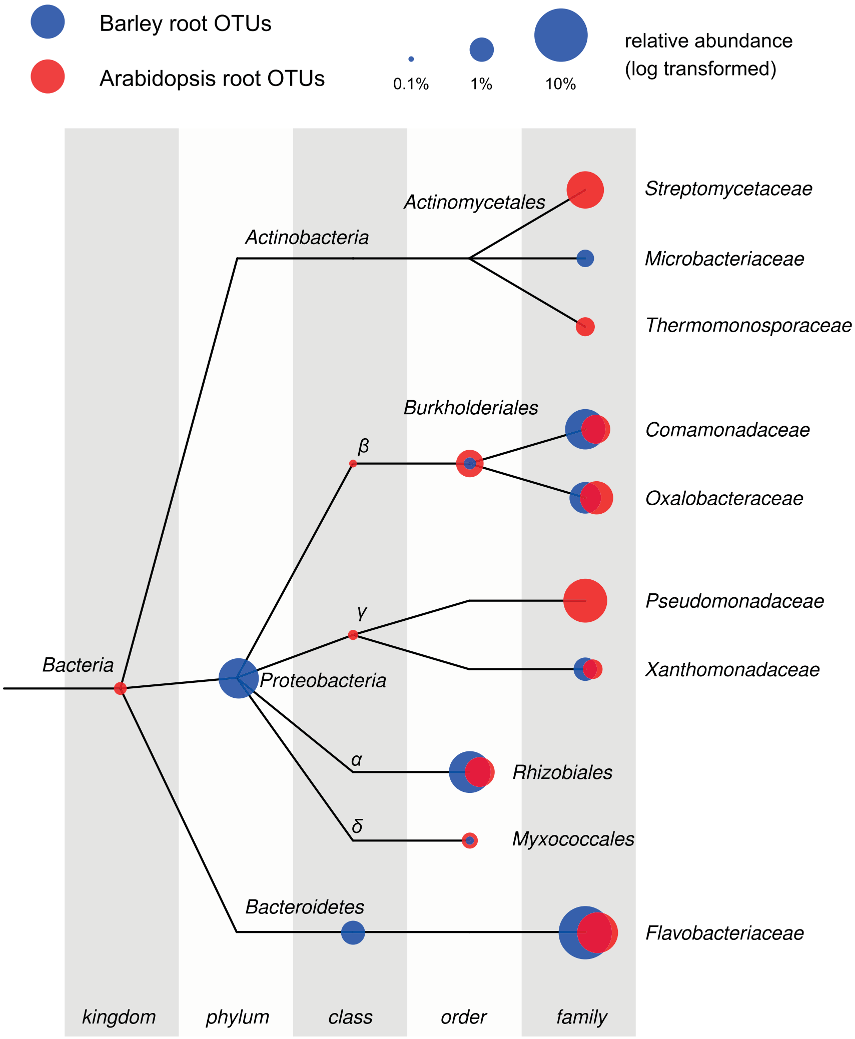


Figure 4. Taxonomic Representation of the Barley and *Arabidopsis* Root-Enriched Bacterial Taxa

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

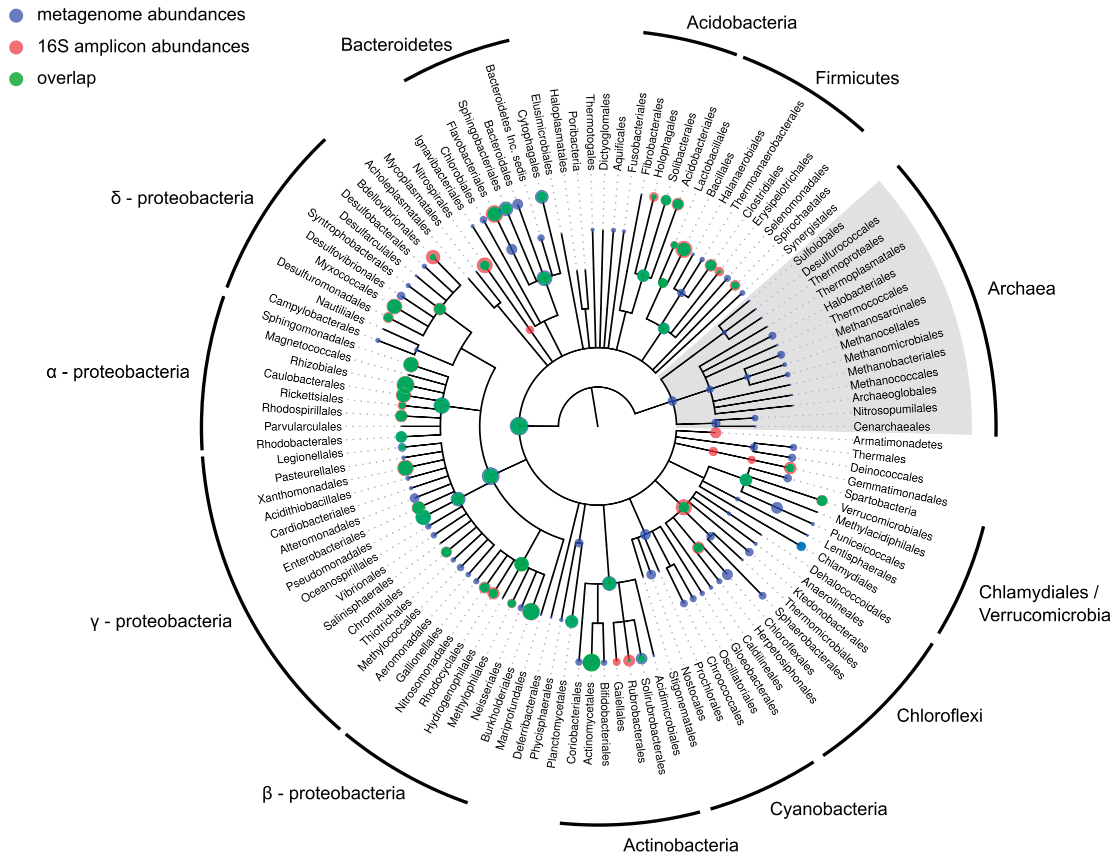


Figure 5. Comparison of 16S rRNA Amplicon and Metagenome Abundances

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

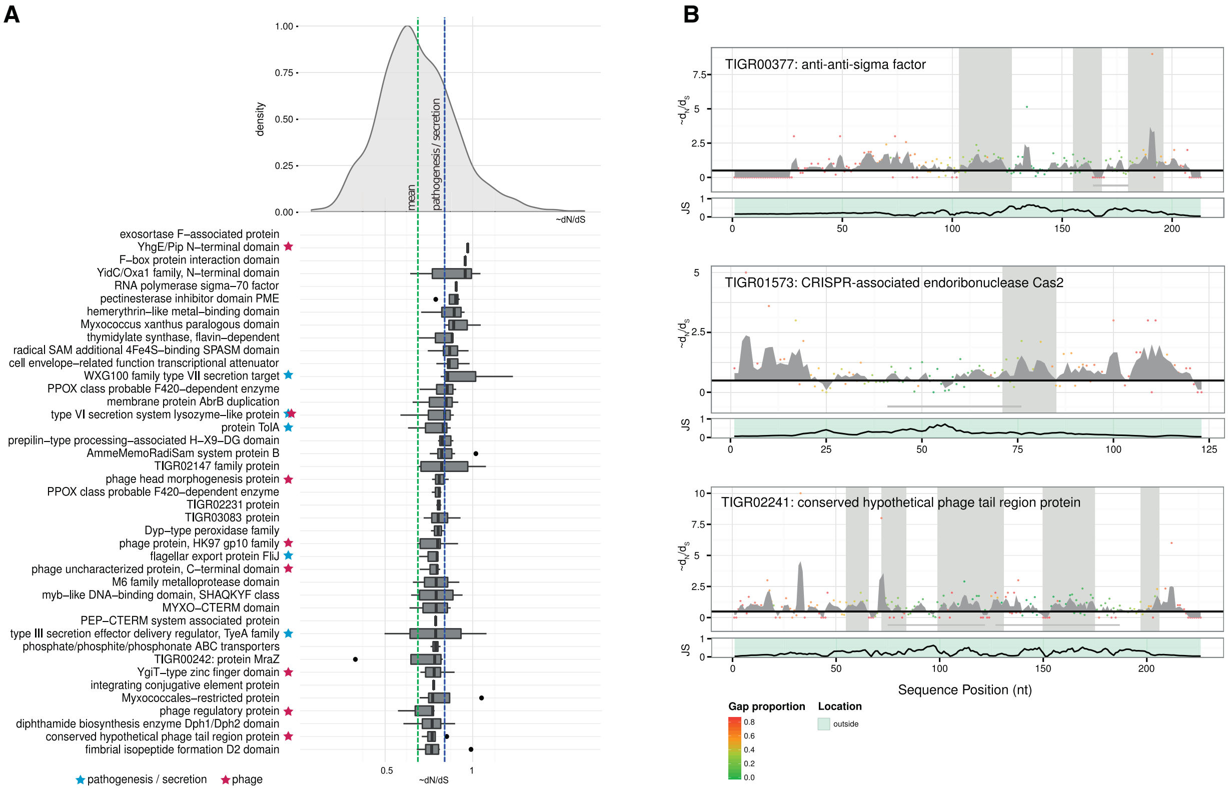


Figure 6. Proteins under Selection in the Barley Rhizosphere Microbiome

(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.

Microbiota and Host Nutrition across Plant and Animal Kingdoms

Stéphane Hacquard,[1,12] Ruben Garrido-Oter,[1,2,3,12] Antonio González,[4,12] Stijn Spaepen,[1,12] Gail Ackermann,[4] Sarah Lebeis,[5] Alice C. McHardy,[2,3,6,*] Jeffrey L. Dangl,[7,8,*] Rob Knight,[4,9,*] Ruth Ley,[10,11,*] and Paul Schulze-Lefert[1,3,*]
[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]Department of Pediatrics, University of California San Diego, La Jolla, CA 92093, USA
[5]Department of Microbiology, University of Tennessee, Knoxville, TN 37996-0845, USA
[6]Computational Biology of Infection Research, Helmholtz Center for Infection Research, 38124 Braunschweig, Germany
[7]Howard Hughes Medical Institute and Department of Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[8]Curriculum in Genetics and Molecular Biology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA
[9]Department of Computer Sciences and Engineering, University of California San Diego, La Jolla, CA 92093, USA
[10]Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853, USA
[11]Department of Microbiology, Cornell University, Ithaca, NY 14853, USA
[12]Co-first author
*Correspondence: alice.mchardy@helmholtz-hzi.de (A.C.M.), dangl@email.unc.edu (J.L.D.), robknight@ucsd.edu (R.K.), rel222@cornell.edu (R.L.), schlef@mpipz.mpg.de (P.S.-L.)
http://dx.doi.org/10.1016/j.chom.2015.04.009

Plants and animals each have evolved specialized organs dedicated to nutrient acquisition, and these harbor specific bacterial communities that extend the host's metabolic repertoire. Similar forces driving microbial community establishment in the gut and plant roots include diet/soil-type, host genotype, and immune system as well as microbe-microbe interactions. Here we show that there is no overlap of abundant bacterial taxa between the microbiotas of the mammalian gut and plant roots, whereas taxa overlap does exist between fish gut and plant root communities. A comparison of root and gut microbiota composition in multiple host species belonging to the same evolutionary lineage reveals host phylogenetic signals in both eukaryotic kingdoms. The reasons underlying striking differences in microbiota composition in independently evolved, yet functionally related, organs in plants and animals remain unclear but might include differences in start inoculum and niche-specific factors such as oxygen levels, temperature, pH, and organic carbon availability.

Physiological Functions of the Vertebrate Gut and Plant Roots

The vertebrate gut and plant roots evolved independently in animal and plant kingdoms but serve a similar primary physiological function in nutrient uptake (Figure 1). One major difference between plant and animal nutritional modes is their distinct energy production strategy. Plants are autotrophs, producing their own energy through photosynthesis (carbohydrate photo-assimilates), while animals rely entirely on the energy originally captured by other living organisms (heterotrophs). Long-distance transport mechanisms ensure the distribution of carbohydrate photo-assimilates from chloroplasts in leaves to all other body parts, including roots. Nutrient acquisition by roots to support plant growth is therefore almost exclusively limited to uptake of mineral ions and water from soil. In contrast, the mammalian gut has evolved to facilitate the uptake of simple sugars, amino acids, lipids, and vitamins in addition to ions. It is typically compartmentalized into sections with low microbial biomass in which the products of host enzymatic activity are absorbed (i.e., the human small intestine, SI) and a section for the uptake of microbe-derived fermentation products (human large intestine or hindgut, LI).

A significant fraction of the soil nutritive complement and of the dietary intake remains unavailable for plants and animals, respectively, and this defines their dietary constraints. Critical nutrients for plant growth and productivity in soil are nitrogen and phosphorus. However, plant roots can absorb only inorganic nitrogen and orthophosphate (Pi), although phosphorus is abundant in soil both in inorganic and organic pools. Pi can be assimilated via low-Pi-inducible (high-affinity) and constitutive Pi uptake systems (low-affinity) (Lambers et al., 2008; López-Arredondo et al., 2014). Plant species adapted to neutral or higher soil pH, and more aerobic soils have a preference for nitrate and deploy two nitrate uptake and transport systems that act in coordination. By contrast, plants adapted to low pH (reducing soil) as found in forests or the arctic tundra appear to assimilate ammonium or amino acids (Maathuis, 2009). Similarly, a fraction of normal human dietary intake remains undigested and therefore non-bioavailable (fiber). These non-digestible components include plant cell wall constituents such as cellulose, hemicellulose, xylan, and pectin, and certain polysaccharides such as β-glucan, inulin, and oligosaccharides that contain bonds that cannot be cleaved by mammalian hydrolytic enzymes (Tungland and Meyer, 2002).

Plant roots and animal guts are colonized by diverse microbial classes, including bacteria and archaea, fungi, oomycetes, as well as viruses (Table 1). These communities can be regarded as the host's extended genome, providing a huge range of potential functional capacities (Berendsen et al., 2012; Gill et al., 2006; Qin et al., 2010; Turner et al., 2013). Here we focus on

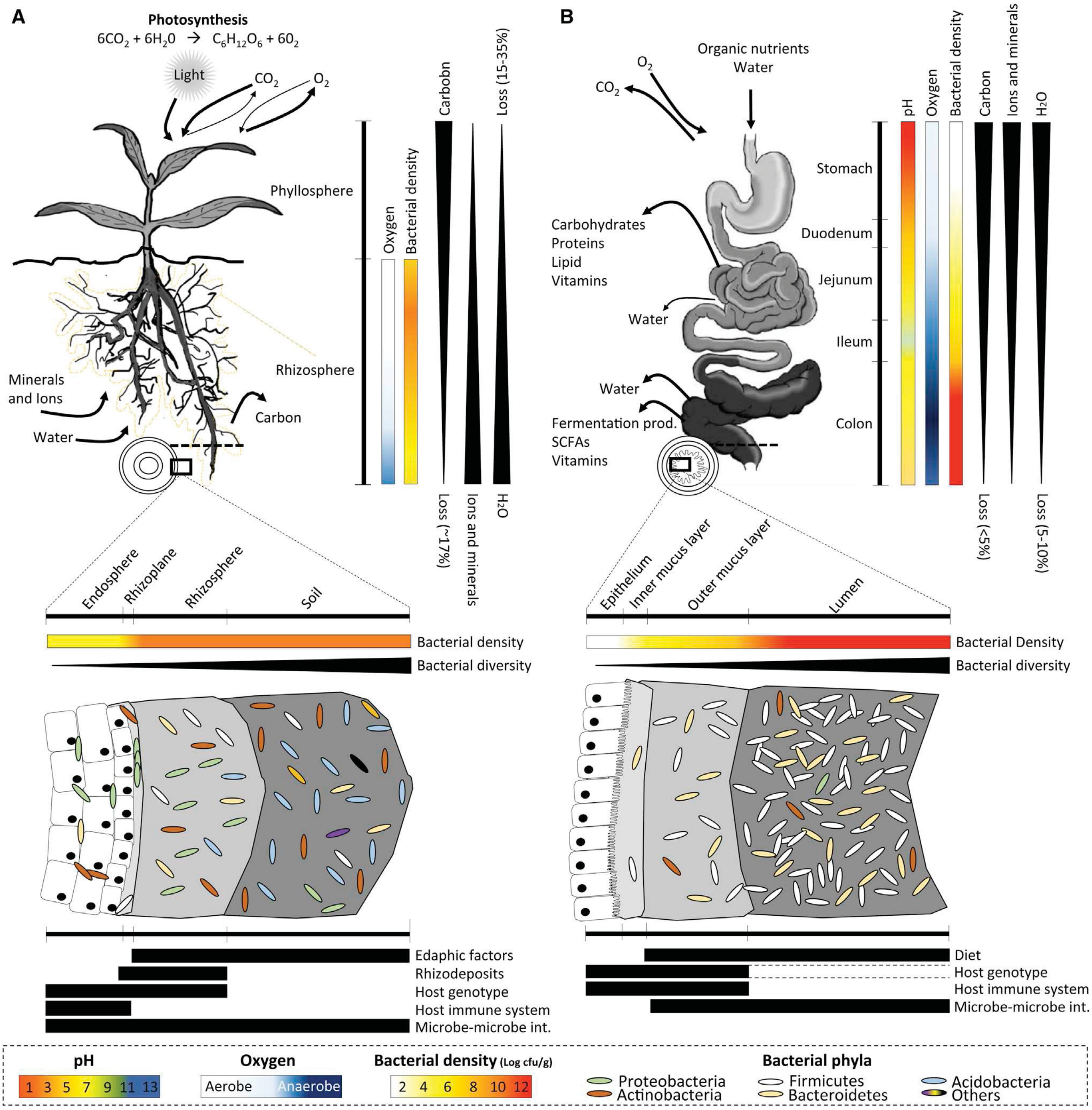


Figure 1. Physiological Functions of the Plant Roots and Human Gut in Nutrient Uptake, Spatial Aspects of Microbiota Composition, and Factors Driving Community Establishment

(A and B) Spatial compartmentalization of the plant root microbiota (A) and the human gut microbiota (B). Upper panels: the major nutrient fluxes are indicated, as well as pH and oxygen gradients in relation with the bacterial density. Lower panels: compartmentalization of the microbiota along the lumen-epithelium continuum in the gut or along the soil-endosphere continuum in the root. For each compartment, the bacterial density, the bacterial diversity, and the major represented phyla are represented for both the gut and the root organs. The main factors driving community establishment in these distinct compartments are depicted with black bars. The gut drawing is adapted from Tsabouri et al. (2014) with permission from the publisher.

bacterial microbiotas because these were shown to form reproducible taxonomic assemblies in animal and plant individuals with well-defined functions.

In plant roots, the microbiota mobilizes and provides nutrients by increasing nutrient bioavailability from soil (Bulgarelli et al., 2013). Non-nutritional functions include increased host tolerance to biotic stresses, e.g., against soil-borne pathogens (Mendes et al., 2011), and likely abiotic stresses. In addition, the root microbiota can also affect plant fitness by impacting flowering plasticity (Panke-Buisse et al., 2015; Wagner et al., 2014).

Table 1. Percentage of Shotgun Metagenome Reads Assigned to Each Kingdom of Life across Metagenome Studies

	Cucumber[a]	Wheat[a]	Soybean[b]	Wheat[c]	Oat[c]	Pea[c]	Barley[d]	Gut[e]
Bacteria	99.36	99.45	96	88.5	77.3	73.7	94.04	99.1
Archaea	0.02	0.02	<1	<0.5	<0.5	<0.5	0.054	
Eukaryotes	0.54	0.48	3	3.3	16.6	20.7	5.90	<0.1

[a] Ofek-Lalzar et al. (2014) (metagenomics of rhizoplane samples).
[b] Mendes et al. (2014) (metagenomics of rhizosphere samples).
[c] Turner et al. (2013) (metatranscriptomics of rhizosphere samples).
[d] Bulgarelli et al. (2015) (metagenomics of rhizosphere samples).
[e] Qin et al. (2010) (metagenomics of gut samples).

Similarly, the gut microbiota has a major role in host nutrition. It contributes nutrients and energy to the host via fermentation of indigestible polysaccharides into short-chain fatty acids (SCFAs) in the colon (Martins dos Santos et al., 2010; Tremaroli and Bäckhed, 2012). The human LI has incomplete peristalsis and a longer retention time, allowing fermentative microbiota to break down complex glycan bonds and liberate additional energy from the diet (Stevens and Hume, 1998). Additionally, gut microbiota provide essential vitamins to the host and modulate the absorptive capacity of the intestinal epithelium. An additional common feature of the gut and root microbiota is their protective role by competitive exclusion against invasion by opportunistic pathogens (Kamada et al., 2013).

Homeostatic balance between both microbe-microbe and host-microbe interactions is critical for a healthy host-microbiota relationship. Alteration of this balance via perturbation of the gut or the plant microbiota composition (microbial dysbiosis) may represent an important mechanism of disease (Martins dos Santos et al., 2010; Kemen, 2014; Sekirov et al., 2010). In plants, a healthy status is the norm, and soil-resident microbes contribute to plant health. This is illustrated by a higher disease severity following pathogen inoculation when plants are grown in pasteurized compared to non-pasteurized soils (Weller et al., 2002). In addition, so-called disease-suppressive soils protect plants against particular soil-borne pathogens. For example, specific bacterial genera belonging to gamma-Proteobacteria were associated with a high level of soil disease suppressiveness. The underlying mechanisms comprise competition between soil-borne microbes for plant-derived nutrients and antimicrobial compound production (Berendsen et al., 2012; Mendes et al., 2011). In the gut, commensal microbes can also suppress pathogen invasion through secretion of antimicrobial compounds, alteration of local pH, or stimulation of host immunity (Kamada et al., 2013).

Compartmentalization of the Gut and Root Microbiota

Relevant biotic and abiotic gradients exist in both the gut and root, leading to microbial compartmentalization (Figure 1). Along the soil-root continuum, four compartments can be distinguished: soil, rhizosphere, rhizoplane, and endosphere (Figure 1A). Bacterial diversity in soil is high, with estimates suggesting that >2,000 species populate 0.5 g of soil (Schloss and Handelsman, 2006). The rhizosphere corresponds to the zone of soil directly influenced by root exudation, while the root compartment can be separated in two distinct niches, rhizoplane and endosphere. The rhizoplane harbors a suite of microbes that tightly adhere to the root surface, while the endosphere is composed of microbes inhabiting the interior of roots. Microbial density is high in the rhizosphere, and species richness gradually decreases along the soil-endosphere continuum (Bulgarelli et al., 2012, 2015; Edwards et al., 2015; Lundberg et al., 2012) (Figure 1A). Therefore, the bacterial community shifts from a dense and diverse soil-borne community to a host-adapted community with reduced diversity.

A spatial heterogeneity of microbial density exists along the digestive track (Stearns et al., 2011). Densities are lowest in the stomach and duodenum (proximal SI) (10^1–10^3 bacteria per gram of content) and increase along the length of the SI with a higher density in the distal ileum (10^4–10^7 bacteria per gram). Cell densities in the LI can reach 10^{12}–10^{13} bacteria per gram of content, representing the highest density recorded so far in any environment and exceeding the density detected in the rhizosphere by 2–3 orders of magnitude. Although the density is high, the diversity is relatively low (Stearns et al., 2011; Walter and Ley, 2011). Using low-error 16S rRNA gene sequencing (LEA-seq) of the human fecal gut microbiota (low depth coverage), the number of bacterial species is estimated at 101 ± 27, which is in alignment with estimates of culture-based techniques (Faith et al., 2013; Mitsuoka, 1992). Compartmentalization exists also from the inside to the outside of the intestinal tube, defined by the intestinal lumen, mucus, and epithelial surface. Similar to the compartmentalization in the root, a decrease in bacterial density is observed from the lumen to the epithelial surface (Swidsinski et al., 2005, Van den Abbeele et al., 2011; Zhang et al., 2014) (Figure 1B). In the LI, the mucus is subdivided into an inner firmly adherent layer largely devoid of bacteria and an outer layer that is looser and non-adherent and allows some microbial colonization (Johansson et al., 2008).

Community Structure of the Vertebrate Gut and Plant Root Microbiota

Where Do They Come from?

A relevant difference for experimentation on the plant root and vertebrate gut microbiota is the ease with which the start inoculum of the root microbiota can be defined. This is due to a predominant horizontal acquisition of root endophytes from the surrounding soil biome, although in some plant species there is evidence for additional vertical transmission of seed-borne endophytes (Barret et al., 2014). These endophytes mainly belong to Proteobacteria and can colonize seeds via different colonization routes, including flowers, fruits as well as roots, leaves, and stems (Truyens et al., 2015). Even though vertical transmission in mammals is not as explicit as in plants (none are transferred with the germline), vertical transmission nevertheless occurs. The transmission from parent to offspring results from the birth process itself, from milk, and from the close contact that comes from parental care (Unger et al., 2015). In humans, vaginal birth inoculates the newborn with a set of strains that can be matched to the mother, whereas caesarean section results in colonization with skin microbes originating from various caregivers (Dominguez-Bello et al., 2010). Breast milk is also an important source of microbiota and antibodies that shape the gut microbiome (Newburg and Morelli, 2015), and introduction of solid foods brings rapid shifts in the bacterial community composition toward an adult-like microbiome (Koenig et al., 2011). Vertical transmission from mother to infant gut microbiota is sometimes behaviorally increased in mammals by feeding mother's fecal matter to their infants. In koalas, for instance, this transmission is believed to participate in the digestion of eucalyptus (Osawa et al., 1993). Additionally, group living is known to aid the transmission of commensal microbes between members of family groups (humans), troupes (primates), and most likely herds as well. Co-habitation in humans leads to sharing of microbiota, which is enhanced when dogs also co-habit in the same house (Song et al., 2013). Ironically, hygiene measures aimed at reducing pathogen transmission may have had broad negative impacts on the transmission of commensals and may underlie the loss of diversity observed in the West (Blaser and Falkow, 2009).

Who Are They?

Despite the vast prokaryotic biodiversity found in the biosphere (currently >80 bacterial phyla are described), the host-associated microbiota is dominated numerically by a few phyla. The rhizosphere and the root endophytic compartment of unrelated plant species is often enriched for bacteria belonging to three main phyla (Proteobacteria, Actinobacteria, and Bacteroidetes). In contrast, abundant soil bacteria belonging to the phylum Acidobacteria are excluded from the endophytic compartment (Bulgarelli et al., 2013). Compared with the surrounding soil, microbiota members belonging to the phylum Proteobacteria are consistently enriched in the rhizosphere/endosphere compartments of monocotyledonous and dicotyledonous plants, including perennial and annual plants (Bulgarelli et al., 2012, 2015; Edwards et al., 2015; Lundberg et al., 2012; Ofek-Lalzar et al., 2014; Peiffer et al., 2013; Schlaeppi et al., 2014; Shakya et al., 2013; Zarraonaindia et al., 2015). This likely reflects niche adaptation (nutrient availability, oxygen levels) and the ability to efficiently invade and persist inside or outside the roots of divergent plant species. Firmicutes and Bacteroidetes are by far the two most-abundant phyla detected in adult human and mouse feces. Other phyla represented include the Actinobacteria, Verrucomicrobia, and a number of less-abundant phyla such as the Proteobacteria, Fusobacteria, and Cyanobacteria (Eckburg et al., 2005). Similar to the rhizosphere compartment, the mucus layer of the gut represents a particular niche favoring the proliferation of specialized inhabitants. It has been estimated that at least 1% of the gut microbiota can degrade mucins as a source for carbon and nitrogen (Hoskins and Boulding, 1981). Select types of bacteria can also attach to mucins, such as *Bifidobacterium bifidum*, which has the ability to stimulate mucin production via butyrate-induced expression of *MUC2*, while others can degrade the nine-carbon sugar sialic acid found in host glycoconjugates (Almagro-Moreno and Boyd, 2009; Gaudier et al., 2004; Leitch et al., 2007).

Are There Structural Similarities across Diverse Host-Associated Microbial Communities?

Striking physiological (dis-)similarities exist between organs dedicated to nutrient acquisition in hosts belonging to different taxonomic lineages. However, the extent to which microbial communities living in association with phylogenetically divergent hosts overlap with each other is largely unknown. In an attempt to unravel host-specific and conserved signatures in the microbiota, we retrieved and re-analyzed the raw sequencing data contributed by 14 previous large-scale 16S rRNA gene survey studies (Table S1). These comprise >3,200 samples from more than 40 different host species, including human, other mammals, and fish gut, as well from the root and rhizosphere of the flowering plant *Arabidopsis thaliana* and relative species, maize, rice, barley, and grapevine. In addition, we included samples from several species of cnidarian hydra, a freshwater basal animal featuring a gut forming a hollow cavity within the body with one opening, the mouth.

To analyze the data, we followed the QIIME (Caporaso et al., 2010) closed-reference protocol and used SortMeRNA (Kopylova et al., 2012) to cluster the sequences into operational taxonomic units (OTUs) at 97% sequence similarity (see Supplemental Experimental Procedures). Analyses of beta-diversity using principal coordinate analysis (PCoA) revealed a clear clustering of samples according to their respective host species (Figure 2A; Supplemental Experimental Procedures). Although all samples are derived from organs with a dedicated function in nutrient uptake, we found striking qualitative differences between their associated microbial communities. This disparity can be explained by the increased abundance of members of the Bacteroidetes phylum in the mammalian stool samples (particularly those belonging to the orders Bacteroidales and Clostridiales) and the enrichment of members of the families Pseudomonadaceae, Streptomycetaceae, and Comamonadaceae in the rhizosphere and plant root compartments (Figure 3). Intriguingly, the bacterial communities in the fish gut are more closely related to those in the root and rhizosphere samples than to the mammalian gut, partially due to an increased abundance in Proteobacteria (45.08% and 54.44% in root-associated samples and fish gut, respectively, compared with 4.20% in the case of the human gut; Figure 4). In addition, the microbial communities from infant gut (from Koenig et al., 2011) are more closely related to those of plant roots (and therefore soil microbiota) than those associated to adults (Figure S1). Together, this suggests that shared environmental and physiological features, rather than phylogenetic relatedness of the hosts, are decisive for community establishment.

Analysis of alpha-diversity (Figures 2B and 3B) shows that the bacterial richness is low in the gut of aquatic organisms and higher in the root and in the rhizosphere of different plant species, consistent with the bacterial diversity detected in their respective surrounding environments (aquatic versus soil environments; Curtis et al., 2002). For all plant species surveyed, the bacterial diversity is lower in the endosphere compartment (root) compared to the rhizosphere compartment (Figures 2B), in concordance with previous studies (Bulgarelli et al., 2012;

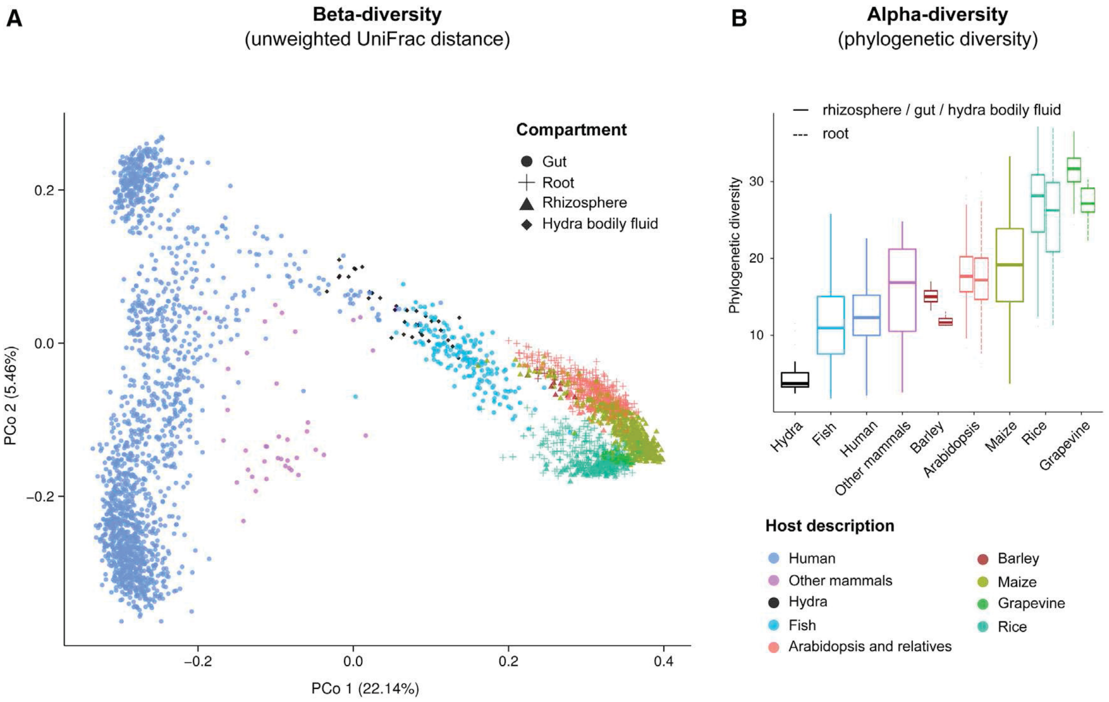


Figure 2. Alpha- and Beta-Diversity Analyses
(A) Principal coordinate analysis (PCoA) of pairwise unweighted UniFrac distances between samples. The color and shape of each point represent the host and compartment, respectively.
(B) Comparison of alpha-diversity between hosts based on the whole tree phylogenetic diversity index (PD), sorted by ascending order of complexity. See Table S1 for more information about the individual host species included in each study.

Edwards et al., 2015; Lundberg et al., 2012). The extent of this gradient in diversity, as well as the differentiation between the two compartments, appears to be dependent on the plant species, indicating a strong host-dependent effect on community establishment.

A phylogenetic comparison of the abundant community members across hosts (OTUs, with a relative abundance higher than 0.1% on average) reveals clear qualitative structural differences between mammalian gut and plant root and rhizosphere samples (Figure 5). These distinct sets of bacterial communities show virtually no overlap even at high taxonomic levels. Samples obtained from human and mammalian guts are dominated by OTUs belonging to the orders Bacteroidales and Clostridiales (34.55% and 51.26% relative abundances, respectively), while these are almost completely absent in the root and rhizosphere samples (0.70% and 0.80%, respectively). This striking difference in community composition in independently evolved, yet functionally related, gut and root organs might be explained by adaptations to specific host and environmental needs, including niche-specific factors such as oxygen levels, pH, and organic carbon availability. Our findings also make a direct transfer and persistence of microbiota members from numerous root-derived dietary plant products in the human gut unlikely.

Do They Fluctuate over Time?

Despite the fact that infancy or the seedling stage for plants are critical windows for microbiota assembly, very little is known about the earliest steps driving host colonization by pioneer bacteria. Assembly of the infant gut microbiome begins at birth (early reports described it as chaotic), and diversity levels slowly increase until ∼2–3 years of age (Koenig et al., 2011; Palmer et al., 2007; Yatsunenko et al., 2012). Sampling from birth to 2.5 years of age revealed the following: (i) community richness increased gradually over time, (ii) the use of antibiotics, changes in diet, and infections led to jumps from one stable consortium of species to another, and (iii) members of the Bacteroidetes phylum were co-dominant with members of the Firmicutes phylum after the introduction of solid foods (Koenig et al., 2011). The adult-like microbiota is characterized by a greater stability (David et al., 2014a; Spor et al., 2011). About 60% of the bacterial strains in the intestine are detected over a 5-year time frame, and Bacteroidetes and Actinobacteria were identified as the most stable phyla (Faith et al., 2013). In contrast to the chaotic microbial succession described for the infant gut, the structure of the root microbiota during the plant life cycle appears rather stable. Despite a higher variability observed during the seedling stage (Chaparro et al., 2014), microbiota acquisition from soil appears to occur relatively rapidly, initiating within 24 hr after sowing and approaching a steady state within 2 weeks (Edwards et al., 2015). Once established, there is little evidence for dramatic changes even late in the life cycle of annual *A. thaliana* plants, when organic carbon and nitrogen are spatially re-allocated during the transition from vegetative to reproductive growth for seed formation (Lundberg et al., 2012). This surprising stability might be explained by the sessile nature of plants, together with a rather stable soil-borne inoculum source, which prevents extreme fluctuations in input communities throughout

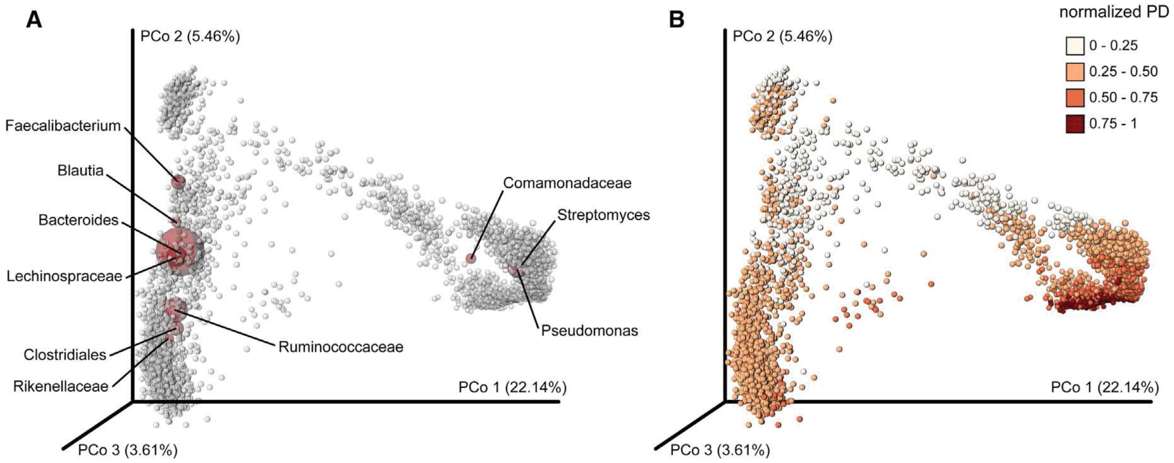


Figure 3. 3D PCoA Plots

(A) Biplots depicting the taxa with the largest contribution to the ordination space (order Clostridiales; families Ruminococcaceae, Rikenellaceae, Lechinospraceae, Comamonadaceae; genera Streptomyces, Pseudomonas, Bacteroides, Blautia, Faecalibacterium).

(B) PCoA plot showing the alpha-diversity variation as measured by the PD index across all samples included in the study.

a rapid annual plant's life cycle. Whether this also applies to longer-lived perennials and to repeated croppings of the same species at the same location remains to be further substantiated (Donn et al., 2015).

Major Factors Driving Community Establishment and Composition

Inter-individual differences in the gut and the plant microbiota are likely to be dictated by many modulating factors, including environmental parameters but also diet/soil-type, microbe-microbe interactions, host genotype, and host immune system (Figure 1).

Environmental Factors

pH. Bacterial community composition is strongly correlated with differences in soil pH, with soils at near-neutral pH showing the highest microbial diversity (Fierer and Jackson, 2006). Roots can acidify the rhizosphere up to two pH units compared to the surrounding soil through release of protons, bicarbonate, organic acids, and CO_2 (Hinsinger et al., 2003). Along the digestive tract, the increase in bacterial titer can be attributed to several factors, such as pH and bile acids. The pH is very low in the stomach (pH 1.5–5), restricting bacterial growth, increases in the SI (duodenum pH 5–7, jejunum 7–9, ileum 7–8) and drops in the colon (pH 5–7) (Walter and Ley, 2011) (Figure 1B). Many types of bacteria, in both the gut and the soil, are sensitive to pH, and this is thought to structure communities to a large degree (Duncan et al., 2009), although it is difficult to disentangle the exact contribution of pH on the overall community structure due to likely interaction with many other factors.

Oxygen. Although both gut and root systems are dedicated for nutrient uptake, O_2 levels are controlled in opposing directions. In the vertebrate gut, luminal microbes generally face anaerobic conditions favoring fermentative metabolism, while in soil and along the root (micro-)aerobic conditions are found (Figure 1). This might be a major factor explaining structural and functional differences between the microbiota of the vertebrate gut and plant roots (Figure 5). The gut microbiota of healthy individuals is dominated by anaerobic bacteria, which outnumber aerobic and facultative anaerobic bacteria by a factor of 100–1,000:1 (Quigley and Quera, 2006), while the root microbiota is enriched for Proteobacteria, a phylum dominated by aerobic species. Consistent with this, genes encoding high-affinity oxidases that use O_2 as a terminal electron acceptor are overrepresented in gut metagenomes, whereas those encoding low-affinity oxidases are enriched in soil metagenomes (Morris and Schmidt, 2013). It is arguably in the host's interest to limit respiration, because (i) limiting respiration will control bacterial growth and (ii) promoting fermentation will result in SCFA availability. Nonetheless, there is a biologically relevant gradient of oxygen levels in both the soil and the gut that is likely to influence microbial community structure at the micro-levels. Despite the fact that plant roots generally face (micro-)aerobic conditions, soil O_2 levels can also fluctuate as a function of soil wetting/drying (Noll et al., 2005), with anoxic niches in the center of soil aggregates. Similarly, a higher O_2 concentration is found at the surface of the epithelium compared with the lumen. Some facultative aerobes can grow along this oxygen gradient by respiring O_2 close to the epithelium using flavins and thiols as electron shuttles to respire at ''long distance'' (Khan et al., 2012).

Temperature. While thermal stability exists in the gut of mammals (endotherm), higher temperature fluctuation is observed for plants or ectothermic animals that rely on the external temperature to regulate their internal body temperature. It has been reported that the bacterial community in soil is modulated by temperature (Bárcen as-Moreno et al., 2009), although plant microbiota functions must remain stable under a wide range of temperatures.

Nutritional Drivers

For both plant roots and vertebrate guts, diet (for plants, soil type defines the diet) is a major driver for microbial community structure (Bulgarelli et al., 2012; Cotillard et al., 2013; Carmody et al., 2015; David et al., 2014b; Edwards et al., 2015; Ley et al., 2008a; Lundberg et al., 2012; Muegge et al., 2011; Schlaeppi et al., 2014; Peiffer et al., 2013; Turnbaugh et al., 2009).

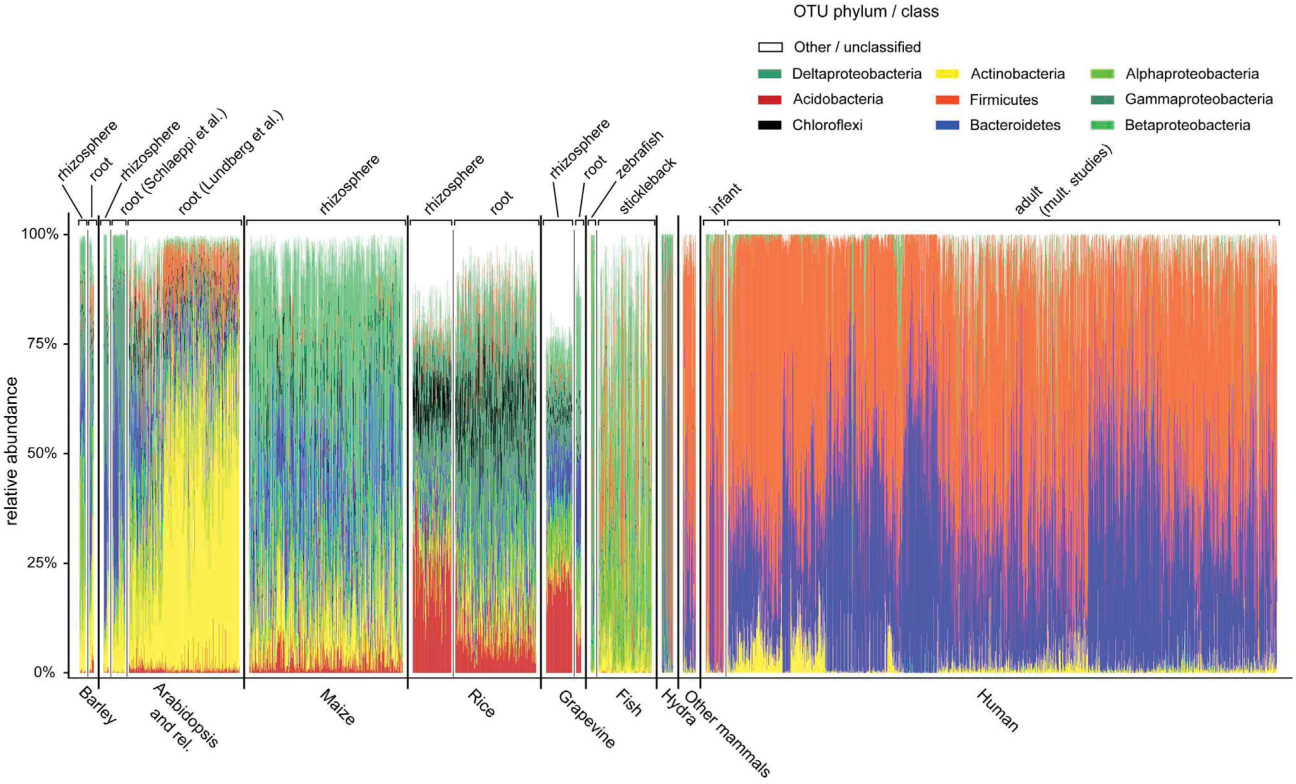


Figure 4. Cumulative Abundance Plots

Relative abundances grouped at the phylum or class taxonomic level for each sample included in the meta-analysis. The bar plots have been arranged along the x axis separating different host groups as well as different species and compartments.

Organic carbon is widely considered to be the most important factor limiting bacterial growth in different soils (Demoling et al., 2007). Isotope probing experiments using different plant species revealed that an average of 17% of all photosynthetically fixed carbon is transferred to the rhizosphere through root exudates (Nguyen, 2003), highlighting a considerable organic carbon deposition in soil. Low molecular weight carbon substrates such as dicarboxylic acids, exuded by roots in large quantities to acidify the rhizosphere, also enhance the availability of Pi and micronutrients such as manganese, iron, and zinc. These dicarboxylic acids are an important driver mediating soil community shifts, leading to an increase in the relative abundance of beta-Proteobacteria, gamma-Proteobacteria, and Actinobacteria (Eilers et al., 2010).

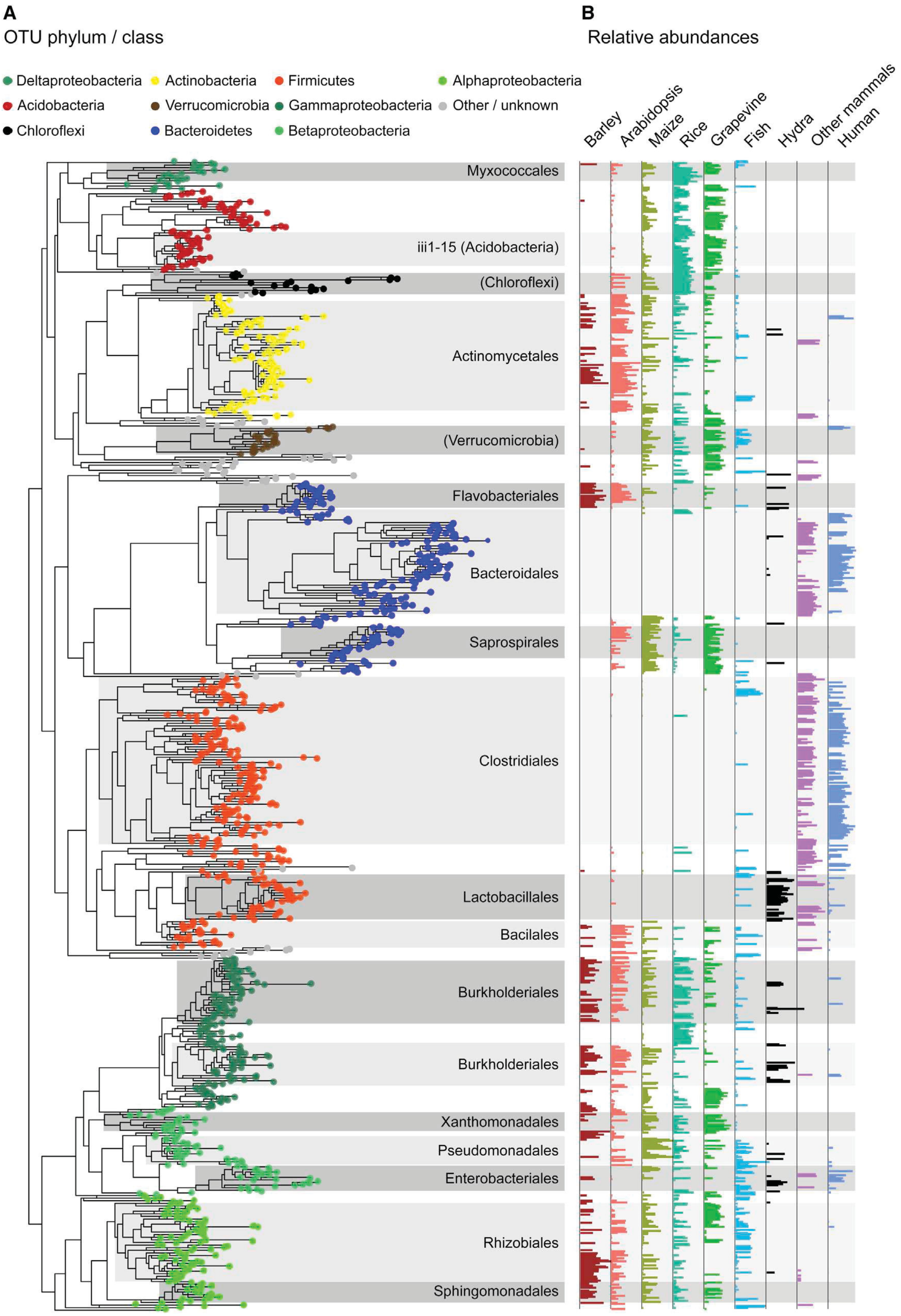
The evolution of the mammalian gut microbiota has been greatly influenced by host diet. Mammals, their gut microbiota, and their diet types are part of a dynamic tripartite coevolution (Ley et al., 2008b). The majority (80%) of extant mammals are herbivorous, which stands in contrast to the early mammals that were most likely carnivorous based on their tooth morphology. The rise in herbivory could only have been accomplished with the necessary changes in gut microbes, since mammalian genomes lack the necessary genes encoding plant cell wall degrading enzymes. Comparisons of microbiomes between host species highlight the specific adaptations of the microbiota to the host diet, such as an increased abundance of genes encoding the necessary enzymes and their respective pathways (Eilam et al., 2014), as exemplified in a comparison between the termite hindgut and the bovine rumen metagenome (Brulc et al., 2009). The latter is enriched for genes encoding glycoside hydrolases, cellulosome enzymes, and nitrogen-related uptake proteins. In contrast, the termite hindgut microbiome showed an enrichment for genes involved in the degradation of the cellulose backbone and nitrogen fixation. This clearly reflects the differences in diet of the hosts (forages and legumes versus nitrogen-poor wood).

Microbe-Microbe Interactions

The role of microbe-microbe interactions is also critical for shaping microbiota structure in both plant and animal systems (Bulgarelli et al., 2015; Fraune et al., 2014; Hacquard and Schadt, 2015; Trosvik et al., 2010). The combination of synergistic, beneficial, and antagonistic interactions among microbiota members colonizing the gut and plants is likely to have a major impact on overall community structure. Therefore, individual members of a community may contribute to the overall stability of the system, and consequently, each community member must be viewed as a potential internal driver of microbial community assemblage. Microbial co-occurrence and co-exclusion patterns are now emerging as important concepts for understanding the rules guiding microbial community assembly (Cardinale et al., 2015; Faust et al., 2012; Zhang et al., 2014)

Host Genotype

Intra-species plant genetic diversity explains less variation in community structure than soil type and root fraction (soil,



(legend on next page)

rhizosphere, and endosphere). Surveys of the bacterial community structure of 27 maize inbred lines, 6 cultivated rice varieties, 3 barley accessions, and several *A. thaliana* accessions each point to a small (~5%–6% of variation) but significant role of the host genotype on community composition (Bulgarelli et al., 2012, 2015; Edwards et al., 2015; Lundberg et al., 2012; Peiffer et al., 2013; Schlaeppi et al., 2014). This suggests a link between host diversification and microbial community establishment (see below).

In humans, family members are often observed to have more similar microbiotas than unrelated individuals (Tims et al., 2013; Turnbaugh et al., 2009; Yatsunenko et al., 2012). Familial similarities are usually attributed to shared environmental influences, such as dietary preference, a powerful shaper of microbiome composition (Cotillard et al., 2013; David et al., 2014b; Wu et al., 2011). However, host genetics also play a small but statistically significant role in shaping the composition and structure of the gut microbiome. Studies comparing microbiota between human subjects differing at specific genetic loci have shown gene-microbiota interactions (Khachatryan et al., 2008; Rehman et al., 2011). A more general approach to this question has linked genetic loci with abundances of gut bacteria in mice (Benson et al., 2010; McKnite et al., 2012), although diet effects outweigh the host genotype effects (Parks et al., 2013). In humans, earlier twin studies failed to reveal significant genotype effects on microbiome diversity (Turnbaugh et al., 2009; Yatsunenko et al., 2012). However, a recent report by Goodrich et al. (2014) comparing monozygotic (MZ) with dizygotic (DZ) twin pairs identified specific taxa as heritable (i.e., the variability in the relative abundances of these taxa across the population was partially driven by host genotype variation). These taxa include health-associated Faecalibacterium and Bifidobacterium and lean phenotype-mediating Christensenella (Goodrich et al., 2014).

Host Immune Systems and Microbiota Homeostasis

Plants and animals each engage structurally related pattern recognition receptors (PRRs) for recognition of evolutionarily conserved non-self microbial structures (i.e., lipopolysaccharides [LPS], lipopeptides, flagellin, chitin) at the cell surface, and activation of these is typically sufficient to halt microbial proliferation. However, successful plant and animal pathogens have evolved mechanisms to dampen or escape PRR-mediated host responses to foster virulence. In response, members of the NLR (nucleotide-binding domain leucine-rich repeat containing) family of intracellular immune receptors in plants and animals are activated by the action of pathogen virulence factors or by direct binding of the virulence factors themselves (Boller and Felix, 2009; Jones and Dangl, 2006; Maekawa et al., 2011). Active animal PRRs and NLR inflammasomes each can instruct the mammalian adaptive immune system and cause spatially dispersed response in plants, as detailed below.

Detection of microbial patterns via PRRs constitutes the first layer of immunity in plants and animals and triggers a variety of output responses. In animals, these include instruction and either activation or suppression of the adaptive immune system via cytokine signaling and cell migration to and from infection sites and lymphoid organs. Because there are no circulating cells in plants, PRR- and NLR-dependent signaling can lead to differential local and systemic signals that result in adequate defense outputs at and directly surrounding the site of infection and a poised defense in distal organs. Analogous to cytokines, plants deploy a handful of defense phytohormones that have variable domains of signaling and instruct cells neighboring an infection site, and even systemically to distal organs, to be ready to respond to infection (Pieterse et al., 2012).

The lack of circulating immunocytes also demands that each plant cell in an organ be capable of recognizing all pathogens adapted to that organ. This drives a complicated requirement for coordination of normal cellular functions, mediated by growth-regulating hormones, and immune output mediated by the defense phytohormones. This coordination is manifested as trade-offs between growth and immunity (Belkhadir and Jaillais, 2015). Thus, systemic acquired resistance in above-ground organs is triggered by biotrophic pathogens and mediated by salicylic acid (SA), while induced systemic resistance, also active in leaves, is triggered in roots by rhizobacteria and is mediated by jasmonic acid (JA) and ethylene (Spoel and Dong, 2008; van Loon et al., 1998).

Because plant defense phytohormones are key signaling molecules between microbial perception and immune system outputs, their production and perception are common pathways targeted by both potential pathogens and beneficial microbes. Hence, there is evidence that during the early stages of colonization both arbuscular mycorrhizal (AM) and Rhizobium species locally suppress SA signaling (García-Garrido and Ocampo, 2002; Stacey et al., 2006), suggesting that defense phytohormones normally act to inhibit microbial survival in the root. Indeed, culture-dependent studies in *A. thaliana* have demonstrated a significantly lower load of culturable bacteria in rhizospheres of plants with either defective JA signaling or, conversely, constitutive SA production (Doornbos et al., 2011). Beyond defense phytohormones, other immune outputs have also been implicated by recent studies. In particular, metagenomic studies in rice uncovered genes present in root endophytic bacteria, notably detoxification of reactive oxygen species (Sessitsch et al., 2012).

The overall structure of the *Arabidopsis* root microbiota remains largely robust to host mutations leading to hypo- or hyper-immunity. However, sets of mutants with altered defense phytohormone biosynthesis and/or perception had specifically altered root microbiome taxonomic compositions compared to wild-type. These alterations were congruent with the known effects of the mutants on immune system outputs in leaves. Experiments using both wild soil and its natural community or synthetic soil microcosms in the presence of a synthetic bacterial community demonstrated that SA and/or SA-dependent processes are major contributors to root microbiome

Figure 5. Phylogenetic Analysis of OTU Abundances

(A) Phylogeny inferred from the representative sequences of all OTUs that had at least 0.1% relative abundance on average for all samples of a host species (1,133 in total). The color of each leaf depicts the taxonomic classification of its corresponding OTU.

(B) Average relative abundances of abundant OTUs across all samples of each host (log-transformed). Note that in the case of plant hosts, abundances are averaged across all root compartments.

composition (S.L., unpublished data). Together, these studies represent some of the insights into mechanisms used by the plant immune system to shape its microbiota.

In the animal gut, a first line of defense consists of the secretion of antimicrobial peptides that are produced deep within the crevices of the epithelial layer, in the crypts between the villi. While some antimicrobial agents are continuously secreted, others are secreted in response to bacterial triggering of specific PRRs (Toll-like receptors, TLRs) on the epithelial cell surfaces. The mucus layer is crucial to prevent systematic activation of these immune responses. When the inner mucus layer is removed chemically (i.e., with dextran sodium sulfate [DSS]) or through gene mutation (*MUC2* mutants), bacteria come into contact with epithelial cells and cause an inflammatory response (Johansson et al., 2010; Van der Sluis et al., 2006). In contrast to plants, the adaptive immune system also plays a role for sequestering symbiotic bacteria in the lumen through the secretion of immunoglobin A (IgA) that target epitopes of intestinal bacteria. Like the antimicrobial activity of the innate immune system, the adaptive immune system can be regulated in parts by TLR signaling (Iwasaki and Medzhitov, 2010). Together, the adaptive and innate immune systems have mechanisms for detecting surface-associated bacteria and work together to reduce inflammation. Because the adaptive immune system is (largely) unique to vertebrates, and based on the observation that vertebrates, notably mammals, harbor microbial communities with much greater complexity than do invertebrates, McFall-Ngai et al. (2013) have proposed that the adaptive immune system itself is important in the shaping and maintenance of high microbial diversity.

Co-diversification of Host-Microbe Communities

By comparing the bacterial communities associated with maize genotypes or other grasses, a significant correlation between rhizobacterial communities and the host phylogenetic distance has been detected, suggesting that the host's evolutionary history can be a good predictor of root microbiota structure (Bouffaud et al., 2014). A comparison of inter-species host phylogeny and microbiota diversification in four Brassicaceae plant species, including *A. thaliana*, which diverged ~35 Ma revealed only quantitative differences. This diversification cannot be explained solely by the phylogenetic distance of these hosts but likely includes plant species-specific ecological adaptations (Schlaeppi et al., 2014). However, qualitative differences can be observed when comparing more distantly related plant species such as *A. thaliana* and barley (dicotyledonous versus monocotyledonous plants), which diverged ~150 Ma (Bulgarelli et al., 2015). Marked differences in microbiota composition were also reported for *Hydra vulgaris* and *Hydra oligactis*, cnidarian animal groups that diverged approximately 100 Ma and have been cultivated under identical laboratory conditions for decades (Franzenburg et al., 2013).

In mammals, similarities in microbial community composition between members of the same species raise the question of whether the bacterial communities track mammalian phylogeny. This would be expected if the bacteria are passed vertically from parent to offspring, which some mammal species encourage behaviorally. Patterns of relatedness of the bacterial communities were compared to the mammalian phylogeny (Ley et al., 2008a). For subsets of the mammalian phylogeny, the trees matched at a rate that is greater than expected by chance. For instance, this pattern was observed in the case of bears, which are an animal group candidate for mother-offspring transmission due to prolonged contact between the cub and the mother, implying that an ancestral microbial population diversified at the same time that bears speciated. A comparison of the microbial communities associated to great ape species, including *Homo sapiens*, also revealed that the host species phylogeny was congruent to the pattern of relatedness of their gut microbial communities, which diverged in a manner consistent with vertical inheritance (Ochman et al., 2010). However, a comparative analysis of the gut microbiota of humans with the ape species indicates an accelerated change in the microbiota composition of humans that cannot be explained by evolutionary distance (Moeller et al., 2014). A recent study of one isolated Amazonian tribe revealed the highly diverse gut microbiota, in both composition and functions, including a broad range of antibiotic resistance genes, suggesting that the Western lifestyle has dramatically reduced bacterial diversity (Clemente et al., 2015).

Taken together, these data indicate generally that a correlation between microbiota and host phylogeny can be explained by co-diversification from common ancestors. Nonetheless, the hugely different generation times of bacteria compared to their associated eukaryotic hosts together with the high density of microbes in the gut or surrounding the root system suggest that the evolution of host-microbe communities is mainly determined by other selective forces, including microbe-microbe and host-microbe-environment interactions.

Metagenome Analysis-Inferred Functions of the Gut and the Plant Microbiota

The gut microbiota is dominated by a few bacterial phyla, but more variation is observed when focusing on lower taxonomic levels. The relative abundance of individual species can vary over a 10-fold range among individual humans (Spor et al., 2011). In contrast, at the level of gene functions, less variability is observed among individuals, pointing to functional redundancy within the bacterial microbiota and the existence of a conserved functional core (Huttenhower et al., 2012; Turnbaugh et al., 2009).

Given the critical function in nutrient acquisition, it is not surprising that gene functions found in the gut microbial community are influenced by both long- and short-term changes in diet (David et al., 2014b; Muegge et al., 2011; Suez et al., 2014; Wu et al., 2011). Pathways found over all human body parts ("core" pathways) include translational machinery, nucleotide charging, ATP synthesis, and glycolysis (Huttenhower et al., 2012). The functional categories found specifically enriched in the gut microbiota are related to metabolism categories (genes involved in starch, sucrose, and monosaccharide metabolism, including many glycoside hydrolase families). More specifically, functions related to fermentation of complex sugars and glycans to SCFAs, methanogenesis, synthesis of essential amino acids and vitamins, and hydrolysis of phenolic glycosidic conjugates are enriched (Gill et al., 2006; Huttenhower et al., 2012; Qin et al., 2010; Turnbaugh et al., 2009). Some of these functions, such as fermentation and carbohydrate metabolism and vitamin biosynthesis, are also highly expressed in the gut microbiome, as assessed by metatranscriptome analysis (Turnbaugh et al., 2010).

For plant studies, experimental design is more standardized across individuals, which often allows for direct or indirect tests of functional enrichment (Bulgarelli et al., 2015; Mendes et al., 2014; Ofek-Lalzar et al., 2014), in contrast to the human gut microbiome. Shared functional categories found across at least two plant rhizosphere studies relate to iron transport and metabolism, nitrogen metabolism, transport and secretion systems, as well as chemotaxis and motility (Mendes et al., 2014; Ofek-Lalzar et al., 2014; Sessitsch et al., 2012). Similar functions were also found in a metaproteogenomics study of the rice rhizosphere, although in addition, a major role for one-carbon compound recycling could be identified (Knief et al., 2012). However, considerable differences were found in these studies, and additionally no specific function can be assigned for a large proportion of annotated genes in metagenomic studies (42%–86% in the gut; 59% in the plant rhizosphere) (Gill et al., 2006; Huttenhower et al., 2012; Ofek-Lalzar et al., 2014; Qin et al., 2010). A striking commonality between the gut and root metagenome studies is the significant enrichment/high abundance of phage-related functions (Bulgarelli et al., 2015; Qin et al., 2010), but the exact role of these functions is not known.

To gain further insight into the evolutionary forces acting on genes in relation to their functional roles, natural selection was assessed using dN/dS ratios for gene families in the barley rhizosphere and human gut microbiomes (Bulgarelli et al., 2015; Schloissnig et al., 2013). Positive selection is a hallmark of protein families implicated in molecular arms races between two competing organisms. In the rhizosphere, proteins involved in host-pathogen interactions showed significant signs of positive selection, such as the type III secretion system and its associated effectors, phage elements, and microbial CRISPR proteins (Bulgarelli et al., 2015). Similarly, CRISPR-related families, as well as transposases and families related to antibiotic resistance, showed signatures of positive selection in the human gut microbiome (Schloissnig et al., 2013).

Concluding Remarks and Perspectives

To complement large-scale community profile and metagenome studies, reference collections of several hundred isolates from different human body sites and their corresponding genome sequences have been generated (Goodman et al., 2011). For plant-associated microbial communities, similar projects aiming to maximize phylogenetic diversity of cultured bacteria through cross-referencing with culture-independent community profiling experiments are about to be concluded (P.S.-L. and J.L.D., unpublished data). In the future, these genome collections may allow determination of multi-locus reference gene collections for the identification of individual strains within a community, as an alternative to lower-resolution 16S rRNA-based taxon identification, as well as comparative analyses of thousands of genomes for association-based analyses, to link genes and genetic variants to particular phenotypes. The construction of defined (synthetic) communities and their assessment under controlled environments with germ-free eukaryotic hosts allows studies of community resilience and responses to perturbation at the level of individual members and simplifies testing of specific hypotheses relating to individual attributes of other community members and the host (Faith et al., 2014; Guttman et al., 2014). Controlled experimental systems will reduce the noise inherent to any natural environmental sample and will drive the next phase of plant and gut microbiota research in which scientific conclusions are based on causation rather than correlations.

For a detailed description of the meta-analysis, see Supplemental Experimental Procedures. The OTU count matrices and taxonomic information as well as 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 Supplemental Experimental Procedures, one figure, and one table and can be found with this article online at http://dx.doi.org/10.1016/j.chom.2015.04.009.

AUTHOR CONTRIBUTIONS

S.H., R.G.-O., S.S., and P.S.-L. designed research. R.G.-O., A.G., and R.K. designed and R.G.-O., A.G., and G.A. performed the computational analysis. S.H., R.G.-O., S.S., S.L., A.C.M., J.L.D., R.K., R.L., and P.S.-L. wrote the paper.

ACKNOWLEDGMENTS

Work in the different research groups was supported by the Max Planck Society (to P.S.-L.), a European Research Council advanced grant (ROOTMICROBIOTA to P.S.-L.), the ‘‘Cluster of Excellence on Plant Sciences’’ program (funded by the Deutsche Forschungsgemeinschaft to P.S.-L. and A.C.M.), NSF Microbial Systems Biology (grant IOS-0958245 to J.L.D.), NSF INSPIRE (grant IOS-1343020 to J.L.D.), and DOE Feedstocks (grant SC0010423 to J.L.D). S.S. was supported by a postdoctoral grant from the Research Foundation – Flanders (FWO-Vlaanderen). J.L.D. is an Investigator of the Howard Hughes Medical Institute, supported by the HHMI and the Gordon and Betty Moore Foundation (GBMF3030).

REFERENCES

Almagro-Moreno, S., and Boyd, E.F. (2009). Sialic acid catabolism confers a competitive advantage to pathogenic vibrio cholerae in the mouse intestine. Infect. Immun. *77*, 3807–3816.

Bárcen as-Moreno, G., Gómez-Brandón, M., Rousk, J., and Bååth, E. (2009). Adaptation of soil microbial communities to temperature: comparison of fungi and bacteria in a laboratory experiment. Glob. Change Biol. *15*, 2950–2957.

Barret, M., Briand, M., Bonneau, S., Préveaux, A., Valière, S., Bouchez, O., Hunault, G., Simoneau, P., and Jacques, M.-A. (2014). Emergence shapes the structure of the seed-microbiota. Appl. Environ. Microbiol. Published online December 12, 2014. http://dx.doi.org/10.1128/AEM.03722-14.

Belkhadir, Y., and Jaillais, Y. (2015). The molecular circuitry of brassinosteroid signaling. New Phytol. *206*, 522–540.

Benson, A.K., Kelly, S.A., Legge, R., Ma, F., Low, S.J., Kim, J., Zhang, M., Oh, P.L., Nehrenberg, D., Hua, K., et al. (2010). Individuality in gut microbiota composition is a complex polygenic trait shaped by multiple environmental and host genetic factors. Proc. Natl. Acad. Sci. USA *107*, 18933–18938.

Berendsen, R.L., Pieterse, C.M.J., and Bakker, P.A.H.M. (2012). The rhizosphere microbiome and plant health. Trends Plant Sci. *17*, 478–486.

Blaser, M.J., and Falkow, S. (2009). What are the consequences of the disappearing human microbiota? Nat. Rev. Microbiol. *7*, 887–894.

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., Poirier, M.-A., Muller, D., and Moënne-Loccoz, Y. (2014). Root microbiome relates to plant host evolution in maize and other Poaceae. Environ. Microbiol. *16*, 2804–2814.

Brulc, J.M., Antonopoulos, D.A., Miller, M.E.B., Wilson, M.K., Yannarell, A.C., Dinsdale, E.A., Edwards, R.E., Frank, E.D., Emerson, J.B., Wacklin, P., et al. (2009). Gene-centric metagenomics of the fiber-adherent bovine rumen microbiome reveals forage specific glycoside hydrolases. Proc. Natl. Acad. Sci. USA *106*, 1948–1953.

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.

Bulgarelli, D., Garrido-Oter, R., Münch, P.C., Weiman, A., Dröge, J., Pan, Y., McHardy, A.C., and Schulze-Lefert, P. (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. Cell Host Microbe *17*, 392–403.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. Nat. Methods *7*, 335–336.

Cardinale, M., Grube, M., Erlacher, A., Quehenberger, J., and Berg, G. (2015). Bacterial networks and co-occurrence relationships in the lettuce root microbiota. Environ. Microbiol. *17*, 239–252.

Carmody, R.N., Gerber, G.K., Luevano, J.M., Jr., Gatti, D.M., Somes, L., Svenson, K.L., and Turnbaugh, P.J. (2015). Diet dominates host genotype in shaping the murine gut microbiota. Cell Host Microbe *17*, 72–84.

Chaparro, J.M., Badri, D.V., and Vivanco, J.M. (2014). Rhizosphere microbiome assemblage is affected by plant development. ISME J. *8*, 790–803.

Clemente, J.C., Pehrsson, E.C., Blaser, M.J., Sandhu, K., Gao, Z., Wang, B., Magris, M., Hidalgo, G., Contreras, M., Noya-Alarcón, Ó., et al. (2015). The microbiome of uncontacted Amerindians. Sci. Adv. *1*, e1500183.

Cotillard, A., Kennedy, S.P., Kong, L.C., Prifti, E., Pons, N., Le Chatelier, E., Almeida, M., Quinquis, B., Levenez, F., Galleron, N., et al.; ANR MicroObes consortium (2013). Dietary intervention impact on gut microbial gene richness. Nature *500*, 585–588.

Curtis, T.P., Sloan, W.T., and Scannell, J.W. (2002). Estimating prokaryotic diversity and its limits. Proc. Natl. Acad. Sci. USA *99*, 10494–10499.

David, L.A., Materna, A.C., Friedman, J., Campos-Baptista, M.I., Blackburn, M.C., Perrotta, A., Erdman, S.E., and Alm, E.J. (2014a). Host lifestyle affects human microbiota on daily timescales. Genome Biol. *15*, R89.

David, L.A., Maurice, C.F., Carmody, R.N., Gootenberg, D.B., Button, J.E., Wolfe, B.E., Ling, A.V., Devlin, A.S., Varma, Y., Fischbach, M.A., et al. (2014b). Diet rapidly and reproducibly alters the human gut microbiome. Nature *505*, 559–563.

Demoling, F., Figueroa, D., and Bååth, E. (2007). Comparison of factors limiting bacterial growth in different soils. Soil Biol. Biochem. *39*, 2485–2495.

Dominguez-Bello, M.G., Costello, E.K., Contreras, M., Magris, M., Hidalgo, G., Fierer, N., and Knight, R. (2010). Delivery mode shapes the acquisition and structure of the initial microbiota across multiple body habitats in newborns. Proc. Natl. Acad. Sci. USA *107*, 11971–11975.

Donn, S., Kirkegaard, J.A., Perera, G., Richardson, A.E., and Watt, M. (2015). Evolution of bacterial communities in the wheat crop rhizosphere. Environ. Microbiol. *17*, 610–621.

Doornbos, R.F., Geraats, B.P.J., Kuramae, E.E., Van Loon, L.C., and Bakker, P.A.H.M. (2011). Effects of jasmonic acid, ethylene, and salicylic acid signaling on the rhizosphere bacterial community of *Arabidopsis thaliana*. Mol. Plant Microbe Interact. *24*, 395–407.

Duncan, S.H., Louis, P., Thomson, J.M., and Flint, H.J. (2009). The role of pH in determining the species composition of the human colonic microbiota. Environ. Microbiol. *11*, 2112–2122.

Eckburg, P.B., Bik, E.M., Bernstein, C.N., Purdom, E., Dethlefsen, L., Sargent, M., Gill, S.R., Nelson, K.E., and Relman, D.A. (2005). Diversity of the human intestinal microbial flora. Science *308*, 1635–1638.

Edwards, J., Johnson, C., Santos-Medellín, C., Lurie, E., Podishetty, N.K., Bhatnagar, S., Eisen, J.A., and Sundaresan, V. (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. Proc. Natl. Acad. Sci. USA *112*, E911–E920.

Eilam, O., Zarecki, R., Oberhardt, M., Ursell, L.K., Kupiec, M., Knight, R., Gophna, U., and Ruppin, E. (2014). Glycan degradation (GlyDeR) analysis predicts mammalian gut microbiota abundance and host diet-specific adaptations. MBio *5*, e01526–e14.

Eilers, K.G., Lauber, C.L., Knight, R., and Fierer, N. (2010). Shifts in bacterial community structure associated with inputs of low molecular weight carbon compounds to soil. Soil Biol. Biochem. *42*, 896–903.

Faith, J.J., Guruge, J.L., Charbonneau, M., Subramanian, S., Seedorf, H., Goodman, A.L., Clemente, J.C., Knight, R., Heath, A.C., Leibel, R.L., et al. (2013). The long-term stability of the human gut microbiota. Science *341*, 1237439.

Faith, J.J., Ahern, P.P., Ridaura, V.K., Cheng, J., and Gordon, J.I. (2014). Identifying gut microbe-host phenotype relationships using combinatorial communities in gnotobiotic mice. Sci. Transl. Med. *6*, 20ra11.

Faust, K., Sathirapongsasuti, J.F., Izard, J., Segata, N., Gevers, D., Raes, J., and Huttenhower, C. (2012). Microbial co-occurrence relationships in the human microbiome. PLoS Comput. Biol. *8*, e1002606.

Fierer, N., and Jackson, R.B. (2006). The diversity and biogeography of soil bacterial communities. Proc. Natl. Acad. Sci. USA *103*, 626–631.

Franzenburg, S., Walter, J., Künzel, S., Wang, J., Baines, J.F., Bosch, T.C., and Fraune, S. (2013). Distinct antimicrobial peptide expression determines host species-specific bacterial associations. Proc. Natl. Acad. Sci. USA *110*, E3730–E3738.

Fraune, S., Anton-Erxleben, F., Augustin, R., Franzenburg, S., Knop, M., Schröder, K., Willoweit-Ohl, D., and Bosch, T.C. (2014). Bacteria-bacteria interactions within the microbiota of the ancestral metazoan Hydra contribute to fungal resistance. ISME J. http://dx.doi.org/10.1038/ismej.2014.239.

García-Garrido, J.M., and Ocampo, J.A. (2002). Regulation of the plant defence response in arbuscular mycorrhizal symbiosis. J. Exp. Bot. *53*, 1377–1386.

Gaudier, E., Jarry, A., Blottière, H.M., de Coppet, P., Buisine, M.P., Aubert, J.P., Laboisse, C., Cherbut, C., and Hoebler, C. (2004). Butyrate specifically modulates MUC gene expression in intestinal epithelial goblet cells deprived of glucose. Am. J. Physiol. Gastrointest. Liver Physiol. *287*, G1168–G1174.

Gill, S.R., Pop, M., Deboy, R.T., Eckburg, P.B., Turnbaugh, P.J., Samuel, B.S., Gordon, J.I., Relman, D.A., Fraser-Liggett, C.M., and Nelson, K.E. (2006). Metagenomic analysis of the human distal gut microbiome. Science *312*, 1355–1359.

Goodman, A.L., Kallstrom, G., Faith, J.J., Reyes, A., Moore, A., Dantas, G., and Gordon, J.I. (2011). Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. Proc. Natl. Acad. Sci. USA *108*, 6252–6257.

Goodrich, J.K., Waters, J.L., Poole, A.C., Sutter, J.L., Koren, O., Blekhman, R., Beaumont, M., Van Treuren, W., Knight, R., Bell, J.T., et al. (2014). Human genetics shape the gut microbiome. Cell *159*, 789–799.

Guttman, D.S., McHardy, A.C., and Schulze-Lefert, P. (2014). Microbial genome-enabled insights into plant-microorganism interactions. Nat. Rev. Genet. *15*, 797–813.

Hacquard, S., and Schadt, C.W. (2015). Towards a holistic understanding of the beneficial interactions across the *Populus* microbiome. New Phytol. *205*, 1424–1430.

Hinsinger, P., Plassard, C., Tang, C., and Jaillard, B. (2003). Origins of root-mediated pH changes in the rhizosphere and their responses to environmental constraints: A review. Plant Soil *248*, 43–59.

Hoskins, L.C., and Boulding, E.T. (1981). Mucin degradation in human colon ecosystems. Evidence for the existence and role of bacterial subpopulations producing glycosidases as extracellular enzymes. J. Clin. Invest. *67*, 163–172.

Huttenhower, C., Gevers, D., Knight, R., Abubucker, S., Badger, J.H., Chinwalla, A.T., Creasy, H.H., Earl, A.M., FitzGerald, M.G., Fulton, R.S., et al.; Human Microbiome Project Consortium (2012). Structure, function and diversity of the healthy human microbiome. Nature *486*, 207–214.

doi:10.1038/nature16192

Functional overlap of the *Arabidopsis* leaf and root microbiota

Yang Bai[1]*, Daniel B. Müller[2]*, Girish Srinivas[1]*, Ruben Garrido-Oter[1,3,4]*, Eva Potthoff[2], Matthias Rott[1], Nina Dombrowski[1], Philipp C. Münch[5,6,7], Stijn Spaepen[1], Mitja Remus-Emsermann[2], Bruno Hüttel[8], Alice C. McHardy[4,5], Julia A. Vorholt[2]* & Paul Schulze-Lefert[1,4]*

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

Plants and animals harbour abundant and diverse bacterial microbiota[1]. These taxonomically structured bacterial communities have important functions for the health of their multicellular eukaryotic hosts[2–4]. The leaf and root microbiota of flowering plants have been extensively studied by culture-independent analyses, which have consistently revealed the co-occurrence of four main bacterial phyla: Actinobacteria, Bacteroidetes, Firmicutes and Proteobacteria[5–15]. Determinants of microbiota composition at lower taxonomic ranks, that is, at genus and species level, are host compartment, environmental factors and host genotype[6,7,12,16].

Soil harbours an extraordinary rich diversity of bacteria and these define the start inoculum of the *Arabidopsis thaliana* root microbiota[6,7]. The inoculum source of the leaf microbiota is thought to be more variable owing to the inherently open nature of the leaf ecosystem, probably involving bacteria transmitted by aerosols, insects, or soil[8,9,17]. A recent study of the grapevine (*Vitis vinifera*) microbiota showed that the root-associated bacterial assemblies differed significantly from aboveground communities, but that microbiota of leaves, flowers, and grapes shared a greater proportion of taxa with soil communities than with each other, suggesting that soil may serve as a common bacterial reservoir for belowground and aboveground plant microbiota[18].

A major limitation of current plant microbiota research is the lack of systematic microbiota culture collections that can be employed in microbiota reconstitution experiments with germ-free plants to address principles underlying community assembly and proposed microbiota functions for plant health under laboratory conditions[19].

Bacterial culture collections from roots and leaves

We employed three bacterial isolation procedures to establish taxonomically diverse culture collections of the *A. thaliana* root and leaf microbiota. Bacterial isolates were recovered from pooled or individual roots or leaves of healthy plants using colony picking from agar plates, limiting dilution in liquid media in 96-well microtitre plates, or microbial cell sorting (see Methods). We adopted a two-step barcoded pyrosequencing protocol[20] for taxonomic classification of the cultured bacteria by determining ≥550 base pairs (bp) 16S ribosomal RNA (rRNA) gene sequences (Supplementary Fig. 1; Methods). In parallel, parts of the root and leaf material was used for cultivation-independent 16S rRNA gene community sequencing to cross-reference Operational Taxonomic Unit (OTU)-defined taxa from the microbiota with individual colony forming units (CFUs) in the culture collections.

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

[1]Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [2]Institute of Microbiology, ETH Zurich, 8093 Zurich, Switzerland. [3]Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, 40225 Düsseldorf, Germany. [4]Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [5]Computational Biology of Infection Research, Helmholtz Center for Infection Research, 38124 Braunschweig, Germany. [6]Max-von-Pettenkofer Institute, Ludwig Maximilian University, German Center for Infection Research (DZIF), partner site LMU Munich, 80336 Munich, Germany. [7]German Center for Infection Research (DZIF), partner site Hannover-Braunschweig, 38124 Braunschweig, Germany. [8]Max Planck Genome Center, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany.
*These authors contributed equally to this work.

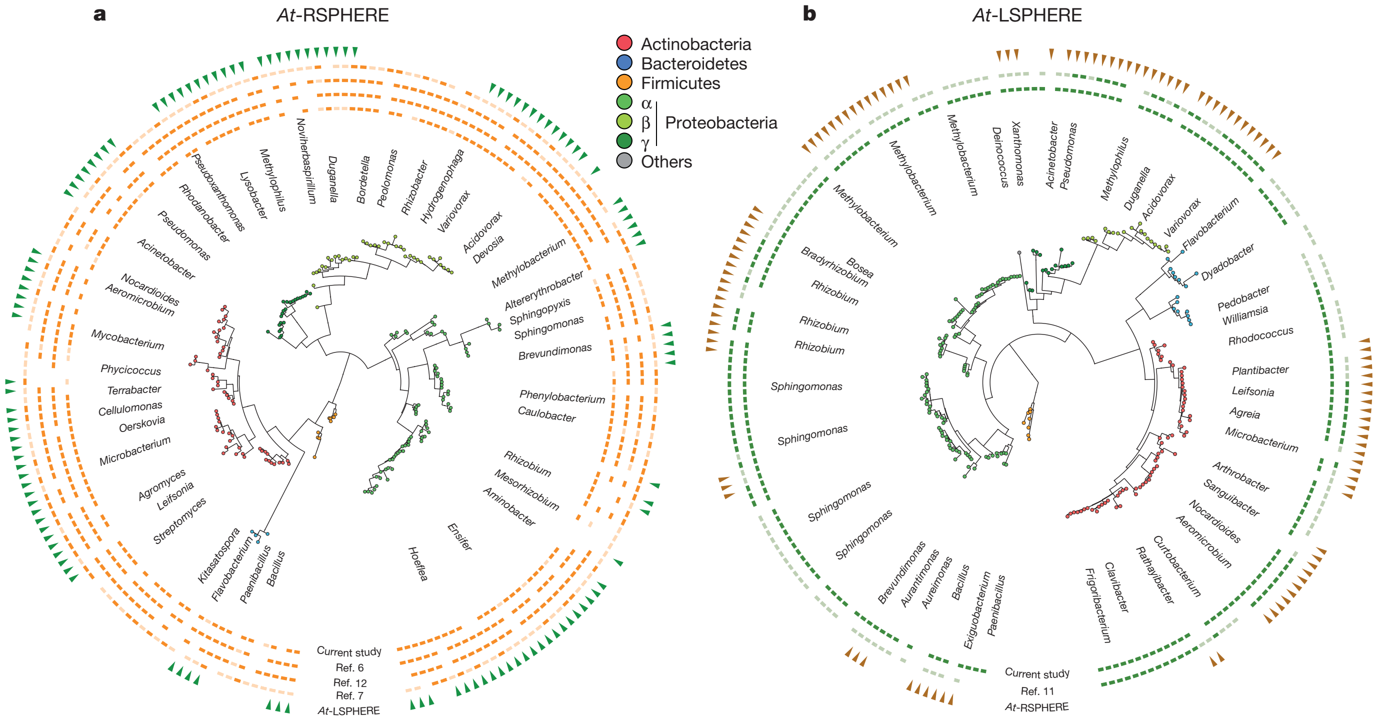


Figure 1 | Taxonomic overlap between *At*-RSPHERE and *At*-LSPHERE isolates and their representation in culture-independent microbiota profiling studies. a, **b**, Phylogenetic trees of *At*-RSPHERE (**a**; $n = 206$ isolates) and *At*-LSPHERE (**b**; $n = 224$ isolates) bacteria. Their taxonomic overlap is shown in the outermost ring (green or brown triangles). **a**, Representation of *At*-RSPHERE bacteria in each of four indicated culture-independent profiling studies of the *A. thaliana* root microbiota; root-associated OTUs with RAs ≥0.1% (dark orange) or ≤0.1% (light orange). **b**, Representation of *At*-LSPHERE bacteria in the two indicated culture-independent phyllosphere profiling studies; leaf-associated OTUs with RAs ≥0.1% (dark green) or <0.1% (light green). Taxonomic assignment and phylogenetic tree inference were based on partial 16S rRNA gene Sanger sequences.

At-RSPHERE and *At*-LSPHERE culture collections

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

Notably, the majority of the *At*-RSPHERE isolates share ≥97% 16S rRNA gene sequence identity matches with root-associated OTUs reported in four independent studies in which *A. thaliana* plants had been grown in Cologne soil[6,12] or other European[6,12] or US soils[7] (inner four circles in Fig. 1a; Methods). Similarly, the bulk of *At*-LSPHERE isolates match leaf-derived OTUs detected in *A. thaliana* populations at the Tübingen/Zurich locations or US-grown plants (innermost two circles in Fig. 1b). This indicates that representatives of the majority of *At*-RSPHERE and *At*-LSPHERE members co-populate the corresponding *A. thaliana* organs in multiple tested environments, including two continents, Europe and North America.

Phylogenetic analysis based on 16S rRNA gene Sanger sequences revealed that 119 out of 206 *At*-RSPHERE isolates (58%) share ≥97% sequence identity matches with corresponding 16S rRNA gene fragments of *At*-LSPHERE members (outermost circle in Fig. 1a). Similarly, 108 out of 224 *At*-LSPHERE isolates (48%) share ≥97% sequence identity matches with *At*-RSPHERE members (outermost circle in Fig. 1b). This extensive overlap both at the rank of bacterial genera and bacterial families (20 out of 38 detectable families) between leaf- and root-derived bacteria is notable because we collected leaf and root specimen from environments that are geographically widely separated (>500 km) and is consistent with a previous report on leaf and root microbiota overlap in *V. vinifera*[18]. This overlap is corroborated by the corresponding culture-independent leaf and root community profiles (Extended Data Fig. 4). As essentially all *A. thaliana* root-associated bacteria are recruited from the surrounding soil biome[6,7,12], this raises the possibility that unplanted soil also defines the start inoculum for a substantial proportion of the leaf microbiota with subsequent selection for niche-adapted organisms.

Comparative genome analysis of the culture collections

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

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

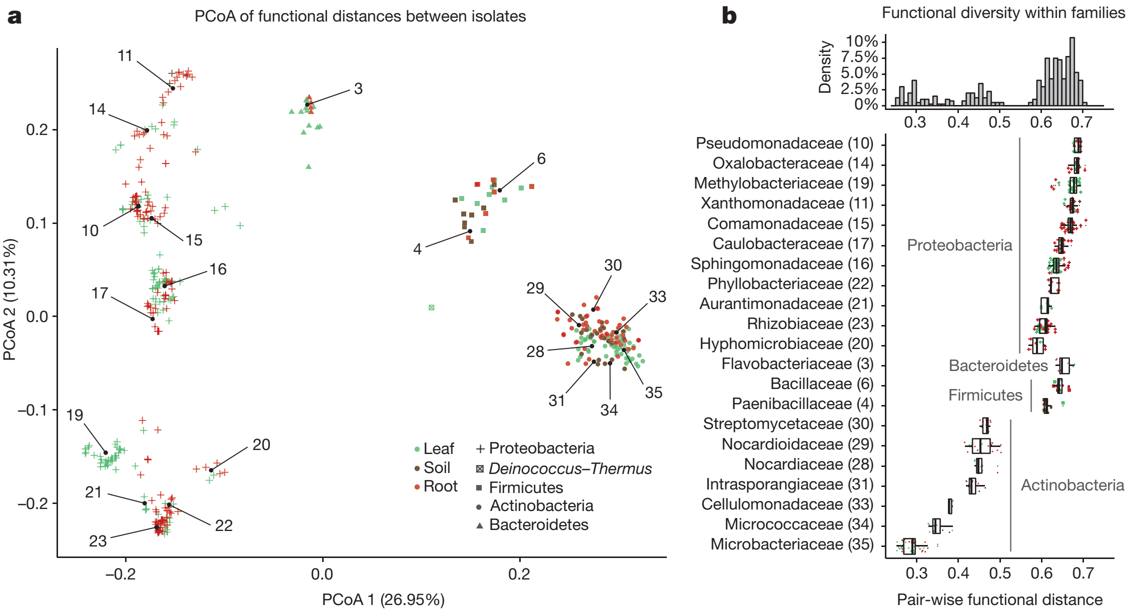


Figure 2 | Analysis of functional diversity between sequenced isolates. **a**, Principal coordinate analysis (PCoA) plot depicting functional distances between sequenced genomes ($n = 432$) based on the KEGG Orthology (KO) database annotation. Each point represents a genome. Colours represent the organ of isolation and shapes correspond to their taxonomy. Numbers inside the plot refer to bacterial families listed in **b**. **b**, Analysis of functional diversity within bacterial families as measured by pair-wise functional distances between genomes (bottom panel; $n = 432$). Higher pairwise distances between members of a family indicate a larger degree of functional diversity. Only families with at least five members are shown. The histogram (top panel) was calculated for the entire data set and the y-axis corresponds to the percentage of data points in each bin. Boxplot whiskers extend to the most extreme data point which is no more than 1.5 times the interquartile range from the upper or lower quartiles.

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

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

We searched for signatures of niche specialization at individual functional categories using enrichment analysis to identify functional categories over-represented in genomes from root and leaf or soil isolates (Fig. 3; Methods). Specifically, we found the category ‘carbohydrate metabolism’ to be enriched in the leaf and soil genomes compared to those isolated from roots (Mann–Whitney test, $P = 1.29 \times 10^{-7}$; Fig. 3b). We speculate that this differential enrichment could reflect the availability of simple carbon sources in roots through the process of root exudation (sugars, amino acids, aliphatic acids)[21,22], whereas bacteria associated with leaves or unplanted soil might rely on a more diverse repertoire of carbohydrate metabolism genes to access scarce and complex organic carbon, for example, polysaccharides and leaf cuticular waxes. The category ‘xenobiotics biodegradation and catabolism’ is enriched in the root genomes with respect to those isolated from leaves ($P = 2.60 \times 10^{-11}$; Fig. 3b), which is consistent with previous evidence that genes for aromatic compound utilization are expressed in the rhizosphere[23]. No single taxon is responsible for these significant differences, but this seems to be a general feature across the sequenced bacterial genomes of the respective ecological niche (Extended Data Figs 5 and 6). Interestingly, we observed the same trends of differential abundance of functional categories in *V. vinifera* root metagenome samples[18] compared to their respective unplanted soil controls (Extended Data Fig. 7).

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

Synthetic community colonization of germ-free plants

We colonized germ-free *A. thaliana* plants with synthetic communities (SynComs) consisting of bacterial isolates from our culture collections to assess their potential for host colonization in a gnotobiotic system containing calcined clay as inert soil substitute (Methods). To mimic the taxonomic diversity of leaf and root microbiota in natural environments we employed mainly two SynComs: ‘L’ comprising 218 leaf-derived bacteria and ‘R+S’ consisting of 188 members of which 158 are root-derived and 30 are soil-derived bacteria (Supplementary Data 6). Input SynComs were either inoculated directly before sowing of surface-sterilized seeds in calcined clay and/or spray-inoculated on leaves of three-week-old germ-free plants. For all defined communities we examined three independent SynCom preparations, each tested in three closed containers containing four plants. We employed 16S rRNA gene community profiling with a method validated for defined communities[24] to detect potential community shifts between input and output

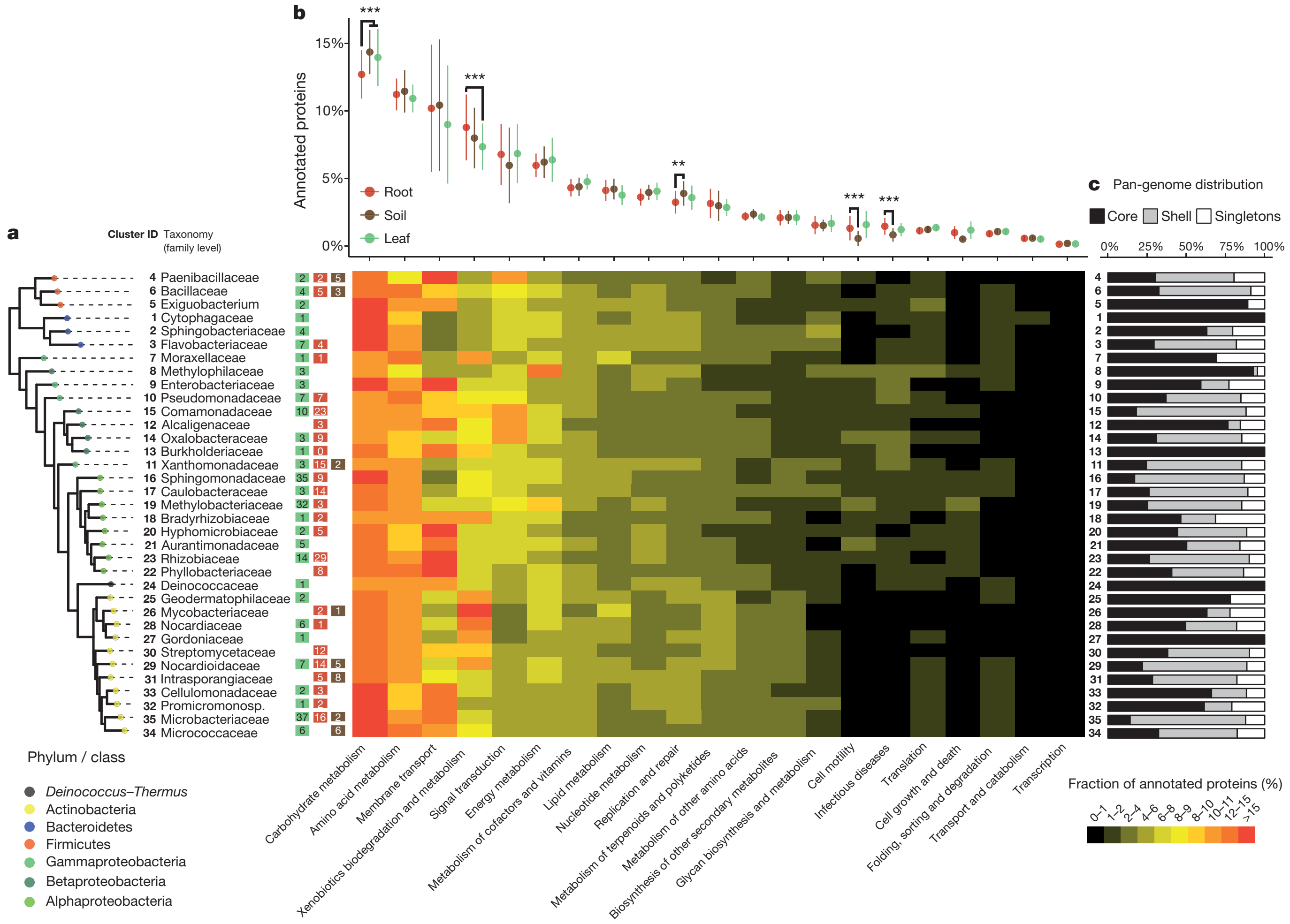


Figure 3 | Functional analysis of sequenced isolates. a, Phylogeny of family-level clusters of bacterial isolates. The tips of the tree are annotated, from left to right, with the cluster ID, taxonomic classification, followed by the number of sequenced isolates from leaf, root or soil that constitute each cluster. The heat map depicts the average percentage of annotated proteins of each cluster belonging to each functional category. **b**, Functional enrichment analysis between leaf ($n = 206$), root ($n = 194$) and soil ($n = 32$) genomes. Points and bars correspond to the mean abundance and standard deviation of each functional category. P values were obtained using the non-parametric Mann–Whitney test corrected by the Bonferroni approach. **c**, Analysis of pan-genome distribution for each cluster of genomes, indicating the percentage of annotated proteins found in only one isolate (singletons), in more than one but not all (shell) or in all genomes within the cluster (core).

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

Upon application of the input R+S SynCom to clay ('R+S in clay') and co-cultivation with *A. thaliana* plants for seven weeks we retrieved reproducible R+S output communities from clay (without host), root, and leaf compartments (Supplementary Fig. 2). These output SynCom profiles were robust against a 75% reduction in RA of Proteobacteria compared to Actinobacteria, Bacteroidetes and Firmicutes in the input R+S SynCom (input ratios 1:1:1:1 or 1:1:1:0.25, respectively), which was confirmed by PCoA (Fig. 4a). PCoA also revealed distinct output communities in each of the three tested compartments (Fig. 4a; $P < 0.001$ Extended Data Fig. 8a, b). This indicates that a marked host-independent community change occurred in clay (without host) as well as host-dependent community shifts that are specific for leaves and roots. Next, we tested the 'L' SynCom of leaf-derived bacteria by spray inoculation on leaves of three week-old plants. After four weeks of L SynCom co-incubation with plants, output communities were detected in leaves and roots (Supplementary Fig. 3). PCoA revealed that these two output communities were different between each other, but robust against a 75% reduction in RA of input Proteobacteria (Fig. 4b; Supplementary Fig. 3; $P < 0.001$; Extended Data Fig. 8c, d). The converging output communities despite varying RAs of input SynComs suggest that the communities have reached a steady state. These experiments also reveal that both R+S and L SynCom members not only colonize cognate host organs, but are capable of ectopic colonization of leaves and roots, which might be linked to the extensive species overlap of *A. thaliana* leaf and root microbiota in natural environments (Fig. 1a, b). Additionally, this provides experimental support for the hypothesis that a subset of leaf-colonizing bacteria originates from unplanted soil and raises the possibility for reciprocal bacterial colonization events between roots and leaves during and/or after the establishment of the respective microbiota, for example, by ascending migration of rhizobacteria from roots to leaves[25]. Upon leaf spray application of SynComs, a small amount of leaf bacteria is likely to land on the clay surface and thereafter colonize roots, which is not fundamentally different from processes occurring in natural environments, for example, during rain showers and/or leaf dehiscence.

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

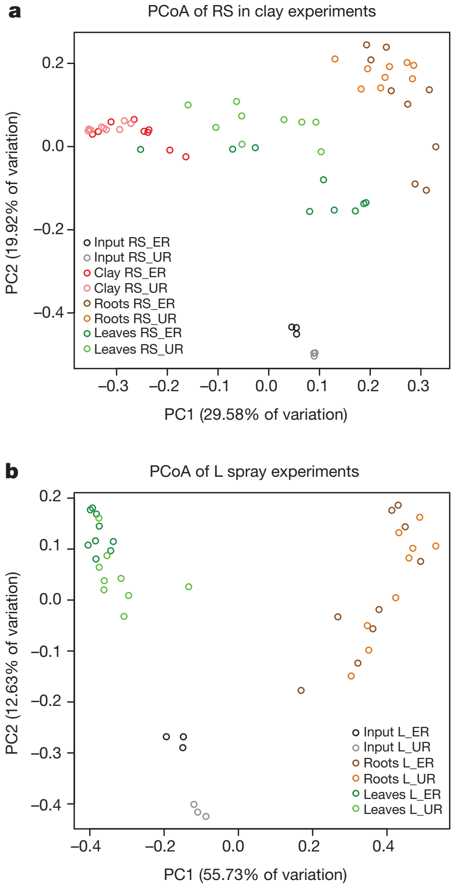


Figure 4 | SynCom colonization of germ-free *A. thaliana* plants. **a**, **b**, Principal coordinate analysis (PCoA) of Bray–Curtis distances of input and output SynCom profiles of RS in clay (**a**; $n = 60$) and L spray (**b**; $n = 42$) experiments. Each condition was tested with 6 independently prepared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

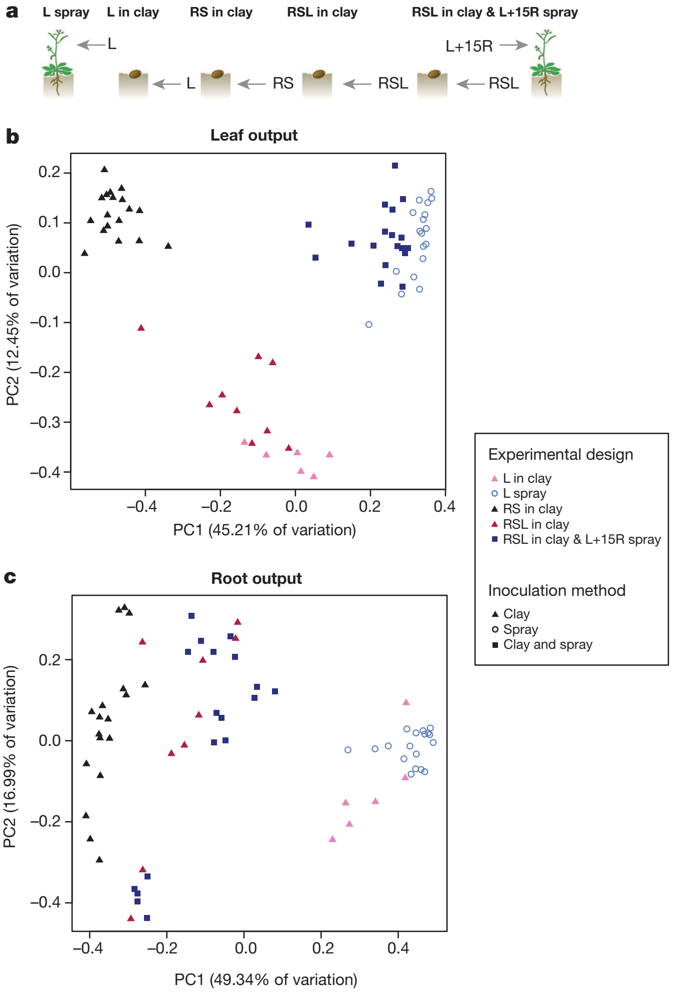


Figure 5 | SynCom competition supports host-organ-specific community assemblies. **a**, Pictograms illustrating 'L spray', 'L in clay', 'RS in clay', 'RSL in clay', and 'RSL in clay & L+15R spray' SynCom experiments. **b**, **c**, PCoA of Bray–Curtis distances of leaf (**b**; $n = 69$) and root (**c**;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

Niche-specific microbiota establishment with SynComs

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

We found significant differences between leaf-associated output communities of the 'RSL in clay' and 'RS in clay' experiments (Fig. 5b; $P < 0.001$, Extended Data Fig. 8f; Supplementary Figs 2 and 4) and that the output community on leaves after 'L in clay' inoculation is similar to the leaf outputs of 'RSL in clay' inoculation (Fig. 5b; $P < 0.001$, Extended Data Fig. 8f; Supplementary Figs 4 and 5), indicating that in this comparison the leaf-derived SynCom has a stronger influence on leaf microbiota structure than root- and soil-derived bacteria. However, both 'RSL in clay' and 'L in clay' leaf outputs are significantly different from the leaf output of the 'L spray' experiment (Fig. 5b; $P < 0.001$, Extended Data Fig. 8e; Supplementary Figs 3–5), showing that many leaf-derived isolates do not successfully colonize leaves when only inoculated in the clay environment. For example, of the top 16 genera a total of three are grossly underrepresented in leaf outputs of the 'RSL in clay' compared to the 'RSL in clay & L+15R spray' experiment (*Chryseobacterium*, *Sphingomonas* and *Variovorax*; Supplementary Fig. 6) and these three genera are abundant in the natural leaf microbiota (Extended Data Fig. 4). Finally, leaf outputs were strikingly similar between 'RSL in clay & L+15R spray' and 'L spray' only experiments (Fig. 5b; Supplementary Figs 3 and 7), indicating that the L+15R SynCom, leaf spray-inoculated three weeks after RSL application to clay, can displace the RSL leaf output. Collectively, these results support the hypothesis that leaf microbiota establishment benefits from air- and soil-borne inoculations[8,17], although we note that our single application of bacteria to leaves does not mimic the continuous exposure of plant leaves to airborne microorganisms in nature.

A comparison of the root-associated community outputs of the experiments described above revealed that the 'RSL in clay' experiment is more similar to root outputs of the 'RS in clay' than 'L in clay' experiments (Fig. 5c; $P < 0.001$ Extended Data Fig. 8g), suggesting that the root- and soil-derived SynCom has a stronger influence on root microbiota structure than the leaf-derived SynCom. In this experiment the fractional contribution of root-specific indicator OTUs increases in the output, but decreases for leaf-specific indicator OTUs, relative to their input, pointing to a potential adaptation of root-derived bacteria for root colonization (Extended Data Fig. 10a; Mann–Whitney; $P < 0.05$). This is further supported by the observation that in the 'RSL in clay' experiment root colonization rates for root-specific indicator OTUs are higher compared to those specific for leaves when applying a 0.1% relative abundance threshold in at least one biological replicate (69% and 33%, respectively). Taken together, this suggests that root-derived bacteria are better adapted to colonize their cognate host niche than leaf-derived bacteria. Further comparisons of the root-associated output communities of the 'L in clay' and 'L spray' experiments (Fig. 5c; Supplementary Figs 3 and 5) revealed similar community composition, indicating convergence of ectopic root-associated community outputs despite different inoculation time points or sites of application. Additional reciprocal transplantation experiments using a 'R' (root strains only) SynCom either applied to clay ('R in clay') or by spray inoculation ('R spray') confirmed the convergence of ectopic community outputs also for root-derived bacteria on leaves (Extended Data Fig. 10 b, c; Supplementary Figs 8 and 9). Convergence of ectopic SynCom outputs is consistent with the hypothesis that a subset of leaf and root colonizing bacteria has the potential to relocate between leaves and roots.

Conclusions

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

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

Received 24 June; accepted 9 November 2015.
Published online 2 December 2015.

1. Rosenberg, E. & Xilber-Rosenberg, I. *The Hologenome Concept: Human, Animal and Plant Microbiota* (Springer, 2013).
2. Spor, A., Koren, O. & Ley, R. Unravelling the effects of the environment and host genotype on the gut microbiome. *Nature Rev. Microbiol*. **9,** 279–290 (2011).
3. Berendsen, R. L., Pieterse, C. M. & Bakker, P. A. The rhizosphere microbiome and plant health. *Trends Plant Sci*. **17,** 478–486 (2012).
4. Subramanian, S. *et al.* Cultivating healthy growth and nutrition through the gut microbiota. *Cell* **161,** 36–48 (2015).
5. Delmotte, N. *et al.* Community proteogenomics reveals insights into the physiology of phyllosphere bacteria. *Proc. Natl Acad. Sci. USA* **106,** 16428–16433 (2009).
6. Bulgarelli, D. *et al.* Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* **488,** 91–95 (2012).
7. Lundberg, D. S. *et al.* Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488,** 86–90 (2012).
8. Vorholt, J. A. Microbial life in the phyllosphere. *Nature Rev. Microbiol*. **10,** 828–840 (2012).
9. Bodenhausen, N., Horton, M. W. & Bergelson, J. Bacterial communities associated with the leaves and the roots of *Arabidopsis thaliana*. *PLoS One* **8,** e56329 (2013).
10. Guttman, D. S., McHardy, A. C. & Schulze-Lefert, P. Microbial genome-enabled insights into plant-microorganism interactions. *Nature Rev. Genet.* **15,** 797–813 (2014).
11. Horton, M. W. *et al.* Genome-wide association study of *Arabidopsis thaliana* leaf microbial community. *Nat. Commun.* **5,** 5320 (2014).
12. Schlaeppi, K., Dombrowski, N., Oter, R. G., Ver Loren van Themaat, E. & Schulze-Lefert, P. Quantitative divergence of the bacterial root microbiota in *Arabidopsis thaliana* relatives. *Proc. Natl Acad. Sci. USA* **111,** 585–592 (2014).
13. Edwards, J. *et al.* Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc. Natl Acad. Sci. USA* **112,** E911–E920 (2015).
14. Hacquard, S. *et al.* Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* **17,** 603–616 (2015).
15. Bulgarelli, D. *et al.* Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17,** 392–403 (2015).
16. Lebeis, S. L. *et al.* Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* **349,** 860–864 (2015).
17. Maignien, L., DeForce, E. A., Chafee, M. E., Eren, A. M. & Simmons, S. L. Ecological succession and stochastic variation in the assembly of *Arabidopsis thaliana* phyllosphere communities. *MBio* **5,** e00682–e13 (2014).
18. Zarraonaindia, I. *et al.* The soil microbiome influences grapevine-associated microbiota. *MBio* **6,** e02527–14 (2015).
19. Lebeis, S. L., Rott, M., Dangl, J. L. & Schulze-Lefert, P. Culturing a plant microbiome community at the cross-Rhodes. *New Phytol.* **196,** 341–344 (2012).
20. Goodman, A. L. *et al.* Extensive personal human gut microbiota culture collections characterized and manipulated in gnotobiotic mice. *Proc. Natl Acad. Sci. USA* **108,** 6252–6257 (2011).
21. Faure, D., Vereecke, D. & Leveau, J. J. Molecular communication in the rhizosphere. *Plant Soil* **321,** 279–303 (2009).
22. Bais, H. P., Weir, T. L., Perry, L. G., Gilroy, S. & Vivanco, J. M. The role of root exudates in rhizosphere interactions with plants and other organisms. *Annu. Rev. Plant Biol.* **57,** 233–266 (2006).
23. Ramachandran, V. K., East, A. K., Karunakaran, R., Downie, J. A. & Poole, P. S. Adaptation of *Rhizobium leguminosarum* to pea, alfalfa and sugar beet rhizospheres investigated by comparative transcriptomics. *Genome Biol.* **12,** R106 (2011).
24. Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* **10,** 996–998 (2013).
25. Chi, F. *et al.* Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. *Appl. Environ. Microbiol.* **71,** 7271–7278 (2005).

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

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

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

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

METHODS

Sampling of *A. thaliana* plants and isolation of root-, leaf- and soil-derived bacteria. *A. thaliana* plants were either harvested from natural populations or grown in different natural soils and used for bacterial isolations by colony picking, limiting dilution or bacterial cell sorting as well as 16S rRNA gene-based community profiling. To obtain a library of representative root colonizing bacteria, *A. thaliana* plants were grown in different soils (50.958 N, 6.856 E, Cologne, Germany; 52.416 N, 12.968 E, Golm, Germany; 50.982 N, 6.827 E, Widdersdorf, Germany; 47.941 N, 04.012 W, Saint-Evarzec, France; 48.725 N, 3.989 W, Roscoff, France) and harvested before bolting. Briefly, *Arabidopsis* roots were washed twice in washing buffers (10 mM $MgCl_2$ for limiting dilution and PBS for colony picking[6]) on a shaking platform for 20 min at 180 rpm and then homogenized twice by Precellys24 tissue lyser (Bertin Technologies) using 3 mM metal beads at 5,600 rpm for 30 s. Homogenates were diluted and used for isolation approaches on several bacterial growth media (Supplementary Data 7). For isolations based on colony picking, diluted cell suspensions were plated on solidified media and incubated, before isolates of plates containing less than 20 colony-forming units (CFUs) were picked after a maximum of two weeks of incubation. For limiting dilution, homogenized roots from each root pool were sedimented for 15 min and the supernatant was empirically diluted, distributed and cultivated in 96-well microtitre plates[20]. In parallel to the isolation of root-derived bacteria, roots of plants grown in Cologne soil were harvested and used to assess bacterial diversity by culture-independent 16S rRNA gene sequencing. Additionally, soil-derived bacteria were extracted from unplanted Cologne soil by washing soil with PBS buffer, supplemented with 0.02% Silwet L-77 and subjected to bacterial isolation as well as 16S rRNA gene community profiling. For the isolation of representative phyllosphere strains, naturally grown *Arabidopsis* plants were collected at eight different sites in southern Germany and Switzerland (six main sampling sites used for bacterial isolations and community profiling: 47.4090306 N, 8.470169444 E, Hoengg, Switzerland; 47.474825 N, 8.305008333 E, Baden, Switzerland; 47.4816806 N, 8.217547222 E, Brugg, Switzerland; 48.5560194 N, 9.134944444 E, Farm, Tuebingen, Germany; 48.5989861 N, 9.201655556 E, Haeslach, Germany; 48.602682 N, 9.213247258 E, Haeslach, Germany; and two additional sites only used for bacterial isolation: 47.4074722 N, 8.50825 E, Zurich, Switzerland; 47.4227222 N, 8.548666667 E, Seebach, Switzerland) during spring and autumn of 2013 and used for bacterial isolations as well as 16S rRNA gene profiling. Leaf-colonizing bacteria of individual leaves were washed off by alternating steps of intense mixing and sonication. The suspension was subsequently filtered (CellTrics filters, 10 μM, Partec GmbH, Görlitz, Germany) in order to remove remaining plant or debris particles as well as cell aggregates and applied to cell sorting on a BD FACS Aria III (BD Biosciences) as well as to plating on different media (Supplementary Data 1 and 7). All isolates were subsequently stored in 30% or 40% glycerol at −80 °C.

Culture-independent bacterial 16S rRNA gene profiling of *A. thaliana* leaf, root and corresponding soil samples. Parts of *A. thaliana* leaves, roots and corresponding unplanted soil samples used for bacterial isolation were also processed for bacterial 16S rRNA gene community profiling using 454 pyrosequencing. Frozen root and corresponding soil samples were homogenized, DNA was extracted with Lysing Matrix E (MP Biomedicals) at 5,600 rpm for 30 s, and DNA was extracted from all samples using the FastDNA SPIN Kit for soil (MP Biomedicals) according to the manufacturer's instructions. Lyophilized leaf samples were transferred into 2 ml microcentrifuge tubes containing one metal bead and subsequently homogenized twice for 2 min at 25 Hz using a Retsch tissue lyser (Retsch, Haan, Germany). Homogenized leaf material was resuspended in lysis buffer of the MO BIO PowerSoil DNA isolation Kit (MO BIO Laboratories Inc., Carlsbad, CA, USA), transferred into lysis tubes, provided by the supplier, and DNA extraction was performed following the manufacturer's protocol. DNA concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), and subsequently diluted to 3.5 ng μl^{-1}. Bacterial16S rRNA genes were subsequently amplified[6] using primers targeting the variable regions V5-V7 (799F[26] and 1193R[6], Supplementary Data 7). Each sample was amplified in triplicate by two independent PCR mixtures (a total of 6 replicates per sample plus respective no template controls). PCR products of triplicate were subsequently combined, purified and subjected to 454 sequencing. Obtained sequences were demultiplexed as well as quality and length filtered (average quality score ≥25, minimum length 319 bp with no ambiguous bases and no errors in the barcode sequences allowed)[27]. High-quality sequences were subsequently processed using the UPARSE[24] pipeline and OTUs were taxonomically classified using the Greengenes database[28] and the PyNAST[29] method.

High-throughput identification of leaf-, root- and soil-derived bacterial isolates by 454 pyrosequencing. We adopted a two-step barcoded PCR protocol[20] in combination with 454 pyrosequencing to define V5-V8 sequences of bacterial 16S rRNA genes of all leaf, root- and soil-derived bacterial (Supplementary Fig. 1). DNA of isolates was extracted by lysis of 6 μl of bacterial cultures in 10 μl of buffer I containing 25 mM NaOH, 0.2 mM EDTA, pH 12 at 95 °C for 30 min, before the pH value was lowered by addition of 10 μl of buffer II containing 40 mM Tris-HCl at pH 7.5. Position and taxonomy of isolates in 96-well microtitre plates were indexed by a two-step PCR protocol using the degenerate primers 799F and 1392R containing well- and plate-specific barcodes (Supplementary Data 7) to amplify the variable regions V5 to V8. During the first step of PCR amplification, DNA from 1.5 μl of lysed cells was amplified using 2 U DSF-Taq DNA polymerase, 1× complete buffer (both Bioron GmbH), 0.2 mM dNTPs (Life technologies), 0.2 μM of 1 of 96 barcoded forward primer with a 18-bp linker sequence (for example, A1_454_799F1_PCR1_wells; Supplementary Data 7) and 0.2 μM reverse primer (454B_1392R) in a 25 μl reaction. PCR amplification was performed under the following conditions: DNA was initially denatured at 95 °C for 2 min, followed by 40 cycles of 95 °C for 30 s, 50 °C for 30 s and 72 °C for 45 s, and a final elongation step at 72 °C for 10 min. PCR products of each 96-well microtitre plate were combined and subsequently purified in a two-step procedure using the Agencourt AMPure XP Kit (Beckman Coulter GmbH, Krefeld, Germany) first, then DNA fragments were excised from a 1% agarose gel using the QIAquick Gel Extraction Kit (Qiagen). DNA concentration was measured by Nanodrop and diluted to 1 ng μl^{-1}.

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

The sequences were quality filtered, demultiplexed according to well and plate identifiers[27]. OTUs were clustered at 97% similarity by UPARSE algorithum[24]. A nucleotide-based blast (v. 2.2.29) was used to align representative sequences of isolated OTUs to culture-independent OTUs and only hits ≥97% sequence identity covering at least 99% of the length of the sequences were considered.

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

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

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

Genome assembly and annotation. Paired-end Illumina reads were subjected to quality and length trimming using Trimmomatic v. 0.33[30] and assembled using two independent methods (A5[31] and SOAPdenovo[32] v. 20.1). In each case, the assembly with the smaller number of scaffolds was selected. Detailed assembly statistics for each sequenced isolate can be found in Supplementary Data 3 and 4. Identification of putative protein-encoding genes and annotation of the genomes were performed using GLIMMER v. 3.02[33]. Functional annotation of genes was conducted using Prokka v. 1.11[34] and the SEED subsystems approach using the RAST server API[35]. Additionally, annotation of KEGG Orthologue (KO) groups was performed by first generating HMM models for each KO in the database[36,37] the HMMER toolkit (v. 3.1b2)[38]. Next, we employed the HMM models to search all predicted ORFs using the hmmsearch tool, with an E value threshold of 10×10^{-5}. Only hits covering at least 70% of the protein sequence were retained and for each gene and the match with the lowest E value was selected.

Analyses of phylogenetic diversity within sequenced isolates. Each proteome was searched for the presence of the 31 well-conserved, single-copy, bacterial AMPHORA genes[39], designed for the purpose of high-resolution phylogeny reconstruction of genomes. Subsequently, a concatenated alignment of these marker genes was performed using Clustal Omega[40] v. 1.2.1. Based on this multiple sequence alignment, a species tree was inferred using FastTree[41] v. 2.1, a maximum likelihood tool for phylogeny inference. Whole-genome taxonomic classification of sequenced isolates was conducting using taxator-tk[42], a homology/based tool for accurate classification of sequences. Analyses of phylogenetic diversity were performed independently for each cluster based on pairwise tree distances between all isolates (Supplementary Data 5).

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

Recolonization experiments of leaf-, root- and soil-derived bacteria on *Arabidopsis*. Calcined clay[16], an inert soil substitute, was washed with water, sterilized twice by autoclaving and heat-incubated until being completely dehydrated. *A. thaliana* Col-0 seeds were surface-sterilized with ethanol and stratified overnight at 4 °C. Leaf-, root- and soil-derived bacteria of the culture collections were cultivated in 96-deep-well plates and subsequently pooled (in equal or unequal ratios) in order to prepare synthetic bacterial communities (SynComs) for inoculations below the carrying capacity of leaves and roots[43,44]. To inoculate SynComs into the calcined clay matrix, OD_{600} was adjusted to 0.5 and 1 ml ($\sim 2.75 \times 10^8$ cells) was added to 70 ml 0.5× MS media (pH 7; including vitamins, without sucrose), and mixed with 100 g calcined clay in Magenta boxes ($\sim 2.75 \times 10^6$ cells per gr calcined clay), directly before sowing of surface-sterilized seeds. Plants were grown at 22 °C, 11 h light, and 54% humidity. Alive cell counts (CFUs) of root-associated bacteria by serial dilutions of root homogenates after seven weeks of co-incubation were $1.4 \times 10^8 \pm 8.4 \times 10^7$ cells per gram root tissue. For leaf spray-inoculation of *A. thaliana* plants, bacterial SynComs were prepared as described above and adjusted to OD_{600} 0.2, before the solution was diluted tenfold and 170 μl ($\sim 1.87 \times 10^6$ cells) were sprayed into each magenta box containing four three-week-old plants using a TLC chromatographic reagent sprayer (BS124.000, Biostep GmbH, Jahnsdorf, Germany). The average volume per spraying event was determined by spraying repeatedly into 50 ml tubes and weighing before and after. All plants and corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. All plants and corresponding unplanted clay samples were harvested under sterile conditions after a total incubation period of seven weeks. During harvest, leaves and roots of individual plants were carefully separated using sterilized tweezers and scissors to avoid cross-contamination and processed separately thereafter. All leaves being obviously contaminated with clay particles or touching the ground were carefully removed and omitted from further processing. Remaining aerial parts of four plants collected from one magenta box were combined and transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at −80 °C until used for DNA extraction. Roots from one Magenta box were pooled, washed twice in 5 ml PBS at 180 rpm for 20 min, dried on sterilized Whatman glass microfibre filters (GE Healthcare Life Sciences), transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at −80 °C until further processing. The corresponding unplanted clay samples were washed in 100 ml PBS supplemented with 0.02% Silwet L-77 at 180 rpm for 10 min, before particles were allowed to settle down for 5 min. The supernatant was collected by centrifugation at 3,220g for 15 min. The pellet was subsequently resuspended in 1 ml water, transferred into lysing matrix E tubes (MP Biomedicals), frozen in liquid nitrogen and stored at −80 °C.

To prepare DNA for bacterial 16S rRNA gene-based community analysis, all samples were homogenized twice by Precellys24 tissue lyser (Bertin Technologies), DNA was extracted and concentrations were measured by PicoGreen dsDNA Assay Kit (Life technologies), before bacterial 16S rRNA genes were amplified by degenerate PCR primers (799F and 1193R) targeting the variable regions V5-V7 (Supplementary Data 7). Each sample was amplified in triplicate (plus respective no template control) in 25 μl reaction volume containing 2 U DFS-Taq DNA polymerase, 1× incomplete buffer (both Bioron GmbH, Ludwigshafen, Germany), 2 mM $MgCl_2$, 0.3% BSA, 0.2 mM dNTPs (Life technologies GmbH, Darmstadt, Germany), 0.3 μM forward and reverse primer and 10 ng of template DNA. After an initial denaturation step at 94 °C for 2 min, the targeted region was amplified by 25 cycles of 94 °C for 30 s, 55 °C for 30 s and 72 °C for 60 s, followed by a final elongation step of 5 min at 72 °C. The three independent PCR reactions were pooled and the remaining primers and nucleotides were removed by addition of 20 U exonuclease I and 5 U Antarctic phosphatase (both New England BioLabs GmbH, Frankfurt, Germany) and incubated for 30 min at 37 °C in the corresponding 1× Antarctic phosphatase buffer. Enzymes were heat-inactivated and the digested mixture was used as template for the 2nd step PCR using the Illumina compatible primers B5-F and 1 of 96 differentially barcoded reverse primers (B5-1 to B5-96, Supplementary Data 7). All samples were amplified in triplicate for 10 cycles using identical conditions of the first-step PCR. Technical replicates of each sample were combined, run on a 1.5% (w/v) agarose gel and the bacterial 16S rRNA gene amplicons were extracted using the QIAquick Gel Extraction Kit (Qiagen) according to the manufacturer's instructions. DNA concentration was subsequently measured using the PicoGreen dsDNA Assay Kit (Life technologies) and 100 ng of each sample were combined. Final amplicon libraries were cleaned twice using the Agencourt AMPure XP Kit (Beckman Coulter GmbH) and subjected to sequencing on the Illumina MiSeq platform using an MiSeq Reagent kit v3 following the 2 × 350 bp paired-end sequencing protocol (Illumina Inc. USA).

Forward and reverse reads were joined, demultiplexed and subjected to quality controls using scripts from the QIIME toolkit[27], v. 180 (Phred ≥ 20). The resulting high quality sequences were further clustered at 97% sequence identity together with Sanger sequences of leaf, root and soil isolates using the UPARSE[24] pipeline as described above. Taxonomic assignments of representative sequences were performed as explained in the previous sections. OTUs only corresponding to one or more Sanger 16S rRNA gene sequence(s) of purified strains in the *At*-RSPHERE, *At*-LSPHERE or soil collection were selected and designated 'indicator OTUs'. The heat maps were generated using the ggplot2 R package.

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

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

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

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

nature COMMUNICATIONS

ARTICLE

Received 21 Sep 2015 | Accepted 16 Mar 2016 | Published 6 May 2016

DOI: 10.1038/ncomms11362

OPEN

Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi

Stéphane Hacquard[1,*], Barbara Kracher[1,*], Kei Hiruma[1,†], Philipp C. Münch[2,3,4], Ruben Garrido-Oter[1,5,6], Michael R. Thon[7], Aaron Weimann[3,5], Ulrike Damm[8,†], Jean-Félix Dallery[9], Matthieu Hainaut[10,11], Bernard Henrissat[10,11,12], Olivier Lespinet[13,14], Soledad Sacristán[15], Emiel Ver Loren van Themaat[1,†], Eric Kemen[1,6], Alice C. McHardy[3,5,6], Paul Schulze-Lefert[1,6] & Richard J. O'Connell[1,9]

The sessile nature of plants forced them to evolve mechanisms to prioritize their responses to simultaneous stresses, including colonization by microbes or nutrient starvation. Here, we compare the genomes of a beneficial root endophyte, *Colletotrichum tofieldiae* and its pathogenic relative *C. incanum*, and examine the transcriptomes of both fungi and their plant host *Arabidopsis* during phosphate starvation. Although the two species diverged only 8.8 million years ago and have similar gene arsenals, we identify genomic signatures indicative of an evolutionary transition from pathogenic to beneficial lifestyles, including a narrowed repertoire of secreted effector proteins, expanded families of chitin-binding and secondary metabolism-related proteins, and limited activation of pathogenicity-related genes *in planta*. We show that beneficial responses are prioritized in *C. tofieldiae*-colonized roots under phosphate-deficient conditions, whereas defense responses are activated under phosphate-sufficient conditions. These immune responses are retained in phosphate-starved roots colonized by pathogenic *C. incanum*, illustrating the ability of plants to maximize survival in response to conflicting stresses.

[1] Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [2] German Center for Infection Research (DZIF), Partner Site Hannover-Braunschweig, 38124 Braunschweig, Germany. [3] Computational Biology of Infection Research, Helmholtz Center for Infection Research, 38124 Braunschweig, Germany. [4] Max-von-Pettenkofer Institute, LMU Munich, German Center for Infection Research (DZIF), Partner Site LMU Munich, 80336 Munich, Germany. [5] Department of Algorithmic Bioinformatics, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany. [6] Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany. [7] Instituto Hispano-Luso de Investigaciones Agrarias (CIALE), Departamento de Microbiología y Genética, Universidad de Salamanca, 37185 Villamayor, Spain. [8] CBS-KNAW Fungal Biodiversity Centre, 3584 CT Utrecht, The Netherlands. [9] UMR BIOGER, INRA, AgroParisTech, Université Paris-Saclay, 78850 Thiverval-Grignon, France. [10] CNRS UMR 7257, Aix-Marseille University, 13288 Marseille, France. [11] INRA, USC 1408 AFMB, 13288 Marseille, France. [12] Department of Biological Sciences, King Abdulaziz University, 21589 Jeddah, Saudi Arabia. [13] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Sud, 91405 Orsay, France. [14] Laboratoire de Recherche en Informatique, CNRS, Université Paris-Sud, 91405 Orsay, France. [15] Centro de Biotecnología y Genómica de Plantas (UPM-INIA) and E.T.S.I. Agrónomos, Universidad Politécnica de Madrid Campus de Montegancedo, 28223 Madrid, Spain. * These authors contributed equally to this work. † Present addresses: Graduate School of Biological Sciences, Nara Institute of Science and Technology, Nara 630-0192, Japan (K.H.); Senckenberg Museum of Natural History Görlitz, 02826 Görlitz, Germany (U.D.); DSM Biotechnology Center, DSM Food Specialties B.V., Delft, The Netherlands (E.V.L.v.T.). Correspondence and requests for materials should be addressed to P.S.-L. (email: schlef@mpipz.mpg.de) or to R.J.O. (email: richard.oconnell@versailles.inra.fr).

Fungal endophytes are a ubiquitous and phylogenetically diverse group of organisms that establish stable associations with living plants, but in most cases their ecophysiological significance is poorly understood[1]. Species of the fungal genus *Colletotrichum* are best known as destructive pathogens on >3,000 species of dicot and monocot plants worldwide, causing anthracnose diseases and blights on leaves, stems, flowers and fruits[2]. However *Colletotrichum* species can also grow benignly as endophytes on symptomless plants[3], and although only few pathogenic members of the genus attack plant roots[4], *Colletotrichum* endophytes are frequently isolated from the roots of healthy plants[5,6]. Moreover, although the genome sequences and *in planta* transcriptomes were recently described for four species pathogenic on above-ground plant parts[2,7], such information is not available for any root-associated *Colletotrichum* pathogens or endophytes.

We found recently that *C. tofieldiae* (*Ct*) is an endophyte in natural populations of *Arabidopsis thaliana* growing in central Spain[8]. The fungus initially penetrates the rhizoderm by means of undifferentiated hyphae, which then ramify through the root cortex both inter- and intracellularly, occasionally spreading systemically into shoots via the root central cylinder without causing visible symptoms. Under phosphate-deficient conditions (50 μM KH_2PO_4), colonization by *Ct* promoted plant growth and fertility and mediated the translocation of phosphate into shoots, as shown by ^{33}P radiotracer experiments[8]. However, neither the plant growth promotion nor phosphate translocation activities were detectable under phosphate-sufficient conditions (625 μM KH_2PO_4), indicating that plant fitness benefits conferred by *Ct* are strictly regulated by phosphate availability. In striking contrast, colonization of *A. thaliana* roots by the closely related pathogenic species *C. incanum* (*Ci*), which attacks members of the Brassicaceae, Fabaceae and Solanaceae, severely inhibited *Arabidopsis* growth and mediated only low levels of ^{33}P translocation into shoots[8]. These findings raise the possibility that in low-phosphate soils, root colonization by the *Ct* endophyte compensates for the absence of key genetic components required for mycorrhizal symbiosis in the Brassicaceae lineage, which is otherwise conserved in ~80–90% of terrestrial plants[9].

In the present study, we report the genomes of five isolates of beneficial *Ct* and one isolate of pathogenic *Ci*, and analyse the transcriptomes of each species during their colonization of *Arabidopsis* roots under phosphate-deficient and phosphate-sufficient conditions. Comparison of the two species allows us to identify fungal adaptations to the endophytic lifestyle at the level of both gene repertoire and gene regulation, and provides insights into the evolutionary transition from parasitism to endophytism within a single fungal genus. On the host side, transcriptional responses of *Arabidopsis* roots to colonization by beneficial *Ct* are modulated by the phosphate status, providing evidence that trade-offs between defense and nutrition control the outcome of the interaction between *Arabidopsis* and *Ct*. Our findings also shed light on the ability of plants to maximize survival by prioritizing their responses to simultaneous biotic and abiotic stresses.

Results

Genome sequencing and evolution of *Ct* and *Ci* lifestyles. We sequenced the genome of the plant growth-promoting fungus *Ct* isolate 0861, a root endophyte isolated from natural populations of *A. thaliana* in Spain[8,10], and those of four other *Ct* isolates isolated from diverse dicot and monocot hosts in Europe (Supplementary Note 1). We also sequenced the broad host-range pathogen *Ci*, isolated from radish (*Raphanus sativus*) leaves in Japan, that strongly impairs plant growth when inoculated onto *Arabidopsis* roots[8,11] (Supplementary Fig. 1, Supplementary Table 1 and Supplementary Note 1). Illumina short reads were used to build high-quality genome assemblies of similar size for all isolates, ranging from 52.8 to 53.6 Mb (Supplementary Table 2 and Supplementary Note 2). Molecular phylogeny, whole-genome alignment and divergence date estimates indicate that *Ct* and *Ci* are closely related taxa within the *Colletotrichum spaethianum* species complex and diverged only ~8.8 million years ago (Fig. 1a, Supplementary Figs 2 and 3, Supplementary Table 3 and Supplementary Note 3). Our phylogenetic analysis suggests that evolution from pathogenic ancestors towards the beneficial endophytic lifestyle in *Ct* is a recent adaptation in *Colletotrichum* fungi.

SNP distribution and reproductive mode of *Ct* isolates. Although the five *Ct* isolates originate from widely separated geographical areas and distantly related plant hosts, they diverged only ~0.29 million years ago and the aligned fractions (>93%) of their genomes share >99% sequence identity (Fig. 1a and Supplementary Tables 1,3 and 4). The overall frequency of single-nucleotide polymorphisms (SNPs) between isolates was similar (2.22–3.04 SNPs per kb) but the SNP distribution within each genome was uneven, with alternating tracts of low (0.22–0.32 SNPs per kb) and high (4.25–5.12 SNPs per kb) SNP density (Fig. 1b, Supplementary Fig. 4 and Supplementary Table 5). This peculiar SNP distribution, also visible in the genomes of other plant-interacting fungi[12,13], is consistent with chromosome recombination events. However, the SNP density profiles are remarkably similar between isolates and large haplotype blocks are conserved between all (21%), four (19%), three (18%) or two (17%) of them, with only 22% being isolate specific (Fig. 1b,c, Supplementary Fig. 4, Supplementary Table 6 and Supplementary Note 4). These conserved SNP signatures in the genomes of geographically distant isolates were likely generated by rare or ancestral sexual/parasexual reproduction and maintained by frequent clonal propagation.

Evolutionary dynamics of multigene families in *Colletotrichum*. Similar numbers of protein-coding genes were predicted in *Ct*0861 and *Ci* (~13,000; Supplementary Table 2), with >11,300 orthologous genes shared between both species. By clustering protein-coding sequences into sets of orthologous genes using OrthoMCL, we identified 7,297 gene families conserved across all six analysed *Colletotrichum* species and 10,519 shared between *Ct*0861 and *Ci* (Fig. 2a,b and Supplementary Note 5). Using a maximum-likelihood approach, we also reconstructed ancestral genomes for each *Colletotrichum* lineage and predicted the number of gene families that were likely gained or lost in each species compared with its corresponding ancestor (Supplementary Fig. 5 and Supplementary Note 6). We found significantly more gene families gained (1,009) than lost (198) on the branch leading to *Ct* compared with other branches of the tree (Fisher's exact test, $P = 3.98 \times 10^{-136}$; Supplementary Fig. 5 and Supplementary Data 1). Functional enrichment analysis among the 1,009 gene families gained (Supplementary Fig. 5) and the 1,486 *Ct*-specific gene families (Fig. 2b) revealed a significant enrichment for genes encoding secondary metabolite biosynthesis-related proteins in *Ct* (Fisher's exact test, $P = 5.89 \times 10^{-3}$ and 3.31×10^{-8}, respectively). This result contrasts with the very low number of secondary metabolite-related genes detected in the genomes of other root-associated fungal endophytes and mycorrhizal fungi[14] and suggests that either fungal secondary metabolites have roles in establishing a beneficial endophytic interaction with host plants or in limiting

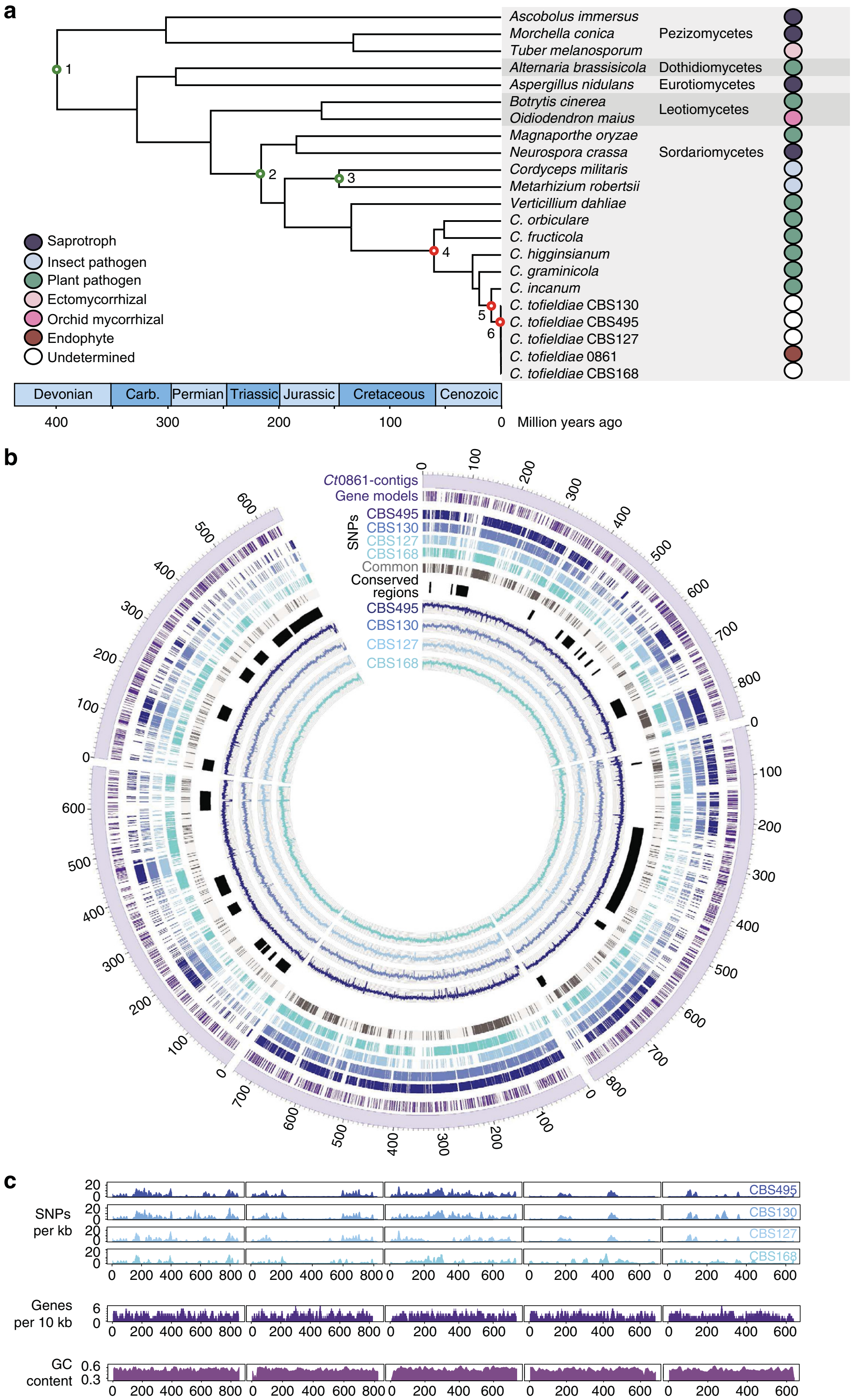


Figure 1 | *Colletotrichum* evolutionary divergence dates and SNP distribution in *C. tofieldiae* isolates. (**a**) Phylogeny of *Colletotrichum* species inferred from analysing 20 single-copy gene families using PhyML and r8s. Nodes 1–3 (green) are calibration points and nodes 4, 5 and 6 (red) represent estimated divergence dates (see Supplementary Note 3). (**b**) Circular visualization of the alignment of genome sequencing reads and SNP locations of four *C. tofieldiae* isolates with respect to the *Ct*0861 reference assembly. Tracks represent (from the outside) the five largest *Ct*0861 contigs (scale: kb); locations of predicted genes; locations of SNPs versus *Ct*0861 in CBS495, CBS130, CBS127, CBS168 (see Supplementary Table 1 for full culture IDs) and SNPs common to these four isolates; conserved regions with low SNP density between all the five isolates; mean read coverage (per 100 bases) for isolates CBS495, CBS130, CBS127 and CBS168. Coverage plot scales are 0 to 1,000 (CBS495) or 0 to 500 (CBS130, 127, 168). (**c**) SNP density (per 1 kb) in isolates CBS495, CBS130, CBS127 and CBS168 versus *Ct*0861, compared with gene density (per 10 kb) and GC content (%) on the five largest *Ct*0861 contigs.

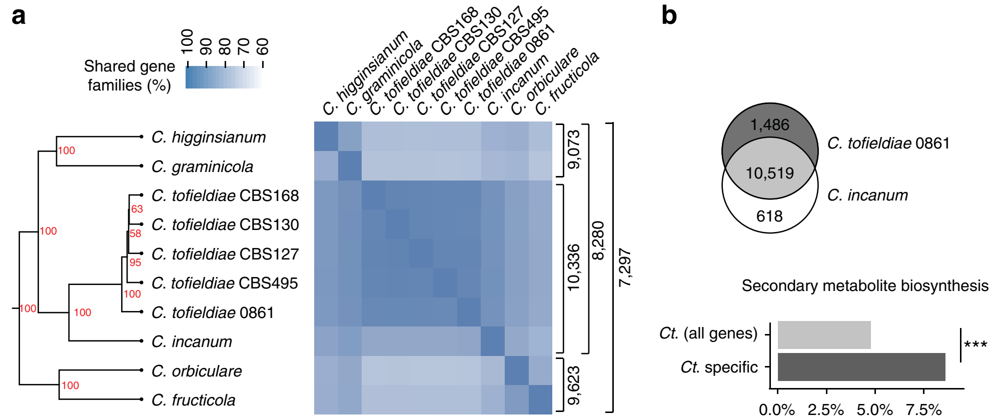


Figure 2 | Conservation of orthoMCL gene families within the proteomes of *Colletotrichum* species. (**a**) Heatmap and hierarchical clustering dendrogram depicting the percentage of gene families shared between 10 *Colletotrichum* genomes. Node labels in the tree indicate bootstrap support after 100 iterations. Brackets (right-hand side) indicate the number of gene families shared between the groups of genomes. (**b**) Upper panel: Venn diagram of gene families shared between the beneficial *C. tofieldiae* 0861 and its close pathogenic relative, C. *incanum*. Lower panel: Barplot showing the over-abundance of proteins related to secondary metabolite biosynthesis among gene families unique to *Ct*0861 compared with all *C. tofieldiae* gene families (Fisher's exact test; ***$P = 3.31\text{E}-08$).

the colonization of microbial competitors inside roots. Evaluation of the selective forces (d_N/d_S ratio) acting on all the protein families in the *Ct* genome revealed that genes involved in 'signal transduction mechanisms', 'RNA processing and modification' and 'lipid transport and metabolism' showed the strongest evidence of adaptive evolution (false discovery rate (FDR) < 0.05, Fisher's test). This contrasts with pathogenic *Colletotrichum* species for which gene families belonging to the categories 'defense mechanisms', 'cell wall/membrane/envelope biogenesis' and 'RNA processing and modification' show the highest d_N/d_S ratios (Supplementary Figs 6 and 7, Supplementary Data 2 and Supplementary Note 7).

Genomic signatures of the pathogenic to beneficial transition. *Ct* encodes large repertoires of transporters, secreted proteins, proteases, carbohydrate-active enzymes (CAZymes) and secondary metabolism key enzymes, very similar to *Ci* and four other pathogenic *Colletotrichum* species (Supplementary Figs 8–12 and Supplementary Note 8). By comparing the *Ct* gene repertoires to those of five other plant-associated fungal endophytes from both ascomycete and basidiomycete lineages, we found no obvious common genomic signatures to indicate the convergent evolution of an endophyte 'toolkit' (Supplementary Figs 8–12). Furthermore, the convergent loss of decay mechanisms characteristic of ectomycorrhizal fungi[15,16] is not a hallmark shared by the non-mycorrhizal root endophytes (Supplementary Fig. 12), suggesting that these fungi have followed different evolutionary trajectories to acquire the ability for intimate growth in living root tissues[14,17].

Despite the overall similar secretome size of all analysed *Colletotrichum* species (13.3–15.9% of the total proteome), the proportion of genes encoding candidate secreted effector proteins (CSEPs), which may promote fungal infection[18], varied considerably between species (6.6–15.8% of the total secretome; Fig. 3a and Supplementary Table 7). The smaller CSEP repertoire in *Ct*0861 (133 versus 189 in *Ci*) is largely explained by the reduction of species-specific CSEPs (34 versus 72 in *Ci*; Fig. 3a, Supplementary Fig. 13, Supplementary Table 7 and Supplementary Data 3). As expected, calculation of d_N/d_S ratios among 331 *CSEP* families derived from all the 10 analysed *Colletotrichum* genomes indicates they are under diversifying selection (median 0.35, interquartile range 0.21–0.49) relative to non-CSEP families (median 0.20, interquartile range 0.07–0.33; Fisher's exact test, $P < 2.2 \times 10^{-16}$; Fig. 3b). Genomes from additional *Ci* isolates are now needed to determine whether there is differential host-selective pressure on the CSEP repertoires of endophytic *Ct* and pathogenic *Ci* that reflect their contrasting lifestyles. Similar to other *Colletotrichum* species[2], *CSEPs* in *Ct* and *Ci* are not organized into large multigene families, possibly due to a low frequency of duplication events in their respective genomes (Fig. 3c,d and Supplementary Table 2).

Both *Ct* and *Ci* genomes encode a very broad range of CAZymes, including large arsenals of pectate lyases, carbohydrate esterases and glycoside hydrolases acting on all major plant cell wall constituents (Fig. 4a, Supplementary Fig. 12 and Supplementary Data 4). However, the number of predicted carbohydrate-binding modules is inflated in *Ct* compared with pathogenic *Colletotrichum* species, especially chitin-binding CBM18 (48 versus 28–40) and CBM50 (57 versus 30–54) modules (Fig. 4a, Supplementary Data 4), though few of the corresponding *Ct* genes were induced *in planta* (Supplementary Fig. 14). These two chitin-binding modules are similarly highly enriched in the genomes of two other non-mycorrhizal root symbionts[19,20] (*Piriformospora indica* and *Harpophora oryzae*; Supplementary Data 4), suggesting this is a genomic signature common to independently evolving root-associated fungal endophytes.

Dual RNAseq of *Arabidopsis* roots and fungal partners. We report elsewhere that *Ct* promotes *Arabidopsis* growth under phosphate-deficient (− P) but not phosphate-sufficient (+ P) conditions and that transfer of radioactive ^{33}P from *Ct* hyphae to host plants is strictly regulated by Pi (inorganic phosphate) availability[8]. To compare the transcriptional dynamics of beneficial *Ct* and pathogenic *Ci* during colonization of *Arabidopsis* roots and study the corresponding host responses, we extensively re-analysed the previously created RNA-seq data for the *Ct*-*Arabidopsis* interaction (6, 10, 16 and 24 days post inoculation (d.p.i.), + P: 625 μM, − P: 50 μM; ref. 8) and included new samples for the *Ci*-Arabidopsis interaction (10 and 24 d.p.i., − P: 50 μM) (Supplementary Figs 15 and 16). After mapping Illumina reads to their respective genomes, we obtained expression data for > 20,000 *Arabidopsis* genes, 8,613 *Ci* genes and 6,693 *Ct* genes (Supplementary Fig. 17, Supplementary Table 8 and Supplementary Note 9). The expression data were validated using quantitative PCR with reverse transcription

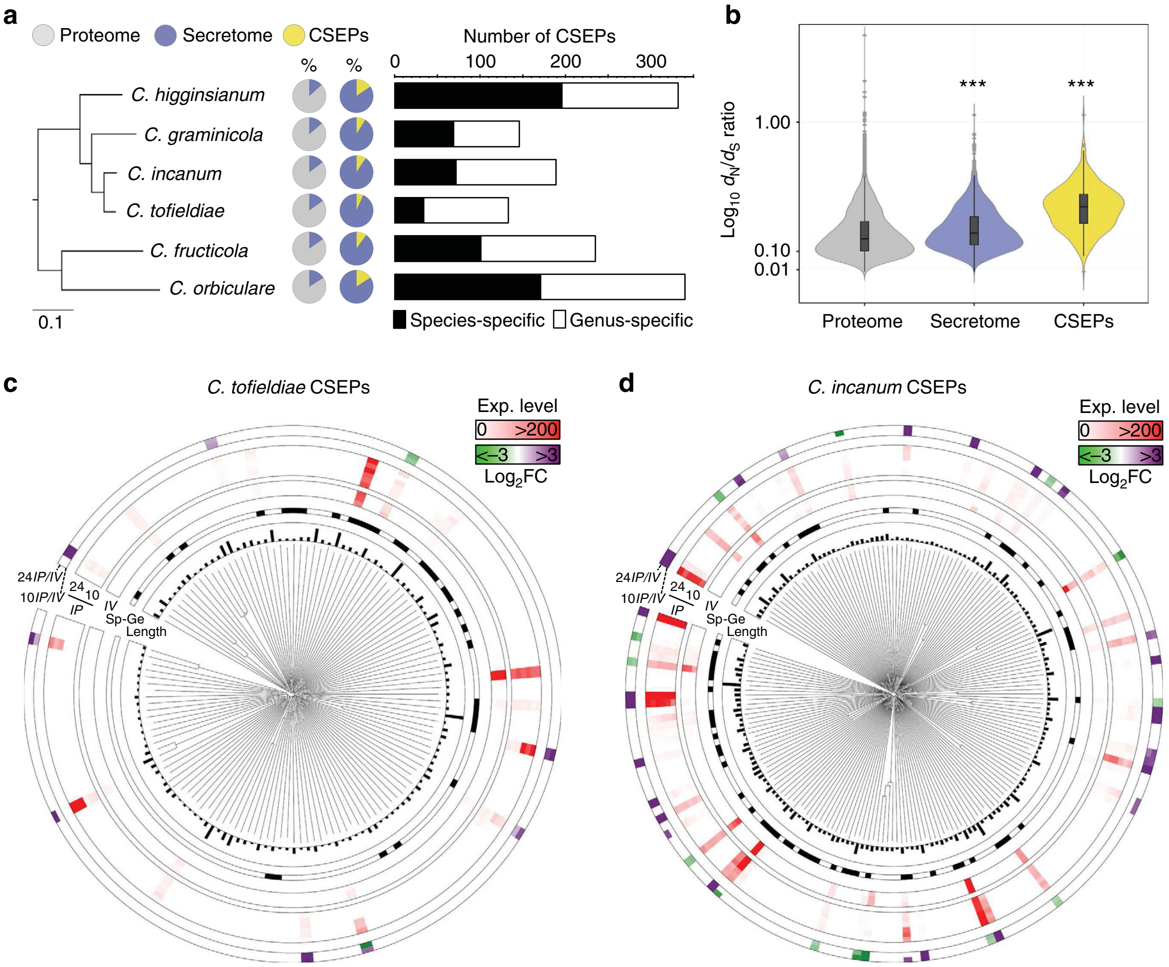


Figure 3 | Conservation and expression of genes encoding candidate secreted effector proteins in *C. tofieldiae* and *C. incanum.* (**a**) Proportions of predicted secreted proteins (circles, violet sectors) and candidate secreted effector proteins CSEPs (circles, yellow sectors) in the proteomes and secretomes of *Colletotrichum* species, respectively. The number of genus- and species-specific CSEPs detected for each species is indicated in the barplot. (**b**) Boxplot with a rotated kernel density on each side showing d_N/d_S ratio ($\log_{10}$) measured in the proteome, the secretome and the CSEP repertoires of 10 *Colletotrichum* isolates using the gene families defined by MCL clustering (see Fig. 2). The overall d_N/d_S ratio is significantly higher for gene families encoding secreted proteins and CSEPs compared with the remaining gene families (One-sided Fisher's test, $^{***}P<0.001$). (**c**,**d**) Expression and regulation of *CSEPs* in *C. tofieldiae* 0861 (**c**) and *C. incanum* (**d**). The circular plots show (from the inside): dendrograms of the CSEPs based on protein sequence alignments, CSEP length (0–500 amino acids), species-specific (Sp, black) and genus-specific (Ge, white) CSEPs, normalized gene expression (Exp.) levels *in vitro* (*IV*) and *in planta* (*IP*) at 10 and 24 days post inoculation, *CSEPs* significantly up- (violet) and downregulated (green) at 10 days post inoculation versus *in vitro* (10*IP/IV*) and 24 days post inoculation versus *in vitro* (24*IP/IV*) ($|\log_2FC|\geq1$, FDR<0.05).

(RT–qPCR) with a subset of *Arabidopsis* and *Ct* genes (Supplementary Fig. 18, Supplementary Table 9 and Supplementary Note 10).

Transcriptional shutdown of pathogenicity genes in *Ct*. Among the 3,885 *Ct* genes significantly regulated (moderated *t*-test, $|\log_2FC|\geq1$, FDR<0.05), only few (61) were impacted by phosphate status (described in ref. 8) or the fungal developmental stage *in planta* (845; Supplementary Data 5 and Supplementary Fig. 19). In contrast, ~80% were induced upon host contact and particularly those encoding CAZymes, for which a dynamic expression pattern was observed (Fig. 4b and Supplementary Figs 19 and 20). A first wave of activation (6–16 d.p.i.) involved few plant cell wall-degrading enzymes (PCWDEs) acting mostly on hemicellulose, while a second wave (24 d.p.i.) involved induction of numerous PCWDEs acting on all major wall polymers, including cellulose, hemicellulose and pectin (Fig. 4b). Thus, at later infection stages, *Ct* displays significant saprotrophic capabilities. However, genes encoding CSEPs, secreted proteases, secondary metabolism key enzymes and transporters showed no clear activation (Supplementary Fig. 21), in contrast to the highly stage-specific deployment of such genes by *C. higginsianum* during infection of *Arabidopsis* leaves[2]. Surprisingly, the activation of *Ct CSEPs* was almost non-existent *in planta*, with only 18/133 expressed during colonization, 8/133 induced *in planta* ($\log_2FC\geq1$) and 4/133 ranking among the 1,000 most highly expressed genes (Fig. 3c). These few expressed *CSEP* genes showed similar d_N/d_S ratios compared with *CSEPs* that were silent *in planta* (Supplementary Data 3). The contracted repertoire and small number of *CSEPs* activated *in planta* suggests *Ct* requires extremely few effectors for host invasion and maintenance of the beneficial relationship.

Gene deployment *in planta* reflects fungal lifestyles. To uncover transcriptional adaptations associated with the evolutionary transition from the ancestral pathogenic lifestyle to beneficial endophytism, we compared the normalized expression levels of 6,804 *Ct* and *Ci* orthologous gene pairs that are expressed

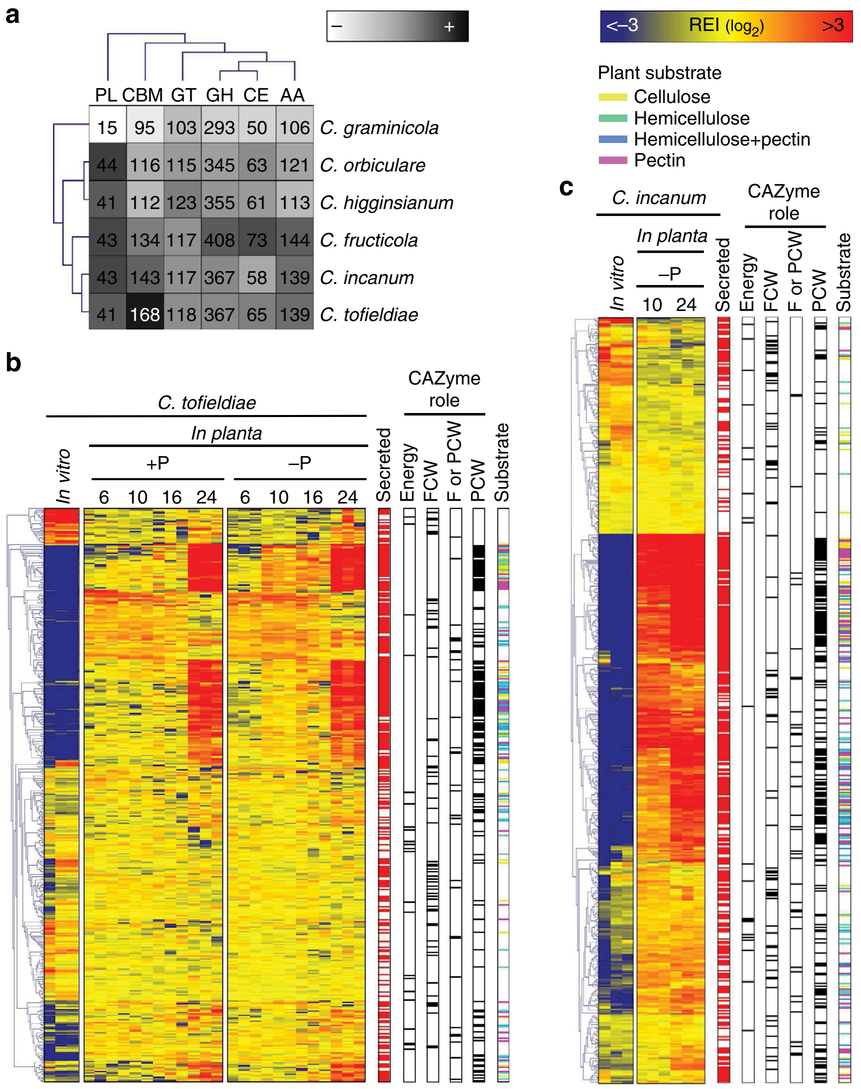


Figure 4 | *Colletotrichum* CAZyme repertoires and their transcriptional regulation in *C. tofieldiae* and *C. incanum*. (**a**) Hierarchical clustering of CAZyme classes from the genomes of four *Colletotrichum* species. AA, auxiliary activities; CBM, carbohydrate-binding module CE, carbohydrate esterase; GH, glycoside hydrolase; GT, glycosyltransferase; PL, polysaccharide lyase. The numbers of enzyme modules in each genome are shown. Overrepresented (dark grey to black) and underrepresented (pale grey to white) modules are depicted as $\log_2$ (fold changes) relative to the class mean. (**b**) Transcript profiling of *C. tofieldiae* CAZyme genes *in vitro* and during colonization of *Arabidopsis* roots at 6, 10, 16 and 24 days post inoculation (d.p.i.) under phosphate sufficient (+ P: [625 μM]) and deficient (− P: [50 μM]) conditions. (**c**) Transcript profiling of *C. incanum* CAZyme genes *in vitro* and during colonization of *Arabidopsis* roots at 10 and 24 d.p.i. under phosphate-deficient conditions (− P: [50 μM]). (**b**,**c**) Overrepresented (yellow to red) and underrepresented transcripts (yellow to blue) are shown as $\log_2$ (fold changes) relative to the mean expression across all the stages. The red marks represent secreted CAZymes and the black marks indicate involvement in metabolic activities linked to energy storage and exchange (Energy), or degradation of fungal cell walls (FCW), plant cell walls (PCW) or both (F or PCW). For CAZymes acting on PCW, the corresponding plant substrates (cellulose, hemicellulose, hemicellulose and pectin, pectin) are indicated by a colour code. REI, relative expression index.

in planta (10, 24 d.p.i.; − P) (Supplementary Data 6). More than twice as many gene pairs were differentially expressed at 10 d.p.i. (621 up, 842 down) than at 24 d.p.i. (306 up, 273 down; moderated *t*-test, $|\log_2 FC| \geq 1$, FDR < 0.05), suggesting that early colonization events are critical for determining the outcome of the interaction. GO term enrichment analysis showed that processes related to melanin biosynthesis were significantly enriched in *Ct*, consistent with the formation of melanized microsclerotia in *Ct* but not *Ci*[8] (Supplementary Table 10). We also found major differences between *Ct* and *Ci* in the expression of gene categories typically associated with fungal pathogenicity. *In planta* activation of *CSEPs* was more pronounced in *Ci* compared with *Ct*, with seven times more *CSEPs* highly expressed (top 1,000 expressed genes) and three times more upregulated *in planta* at 10 d.p.i. (Fig. 3c,d and Supplementary Data 5 and 7). Likewise genes encoding CAZymes and secondary metabolism enzymes displayed earlier and stronger transcriptional activation *in planta* and broader diversity in *Ci* (Fig. 4b,c and Supplementary Fig. 22). Consistent with this, we observed a reduced number of living cells and a depletion of beta-linked polysaccharides (including cellulose) from host cell walls in *Ci*-colonized roots at 10 d.p.i., but not in *Ct*-colonized roots (Supplementary Fig. 1). This finding suggests that pathogenic *Ci* harvests carbon from plant cell walls more aggressively than *Ct*. Thus, despite their phylogenetic proximity and similar gene arsenals, gene deployment during infection was strikingly different between *Ct* and *Ci*. The *in planta* transcriptome of *Ci* resembles that of other pathogenic *Colletotrichum* species[2], whereas the less dynamic transcriptome of *Ct* might contribute to, or be a consequence of, the beneficial relationship. Overall, our results suggest that the recent transition from pathogenic to beneficial lifestyles might be partly controlled through transcriptional downregulation of pathogenicity-related genes in *Ct*.

Host responses to *Ct* are phosphate-status dependent. To disentangle how Pi-starved and non-starved *Arabidopsis* roots respond to *Ct* colonization over time, we compared *Ct*-colonized and mock-inoculated roots under +P and −P conditions. In total, 5,661 *Arabidopsis* genes were differentially expressed in at least one of the 16 pair-wise comparisons (moderated *t*-test, $|\log_2 FC| \geq 1$, $FDR < 0.05$) and grouped into 20 major gene expression clusters (Fig. 5a and Supplementary Data 8). GO term enrichment analysis among these clusters indicated that the phosphate level used in our study (50 μM) was sufficient to provoke a phosphate starvation response in *Arabidopsis* roots (clusters 2 and 4; Fig. 5b). Furthermore, our analysis indicates that 'response to stimulus', 'indole glucosinolate metabolic process', 'defense response' and 'ethylene metabolic process' are activated in *Ct*-colonized roots under +P but not −P conditions (cluster 9) (Fig. 5b and Supplementary Data 9). In contrast, the genes related to 'root cell differentiation' (cluster 8, Fig. 5b) and phosphate uptake[8] were preferentially activated in Pi-starved *Arabidopsis* roots during *Ct* colonization, similar to mycorrhizal symbiont–host interactions[21]. To identify key regulatory genes (hub genes) that might orchestrate transcriptional reprogramming in the contrasting directions seen in clusters 8 and 9, we checked which of these genes are often co-regulated in other expression data sets using the ATTED-II gene co-expression database (Fig. 5c). Among the hub genes that showed high connectivity within cluster 8 (highlighted with black dots), many encode proteins involved in cell wall remodelling and root hair development. Particularly, genes encoding the root hair-specific proteins RHS8, RHS12, RHS13, RHS15 and RHS19 (ref. 22) are upregulated (moderated *t*-test, $|\log_2 FC| \geq 1$, $FDR < 0.05$) in *Ct*-colonized versus mock-treated roots under −P conditions, which was validated by RT–qPCR (Fig. 5d and Supplementary Fig. 18). This expression pattern suggests that *Ct*-dependent remodelling of root architecture might play a key role to enhance phosphate uptake during starvation (Supplementary Note 9). Similarly, we identified 27 hub genes within cluster 9 (Fig. 5c, black dots), encoding well-characterized defense-related proteins such as the transcription factors WRKY33 and WRKY40 (ref. 23), the ethylene-responsive factors ERF11 and ERF13 (ref. 24), as well as MYB51 (ref. 25), a transcription factor regulating Tryptophan (Trp)-derived indole glucosinolate metabolism. Four other genes involved in indole glucosinolate metabolism were also highly differentially regulated in cluster 9, including the myrosinase *PEN2* and the P450 monooxygenase *CYP81F2* required for the biosynthesis of 4-methoxy-indol-3-ylmethylglucosinolate, the substrate of PEN2 myrosinase[26,27] (Supplementary Data 9). The PEN2-dependent metabolism of Trp-derived indole glucosinolates in *A. thaliana* is activated upon perception of pathogen-associated molecular

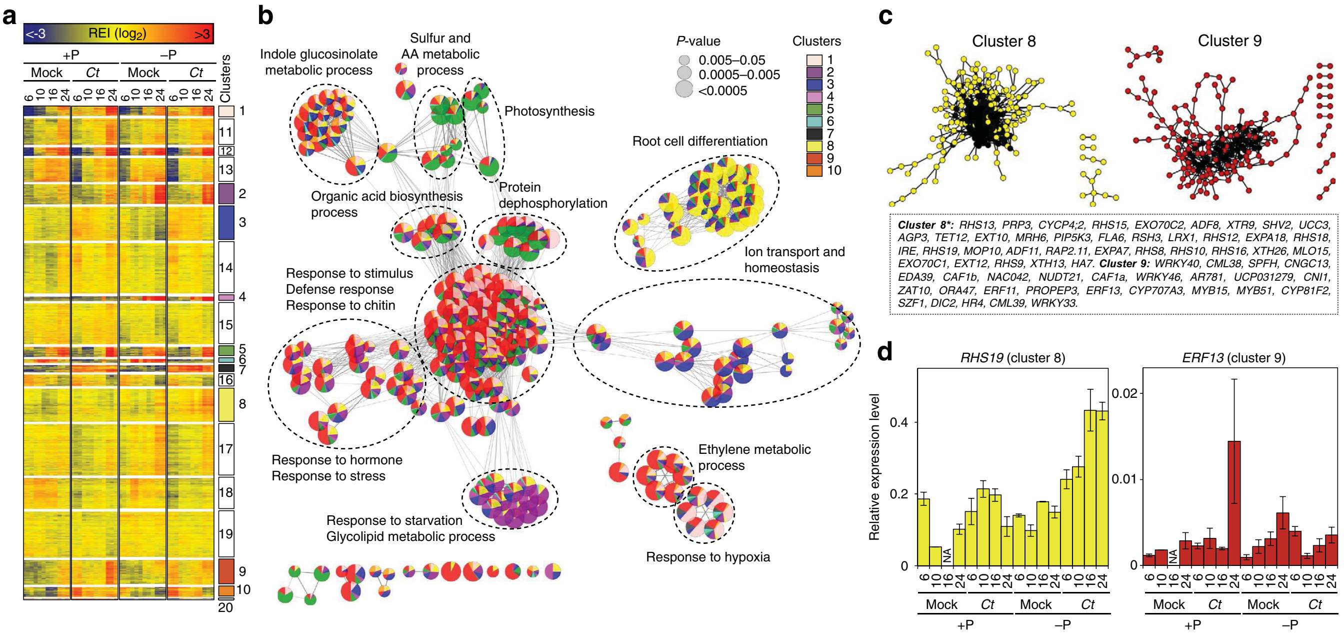


Figure 5 | Transcriptional reprogramming of Pi-starved and non-starved *Arabidopsis* roots in response to *C. tofieldiae*. (**a**) Transcript profiling of 5,561 *Arabidopsis* genes significantly regulated (moderated *t*-test, $|\log_2 FC| \geq 1$, $FDR < 0.05$) between colonized versus mock-treated roots and phosphate-starved (−P: [50 μM]) versus non-starved roots (+P: [625 μM]) at 6, 10, 16 and 24 days post inoculation. Overrepresented (yellow to red) and underrepresented transcripts (yellow to blue) are shown as $\log_2$ (fold changes) relative to the mean expression across all stages. Using *k*-means partitioning, the gene set was split into 20 major gene expression clusters. (**b**) Gene Ontology term enrichment network analysis among the 10 clusters highlighted in **a**. Each significantly enriched GO term ($P < 0.05$, hypergeometric test, Bonferroni step-down correction) is represented with a circle and the contribution (%) of each cluster to the overall GO term enrichment is represented using the same colour code as in **a**. As tightly connected GO terms are functionally linked, only the major host responses outputs are indicated (dotted line). (**c**) For cluster 8 and cluster 9, gene relationships based on co-regulation were assessed using other *Arabidopsis* expression data sets (see Supplementary Note 9). The genes within each cluster that show strong expression relationships in other expression data sets are likely to encode key regulatory hubs. Hub genes (cluster 9: ≥5 connections, *cluster 8: ≥10 connections) are highlighted in black. The corresponding characterized *Arabidopsis* genes are indicated below the co-expression networks. (**d**) Validation of the expression profiles of the hub genes *RHS19* (cluster 8) and *ERF13* (cluster 9) using RT–qPCR (see Supplementary Note 10). Error bars indicate standard error ($n = 3$ biological replicates), NA, data not available; REI, Relative Expression Index.

patterns by receptors of the innate immune system and is needed for broad-spectrum defence to restrict the growth of fungal pathogens[26,27]. Notably, in *Arabidopsis* mutants that cannot activate PEN2-mediated antifungal defense, the promotion of plant growth by *Ct* is impaired, while the depletion of all Trp-derived secondary metabolites renders *Ct* a pathogen on *Arabidopsis*[8]. These findings strongly suggest that the phosphate starvation response and Trp-derived indole glucosinolate metabolism are interconnected to control fungal colonization of *Arabidopsis* roots[28]. Phosphate status-dependent activation of defense responses was also observed among the 411 expressed *Arabidopsis* genes annotated as 'chitin-responsive' (Supplementary Fig. 23), based on GO term enrichment among all significantly regulated genes (Supplementary Fig. 24) and this was validated by RT–qPCR (Fig. 5d and Supplementary Fig. 18). These data reveal a remarkable capacity of *Arabidopsis* roots to prioritize different transcriptional outputs in response to *Ct*, favouring either defense responses under +P conditions or root growth and phosphate metabolism under −P conditions.

Phosphate-starved roots activate defense responses to *Ci*. To clarify whether the reduced activation of defense responses observed in *Ct*-colonized roots under −P conditions is not simply due to phosphate deficiency, we compared the transcriptomes of Pi-starved *Arabidopsis* roots in response to either *Ci* or *Ct* at 10 d.p.i. In total, 2,009 differentially expressed genes were identified (moderated *t*-test, $|\log_2 FC| \geq 1$, FDR < 0.05), including 988 genes induced in *Ct*-colonized roots (cluster 1) and 1,021 genes in *Ci*-colonized roots (cluster 2; Fig. 6a and Supplementary Data 10). GO term enrichment analysis revealed that ion transport and root cell differentiation mechanisms were activated in *Ct*-colonized roots, whereas strong defense responses were triggered in *Ci*-colonized roots (Fig. 6b). Thus, although Pi-starved *Arabidopsis* roots remain able to mount immune responses against pathogenic *Ci*, transport and root growth are instead prioritized during interaction with beneficial *Ct*.

Discussion

Deciphering the genetic basis of the transition from pathogenic to beneficial plant-fungal interactions is crucial for a better understanding of the evolutionary history of fungal lifestyles[20,29]. It was recently shown that the ectomycorrhizal lifestyle arose independently multiple times during evolution and that the transition was associated with (1) convergent loss of genes encoding PCWDEs present in their saprotrophic ancestors and (2) the repeated evolution of lineage-specific 'toolkits' of mycorrhiza-induced genes[15]. However in striking contrast with ectomycorrhizal fungi, this transition in *Ct*, *P. indica* and *H. oryzae* was not accompanied by contraction of their PCWDE repertoires[19,20]. In our study, the close phylogenetic relatedness of beneficial *Ct* and pathogenic *Ci*, and their ability to infect the same plant host, allowed us to resolve both genomic and transcriptomic signatures associated with this evolutionary transition. The overall high genomic similarity between *Ct* and *Ci* suggests that this transition involved only subtle remodelling of the gene repertoire (that is, a reduced set of CSEPs and expansion of chitin-binding and secondary metabolism-related protein families). The retention of abundant pathogenicity- or saprotrophy-related genes implies that they are still needed by *Ct*, perhaps for exploitation of other plant hosts or during plant senescence when *Arabidopsis* leaves are extensively colonized by *Ct* mycelium[8]. Our results also suggest that changes in fungal

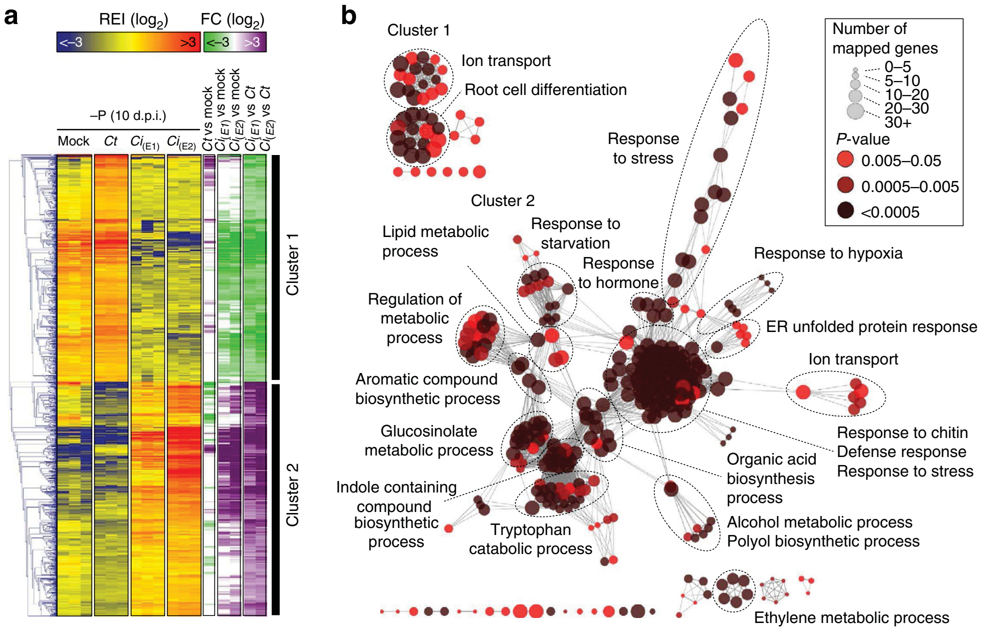


Figure 6 | Comparative transcriptome analysis of *Arabidopsis* roots in response to beneficial *C. tofieldiae* and pathogenic *C. incanum*. (**a**) Transcript profiling of 2,009 *Arabidopsis* genes significantly regulated (moderated *t*-test, $|\log_2 FC| \geq 1$, FDR < 0.05) between *C. incanum*- versus (vs) *C. tofieldiae*-colonized roots at 10 days post inoculation (d.p.i.) under phosphate-deficient conditions (−P: 50 μM). Overrepresented (yellow to red) and underrepresented transcripts (yellow to blue) are shown as log$_2$ (fold changes) relative to the mean expression across all stages. E1 and E2 correspond to two fully independent experiments (see Supplementary Note 9). Gene expression fold changes (green: downregulated; violet: upregulated) were calculated between *C. tofieldiae*-colonized versus mock-treated roots, *C. incanum*-colonized versus mock-treated roots or *C. incanum*-colonized versus *C. tofieldiae*-colonized roots. (**b**) GO term enrichment analysis of *Arabidopsis* genes preferentially expressed in response to *C. tofieldiae* (Cluster 1) or in response to *C. incanum* (cluster 2). Each circle corresponds to a significantly enriched GO term ($P < 0.05$, hypergeometric test, Bonferroni step-down correction). The colour code reflects *P* values and the circle size the number of genes associated to each GO term. Similar to Fig. 5b, the GO terms that are tightly connected are functionally linked and therefore only the major host-response outputs are indicated (dotted line). REI, Relative Expression Index.

gene expression patterns during host colonization, rather than extensive remodelling of the gene repertoire, provides an alternative and probably transient adaptation to a beneficial endophytic lifestyle. This may reflect the relatively recent transition from pathogenic to non-pathogenic lifestyles in *Ct* and, consequently, a latent capacity to revert to a pathogenic lifestyle.

During the last decade, the molecular mechanisms by which plants respond to colonization by pathogenic or mutualistic fungi have been extensively studied[30]. However, it remains unclear how plants discriminate and respond appropriately to closely related fungal partners with different lifestyles. The sedentary nature of plants suggests they have evolved regulatory systems to integrate exposure to conflicting biotic and abiotic stresses and balance their resource allocation strategically to maximize growth and survival. A recent report showed that plant responses to multiple stresses are not cumulative and suggested that prioritization of stress responses does take place[31]. For plant–mycorrhizal associations, an inverse correlation was observed between phosphate levels and the number of arbuscules formed in roots[32]. Although the detailed molecular mechanism remains unclear, this suggests that the nutritional status of the plant impacts fungal colonization efficiency. Here, we show that host transcriptional responses to *Ct* are dependent on phosphate availability, with defense responses activated or suppressed under high- or low-phosphate conditions, respectively. The fact that immune responses are retained in phosphate-starved roots colonized by *Ci* makes it unlikely that metabolic competition between phosphate starvation and defense response systems attenuates defense gene activation during interactions with *Ct* under P-limiting conditions. Recently, a metabolic link between the phosphate starvation response and glucosinolate biosynthesis was described[28] and the functional relevance of this link is supported by our observation that *Ct*-mediated plant growth promotion is impaired in *Arabidopsis* mutants lacking regulatory components of indole glucosinolate metabolism or the phosphate starvation response[8]. Therefore, we hypothesize that connectivity between nutrient sensing and innate immunity systems in the host, combined with subtle genomic adaptations in *Ct*, has enabled the transition from pathogenic to beneficial *Arabidopsis*–*Colletotrichum* interactions (Supplementary Fig. 25). Consequently, the interaction with beneficial *Ct*, but not with pathogenic *Ci*, is tightly controlled in plant roots by trade-offs between nutrition and defense. Whether phosphate stress-dependent defense attenuation renders *Ct*-colonized plants super-susceptible to other microbial pathogens remains to be tested. Our results are consistent with the fact that transfer of Pi from ramifying fungal hyphae to roots, and subsequent allocation to shoots for plant growth, occurs only under phosphate-deficient conditions[8]. Notably, where *Ct* naturally associates with *Arabidopsis* in central Spain, the level of bioavailable phosphate in soil at those locations is very low (5.5 to 17 p.p.m., Supplementary Table 11). Our findings suggest that both innate immune responses (that is, indole glucosinolate metabolism) and soil phosphate availability are important selective forces driving fungal adaptation and contributing to the evolutionary transition from parasitic to beneficial *Arabidopsis*–fungal associations.

Methods

Genome sequencing and assembly. *C. incanum* and the five *C. tofiediae* isolates were grown in liquid Mathur's medium (2.8 g glucose, 1.22 g $MgSO_4.7H_2O$, 2.72 g KH_2PO_4 and 2.18 g Oxoid mycological peptone in 1 l deionized water) supplemented with 100 $\mu g\,ml^{-1}$ rifampicin and 125 $\mu g\,ml^{-1}$ streptomycin. Genomic DNA was isolated using the DNeasy Plant Mini Kit (Qiagen) from 100 mg of fungal mycelium. Library construction, quality control and DNA sequencing for 454 GFLX+ or Illumina Hiseq sequencing were performed at the Max Planck Genome Centre Cologne (http://mpgc.mpipz.mpg.de) using 1 µg genomic DNA. After the preparation of genomic DNA libraries, 454 reads (557 bp on average) and Illumina paired-end reads (100 bp) were obtained from Roche 454 FLX+ and Illumina HiSeq2500 sequencers, respectively. For the *Ct*0861 reference genome, a hybrid assembly strategy was used combining 454 and Illumina data. Unpaired 454 reads were first assembled using MIRA 4.0 (ref. 33) and filtered MIRA-contigs (> 5,000 bp) were further used for scaffolding of Illumina paired read assemblies from SPAdes 3.0 (ref. 34). The established SPAdes 3.0 pipeline was used in 'careful' mode providing 454 MIRA assemblies as untrusted-contigs for scaffolding only and a kmer scan using 21, 31, 41, 61, 75 and 81. All other assemblies were constructed only from Illumina data using a combination of VELVET 1.2.1 (ref. 35) and SPAdes[34]. Using BLASTN searches, contigs were identified that were missing from combined SPAdes assemblies but present in VELVET assemblies. To integrate those contigs and extend further where possible, SPAdes was re-run as described above but in 'trusted-contigs' mode where trusted contigs were provided as fasta files with absent contigs only. All the assemblies were generated using 'careful' mode in SPAdes to avoid miss-pairing of contigs by scaffolding and for further analyses, contigs < 100 bp were removed. To identify and remove potential contaminating sequences, assemblies were aligned to the genomes of *A. thaliana*, *H. sapiens* and PhiX (sequencing spike-in control) using MUMmer[36] with default parameter settings. Contigs that aligned with more than 50% of their sequence (coverage; 'COV') and at least 85% sequence identity ('IDY') to any of the tested contaminants were removed from the assemblies. In addition, contigs that aligned with 75–85% identity (and > 50% coverage) or with 10–50% coverage (and > 85% identity) were also removed, if the judgment of the sequence being non-fungal was confirmed through BLASTN searches in the NCBI nr database (with default settings). For the *Ct*0861 assembly, RNA-sequencing data were used for further clean-up. Finally, assembly quality was assessed on the basis of L50/75/90 and N50/75/90 values, percentage of error-free bases estimated with REAPR[37] (version 1.0.16, default settings) and gene space coverage estimated with CEGMA[38] (version 2.0, default settings).

Repetitive DNA analysis. We identified repetitive DNA in the genome assemblies using either *de novo* or homology approaches. For *de novo* searches, we used PILER and PALS[39] to identify repetitive sequences and classify them into families. The resulting libraries of consensus sequences were then used to scan the genome sequences using RepeatMasker[40] (version 4.0.3) to identify individual repetitive elements. For homology-based searches, we used RepeatMasker using a library of all fungal elements in the Repbase database[41] (version 20140131).

Phylogeny and divergence date estimation. All phylogenetic analyses performed in this study are described in the Supplementary Note 11. For evolutionary divergence date estimation, clustering, protein family selection and phylogenetic analyses were performed with scripts in the Mirlo package (https://github.com/mthon/mirlo). The phylogeny was calibrated using the penalized-likelihood method implemented in r8s (ref. 42) using one primary and two secondary calibration points (Supplementary Note 11).

Short-read alignment and SNP analysis. To compare the genome sequences of *Ct* isolates, Illumina short reads of the four other isolates were mapped onto the genome assembly of *Ct*0861 using Bowtie2 (ref. 43) (default settings for paired-end data). Subsequently, duplicate reads were removed using the rmdups function from the SAMtools toolkit[44] (default settings). On the basis of the mapped genome sequencing reads, single-nucleotide polymorphisms (SNPs) were identified using the mpileup function in SAMtools[44] (version 0.1.18; with option -u) The obtained SNP sets were filtered by applying the bcftools script vcfutils.pl varFilter (SAMtools) with adjusted read depth settings according to the respective sequencing read coverage to -d 80 and -D 800 for CBS495 and to -d 40 and -D 400 for CBS130, CBS127 and CBS168. The SNP locations, read coverage for each isolate and locations of conserved regions were visualized using the Circos software package[45] (version 0.62.1). In addition, we also calculated SNP densities (SNPs per kb) relative to *Ct*0861 for each isolate as a function of the genomic location on all *Ct*0861 contigs larger than 50 kb, using a 10-kb sliding window that moved 1 kb at each step. For visualization of the SNP densities, these windows were sorted in the increasing order by contig number and position on the contig. To identify windows with a low SNP density, that is, a common haplogroup, between isolates we classified the SNP density in each window as either 'low' or 'high' using a two-state hidden Markov model (HMM). This HMM was created and fitted on the observed 10 kb SNP densities by the expectation-maximization algorithm using functions 'depmix' and 'fit' (R package depmixS4), and subsequently the posterior state sequence (with states 'low' and 'high'), computed via the Viterbi algorithm, was extracted with function 'posterior' (R package depmixS4).

Gene annotation. The prediction of *Ct* and *Ci* gene models was performed using the MAKER pipeline[46] (version 2.28) , which integrates different *ab initio* gene prediction tools together with evidence from EST and protein alignments. In a first step, for each genome, the pipeline was run using Augustus[47] (with species model *Fusarium graminearum*) and GeneMark-ES[48] for *ab initio* gene prediction together with transcript and protein alignment evidence. The resulting gene models from this first run were used as training set for a third *ab initio* prediction tool, SNAP[49], and subsequently the annotation pipeline was re-run, this time including all three *ab initio* prediction tools together with the transcript and protein alignment evidence to yield the final gene models. The alignment evidence was created from BLAST and Exonerate[50] alignments of both protein and transcript sequences of each respective fungus (*Ct*/*Ci*) and protein sequences of *C. higginsianum* and *C. graminicola*. *Ct* (isolate 0861) and *Ci* transcript and protein sequences were obtained from the corresponding RNA-seq data via a transcriptome *de novo* assembly. For this purpose, we extracted all RNA-seq read pairs that did not align to the host plant genome from four (*Ct*) to nine (*Ci*) *in planta* samples and combined these with the read pairs from one *in vitro* sample of the respective fungus. The combined RNA-seq reads were then used as input for Trinity[51] (with default parameter settings for paired-end reads) to assemble transcripts and extract peptide sequences of the best-scoring ORFs (using the Perl script 'transcripts_to_best_scoring_ORFs.pl' provided with the Trinity software). General functional annotations for the predicted gene models were obtained using Blast2GO (ref. 52). To perform Blast2GO searches and ensure stable databases over time for multiple genome annotations, the NCBI nr database was downloaded locally (version: 8 January 2015). In addition, a local b2gdb mysql database was generated (version 201402) and connected to the Blast2GO java tool. For each genome annotation, BLASTP was performed against the local NCBI nr database (− e 1E − 3, -v 10 − b 10) and tabular BLAST output was loaded into Blast2GO using graphical java interface. Further analyses were performed according to the Blast2GO user manual.

MCL analysis. Gene families and clusters of orthologous genes were inferred using OrthoMCL[53] (version 2.0) with standard parameters and granularity 1.5 for the MCL clustering step. Functional enrichment and overrepresentation analyses were performed using a Fisher's exact test, adjusting for FDR. For each gene family inferred with orthoMCL, a multiple sequence alignment of the protein sequences was obtained using Clustal Omega[54] and an HMM model was generated with the hhmake program of the HHSuite toolkit[55]. Sequences from the fungal database fuNOG[56] were similarly aligned and HMM models generated. To annotate whole gene families, the hhsearch program was used to obtain matches between the gene family and the fuNOG HMMs and only hits with a probability equal to or higher than 0.99 were considered. To annotate whole gene families, the hhsearch program was used to obtain matches between the gene family and the fuNOG HMMs and only hits with a probability ≥ 0.99 were considered.

Ancestral genome reconstruction. Gene families inferred with OrthoMCL were used to reconstruct the ancestral genomes of each *Colletotrichum* lineage. GLOOME[57] (maximum-likelihood approach) was used to infer ancestral gene gains and losses (GGLs) and to reconstruct the ancestral GGLs of gene families on the species tree of *Ct*0861 and the other five genomes available for this genus. Evolution of the GGLs along the branches of a phylogenetic tree was modelled as a continuous time Markov process using a binary character alphabet corresponding to gene family presence or absence. Default parameters were used, corresponding to a mixture model that allows varying GGL rates across gene families. We approximated the total number of gene families that were gained or lost on a branch by summing up the individual posterior probabilities for each gene family to be gained or lost on that branch and rounding this number to the closest integer. The number of genes either gained or lost (annotated with one specific category) was compared with the respective numbers detected for all other branches of the tree. The significance was assessed using Fisher's exact test and FDR corrected.

d_N/d_S analysis. A multiple-sequence alignment (MSA) of orthologous groups of coding sequences (CDSs) was created with Clearcut[58]. Based on the MSA and the CDSs, a codon alignment was constructed for each protein family with pal2nal (ref. 59; version 14) using default parameters. Because of the data set size and the shorter runtime of neighbour joining algorithms compared with maximum-likelihood methods, Clearcut, a relaxed neighbour joining algorithm[58], was chosen for reconstructing phylogenetic trees from the MSA of each protein family with slightly modified additive pairwise distances whereby gaps are not counted as mismatches. Gaps in this alignment were mostly of technical origin due to the alignment of short contigs to longer reference sequences. Using an in-house tool (phylorecon), CDSs and amino acid sequences were reconstructed for the internal nodes of each phylogenetic tree using maximum parsimony as a criterion[60], and the synonymous and non-synonymous substitution rates per site were inferred with correction for multiple substitutions. The average d_N/d_S ratio was calculated for each protein family and a one-sided Fisher's test (FDR corrected) was performed to identify protein families with a significant enrichment of synonymous mutations per synonymous site versus non-synonymous mutations per non-synonymous site.

Annotation of specific gene categories. Secretomes of all species were predicted using WoLF-PSORT[61] with default settings. *Colletotrichum* CSEPs were defined as extracellular proteins with no significant BLAST homology (*E*-value $<1 \times 10^{-3}$) to sequences outside the genus *Colletotrichum* in the UniProt database (SwissProt and TrEMBL components). To identify secreted proteases, sequences of predicted extracellular proteins were subjected to a MEROPS Batch BLAST analysis[62]. Membrane transporters were identified and classified through BLAST searches against the Transporter Collection Database (http://www.tcdb.org/). To predict the repertoire of carbohydrate-active enzymes encoded by *Colletotrichum* species, we scanned their genomes using the CAZy annotation pipeline[63] (http://www.cazy.org). For annotating genes encoding secondary metabolism key enzymes in *Colletotrichum* species, we used an in-house bioinformatics pipeline that was developed as described in Supplementary Note 12.

RNA sequencing. The RNA-seq samples presented in Hiruma *et al.*[8] and the new samples presented here were prepared as follows. Fungal cultures were maintained on Mathur's agar medium at 25°C, and conidia were harvested from 7- to 10-day-old cultures. For sample preparation, *Arabidopsis* Col-0 seeds were surface sterilized in 70% ethanol and subsequently in 2% hypochlorous acid (v/v) containing 0.05% (v/v) Triton. We inoculated *A. thaliana* Col-0 seeds with spores (5×10^4 spores ml^{-1}) of *Ct* 0861 or *Ci* and transferred the inoculated seeds onto solid half-strength Murashige and Skoog medium (pH = 5.1) either in normal [625 μM] or low phosphate [50 μM] conditions. For each biological replicate ($n = 3$), the entire root system of at least 10 plants was collected at time intervals (6, 10, 16 or 24 d.p.i.) and pooled before RNA extraction. In addition, we grew *Ct* and *Ci* in liquid Mathur's medium (*in vitro* samples) for 2 days at 24 °C with shaking at 50 r.p.m. and collected the hyphae by filtration. Total RNA was purified with the NucleoSpin RNA plant kit (Macherey-Nagel) according to the manufacturer's protocol. RNA-seq libraries were prepared from an input of 1 μg total RNA using the Illumina TruSeq stranded RNA sample preparation kit. Libraries were subjected to paired-end sequencing (100 bp reads) using the Illumina HiSeq2500 Sequencing System. To make sure the sequenced reads were of sufficiently high quality, an initial quality check was performed using the FastQC suite (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Subsequently, the RNA-seq reads were mapped to the assembled and annotated genomes of either *Ct* 0861 or *Ci*, and in parallel to the annotated genome of the host plant *A. thaliana* (TAIR10) using Tophat2 (ref. 64; $a = 10$, $g = 10$, $r = 100$, mate-std-dev = 40). The mapped RNA-seq reads were then transformed into a fragment count per gene per sample using the htseq-count script (s = reverse, t = exon) in the package HTSeq[65]. The complete RNA-Seq data presented by Hiruma *et al.*[8] and in this manuscript have been deposited under the GEO series accession number GSE70094.

Statistical analysis of differential gene expression. All statistical analyses of plant and fungal gene expression were performed in R (codes are available upon request). For the analyses of plant gene expression, genes with less than 100 mapped fragments in total (that is, across all the analysed samples) were rated as 'not expressed' and therefore excluded. For analyses of fungal gene expression, we excluded genes that were not sufficiently expressed in the *in planta* samples, that is, genes with less than 100 (*Ct*, 24 samples) or less than 50 (*Ci*, 6 samples) mapped fragments across all the analysed samples. Subsequently, the count data for all expressed genes was TMM-normalized and log-transformed using the functions 'calcNormFactors' (R package EdgeR[66]) and 'voom' (R package limma[67]) to yield $\log_2$ counts per million ($\log_2$cpm). To analyse the aspects of differential gene expression in *Ct*0861, *Ci* and their host plant *Arabidopsis*, we fitted for each analysis a distinct linear model to the respective $\log_2$-transformed count data using the function lmFit (R package limma[67]) and subsequently performed moderated *t*-tests for specific comparisons of interest. Resulting *P* values were adjusted for false discoveries due to multiple hypotheses testing via the Benjamini–Hochberg procedure (FDR). To extract genes with significant expression differences, a cutoff of FDR <0.05 and $|\log_2\text{FC}| \geq 1$ was applied. Heatmaps of gene expression profiles were generated with the Genesis expression analysis package[68] and interactive Tree Of Life[69] was used to visualize CSEP gene expression data. To derive *Arabidopsis*, *Ct* and *Ci* gene expression profiles during the time-course experiment, $\log_2$ expression ratios were calculated between the normalized number of reads detected for a given gene at a given developmental stage and the geometrical mean of the number of reads calculated across all developmental stages. This $\log_2$ ratio is referred as the 'Relative Expression Index'. The Cytoscape plug-in ClueGO + CluePedia[70] was used to construct GO term enrichment networks and to visualize functionally grouped terms among significantly regulated genes. Significant enrichments were determined using the hypergeometric test and Bonferroni step-down corrected *P* values are represented. Co-regulated genes that were also co-expressed in other *Arabidopsis* expression data sets were identified using ATTED-II (http://atted.jp/) and co-expression networks were generated using Cytoscape[71] (version 3.1.1).

RT-qPCR analysis. First-strand cDNA was synthesized from 1 μg DNase-treated total RNA using the iScript cDNA synthesis kit (Bio-Rad) and PCR amplification was performed using the iQ5 real-time PCR detection system (Bio-Rad). For each gene, specific primers were designed with the Primer 3 and AmplifX programs. BLASTN searches against the *Ct* and *A. thaliana* genomes were performed to rule out cross-annealing artefacts. Gene expression levels were normalized using the reference gene actin (*ACT2*, AT3G18780) for *A. thaliana* and the reference gene tubulin beta-1 chain (CT04_12898) for *Ct*. These genes were used to normalize gene expression levels using the Pfaffl calculation method[72].

Microscopy methods. For cytology experiments, surface-sterilized *A. thaliana* Col-0 seeds were inoculated with either *Ct* or *Ci* conidia (5×10^4 spores ml^{-1}). The seeds were then transferred to half-strength Murashige and Skoog agarose medium without sucrose and low-phosphate content (50 μM). Inoculated plants were grown at 22 °C with a 10-h photoperiod (80 μE m^{-2} s^{-1}) for 1 to 24 days. The roots were either mounted in water for viewing GFP or first stained with Calcofluor white (0.01 %, Sigma) or fluorescein diacetate (10 μg ml^{-1}, Sigma). For visualizing GFP and FDA fluorescence, we used an Olympus FV1000 confocal microscope equipped with dry ×20 and ×40 objectives, using the 488-nm line of an Argon laser for excitation and fluorescence was collected at 490–520 nm. For imaging Calcofluor fluorescence, we used a Zeiss Axiophot epifluorescence microscope (filter set BP 365, FT 395, LP 397).

References

1. Rodriguez, R. J., White, Jr J. F., Arnold, A. E. & Redman, R. S. Fungal endophytes: diversity and functional roles. *New Phytol.* **182,** 314–330 (2009).
2. O'Connell, R. J. *et al.* Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. *Nat. Genet.* **44,** 1060–1065 (2012).
3. Hyde, K. D. *et al. Colletotrichum*—names in current use. *Fungal Divers.* **39,** 147–182 (2009).
4. Sukno, S. A., Garcia, V. M., Shaw, B. D. & Thon, M. R. Root infection and systemic colonization of maize by *Colletotrichum graminicola*. *Appl. Environ. Microbiol.* **4,** 823–832 (2008).
5. Götz, M. *et al.* Fungal endophytes in potato roots studied by traditional isolation and cultivation-independent DNA-based methods. *FEMS Microbiol. Ecol.* **58,** 404–413 (2006).
6. Keim, J., Mishra, B., Sharma, R., Ploch, S. & Thines, M. Root-associated fungi of *Arabidopsis thaliana* and *Microthlaspi perfoliatum*. *Fungal Divers.* **66,** 99–111 (2014).
7. Gan, P. *et al.* Comparative genomic and transcriptomic analyses reveal the hemibiotrophic stage shift of *Colletotrichum* fungi. *New Phytol.* **197,** 1236–1249 (2012).
8. Hiruma, K. *et al.* Root endophyte *Colletotrichum tofieldiae* confers plant fitness benefits that are phosphate status-dependent. *Cell* **165,** 1–11 (2016).
9. Delaux, P. M. *et al.* Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet.* **10,** e1004487 (2014).
10. García, E., Alonso, Á., Platas, G. & Sacristán, S. The endophytic mycobiota of *Arabidopsis thaliana*. *Fungal Divers.* **60,** 71–89 (2013).
11. Sato, T. *et al.* Anthracnose of Japanese radish caused by *Colletotrichum dematium*. *J. Gen. Plant Pathol.* **71,** 380–383 (2005).
12. Stukenbrock, E. H., Christiansen, F. B., Hansen, T. T., Dutheil, J. Y. & Schierup, M. H. Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species. *Proc. Natl Acad. Sci. USA* **109,** 10954–10959 (2012).
13. Hacquard, S. *et al.* Mosaic genome structure of the barley powdery mildew pathogen and conservation of transcriptional programs in divergent hosts. *Proc. Natl Acad. Sci. USA* **110,** E2219–E2228 (2013).
14. Zuccaro, A., Lahrmann, U. & Langen, G. Broad compatibility in fungal root symbioses. *Curr. Opin. Plant Biol.* **20,** 135–145 (2014).
15. Kohler, A. *et al.* Convergent losses of decay mechanisms and rapid turnover of symbiosis genes in mycorrhizal mutualists. *Nat. Genet.* **47,** 410–415 (2015).
16. Martin, F. *et al.* Périgord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. *Nature* **464,** 1033–1038 (2010).
17. Lahrmann, U. *et al.* Mutualistic root endophytism is not associated with the reduction of saprotrophic traits and requires a noncompromised plant innate immunity. *New Phytol.* **207,** 841–857 (2015).
18. Lo Presti, L. *et al.* Fungal effectors and plant susceptibility. *Annu. Rev. Plant Biol.* **66,** 513–545 (2015).
19. Zuccaro, A. *et al.* Endophytic life strategies decoded by genome and transcriptome analyses of the mutualistic root symbiont *Piriformospora indica*. *PLoS Pathog.* **7,** e1002290 (2011).
20. Xu, X. H. *et al.* The rice endophyte *Harpophora oryzae* genome reveals evolution from a pathogen to a mutualistic endophyte. *Sci. Rep.* **4,** 5783 (2014).
21. Bonfante, P. & Genre, A. Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. *Nat. Commun.* **1,** 48 (2010).
22. Won, S. K. *et al.* Cis-element- and transcriptome-based screening of root hair-specific genes and their functional characterization in Arabidopsis. *Plant Physiol.* **150,** 1459–1473 (2009).
23. Pandey, S. P. & Somssich, I. E. The role of WRKY transcription factors in plant immunity. *Plant Physiol.* **150,** 1648–1655 (2009).
24. Nakano, T., Suzuki, K., Fujimura, T. & Shinshi, H. Genome-wide analysis of the ERF gene family in Arabidopsis and rice. *Plant Physiol.* **140,** 411–432 (2006).
25. Gigolashvili, T. *et al.* The transcription factor HIG1/MYB51 regulates indolic glucosinolate biosynthesis in *Arabidopsis thaliana*. *Plant J.* **50,** 886–901 (2007).
26. Bednarek, P. *et al.* A glucosinolate metabolism pathway in living plant cells mediates broad-spectrum antifungal defense. *Science* **323,** 101–106 (2009).
27. Clay, N. K., Adio, A. M., Denoux, C., Jander, G. & Ausubel, F. M. Glucosinolate metabolites required for an Arabidopsis innate immune response. *Science* **323,** 95–101 (2009).
28. Pant, B. D. *et al.* Identification of primary and secondary metabolites with phosphorus status-dependent abundance in *Arabidopsis*, and of the transcription factor PHR1 as a major regulator of metabolic changes during phosphorus limitation. *Plant Cell Environ.* **38,** 172–187 (2015).
29. Freeman, S. & Rodriguez, R. J. Genetic conversion of a fungal plant pathogen to a nonpathogenic, endophytic mutualist. *Science* **260,** 75–78 (1993).
30. De Coninck, B., Timmermans, P., Vos, C., Cammue, B. P. & Kazan, K. What lies beneath: belowground defense strategies in plants. *Trends Plant Sci.* **20,** 91–101 (2015).
31. Rasmussen, S. *et al.* Transcriptome responses to combinations of stresses in Arabidopsis. *Plant Physiol.* **161,** 1783–1794 (2013).
32. Bruce, A., Smith, S. E. & Tester, M. The development of mycorrhizal infection in cucumber: effects of P supply on root growth, formation of entry points and growth of infection units. *New Phytol.* **127,** 507–514 (1994).
33. Chevreux, B. *et al.* Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res.* **14,** 1147–1159 (2004).
34. Bankevich, A. *et al.* SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
35. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
36. Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol.* **5,** R12 (2004).
37. Hunt, M. *et al.* REAPR: a universal tool for genome assembly evaluation. *Genome Biol.* **14,** R47 (2013).
38. Parra, G., Bradnam, K. & Korf, I. CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* **23,** 1061–1067 (2007).
39. Edgar, R. C. & Myers, E. W. PILER: identification and classification of genomic repeats. *Bioinformatics* **21,** i152–i158 (2005).
40. Smit, A. F. A., Hubley, R. & Green, P. RepeatMasker Open-4.0 http://www.repeatmasker.org http://www.repeatmasker.org (2010).
41. Jurka, J. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).
42. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19,** 301–302 (2003).
43. Langmead, B. & Salzberg, S. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9,** 357–359 (2012).
44. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
45. Krzywinski, M. *et al.* Circos: an information aesthetic for comparative genomics. *Genome Res.* **19,** 1639–1645 (2009).
46. Holt, C. & Yandell, M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12,** 491 (2011).
47. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34,** W435–W439 (2006).
48. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. & Borodovsky, M. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Res.* **18,** 1979–1990 (2008).
49. Korf, I. Gene finding in novel genomes. *BMC Bioinformatics* **5,** 59 (2004).
50. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6,** 31 (2005).
51. Grabherr, M. G. *et al.* Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29,** 644–652 (2011).
52. Conesa, A. *et al.* Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **21,** 3674–3676 (2005).
53. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13,** 2178–2189 (2003).
54. Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7,** 539 (2011).
55. Söding, J. Protein homology detection by HMM-HMM comparison. *Bioinformatics* **21,** 951–960 (2005).
56. Powell, S. *et al.* eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42,** D231–D239 (2014).
57. Ofir, C. & Pupko, T. Inference of gain and loss events from phyletic patterns using stochastic mapping and maximum parsimony—a simulation study. *Genome Biol. Evol.* **3,** 1265–1275 (2011).
58. Sheneman, L., Evans, J. & Foster, J. A. Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics* **22,** 2823–2824 (2006).
59. Suyama, M., Torrents, D. & Bork, P. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res.* **34,** W609–W612 (2006).

60. Tusche, C., Steinbrück, L. & McHardy, A. C. Detecting patches of protein sites of influenza A viruses under positive selection. *Mol. Biol. Evol.* **29,** 2063–2071 (2012).
61. Horton, P. *et al.* WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* **35,** W585–W587 (2007).
62. Rawlings, N. D., Barrett, A. J. & Bateman, A. MEROPS: the peptidase database. *Nucleic Acids Res.* **38,** D227–D233 (2010).
63. Lombard, V., Ramulu, H. G., Drula, E., Coutinho, P. M. & Henrissat, B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* **42,** D490–D495 (2014).
64. Kim, D. *et al.* TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14,** R36 (2013).
65. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31,** 166–169 (2014).
66. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26,** 139–140 (2010).
67. Smyth, G. K., Michaud, J. & Scott, H. S. Use of within-array replicate spots for assessing differential expression in microarray experiments. *Bioinformatics* **21,** 2067–2075 (2005).
68. Sturn, A., Quackenbush, J. & Trajanoski, Z. Genesis: cluster analysis of microarray data. *Bioinformatics* **18,** 207–208 (2002).
69. Letunic, I. & Bork, P. Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23,** 127–128 (2007).
70. Bindea, G. *et al.* ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. *Bioinformatics* **25,** 1091–1093 (2009).
71. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).
72. Pfaffl, M. W. A new mathematical model for relative quantification in real-time RT-PCR. *Nucleic Acids Res.* **29,** e45 (2001).

Acknowledgements

This work was primarily supported by the Max Planck Society (S.H., B.K., K.H., R.G.-O, E.V.L.v.T., E.K, P.S.-L and R.J.O.), the Agence Nationale de la Recherche grant ANR-12-CHEX-0008-01 (R.J.O.), the European Research Council advanced grant (ROOTMICROBIOTA) (P.S.-L), the 'Cluster of Excellence on Plant Sciences' program (funded by the Deutsche Forschungsgemeinschaft; R.G.-O, E.K., A.C.M, P.S.-L). Other funding sources included the German Center for Infection Research, DZIF (P.C.M.) and the Japan Society for the Promotion of Science (Post-doctoral Fellowship for Research Abroad; K.H.). We are grateful to Lisa Vaillancourt (University of Kentucky) and Pedro Crous (CBS-KNAW, Utrecht) for providing fungal cultures. We thank the staff of the Max Planck Genome Centre Cologne for their expertise in generating the sequencing data. We thank Anthony Levasseur (Aix-Marseille University) for assistance with annotation of CAZyme genes and Antonios Zampounis for bioinformatic assistance with secondary metabolism gene analysis.

Author contributions

P.S.-L., R.J.O. and K.H. initiated the project. S.H., R.J.O. and P.S.-L. coordinated the project. B.K., E.K. and A.C.M. coordinated the bioinformatics. K.H. performed the inoculation experiments and isolated the DNA/RNA. B.K, E.V.L.v.T. and E.K. performed the sequence assembly and annotation. B.K. and E.K. performed the gene predictions and managed the data. U.D., R.G.-O. and M.R.T. performed the phylogeny and evolution analyses. B.K. and R.G.-O. performed the comparative genomics and MCL analyses. M.R.T analysed the repeats. B.K. and S.H. analysed RNA-Seq data. S.H., B.K., R.J.O., J.-F.D., M.H., B.H. and O.L. annotated gene families. A.W. and A.C.M performed ancestral genome reconstruction. P.C.M. and A.C.M performed d_N/d_S analysis. B.K. performed SNP analysis. S.S. contributed *C. tofieldiae* isolate 0861 and performed soil analyses. S.H., B.K., A.W., P.M., U.D., O.L., J.-F.D., B.H., M.R.T., R.G.-O., A.C.M. and E.K. prepared the tables, figures and text. S.H, B.K, R.J.O and P.S.-L wrote and edited the paper.

Additional information

Accession codes: The genome assemblies have been deposited at DDBJ/EMBL/GenBank with accession numbers LFIW01000000 (*Ci*), LFIV01000000 (*Ct*0861), LFHR01000000 (CBS127), LFHS01000000 (CBS130), LFHP01000000 (CBS495), LFHRQ01000000 (CBS168). The RNA-Seq data have been deposited in the NCBI Gene Expression Omnibus under GEO Series accession number GSE70094.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

How to cite this article: Hacquard, S. *et al.* Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nat. Commun.* 7:11362 doi: 10.1038/ncomms11362 (2016).

OPEN

The ISME Journal (2016), 1–13
© 2016 International Society for Microbial Ecology All rights reserved 1751-7362/16
www.nature.com/ismej

ORIGINAL ARTICLE

Root microbiota dynamics of perennial *Arabis alpina* are dependent on soil residence time but independent of flowering time

Nina Dombrowski[1], Klaus Schlaeppi[2], Matthew T Agler[1], Stéphane Hacquard[1], Eric Kemen[1,3], Ruben Garrido-Oter[1,3,4], Jörg Wunder[5], George Coupland[5] and Paul Schulze-Lefert[1,3]

[1]*Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, Cologne, Germany;* [2]*Plant–Soil-Interactions, Institute for Sustainability Sciences, Agroscope, Reckenholzstrasse 191, Zurich, Switzerland;* [3]*Cluster of Excellence on Plant Sciences (CEPLAS), Max Planck Institute for Plant Breeding Research, Cologne, Germany;* [4]*Department of Algorithmic Bioinformatics, Heinrich Heine University Düsseldorf, Düsseldorf, Germany and* [5]*Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Cologne, Germany*

Recent field and laboratory experiments with perennial *Boechera stricta* and annual *Arabidopsis thaliana* suggest that the root microbiota influences flowering time. Here we examined in long-term time-course experiments the bacterial root microbiota of the arctic-alpine perennial *Arabis alpina* in natural and controlled environments by 16S rRNA gene profiling. We identified soil type and residence time of plants in soil as major determinants explaining up to 15% of root microbiota variation, whereas environmental conditions and host genotype explain maximally 11% of variation. When grown in the same soil, the root microbiota composition of perennial *A. alpina* is largely similar to those of its annual relatives *A. thaliana* and *Cardamine hirsuta*. Non-flowering wild-type *A. alpina* and flowering *pep1* mutant plants assemble an essentially indistinguishable root microbiota, thereby uncoupling flowering time from plant residence time-dependent microbiota changes. This reveals the robustness of the root microbiota against the onset and perpetual flowering of *A. alpina*. Together with previous studies, this implies a model in which parts of the root microbiota modulate flowering time, whereas, after microbiota acquisition during vegetative growth, the established root-associated bacterial assemblage is structurally robust to perturbations caused by flowering and drastic changes in plant stature.

The ISME Journal advance online publication, 2 August 2016; doi:10.1038/ismej.2016.109

Introduction

Plants host taxonomically structured bacterial consortia on and inside roots and leaves, designated the root and leaf microbiota (Vorholt, 2012; Bulgarelli *et al.*, 2013; Hacquard *et al.*, 2015). The start inoculum of the leaf-associated microbiota is derived from multiple sources, likely involving bacteria transmitted by aerosols, insects or soil particles (Vorholt, 2012; Bodenhausen *et al.*, 2013; Maignien *et al.*, 2014; Bai *et al.*, 2015). Conversely, root-associated bacterial consortia are mostly derived from the bacterial soil biome surrounding roots and establish rapidly. For example, a stable taxonomic structure in rice roots developed within 14 days after seed germination (Edwards *et al.*, 2015). Soil represents the most diverse ecosystem on earth with an exceptionally high bacterial species diversity that varies greatly between different soil types (Fierer and Jackson, 2006; Lauber *et al.*, 2009). Numerous studies employing next generation sequencing technologies have shown that soil type is a major determinant of root microbiota composition, most likely reflecting the different bacterial start inocula present in each soil type (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Peiffer *et al.*, 2013; Schlaeppi *et al.*, 2014; Edwards *et al.*, 2015). Despite significant variation in microbiota composition at low taxonomic ranks, for example, genus- or species-level, a recent direct comparison of the root microbiota of eight flowering plant species, including monocots and dicots, revealed a co-occurrence of three main bacterial phyla comprising Actinobacteria, Bacteroidetes and Proteobacteria (Hacquard *et al.*, 2015). This finding suggests that the root microbiota and its overall

Correspondence: Dr P Schulze-Lefert, Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, Cologne 50829, Germany.
E-mail: schlef@mpipz.mpg.de

Received 4 February 2016; revised 27 June 2016; accepted 3 July 2016

taxonomic structure is a conserved plant trait across flowering plants.

An unresolved question in microbiota research is whether plant-associated bacterial assemblages contribute to the plasticity of complex plant traits. For example, field experiments with perennial *Boechera stricta* and laboratory experiments with annual *A. thaliana* have indicated that the bacterial root microbiota modulates flowering time (Wagner *et al.*, 2014; Panke-Buisse *et al.*, 2015). Conversely, root-associated bacterial consortia of *A. thaliana* appear to be affected by plant development stages and metatranscriptome studies revealed bacterial transcripts induced at bolting and flowering stages (Chaparro *et al.*, 2014).

Plant growth in the arctic-alpine environment requires an adaptation to a range of abiotic stresses, including water limitation, extreme temperature shifts and low nutrient availability (Billings and Mooney, 1968; Chapin, 1983; Chapin and Shaver, 1989; Margesin and Miteva, 2011). Root-associated bacterial members from three arctic-alpine plant species (*Oxyria digyna*, *Diapensia lapponica* and *Juncus trifidus*) appear to be enriched for potent microbial solubilizers of mineral phosphorus, a common but plant-inaccessible source of phosphorus in arctic-alpine environments (Nissinen *et al.*, 2012). This suggests the potential importance of microbiota members for plant growth and health in arctic-alpine environments (Richardson *et al.*, 2009). Perennial *Arabis alpina* is an arctic-alpine plant closely related to the annual *Arabidopsis thaliana*, which allows comparative studies of inter-species trait diversification such as flowering time (Beilstein *et al.*, 2010; Wang *et al.*, 2009). For example, the orthologue of the *A. thaliana* gene *FLOWERING LOCUS C* (*FLC*) that inhibits flowering until *A. thaliana* is exposed to winter temperature, was shown to be *A. alpina PEP1* (*PERPETUAL FLOWERING 1*), which limits flowering duration (Wang *et al.*, 2009). Perpetual flowering is characteristic for *A. alpina pep1* mutant plants, resulting in a drastic difference in plant stature due to the presence of reproductive shoots compared with wild-type (WT) plants (Wang *et al.*, 2009). Extensive allelic variation at *PEP1* also exists in natural populations, including loss of *PEP1* function alleles (Albani *et al.*, 2012). The adaptation of *A. alpina* to arctic-alpine environments, its perennial nature, the availability of genetic lines, as well as the close evolutionary relationship to *A. thaliana* make this plant species a suitable model to investigate the interplay of determinants for root microbiota composition and diversification.

Here we characterized *A. alpina* root-associated bacterial communities using several experimental approaches. We employed culture-independent community profiling by 16S rRNA gene sequencing to assess bacterial community composition of *A. alpina* plants grown in their natural habitat compared with plants grown under controlled environmental conditions in native or non-native soils ('soil type and environment' experiment). We then investigated potential changes of the *A. alpina* root microbiota at two developmental stages (flowering vs vegetative) during an extended residence time of WT and *pep1* mutant plants in soil ('time-course' experiment). Finally, we examined the diversification of the *A. alpina* bacterial microbiota by comparison with root microbial communities collected from two other Brassicaceae species, *A. thaliana* and *Cardamine hirsuta* ('diversification' experiment).

Materials and methods

Soil types used and plant material

French soil was harvested in fall 2012 ('FS Fall-12') at the Col du Galibier, France (45.061 N/6.402 E). Similar to earlier work (Bulgarelli *et al.*, 2012), Cologne soil batches were collected in fall 2010, in spring 2013 and in fall 2013 (termed 'CS Fall-10', 'CS Spring-13' and 'CS Fall-13') at the Max Planck Institute for Plant Breeding Research in Cologne, Germany (50.958 N/6.856 E). Geochemical characterization of soil types was carried out by the 'Labor für Boden- und Umweltanalytik' (Eric Schweizer AG, Thun, Switzerland).

Three Brassicaceae plant species were investigated during this study. Experiments with *A. alpina* were conducted using the Spanish reference ecotype Pajares (Paj), the two French ecotypes F1-Gal5 and Gal60, as well as the mutant line *pep1* (Paj background; Wang *et al.*, 2009). *A. alpina* Gal60 was harvested during the course of this study at the Col du Galibier, France. In addition, *A. thaliana* (ecotype Columbia, Col-0) and *C. hirsuta* (ecotype Oxford, Ox) were investigated.

Plant growth

We characterized *A. alpina* root-associated bacterial communities using three sets of experiments (Table 1, Supplementary Figure S1). First, for the 'soil type and environment' experiment, individual flowering *A. alpina* plants of unknown age were excavated (designated Gal60) from their native habitat at the Col du Galibier (France) in fall 2012 to evaluate their natural root microbial communities (Figure 1a). To evaluate the effect of environmental factors on community assembly, we also collected soil from this site and transported it to Cologne to compare microbial communities from *A. alpina* Gal60 grown its native French soil in the natural environment with controlled environmental conditions in the greenhouse. To investigate possible host genotype-dependent effects on microbiota composition, we grew Gal60 alongside with a French genotype from a nearby location (Gal5) and a Spanish reference genotype (Paj) in the French soil in the greenhouse for 3 months. During harvest, Gal60 and Paj resided in the vegetative growth stage, while Gal5 plants started to flower (Figure 1b). Apart

Table 1 Experimental design and numbers of replicates per DNA sample across experimental setups

Compartment	*Host plant*	*Soil type and environment*			*Time course*						*Diversification*			*454 vs Illumina*
		CS_C	*FS_C*	*FS_N*	*6 weeks*		*12 weeks*		*28 weeks*					
					rep1	*rep2*	*rep1*	*rep2*	*rep1*	*rep2*	*rep1*	*rep2*^{a}	*rep3*	*rep2*^{a}
Soil	—	4	4	3	3	3	3	3	3	3	3	3	3	3
Rhizosphere	*A. alpina* (Paj)	4	5	—	3	3	4	4	4	4	—	—	—	—
	A. alpina (*pep1*)	—	—	—	3	3	4	4	4	4	—	—	—	—
	A. alpina (Gal5)	—	5	—	—	—	—	—	—	—	—	—	—	—
	A. alpina (Gal60)	—	5	5	—	—	—	—	—	—	—	—	—	—
Root	*A. thaliana* (Col-0)	—	—	—	—	—	—	—	—	—	3	3	3	3
	C. hirsuta (Ox)	—	—	—	—	—	—	—	—	—	3	3	3	2
	A. alpina (Paj)	4	5	—	3	3	4	4	4	4	3	3	3	3
	A. alpina (*pep1*)	—	—	—	3	3	4	4	4	4	—	—	—	—
	A. alpina (Gal5)	—	5	—	—	—	—	—	—	—	—	—	—	—
	A. alpina (Gal60)	—	5	5	—	—	—	—	—	—	—	—	—	—
	Soil type	Cologne	France		Cologne						Cologne			Cologne
	Residence time	12 weeks	12 weeks	ND	6 weeks	6 weeks	12 weeks	12 weeks	28 weeks	28 weeks	6 weeks	6 weeks	6 weeks	6 weeks
	Harvest date soil	Spring13	Fall12		Spring13	Fall13	Spring13	Fall13	Spring13	Fall13	Fall10	Spring13	Fall13	Spring13
	Environment	Controlled		Native	Controlled						Controlled			Controlled
	Seq. technology	Illumina			Illumina						454			Illumina

Abbreviations: CS, Cologne soil; FS, French soil; ND, not determined.
Experimental setup comparing bacterial community composition under natural growth conditions ('soil type and environment' experiments), prolonged residence time of plants in soil ('time-course' experiment) on three different plant species ('diversification' experiment). The first two columns characterize the tested compartment, host plant and plant genotype. The numbers indicate the number of sequenced DNA samples. 'Soil type and environment' experiment: plants were grown in CS or FS under controlled environmental conditions (C) or in their native habitat in France (N). 'Time course' experiment: *A. alpina* plants were grown under controlled environmental conditions in CS for 6, 12 and 28 weeks. 'Diversification' experiment: *A. alpina*, *C. hirsuta* and *A. thaliana* were grown under controlled environmental conditions in CS for 6 weeks. Indicated are the number of independent biological replicates (rep1–rep3), the soil type used for growing plants (see also Supplementary Tables S3 and S4), the harvest date of the soil, the environmental conditions used for plant growth and the employed sequencing technology. ND, not determined.
[a]Represents the same DNA samples that were processed with both 454 and Illumina sequencing technologies.

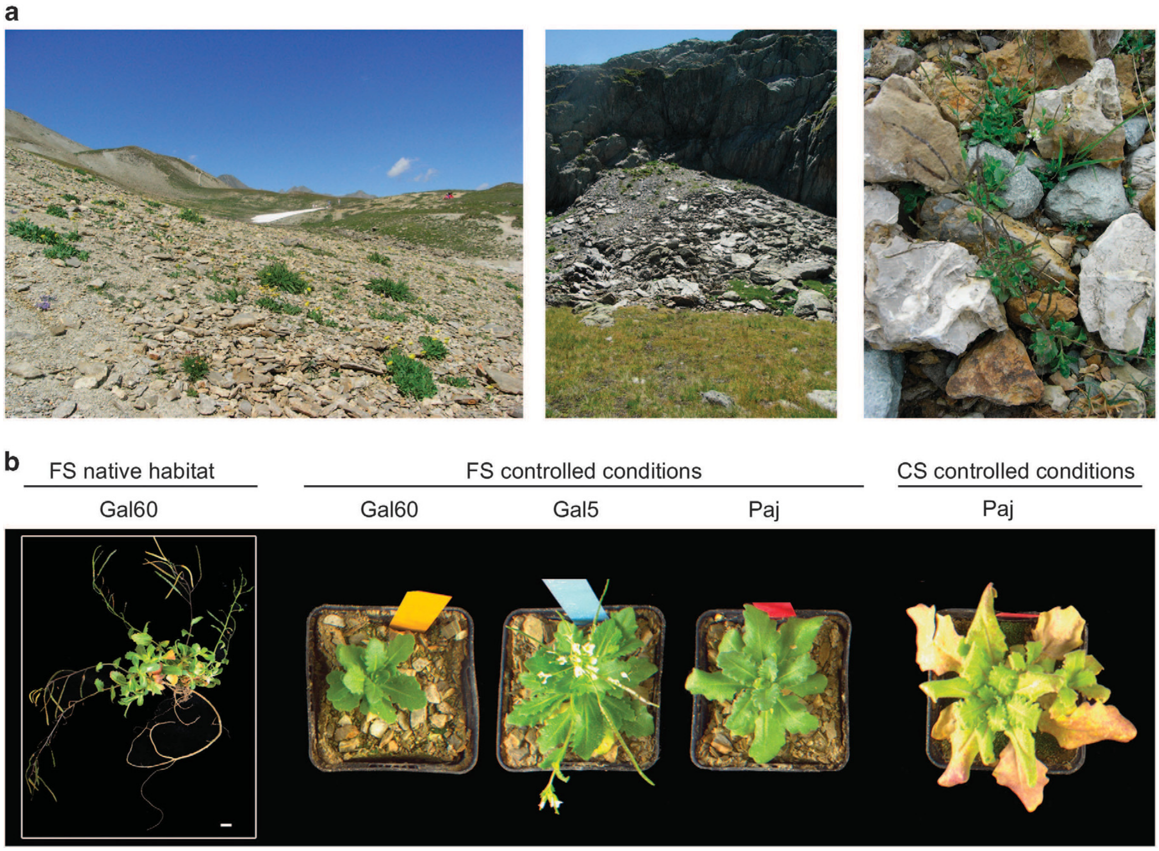


Figure 1 Growth phenotype of *A. alpina* in a natural French habitat and under controlled environmental conditions. (**a**) Natural habitat of *A. alpina* in the French Alps. (**b**) Growth morphology of different *A. alpina* accessions in a French soil (FS) in its native habitat in the French Alps, in FS under controlled environmental conditions and Cologne soil (CS) under controlled environmental conditions. Gal60, French *A. alpina* accession collected in its natural habitat in the French Alps, Col du Galibier; Gal5, French *A. alpina* accession; Paj, Spanish *A. alpina* accession.

from their geographic origin, the three genotypes differ from each other in that the genotype Paj requires a vernalization treatment to flower, while Gal5 and Gal60 do not. In an additional subset of this experimental setup, we planted the *A. alpina* genotype Paj not only in French but also Cologne soil to study the effect of soil type on community composition. For the 'soil type and environment' experiment conducted in the greenhouse, *A. alpina* seeds were surface-sterilized, sown onto the respective soil type, stratified for 4 days and a single plant per pot (at least four to five individuals per genotype) was grown for 3 months until harvest. In a second experimental setup, denoted 'time course' experiment, individual *A. alpina* Paj plants were grown in two batches of Cologne soil, that is, two full factorial biological replicates, under controlled environmental conditions and harvested after 6, 12 and 28 weeks. Finally, in the 'diversification' experiment, *A. alpina* (Paj), *C. hirsuta* (Oxford) and *A. thaliana* (Col-0) plants (four plants per pot) were grown in three batches of Cologne soil, that is, three full factorial biological replicates, under controlled environmental conditions for 6 weeks until harvest.

16S rRNA gene sample preparation, sequencing and analysis

We collected soil, rhizosphere and root compartments and prepared a DNA template for further processing using established protocols (Schlaeppi *et al*., 2014). Briefly, amplicon libraries were prepared using the primers 799F (Chelius and Triplett, 2001) and 1193R (Bodenhausen *et al*., 2013). Sequencing of samples with 454 pyrosequencing (Branford, CT, USA; 'diversification' experiment) was conducted as previously described (Schlaeppi *et al*., 2014), while the majority of the samples ('soil type and environment' and 'time course' experiment) were sequenced using Illumina (MiSeq) paired-end sequencing (San Diego, CA, USA; using a slightly modified PCR-amplification protocol, but targeting the same 16S rRNA gene region, see Supplementary Notes for details). We prepared a total of 201 DNA samples for sequencing: 59, 106 and 36 samples belonging to the 'soil type and environment', 'time course' and 'diversification' experiment, respectively. In addition, 11 samples of the 'diversification' experiment sequenced by the 454 technology, were re-sequenced using the Illumina protocol to investigate potential methodological biases (Table 1, Supplementary Figure S1). The bioinformatic analysis of sequence reads included de-multiplexing of samples using the QIIME software package (Caporaso *et al.*, 2010) and determination of operational taxonomic units (OTUs) on the three concatenated datasets using the UPARSE pipeline (Edgar, 2013). Afterwards, OTU tables were separated and investigated independently based on the three experimental setups (if not stated otherwise) and closer investigated by employing Principal Coordinate Analysis (PCoA) analyses and established linear model statistics using analysis of variance (ANOVA) and Bayesian model-based moderated *t*-tests to calculate differentially enriched OTUs (Bulgarelli *et al.*, 2012, 2015).

Data deposition

Raw sequencing data were deposited in the NCBI Short Read Archive (SRA), BioProjectID PRJNA317760, accession number SRP073035. The reference numbers for the original raw sequencing data are SRR3350817 (L1151), SRR3351958 (L1264), SRR3353796 (L1291), L35 (SRR3355061), L905 (SRR3355062) and L1118 (SRR3355063). Custom R scripts can be found at http://www.mpipz.mpg.de/R_scripts. This provides a downloadable folder including the used R script, input data files and custom R-packages that were separated for the individual parts of the analysis.

Results

Marked bacterial community shifts in both A. alpina *rhizosphere and root compartments*

Plants were grown in microbe-containing soils and root, rhizosphere and soil compartments sampled as described previously (Figure 1; Bulgarelli *et al.*, 2012; Schlaeppi *et al.*, 2014). Briefly, the 'soil compartment' refers to soil collected from unplanted pots and the 'rhizosphere compartment' defines soil particles tightly adhering to roots that were collected by two washing steps. Washed roots were additionally sonicated to deplete bacterial epiphytes and enrich for root-inhabiting bacteria, designated 'root compartment'. Bacterial community profiles for each compartment were generated by PCR amplification of the 16S rRNA gene by targeting region V5-V7 using PCR primers 799F (Chelius and Triplett, 2001) and 1193R (Bodenhausen *et al.*, 2013) followed by 454 or Illumina sequencing (Table 1 and Supplementary Information). We generated a total of 10 138 758 high-quality sequences from 201 samples (average of 50 342 and 2804-155 004 sequences per sample, individual read counts supplied in Supplementary Data S1). In addition, 11 samples of the 'diversification' experiment that were sequenced with 454 were re-sequenced using the Illumina methodology and protocol, to estimate a potential methodological bias (533 801 total reads, Table 1, Supplementary Figure S1). The separation pattern between compartments was consistent between 454 and Illumina methodologies (Supplementary Figure S2) and we did not detect marked differences in taxonomic assignments (Supplementary Figure S3, see Supplementary Notes for details). This shows that the two sequencing technologies produce comparable results. For the 201 samples included in the main analysis of this study, we defined 8969 OTUs (⩾97% sequence similarity of 16S rRNA gene sequences), all belonging to the kingdom Bacteria. Across experiments, the tested compartment (soil, rhizosphere or root) explained 18–36% of community variation among samples (Figure 2a, Table 2). At phylum rank, unplanted soil communities were dominated with decreasing relative abundance by Proteobacteria, Actinobacteria and Gemmatimonadetes (Supplementary Figures S4–S7). In comparison to unplanted soil, root microbial communities displayed a significantly reduced alpha-diversity (Supplementary Figure S8, Supplementary Table S1) and were mainly inhabited by Proteobacteria, Actinobacteria and Bacteroidetes in all soil types tested. This co-occurrence of three bacterial phyla corresponds well with a recent assessment of the root microbiota of eight flowering plant species (Hacquard *et al.*, 2015). Taken together, the tested compartments had a pronounced effect on community structure, with the rhizosphere community displaying an alpha-diversity that is comparable to unplanted soil, whereas its taxonomic composition shared features with the soil and the highly distinctive root microbiota (Figure 2a, Supplementary Figures S4–S8).

The A. alpina *root microbiota is dependent on soil type but exhibits a converging taxonomic structure*

We analyzed the 59 samples from the 'soil type and environment' experiments separately, enabling us to dissect the relative contribution of the factors soil type, environment and host genotype on bacterial community composition (Table 1). We compared the taxonomic composition of bacterial assemblages in root and rhizosphere compartments with the soil biome of the *A. alpina* accession Gal60 in its native arctic-alpine environment at the Col du Galibier (France, Figure 1a). In addition, the Gal60 accession was grown in the same native soil, but under controlled environmental conditions, together with the accessions Gal5 and Paj that are derived from a different site at the Col du Galibier and the Cordillera Cantábrica in Spain, respectively. In a further subset of this experiment, the accession Paj was grown under controlled environmental conditions in Cologne soil (Figure 1b, Supplementary Tables S3 and S4).

To quantify the contribution of individual factors to the observed variation in community profiles, we used Bray–Curtis (accounting for OTU abundances but not phylogenetic distances), unweighted

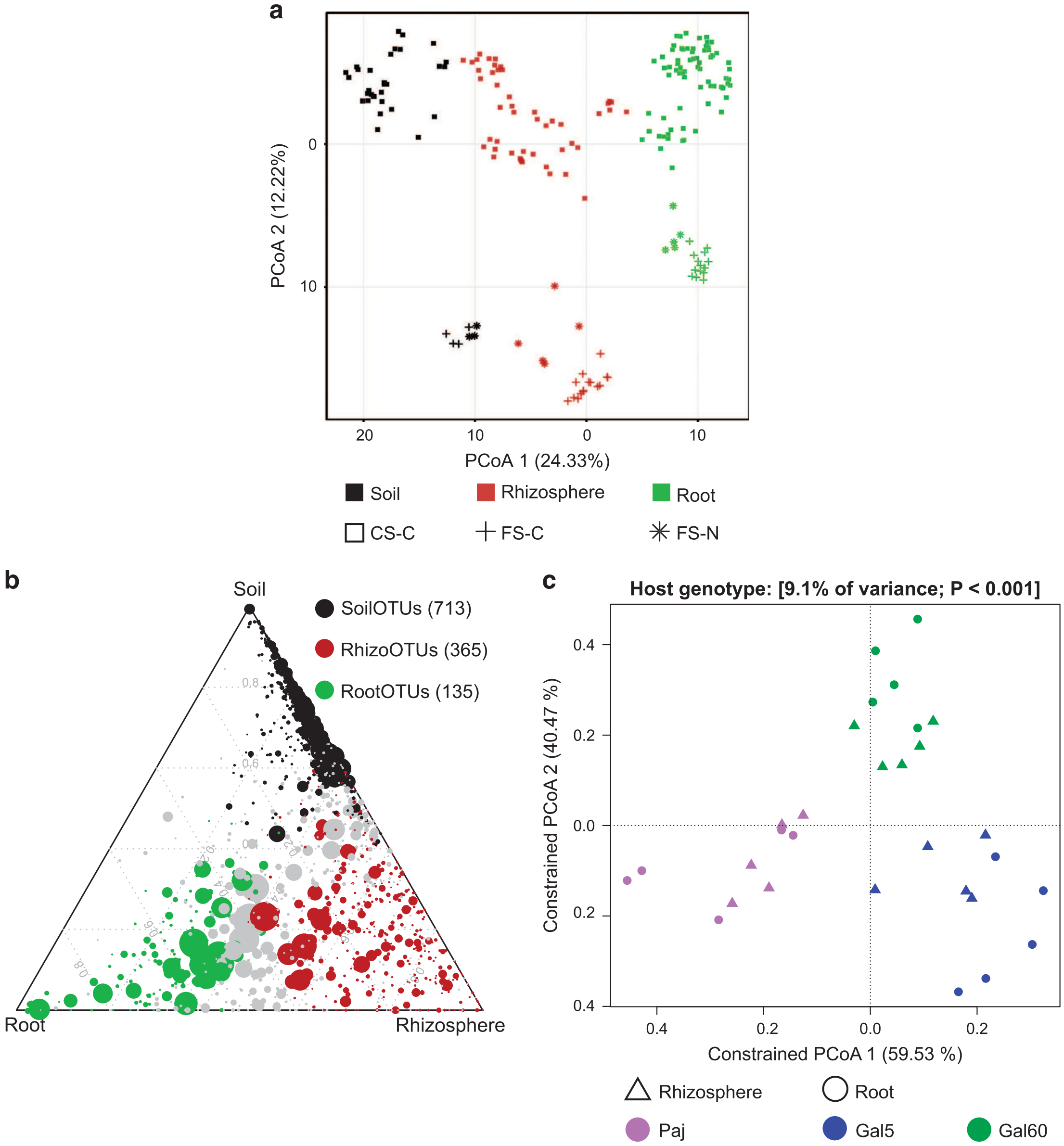


Figure 2 Bacterial community shifts in *A. alpina* rhizosphere and root compartments. (**a**) Unconstrained ordination revealing that most of the variation among all 201 samples from the 3 experimental setups is explained by the factor compartment (first principal coordinate axis) and soil type (second principal coordinate axis) based on the Bray–Curtis distance metric. (**b**) Ternary plot depicting the number of OTUs enriched in the soil, rhizosphere and root compartments of the 59 samples of the 'soil type and environment' experiment (SoilOTUs (black), RhizoOTUs (brown), RootOTUs (green), respectively). Each circle depicts one individual OTU. The size of the circle reflects the relative abundance (RA). The position of each circle is determined by the contribution of the indicated compartments to the RA. Number in brackets: Enriched OTUs based on a Bayes moderated *t*-test; $P < 0.05$ (FDR-corrected). (**c**) Constrained principal coordinates analysis on the genotypes Paj, Gal60 and Gal5 grown in the French Soil in the greenhouse using the Bray–Curtis dissimilarity and constraining by host genotype. In each case, the percentage of variation explained by each axis refers to the fraction of the total variance of the data explained by the constrained factor. CS-C, Cologne soil grown under controlled environmental conditions' FS-C, French soil grown under controlled environmental conditions; FS-N, French soil grown under native environmental conditions.

(accounting for phylogenetic distances but not abundances) and weighted (accounting for OTU abundances and phylogenetic distances) UniFrac distance metrics as β-diversity (between-sample diversity) estimates (Lozupone and Knight, 2005). A constrained Canonical Analysis of Principal coordinates (CAP), followed by a permutation-based ANOVA (PERMANOVA, Table 2, Figure 2c, Supplementary Figure S9) revealed that the factor soil type explained 11–15% of the observed variation, followed by environment (8–11%) and host genotype (5–12%). Both soil type and environment significantly influenced community composition, and we found no significant interaction between the soil type and environment variables (Supplementary Figure S9a). Notably, host genotype

Table 2 Determining drivers of bacterial community assembly using CAP

	Constrained factor	*Bray–Curtis*			*Weighted Unifrac*			*Unweighted Unifrac*		
		%	P-*value*	*CI*	%	P-*value*	*CI*	%	P-*value*	*CI*
Soil type+environment	Compartment[a]	31	***	21, 48	33	***	20, 52	36	***	23, 64
	Environment[a]	11	***	8, 16	8.2	***	5, 13	9.8	***	7, 14
	Compartment[b]	25	***	15, 47	24	***	13, 46	37	***	19, 78
	Genotype[b]	9.1	***	7.3, 11	12	***	8.5, 16	4.8	*	4.2, 5.5
	Compartment[c]	31	***	20, 50	39	***	22, 67	33	***	21, 53
	Soil type[c]	15	***	9, 26	11	***	6, 20	15	***	10, 25
Time course	Compartment	19	***	14, 27	25	***	17, 39	18	***	14, 26
	Time point	9.9	***	7.8, 13	12	***	8.5, 18	6.6	***	5.7, 7.8
	Soil batch	6.6	***	4.9, 9.1	5.3	***	3.6, 8	8.4	***	6.6, 11
	Flowering stage	1.5	NS	1.2, 1.9	1.3	NS	1.0, 1.9	1.8	*	1.3, 2.5
	Mutant background	0.6	NS	0.4, 0.8	0.8	NS	0.5, 1.2	0.7	NS	0.5, 1.1
Diversification	Compartment	21	***	12, 37	26	***	12, 51	18	***	12, 31
	Soil batch	19	***	14, 28	20	***	12, 32	18	***	14, 24
	Plant species	10	***	8, 14	10	***	7, 15	7.3	***	6, 9

Abbreviations: CAP, constrained analysis of principal coordinates; CI, confidence interval; NS, non-significant; PERMANOVA, permutation-based analysis of variance.
Variation (in %) between samples in the 'soil type and environment', 'time course' and 'diversification' experiments based on Bray–Curtis, weighted and unweighted UniFrac distances, constraining for the indicated factors. *P*-value based on PERMANOVA (999 permutations). *$P<0.05$, **$P<0.01$, ***$P<0.001$. Samples were analyzed separately according to the three different experimental setups as stated in Table 1. In addition, the 'soil type and environment' samples were separated as follows:
[a] Gal60 grown at the natural site and the greenhouse.
[b] Gal60, Gal5 and Paj grown in French soil in the greenhouse.
[c] Paj grown in the French and Cologne soil in the greenhouse.

affected community composition with a higher significance using weighted compared with unweighted UniFrac, suggesting that host genotype preferentially acts on abundant community members. At the time of harvest, the accession Gal5 was flowering while Gal60 and Paj resided in the vegetative stage (Figure 1b). Because only Paj requires a vernalization treatment to flower (Gal5 and Gal60 do not), the differences in developmental stages could potentially confound the observed host genotype-dependent effects (see below). In addition, the three genotypes were only grown in one soil type; therefore, we cannot make conclusions on potential environment-specific adaptations of individual genotypes. The impact of soil type on community profiles was apparent at the highest taxonomic rank. For example, differential enrichments of the phyla Actinobacteria and Firmicutes were found in French and Cologne soils (Supplementary Table S5 and Supplementary Figure S4). In addition, soil type-dependent differences of the bacterial soil biome at phylum rank converged progressively towards more soil type-independent community profiles in rhizosphere and root compartments (Supplementary Table S5), indicating that roots provide a niche for the assembly of a stable bacterial consortium despite contrasting soil types and environments.

Next, we quantified the number of OTUs that were enriched in a compartment across soil types and environments (Supplementary Information; Bulgarelli *et al.*, 2012). We refer to OTUs significantly enriched in soil, rhizosphere and root compartments as SoilOTUs, RhizoOTUs and RootOTUs, respectively. Of a total of 1676 OTUs, the compartment-enriched OTUs gradually declined from 713 SoilOTUs to 365 RhizoOTUs and 135 RootOTUs, with the remaining 463 OTUs being shared among compartments (Figure 2b). Of the RootOTUs, ~30% were shared across all tested soil types and environmental conditions (Supplementary Figure S10). Taken together, this analysis shows that *A. alpina* roots accommodate a taxonomically congruent bacterial community in contrasting environments and produce a marked rhizosphere effect, as evidenced by the identification of 365 RhizoOTUs.

The A. alpina *root microbiota is dependent on soil residence time and independent of flowering time*

To evaluate potential changes of *A. alpina* root- and rhizosphere-associated microbiota composition over time, *A. alpina* samples were collected at 6, 12 and 28 weeks after sowing in Cologne soil under controlled environmental conditions (Table 1, Supplementary Information). In the context of this 'time-course' experiment we also investigated whether host developmental stage (non-flowering and flowering) affects bacterial community structure. We planted the WT *A. alpina* accession Paj together with the perpetually flowering mutant *pep1* (Wang *et al.*, 2009). This mutant started to flower after

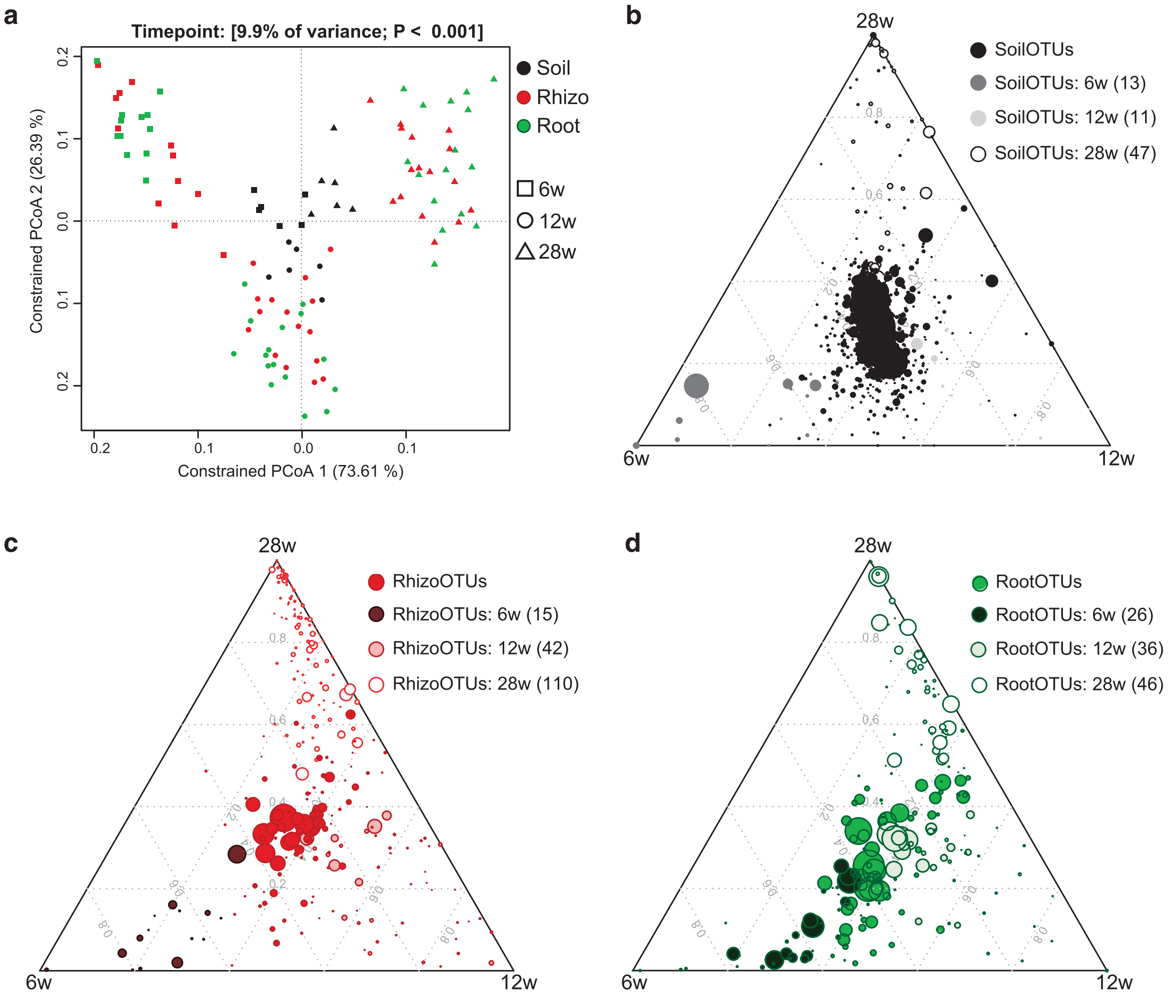


Figure 3 *A. alpina* root-associated bacterial communities depend on soil residence time. (**a**) Principal coordinate analysis on samples from the 'time course' experiment (106 samples) constrained for the factor time point for soil, rhizosphere and root compartments at 6, 12 and 28 weeks, respectively. 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–d**) Soil-, rhizosphere- and root-enriched OTUs (RootOTUs, RhizoOTUs, SoilOTUs) including OTUs observed across all tested time points and OTUs specific for individual time points. Time point-specific SoilOTUs (**b**), RhizoOTUs (**c**) and RootOTUs (**d**). The mean of soil (**b**), rhizosphere (**c**) or root (**d**) samples at each time point is plotted. Number in brackets: total number of OTUs enriched in the respective compartment or time point. Enriched OTUs are based on a Bayes moderated *t*-test; $P<0.05$ (FDR-corrected).

12 weeks, whereas WT Paj plants remained in the vegetative stage until the end of the experiment.

Within this experimental setup, the factor time point explained 7–12% of the observed variation (Table 2, Figure 3a). Considering the interaction between time point and compartment, we found a strong interaction between those two factors, explaining 11% of the variance of the data (Supplementary Figure S9b). Interestingly, only root and rhizosphere community profiles were affected over time, while the initial bacterial community present in unplanted soil remained largely stable over the tested time frame from 6 to 28 weeks across the used soil batches (Figure 3, Supplementary Figure S5). The stability of soil samples over time was apparent since these samples did not separate within the PCoA and most SoilOTUs remained in the center of the ternary plot, reporting their equal abundance across all tested time points. The observed strong impact of residence time of plants in soil on bacterial root and rhizosphere communities was apparent at high taxonomic rank, for example, by an increase in the relative abundance of Proteobacteria and decrease of Bacteroidetes in roots over time (Supplementary Figure S5 and Supplementary Table S6). To identify OTUs that explain these time point-dependent changes, we first determined all SoilOTUs (745), RhizoOTUs (271) and RootOTUs (200). Around 62% of RhizoOTUs and 54% RootOTUs significantly changed in relative abundance over time, whereas this was only the case for ~10% of SoilOTUs (Figures 3b–d). Closer inspection showed that changes in rhizosphere and root microbiota profiles were characterized by a continuous increase in the total number of Rhizo- and RootOTUs over time (Figures 3c and d).

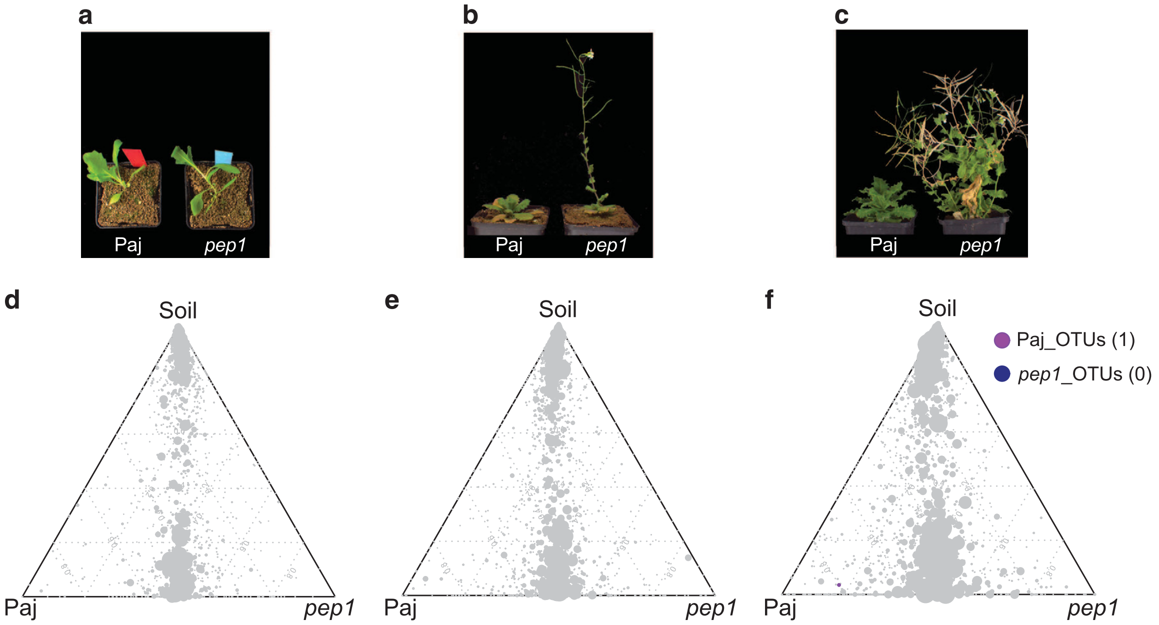


Figure 4 The *A. alpina* root microbiota is independent of flowering time. (**a–c**) Growth morphology of *A. alpina* wild-type (Paj) and *pep1* mutant plants across the 'time course' experiment. (**d–f**) Ternary plots of OTUs enriched in the 'time course' experiment across the soil and root compartments of the *A. alpina* wild-type (Paj) and *pep1* mutant plants after 6 weeks (**a**, **d**), 12 weeks (**b**, **e**) and 28 weeks (**c**, **f**). Each circle depicts one individual OTU. The size of the circle reflects the relative abundance (RA). The position of each circle is determined by the contribution of the indicated compartments to the RA. Number in brackets: total number of OTUs enriched in the respective plant line. OTUs enriched in the two plant lines are based on a Bayes moderated *t*-test; $P<0.05$ (FDR-corrected).

Remarkably, despite drastic differences in plant stature, we did not detect significant changes between the root and rhizosphere microbiota of WT and *pep1* mutant plants (Table 2, Figure 4). None of the Root- or RhizoOTUs differentially accumulated in 6- or 12-week-old WT and *pep1* plants, when 12-week-old *pep1* plants were flowering (Figures 4d and e). A single low-abundant OTU (OTU927, order Myxococcales) was enriched in 28-week-old WT plants when *pep1* plants were flowering and WT plants remained in the vegetative phase (Figure 4f). In addition, we detected stable nutrient contents and soil parameters in unplanted soil compared with soil samples recovered from pots containing 28-week-old plants (Supplementary Tables S3 and S4). In summary, our findings reveal marked residence-time-dependent changes of *A. alpina* root- and rhizosphere-associated bacterial consortia, but a stability of the root microbiota against the onset and perpetual flowering of *A. alpina* and concomitant drastic changes in plant stature.

The composition of the root microbiota of perennial A. alpina *is similar compared with the annuals* A. thaliana *and* C. hirsuta

Next, we planted perennial *A. alpina* side-by-side with the annuals *C. hirsuta and A. thaliana* in Cologne soil for 6 weeks to examine potential diversification of the corresponding root microbiota (Table 1, Supplementary Figures S11a and b and Supplementary Information). We found evidence for host species-specific root microbiota profiles in which the factor host species explained 7–10% of the observed variation (Table 2). PCoA, constrained by host species, revealed that samples from *A. alpina* roots clustered closely together with those of *A. thaliana* although these plant species diverged ~45 Myr ago (Figure 5, Supplementary Figure S11a; Beilstein *et al.*, 2010). Root samples of *C. hirsuta* clustered separately from the former two hosts although *A. alpina* diverged from *C. hirsuta* only ~10 Myr ago (Figure 5a). Thus, for the tested plant species host divergence time appears to be incongruent with root microbiota diversification. The impact of host species on bacterial root communities was detectable at phylum rank by a significant reduction in the relative abundance of Bacteroidetes in root samples of *A. alpina* (Supplementary Figure S7a). In addition, significant differences were detected for one, three and five bacterial families in *A. thaliana*, *C. hirsuta* and *A. alpina*, respectively (Supplementary Figure S7b, Supplementary Table S7). Each tested host species was also enriched for a few species-characteristic OTUs that displayed quantitative differences in relative abundance (designated *At*OTUs, *Ch*OTUs and *Aa*OTUs; Figure 5b). However, the majority of RootOTUs (62) were similarly detected in roots of all three host species (Figure 5b). In addition, a total of 18 RootOTUs were significantly enriched in roots of all 3 plant species across the tested soil batches (Figure 5c, Supplementary Figure S12a). These shared OTUs were barely detectable in unplanted soil and together represented almost 50% of the root-associated consortia (Supplementary Figure S12b). The shared OTUs belonged to 3 phyla, with 12 OTUs assigned to Proteobacteria and 3 each to Actinobacteria and Bacteroidetes (Figure 5c). In summary, we found few host species-characteristic OTUs, while the majority of the root microbiota was shared among the tested host species.

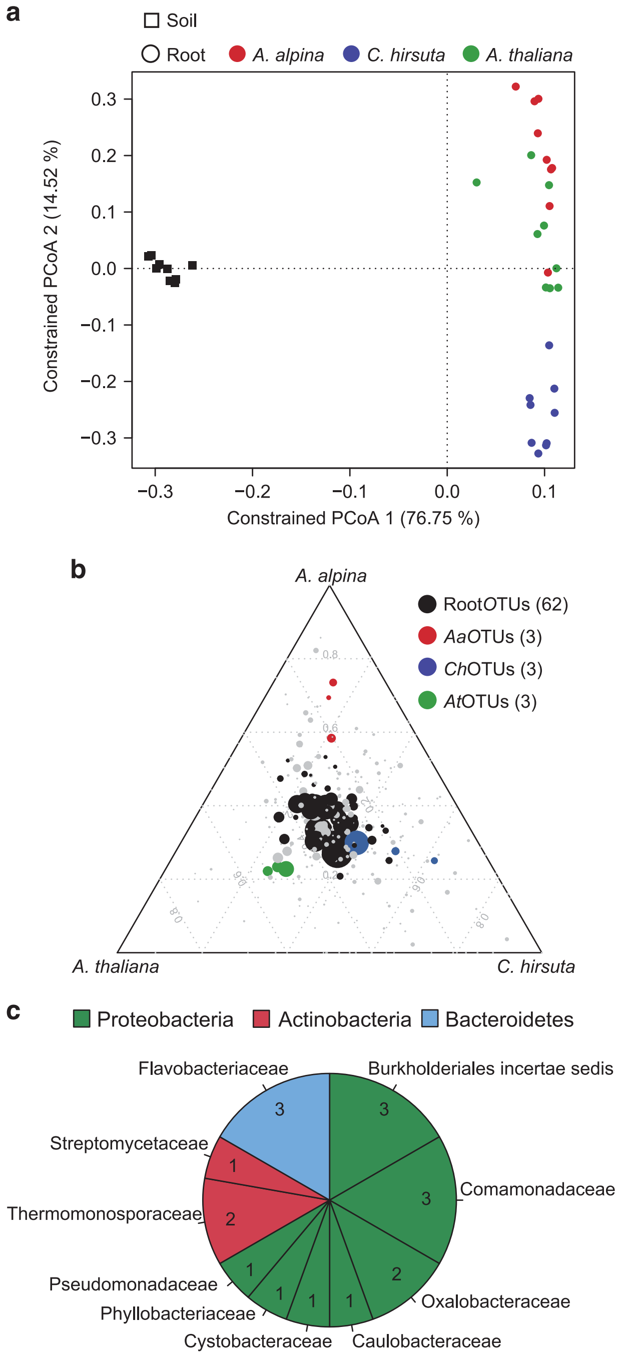


Figure 5 The bacterial root microbiota of *A. alpina* is similar compared with *A. thaliana* and *Cardamine hirsuta*. (**a**) Constrained principal coordinates analysis (PCoA) based on the Bray–Curtis distances on the 36 samples of the 'diversification' experiment constrained by sample groups. (**b**) OTUs enriched on roots of *A. alpina*, *C. hirsuta* and *A. thaliana*. Color-coded in black are root-enriched OTUs (RootOTUs). All enriched OTUs are based on a Bayes moderated *t*-test, $P<0.05$ (FDR-corrected). Number in brackets: total number of OTUs enriched in the respective plant species. (**c**) Pie chart reporting family distribution of 18 shared OTUs based on parametric Tukey's honest significant difference and Bayesian and non-parametric Mann-Whitney statistics.

Discussion

Here, we quantified the relative contributions of factors explaining variation in *A. alpina* root-associated bacterial consortia by analyzing high-quality 16S rRNA reads from >200 samples and multiple environmental factors. This revealed the variable compartment as strongest determinant of community variation (18–36%). Additional variables influencing community variation were soil type (11–15%), residence time of plants in soil (7–12%), environmental conditions (8–11%), host species (7–10%) and host genotype (5–12%). The relative contributions of the factors compartment, soil type and host species in this study are comparable to an earlier report on root microbiota diversification between *A. thaliana* relatives (Schlaeppi *et al.*, 2014). Despite the observed host species-dependent differences in root microbiota profiles, we identified a largely shared bacterial assemblage whose taxonomic structure is similar to the microbiota intersection between *A. thaliana*, *A. halleri*, *A. lyrata* and *C. hirsuta* (Schlaeppi *et al.*, 2014). This points to potentially common services provided by these bacterial taxa for plant growth and health across five tested Brassicaceae host species. However, in contrast to previous work showing only weakly differentiated bacterial rhizosphere assemblages compared with unplanted soil biomes in annual *A. thaliana* and *C. hirsuta* plants that were partially performed in the same soil type and growth conditions (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Schlaeppi *et al.*, 2014), we detected a pronounced rhizosphere effect for *A. alpina* at all tested time points and across soil types and environments. This suggests that within the Brassicaceae, evolutionary diversification has apparently produced host species-specific differences in bacterial community differentiation in the rhizosphere compartment despite an overall convergent community composition inside roots (Bulgarelli *et al.*, 2012; Lundberg *et al.*, 2012; Schlaeppi *et al.*, 2014). This striking feature of *A. alpina* may be due to its phylogenetic distance to *A. thaliana* and *C. hirsuta*, adaptation to the arctic-alpine habitat or its perennial lifestyle.

Based on similar CAP analyses, the variation of root microbiota profiles explained by intra-species genetic diversity in *A. alpina* (7–10%) is remarkably similar to those reported for corn (5–8%; *Zea mays*), barley (~6%; *Hordeum vulgare*) and rice (3–5%; *Oryza sativa*), suggesting that host genotype determines a small, but significant amount of microbiota variation across all tested flowering plants (Peiffer *et al.*, 2013; Bulgarelli *et al.*, 2015; Edwards *et al.*, 2015). In addition, our study supports and extends earlier observations that the Brassicaceae host phylogeny is incongruent with root microbiota diversification (Schlaeppi *et al.*, 2014), suggesting that host and root microbiota structure are subject to independent diversification processes. However, our experimental approach does not allow us to assess whether inter-exceeds intra-species root microbiota diversification, as this requires information on population structure and genetic diversity of each host and sampling in more diverse environments. This experimental limitation does not permit us to completely exclude correlated evolutionary processes driving host and microbiota diversification.

Previous work with *A. thaliana* suggested that the root-associated bacterial assemblage is affected by plant development, showing differences in community structure between seedling and vegetative stage and bacterial transcripts induced at bolting and flowering stages (Chaparro *et al.*, 2014). However, these experiments were conducted with WT *A. thaliana* only and do not exclude the possibility that changes in root microbiota structure or activity are influenced by differences in the residence time of the plants in soil. While we cannot exclude differences in root microbiota activity, bacterial community structure on WT *A. alpina* and *pep1* mutant plants was essentially indistinguishable throughout the analyzed time course, even though the *pep1* mutant started to flower during the analysis whereas the WT did not. This strong, genetically determined difference in flowering time enabled us to uncouple flowering from plant residence time-dependent changes in the microbiota. This strongly suggests that the bacterial root assemblage, once established during the vegetative plant growth phase, is structurally robust to the onset and perpetual flowering of *A. alpina*. During the transition to flowering, the shoot apex becomes a sugar sink (Bernier, 1988) and such changes might reduce photosynthates available for transport to the roots. In addition, enhanced root secretion of defense-related proteins has been reported during flowering of *A. thaliana* (De-la-Peña *et al.*, 2010). Such metabolic changes do not detectably affect the *A. alpina* root microbiota as evidenced by an indistinguishable root-associated bacterial community composition of non-flowering WT and flowering *pep1* mutant plants. Conversely, differences in soil bacteria between soil types can influence flowering time as recently shown for perennial *B. stricta* and annual *A. thaliana* although the underlying mechanisms and specific bacterial taxa causing these flowering time shifts remain unidentified (Wagner *et al.*, 2014; Panke-Buisse *et al.*, 2015). Collectively, this suggests a unidirectional rather than bidirectional interference model in which particular root microbiota members modulate flowering time, whereas the established root-associated microbial community remains robust to perturbations caused by flowering onset and drastic changes in plant stature. This stability of root-associated bacterial assemblages against changes in host developmental stage or physiological status is supported by the finding that community profiles of *A. thaliana* roots at bolting and leaf senescence stage were indistinguishable (Lundberg *et al.*, 2012). Acquisition of the root microbiota in rice plants was shown to initiate within 24 h and approach steady-state within 14 days (Edwards *et al.*, 2015). Thus, it is possible that after microbiota acquisition at the seedling stage, microbe–microbe interactions stabilize the root-associated community composition, thereby rendering the assemblage resistant to flowering time-associated metabolic changes in the host.

We monitored microbiota profiles of *A. alpina* roots for up to 7 months and found marked changes over time at all tested taxonomic ranks (7–12% of variation) that are dependent on the residence time of plants in soil rather than plant developmental stage or plant stature. These long-term dynamics of the *A. alpina* root microbiota contrasts with the stability of the bacterial community in unplanted soil, monitored in parallel over the same time period. Comparable long-term time-course experiments are not possible with short-lived annual *A. thaliana* and time-course experiments covering most of the life cycle of other perennial or long-lived annual plants are, to our knowledge, currently unavailable. Given the robustness of the *A. alpina* root-associated bacterial assemblage to flowering onset and changes in plant stature (see above), we speculate that the observed long-term root microbiota dynamics are a consequence of competition among bacteria rather than altered host-bacteria interactions. Alternatively, unknown host factors linked to the longevity of *A. alpina* such as a local depletion of soil-borne mineral micronutrients due to prolonged plant growth might contribute to the observed microbiota dynamics. It will be interesting to investigate in future work whether other perennial plant species exhibit a similar soil residence time-dependent dynamics of the root microbiota and whether the observed community shift is linked to altered community functions.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We thank Dr Maria Albani for providing seed material from *A. alpina*, Dr Girish Srinivas for bioinformatic support and Seema Thaliyath for processing DNA samples. This work was supported by funds to PS-L from the Max Planck Society, a European Research Council advanced grant (ROOT MICROBIOTA), and by funds to PS-L and GC from the ''Cluster of Excellence on Plant Sciences'' program funded by the Deutsche Forschungsgemeinschaft.

Author contributions

ND, KS, GC and PS-L conceived this study and supervised the experiments and analyses. ND performed the experiments. MA, SH and EK developed the Illumina sequencing protocol and RGO provided computational tools for quantitative community analysis. ND and JW collected *A. alpina* plants and soil material from natural environments. ND, KS and PS-L wrote the paper with contributions from all authors.

References

Albani MC, Castaings L, Wötzel S, Mateos JL, Wunder J, Wang R *et al.* (2012). PEP1 of *Arabis alpina* is encoded by two overlapping genes that contribute to natural genetic variation in perennial flowering. *PLoS Genet* **8**: e1003130.

Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M *et al.* (2015). Functional overlap of the Arabidopsis leaf and root microbiota. *Nature* **528**: 364–369.

Beilstein MA, Nagalingum NS, Clements MD, Manchester SR, Mathews S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. *Proc Natl Acad Sci USA* **107**: 18724–18728.

Bernier G. (1988). The control of floral evocation and morphogenesis. *Annu Rev Plant Physiol Plant Mol Biol* **39**: 175–219.

Billings WD, Mooney HA. (1968). The ecology of arctic and alpine plants. *Biol Rev* **43**: 481–529.

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

Bulgarelli D, Garrido-Oter R, Münch PC, Weiman A, Dröge J, Pan Y *et al.* (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe* **17**: 392–403.

Bulgarelli D, Rott M, Schlaeppi K, Themaat EVL, van, Ahmadinejad N, Assenza F *et al.* (2012). Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature* **488**: 91–95.

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

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK *et al.* (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Chaparro JM, Badri DV, Vivanco JM. (2014). Rhizosphere microbiome assemblage is affected by plant development. *ISME J* **8**: 790–803.

Chapin FS. (1983). Direct and indirect effects of temperature on arctic plants. *Polar Biol* **2**: 47–52.

Chapin FS, Shaver GR. (1989). Differences in growth and nutrient use among arctic plant growth forms. *Funct Ecol* **3**: 73.

Chelius MK, Triplett EW. (2001). The diversity of Archaea and Bacteria in association with the roots of *Zea mays* L. *Microb Ecol* **41**: 252–263.

De-la-Peña C, Badri DV, Lei Z, Watson BS, Brandão MM, Silva-Filho MC *et al.* (2010). Root secretion of defense-related proteins is development-dependent and correlated with flowering time. *J Biol Chem* **285**: 30654–30665.

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

Edwards J, Johnson C, Santos-Medellín C, Lurie E, Podishetty NK, Bhatnagar S *et al.* (2015). Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc Natl Acad Sci USA* **112**: E911–E920.

Fierer N, Jackson RB. (2006). The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**: 626–631.

Hacquard S, Garrido-Oter R, González A, Spaepen S, Ackermann G, Lebeis S *et al.* (2015). Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* **17**: 603–616.

Lauber CL, Hamady M, Knight R, Fierer N. (2009). Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Appl Environ Microbiol* **75**: 5111–5120.

Lozupone C, Knight R. (2005). UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**: 8228–8235.

Lundberg DS, Lebeis SL, Paredes SH, Yourstone S, Gehring J, Malfatti S *et al.* (2012). Defining the core *Arabidopsis thaliana* root microbiome. *Nature* **488**: 86–90.

Maignien L, DeForce EA, Chafee ME, Eren AM, Simmons SL. (2014). Ecological succession and stochastic variation in the assembly of *Arabidopsis thaliana* phyllosphere communities. *mBio* **5**: e00682–13.

Margesin R, Miteva V. (2011). Diversity and ecology of psychrophilic microorganisms. *Res Microbiol* **162**: 346–361.

Nissinen RM, Männistö MK, van Elsas JD. (2012). Endophytic bacterial communities in three arctic plants from low arctic fell tundra are cold-adapted and host-plant specific. *FEMS Microbiol Ecol* **82**: 510–522.

Panke-Buisse K, Poole AC, Goodrich JK, Ley RE, Kao-Kniffin J. (2015). Selection on soil microbiomes reveals reproducible impacts on plant function. *ISME J* **9**: 980–989.

Peiffer JA, Spor A, Koren O, Jin Z, Tringe SG, Dangl JL *et al.* (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions. *Proc Natl Acad Sci USA* **110**: 6548–6553.

Richardson AE, Barea J-M, McNeill AM, Prigent-Combaret C. (2009). Acquisition of phosphorus and nitrogen in the rhizosphere and plant growth promotion by microorganisms. *Plant Soil* **321**: 305–339.

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

Vorholt JA. (2012). Microbial life in the phyllosphere. *Nat Rev Microbiol* **10**: 828–840.

Wagner MR, Lundberg DS, Coleman-Derr D, Tringe SG, Dangl JL, Mitchell-Olds T. (2014). Natural soil microbes alter flowering phenology and the intensity of selection on flowering time in a wild *Arabidopsis* relative. *Ecol Lett* **17**: 717–726.

Wang R, Farrona S, Vincent C, Joecker A, Schoof H, Turck F *et al.* (2009). PEP1 regulates perennial flowering in *Arabis alpina*. *Nature* **459**: 423–427.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)

Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities

Rafal Zgadzaj[a,b,c,1], Ruben Garrido-Oter[a,d,e,1], Dorthe Bodker Jensen[b,c], Anna Koprivova[d,f], Paul Schulze-Lefert[a,d,2], and Simona Radutoiu[b,c,2]

[a]Department of Plant Microbe Interactions, Max Planck Institute for Plant Breeding Research, 50829 Cologne, Germany; [b]Department of Molecular Biology and Genetics, Faculty of Science and Technology, Aarhus University, 8000 C Aarhus, Denmark; [c]Carbohydrate Recognition and Signaling Centre, 8000 C Aarhus, Denmark; [d]Cluster of Excellence on Plant Sciences, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany; [e]Department of Algorithmic Bioinformatics, Heinrich Heine University Duesseldorf, 40225 Duesseldorf, Germany; and [f]Botanical Institute, Cologne Biocenter, University of Cologne, 50674 Cologne, Germany

Contributed by Paul Schulze-Lefert, October 13, 2016 (sent for review June 3, 2016; reviewed by Philip N. Benfey and Sharon R. Long)

Lotus japonicus **has been used for decades as a model legume to study the establishment of binary symbiotic relationships with nitrogen-fixing rhizobia that trigger root nodule organogenesis for bacterial accommodation. Using community profiling of 16S rRNA gene amplicons, we reveal that in *Lotus*, distinctive nodule- and root-inhabiting communities are established by parallel, rather than consecutive, selection of bacteria from the rhizosphere and root compartments. Comparative analyses of wild-type (WT) and symbiotic mutants in Nod factor receptor5 (*nfr5*), Nodule inception (*nin*) and *Lotus* histidine kinase1 (*lhk1*) genes identified a previously unsuspected role of the nodulation pathway in the establishment of different bacterial assemblages in the root and rhizosphere. We found that the loss of nitrogen-fixing symbiosis dramatically alters community structure in the latter two compartments, affecting at least 14 bacterial orders. The differential plant growth phenotypes seen between WT and the symbiotic mutants in nonsupplemented soil were retained under nitrogen-supplemented conditions that blocked the formation of functional nodules in WT, whereas the symbiosis-impaired mutants maintain an altered community structure in the nitrogen-supplemented soil. This finding provides strong evidence that the root-associated community shift in the symbiotic mutants is a direct consequence of the disabled symbiosis pathway rather than an indirect effect resulting from abolished symbiotic nitrogen fixation. Our findings imply a role of the legume host in selecting a broad taxonomic range of root-associated bacteria that, in addition to rhizobia, likely contribute to plant growth and ecological performance.**

Lotus japonicus | microbiota | symbiosis | 16S | nitrogen fixation

The transition from an aquatic to terrestrial lifestyle during plant evolution required the formation of roots as organs for water, macronutrient, and micronutrient retrieval from soil. Nutrient-uptake systems of roots are usually specific for plant-available forms of nutrients, for example, inorganic nitrogen such as nitrate (NO_3^-) or inorganic orthophosphate (Pi) (1). However, phosphorus, per se, is abundant in soil in plant-inaccessible pools, and, likewise, atmospheric dinitrogen (N_2) is abundant in aerobic soil [78% (vol/vol)] but cannot be accessed by plants. Soil-resident microbes play important roles in the solubilization and conversion of mineral nutrients into root-available forms, and a subset of these microbes have acquired the capacity to engage in mutualistic interactions with plant roots to trade soil-derived bioavailable macronutrients for plant-derived photoassimilates (2–4).

Healthy, asymptomatic plants live in association with diverse microbes, including bacteria, fungi, viruses, and protists, collectively called the plant microbiota (3, 5). The bacterial root microbiota is taxonomically structured and characterized by the co-occurrence of three main phyla comprising Actinobacteria, Bacteroidetes, and Proteobacteria across different soil types and divergent plant hosts (6, 7). This root-associated bacterial assemblage is mostly derived from the highly diverse bacterial soil biome surrounding roots and is established rapidly within a few days after seed germination (6, 8). Soil type is the main driver of diversification of the bacterial root microbiota at low taxonomic ranks (i.e., at genus and species level), with less variation detectable at the higher phylum rank (8–11). However, root exudates are thought to play an important role as cues to initiate a substrate-driven competition between, and differential proliferation of, soil-resident microbes for root colonization (3, 12). An estimated 17% of photosynthetically fixed carbon is transferred to the rhizosphere, the thin layer of soil surrounding the root, through root exudation (13). These carbon substrates likely contribute to the bacterial community shifts that are often detected in the rhizosphere. A fraction of the bacterial taxa present in the rhizosphere colonize roots either as epiphytes on the root surface (rhizoplane) or as bacterial endophytes inside roots (3, 8). In particular, members of Proteobacteria are consistently found enriched in root and rhizosphere compartments, and diazotrophs within this phylum have evolved the capacity to establish a sophisticated form of mutualistic interaction with plant

Significance

Legumes are known as pioneer plants colonizing marginal soils, and as enhancers of the nutritional status in cultivated soils. This beneficial activity has been explained by their capacity to engage in symbiotic relationship with nitrogen-fixing rhizobia. We performed a community profiling analysis of *Lotus japonicus* wild type and mutants to investigate the role of the nodulation pathway on the structure of the root-associated bacterial microbiota. We found that several bacterial orders were almost entirely depleted from the mutant roots, and that an intact symbiosis is needed for the establishment of taxonomically diverse and distinctive bacterial communities in the root and rhizosphere. Our findings imply that a symbiosis-linked bacterial community, rather than dinitrogen-fixing rhizobia alone, contributes to legume growth and ecological performance.

Author contributions: R.Z., R.G.-O., P.S.-L., and S.R. designed research; R.Z., R.G.-O., D.B.J., and A.K. performed research; R.Z., R.G.-O., A.K., P.S.-L., and S.R. analyzed data; and R.Z., R.G.-O., P.S.-L., and S.R. wrote the paper.

Reviewers: P.N.B., Duke University; and S.R.L., Stanford University.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequence reported in this paper has been deposited in the European Nucleotide Archive (accession no. PRJEB15623).

[1]R.Z. and R.G.-O. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: schlef@mpipz.mpg.de or radutoiu@mbg.au.dk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1616564113/-/DCSupplemental.

roots, designated root nodule symbiosis. Unlike the taxonomically diverse root- and rhizosphere-associated bacterial communities that comprise a network of microbe–microbe and plant–microbe associations, the root nodule symbiosis defines a highly specific binary plant–microbe interaction where the compatible nitrogen-fixing soil bacterium is selected by the host for intracellular infection often via plant-derived infection threads and subsequent accommodation and amplification inside nodule cells.

Decades of bacterial and legume genetics allowed a detailed dissection of the regulatory networks behind the stepwise symbiotic association with diazotrophic Alphaproteobacteria. A two-way signal recognition initiates the interaction. Root-secreted flavonoids are perceived by the compatible soil bacteria, which start the production and secretion of the rhizobial symbiotic signal, the Nod factor. On the host side, lysin motif (LysM) receptor kinases, like Nod factor receptor1 (NFR1) and NFR5 in *Lotus japonicus*, specifically recognize and bind the compatible Nod factors (14, 15) and initiate the symbiotic signaling cascade. Nodule inception (NIN) was identified as an early key regulator of both nodule organogenesis and infection thread formation (16), whereas cytokinin signaling proteins involving *Lotus* histidine kinase1 (LHK1) receptor (17) control progression of the signaling events from root epidermis into the cortex (18). Inside nodules, a low-oxygen, carbon-rich environment is established by the host, allowing bacteria, upon endocytosis, to start the nitrogen fixation (19). Symbiotic nitrogen fixation reprograms the whole-root transcriptional and metabolic landscape (20–23). Moreover, the process is reiterative and highly asynchronous, because rhizobia from the rhizosphere recapitulate the infection on newly formed, competent root hairs. Nevertheless, the legume host controls the number of infection events and nodule primordia via shoot-derived signal(s) (24, 25).

Symbiotic nitrogen fixation allows legumes to thrive in habitats with limited nitrogen availability (26–28). The beneficial effect of this symbiosis is not limited to legume hosts, but extends to subsequent or concurrent plantings with nonlegumes as exemplified by ancient agricultural practices with legume cropping sequences or intercropping systems. This symbiosis likely involves a beneficial activity of legume roots and their associated microbes on the nutritional status of the soil as well as the soil biome. However, the mechanisms underpinning these symbiotic interactions in a community context and their impact on the complex microbial assemblages associated with roots remain largely unknown. Integrating these highly specific binary interactions into an ecological community context is critical for understanding the evolution of symbiosis and efficient use of rhizobia inoculum in agricultural systems.

Here, we investigated the role of symbiotic nitrogen fixation on the structure of the root-associated bacterial microbiota of the model legume *L. japonicus*. We performed bacterial 16S rRNA gene-based community profiling experiments of wild-type (WT) plants, grown in natural soil, and symbiotic mutants impaired at different stages of the symbiotic process. We have found that an intact nitrogen-fixing symbiosis in WT *Lotus* plants is needed for the establishment of taxonomically diverse and distinctive bacterial communities in root and rhizosphere compartments. This finding raises the possibility that the influence of legumes on soil performance in agricultural and ecological contexts is mediated by the enrichment of a symbiosis-linked bacterial community rather than dinitrogen-fixing rhizobia alone.

Results

Characterization of the *L. japonicus* Root, Nodule, and Rhizosphere Microbiota. We established a root fractionation protocol for 10-wk-old *L. japonicus* plants (accession Gifu, designated WT), grown in three batches of natural Cologne soil (10) to account for batch-to-batch and seasonal variation at the soil sampling site (Fig. 1*A* and *Materials and Methods*). This fractionation enabled us to compare the structure of bacterial communities present in nodules, roots without nodules (denoted hereafter as "root compartment"), the rhizosphere, and unplanted soil (*Materials and Methods* and *SI Appendix*, Fig. S1). Briefly, the "rhizosphere compartment" defines soil particles tightly adhering to *Lotus* roots that were collected after the first of two successive washing steps. Macroscopically visible nodules and nodule initials were excised from roots with a scalpel and designated the "nodule compartment." Pooled nodules and washed roots without nodules were separately subjected to a sonication treatment to deplete epiphytes and enrich for endophytic bacteria. Abundant nodulation (~20 nodules per plant) of healthy WT plants demonstrates that this soil is conducive for nodule formation and contains *Lotus*-compatible rhizobia (Fig. 1*A*, *Inset*). We subjected a total of 27 unplanted soil, 73 rhizosphere, 75 root, and 27 nodule samples to amplification of the 16S rRNA gene with PCR primers targeting the V5–V7 hypervariable regions (29) (*Materials and Methods*) and generated ~1 M high-quality sequencing reads (4,670 reads per sample on average). After removal of low-quality reads, chimeras, and sequences assigned to plant-derived organellar DNA, we clustered the data into 1,834 operational taxonomic units (OTUs) at 97% sequence similarity (*Materials and Methods* and Dataset S1).

To assess the effect of the different compartments on the assembly of bacterial communities, we compared the β-diversity (between-samples diversity) using Bray–Curtis distances and performed a canonical analysis of principal coordinates (CAP) (30) (*Materials and Methods*). This analysis revealed a clear differentiation of samples belonging to the root, rhizosphere, nodule, and soil compartments that explains as much as 19.97% of the overall variance of the data (Fig. 2*A*; $P < 0.001$), whereas the effect attributable to the soil batch was comparatively small (8.01% of the variance, $P < 0.001$). Analysis of α-diversity (within-samples diversity) using the Shannon index indicated a decreasing gradient of complexity from the soil bacterial communities (highest richness) to the rhizosphere, root, and, finally, the nodule microbiota (*SI Appendix*, Fig. S2).

Our finding of a bacterial community shift in the *Lotus* rhizosphere compared with the bulk soil reservoir is consistent with previous reports from WT pea (31), soybean (32), and peanut (33), in which a similar enrichment of members of Burkholderiales, Flavobacteriales, and Rhizobiales has been shown, whereas information on the community structure of the root microbiota is unavailable for other legumes.

Parallel Selection of Nodule- and Root-Specific Bacteria from the Rhizosphere Compartment. Legume nodules represent a unique environmental niche derived from differentiated cortical root cells where both symbiotic and nonsymbiotic bacteria are allowed accommodation and proliferation. Laboratory studies with single WT or mutant symbiotic strains demonstrated a stepwise, host-controlled colonization process ensuring symbiont selection (34). By contrast, little is known about the extent or the diversity of nodule and root colonization by nonsymbionts (35).

We took advantage of the compatible symbiotic association between *Lotus* and rhizobia present in Cologne soil and performed an analysis of the bacterial community of epiphyte-depleted, functional nodules of WT plants grown in this soil (*Materials and Methods*). We found that nodules were inhabited by a distinctive bacterial community compared with those present in the root and rhizosphere (Fig. 2*A*). Only a small number of the 1,834 OTUs were identified to be nodule-enriched (12 red circles in Fig. 3*A*) with one dominant member classified as belonging to the *Mesorhizobium* genus, substantiating that nodules also represent a highly selective bacterial niche for soil-grown *Lotus* plants. Importantly, the nodule and root communities were similarly divergent from the rhizosphere (separation by second component; Fig. 2*A*), and nodule-enriched OTUs were found in similar abundances in the root and rhizosphere samples (red circles in Fig. 3*A*). These findings suggest a parallel rather than consecutive selection of bacterial taxa from the rhizosphere assemblage for enrichment in

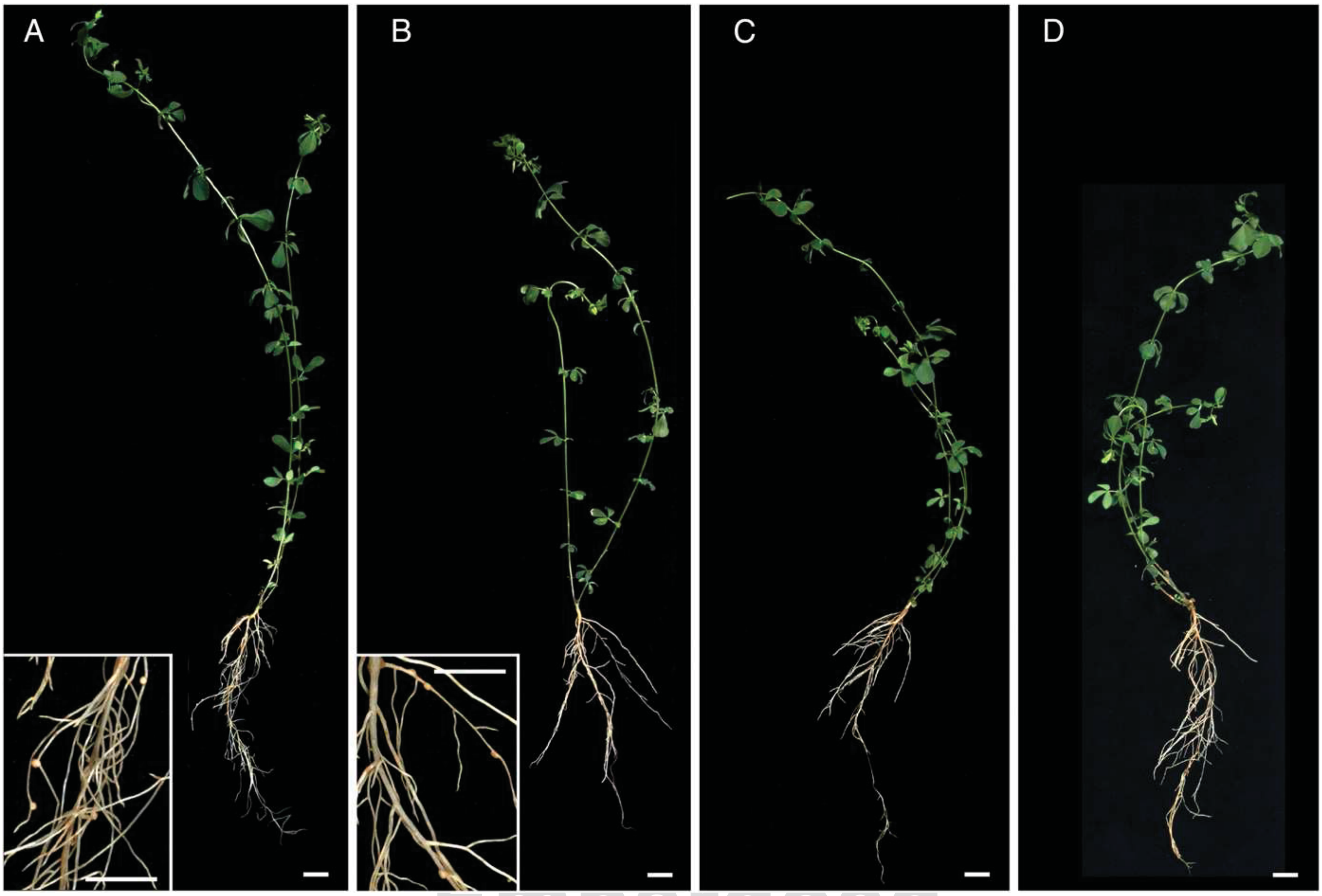


Fig. 1. Images depicting *L. japonicus* WT (*A*) and nodule symbiosis-deficient mutant plants *lhk1*-1 (*B*), *nfr5*-3 (*C*), and *nin*-2 (*D*) following harvest. (*A* and *B*, *Insets*) For nodulating genotypes, close-up views of nodules are shown. (Scale bars: 1 cm.)

the two endocompartments, most likely via host-induced infection threads. If there were a sequential selection, then nodule-enriched OTUs would be expected to be more abundant in the root compared with the rhizosphere. Taxonomic assignments at the order level for all OTUs with a relative abundance (RA) >5‰ revealed that the *Lotus* nodule community is dominated by bacteria belonging to the order Rhizobiales (88.01% average RA; Fig. 3*B*), which is mainly due to a selective enrichment of *Mesorhizobium* members (Fig. 3*C*). This analysis also revealed the presence of Burkholderiales, Flavobacteriales, Pseudomonadales, and Actinobacteridae at detectable abundances (>1% RA; Fig. 3*B*), showing that nodules of soil-grown *Lotus* are primarily, but not exclusively, colonized by symbiotic rhizobia.

Impairment of Nitrogen-Fixing Symbiosis Dramatically Alters Bacterial Community Structure in the *Lotus* Root and Rhizosphere Compartments. Next, we applied the same growth conditions, fractionation protocol, and bacterial community analysis to four symbiotic *Lotus* mutants (*nfr5*-2, *nfr5*-3, *nin*-2, and *lhk1*-1) to identify the role of the nodulation signaling pathway on bacterial assemblages. In *nfr5* and *nin* mutants, the infection process is either not initiated or terminated at the microcolony stage, respectively, whereas *lhk1* plants develop a large number of root hair infection threads that subsequently fail to infect cortical cells (16, 18, 36). Similar to WT, symbiotic mutant plants appeared healthy, but were smaller and had slightly pale green leaves (Fig. 1 *B–D*). With the exception of occasional nodules on *lhk1* roots (18), no nodules were found on *nfr5* or *nin* root systems (Fig. 1 *B–D*). Remarkably, we found that communities associated with the roots and rhizospheres of each of the four symbiosis mutants were similar to each other, but significantly different from the communities of WT plants (Fig. 2*B* and *SI Appendix*, Fig. S3). This separation between the mutant and WT samples was found to be robust, as indicated by unconstrained principal coordinate analysis (PCoA) performed independently for each soil batch (*SI Appendix*, Fig. S3). Furthermore, CAP performed on the entire dataset revealed a prominent effect of the host genotype on bacterial communities, explaining 9.82% of the variance (Fig. 2*B*).

Nodules are root-derived and -anchored structures, and yet the two organs host distinctive bacterial assemblages (Fig. 2*A*). As a consequence, despite rigorous preparation of root compartments, WT root segments might contain incipient root-concealed nodule primordia and, vice versa, the nodule samples might be contaminated with surrounding root tissue. To clarify whether these potential limitations of our sampling protocol confound the observed host genotype-dependent community differentiation, we performed an in silico depletion of all nodule-enriched OTUs from the WT root dataset and repeated the PCoA and CAP (*SI Appendix*, Fig. S4). This experiment revealed only a negligible reduction in the portion of the community variance explained by the host genotype (9.82% versus 9.72%), indicating that the differences in the root-associated assemblages caused by the impairment of nitrogen-fixing symbiosis are largely robust against residual cross-contamination between the two compartments.

To understand better how the *Lotus* nodulation pathway influences bacterial community composition, we identified OTUs that are specifically enriched in the root and rhizosphere of WT or

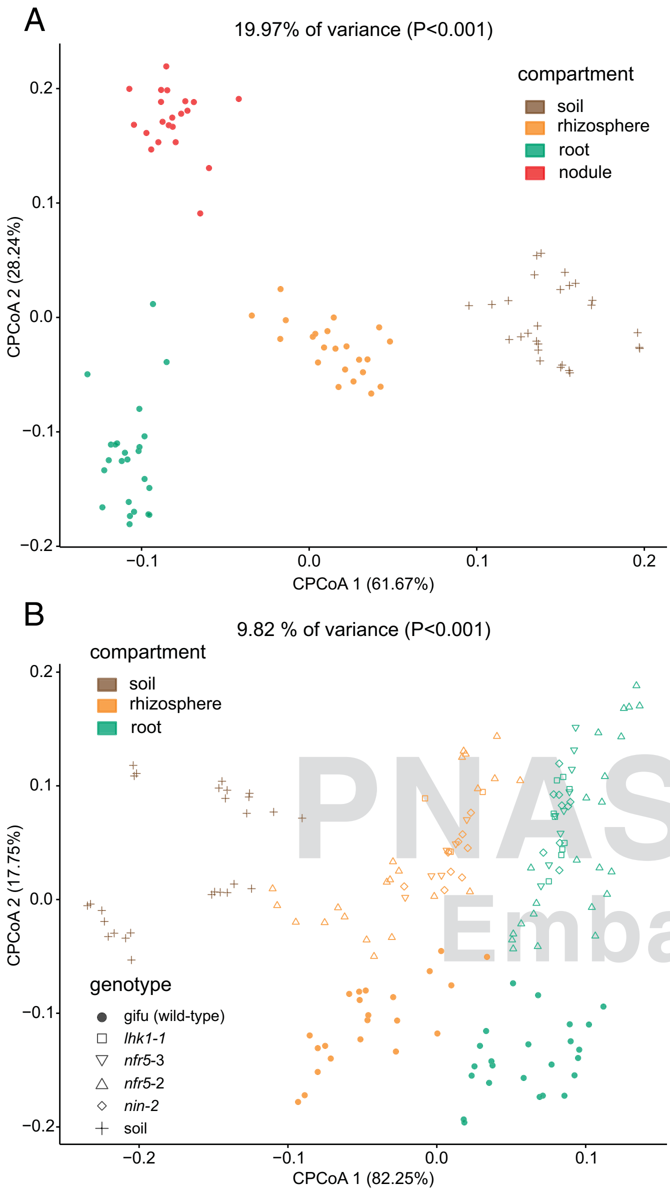


Fig. 2. (*A*) Constrained PCoA plot of Bray–Curtis distances between samples including only the WT constrained by compartment (19.97% of variance, $P >$ 0.001; n = 94). (*B*) Constrained PCoA plot of Bray–Curtis distances constrained by genotype (9.82% of variance explained, $P < 0.001$; n = 164). Each point corresponds to a different sample colored by compartment, and each host genotype is represented by a different shape. The percentage of variation indicated in each axis corresponds to the fraction of the total variance explained by the projection. Corresponding unconstrained PCoA plots for each soil batch are shown in *SI Appendix*, Fig. S3.

mutants compared with unplanted soil (*Materials and Methods*). Due to the fact that the bacterial assemblages of the tested symbiosis mutants do not significantly differ among each other (Fig. 2*B*), we performed our analyses using the combined samples from all mutant genotypes across all soil batches (Fig. 4). The *Lotus* WT root microbiota is characterized by a large number of root-enriched OTUs, mostly belonging to Proteobacteria, Actinobacteria, and Bacteroidetes (105 green circles in Fig. 4*A* and *SI Appendix*, Fig. S12). By contrast, only a small number of OTUs were found specifically enriched in the WT rhizosphere samples (8 orange circles in Fig. 4*A*). Compared with WT, roots of the symbiotic mutants are dramatically depleted of root-enriched OTUs (28 green circles in Fig. 4*B*), whereas the number of rhizosphere-enriched OTUs increased by a factor of 8 (68 orange circles in Fig. 4*B*). This pattern was reproducible when we performed the same analysis for each soil batch and mutant genotype independently (*SI Appendix*, Figs. S5–S9).

To characterize the bacterial community shifts further, we calculated separately for WT and symbiotic mutant plants aggregated RAs of OTUs that are specifically enriched in one compartment. As expected, this calculation revealed a decreasing contribution of the soil-enriched OTUs in soil, rhizosphere, and root samples (69.40%, 17.03%, and 2.40% mean aggregated RA, respectively; dark brown box plots in Fig. 4 *C* and *D*) in both WT and mutant samples. This finding suggests that impairment of the symbiosis pathway does not affect the capacity of *Lotus* to exclude colonization by the majority of the detectable bacterial soil biome and the formation of characteristic root-associated microbiota, fully differentiated from the root-associated microbiota present in bulk soil. We observed an inverse pattern for the root-enriched OTUs across the three WT compartments (green box plots in Fig. 4*C*). The steep increase in the aggregated RAs, from 8.76% in the soil, to 35.72% in rhizosphere, and to 72.34% in roots for WT samples, was almost completely abolished in the mutants (1.49%, 3.63%, and 17.74%, respectively; green box plots in Fig. 4*D*). Conversely, the aggregated RAs of rhizosphere-specific OTUs are only slightly higher in the rhizosphere samples of WT plants compared with roots and soil, which are influenced by the low number of rhizosphere-specific OTUs (orange box plots in Fig. 4*C*). However, RAs are significantly higher in the mutant rhizosphere samples with respect to the other compartments (3.29% in soil samples, 22.09% in rhizosphere samples, and 9.94% in root samples; orange box plots in Fig. 4*D*). Taken together, these data support the hypothesis that the *Lotus* symbiosis pathway is a key component for the progressive enrichment/selection of specific soil-derived OTUs and the establishment of fully differentiated microbiota in rhizosphere and root compartments.

The Symbiosis Pathway Drives Root and Rhizosphere Differentiation Across Multiple Bacterial Orders. We dissected the observed bacterial community shifts by arranging OTUs according to their taxonomy and displaying their enrichment in the root or rhizosphere of WT and symbiotic mutants in a set of Manhattan plots (*Materials and Methods*). The results revealed unexpectedly nuanced taxonomic alterations underlying the community shifts in the plant-associated compartments (Fig. 5 *A* and *B*). Whereas WT plants host root-enriched OTUs belonging to a wide range of bacterial orders, mutants roots fail to enrich any member of the orders Flavobacteriales, Myxococcales, Pseudomonadales, Rhizobiales, and Sphingomonadales above a threshold of significance (false discovery rate-corrected *P* values, $\alpha = 0.05$; *Materials and Methods*). In addition, a striking enrichment of more than 15 Burkholderiales OTUs in WT roots contrasts with a marginal enrichment of this order in the symbiotic mutants. However, the mutant roots retain the capacity to enrich OTUs belonging to the orders Actinobacteridae, Rhodospirillales, Sphingobacteriales, and Xanthomonadales (Fig. 5 *A* and *B*). Strikingly, we found an almost inverse pattern when we considered the rhizosphere-enriched OTUs in WT and mutant plants: Both the number and the taxonomic diversity of significantly enriched OTUs increased dramatically in the mutants compared with WT (Fig. 5 *C* and *D*).

Next, we compared directly the WT and mutants to identify OTUs differentially abundant in the root or rhizosphere (*SI Appendix*, Figs. S10 and S11). We found that the community shift that separates host genotypes (Fig. 2*B*) is largely caused by numerous OTUs that are specifically enriched (n = 45) or depleted (n = 15) in WT roots with respect to mutant root samples (*SI Appendix*, Fig. S10*A*) belonging to at least 14 bacterial orders (*SI Appendix*, Fig. S11 *A* and *B*). We observed a parallel effect on OTUs of a similar taxonomic profile when comparing rhizosphere samples across genotypes, and identified numerous OTUs enriched (n = 27) or

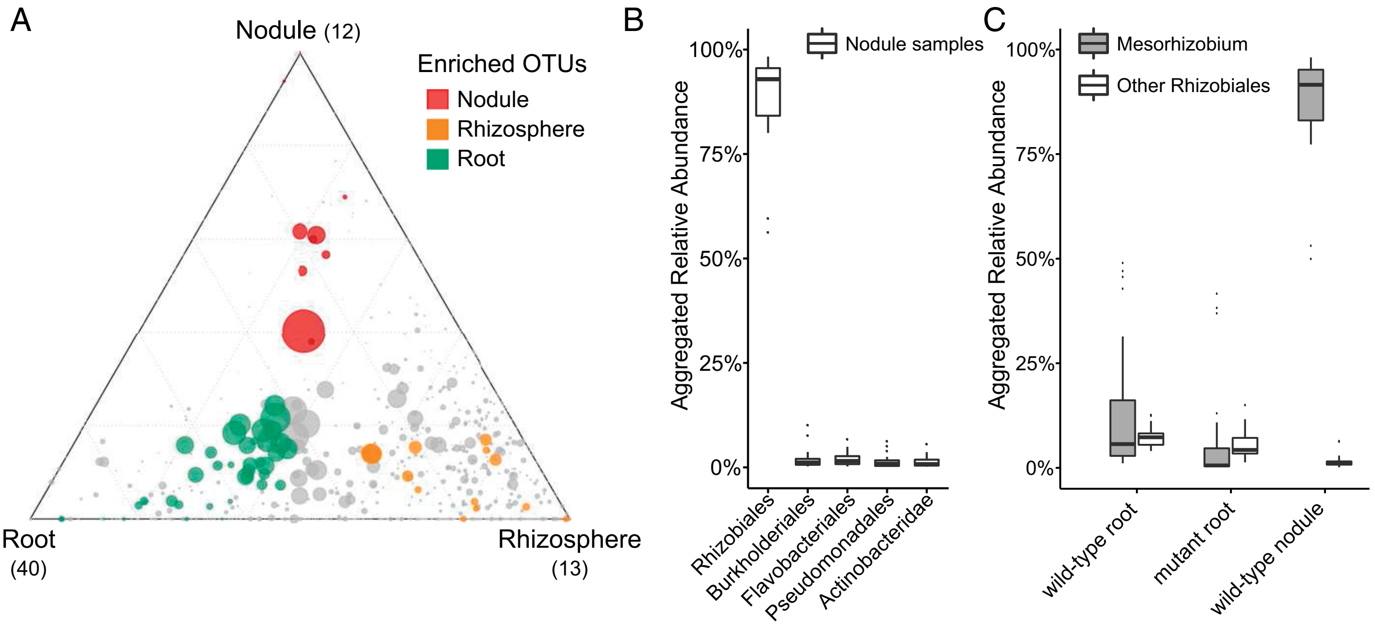


Fig. 3. (*A*) Ternary plot depicting compartment RAs of all OTUs (>5 ‰) for WT root, rhizosphere, and nodule samples ($n = 67$) across three soil batches (CAS8–CAS10). CAS, Cologne agricultural soil. Each point corresponds to an OTU. Its position represents its RA with respect to each compartment, and its size represents the average across all three compartments. Colored circles represent OTUs enriched in one compartment compared with the others (green in root, orange in rhizosphere, and red in nodule samples), whereas gray circles represent OTUs that are not significantly enriched in a specific compartment. (*B*) Rank abundance plot depicting RAs aggregated to the order taxonomic level for the top abundant taxa found in the WT nodule samples ($n = 21$). (*C*) Comparison of abundances between *Mesorhizobium* and other Rhizobiales genera in WT roots ($n = 48$), mutant roots ($n = 100$), and WT nodule samples ($n = 21$).

depleted ($n = 6$) in the WT rhizosphere (*SI Appendix*, Figs. S10*B* and S11 *C* and *D* and Dataset S2).

Next, we examined whether the complex community shifts observed at the order taxonomic rank were also detectable at the higher phylum rank. Interestingly, we observed clear differences between the root and rhizosphere samples of WT and mutant *Lotus* plants for Proteobacteria (66.78% and 46.97%, respectively) and Bacteroidetes (14.57% and 38.10%, respectively), which were largely explained by variation in abundances of OTUs belonging to the Rhizobiales, Burkholderiales, and Flavobacteriales bacterial orders (*SI Appendix*, Figs. S10–S12). Conversely, we found no significant differences between WT and mutant roots and rhizospheres for Actinobacteria (13.42% and 11.62% average RA, respectively) and Firmicutes (~1% average RA). These results illustrate that the large shifts observed between the WT and mutant *Lotus* plants affect root and rhizosphere communities similarly, even at higher taxonomic levels (*SI Appendix*, Figs. S10–S12).

Comparable Immune- and Symbiosis-Related Metabolic Responses in Soil-Grown WT and Symbiotic Mutant Roots. The extensive changes of root microbiota structure across multiple bacterial orders in the symbiosis mutants prompted us to investigate whether mutant roots display altered immune- or symbiosis-related metabolic responses that indirectly perturb an orderly microbiota establishment. We quantified relative transcript levels for a panel of defense and symbiotic marker genes using WT and mutant root tissue samples that were processed as for the 16S rRNA gene community profiling (*SI Appendix* and Dataset S3). Analysis of eight genes induced during pathogen defense in *Lotus* or likely representing *Lotus* orthologs of *Arabidopsis* defense marker genes revealed that WT and mutant roots accumulate similar transcript levels, indicative of a comparable immune status rather than an induced defense in the mutants (*SI Appendix*, Fig. S13*A*). We also tested whether WT and mutant roots differed in expression levels of genes that have been reported to contribute to the metabolic state established between host and nitrogen-fixing symbiont (37–40). We found comparable transcript levels of *Nodulin26*, *Nodulin70*, *Sucrose transporter4*, and *Invertase1* in the tested genotypes, suggesting similar metabolic responses in the roots of the WT and mutants (*SI Appendix*, Fig. S13*B*). On the other hand, early symbiotic genes like *Nin*, *Peroxidase*, and *Thaumatin* were induced in WT or *lhk1* and *nin*-2 mutants, but not in *nfr5*-2 roots, indicating that soil-grown symbiotic mutants maintain their previously described, gradually impaired root response to nitrogen-fixing rhizobia (16, 18, 22, 36) (*SI Appendix*, Fig. S13*C*). Direct measurements of total protein content revealed comparable levels in WT and symbiotic mutants (*SI Appendix*, Fig. S14), whereas quantification of nitrate levels revealed significant differences between *nfr5*, *nin*, and *lhk1* or WT, indicating that regulation of nitrate uptake, which has a known inhibitory effect on nodulation (41, 42), operates downstream of *Nin* (*SI Appendix*, Fig. S14). Together, these results suggest that a nitrogen-sufficient status is reached in all tested genotypes, but that the nitrogen source, N_2 or nitrate, might differ among them.

***L. japonicus* and Various Brassicaceae Species Assemble Highly Diverged Root-Inhabiting Bacterial Communities.** We have previously shown that *Arabidopsis thaliana* and three other Brassicaceae species (*Cardamine hirsuta*, *Arabidopsis halleri*, and *Arabidopsis lyrata*), grown in Cologne soil, assemble a highly similar root microbiota, characterized only by small quantitative differences of community profiles (29). We retrieved the corresponding raw 16S sequence reads and performed de novo OTU clustering together with the amplicon data of WT and symbiotic *Lotus* mutants (Fig. 6). PCoA of Bray–Curtis distances revealed a clear separation of root and soil samples, but also a marked distinction between all *Lotus* and Brassicaceae samples (Fig. 6*A*), indicating that both WT and symbiosis-impaired *Lotus* plants harbor strikingly distinctive root microbiota compared with root microbiota of the four tested Brassicaceae species. Quantitative analysis of WT *L. japonicus* and *A. thaliana* root-enriched OTUs revealed significant and contrasting differences already at the phylum level primarily reflected in the abundances of Proteobacteria and Actinobacteria (Fig. 6*B*). Similar rank abundance analysis performed at the order level identified particular taxonomic lineages contributing to the differences

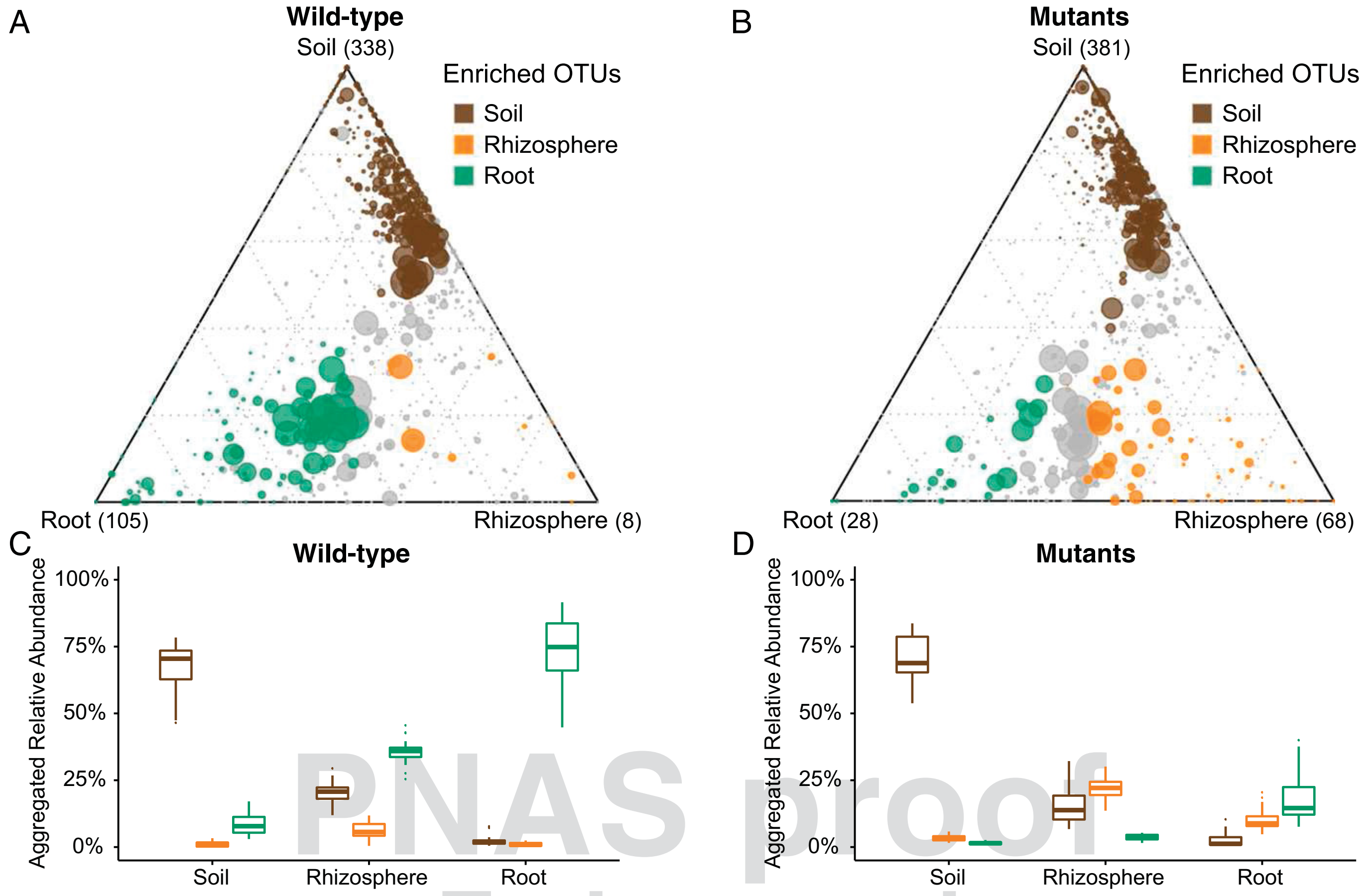


Fig. 4. Ternary plots depicting compartment RA of all OTUs (>5 ‰) for WT samples (*A*; WT; *n* = 73) and mutant samples (*B*; *nfr5*-2, *nfr5*-3, *nin*-2, and *lhk1*-1; *n* = 118) across three soil batches (CAS8–CAS10). Each point corresponds to an OTU. Its position represents its RA with respect to each compartment, and its size represents the average across all three compartments. Colored circles represent OTUs enriched in one compartment compared with the others (green in root, orange in rhizosphere, and brown in root samples). Aggregated RAs of each group of enriched OTUs (root-, rhizosphere- and soil-enriched OTUs) in each compartment for the WT samples (*C*; WT; *n* = 73) and mutant samples (*D*; *nfr5*-2, *nfr5*-3, *nin*-2, *lhk1*-1; *n* = 118) are shown. In each compartment, the difference from 100% RA is explained by OTUs that are not significantly enriched in a specific compartment.

between these two plant species; Rhizobiales, Caulobacterales, Rickettsiales, and Sphingobacteriales were found in larger abundances in *Lotus* roots, whereas Burkholderiales, Actinomycetales, Myxococcales, and Pseudomonadales were more abundant in *Arabidopsis* roots (Fig. 6*C*).

Symbiosis-Impaired Mutants Maintain an Altered Community Structure in Nitrogen-Supplemented Soil. Nitrogen-fixing symbiosis is nitrogen-sensitive, and already from 2 mM KNO_3 concentration, reduced nodulation and infection were observed in *Lotus* (42). To determine if the community shifts observed in *Lotus* mutant roots and the rhizosphere were caused by a potentially differential nitrogen status, we performed a similar community analysis using plants grown in Cologne soil (different soil batch) supplemented with 10 mM KNO_3. In these conditions, the symbiotic mutants no long had a pale leaf phenotype as observed in nonsupplemented soil (*SI Appendix*, Fig. S15) and the WT plants developed no functional nodules, based on their low number and small and white appearance at the time of harvest. Despite similar nitrogen content in WT and mutant roots (*SI Appendix*, Fig. S15), we found that the differential phenotypes (stature and fresh weight) seen in nonsupplemented soil were retained under nitrogen-supplemented conditions (*SI Appendix*, Fig. S15), indicating that an intact symbiosis pathway promotes plant growth irrespective of the presence of functional nodules. Based on the similar macroscopic phenotypes of the symbiotic mutants in response to the nitrogen supplementation, we then analyzed the composition of bacterial communities in WT and two Nod factor receptor mutants, *nfr5*-2 and *nfr5*-3. Remarkably, the PCoA revealed a similar shift in the root and rhizosphere communities of the mutants relative to the corresponding WT compartments for plants grown in nitrogen-supplemented soil (21.20% of the variance, $P < 0.001$; *SI Appendix*, Fig. S16*A*) as in the nonsupplemented Cologne soil (21.80% of the variance, $P < 0.001$; *SI Appendix*, Fig. S16*B*). Finally, no detectable differences in soil biome composition were seen between nitrogen-supplemented and nonsupplemented unplanted soil samples (Dataset S4). Together, these results provide evidence for a direct impact of the disabled symbiosis pathway on the root-associated community structure rather than an indirect effect resulting from abolished symbiotic nitrogen fixation.

Discussion

Here, we have characterized the root microbiota of the model legume *L. japonicus* using a 16S rRNA amplicon survey. By using a panel of symbiosis-impaired mutants, we have investigated the role of host genes with known functions in the establishment of a highly specific and binary symbiotic plant–microbe association in the context of the root-associated bacterial community. Our study reveals that key symbiotic genes play a major role in the establishment of taxonomically structured bacterial communities in the root and rhizosphere of *L. japonicus* (Fig. 4), which extends their role beyond the perception and selection of nitrogen-fixing rhizobia for

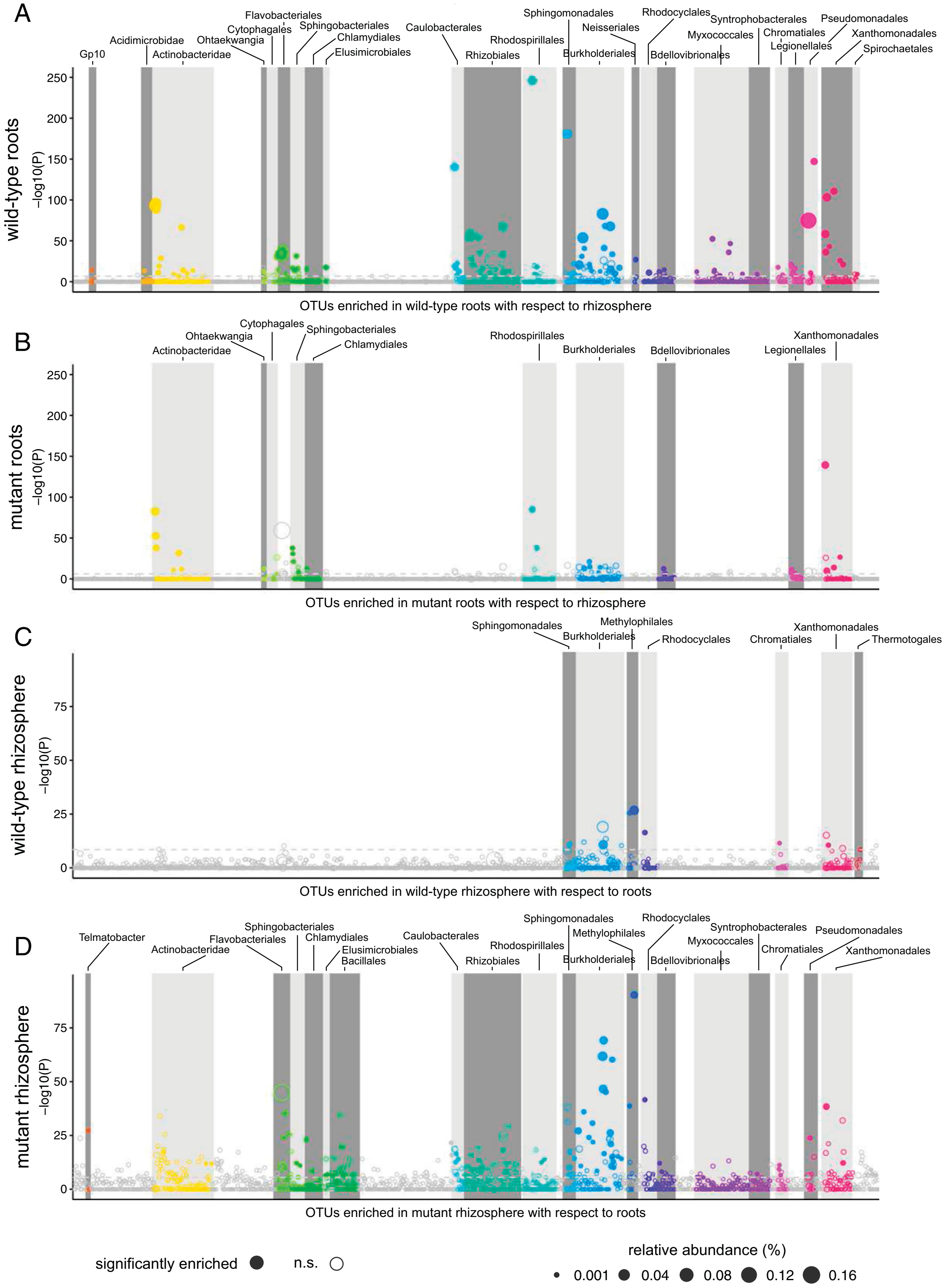


Fig. 5. Manhattan plots showing root-enriched OTUs in WT (*A*) or in the mutants (*B*) with respect to rhizosphere and rhizosphere-enriched OTUs in WT (*C*) or in the mutants (*D*) with respect to root. OTUs that are significantly enriched (also with respect to soil) are depicted as full circles. The dashed line corresponds to the false discovery rate-corrected *P* value threshold of significance (α = 0.05). The color of each dot represents the different taxonomic affiliation of the OTUs (order level), and the size corresponds to their RAs in the respective samples [WT root samples (*A*), mutant root samples (*B*), WT rhizosphere samples (*C*), and mutant rhizosphere samples (*D*)]. Gray boxes are used to denote the different taxonomic groups (order level).

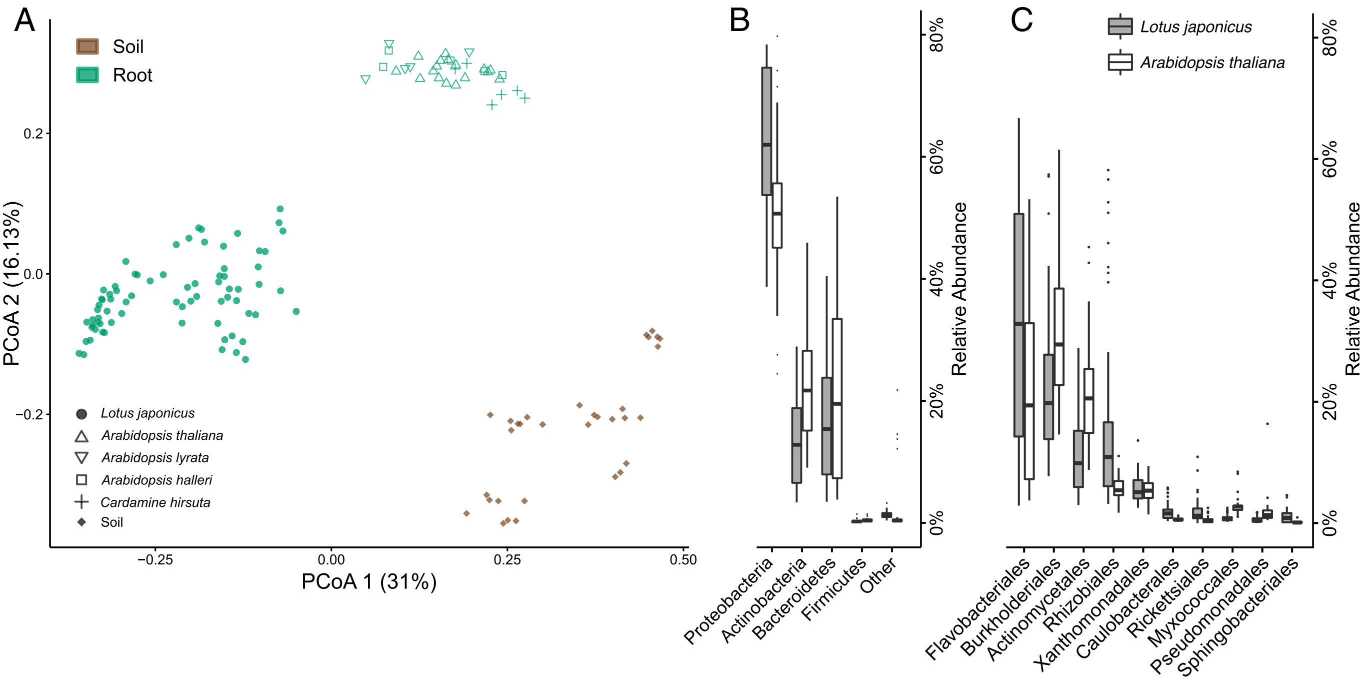


Fig. 6. (*A*) PCoA plot of Bray–Curtis distances for root and soil samples showing a clear separation between the roots of all *Lotus* genotypes (circles) compared with the roots of *Arabidopsis* and relative species (hollow shapes) grown in Cologne soil and sequenced using the same primer set. RAs aggregated to the phylum (*B*) and order taxonomic levels (*C*, 10 most abundant orders) showing a comparison between *A. thaliana* (*n* = 26) and *L. japonicus* root samples (*n* = 74).

intracellular accommodation in nodules. The observed impact of disenabling symbiosis in *Lotus* seemingly differs from analogous perturbations in insects and mammals, where impairment of symbiosis in the gut leads to the accumulation of pathogens and dramatic consequences for the health of the host (43–46). Cologne soil-grown *Lotus* symbiotic mutants showed no signs of disease or altered immune response in comparison to WT (Fig. 1 and *SI Appendix*, Fig. S13). However, it remains possible that the observed reduced plant growth (Fig. 1) leads to lower reproductive fitness of the mutant plants in a natural ecosystem.

Genetic disruption of the nodulation pathway resulted in depletion of six bacterial orders from the root compartment, including the two most abundant orders identified in WT, Flavobacteriales and Burkholderiales (Fig. 5 *A* and *B*). Rhizobium therefore acts as a bacterial "hub" for *Lotus* roots, in analogy to the *Albugo* oomycete pathogen that largely affected the phyllosphere microbiome of *Arabidopsis* after infection (47). Two mutually nonexclusive scenarios could account for the observed dramatic community shift inside roots: (*i*) cooperative microbe–microbe interactions between symbiotic nitrogen-fixing bacteria and a subset of root microbiota members or (*ii*) direct use of the nodulation pathway by soil-borne bacteria other than nodulating rhizobia for entry and proliferation inside legume roots. Quantification of selected transcripts in WT and mutant roots provided evidence for a similar immune status and symbiosis-specific metabolic responses, but, as expected, differential expression of the symbiotic genes (*SI Appendix*, Fig. S13). These results support the hypothesis of a direct rather than indirect engagement of the nitrogen-fixing symbiosis pathway in the selection of specific bacterial taxa other than nodulating rhizobia.

The beneficial association with Burkholderiales is habitual for legumes. For example, *Mimosa* genera within South America and legumes from the Papilionoideae subfamily in South Africa are known to form an ancient and stable symbiosis with nitrogen-fixing Burkholderia lineages inside nodules (48–50). Members of the order Burkholderiales, which are dramatically depleted in the *Lotus* mutants, are also known for potent plant growth promotion activities in nonleguminous plants (51). The inclusion of legumes in a cropping rotation sequence in agriculture generally enriches the soil, but the symbiotic nitrogen fixation alone cannot explain the productivity increase in the subsequent crop (26). Thus, our study raises the intriguing possibility that besides the activity of nitrogen-fixing bacteria, the selective enrichment of other symbiosis-linked root microbiota members influences the soil biome and, consequently, soil biofertilization by legumes might involve a much wider taxonomic range than currently thought.

Our statistical analyses revealed a major impact of the host genotype on the *Lotus* root microbiota, which explains 9.82% of the variance in the data (Fig. 2*B*). This genotype effect is almost twofold larger than the effect identified when comparing the root-associated communities of WT and severely immunocompromised *A. thaliana* mutants (52) or between *A. thaliana* and three Brassicaceae species that diverged up to 35 Mya (29). Interestingly, the community shifts detected in *nfr5*, *nin*, and *lhk1* mutant roots are highly comparable (Fig. 2*B*). This similarity is likely related to our stringent root fractionation protocol (10, 11), depleting both epiphytes and epidermal endophytes. However, we cannot rule out differences between the communities of the tested mutants of a magnitude smaller than the effect caused by soil batch variation and technical noise present in the pyrosequencing data.

Recent studies have shown that upon microbiota acquisition, the root-associated bacterial assemblage remains robust against major changes in plant stature and source-sink relationships (9, 11). This microbiota stability is also observed here, where a clear separation between the root and rhizosphere of WT and symbiosis-impaired *Lotus* plants is retained when plants are grown under different nitrate conditions (*SI Appendix*, Fig. S16). The latter observation is also strong evidence that the root-associated community shift in the symbiotic mutants is a direct consequence of the disabled symbiosis pathway rather than an indirect effect resulting from abolished symbiotic nitrogen fixation.

Previous controlled coinoculation experiments with *Mesorhizobium loti* and root endophytes have shown that *Lotus* can selectively guide endophytic bacteria toward nodule primordia via symbiont-induced infection threads, and that endophytes and symbionts can promote each other's infection of the host (35). These experiments focused on nodule colonization and were conducted using a gnotobiotic plant system with a limited number of endophytes. Our bacterial community profiling data obtained with soil-grown symbiotic mutants allowed testing the contribution of infection thread-dependent root colonization by natural populations of compatible endophytes present in soil. Indeed, the large number of bacterial taxa found depleted in *nfr5*, *nin*, and *lhk1* genotypes (*SI Appendix*, Fig. S11) suggests that infection threads arrested in WT may facilitate root colonization by endophytes. This finding implies additional functions of host genes active in early symbiosis for efficient root colonization by a subset of the root microbiota. Our community profiling data also identified bacterial orders enriched in the mutant roots; thus, their root colonization takes place independent of nitrogen-fixing symbiosis. These endophytic taxa may use alternative entry routes [i.e., crack entry, which occurs at the base of emerging lateral roots and is likely used as an entry portal for root endophytes in nonleguminous plants (53–55)].

Our study revealed that *Lotus* growth in Cologne soil did not interfere with the host–symbiont compatibility described previously in monoassociations with plants grown in artificial media (56). This high selectivity is evidenced by only 12 nodule-enriched OTUs among a total of 1,834 OTUs in our dataset and by *Mesorhizobium* members representing the most abundant taxa (Fig. 3*A*). Remarkably, nodule-enriched OTUs had a similar RA in the root and rhizosphere compartments, suggesting linked selection process(es) in all three compartments. Furthermore, nodule-enriched OTUs were depleted from the mutant root samples, which corroborates our hypothesis that an intact symbiotic pathway is selecting rhizobia during infection in both the root cortex and nodules. Interestingly, nodule-enriched *Mesorhizobium* OTUs were depleted from both mutant root and rhizosphere compartments, indicating that root-derived diffusible compounds produced by WT exert a role in enriching symbionts in soil that adheres to the legume root surface. Thus, our findings support previous observations that maintenance of highly symbiotic isolates in the soil is not only a function of rhizobia release from decaying nodules but is also dependent on persistent host selective pressure (57, 58). A likely scenario for this enrichment is a positive feedback mechanism in which host-initiated signaling leads to enrichment of symbionts in the root and rhizosphere, a hypothesis that is supported by the markedly similar patterns observed in plants grown under nitrogen-supplemented conditions that impede nitrogen fixation. Legume root-derived flavonoids are candidate diffusible signaling molecules in such a feedback mechanism because their profile was shown to change during root–nodule symbiosis (59, 60) and their broad impact on soil bacterial communities (61, 62), especially on symbiotic rhizobia, has been documented (63, 64).

Our comparative analyses of the root microbiota from *L. japonicus* and four Brassicaceae species grown in Cologne soil (29) revealed a highly distinctive community composition irrespective of an intact or dysfunctional symbiosis pathway (Fig. 6). *Lotus* and *Arabidopsis* ancestors diverged ∼118 Mya (65), and two major evolutionary events took place soon after their separation: loss of arbuscular mycorrhiza (AM) symbiosis in the Brassicaceae (66) and gain of nitrogen-fixation predisposition in the legume predecessor (67). Thus, our findings suggest that the marked distinctiveness of the *Lotus*-specific root microbiota is not governed by the evolved functions of *Nfr5*, *Nin*, or *Lhk1*, but is possibly linked to the loss of AM symbiosis in the Brassicaceae lineage. This hypothesis can be tested by future experiments with mutants affecting *Lotus* "common symbiosis genes" that fail to establish both symbiotic relationships with AM fungi and nodulating rhizobia (68–70).

Materials and Methods

A detailed description of the methods used in this study can be found in *SI Appendix, Supplementary Materials and Methods*.

Soil and Plant Material. Seeds of *L. japonicus* WT, ecotype Gifu B-129, and the corresponding symbiosis-deficient mutants (*nfr5*-2, *nfr5*-3, *lhk1*-1, and *nin*-2) were grown in soil batches collected in the successive seasons. Plants were grown in the greenhouse under long-day conditions (16-h photoperiod), watered with tap water (optionally supplemented with 10 mM KNO_3), and harvested after 10 wk.

Sample and 16S Library Preparations. Fragments of the root systems were washed, and the rhizosphere, root, and nodules were separated. A first wash containing the root-adhering soil layer defined the rhizosphere compartment. Nodules and visible primordia were separated from root fragments of nodulating genotypes (WT and *lhk1*-1) with a scalpel. Root and nodule samples were ultrasound-treated. DNA extraction was performed using a FastDNA SPIN Kit for Soil (MP Biomedicals). Barcoded primers targeting the variable V5–V7 regions of bacterial 16S rRNA genes (799F and 1193R) (10, 71) were used for amplification. Amplicons were purified (Qiagen), combined, and subjected to 454 sequencing.

Metabolite Analysis. Root nitrate contents were determined by the ion chromatography method as previously described (72). Proteins were extracted in 10 mM Tris·HCl buffer (pH 8) and determined with a Bio-Rad Protein Assay Kit using BSA as the standard.

Computational Analyses. The 16S rRNA gene sequences were processed using a combination of custom scripts as well as tools from the QIIME (73) and USEARCH (74) pipelines (QIIIME-ready mapping files are provided in Datasets S4 and S5). The resulting OTU table was used in all subsequent statistical analyses of differentially abundant taxa as well as analyses of α- and β-diversity. Indices of α-diversity were calculated after subsampling to an even depth of 1,000 reads. Measures of β-diversity were calculated on a normalized OTU table. The PCoA was done by classical multidimensional scaling of β-diversity distance matrices using the *cmdscale* function in R. CAP (30) was computed using the *capscale* function implemented in the vegan R library (75), by constraining for the variable of interest and conditioning for the remaining factors. Statistical analyses of differentially abundant OTUs were performed using the edgeR library (76) by fitting a negative binomial generalized linear model to the OTUs.

Code Availability. All scripts required for the computational analyses performed in this study as well as the corresponding raw sequencing and intermediate data are available at www.mpipz.mpg.de/R_scripts.

ACKNOWLEDGMENTS. We thank Diana Kuehn and Bruno Huettel (Max Planck Genome Centre Cologne) for the preparation and sequencing of the 454 libraries and Irene Klinkhammer (Botanical Institute Cologne) for assistance with nitrate analysis. We thank Omri Finkel, Jeffery L. Dangl, Jens Stougaard, Lene H. Madsen, Stig U. Andersen, Ryohei Nakano, Stephane Hacquard, and Thorsten Thiergart for valuable comments on the manuscript. This work was supported by Danish Council for Independent Research Grant 09-070023 and Danish National Research Foundation Grant DNRF79 (to S.R.) and by the Max Planck Society, European Research Council Grant ROOTMICROBIOTA, and the "Cluster of Excellence on Plant Sciences" program funded by the Deutsche Forschungsgemeinschaft (P.S.-L.).

1. Maathuis FJ (2009) Physiological functions of mineral macronutrients. *Curr Opin Plant Biol* 12(3):250–258.
2. Fierer N, Jackson RB (2006) The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* 103(3):626–631.
3. Bulgarelli D, Schlaeppi K, Spaepen S, Ver Loren van Themaat E, Schulze-Lefert P (2013) Structure and functions of the bacterial microbiota of plants. *Annu Rev Plant Biol* 64:807–838.
4. Parniske M (2008) Arbuscular mycorrhiza: The mother of plant root endosymbioses. *Nat Rev Microbiol* 6(10):763–775.
5. Vorholt JA (2012) Microbial life in the phyllosphere. *Nat Rev Microbiol* 10(12): 828–840.
6. Hacquard S, et al. (2015) Microbiota and host nutrition across plant and animal kingdoms. *Cell Host Microbe* 17(5):603–616.
7. Guttman DS, McHardy AC, Schulze-Lefert P (2014) Microbial genome-enabled insights into plant-microorganism interactions. *Nat Rev Genet* 15(12):797–813.
8. Edwards J, et al. (2015) Structure, variation, and assembly of the root-associated microbiomes of rice. *Proc Natl Acad Sci USA* 112(8):E911–E920.
9. Lundberg DS, et al. (2012) Defining the core Arabidopsis thaliana root microbiome. *Nature* 488(7409):86–90.
10. Bulgarelli D, et al. (2012) Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota. *Nature* 488(7409):91–95.

11. Dombrowski N, et al. (August 2, 2016) Root microbiota dynamics of perennial Arabis alpina are dependent on soil residence time but independent of flowering time. *ISME J*, 10.1038/ismej.2016.109.
12. Eilers KG, Lauber CL, Knight R, Fierer N (2010) Shifts in bacterial community structure associated with inputs of low molecular weight carbon compounds to soil. *Soil Biol Biochem* 42(6):896–903.
13. Nguyen C (2003) Rhizodeposition of organic C by plants: Mechanisms and controls. *Agronomie* 23(5-6):375–396.
14. Radutoiu S, et al. (2007) LysM domains mediate lipochitin-oligosaccharide recognition and Nfr genes extend the symbiotic host range. *EMBO J* 26(17):3923–3935.
15. Broghammer A, et al. (2012) Legume receptors perceive the rhizobial lipochitin oligosaccharide signal molecules by direct binding. *Proc Natl Acad Sci USA* 109(34): 13859–13864.
16. Schauser L, Roussis A, Stiller J, Stougaard J (1999) A plant regulator controlling development of symbiotic root nodules. *Nature* 402(6758):191–195.
17. Held M, et al. (2014) Lotus japonicus cytokinin receptors work partially redundantly to mediate nodule formation. *Plant Cell* 26(2):678–694.
18. Murray JD, et al. (2007) A cytokinin perception mutant colonized by Rhizobium in the absence of nodule organogenesis. *Science* 315(5808):101–104.
19. Udvardi M, Poole PS (2013) Transport and metabolism in legume-rhizobia symbioses. *Annu Rev Plant Biol* 64(64):781–805.
20. El Yahyaoui F, et al. (2004) Expression profiling in Medicago truncatula identifies more than 750 genes differentially expressed during nodulation, including many potential regulators of the symbiotic program. *Plant Physiol* 136(2):3159–3176.
21. Colebatch G, et al. (2004) Global changes in transcription orchestrate metabolic differentiation during symbiotic nitrogen fixation in L. japonicus. *Plant J* 39(4):487–512.
22. Høgslund N, et al. (2009) Dissection of symbiosis and organ development by integrated transcriptome analysis of lotus japonicus mutant and wild-type plants. *PLoS One* 4(8):e6556.
23. Nakagawa T, et al. (2011) From defense to symbiosis: Limited alterations in the kinase domain of LysM receptor-like kinases are crucial for evolution of legume-Rhizobium symbiosis. *Plant J* 65(2):169–180.
24. Krusell L, et al. (2002) Shoot control of root development and nodulation is mediated by a receptor-like kinase. *Nature* 420(6914):422–426.
25. Searle IR, et al. (2003) Long-distance signaling in nodulation directed by a CLAVATA1-like receptor kinase. *Science* 299(5603):109–112.
26. Peoples MB, et al. (2009) The contributions of nitrogen-fixing crop legumes to the productivity of agricultural systems. *Symbiosis* 48(1):1–17.
27. Batterman SA, Wurzburger N, Hedin LO (2013) Nitrogen and phosphorus interact to control tropical symbiotic N_2 fixation: A test in Inga punctata. *J Ecol* 101(6):1400–1408.
28. Adams MA, Turnbull TL, Sprent JI, Buchmann N (2016) Legumes are different: Leaf nitrogen, photosynthesis, and water use efficiency. *Proc Natl Acad Sci USA* 113(15): 4098–4103.
29. Schlaeppi K, Dombrowski N, Oter RG, Ver Loren van Themaat E, Schulze-Lefert P (2014) Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. *Proc Natl Acad Sci USA* 111(2):585–592.
30. Anderson MJ, Willis TJ (2003) Canonical analysis of principal coordinates: A useful method of constrained ordination for ecology. *Ecology* 84(2):511–525.
31. Turner TR, James EK, Poole PS (2013) The plant microbiome. *Genome Biol* 14(6):209.
32. Mendes LW, Kuramae EE, Navarrete AA, van Veen JA, Tsai SM (2014) Taxonomical and functional microbial community selection in soybean rhizosphere. *ISME J* 8(8): 1577–1587.
33. Chen M, et al. (2014) Dynamic succession of soil bacterial community during continuous cropping of peanut (Arachis hypogaea L.). *PLoS One* 9(7):e101355.
34. Suzaki T, Yoro E, Kawaguchi M (2015) Leguminous plants: Inventors of root nodules to accommodate symbiotic bacteria. *Int Rev Cell Mol Biol* 316:111–158.
35. Zgadzaj R, et al. (2015) A legume genetic framework controls infection of nodules by symbiotic and endophytic bacteria. *PLoS Genet* 11(6):e1005280.
36. Madsen EB, et al. (2003) A receptor kinase gene of the LysM type is involved in legume perception of rhizobial signals. *Nature* 425(6958):637–640.
37. Szczyglowski K, Kapranov P, Hamburger D, de Bruijn FJ (1998) The Lotus japonicus LjNOD70 nodulin gene encodes a protein with similarities to transporters. *Plant Mol Biol* 37(4):651–661.
38. Hwang JH, Ellingson SR, Roberts DM (2010) Ammonia permeability of the soybean nodulin 26 channel. *FEBS Lett* 584(20):4339–4343.
39. Flemetakis E, et al. (2003) A sucrose transporter, LjSUT4, is up-regulated during Lotus japonicus nodule development. *J Exp Bot* 54(388):1789–1791.
40. Welham T, et al. (2009) A cytosolic invertase is required for normal growth and cell development in the model legume, Lotus japonicus. *J Exp Bot* 60(12):3353–3365.
41. Carroll BJ, Gresshoff PM (1983) Nitrate inhibition of nodulation and nitrogen-fixation in white clover. *Z Pflanzenphysiol* 110(1):77–88.
42. Reid DE, Heckmann AB, Novák O, Kelly S, Stougaard J (2016) Cytokinin oxidase/dehydrogenase3 maintains cytokinin homeostasis during root and nodule development in L. japonicus. *Plant Physiol* 170(2):1060–1074.
43. Brandl K, et al. (2008) Vancomycin-resistant enterococci exploit antibiotic-induced innate immune deficits. *Nature* 455(7214):804–807.
44. Dessein R, et al. (2009) Toll-like receptor 2 is critical for induction of Reg3 beta expression and intestinal clearance of Yersinia pseudotuberculosis. *Gut* 58(6):771–776.
45. Round JL, Mazmanian SK (2009) The gut microbiota shapes intestinal immune responses during health and disease. *Nat Rev Immunol* 9(5):313–323.
46. Johnston PR, Rolff J (2015) Host and symbiont jointly control gut microbiota during complete metamorphosis. *PLoS Pathog* 11(11):e1005246.
47. Agler MT, et al. (2016) Microbial hub taxa link host and abiotic factors to plant microbiome variation. *PLoS Biol* 14(1):e1002352.
48. Bontemps C, et al. (2010) Burkholderia species are ancient symbionts of legumes. *Mol Ecol* 19(1):44–52.
49. Garau G, Yates RJ, Deiana P, Howieson JG (2009) Novel strains of nodulating Burkholderia have a role in nitrogen fixation with papilionoid herbaceous legumes adapted to acid, infertile soils. *Soil Biol Biochem* 41(1):125–134.
50. Angus AA, Hirsch AM (2010) Insights into the history of the legume-betaproteobacterial symbiosis. *Mol Ecol* 19(1):28–30.
51. Touceda-Gonzalez M, et al. (2015) Combined amendment of immobilizers and the plant growth-promoting strain Burkholderia phytofirmans PsJN favours plant growth and reduces heavy metal uptake. *Soil Biol Biochem* 91:140–150.
52. Lebeis SL, et al. (2015) PLANT MICROBIOME. Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* 349(6250):860–864.
53. Patriquin DG, Döbereiner J, Jain DK (1983) Sites and processes of association between diazotrophs and grasses. *Can J Microbiol* 29(8):900–915.
54. Egener T, Hurek T, Reinhold-Hurek B (1998) Use of green fluorescent protein to detect expression of nif genes of Azoarcus sp. BH72, a grass-associated diazotroph, on rice roots. *Mol Plant Microbe Interact* 11(1):71–75.
55. Chi F, et al. (2005) Ascending migration of endophytic rhizobia, from roots to leaves, inside rice plants and assessment of benefits to rice growth physiology. *Appl Environ Microbiol* 71(11):7271–7278.
56. Handberg K, Stougaard J (1992) L. japonicus, an autogamous, diploid legume species for classical and molecular genetics. *Plant J* 2(4):487–496.
57. Triplett EW, Albrecht KA, Oplinger ES (1993) Crop-rotation effects on populations of Bradyrhizobium japonicum and Rhizobium meliloti. *Soil Biol Biochem* 25(6):781–784.
58. Thies JE, Woomer PL, Singleton PW (1995) Enrichment of Bradyrhizobium spp populations in soil due to cropping of the homologous host legume. *Soil Biol Biochem* 27(4-5):633–636.
59. Zuanazzi JAS, et al. (1998) Production of Sinorhizobium meliloti nod gene activator and repressor flavonoids from Medicago sativa roots. *Mol Plant Microbe Interact* 11(8):784–794.
60. Rispail N, et al. (2010) Secondary metabolite profiling of the model legume L. japonicus during its symbiotic interaction with Mesorhizobium loti. *Symbiosis* 50(3): 119–128.
61. White LJ, Jothibasu K, Reese RN, Brözel VS, Subramanian S (2015) Spatio-temporal influence of isoflavonoids on bacterial diversity in the soybean rhizosphere. Molecular plant-microbe interactions. *Mol Plant Microbe Interact* 28(1):22–29.
62. Szoboszlay M, White-Monsant A, Moe LA (2016) The effect of root exudate 7,4′-dihydroxyflavone and naringenin on soil bacterial community structure. *PLoS One* 11(1):e0146555.
63. Hartwig UA, Joseph CM, Phillips DA (1991) Flavonoids released naturally from alfalfa seeds enhance growth rate of Rhizobium meliloti. *Plant Physiol* 95(3):797–803.
64. Dakora FD, Phillips DA (1996) Diverse functions of isoflavonoids in legumes transcend anti-microbial definitions of phytoalexins. *Physiol Mol Plant Pathol* 49(1):1–20.
65. Magallón S, Gómez-Acevedo S, Sánchez-Reyes LL, Hernández-Hernández T (2015) A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. *New Phytol* 207(2):437–453.
66. Delaux PM, et al. (2014) Comparative phylogenomics uncovers the impact of symbiotic associations on host genome evolution. *PLoS Genet* 10(7):e1004487.
67. Werner GD, Cornwell WK, Sprent JI, Kattge J, Kiers ET (2014) A single evolutionary innovation drives the deep evolution of symbiotic N_2-fixation in angiosperms. *Nat Commun* 5:4087.
68. Stracke S, et al. (2002) A plant receptor-like kinase required for both bacterial and fungal symbiosis. *Nature* 417(6892):959–962.
69. Kanamori N, et al. (2006) A nucleoporin is required for induction of Ca2+ spiking in legume nodule development and essential for rhizobial and fungal symbiosis. *Proc Natl Acad Sci USA* 103(2):359–364.
70. Charpentier M, et al. (2008) Lotus japonicus CASTOR and POLLUX are ion channels essential for perinuclear calcium spiking in legume root endosymbiosis. *Plant Cell* 20(12):3467–3479.
71. Chelius MK, Triplett EW (2001) The diversity of archaea and bacteria in association with the roots of Zea mays L. *Microb Ecol* 41(3):252–263.
72. Koprivova A, Harper AL, Trick M, Bancroft I, Kopriva S (2014) Dissection of the control of anion homeostasis by associative transcriptomics in Brassica napus. *Plant Physiol* 166(1):442–450.
73. Caporaso JG, et al. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* 7(5):335–336.
74. Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* 26(19):2460–2461.
75. Oksanen J, et al. (2015) R Package vegan: Community Ecology Package (R Foundation, Vienna), Version 2.3-3.
76. Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* 26(1): 139–140.