Computational analyses of the plant-associated microbiota

Inaugural Dissertation

zur

Erlangung des Doktorgrades

der Mathematisch-Naturwissenschaftlichen Fakultät

der Universität zu Köln

vorgelegt von

Pengfan Zhang

aus

Zhejiang, China

Köln, April 2023

Die in dieser Arbeit beschriebenen Arbeiten wurden unter der Leitung von Dr. Ruben Garrido-Oter am Max-Planck-Institut für Pflanzenzüchtungsforschung und Prof. Dr. Alga Zuccaro an der Universität zu Köln durchgeführt.

The work described in this thesis was conducted under the supervision of Dr. Ruben Garrido-Oter at the Max Planck Institute for Plant Breeding Research and Prof. Dr. Alga Zuccaro at the University of Cologne.

Berichterstatter:	Prof. Dr. Alga Zuccaro
	Prof. Dr. Eric Kemen
Prüfungsvorsitzender:	Prof. Dr. Berenike Maier
Tag der Disputation:	

Summary

Plants harbor phylogenetically diverse microbes on the exterior and interior of all organs and they form intimate relationships with the colonized microbiota. Multi-omics dramatically facilitates and expands our knowledge in plant-microbiota interactions and associations. To establish causalities, manipulation of microbiota populating plants under strictly controlled conditions is a necessity, which forged the development of reductionist approaches for studying plant-microbiota interactions, including the process of deconstruction and reconstruction of the plant microbiota. Deconstruction of the plant microbiota requires the establishment of genome-indexed microbial culture collections representing the plant microbiota of interests. The reconstruction step is to design synthetic microbial communities (SynComs) by mixing the strains from the culture collections and inoculate onto the plants.

In this dissertation, I introduced a software named Rbec that is developed to exclusively characterize the accurate microbial composition in SynComs subject to amplicon sequencing by both correcting PCR/sequencing errors and identifying maker gene paralogues within the same strain. Rbec also provides a novel feature for contamination identification in the SynCom experiments, which has been overlooked in previous studies but is a necessity to verify the robustness of the readouts from SynCom experimentations. Further, with the established pipelines for analyzing amplicon sequencing data from either natural or synthetic communities, I analyzed the microbial compositions from different studies including the study of the host preference of *Arabidopsis thaliana* and *Lotus Japonicus* commensals, the phycosphere microbiota, the effects of plant metabolites on soil microbiota and how bacterial antibiotics shape root microbiota.

Genome-indexed microbial culture collections allow us to study the functional capacities of microbiota. We systematically analyzed the biosynthetic gene clusters and the spread of antimicrobial 2,4-diacetylphloroglucinol synthetic gene clusters in *Pseudomonas* in established culture collections. Moreover, I studied the recent horizontal gene transfer (HGT) in bacteria from different culture collections assembled from different host plants and sites. This provides an atlas of the active taxa involved in HGT and the frequently transferred functional orthologues in plant-associated niches. In addition, it reveals the selection forces exerted on different taxa in the relevant environments.

In summary, our work tried to move the reductionist approaches forward in the aspect of computational analyses. We not only introduced a new computational method for accurately

profiling microbial compositions in SynComs, but also digged deeper into the genome-indexed culture collections by making full use of genome sequences. With the valuable integrated genome information of the plant microbiota, it'll provide the opportunity to study the functional diversities, evolutionary trajectories, genomic contents related to adaptations to hosts. However, with the increased volume of available genomes, novel methodology will be required to fast processing large datasets in a computational-efficient way.

Zusammenfassung

Pflanzen beherbergen phylogenetisch vielfältige Mikroben auf der Außenseite und im Inneren aller Organe, und sie gehen enge Beziehungen mit diesen Mikroorganismen ein. Die Fortschritte von „Multi-Omics“ Methoden der letzten Jahre haben erheblich dazu beigetragen, unser Verständnis von den Interaktionen von Pflanzen und deren Mikroorganismen zu erweitern. Um kausale Zusammenhänge besser verstehen zu können, ist es notwendig, die Interaktion zwischen Pflanzen und Mikroorganismen unter sterilen Bedingungen mit Hilfe von reduktionistischen Ansätzen, bestehend aus Dekonstruktion und Rekonstruktion, zu untersuchen. Die Dekonstruktion der pflanzlichen Mikrobiota bedeutet dabei das Erstellen von mikrobiellen Kultur Sammlungen, sowie die Sequenzierung der mikrobiellen Genome. Der Rekonstruktionsschritt besteht darin, synthetische mikrobielle Gemeinschaften (SynComs) zu entwerfen, indem die Stämme aus den Kultur Sammlungen gemischt und auf die Pflanzen beimpft werden.

In dieser Dissertation präsentiere ich das von mir entwickelte Tool Rbec, das ausschließlich dazu dient, die genaue mikrobielle Zusammensetzung, basierend auf Amplikon-Sequenzierung, in einer SynCom zu charakterisieren. Dabei korrigiert Rbec auch PCR- und Sequenzierfehler und identifiziert paraloge Gene innerhalb eines einzelnen Bakterienstammes. Als zusätzliche Neuerung kann man mit Rbec die Daten vom Amplicon-Sequenzieren auf mögliche Kontaminationen hin testen. Mit Hilfe von Rbec habe ich Amplikon-Sequenzierung Daten aus Studien mit natürlichen mikrobiellen Gemeinschaften sowie SynComs analysiert, darunter die Untersuchung die Präferenz von kommensalen Bakterien, ihren natürlichen Wirt – Arabidopsis thaliana und Lotus Japonicus - zu besiedeln, die mikrobiellen Gemeinschaften in der Phycosphäre von Chlamydomonas reinhardtii, die Auswirkung von pflanzlichen Metaboliten auf das Mikrobiota und die Frage, wie bakterielle Antibiotika das Wurzel Mikrobiota beeinflussen.

Mikrobielle Kultur-Sammlungen mit sequenzierten Genomen ermöglichen es uns, die funktionellen Fähigkeiten der Mikrobiota zu untersuchen. Wir analysierten systematisch die biosynthetischen Gencluster und die Verbreitung von antimikrobiellen 2,4-Diacetylphloroglucinol-Synthesegenclustern in Pseudomonas Stämmen. Darüber hinaus untersuchte ich den horizontalen Gentransfer (HGT) in Bakterien aus verschiedenen Kultur-Sammlungen, die von unterschiedlichen Wirtspflanzen und Standorten stammen. Dies liefert einen Atlas der aktiven Taxa, die am HGT beteiligt sind, und den häufig übertragenen funktionellen orthologen Genen in Pflanzen assoziierten Nischen. Darüber hinaus zeigt es die

Selektion Kräfte auf, die auf verschiedene Taxa in den jeweiligen Umgebungen ausgeübt werden.

Zusammenfassend lässt sich sagen, dass wir mit unserer Arbeit versucht haben, die reduktionistischen Ansätze unter dem Aspekt der computergestützten Analysen voranzubringen. Wir haben nicht nur eine neue Berechnungsmethode für die genaue Erstellung von Profilen der mikrobiellen Zusammensetzung in SynComs eingeführt, sondern sind auch tiefer in die mikrobiellen Kultur-Sammlungen eingedrungen, indem wir die bakteriellen Genome analysiert haben. Die wertvollen Informationen der bakteriellen Genome der pflanzlichen Mikroorganismen bieten die Möglichkeit, die funktionelle Vielfalt, mögliche Evolution und die genomischen Inhalte im Zusammenhang mit den Anpassungen an den Wirt zu untersuchen. Angesichts der zunehmenden Menge an verfügbaren Genomen sind jedoch neue Methoden erforderlich, um große Datensätze schnell und effizient zu verarbeiten.

Acknowledgement

First of all, I could not thank more to my actual supervisor Dr. Ruben Garrido-Oter for his supervision on my PhD study. Without his supervision and expertise, I could not have accomplished all the scientific researches in time. I would also like to thank him for offering me an independent research environment in which I have the opportunity to learn and explore my own scientific research interests. I learnt a lot from him about how to draft scientific manuscripts more reasonably and response to reviewers' comments precisely and concisely. Furthermore, I would like to thank my official supervisor Prof. Dr. Alga Zuccaro and Prof. Dr. Eric Kemen, one of my TAC members, for providing valuable suggestions to my projects and thank the SPP foundation led by Alga Zuccaro for supporting my study and livelihood in Germany.

I would also like to thank all my group members, Rui Guan, Yulong Niu, José Flores-Uribe, Jia Yu, Paloma Duran, Eik Dahms, Niklas Kiel and Magdalena Slawinska, for the wonderful discussions with them and their inputs into my projects. As an old Chinese idiom saying "If three of us are walking together, at least one of the two others is good enough to be my teacher", we can always learn from each other and push our projects forward in the right direction.

To be honest, we all suffered a lot in the past three years because of the COVID pandemics. Because of the lockdown and restrictions on entering China from abroad, I have not visited my parents for more than 3 years. I would sincerely thank them for their support and understanding that I pursued my PhD degree in a country that is thousands of miles away from my hometown. I have to thank my sister for accompanying my parents and taking care of them when I am absent. I feel deeply sorry for being absent along the ways of growth of my nephew and niece. I truly hope they can understand and forgive me. Last but not least, I would like to thank all my friends for accompanying me all along the way during my PhD study, especially my best friend Qi Zhang.

Contents

List of abbreviations

ANI	Average nucleotide identity
ARGs	Antimicrobial resistance genes
ASV	Amplicon sequence variant
At	*Arabidopsis thaliana*
BGCs	Biosynthetic gene clusters
BOA	Benzoxazolinone
CAS	Cologne agriculture soil
CFU	Colony forming units
CPCoA	Constrained principle coordinate analysis
Cr	*Chlamydomonas reinhardtii*
DAPG	Diacetylphloroglucinol
GC-MS	Gas chromatography–mass spectrometry
GWAS	Genome-wide association analysis
HGT	Horizontal gene transfer
HPLC	High-performance liquid chromatography
ITS	Internal Transcribed Spacer
Lj	*Lotus japonicus*
mBA	Modified Burkholder plate-based assay
OUT	Operational taxonomic unit
PCoA	Principle coordinate analysis
PLFA	Phospholipid fatty acid
qPCR	Quantitative polymerase chain reaction
Rbec	Reference-based error correction
Rs	*Ralstonia solanacearum*
SD	Standard deviation

SNP	Single nucleotide polymorphism
SSA	Succinate semialdehyde
SynComs	Synthetic communities
T4SS	Type IV secretion system
UPLC-MS	Ultra performance liquid chromatography - tandem mass spectrometer
WT	Wide type
AtGr-sphere	Root culture collection of *Arabidopsis thaliana* from Germany
AtIr-sphere	Root culture collection of *Arabidopsis thaliana* from Italy
AtUr-sphere	Root culture collection of *Arabidopsis thaliana* from the United States
LjGr-sphere	Root culture collection of *Lotus Japonicus* from Germany
AtSl-sphere	Leaf culture collection of *Arabidopsis thaliana* from Switzerland and Germany
CrG-sphere	Phycosphere culture collection of *Chlamydomonas reinhardtii* from Germany

Chapter 1

Introduction

1.1 Interactions between plants and the soil microbiota

In nature, terrestrial plants are populated by phylogenetically diverse microbes on the surface and in the interior of all their organs (Figure 1.1). These microorganisms include bacteria, fungi and oomycetes who can establish mutualistic, commensal and pathogenic relationships with plants (Berendsen et al. 2012; Bulgarelli et al. 2013; Müller et al. 2016b; Trivedi et al. 2020), and are collectively termed the plant microbiota. Plants provide a specific niche for the colonization of microbes by releasing rich nutrients into their surroundings, and microbial taxa favored by these chemicals are recruited to inhabit plant-associated niches from the soil and atmosphere. In turn, indigenous microbes can modulate plant growth and health, e.g., by protecting plants against soil-borne pathogens (Berendsen et al. 2018; Ciancio et al. 2019) and participating in nutrient mobilization (Harbort et al. 2020). The establishment of associations between soil microbes and plants may date back to the rise of terrestrial plants (Lambers et al. 2009). Analysis of the rhizosphere microbiota of *Arabidopsis thaliana* and its relatives, which diverged from one another ~35 million years ago revealed a strong co-evolutionary signal between microbiota and plants (Schlaeppi et al. 2014). This study highlights that plants fine-tune their microbiota to sustain these interactions over long-term evolutionary time scales. However, recent studies also show the turnover of microbial communities from domesticated crops and wild accessions, suggesting a short-term evolutionary relationship. That being said, a lower contribution of domesticated crops to the variation of microbial composition is also observed compared to the wild accessions, probably as a consequence of reduced microbial diversity after plant domestication owing to external input of chemical fertilizers (Bouffaud et al. 2012; Peiffer et al. 2013). This reduced microbial diversity would undermine microbial resilience upon encountering stressors and further disrupt the diversity of soil microbiota. To reduce the usage of chemical fertilizers and the footprint of domestication on the environment, the use of engineered microbiota is a promising alternative. Understanding the assembly rules of plant microbiota would be an essential step towards engineered microbiota for sustainable agriculture. Many studies have shown that plant genotypes, root-secreted secondary metabolites, developmental stage of plants and edaphic factors contribute to the variation of plant microbiota (Pascale et al. 2020; Schlaeppi et al. 2014; Zhang et al. 2018). More specifically, edaphic factors have been identified as the most important factors in shaping plant bacterial communities and climate is more important for variation in root-associated filamentous eukaryotic communities in well-designed experimentations with combinations of multiple variables (Schlaeppi et al. 2014;

Thiergart et al. 2019). Furthermore, GWAS analysis has been employed to systematically examine the plant genes controlling the overall indigenous microbial communities and individual taxa (Beilsmith et al. 2019; Deng et al. 2021; Escudero-Martinez et al. 2022; Wang et al. 2022), which is a promising strategy to survey plant genetics and microbes inducing phenotypes of interest. Though extensive variations have been observed at the finer level (species/strain) among plant microbiota from different compartments, host species and environments, a relatively conserved abundant core microbiota has been reproducibly found across different studies, which comprises *Proteobacteria*, *Actinobacteria* and *Bacteroidetes* (Müller et al. 2016b). This consistent pattern suggests that microbial members from those taxonomic groups are highly adapted to plant-associated niches and may be important for plant physiology.

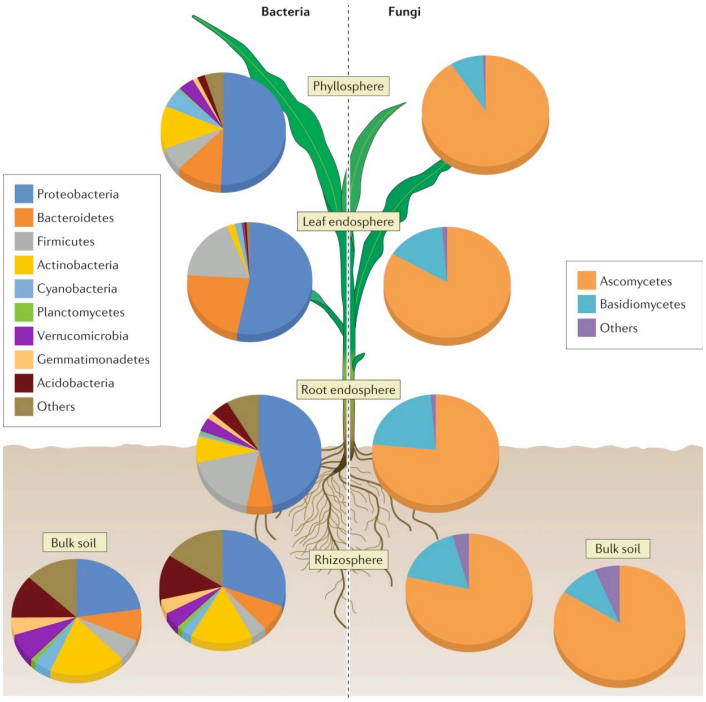


Figure 1.1 Different compartments of plant microbiota and the corresponding microbial compositions. Adapted from Trivedi et al. 2020.

1.2 Reductionist approaches for studying plant-associated microbiota

Association studies can identify multiple microbial groups or taxa correlated with specific plant phenotypes; nevertheless, causative insights into plant-microbe associations are important for understanding the molecular mechanisms underpinning the modulation of plant phenotypes by microbiota or *vice versa*. To establish causality, studying plant-microbiota interactions in controlled environments to rule out other unexpected variables is important, which has led to pioneering reductionist approaches for studying plant-microbiota

interactions. Reductionist approaches were first exploited by studying the root microbiota of *Arabidopsis thaliana* under well-controlled laboratory conditions to explicitly monitor the moisture, temperature and wind in the growth chambers (Bulgarelli et al. 2012; Lundberg et al. 2012). Besides the control of environmental factors, manipulation of microbial communities was also developed in the last decade through the use of Synthetic Microbial Communities (SynComs) (Vorholt et al. 2017), through which it is possible to either induce or deplete specific microbial members in the indigenous communities. The development of SynComs undoubtedly owes a lot to the establishment of diverse microbial culture collections isolated from plants.

1.2.1 Deconstruction of plant-associated microbiota

The plant microbiota encompasses phylogenetically diverse microbes, so efforts to establish a representative cross-section of microbial isolates that covers the diversity of natural microbial communities is labor-intensive. Zhang et al. proposed a streamlined high-throughput cultivation and identification of bacteria from the root microbiota via limiting dilution (Zhang et al. 2021). They succeeded in characterizing 13,512 isolates from rice roots covering 70% of the rice root bacterial microbiota members, which suggests that this method can be broadly applied to bacterial cultivation in the future. With the continuous efforts and inputs from the scientific community, several bacterial stocks have been established from a variety of plants, including the model plant *Arabidopsis thaliana*, rice (*Oryza sativa*), barley (H*ordeum vulgare*), maize (*Zea mays*), tomato (*Solanum lycopersicum*), potato (*Solonum tuberosum*), clover (*Trifolium pratense*), sugarcane (*Saccharum sp.*), green alga (*Chlamydomonas reinhardtii*), and *Lotus Japonicus* (Armanhi et al. 2018; Bai et al. 2015; Durán et al. 2022a; Hartman et al. 2017; Khan Chowdhury et al. 2017; Kwak et al. 2018; Levy et al. 2018; Wippel et al. 2021; Zhang et al. 2019). The majority of the isolates are genome-annotated, which represents a valuable resource for the study of the functions, genetic diversity and evolution of plant microbiota. Two recent studies showed the phylogenetic diversity of *Pseudomonas* and the onset of a commensal lifestyle in *Rhizobiales*, respectively, by leveraging the genomes from thousands of plant-derived isolates (Garrido-Oter et al. 2018; Karasov et al. 2018). In general, the aim of bacterial cultivation is to study the plant-microbiota interactions and the assembly patterns of the plant microbiota in controlled environments.

1.2.2 Reconstruction of the plant-associated microbiota

With access to enormous numbers of microbial isolates, mainly bacteria, collected from the same plant species grown in the same field, we can use bottom-up combination approaches to mix the selected microbes into artificial SynComs with different complexities, i.e., different numbers of strains. Inoculation of plants with these SynComs then facilitates the study the interactions between plants and microbiota in gnotobiotic systems. By utilizing SynComs, researchers can perform targeted manipulation of biotic factors, such as the presence/absence of specific microbes, the abundance ratios between different microbes, the order of introduction of some microbes and the inclusion of genetically modified microbes (Carlström et al. 2019; Wippel et al. 2021). Though SynComs cannot recapture the full phylogenetic diversity of natural communities, it has been shown that the SynComs reduced in alpha diverisities could recapitulate the microbial higher-taxonomic-level diversity observed in natural communities (Bai et al. 2015). More importantly, SynCom experiments can enable establishment of causality (Vorholt et al. 2017), which is infeasible to test under natural environments. A milestone in studies of the reconstruction of the plant microbiota showed that survival of *Arabidopsis thaliana* was enhanced by inter-kingdom microbial interactions compared to the germ-free *Arabidopsis* (Durán et al. 2018), which highlights the essential role of consortia of phylogenetically diverse microbes on plant health. A recent study that studied two SynComs with opposite impacts on the plant immune system proposed a 'rheostat model' in which the balance between immune-suppressive and non-suppressive strains could contribute to the susceptibility of plants to pathogens (Ma et al. 2021). These two studies emphasize that SynCom experiments can either focus on the complexity of microbial interactions or on physiological function.

1.2.3 Computational analyses of community compositions in synthetic microbial communities

Characterization of community compositions in SynComs is relatively easy compared to in natural communities due to the reduced complexity and because the identities of the microbes included in the communities are known. Amplicon sequencing of marker genes (e.g., 16S rRNA gene for bacteria and ITS for fungi) has been widely used to quantify strain abundance in both natural and synthetic communities. Genome-indexed culture collections simplify the data processing procedure for amplicon sequencing. Closed operational taxonomic unit (OTU) picking or the exact match method has been employed to dissect community compositions in SynComs via the alignment of sequencing outputs against the reference sequence of each strain in the corresponding community (Carlström et al. 2019; Durán et al.

2018; Ma et al. 2021; Wippel et al. 2021). Biases in strain abundance can be introduced by either of the methods, which are attributed to the arbitrary clustering threshold, misclustering, sequencing errors and the presence of polymorphic paralogues of marker genes within the same strain. With the inclusion of close relatives in the SynComs that cannot be distinguished at the OTU level, the closed OTU picking method is inapplicable. Moreover, the strain abundance cannot be directly translated into cell counts of the corresponding strain because of variations in DNA extraction and PCR amplification efficiencies and marker gene copy numbers (Sun et al. 2021; Vorholt et al. 2017). By taking advantage of genome-indexed culture collections, it has become feasible to predict marker gene copy numbers, such as of the 16S rRNA gene, from whole-genome assemblies (Perisin et al. 2016). Further, to circumvent shortcomings of relative abundance estimations and perform accurate comparison across samples, qPCR and spike-in sequences can be deployed to quantify the absolute abundance of each strain (Bodenhausen et al. 2014; Guo et al. 2020; Tkacz et al. 2018). One of the prerequisites for SynCom experimentation is the absence of biological contamination in the gnotobiotic systems; thus, it is critical to evaluate contaminants in the system, an issue which has been largely disregarded in previous studies. The establishment of a standardized protocol to check contamination should be considered, given the expected explosion of SynCom studies in the near future.

1.3 Genome-scale analysis with microbial culture collections

Microbial culture collections allow researchers to recapitulate the microbiota-responsive traits of plants observed in nature via mono-associations or a multitude of synthetic community experiments in planta. Genome-indexed plant microbiota have provided further functional insights into the evolution and adaptation of microbes to plants. In this section, I summarize the computational analyses that have been applied to genome-indexed microbial culture collections and discuss how genome-scale analyses facilitate our understanding of genetic and functional diversities in the plant microbiota.

1.3.1 Population-level genome analysis of commensals

Surveys on the microbial composition of the plant microbiota have revealed the persistent colonization of some abundant taxa, including *Burkholderiaceae*, *Xanthomonadaceae*, *Rhizobiaceae* and *Pseudomonadaceae* (Müller et al. 2016b). Given their high cell densities, isolates from the culture collections are taxonomically biased towards the dominant taxa. Genetically diverse strains from the same taxa in the culture collections provide the

opportunity of studying the dynamics and evolution of the taxa in plant-associated niches. Karasov et al. established a *Pseudomonas* culture collection specifically consisting of OTU5 strains, which have been found to be prevalent and dominant on the leaves of different *Arabidopsis* populations across seasons (Karasov et al. 2018). Those strains are pathogenic to *Arabidopsis thaliana* and some of them are found to co-exist within a single plant though they are predicted to diverge 300,000 years ago. The large genomic variations found within those pathogenic strains challenge the boom-bust concept that a single-lineage pathogenic clade contributes to epidemics in agricultural systems (Butler et al. 2013; Kolmer 2005; Yoshida et al. 2013). A similar genomic study of 944 representative *Rhizobiales* strains isolated from legume and non-legume host plants revealed that nodulation-related genes are absent in the most recent common ancestor and that they must have been acquired via horizontal gene transfer during co-incubation with legume plants (Garrido-Oter et al. 2018). Though *Rhizobiales* is important for nodulation in legume plants, this study reports that the commensal lifestyle of those strains predates the acquisition of nodulation-related genes. These two studies illustrate that population-level surveys of the genomic diversity in plant-associated taxa provide insights into plant-bacteria population dynamics and evolution.

1.3.2 Detection of genomic features related to adaptation to the host environment

Understanding how microbes switch from free-living soil microbes to commensals of plants is a long-standing question in the area of plant-microbe interactions. Great efforts have been put into identifying the genetic elements responsible for microbial adaptation to host-associated niches based on comparative genomics (Ailloud et al. 2015; Langridge et al. 2015; Thomas et al. 2021; Thomson et al. 2008). For instance, Levy et al. constructed comprehensive plant-associated bacterial genomes with the inclusion of publicly available genomes and newly sequenced genomes from microbial culture collections (Levy et al. 2018). This database allowed them to identify the genomic contents that are associated with bacterial adaptation to host plants by comparison with genomes isolated from other environments. Miyauchi et al. performed comprehensive comparative genomic analysis on mycorrhizal fungi and observed some genomic features related to the transition from saprotrophy to symbiosis, which included the widespread loss of some carbohydrate-degrading genes (Miyauchi et al. 2020). Another recent study also identified the genomic traits that are potentially related to the endophytic lifestyle of fungi by performing comparative genomic analysis between genomes of fungi isolated from *Arabidopsis thaliana*

and those with other lifestyles (Mesny et al. 2021). These studies highlight the importance of integrated microbial genome datasets for assessing genomic contents associated with traits of interest coupled with well-documented metadata.

1.3.3 Horizontal gene transfer prediction in microbiomes

Horizontal gene transfer is a common phenomenon in species across the tree of life and is a major force driving the evolution and adaptation of bacteria in changing environments. One widely known example is the dissemination of antibiotic resistance genes in important pathogens that results in the emergence of multidrug-resistance bacteria that pose risks to human health (Kent et al. 2020). Dissection of HGT events in microbes can provide insights into the adaptation mechanisms, evolutionary tracks and potential selection forces exerted on microbes. With improvements in culturomics and the sharp decrease in genome sequencing costs, high-quality genome assemblies make computational prediction of HGT feasible (Brito 2021). The direct detection of mobile genetic elements (MGEs) in genomes is a promising strategy for detecting signatures of HGT. However, the *de novo* assembly of MGEs is difficult given the presence of direct or inverted repeats flanking MGRs (Pop 2009). Pairwise whole-genome comparison has become the gold standard for HGT detection, and several different strategies have been described for examining HGT based on genome comparison. MGEfinder (Durrant et al. 2020) searches for the insertion sites by mapping the sequencing reads of isolates against a reference genome. MetaCHIP (Song et al. 2019) aims to detect the transferred protein-coding genes in a set of phylogenetically diverse microbial genomes that show the best BLAST hits to genomes from other taxon groups rather than from the self-group, but this method is sensitive to the diversity of available genomes and the taxonomy level specified by users. Identical or near-identical genomic regions between any pair of distant species also serve as a signature of recent HGT, which has been deployed in human microbiome and cheese microbiome studies (Groussin et al. 2021; Handley et al. 2017; Smillie et al. 2011). Those reports revealed that shared environments, phylogenetic boundaries, cell densities and cell wall properties can influence the HGT frequency in microbiota. Condon usage bias, base composition and incongruence between the species tree and the gene tree can further strengthen the robustness of predicted HGT events.

1.4 Research aims

In the era of culturomics, we are drowned with explosive growth of microbial culture collections collected from diverse plant-associated environments and their corresponding

genome sequences benefiting from the profound decrease in the cost of sequencing. On one hand, we can take advantages of those isolates to study the plant-microbiota interactions under controlled conditions by using SynComs with reduced complexity; on the other hand, we can dig deeper into the *in silico* sequences to uncover the evolution and adaptations of bacteria to plant-associated niches.

The aim of this work was to develop and apply new and well-established computational approaches to analyze the sequencing data mostly generated during the process of applying reductionist approaches to study plant-microbiota interactions and partially came from natural environments. Culture-independent amplicon sequencing is a well-established method to profile the microbial communities in both natural communities and SynComs, however, given the shortcomings of the current prevalent analysis tools, they cannot produce accurate abundance tables from amplicon sequencing data. Those existing methodological pitfalls can be improved in principle when analyzing amplicon sequencing data from SynComs, so I developed a software named Rbec to precisely characterize the microbial compositions exclusively in SynComs. Furthermore, with the access to genome-indexed bacterial culture collections assembled from plants, genomic analyses can be applied to investigate the potential functional genes essential for bacterial adaptations to plants. Here, I used comparative genomics to identify near identical genes between any genome pairs, which represents potential recent HGT events, to reveal the evolutionary trajectories of microbes in plant-associated niches.

1.5 Outline of the research chapters

This work is a cumulative dissertation that contains four peer-reviewed articles published in scientific journals, one articles that is under review in a scientific journal and one article presented here before submission for scientific review. The first chapter, as presented above, includes an overarching introduction to the plant microbiota, reductionist approaches for studying plant-microbiota interactions and the computational approaches to analyze the sequencing data generated during applying the reductionist approaches (*Chapter 1*). The second chapter presents a novel computational method, named Rbec, to characterize the microbial compositions in SynComs (*Chapter 2*). The third chapter includes the brief reviews of the four scientific articles in which I applied Rbec and the in-house established pipeline to analyze the microbial compositions in SynComs and natural communities respectively (*Chapter 3*). The fourth chapter represents an atlas of the recent HGT in the plant microbiota and provides insights into bacterial adaptations to the plant-associated niches (*Chapter 4*).

Supplementary materials such as raw and intermediate data, supplementary figures and tables necessary to reproduce each of the figures and statistical tests presented in the published articles can be accessed from the online version of each article.

Chapter 2

Rbec: a tool to analyze amplicon sequencing data from the microbial synthetic communities

2.1 Abstract

Synthetic microbial communities (SynComs) are emerging as a powerful tool in biological, biomedical, and biotechnological research. However, despite recent advances in algorithms for the analysis of culture-independent amplicon sequencing data from microbial communities, there is a lack of tools specifically designed for analysing SynCom data, for which reference sequences for each strain are available. Here we present Rbec, a tool specifically designed for the analysis of SynCom data that accurately corrects PCR and sequencing errors in amplicon sequences and identifies intra-strain polymorphic variation. Extensive evaluation using mock bacterial and fungal communities shows that our tool outperforms current methods for samples of varying complexity, diversity, and sequencing depth. Furthermore, Rbec also allows accurate detection of contaminants in SynCom experiments.

2.2 Introduction

Amplicon sequencing is a powerful technique for characterizing the composition of microbial communities from environmental samples. Recent advances in algorithms and tools for the analysis of marker gene amplicon data have driven a shift from clustering approaches, based on operational taxonomic units (OTUs) and arbitrary sequence similarity thresholds, toward error correction methods (Amir et al. 2017; Callahan et al. 2016; Edgar and Flyvbjerg 2015; Peng and Dorman 2021) that seek to estimate abundances of individual amplicon sequence variants (ASVs). A new generation of integrated pipelines (Bolyen et al. 2019) allows researchers from a variety of fields in the environmental, biological, and medical sciences to reproducibly analyse marker gene sequencing data.

Synthetic microbial communities (SynComs) constitute a powerful emerging tool for building experimentally tractable, reproducible microbial systems in the laboratory that enable controlled perturbation experiments and testing of falsifiable hypotheses. These bottom-up, reductionist approaches are increasingly employed in studies of microbial ecology and evolution (Cairns et al. 2020), the plant and animal microbiota (Bai et al. 2015; Vrancken et al. 2019; Zhang et al. 2019), and biotechnology (McCarty and Ledesma-Amaro 2019). However, the lack of bioinformatic tools specifically designed for the analysis of sequencing data obtained from gnotobiotic systems and SynComs is preventing these innovative experimental approaches from developing to their full potential. As a result, researchers typically employ standard clustering, error-correcting or mapping approaches that do not take

full advantage of these tractable experimental systems (e.g., reduced community complexity and the availability of reference sequences for classification), resulting in reduced resolution and accuracy or data loss. To address this limitation, we have developed a reference-based error correction algorithm that is able to accurately and precisely correct PCR and sequencing errors in SynCom amplicon data, identify intra-strain polymorphism, and detect the presence of contaminants in gnotobiotic systems.

The algorithm that we have developed, Rbec, is an easy-to-use tool, freely available as an R package. Rbec corrects amplicon sequencing errors by implementing a modified version of the quality-aware model implemented in the DADA2 tool: the error matrix in DADA2 is expanded to a 20*43 matrix in Rbec by including insertion cases in amplicon sequencing reads (Callahan et al. 2016). Further, Rbec also identifies intra-strain polymorphic variation and contaminants in samples of SynComs. Rbec is specifically designed to efficiently and accurately process data from SynComs, for which reference sequences of individual community members are available. A detailed description of the Rbec algorithm is provided in the Materials and Methods and an overview is given below. Rbec is freely available as an open-source multi-platform R package. Release versions can be obtained via Bioconductor. The developer version is maintained and can be downloaded at: https://github.com/PengfanZhang/Rbec.

2.3 Results

2.3.1 Extensive sequencing errors in amplicon sequencing data

It has long been acknowledged that errors are expected in the outputs from sequencing as a result of PCR and sequencing errors, which can strongly impact computational analysis of sequences. In the case of microbial community characterization with amplicon sequencing, each sequence is assumed to be from a single strain, and different sequences can be distinguished as coming from different strains by ignoring marker gene paralogues in the same strain. Under this scenario, sequencing errors would strongly influence the microbial communities predicted from computational analysis. To evaluate the error distribution in amplicon sequencing, we performed amplicon sequencing with V5–V7 variable regions on 236 bacterial strains isolated from root samples separately and *in silico*-mapped the reads against the reference sequence of each strain. Surprisingly, only 31.8% of all reads per sample, on average, had a perfect match (100% identity) in the database (Figure 2.1). Around half of the remaining reads showed ≥99% similarity and 5% of reads showed similarities

ranging from 97% to 99% to the corresponding reference sequence. The remaining 10% of reads were less identical to the reference sequences with similarities <97%. This result indicates the presence of extensive sequencing and PCR errors as well as polymorphic copies. To rule out the possibility that this outcome was a laboratory-specific issue, we also analyzed published data generated on a different sequencing platform by another laboratory. The error distribution pattern identified in this data showed a very similar pattern to our original finding.

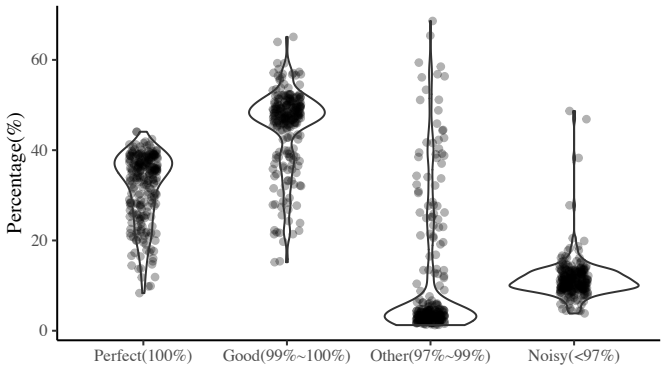


Figure 2.1: The error distribution of amplicon sequencing data. Amplicon sequencing of the 16s V5–V7 region from each of 236 strains isolated from the roots of *At* was mapped to the reference sequence of the corresponding strain. 'Perfect' reads represent those exactly matching the reference sequences. Y axis shows the relative abundance of reads from each category in each strain.

2.3.2 Design and workflow of Rbec

To overcome the extensive mismatches between the reads and corresponding reference sequences, we set out to develop a framework to simultaneously correct erroneous reads and identify polymorphic copies, enabling us to carry out precise microbial community profiling in SynComs with amplicon sequencing. By taking advantage of tacks from the prevalent error-correcting software for amplicon sequencing data from natural communities, e.g., DADA2 and Unoise (Callahan et al. 2016; Edgar and Flyvbjerg 2015), we can adapt the algorithms to fit the data from SynComs in which the strain identities and reference sequence for each member in the community are known. The schematic workflow is described in Figure 2.2.

First, reads are de-replicated into unique tags and subsequently aligned to the reference database containing amplicon sequences from SynCom members, typically generated from sequencing of clonal cultures. Initial abundances are then assigned to each strain according to the copy number of each exactly aligned tag. Next, tags that are not exactly matched to any sequence in the database are assigned a candidate error-producing reference based on *k*-mer

distances. Sequencing reads are then subsampled, and an error matrix is calculated using the mapping between subsampled reads and candidate error-producing sequences. The probability that a unique tag is erroneously produced by a given candidate error-producing sequence is then calculated using a Poisson distribution. The probability and expectation values of this distribution are then used to determine whether a unique tag can be corrected from a reference sequence or whether it can be identified as originating from a paralogous sequence. Tags that cannot be corrected are subsequently removed. The parameters of the error model are recomputed iteratively until the number of re-assignments falls below a set threshold. Strain abundances are then estimated from the number of error-corrected reads mapped to each reference sequence. Finally, potentially contaminated samples are identified by assessing a significant deviation from the expected proportion of corrected reads. Sequences of putative contaminants are then provided as an output for further examination.

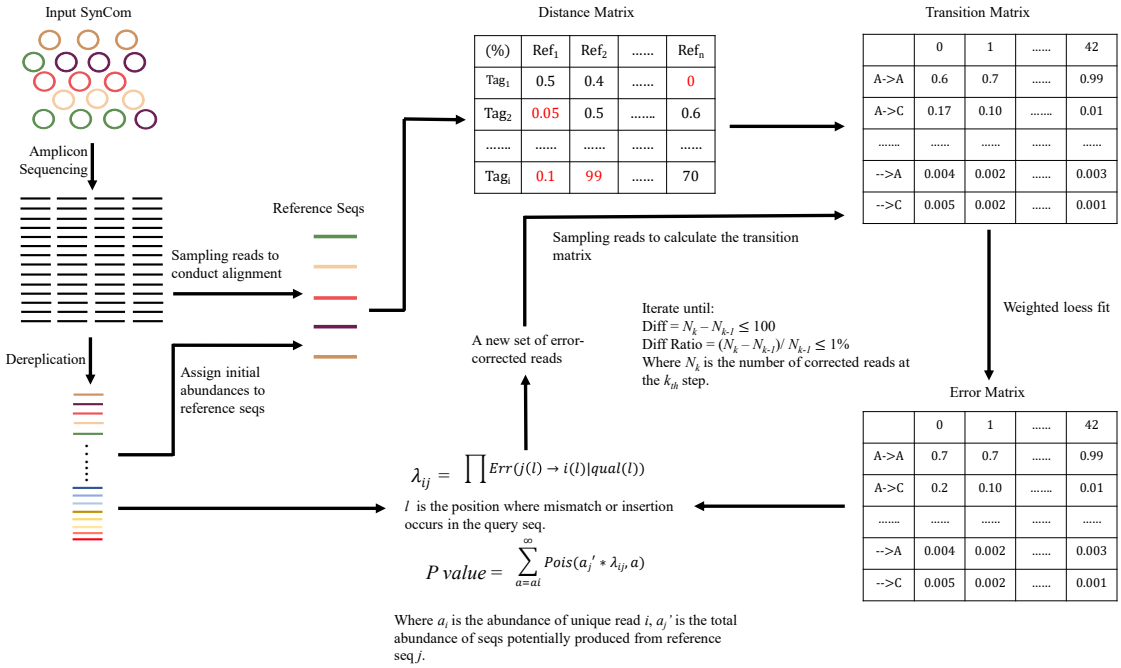


Figure 2.2: Schematic diagram of the Rbec algorithm. Rbec consists of two main steps: error matrix estimation and abundance probability calculation. For the error matrix estimation, Rbec traverses through all query reads and reference sequences and matches each read with a unique candidate error-producing reference. Alignments between input and reference sequences are then used to calculate the error matrix. Abundance probabilities are then estimated by fitting a Poisson distribution.

2.3.3 Rbec corrects erroneous reads and identifies 16S rRNA gene paralogues in a single strain

Next, we tested the performance of Rbec in correcting the erroneous reads and identifying polymorphic copies by applying it to the amplicon sequencing data from individual strains (Durán et al. 2018). Our implementation of the Rbec algorithm successfully corrected most erroneous reads (89.2% on average), outperforming all other tested *de novo* correction methods. We asked whether this improvement is largely due to higher error correction efficiency or polymorphic copy identification. To this end, we split the strains into two sets: without 16S polymorphism and with 16S polymorphism, such that the overall improvement can be decomposed into either a complete improvement in error correction or a combined improvement in both aspects. For the strains without polymorphism, we noticed that Rbec showed slightly higher efficiency in correcting erroneous reads compared to DADA2 and Unoise but had overwhelmingly better performance than Deblur and AmpliCI (Figure 2.3A). For the strains with polymorphism, Rbec exhibited superior performance compared to all other methods and correctly allocated >20% more reads to the corresponding reference sequence compared to DADA2, Unoise and other software (Figure 2.3A), which could be attributed mainly to the identification of polymorphic copies in the same strain. To further justify this conclusion, we examined the sequence similarity and abundance ratio between the most abundant and the second most abundant ASVs after error correction by different methods. As expected, almost all of the samples processed with Rbec showed lower abundance ratios and sequence similarities compared to other methods (Figure 2.3B), which supports the capacity of Rbec to identify polymorphic copies in the same strain. Taken together, the implementation of Rbec on amplicon sequencing data from a single strain suggests that Rbec outperforms other error-correcting methods in data processing through its ability to correct more erroneous reads and innovatively identify polymorphic copies.

2.3.4 Rbec outperforms other methods in characterizing synthetic microbial communities

Rbec shows excellent performance when used on amplicon data from single strains, suggesting that it could be effectively applied to SynComs. To evaluate the accuracy of Rbec in characterizing community composition, we simulated *in silico* bacterial and fungal mock samples by mixing reads generated from sequencing individual isolates separately. For these simulations, we varied community complexity, strain similarity and sequencing depth. Across these three parameters, Rbec consistently performed better than all other tested methods in characterizing microbial composition in terms of deviation from the ground truth (Figure 2.4). In further support of its stable performance, Rbec always showed smaller variance in

terms of deviation from the ground truth across different simulations than other methods. Closed OTU picking method exhibited large variance and lower accuracy when the strains in the community were phylogenetically close, probably due to ambiguous clustering among reads from close relatives. These results show that Rbec can be applied to amplicon sequencing data from SynComs with any kinds of combination parameters and even with lower sequencing depths.

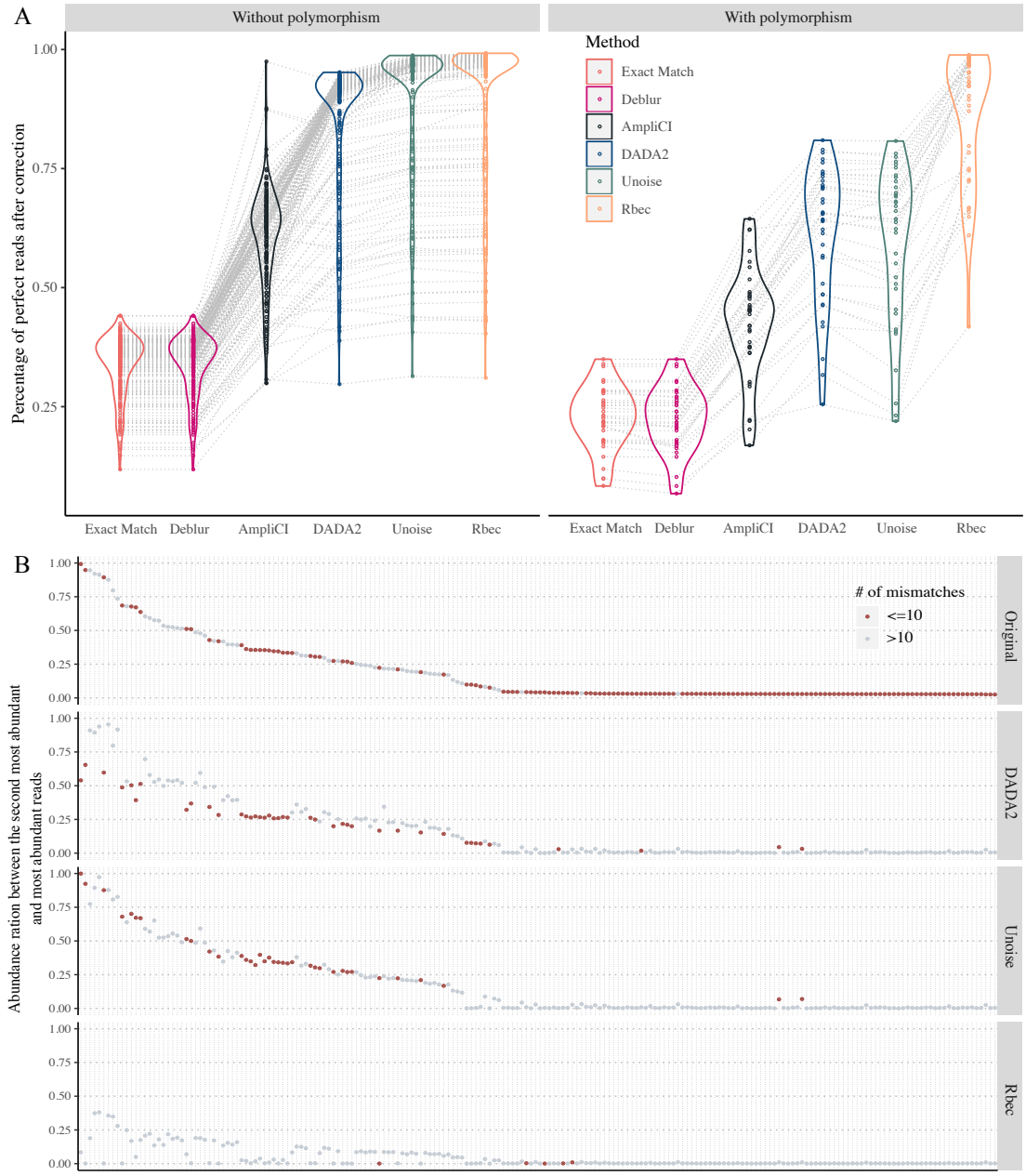


Figure 2.3: Performance of Rbec in correcting errors and identifying polymorphic copies in the same strain. (A) The percentages of corrected reads by different software in strains without and with 16S polymorphisms. (B) Abundance ratios between second-most abundant and most abundant unique tags in each sample before and after correction with different methods. The x axis represents different strains and the y axis represents the abundance ratio calculated as $\frac{\text{Abundance of second-most abundant unique tag}}{\text{Abundance of most abundant unique tag}}$. The color of each dot

indicates the sequence dissimilarity between the two unique tags. Data points are depicted in dark brown if the two tags are close (≤ 10 base mismatches).

2.3.5 Contamination detection with Rbec

SynCom experiments are strictly controlled in gnotobiotic systems and are susceptible to accidentally introduced microbial contaminants. Without clear verification of contaminants, the readouts of experiments can be misleading. Given that Rbec not only corrects most

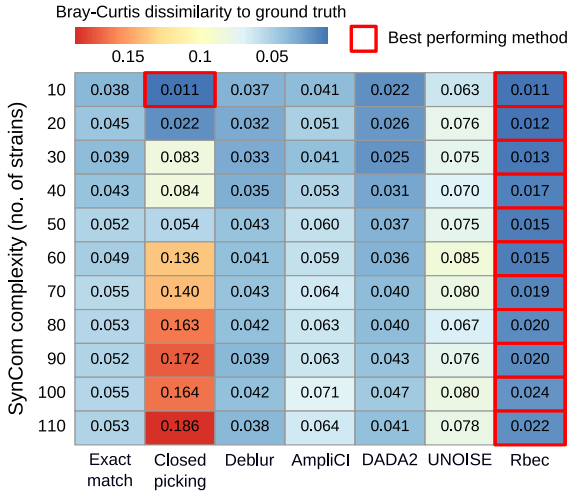


Figure 2.4: Evaluation of the influence of community complexity on the performance of different methods, measured as a deviation from the ground truth using Bray-Curtis dissimilarities. The column labels represent methods and the row labels represent the number of strains in the mock community implemented to analyse the mock data. The values inside the heatmap refer to the averaged Bray-Curtis dissimilarities over 20 replicates for each SynCom combination.

erroneous amplicon sequencing reads, but also successfully identifies paralogous sequences, we can assume that a high proportion of uncorrected reads is likely the result of contamination. We evaluated the capacity of Rbec to identify contaminated samples by performing an *in silico*-simulated dataset, where we included amplicon sequences from *Escherichia coli* with 5% relative abundance per 'contaminated' mock sample. When examining the percentage of reads successfully corrected per sample (Figure 2.5), we observed a clear separation between 'clean' and 'contaminated' mock samples. This observation motivated us to include a function in Rbec to identify the outlier samples with very low percentage of corrected reads compared to other samples and flag them as 'contaminated' samples. The sequences from contaminants can be isolated for further identification.

2.4 Discussion

In this work, we developed a software, Rbec, to precisely characterize the microbial compositions of SynComs through correction of erroneous reads and identification of reads from polymorphic copies. When comparing Rbec with other methods that can be used for SynCom analysis, we found that the performance of the conventional closed OTU picking

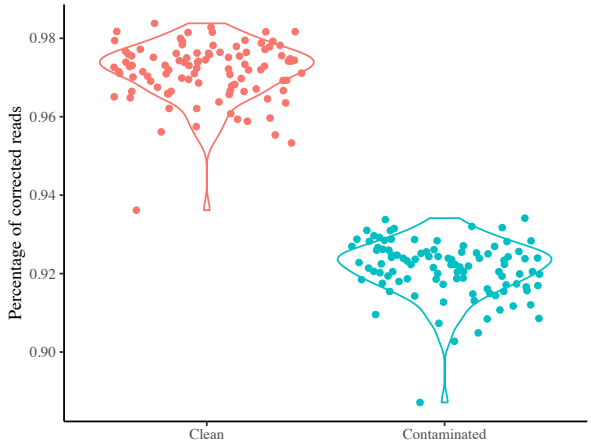


Figure 2.5: Segregation of percentages of corrected reads between clean and contaminated SynCom samples. A set of 100 'clean' mock communities was generated by randomly picking up 50 bacteria from the bacterial seed pool and mixing the subsampled reads from corresponding strains. To generate a comparable set of contaminated samples, amplicon reads from the *E. coli* K12 *16S* rRNA sequence were added to each clean mock community to make up 5% relative abundance.

method is diminished as the number of strains and phylogenetic relatedness increases in SynComs. Though we cannot conclude what potential biases and misinterpretations have been introduced by the closed OTU picking method, this points to the importance of careful selection of software and evaluation when dealing with data from SynCom samples as well as scrupulous surveillance of contamination. We recommend Rbec for the analysis of amplicon sequencing data from SynComs, given it is the best-performing and most stable tool currently available. Rbec is easy to use and highly customizable. As well as being parallelizable, it can also be run on a standard modern desktop or laptop computer and process amplicon samples containing thousands of sequencing reads within minutes using a single CPU core.

2.5 Materials and methods

Detailed description of the Rbec algorithm

A schematic workflow of Rbec is shown in Figure 2.2. A detailed, step-by-step description of the algorithm is detailed below.

Dereplication and assignment of initial abundances

Initial abundances for each sequence present in the reference database are inferred on the basis of the number of exact matches found in the uncorrected sequencing reads. First, merged reads are dereplicated into unique tags in a sample-wise manner. Sequence quality scores for each unique tag are averaged over the scores of all identical copies of that sequence and for every residue. Each reference sequence is assigned an initial abundance equal to the number of identical copies of that unique tag found in the sample. If a reference sequence does not have any tags that exactly match it, the strain from which the reference sequence is derived is marked as 'absent' from the sample.

Assignment of erroneous reads to error-generating reference sequences

Unique tags that do not exactly match any sequence in the reference database are initially assumed to originate from erroneous sequencing reads generated by a given reference sequence. In order to identify the most likely error-generating reference sequence for each unique tag, *k*-mer (*k*=7) distances between each unique tag and each reference sequence are calculated. We use *k*-mer distances for pairwise comparisons between unique tag queries and reference sequences instead of computationally demanding global alignments to improve the time performance of the algorithm. The reference sequence showing the lowest *k*-mer distance to the query unique tag is marked as the candidate error-producing sequence from which the corresponding erroneous sequences originate. If multiple candidates with the same *k*-mer distance are found, only the reference sequence with the highest initial abundance is considered as the original error-generating sequence, as sequences with higher abundances are more likely to generate erroneous sequences.

Estimation of the transition and error matrices

To calculate the probability that a unique tag is produced by a given error-generating reference sequence, transition and error matrices need to be estimated. The transition matrix is a 20 by 43 matrix where the rows represent the transition combinations (e.g., A→A, A→T, A→G, A→C, T→T, …, C→G, C→C; including insertions), and the columns represent the sequence quality scores. This transition matrix can be estimated by performing a global alignment between a random set of subsampled unique sequences (or 'tags'; 5,000 reads by default) and the reference sequences. Entries in the transition matrix are calculated by counting the number of each transition combination along the length of the alignment. The log-transformed transition matrix is then fitted with a weighted loess function to generate the error matrix.

Calculation of error-generation probabilities

We assume that the mismatches between query and reference are generated independently, so the rate at which a unique tag *i* is produced from the error-generating reference *j*, designated λ, is calculated by the product over the error probabilities at each position of the alignment *l*:

$$\lambda_{ij} = \prod_{l=1}^{L} Err(j(l) \rightarrow i(l)|qual(l))$$

Where *L* is the total length of the alignment.

Similar to the error-aware model implemented in DADA2 (Callahan et al. 2016), the abundance probability of each unique tag is calculated using the *Poisson* distribution:

$$E = a_j^{'} \lambda_{ij}$$

$$Pvalue = \sum_{a=a_i}^{\infty} Pois(E, a)$$

Where *E* is the expectation of the *Poisson* distribution, a_i is the abundance of a unique tag *i*, and a_j' is the aggregated abundances of all unique tags assigned to a reference sequence *j*.

Unique tags with a *P*-value lower than 10^{-40}, and an expectation lower than 0.05 are discarded. This expectation cut-off is intended to retain tags that could be produced at least once by the reference with the probability above 5%. The aim of this step is to retain tags that are generated from intra-strain amplicon sequence variants, which show high abundance relative to the reference sequence but do not exceed the *P*-value cut-off. It is possible that, for certain experiments, modifying these parameters could be useful. For instance, in a community containing very low-abundance strains, lowering the minimum expectation threshold might increase the sensitivity of the algorithm. Similarly, if the presence of low-abundance contaminants that are closely related to a reference strain is of particular concern, increasing the minimum P-value threshold will help identify potential contaminants, albeit at the risk of generating a larger number of false positives.

Since Rbec identifies the paralogues of the marker gene from the same strain, this opens up the possibility of estimating the copy numbers of a marker gene in a given genome by summing up the numbers of paralogues identified for each reference sequence from each strain. Normalization by the internally inferred copy number is conducted by Rbec by default. However, the copy number inferred by Rbec may underestimate the true copy number due to

the potential identical copies of the marker gene in the genome; therefore, in addition, Rbec supports a user-provided copy number table for abundance normalization.

Iterative correction of unique tags

Tags above the *P*-value or *E* threshold are then randomly subsampled (5,000 reads by default) and aligned to the reference sequences in an iterative process. In each iteration, the error matrix is updated with tags corrected during the last iteration. The iterations continue until the number of corrected reads falls below two fixed thresholds, which we set to determine whether the iteration should stop or not. These two thresholds correspond to the absolute and relative differences in the number of corrected reads between the present and previous iterations, and are calculated as follows:

$$N_k - N_{k-1} \leqslant 100$$

$$(N_k - N_{k-1})/N_{k-1} \leqslant 1\%$$

N_k and N_{k-1} denote the number of corrected reads in the k_{th} and $k\text{-}1_{th}$ iteration, respectively. The threshold based on relative differences is used to appropriately stop the iterative process for samples with low sequencing depths, since they can easily satisfy the cut-off based on absolute differences. Once both of these two conditions are met, iterations stop, and each reference sequence is assigned an abundance equal to the aggregated abundance of all its assigned unique tags.

Detection of contamination

Existing error-correction algorithms designed for culture-independent community profiling data cannot accurately estimate the abundances of strains with marker gene paralogs and show a strong bias towards underestimation of their abundances. In addition, paralog sequences are typically classified as sequence variants originating from different strains, resulting in an inflation of alpha-diversity (within-sample diversity). When amplicon sequencing data obtained from synthetic communities is analysed using approaches such as closed OTU-picking, reads from paralog sequences are discarded, leading to low percentages of aligned reads per sample. Samples with a high abundance of strains containing polymorphic paralogous marker sequences can thus be erroneously considered as contaminated. Given that Rbec not only corrects most erroneous amplicon sequencing reads but also successfully identifies paralogous sequences, we can assume that a high proportion of uncorrected reads is likely the result of contamination.

We evaluated the capacity of Rbec to identify contaminated samples by testing it on an *in silico*-simulated dataset, where we included amplicon sequences from *Escherichia coli* with 5% relative abundance per 'contaminated' mock sample. When examining the percentage of reads successfully corrected per sample (Figure 2.5), we observed a clear separation between 'clean' and 'contaminated' mock samples. Based on these results, we included a function in Rbec that can be used to flag potentially contaminated samples and output the amplicon sequences of putative contaminants.

A sample *i* is flagged as contaminated, if

$$R_i < \mu - 1.5IQR$$

Where R_i is the recruitment ratio of reads of sample *i*, μ is the mean of recruitment ratio of reads across all samples in the dataset, and IQR is the interquartile range of the recruitment ratio. If a sequence accounts for more than 3% of total reads after error correction, we assume this sequence originates from a contaminant strain. The accuracy of this heuristic approach depends on the abundance of the contaminant as well as on its prevalence across samples within a dataset. When contamination occurs in the majority of the samples in a dataset, low read requirement ratios would be observed in general, and no individual samples will be flagged. Based on our analyses, samples with less than 90% input reads successfully corrected should be considered as potentially contaminated and further examined. Putative contaminant sequences provided by Rbec can then be used for further check and analysis. If no prevalent contaminant sequences are identified across samples, low percentages of error-corrected reads can be attributed to other technical factors, such as free DNA from the environment or an unusually high level of PCR or sequencing errors.

Construction of accurate reference sequence databases

Analysis of amplicon data derived from SynCom experiments relies on the use of accurate reference sequences derived from each strain. As explained above, Rbec uses this reference database for error correction, identification of polymorphic paralogs, and contaminant detection. Generally, these reference sequences have been previously generated from clonal cultures of individual SynCom members, either by extraction from their respective whole-genome sequences or by Sanger sequencing of the specific marker sequence.

Independently of the method, inaccuracies in the references are likely to occur. For instance, assembly and subsequent extraction of rRNA gene sequences from genomes obtained using short-read technologies are common, and PCR or sequencing errors from targeted

amplification could likewise result in erroneous sequences. Whenever errors are present in the reference sequence, no identical unique tags will be identified during the first step of the Rbec algorithm, and the corresponding strain will be labeled as 'absent' from all samples. Other methods, such as closed OTU picking suffer from similar pitfalls, making this a general problem for SynCom data analysis.

However, errors in the reference database can be corrected by leveraging the feature described above, which allows Rbec to predict potential contaminants. If a reference sequence is inaccurate but its corresponding strain is found in a given SynCom sample with at least a 3% relative abundance, the ratio of corrected reads will decrease and Rbec will output the correct sequence as a putative contaminant. By identifying contaminant sequences with high similarity to reference sequences from strains labeled as 'absent', these inaccuracies can be readily corrected. Once the reference database has been updated, Rbec can be re-run on the same dataset to obtain accurate strain abundances.

Major difference to DADA2

Since we compare Rbec to DADA2 throughout, here we provide an exhaustive description of the differences between Rbec and DADA2.

-Purpose of method: Rbec is the first algorithm exclusively designed for SynComs where there is prior knowledge of the strains existing in the samples, whereas DADA2 is designed for the analysis of natural microbial communities.

-Dependence on reference sequences: Currently, error-correcting methods (DADA2, Unoise, Deblur and AmpliCI) for amplicon sequencing data are designed for natural communities, for which we have a paucity of prior knowledge on which strains are truly present. Rbec has a stringent requirement for accurate reference sequences as input, which also results in an improved capacity in detecting the sequences from low-abundance strains that might be falsely corrected by the error-correcting methods for natural communities. However, recently, a new function has been added to DADA2 to support reference-guided search for low-abundance strains.

-Identification of paralogues: None of the well-established error-correcting methods can deal with paralogues since they are too abundant to be corrected by one of the paralogues from the same strain, so they will be treated as sequences from different strains. However, Rbec does infer the paralogues based on the *E* value of the *Poisson* distribution.

-An expanded error matrix: DADA2 does not take the insertion case into account in the error matrix, but in Rbec, we have expanded the error matrix to include the probability of insertion cases.

Simulation of mock communities

To evaluate the performance of the different algorithms when analysing SynCom data with varying complexities, strain similarities, and sequencing depths, we simulated mock samples using data obtained from sequencing of clonal cultures individually. Firstly, reference sequences of the V5–V7 region from all the bacterial strains were dereplicated, resulting in 114 strains with unique sequences in the V5–V7 region. To generate mock samples with different complexities, 10 to 110 strains from the candidate list containing 114 strains were randomly picked for each mock sample, with a step size of 10 strains. The relative abundance of each strain was simulated using a log normal distribution (s.d. = 2). The total number of reads in each mock sample was fixed at 10,000 reads, and the reads for each strain were subsampled from the amplicon sequencing output of each individual strain using Seqkit (Shen et al. 2016). For instance, to generate a mock community with 3 strains, with relative abundances of 70%, 20%, and 10% respectively, 7,000, 2,000, and 1,000 reads would be sampled from each of the 3 individual amplicon samples and subsequently mixed to generate a mock sample.

To generate mock samples with different strain similarities, we set a maximum pairwise similarity threshold between each pair of strains in each mock community, ranging from 85% to 100%. To alleviate the influence of uneven abundance distribution of each strain on the evaluation, only 20 strains with equal abundances (5% relative abundance for each strain) were included in mock communities. Similarly, to evaluate the impact of different sequencing depths on the performance of the different algorithms, we simulated mock data with 50 strains at different depths, ranging from 500 to 10,000 reads. For each evaluation category, we generated 20 replicates for each parameter combination.

In addition to mock samples of bacterial communities, we also generated fungal mock communities for the purpose of evaluating our algorithm on a different marker gene. Ninety-seven fungal strains with unique ITS1 sequences were set as seeds and randomly chosen to generate the fungal mock communities.

Data processing with different methods

Raw reads were merged using Flash2 (Magoc and Salzberg 2011) with parameters '-m 0.25 -M 250'. Merged reads with ambiguous bases were excluded with USEARCH (Edgar 2010). DADA2 and Deblur plug-ins in QIIME2 (Bolyen et al. 2019) were applied to the filtered data by following the protocols indicated on the QIIME2 website (https://docs.qiime2.org/2019.7/tutorials/), to correct the reads and generate ASV tables. We also included the recently published algorithm AmpliCI (Peng and Dorman 2021) for comparison purposes. We followed the instructions on the corresponding Github website (https://github.com/DormanLab/AmpliCI) to process the data and generate the ASV table. The abundances of ASVs showing exact matches to the reference sequences were extracted from this feature table. For the UNOISE tool (Edgar and Flyvbjerg 2015), filtered reads were dereplicated by USEARCH and subsequently denoised using the -unoise3 function in USEARCH. For the exact match method, the filtered sequencing reads were aligned to the reference database by running the -uparse_ref command in USEARCH. Only hits with 100% identities were retained for generating the profiling table. The Deblur and AmpliCI methods were not applied to fungal data, since the two methods require the equal length of input data, while the length of ITS sequences shows a large variation among strains.

We applied the DADA2, UNOISE, Deblur, exact match, closed OTU picking, and Rbec methods to the simulated data sets. Finally, we calculated the Bray-Curtis dissimilarities between the predicted profiling tables from different methods and the real composition for each mock sample (ground truth) using the *vegan* R package (Oksanen et al. 2020) to evaluate the performance of each algorithm. Significance differences between methods were assessed using a pairwise Wilcoxon test.

We also compared the precision and recall of the two second-best-performing methods, namely, DADA2 and exact match classification with Rbec, using the following formulas:

$$Recall = \frac{No.\ of\ precisely\ corrected\ \text{reads}}{Total\ No.\ of\ reads}$$

$$Precision = \frac{No.of\ precisely\ corrected\ reads}{No.of\ corrected\ reads}$$

Time performance

The CPU time of Rbec on a single sample was tested on an Intel processor (Intel(R) Xeon(R) CPU E5-4657L v2 @ 2.40GHz) with 48 CPUs and 756 GB RAMs. Rbec can analyse 10,000 reads from a SynCom sample comprising 100 strains within 3 minutes using a single CPU.

The comparison of time performance among different methods can be found in online supplementary materials. All the methods were tested on 5 simulated bacterial SynComs containing 100 strains at a depth of 10,000 sequencing reads each using 1 CPU.

Description of output files

strain_table.txt: the strain composition of the sample

strain_table_normalized.txt: the copy-number-normalized strain composition of the sample

contamination_seq.fna: the potential sequences generated by contaminants

rbec.log: percentage of corrected reads, which can be used to predict contaminated samples

paralogue_seq.fna: paralogue sequences found in each strain except for the reference provided

lambda_final.out: the lambda value and *P-value* of the *Poisson* distribution for each unique tag

error_matrix_final.out: the error matrix in the final iteration

Sequencing library construction and preparation

To evaluate the errors in the output from sequencing and test this algorithm, we performed amplicon sequencing on clonal cultures of 236 individual bacterial strains and 97 fungal strains isolated from the roots of *Arabidopsis thaliana* separately (Durán et al. 2018). Genomic DNA was isolated from each strain using the MP Biomedicals FastDNATM Spin Kit for Soil. DNA concentration was determined fluorometrically using the Quant-iT PicoGreen dsDNA Assay Kit (Thermo Fisher Scientific). The V5–V7 region of the *16S* rRNA gene in bacteria and ITS1 region in fungi was amplified using the AACMGGATTAGATACCCKG (799F) and ACGTCATCCCCACCTTCC (1192R) primers, and CTTGGTCATTTAGAGGAAGTAA (ITS1F) and GCTGCGTTCTTCATCGATGC (ITS2R) primers, respectively. Indexing was done using Illumina-barcoded primers. The indexed amplicons were subsequently pooled, purified, and sequenced on the Illumina MiSeq platform.

To exclude the possibility that the observed error distribution of amplicon sequencing is specific to the MiSeq platform, we also analysed the amplicon sequencing data obtained using a HiSeq platform (Guo et al. 2020).

2.6 Author contributions

R.G.-O. and P.Z. conceived the statistical framework of this algorithm. S.S. conducted the amplicon sequencing of bacterial strains. Y.B. and S.H. provided the amplicon sequencing data

from bacterial strains sequenced on a HiSeq platform and fungal strains respectively. P.Z. developed the R package and performed the analysis. R.G.-O. and P.Z. drafted the manuscript.

2.7 Acknowledgements

We would like to acknowledge the useful feedback provided by Prof. Alga Zuccaro, Prof. Eric Kemen, and Dr. Yulong Niu during the development of this algorithm. We also would like to thank Anna Lisa Roth for her help in generating amplicon data from individual strains. This work was Funded by the Max Planck Society and Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC-Nummer 2048/1– project 390686111 and the '2125 DECRyPT' Priority Programme.

2.8 Supporting materials

The supporting material corresponding to this section, including all supplementary tables and figures can be accessed via the online version of the published article (https://www.nature.com/articles/s43705-021-00077-1) and have not been included in this thesis due to space limitations.

Chapter 3

Analysis of amplicon sequencing data from both natural and synthetic microbial communities

3.1 Host preference and invasiveness of commensal bacteria in the *Lotus* and *Arabidopsis* root microbiota

Authors: Kathrin Wippel, Ke Tao, Yulong Niu, Rafal Zgadzaj, Niklas Kiel, Rui Guan, Eik Dahms, Pengfan Zhang, Dorthe B. Jensen, Elke Logemann, Simona Radutoiu, Paul Schulze-Lefert & Ruben Garrido-Oter

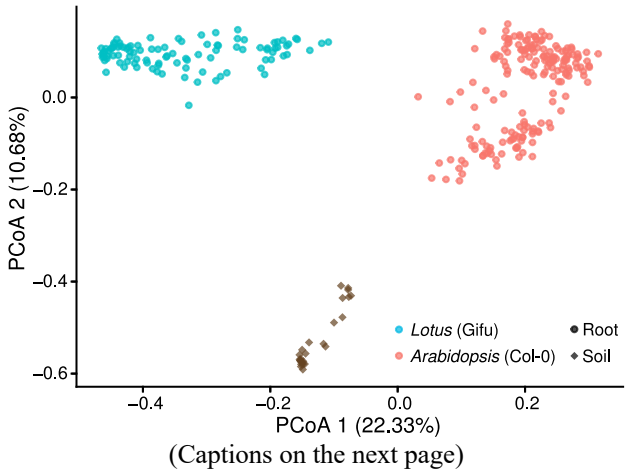
Publisher: *Nature Microbiology*

Own contribution: Genome decontamination and annotation; Amplicon sequencing data analysis

3.1 Host preference and invasiveness of commensal bacteria in the *Lotus* and *Arabidopsis* root microbiota

Statistical analysis shows that host genetics and phylogeny contribute to plant-associated microbiota assemblages (Schlaeppi et al. 2014), representing the co-evolutionary routes of plant-associated microbiota and their hosts. Soil microbes are expected to rapidly evolve at ecological timescales as a consequence of selection by host plants. A recent study (Batstone et al. 2020a) showed that nitrogen-fixing *Ensifer meliloti* can become more beneficial to its original legume host than non-native hosts through continuous experimental evolution, which suggests that a shared evolutionary history with the host plant is important for inducing host-specific microbial evolution. However, at a broader scale, the extent to which microbial communities from different hosts are also selected by corresponding native hosts is unknown. To address this question, we used taxonomically paired SynComs containing bacterial strains isolated from roots of *Arabidopsis thaliana* (*At*) and *Lotus japonicus* (*Lj*), respectively, and performed reciprocal inoculations to examine the growth of bacterial communities collected from different hosts.

We first characterized the variation in bacterial communities in the roots of *At* and *Lj* grown in the same field. A clear separation between the root microbiota of the two host plants was observed and both were distinct from the bulk soil microbiota (Figure 3.1). To understand the mechanism by which distinct microbiota are associated with different host plant species, we established a representative bacterial culture collection containing 3,960 colony forming units (CFUs) from the root and nodules of *Lj*, which accounted for 53% of the cumulative relative abundance of the entire bacterial community in the roots of *Lj*. In total, 294 representative isolates were genome-sequenced. Compared to the bacterial culture collection from the roots of *At*, an extensive taxonomic and functional overlap between the two genome-indexed culture collections was detected.



(Captions on the next page)

Figure 3.1: Differentiation of root microbiota between *At* and *Lj*. PCoA of *Bray-Curtis* dissimilarities of root microbiota from *At* and *Lj*, and corresponding unplanted soil microbiota.

The taxonomic overlap between the bacteria isolated from different hosts allows the evaluation of host specificity. To test host preference, we generated a mixed SynCom including 32 representative bacteria from families common to both culture collections, with 16 strains from each. After inoculating the mixed SynCom on either *Lj* or *At*, the community composition of the root microbiota was profiled with 16S amplicon sequencing. As observed in the natural bacterial communities in the roots of the two hosts, the composition of input SynCom shifted into two clusters according to host species (Figure 3.2A). Interestingly, the aggregated abundance of native strains in the mixed SynCom was higher when colonizing the cognate host than the non-cognate host (Figure 3.2B). Within the same host, the SynCom was also dominated by the native strains (Figure 3.2B). These results reveal a clear host preference of root commensal bacteria. However, mono-association experiments showed that individual strains did not maintain the host preference observed in the community context. Moreover, the native strains were incapable of providing a growth benefit to the cognate host relative to the non-cognate host. The community shift in the root microbiota and host preference was preserved in two additional plant species, *Lotus corniculatus* and *Arabidopsis lyrate*, suggesting that the trait of host preference is induced by conserved root niches in the corresponding plant lineages. Given the host preference, we designed an experiment to test the ecological theory of priority effects. We inoculated the *At*- or *Lj*-derived SynCom onto either *At* or *Lj* and then allowed invasion by the other SynCom. Intriguingly, we found that the later-arriving SynCom grew better in the root compartment of the cognate host than the non-cognate host but not in the rhizosphere (Figure 3.2C and 3.2D). This result suggests that host preference can reduce the impact of priority effects on community assemblage.

Next, we sought to understand the mechanisms driving host preference. The community phenotype of host preference was maintained in plants impaired in immunity and glucosinolate biosynthesis. However, host preference was lost upon growth in root exudates, implying that root exudates could not induce host preference and that physical contact with live roots might be required for host preference. As expected, a loss of host preference was observed when the mixed SynCom grew on dead roots. Further, we compared the transcriptomic profiles of plants inoculated with native or non-native SynComs to identify potential genes regulating host preference. *K*-means clustering separated the transcriptomes from samples treated with native or non-native SynComs (Figure 3.3). Interestingly, several

immunity-related genes were specifically up-regulated by native SynComs in both host species.

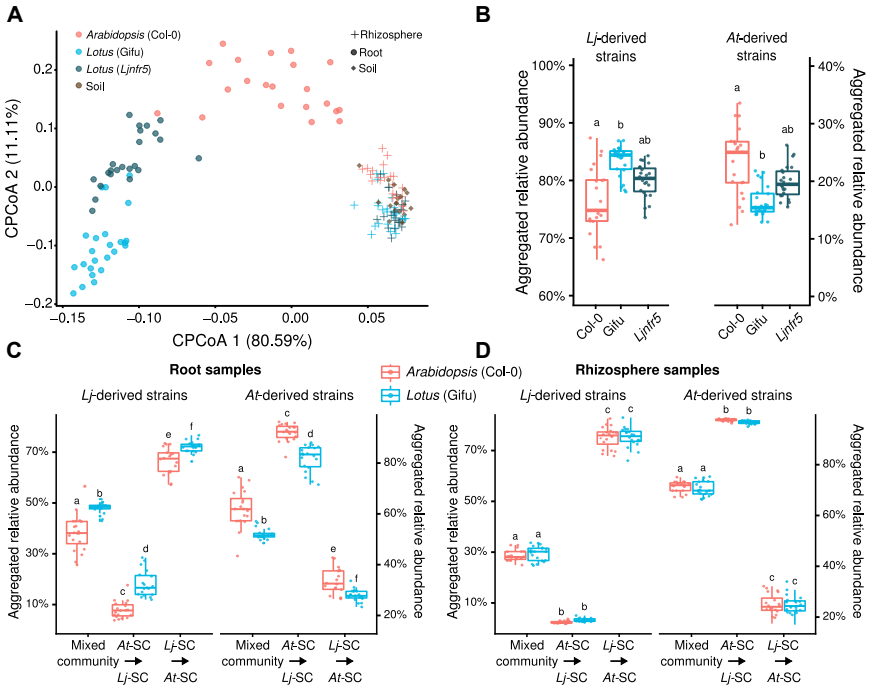


Figure 3.2: Host preference and invasion of commensal microbes in cognate hosts. (A) *Lj* wild-type *Gifu*, *nfr5* mutant and *At* wild-type Col-0 plants co-cultivated with the mixed community *LjAt-SC1* (exp. B, n = 155, variance explained 53.8%, P = 0.001). (B) Aggregated RA of the 16 *Lj*-derived and the 16 *At*-derived strains in roots of *At* and *Lj*. (C, D) Aggregated RA of the 16 *Lj*-derived and the 16 *At*-derived strains in Lj and *At* roots (C) (n = 120) and rhizosphere (D) (n = 120) samples in the indicated treatments. Different letters above boxes indicate different significance groups according to a Kruskal-Wallis test, followed by a Dunn's post hoc.

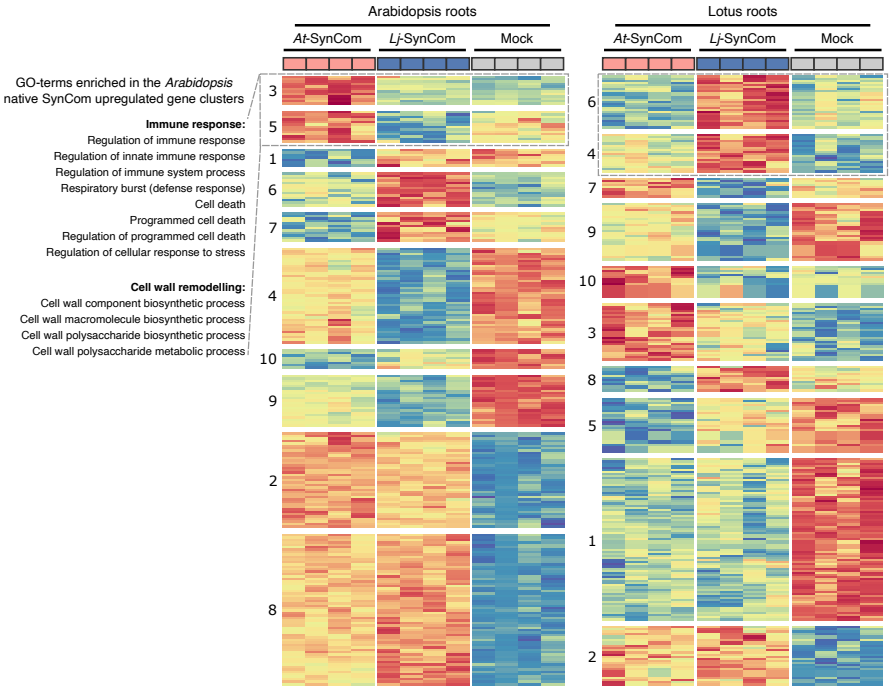


Figure 3.3: Transcriptomic analysis of root responses to different native or non-native SynComs. Heatmaps showing scaled counts of genes arranged according to k-means clustering results (only differentially expressed genes shown) for *At* and *Lj*.

Host preference has been widely observed in pathogenic and mutualistic microbes (Mallott and Amato 2021; Pérez-Carrascal et al. 2022). In our experiments, we could show that host preference can also be maintained at the community level, though that phenotype could not be perceived in individual tested strains. Moreover, the phenotype of host preference could also be detected in two additional hosts that are from the same genera as the native hosts, suggesting that hosts from the same genera can partially form conserved niches for commensal bacteria. Co-habitation of indigenous microbes and plants can induce rapid evolution of microbes, which can potentially facilitate bacterial mutualism that increases bacterial growth and plant fitness (Batstone et al. 2020a; Li et al. 2021). However, our results showed that host preference exhibited by native SynComs could not provide a growth benefit to plants, indicating this is not a consequence of co-adaptation of microbiota and the host. Our experiments were performed in a rich medium where plants can grow quite well and we expect that the benefits to the host provided by the native SynCom may be more apparent in stressful environments, which requires further well-designed verification experiments. We showed that host preference reduced the impact of priority effects on community assembly by allowing the invasion of the native SynCom into the standing non-native SynCom in the roots of the corresponding cognate host, which can be promisingly transformed in fields of agriculture and medicine.

3.2 Shared features and reciprocal complementation of the *Chlamydomonas* and *Arabidopsis* microbiota

Authors: Paloma Durán, José Flores-Uribe, Kathrin Wippel, Pengfan Zhang, Rui Guan, Barbara Melkonian, Michael Melkonian & Ruben Garrido-Oter

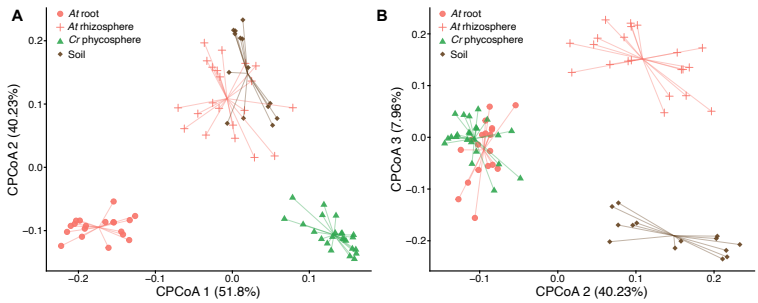
Publisher: *Nature Communications*

Own contribution: Genome decontamination; Amplicon sequencing data analysis

3.2 Shared features and reciprocal complementation of the *Chlamydomonas* and *Arabidopsis* microbiota

Land plants possess the ability to form intimate relationships with the surrounding soil microbiota via the exchange of secreted metabolites, a property which might date back to the colonization of the first land plants (Lambers et al. 2009). In aquatic environments, algae are also known to be colonized by complex microbial communities termed phycosphere microbiota (Amin et al. 2015; Kim et al. 2014; Seymour et al. 2017). However, it's still unclear if the algal ancestors of land plants could already recruit a subset of diverse microbes from the surrounding soil, which is a common trait of embryophytes. To address this question, we characterized the microbiota of a selection of taxonomically diverse sub-aerial green algae, including the model organism *Chlamydomonas reinhardtii* (*Cr*).

By inoculating axenic *Cr* into pots containing CAS soil, we found a clear separation between phycosphere and soil microbiota, which was also captured in the root microbiota of *At* (Figure 3.4A). Though we found that the microbial compositions of the phycopshere of *Cr* and the roots of *At* were distinct on the first two PCoA axes, further inspection of the second and third axes revealed an overlap between the two microbial compositions (Figure 3.4B). Importantly, a significant overlap of abundant OTUs (≥0.1% relative abundance) between phycosphere and root compartment was observed. These results suggest that *Cr*, similarly to higher land plants, can recruit microbes from the species pool in the surrounding soil. To test whether the overlapping pattern of microbial composition in the phycosphere of *Cr* and roots of *At* can be extended to other land plant lineages, we performed a meta-analysis by including the amplicon sequencing data from the roots of plant species spanning lycopods, ferns, gymnosperms, and angiosperms (Yeoh et al. 2017). Six orders encompassing *Caulobacterales*, *Rhizobiales*, *Sphingomonadales*, *Burkholderiales*, *Xanthomonadales* and *Chitinophagales* consistently colonized the roots of every plant species and algae, accounting for 39% of the total bacteria in their respective communities on average (Figures 3.5). This conserved pattern suggests that the ability to form associations between bacteria from the six orders and photosynthetic hosts is potentially inherited from the ancestor of land plants.



(Figure in the previous page) **Figure 3.4: The assembly of phycosphere microbiota from soil.** PCoA of Bray-Curtis dissimilarities (ASV-level) constrained by compartment (22.4% of variance explained; $P < 0.001$).

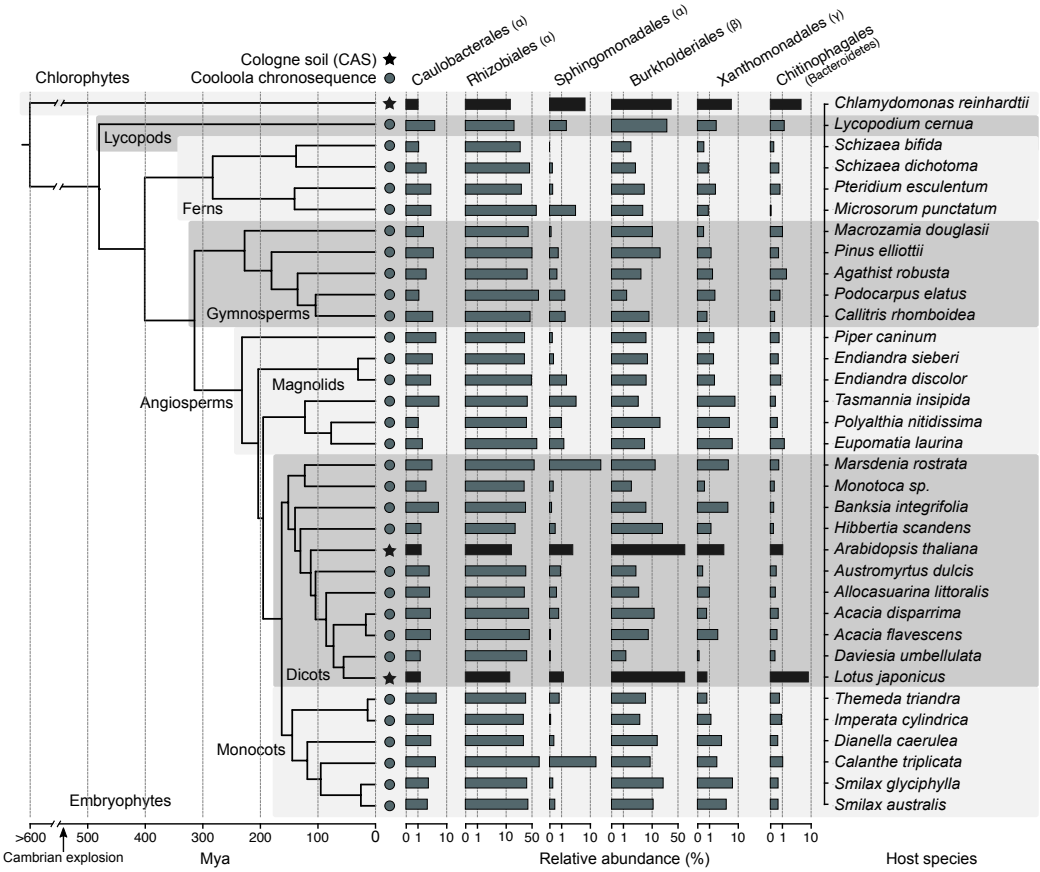


Figure 3.5: Conserved features in the root-associated microbiota across the phylogeny of photosynthetic hosts. Phylogeny was inferred from a multiple sequence alignment of the ribulose-bisphosphate carboxylase gene (*rbcL*) of 35 plant species and *Cr*. The bar plots represent the average aggregated relative abundance of the six bacterial orders found to be present in the root microbiota of each plant species (80% occupancy and ≥0.1% average relative abundance). Leaf nodes depicted with a star symbol denote community profiles of plants grown in CAS soil in the greenhouse, whereas those marked with a circle were obtained from plants sampled at the Cooloola natural site chronosequence.

To further study the assemblage of phycosphere microbiota and alga-microbe interactions in a controlled environment in which we can monitor and perturb environmental factors and manipulate microbial community compositions, we established a mesocosm system using microbial extracts from soil as starting inocula. *Cr* is co-inoculated with microbes in carbon-free medium such that microbes can only consume the carbons photosynthesized and secreted by *Cr*. We found that *Cr* was able to re-wire the soil microbiota within the first four days in this system (Figure 3.6A). Importantly, cultivation of soil microbiota in the absence of carbon sources or upon supplementation with artificial photosynthates led to distinct microbial compositions compared to those that ensued upon co-cultivation with live and metabolically active *Cr* (Figure 3.6A**).** In order to perform targeted manipulation of microbial communities,

185 genome-indexed bacterial isolates, spanning 42 species, were collected from the phycosphere of *Cr*,

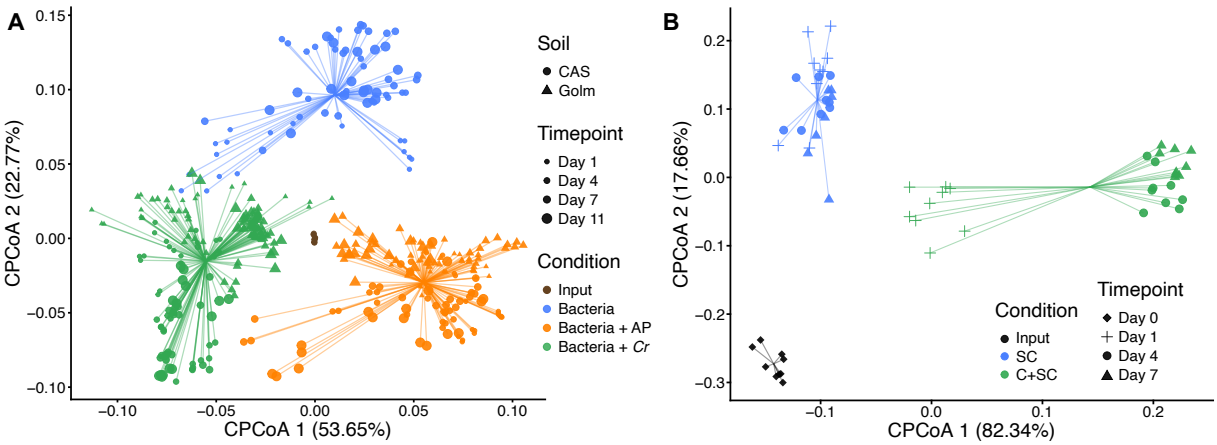


Figure 3.6: Establishment of phycosphere mesocosm to study algae-microbiota interactions. (A) PCoA analysis of Bray-Curtis dissimilarities of bacterial community profiles (ASV-level), constrained by condition (17.9% of the variance; $P < 0.001$), showing a significant separation between starting inocula (soil washes, depicted in brown; n = 12), phycosphere communities (green; n = 158), and soil washes incubated in minimal media (blue; n = 112), or media supplemented with artificial photoassimilates (APs, depicted in orange; n = 144). (B) Strain-level beta-diversity analysis (CPCoA of Bray-Curtis dissimilarities; 40.4% of the variance; $P < 0.001$) of bacterial communities from samples obtained from a liquid-based gnotobiotic system. Samples are color-coded based on the experimental condition: input SynCom samples (black; n = 9), synthetic phycospheres (light green; n = 27), and SynCom-only controls (blue; n = 25).

accounting for 63% of relative abundance of the entire phycosphere bacterial community. We generated a 26-member SynCom and co-inoculated the SynCom with *Cr* in the mesocosm system. The community compositions of SynCom shifted and saturated in four days compared to the starting inocula and SynCom alone (Figure 3.6B), demonstrating that our gnotobiotic system can recapitulate the impact of *Cr* on the soil microbiota found by the culture-independent method. We also established a gnotobiotic system to test the impact of physical proximity on phycosphere microbiota assembly. In this system, two growth chambers are connected through a membrane that only allows diffusion of compounds but not the passage of bacterial or algal cells. By cultivating SynComs and *Cr* on separate sides or on the same side of growth chambers, we found that SynComs showed distinct compositions, suggesting that physical contact with the host is indispensable for the establishment of the phycosphere microbiota (Figure 3.7).

Given the features shared by bacterial communities from the phycosphere and root microbiota, we hypothesized that SynComs derived from cognate hosts can assemble similar communities on the other host because of the functional similarities of bacteria isolated from *Cr* and *At*. We generated taxonomically paired 9-member SynComs for *Cr* and *At*, in which

bacteria from families shared between *Cr* and *At* culture collections were selected. A mixed SynCom was also generated to test host preference. As expected, SynComs inoculated on the same host showed similar compositions at the family level irrespective of their isolation origin together with a retention of host preference (Figure 3.8).

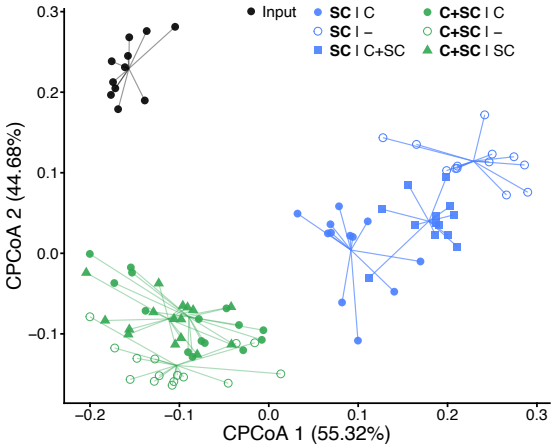


Figure 3.7: Physical proximity to Cr is required for the establishment of phycosphere bacterial communities. Strain-level beta-diversity analyses of Bray-Curtis dissimilarities of bacterial SynComs (SC, n = 37), and synthetic phycospheres (SC + C, n = 46), grown in a split gnotobiotic system. Constrained PCoA is shown for all samples (21% of variance; P < 0.001).

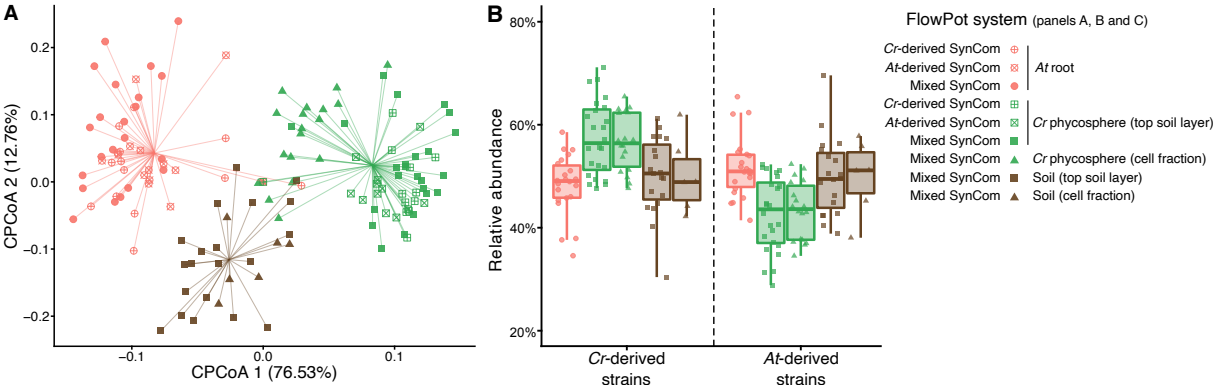


Figure 3.8: Root and phycosphere bacteria colonize *At* and *Cr* and assemble into taxonomically equivalent communities. (A) Strain-level beta diversity analysis of soil (n = 26), root (n = 57), and phycosphere (n = 66) bacterial community profiles, from gnotobiotic *At* and *Cr*, inoculated with bacterial SynComs derived from *At* roots (*At*-SPHERE), *Cr* (*Cr*-SPHERE), or mixed (*At*- and *Cr*-SPHERE), grown in the FlowPot system. (B) Aggregated relative abundances of *At*- and *Cr*-derived strains in the mixed SynCom (n = 149).

Microalgae-microbe interactions have been extensively studied in aquatic environments owing to the importance of phytoplankton-microbe interactions in carbon and energy fluxes in aquatic environments, especially in marine settings (Horňák et al. 2017). In this study, we tried to extend these insights to terrestrial systems. We showed that algae could recruit a subset of microbes from the surrounding soil to colonize the phycosphere, which is similar to the process of the assembly of the root microbiota in land plants. We found that six abundant bacterial groups were consistently present both in phycosphere and root microbiota, though

differences in microbiota composition were found. This suggests that adaptations to photosynthetic hosts of those bacteria predate the diversification of land plants and imply the presence of shared principles of microbiota establishment between phycosphere and root microbiota.

In aquatic environments, algae release photosynthesized carbon and other forms of chemicals to the surrounding environment, which constitutes a niche for heterotrophic bacteria (Seymour et al. 2017). Bacteria compete for nutrients and form complex phycosphere microbiota. To systematically study the assembly and functionality of phycosphere microbiota, we established a gnotobiotic system in which environmental factors and microbial compositions can be manipulated simultaneously. By taking advantage of this mesocosm, we found that physical contact of bacteria with living algae was required for the formation of phycosphere microbiota and that growing bacteria with algal metabolites was not sufficient for establishment of the normal phycosphere microbiota. This observation runs counter to that of a recent study where the authors showed that phycosphere microbiota composition could be partially predicted by the phytoplankton metabolites (Fu et al. 2020). The shared features between the phycosphere and root microbiota, the mesocosm we have established to study algae-microbiota interactions, and the short generation time of algae, will, taken together, allow us to further study the potential molecular and ecological principles governing soil microbiota colonization of photosynthetic hosts and the co-evolution of microbiota and hosts.

3.3 Differential Impact of Plant Secondary Metabolites on the Soil Microbiota

Authors: Vadim Schütz, Katharina Frindte, Jiaxin Cui, Pengfan Zhang, Stéphane Hacquard, Paul Schulze-Lefert, Claudia Knief, Margot Schulz & Peter Dörmann

Publisher: *Frontiers in Microbiology*

Own contribution: Amplicon sequencing data analysis

3.3 Differential Impact of Plant Secondary Metabolites on the Soil Microbiota

The roots of land plants recruit a subset of the soil microbiota to form distinct root microbiota. Root exudates are one of the key factors that induce specific bacteria to migrate to roots from soil via chemotaxis mechanisms (Feng et al. 2021). Cumulative evidence has shown that plant metabolites can modulate the soil microbiota to form complex root microbiota (Hu et al. 2018; Kudjordjie et al. 2019). For example, salicylic acid secreted from willow tree roots alters microbial composition in the rhizosphere (Schmidt et al. 2000). Given the extensive diversity of root exudates from a single plant, systematic analysis of the impact of each chemical on microbial composition is labor-intensive and time-consuming. In this work, we specifically focused on three plant metabolites, benzoxazolinone (BOA), gramine and quercetin, and described their impacts on root microbiota establishment. BOA and gramine are indole-derived metabolites that are abundant in the root exudates of *Poaceae*, being produced in a mutually exclusive way by these species. Quercetin is from the flavonoid group and is one of the most abundant flavonoids in root exudates (Mathesius 2018).

We inoculated CAS soil in pots with the addition of one of the three metabolites, adhering to the concentrations that occur in natural conditions. The addition of metabolites was performed every other day to keep their concentrations stable in soil over the duration of the experiments, which lasted 28 days. HPLC analysis revealed that the metabolites were markedly decreased in soil after two days. However, by immediately extracting chemicals after mixing the metabolites with soil, we found that the decrease in gramine and quercetin was actually caused by firm binding to soil particles, which inhibits efficient extraction of the two metabolites.

To examine the impacts of different metabolites on soil microbiota, we first evaluated the content of phospholipid fatty acid (PLFA) with GC-MS. No significant differences in the total amount of PLFA were found among any of the treatments and soil control, suggesting that addition of metabolites did not influence the total microbial biomass (Figure 3.9). However, we found that different types of PLFA were enriched in different treatments compared to the control. Intriguingly, the amount of the cyclopropane 17:0 cyclo acid increased upon treatment, indicating that the bacteria were subject to stress after metabolite supplementation (Figure 3.9). The fatty acids typically found in fungi and algae or cyanobacteria were almost absent in soil, indicating that the agricultural soil had a low amount of fungal biomass and was dominated by bacteria. This finding prompted us to examine how the soil bacterial

community shifted after exposure to the different metabolites. The PCoA plot revealed a clear separation of microbial composition among different treatments independent of time points (Figure 3.10), suggesting differential impacts of plant metabolites on the soil microbiota. Different ASVs were found to be enriched or depleted in response to different treatments compared to the control. We further attempted to isolate bacteria from different treatments along with the soil control. In line with the differences in microbial composition among treatments, some groups of bacterial isolates were exclusively found in one of the treatments and not in the others.

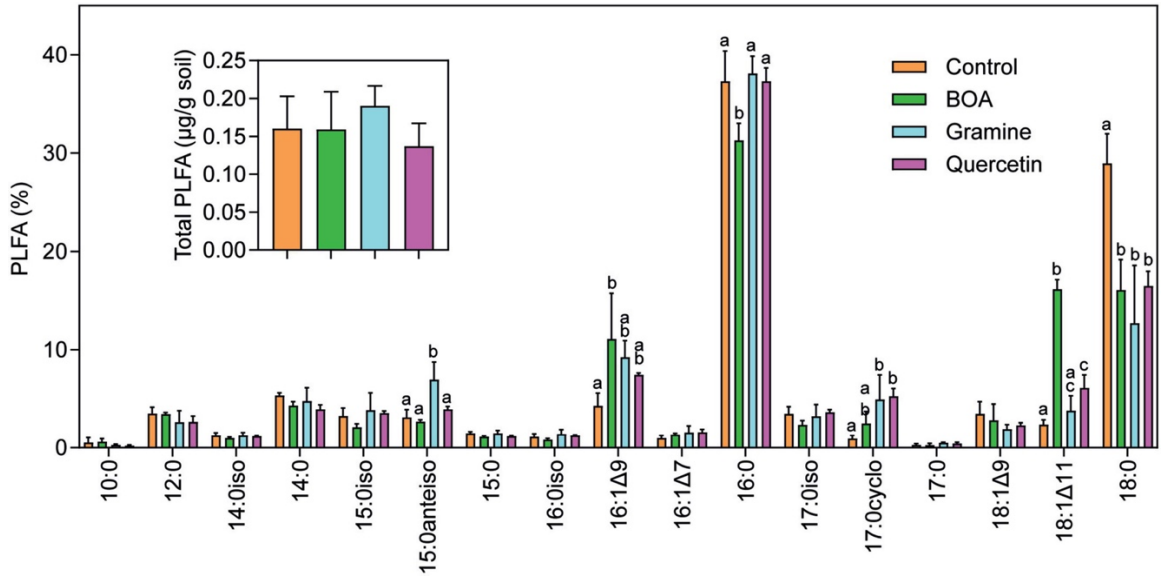


Figure 3.9: Phospholipid fatty acid (PLFA) analysis of soil samples. Soil samples harvested after 28 days of treatment with BOA, gramine or quercetin were extracted and phospholipid fatty acids determined by GC-MS. The inset shows total PLFA in µg/g of dried soil. 10:0, decanoic acid; 12:0, lauric acid; 14:0iso, 11-methyl-tridecanoic acid; 15:0iso, 13-methyl-myristic acid; 15:0anteiso, 12-methyl-myristic acid; 16:0iso, 13-methyl-pentadecanoic acid; 16:1Δ9, palmitoleic acid; 16:1Δ7, Δ7-hexadecenoic acid; 16:0, palmitic acid; 17:0iso, 15-methyl-palmitic acid; 17:0cyclo, 7,8-cyclopropane-palmitic acid; 18:1Δ9, oleic acid; 18:1Δ11, vaccenic acid; 18:0, stearic acid. (ANOVA, post hoc Tukey; $p < 0.05$; $n = 3$; mean ± SD; different letters indicate significant differences).

The establishment of the root microbiota is a consequence of the soil microbiota interacting with root exudates and the plant immune system. In the present study, we found that different bacteria can be enriched in soil by different plant metabolites leading to the formation of differentiable microbial communities; therefore, the assembly of the root microbiota could be regulated by the additive control of different soil microbes by different root exudates. However, due to the lack of fungal biomass in CAS soil, it was not possible to evaluate the impacts of plant metabolites on fungal growth and community composition in our current work. Additional soil conditions could be tested to address the shortcomings of the current

study and to determine whether impacts of plant metabolites on soil microbiota follow conserved patterns, independently of edaphic factors.

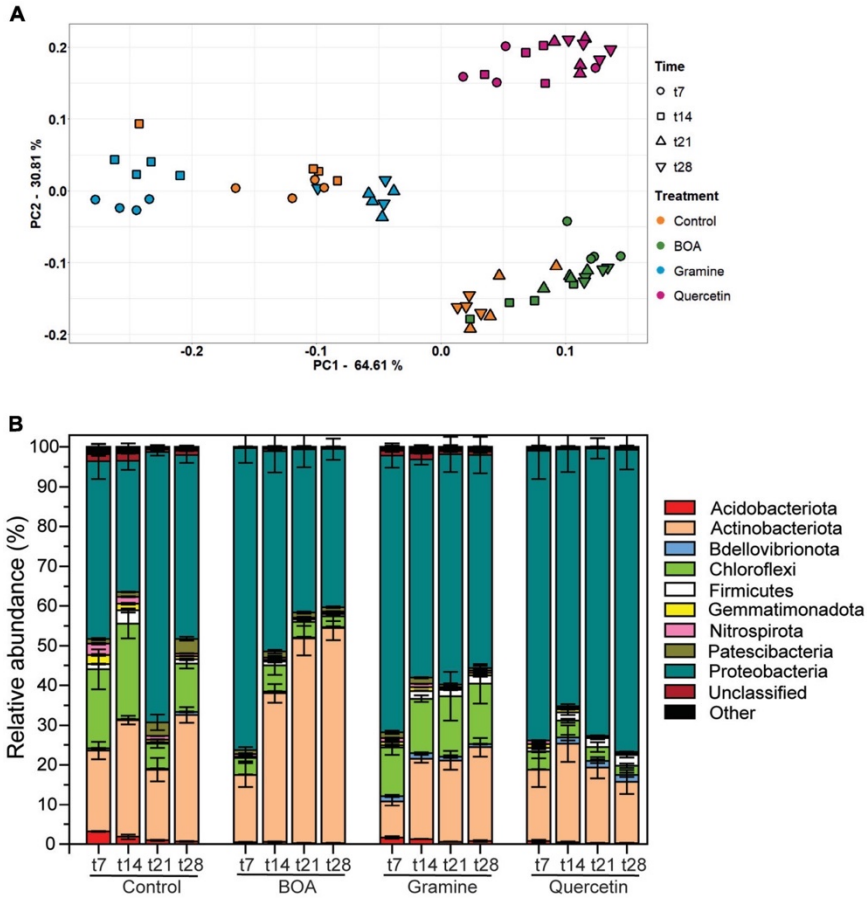


Figure 3.10: Changes in the soil bacterial community structure after treatment with BOA, gramine or quercetin. (A) Differences in bacterial community structure between samples are illustrated in principle component (PCA) plots. The time points are shown by different symbols, and the color code depicts the different treatments. Each measurement is represented by four replicates. (B) Relative abundance of bacterial phyla in the soil of the control or after treatment with BOA, gramine or quercetin. Low abundant groups with <2% of the total reads are summarized as "Other". (n = 4, mean ± SD).

3.4 Co-functioning of bacterial exometabolites drives root microbiota establishment

Authors: Felix Getzke, Amine Hassani, Max Crüsemann, Milena Malisic, Pengfan Zhang, Yuji Ishigaki, Nils Böhringer, Alicia Jiménez Fernández, Lei Wang, Jana Ordon, Ka-Wai Ma, Hidde Wesseler, Shingo Miyauchi, Ruben Garrido-Oter, Ken Shirasu, Till Schäberle, Stéphane Hacquard & Paul Schulze-Lefert

Publisher: *Under review*

Own contributions: BGC prediction for At-Sphere culture collections and *Pseudomonas* genomes; Amplicon sequencing data analysis

3.4 Co-functioning of bacterial exometabolites drives root microbiota establishment

The establishment of complex microbial communities in host environments is driven by multiple factors, including host metabolites, environmental factors, and interspecies interactions between microbes (Carlström et al. 2019; Hu et al. 2018; Kudjordjie et al. 2019; Schlaeppi et al. 2014). A full understanding of assembly rules can aid in the manipulation of microbial communities for sustainable agriculture and human health. Plant roots harbor phylogenetically diverse microbes, which are largely recruited from the species pool of the soil microbiota (Bulgarelli et al. 2012). These microbes provide different forms of benefits to hosts encompassing nutrition mobilization, pathogen protection and tolerance to abiotic stresses (Castrillo et al. 2017; Hou et al. 2021; Vogel et al. 2021). Recent studies showed that many soil-derived microbes actually compete with each other in rich medium (Palmer and Foster 2022), potentially via a wide variety of inhibitory exometabolites with antimicrobial functions. However, how antagonistic activity between microbes contributes to microbial community establishment in the root compartment remains elusive. We used a binary interaction assay, mass spectrum analysis, genome mining, and experimental validation to address this question.

We tested 39,204 binary interbacterial interactions including 198 strains isolated from the roots of *At* via plate-based assays, in which 1,011 inhibitory interactions were observed involving 66% of strains. *Actinobacteria* isolates were most sensitive to all other classes, especially to *Gammaproteobacteria* (Figure 3.11A). Strains (R569, R9, R562, R401, R329, R71 and R68) from *Pseudomonadaceae* showed broad inhibitory activity towards phylogenetically diverse strains (Figure 3.11B). Interestingly, closely related strains also exhibited varied inhibitory spectrums, suggesting that differences in accessory gene repertoires among close bacteria could be important for the production of diverse bacterial inhibitory exometabolites in root-associated bacteria. Given the prevalent inhibitory activity of rhizobacteria, we hypothesized that rhizobacteria isolated from healthy plants can contribute to the suppression of plant pathogens. We found that 10.9% of tested strains, mainly from *Pseudomonas*, *Streptomyces* and *Bacillus*, inhibited pathogenic *Rs* GMI1000 (Figure 3.11B).

Inhibitory exometabolites are often produced by biosynthetic gene clusters (BGCs) (Crits-Christoph et al. 2018). We systematically examined the BGCs in our root-derived strains and

observed that strains with higher inhibitory activity possessed significantly higher numbers of BGCs and BGC families than the average (Figure 3.12). Using UPLC-MS, we observed 224 new nodes that were only secreted and detected upon interaction with other isolates compared to the exometabolite profile of individually grown bacteria. Dereplication of the interaction metabolome revealed that the biosynthesis of several polyketides was exclusively

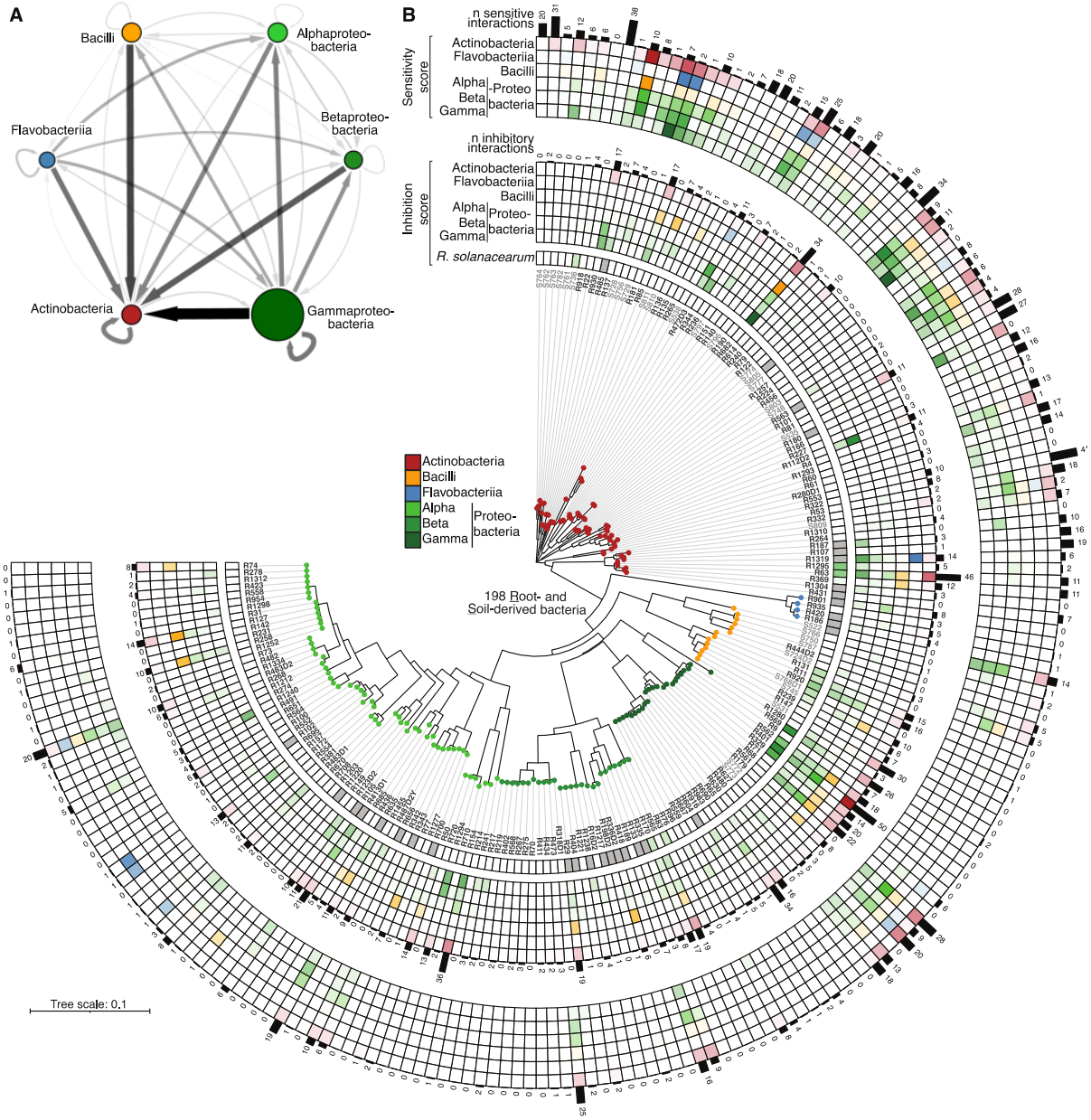


Figure 3.11: Widespread production of specialised exometabolites among root-associated bacteria. (A) Inhibitory interaction network computed based on binary interaction data from a modified Burkholder assay (mBA; n = 198 strains, 39,204 pairwise combinations tested, 1–2 biological replicates). Edge width and color depicts the aggregated frequency of inhibitions at the class level while node size indicates the mean halo size, measured at 4 days post-inoculation (dpi). Arrows reflect inhibition directionality and intensity. (B) Phylogenetic tree showing inhibition and sensitivity scores for each strain. The tree was built based on the full-length 16S rRNA gene sequences of 167 and root- (dark grey) and 31 soil-derived (light grey) bacteria. All

bacteria were reciprocally screened against each other and against *Ralstonia solanacearum* GMI1000. For each strain, sensitivity scores (average sensitivity of a strain to exometabolites produced by a given bacterial class) and inhibition scores (average exometabolite-dependent inhibition of a given strain to a bacterial class), are shown in the outer and inner heatmap, respectively. These scores represent the average halo size for a given strain at the class level and were determined by a mBA. The number of sensitivities per strain and the number of inhibitory interactions are indicated by black bars.

induced by interactions with other bacterial strains, pointing to a possible role of these molecules in root microbiota establishment and pathogen suppression in natural environments.

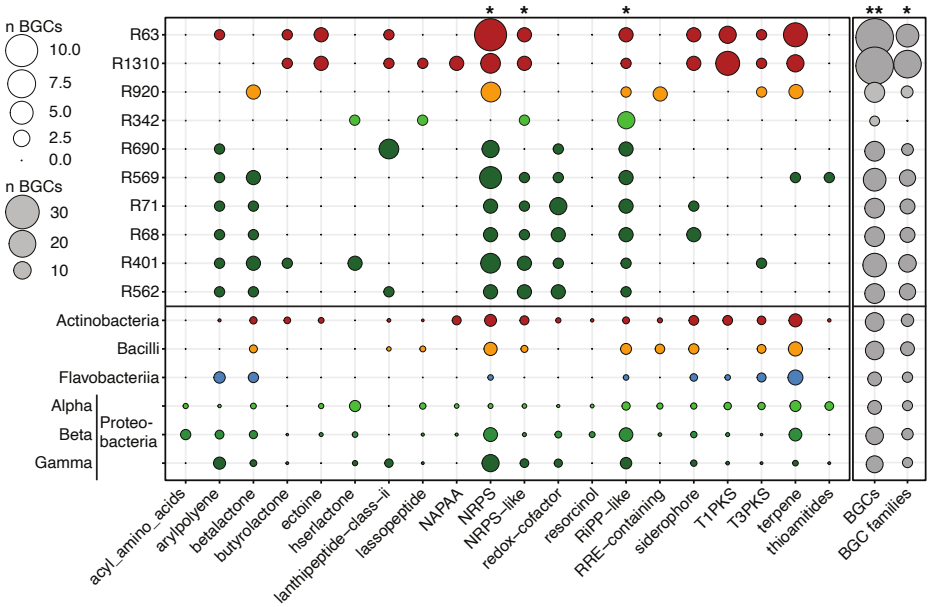


Figure 3.12: Genomic capacity for specialised metabolite production explains pronounced inhibitory activity. Balloon plot depicting the genetic potential for specialized metabolite production. Using antiSMASH 6.0, we predicted BGCs for the genomes of all analysed bacteria. On the x-axis, the 20 most abundant BGC families across all tested bacterial taxa are depicted. Shown are the number of different BGC families within the 10 strains with the highest halo production capabilities, based on **Figure 3.11**. Furthermore, average numbers for all six bacterial classes are depicted. Spheres are colored by bacterial classes. In grey, the total number of BGCs and BGC families is depicted. The full data set can be found in Table 2. Statistical significance was determined by Kruskal-Wallis followed by Dunn's post-hoc test and BH adjustment. Significance compared to WT is indicated by black asterisks (∗, ∗∗, indicate $p < 0.05$, and 0.01, respectively).

Next, we sought to identify the mechanisms underpinning the strong inhibitory activity of R401, which exhibited the greatest inhibitory interactions and the largest average halo size across all the tested strains. We found that the gene *phlD*, which controls the biosynthesis of 2,4-diacetylphloroglucinol (DAPG), partially contributed to the inhibitory activity of R401 towards *Rs* (Figure 3.13A and 3.13B). Using the mini-Tn5 transposon insertion library of R401, we identified another two genes, *pvdY* and *pvdL*, involved in the biosynthesis of the siderophore pyoverdine, to be responsible for the suppression of *Rs* both in liquid medium and on solid agar

by chelating iron (Figure 3.13C and 3D). Double mutants of R401 deficient in the synthesis of both metabolites showed a pronounced reduction (70%) of inhibitory activity against *Rs* and other tested strains compared to single mutants (Figure 3.13E), suggesting an additive effect of DAPG and pyoverdine.

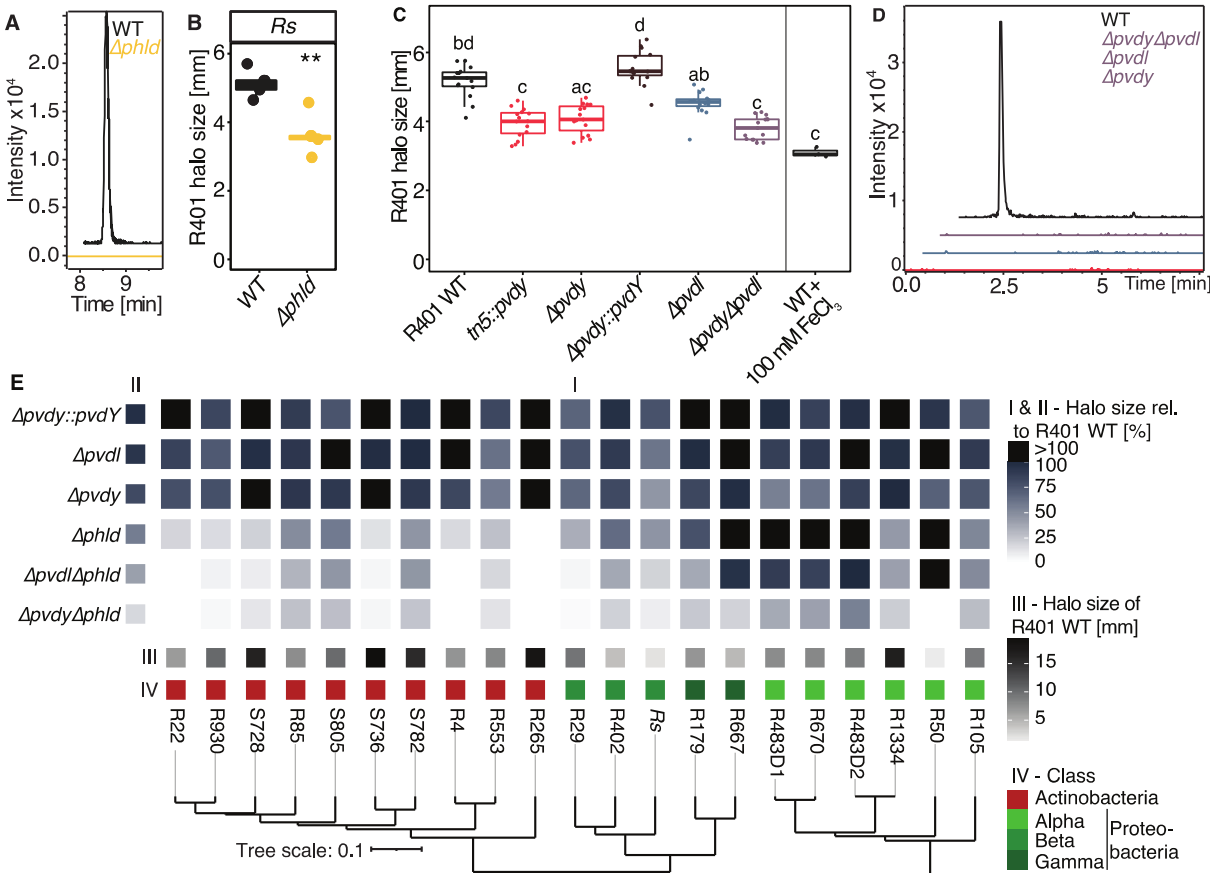


Figure 3.13: DAPG and pyoverdine act additively to inhibit taxonomically distinct root microbiota members. (A) Extracted ion chromatograms for R401 DAPG (EICs: 211.0601 m/z ± 0.01 [M+2H]+) of the WT and mutant extracts, confirming complete lack of DAPG production in the tested mutant. (B) Halo production of R401 WT and Δ*phld* using *Rs* as target bacterium as measured in mBA. Statistical significance was determined by Kruskal-Wallis followed by Dunn's post-hoc test and Benjamini-Hochberg (BH) adjustment. Significance compared to WT is indicated by black asterisks (** indicate $p < 0.01$; n=5). (C) Halo production of R401 WT and four different mutants that are impaired in the production of pyoverdine (*tn5::pvdy*, *Δpvdy*, *Δpvdl*, *ΔpvdyΔpvdl*) as measured in mBA. Mutant names and colors are depicted as in panel C. Δ*pvdy::pvdY* is a complementation line of Δ*pvdy*. Halo production of R401 WT strains after medium supplementation with 100 µM FeCl3. *Rs* was used as a target strain. Halo size measurements were taken after 3 days of interaction. Letters indicate statistically significant differences as determined by Kruskal-Wallis followed by Dunn's post-hoc test and BH adjustment with $p < 0.05$ (n=15). (D) Extracted ion chromatograms for the R401 Dihydropyoverdine (EICs: 622.2764 m/z ± 0.1 [M+2H]2+) of the WT and mutant extracts, confirming complete lack of production in all tested mutants. (E) Heatmap depicting a halo of mBA screen of R401 WT and single and double mutants that are impaired in DAPG (Δ*phld*) and/or pyoverdine (Δ*pvdl* or Δ*pvdy*) production. Δ*pvdy::pvdY* is a Δ*pvdy* complementation line. A taxonomically diverse set of root- and soil-derived bacteria (comprising *Rs*) has been

used as target bacteria. Halo sizes have been normalized to the respective WT-halo sizes. Average, relative halo sizes are depicted in (I). (II) shows the average thereof across all tested strains, while (III) shows the average absolute halo size of R401 WT on a given target strain; n=5. (IV), is a phylogenetic tree based on v5v7 16S rRNA genes, depicting strain taxonomy at the class level.

Given that R401 exerted strong antagonistic activity against a wide variety of strains, we speculated that it may play an important role in mediating the establishment of root microbiota. By inoculating an 18-member SynCom and wild-type/mutant R401 onto the roots of axenic *At*, we found an increased alpha diversity and altered microbial composition in SynComs treated with single/double mutants of R401 compared to the wild-type (Figure 3.14A and 3.14B). The shift in microbial composition was not observed in soil condition where no *At* was growing (Figure 3.14C and 3.14D), indicating a niche-specific function of DAPG and pyoverdine. In line with results from binary interaction assays, strains sensitive to R401 were depleted in the SynCom treated with the wild-type R401 (Figure 3.14E). An increased relative abundance of wild-type R401 in root compartments was observed in the SynCom context, but the accumulation was gradually reduced for single and double mutants (Figure 3.15A), which was not a consequence of impaired colonization of *At* roots by mutants *per se* (Figure 3.15C). However, the accumulation was not found in soil compartments (Figure 3.15B). These results indicate that its antagonistic ability provides R401 with root competence advantages.

In contrast to the root compartment that is rich in nutrients continuously secreted by plants, unplanted soils are oligotrophic habitats for microbiota (Eilers et al. 2010). Biosynthesis of secondary metabolites in organisms is energetically costly (Wink 2016). This potentially explained the loss of antagonistic activity of R401 towards other SynCom members in soil compartments. The sequential reduction in alpha diversity from soil to rhizosphere to root microbiota has been reported across multiple hosts (Bulgarelli et al. 2012; Lundberg et al. 2012; Thiergart et al. 2019), which can be explained by the selection of the plant immune system and root exudates either with antimicrobial functions or exclusively assimilated by a subset of soil microbes. Our results highlight that interspecies antagonism can also contribute to the reduction of alpha diversity in root compartments. Moreover, the shift of microbial composition in root compartments can also be forecasted from binary interaction assays. This might have relevance for future interventions in the root microbiota with rational biologicals that provide benefits to the host, including indirect pathogen protection and mineral nutrition.

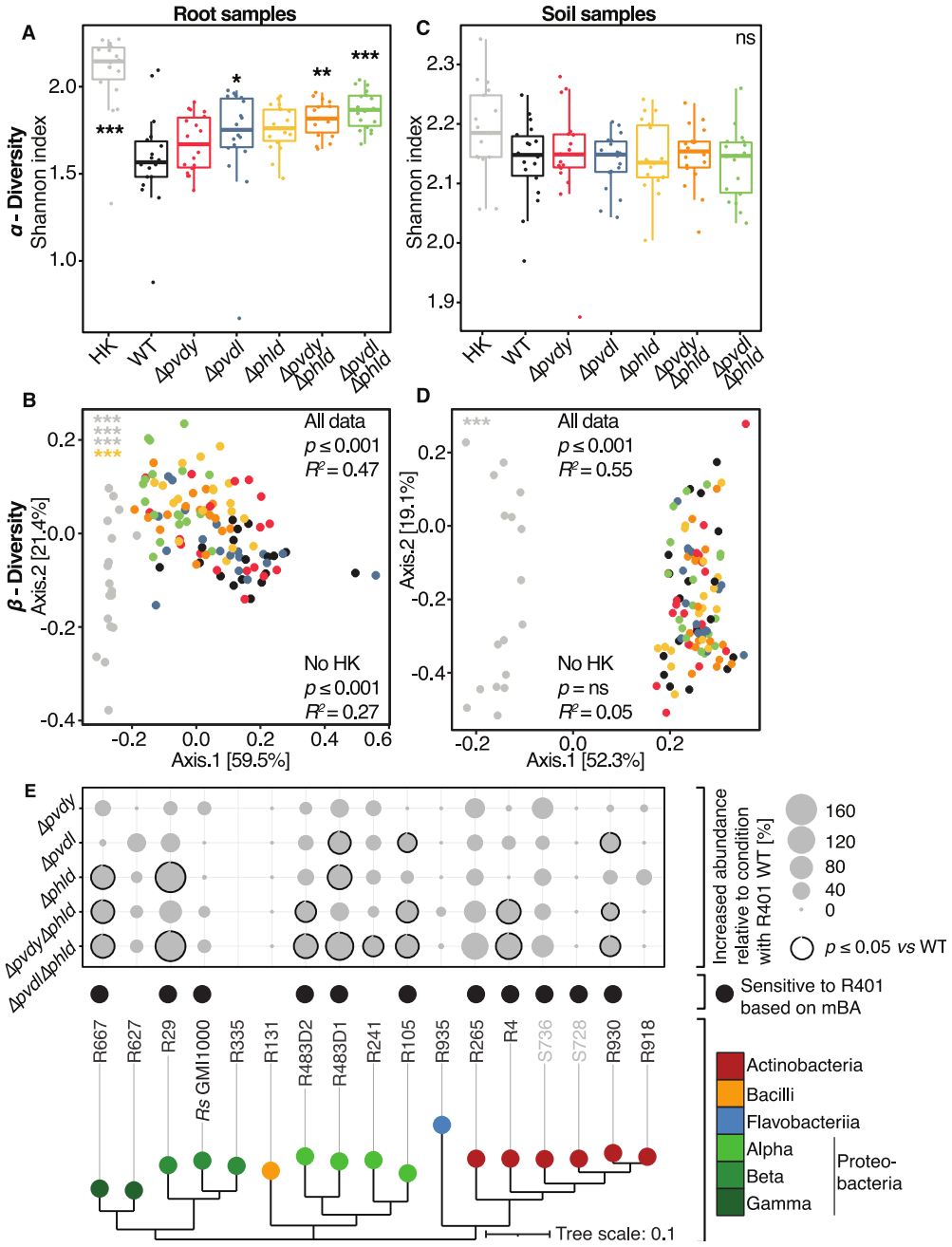


Figure 3.14: DAPG and pyoverdine modulate root microbiota assembly and restrict bacterial diversity. (A, C) Alpha diversity (Shannon index) of root (A) and soil (C) samples in response to R401 or its mutants. Statistical significance was determined by Kruskal-Wallis followed by Dunn's post-hoc test and BH adjustment. Significance compared to WT is indicated by black asterisks (**, ***, indicate $p < 0.01$, and 0.001, respectively; "ns": not significant; n=18). (B, D) PCoA based on Bray-Curtis community dissimilarities between samples in root (B) and soil (D) in response to R401 or its mutants. Density plots represent the clustering of samples on the first two axes. PERMANOVA-derived p-values are represented as asterisks (***, indicate $p < 0.001$; n=18), colored by the respective condition. PERMANOVA analysis on the full data set before (All data) or after (No HK) *in silico* depletion of HK samples are indicated in black; R^2 represents the variance explained by R401 genotype. (E) Balloon plot depicting the increase in relative abundance of SynCom members at the root relative to the condition in which R401 WT has been inoculated. Statistical significance was determined by Kruskal-Wallis followed by Dunn's post-hoc test and BH adjustment. Significance compared to WT in the non-

normalized dataset is indicated by black circles, indicating $p < 0.05$ (n=18). Susceptibility towards R401 wild type in the halo assay is depicted as black spheres. Black spheres indicate sensitivity towards R401 while colored spheres represent the respective bacterial class.

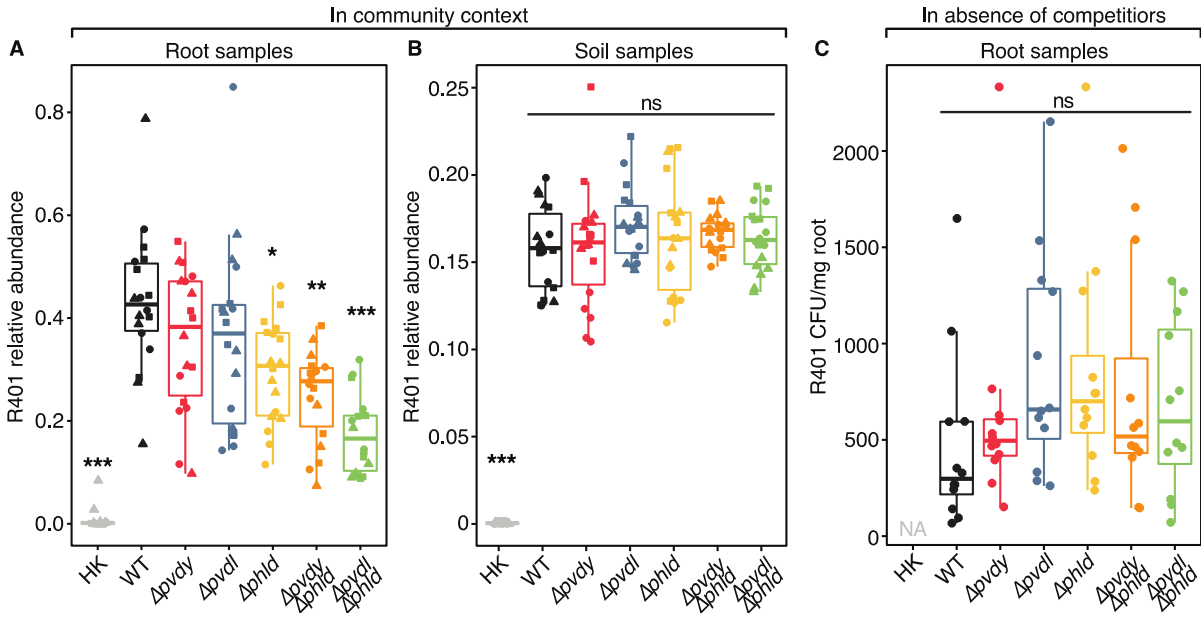


Figure 3.15: DAPG and pyoverdine act as root competence determinants in a community context. (A, B) Relative abundance of R401 WT or mutants in root (A) and soil (B) samples in competition with 18-member SynCom, as in Figure 5; n=18. (C) Colonization capability of R401 or its mutants in mono-associations on axenically grown *A. thaliana* Col-0 roots. Plants were grown on ½ MS-agar plates for 14 days. Colony-forming units have been determined and normalized to root fresh weight; n=12. Statistical significance was determined by ANOVA followed by Tukey's HSD test. No significant differences were detected as indicated by "ns". Statistical significance was determined by Kruskal-Wallis followed by Dunn's post-hoc test and BH adjustment. Significance compared to WT is indicated by black asterisks (∗, ∗∗, ∗∗∗, indicate p < 0.05, 0.01, and 0.001, respectively; ns, not significant).

Chapter 4

Horizontal gene transfer in the plant-associated microbiota

4.1 Abstract

Horizontal gene transfer (HGT) is one of the main factors driving microbial evolution in rapidly changing environments. Phylogenetically diverse soil-borne microbes colonize exterior and interior of plants. However, there is a lack of knowledge about the role of HGT in helping free-living microbes to successfully adapt to plant-associated niches and persist in these complex communities. In this study, we systematically analyzed HGT events in the plant-associated microbiome. We observed that 42% of the genomes showed a signature of recent HGT, and many of the genes were transferred in the form of gene clusters *via* conjugative plasmids. Here, we provid an atlas of the taxonomy and functions of HGT events in the plant-associated microbiota, which shows dynamic patterns across microhabitats and is strongly influenced by environmental parameters. Functional interpretation of transferred genes prompted the potential selection force in and adaptation to plant-associated niches. We also examined the gain of novel functions that would potentially contribute to the fitness of the recipient organisms by incorporating into our analyses strain-level microbial community composition data. Our results illustrate the importance of HGT in bacterial adaptation to plants and potential beneficial outcome of HGT in recipient microbes, which are relevant to understand the establishment of the plant-associated microbiota and may have transferrable applications to sustainable agriculture.

4.2 Introduction

The plant microbiota is composed by phylogenetically diverse microbes that may play an important role in plant fitness. Innovations in multi-omics techniques have advanced our understanding of the complex community structures and functions of the plant microbiota from different compartments across diverse host species (Bulgarelli et al. 2013; Bulgarelli et al. 2012; Ling et al. 2022; Xu et al. 2018), however there is limited knowledge of the mechanisms driving microbial adaptation to plants and the evolution of bacteria in the context of complex communities. Recent studies have found a large number of genes and genetic variants responsible for bacterial plant colonization by using comparative genomics, Tn-seq mutant libraries and artificial evolution strategies (Batstone et al. 2020b; Levy et al. 2018; Wheatley et al. 2020), but those studies largely disregard the role of interactions among microbiota in driving microbial evolution and adaptation to plants. Understanding the transfer from free-living bacteria to plant symbionts during co-habitation with plants is a long-standing question in plant-microbe interaction.

In the context of microbial communities, bacteria can share genetic elements through horizontal gene transfer (HGT), which is a major force in microbial evolution and adaptation to changing environments. Diazotrophic *Rhizobiales* is often studied for their symbiotic genes with leguminous plants in consideration of their nitrogen-fixation and nodulation capacity (Carvalho et al. 2010). A large-scale genome analysis reported that the ancestor of rhizobia likely lacked the nitrogen fixation and nodulation traits, which suggested the acquisition of these functions through HGT in the descendants (Garrido-Oter et al. 2018). Given the nitrogen fixation ability in free-living bacteria that lack the symbiotic capacity, it is believed that both the nitrogen fixing genes and nodulation genes were acquired via HGT in batches and integrated into an operon when bacteria evolved upon colonization on leguminous plants (Carvalho et al. 2010). Moreover, the symbiosis genes in *Rhizobiales* tend to be mobile (Wardell et al. 2022). Accumulative evidence has proved that plant-associated niches can facilitate HGT among colonizing microbes. For instance, Ling *et. al.* provided the evidence that the plant-derived flavonoids can enhance the transfer frequency of a symbiotic island between rhizobia and convert the non-symbiotic bacterium into symbiont (Ling et al. 2016). Nonetheless, plant-associated microbiota is composed of diverse abundant microbes for which patterns of HGT and its impact in community assembly and dynamics have never been elucidated.

To obtain a holistic overview of the HGT in plant-associated microbiota and understand the role of HGT in bacterial adaptation to plants and community assemblage, we systematically analyzed the HGT events in 6 genome-indexed bacterial culture collections assembled from the root and leaf of *Arabidopsis thaliana*, root of *Lotus japonicus* and phycosphere of *Chlamydomonas reinhardtii* collected across four countries (Bai et al. 2015; Durán et al. 2022b; Harbort et al. 2020; Levy et al. 2018; Wippel et al. 2021). Phycosphere isolates were included after a recent study showed that the photosynthetic green algae assembled a microbiota that resembled the root microbiota of *Arabidopsis thaliana* (Durán et al. 2022b), illustrating that this system can be used as a potential model to study the interaction between microbiota and photosynthetic eukaryotic hosts. By incorporating the microbial community structures of the samples where the strains were isolated, we can associate HGT features with strain fitness in the specific niche to probe the potential outcome of HGT in microbial adaptation and community assemblage.

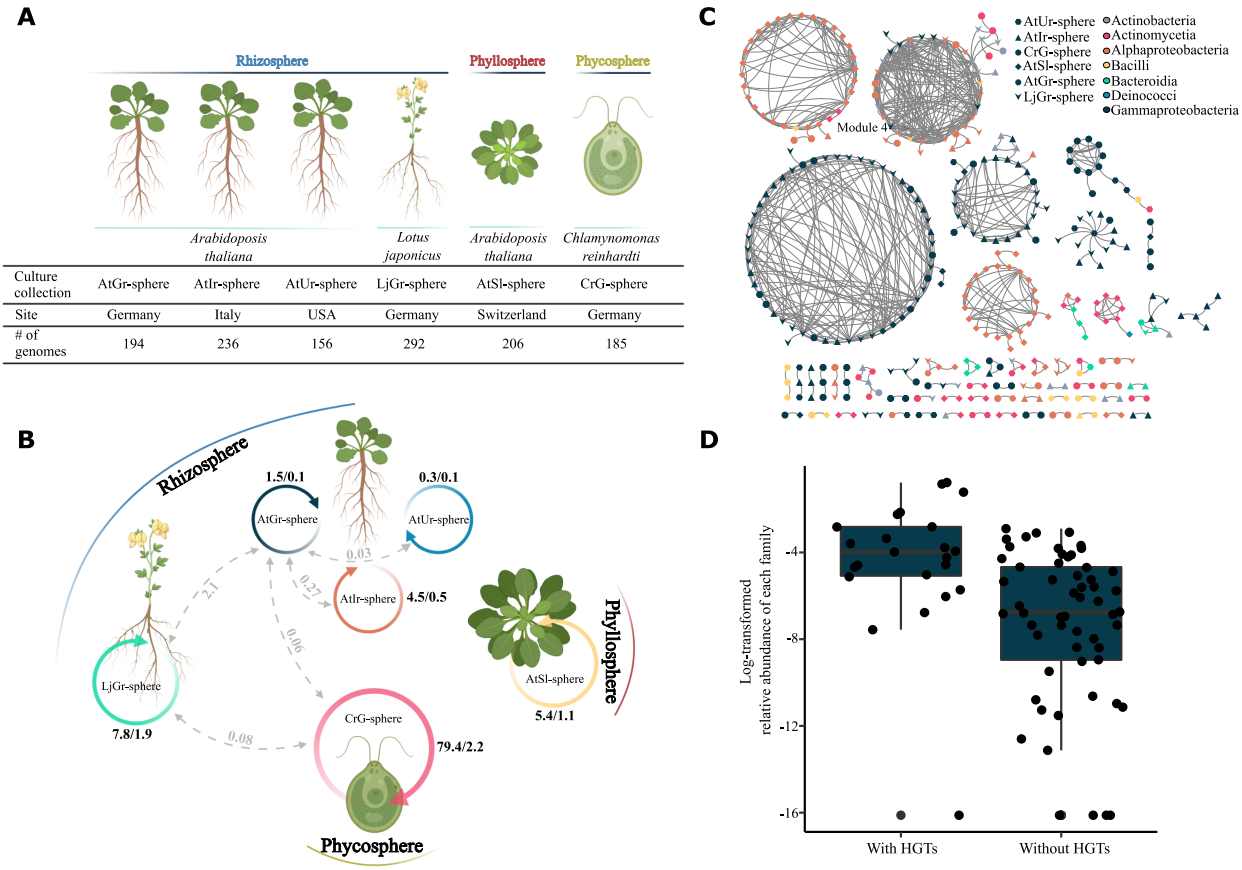
4.3 Results

4.3.1 HGT is prevalent in the plant-associated microbiota

We retrieved 168, 140 and 142 non-redundant genomes of root bacterial isolates from *Arabidopsis thaliana* collected from Germany (*AtGr*-sphere), Italy (*AtIr*-sphere) and the USA (*AtUr*-sphere) respectively (Bai et al. 2015; Harbort et al. 2020; Levy et al. 2018), 191 non-redundant genomes of leaf isolates from *Arabidopsis thaliana* sampled from Switzerland and Germany (*AtSl*-sphere) (Bai et al. 2015), 180 non-redundant genomes of root isolates from *Lotus Japonicus* grown in Germany (*LjGr*-sphere) and 47 non-redundant genomes from phycosphere of *Chlamydomonas reinhardtii* (*CrG*-sphere) (Durán et al. 2022b; Wippel et al. 2021). These culture collections spanned 3 distinct photosynthetic host-associated microhabitats: rhizosphere, phyllosphere and phycosphere. Bacteria from the same culture collections were isolated simultaneously from the same host species grown in the same field, which we hypothesized would enable the detection of very recent HGT events in the context of microbial communities. We clustered the non-redundant genomes into species clusters based on 95% ANI and retained 136, 105, 100, 147, 125, 29 species in *AtGr*-sphere, *AtIr*-sphere, *AtUr*-sphere, *AtSl*-sphere, *LjGr*-sphere and *CrG*-sphere culture collection respectively (Figure 4.1A).

To identify HGT events, we clustered genes that are at least 500bp in length based on 99.9% identity between any pair of genomes from different species, which is thought to be potential recent HGT (Groussin et al. 2021; Handley et al. 2017; Smillie et al. 2011). Here, each HGT event refers to a pair of potentially transferred genes between two genomes. We identified that 5,193 genes pairs from 719 genome pairs were potentially recently transferred, including 213 HGTs from *AtGr*-sphere, 436 from *AtIr*-sphere, 34 from *AtUr*-sphere, 982 from *AtSl*-sphere, 1,315 from *LjGr*-sphere, 798 from *CrG*-sphere and 1,415 transfers across culture collections. To compare the HGT frequencies between different culture collections, we normalized the number of HGTs by the total number of comparable genome pairs in each culture collection. Intriguingly, we found that the *CrG*-sphere showed unexpectedly higher HGT frequency than other culture collections, which is 79.4 HGT events per 100 genome pairs, followed by 7.8 in *LjGr*-sphere, 5.4 in *AtSl*-sphere, 4.5 in *AtIr*-sphere, 1.5 in *AtGr*-sphere and 0.3 in *AtUr*-sphere (Figure 4.1B). On average, the number of HGT across culture collections was less than within culture collection transfer, which implies that shared environments limit the occurrence of HGT in the plant-associated microbiota. Qualitatively,

the number of genome pairs involved in HGT varied by 20 times for microbiota dwelling in different microhabitats, 0.1~0.5 out of 100 genome pairs for all *Arabidopsis thaliana* root culture collections, 1.1 for *AtSl*-sphere, 1.9 for *LjGr*-sphere, and 2.2 for *CrG*-sphere (Figure 4.1B).



Culture collection	AtGr-sphere	AtIr-sphere	AtUr-sphere	LjGr-sphere	AtSl-sphere	CrG-sphere
Site	Germany	Italy	USA	Germany	Switzerland	Germany
# of genomes	194	236	156	292	206	185

Figure 4.1 The prevalence of HGT in plant-associated and constraining factors. (A) The description of culture collections used in this study. All the bacteria from the same culture collections were isolated from the same host species. (B) Shared environments constrain the HGT frequency. The self-loop represents the HGT within the same culture collection and dashed grey line represents the HGT across different culture collections. The index left to the slash indicates the normalized HGT frequency in 100 genome pairs within the same culture collection and the right index represents the normalized number of genome pairs involved in HGT out of 100 genome pairs. Only normalized HGT frequency is shown for cross-culture collection transfer. (C) Phylogenetic relatedness constrains HGT frequency. Each node represents each genome and any genome pairs are connected by an edge if they share genes. The width of the edge indicates the HGT frequency between the genome pair. (D) Bacterial abundance constrains HGT frequency. The relative abundance of each family was calculated from the amplicon sequencing data of samples where the strains from *AtGr*-sphere, *AtIr*-sphere, *LjGr*-sphere and *CrG*-sphere were isolated. The x axis indicates whether strains from the given family are involved in HGT with others (P value < 0.05, Wilcox test).

To obtain a comprehensive insight into the HGT in plant-associated microbiota, we built a network of HGT events. Analyzing this network, we observed that strains from the same class were more

likely to be involved in HGT than across classes in general (4,958 within-class transfer vs. 235 between-class transfer, Figure 4.1C). However, we found an exceptional case, where strains in network module 4 tend to exchange genes across phylogenetically distant bacteria, which were dominated by HGTs between *Alphaproteobacteria* and *Gammaproteobacteria*, *Actinobacteria* and *Gammaproteobacteria*, and *Actinobacteria* and *Alphaproteobacteria*. This result highlights that phylogenetic relatedness between strains serves as another constraint in HGT, which was also supported by studies of HGT in the human and cheese microbiota (Handley et al. 2017; Smillie et al. 2011). By comparing the HGT network with the co-occurrence network of the strains in the corresponding natural environments, we found many of the strains sharing genes do form neither positive nor negative correlations. However, we found an enrichment of positive correlations among strains exchanging genes in *AtIr*-sphere and *AtGr*-sphere (P=0.03 and 0.13, respectively), and a depletion of negative correlations in *LjGr*-sphere and *AtGr*-sphere (P=0.002 and 0.099, respectively) (Figure 4.2), which implies that strains exchanging genes are less likely to be antagonistic to each other. Given the finding that close strains are preferentially engaged in HGT, we questioned whether the high frequency of HGT found in phycosphere can be explained by the close phylogenetic relationships between strains in *CrG*-sphere. We calculated the phylogenetic distances between each pair of strains in each culture collection. We found that the phylogenetic distances between strains were comparable across culture collections, which indicates phylogenetic relationship cannot explain the high HGT frequency in phycosphere. Further, we also found a positive correlation between the abundance of bacteria families and the elevated HGT incidence (Figure 4.1D), suggesting that high cell densities potentially increase the possibility of the formation of physical contact between bacteria and initiation of gene transfer via conjugation. Alternatively, this finding may also suggest HGT mediates the acquisition of genetic traits that are beneficial in bacterial associations with plants.

4.3.2 HTG in the plant-associated microbiota depends on the genomic context

We found multiple genes can be transferred between genomes (ranging from 1~383 HGT events in a genome pair), especially in *CrG*-sphere. We sought to answer the question of whether those genes were transferred concurrently or at independent evolutionary time points. One hypothesis is that if the transferred genes are proximal to each other in the genome, they are more likely to be transferred at the same time. To test this hypothesis, transferred genes were classified into gene

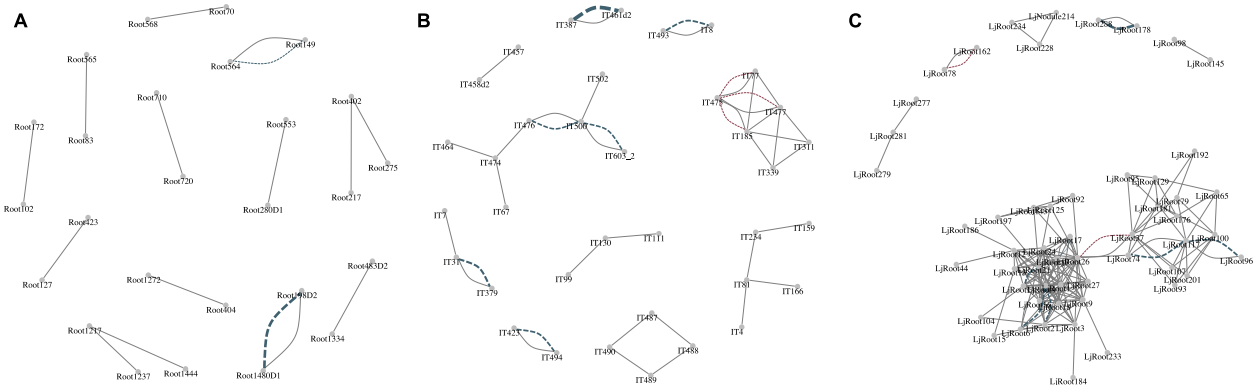


Figure 4.2 Mapping HGT network onto the microbial co-occurrence network. Each panel represents network from one culture collection: (A) AtGr-sphere, (B) AtIr-sphere and (C) LjGr-sphere. In the network, each node represents one strain. The solid grey edges connect two genomes sharing genes. Dashed blue and brown edges represent positive and negative correlation between two strains respectively. The width of dashed edges is proportional to the correlation coefficient.

clusters based on their genomic proximity, which resulted in 5,546 (90.9%) genes falling into 545 transferred gene clusters encompassing at least 2 transferred genes. For each culture collection, the percentage of transferred genes falling into transferred gene clusters ranged from 78.0%~98.4% (Figure 4.3A). The average number of genes in gene clusters was 5.6, with the maximum number of genes reaching to 272. Specifically, the average size of transferred gene clusters in *CrG*-sphere was 41.5, which was 4~8 times larger than other culture collections, a difference which could explain the high HGT frequency in phycosphere. However, we have to acknowledge that some of the genome assemblies are fragmented, so transferred genes from the fragmented contigs cannot be well clustered and larger gene clusters would be expected if high-quality genome assemblies were available.

How were the large gene clusters transferred? Plasmids can be transferred between bacteria via conjugative type IV secretion system (T4SS), so we expect that plasmids can contribute to the transfer of gene clusters in plant-associated microbiota. We predicted the plasmid contigs in all non-redundant genomes and identified the transferred genes from plasmid contigs. Intriguingly, we found 1,590 (30.6%) HGTs in which both genes in the transferred gene pair were located on plasmids (between-plasmid), 1,380 (26.6%) HGTs containing one gene in a gene pair on the plasmid and the other one on the chromosome, and neither of the paired genes coming from plasmids (between-chromosome) in the remaining 2,223 (42.8%) HGT events (Figure 4.3B). Generally, plasmid-involved HGT accounted for >50% of entire HGT events in most culture

collections. However, we observed two opposite cases where between-plasmid transfer was nearly dominant in *CrG*-sphere (94.1%), in line with the larger transferred gene clusters found in phycosphere, while between-chromosome transfer was high in *AtIr*-sphere (86.9%). For transferred gene clusters including at least 20 genes, 110 out of 253 (43.5%) clusters were from plasmids, especially for the large transferred gene clusters (>100 genes) (Figure 4.3C). As mentioned above, due to the fragmented genome assemblies for some strains, not all the contigs could be correctly distinguished as plasmid contigs, especially for IT bacterial genomes, so the finding in this study should serve as a lower boundary, and we can expect more transferred gene clusters originating from plasmids. In congruent with the high proportion of plasmid-dependent HGT, for the 205 strains carrying plasmid-involved HGT events, we found that 73 (35.6%) of them possessed complete conjugative T4SS. Taken together, these results illustrate that plasmids play an important role in initiating the HGT in plant-associated microbiota.

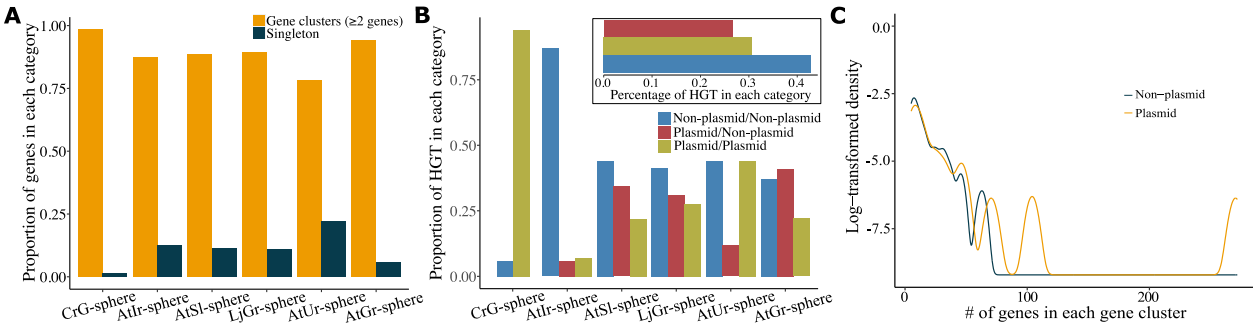


Figure 4.3 Plasmids participate in transfer of gene clusters. (A) The proportion of transferred gene clusters in each culture collection. Gene clusters were identified based on genomic proximity of transferred genes. (B) The proportion of plasmid-involved HGT events in each culture collection. 'Plasmid/Plasmid' represents both of the genes in a gene pair are located on plasmid contigs; 'Plasmid/Non-plasmid' represents one of the genes in a gene pair is located on plasmid; 'Non-plasmid/Non-plasmid' represents none of the genes in a gene pair is located on plasmids or both of them are located on chromosomes. The embedded barplot shows the overall distribution of HGT events from the three categories across the whole dataset. (C) The distribution of plasmid-based gene clusters. The density plot shows the number of plasmid-based or chromosome-based gene clusters as a function of the size of clusters.

Though extensive HGT have been reported in several host-associated microbiomes (Groussin et al. 2021; Smillie et al. 2011), the fate of the transferred genetic elements in the microbe recipients are rarely studied. We found transfer of plasmids to other cells, which makes it possible to detect the evolution of plasmids after transfer in recipients. Here, we specifically scrutinized the two largest transferred gene clusters from two phycosphere strains (Chlamy76 and Chlamy187) belonging to *Burkholderiaceae*. They came from two mega-plasmid contigs with a length of

425,524bp and 423,742bp respectively. One of the two mega-plasmids was potentially transferred from one strain to another because of the high sequence similarity between them and absence of isogenic plasmids in their close relatives (Figure 4.4A). According to the synteny plot between the two mega-plasmids, they were well aligned except several genetic variations (Figure 4.4B), including 28 SNPs and 3 indels. Assuming that the mega-plasmid evolves with a mutation rate of 10^{-9} SNP per site per generation and the generation time of the strain is 20 mins, this led to the estimated time of transfer around 2.5 years ago, which was a very recent horizontal transfer. However, given the fact that the plasmid can present with multiple copies in the genome, we can expect a more recent time of the transfer. 5 out of the 28 SNPs were intragenic, in which 4 SNPs caused nonsynonymous mutations in 4 genes. The 3 indels were all transposase-related, including two deletions of transposases in the mega-plasmid from Chlamy76 and one insertion of transposase and two neighboring genes. This finding indicates the transferred genetic elements can experience rapid evolution via insertion or deletion of other mobile genetic elements in recipient microbial genomes.

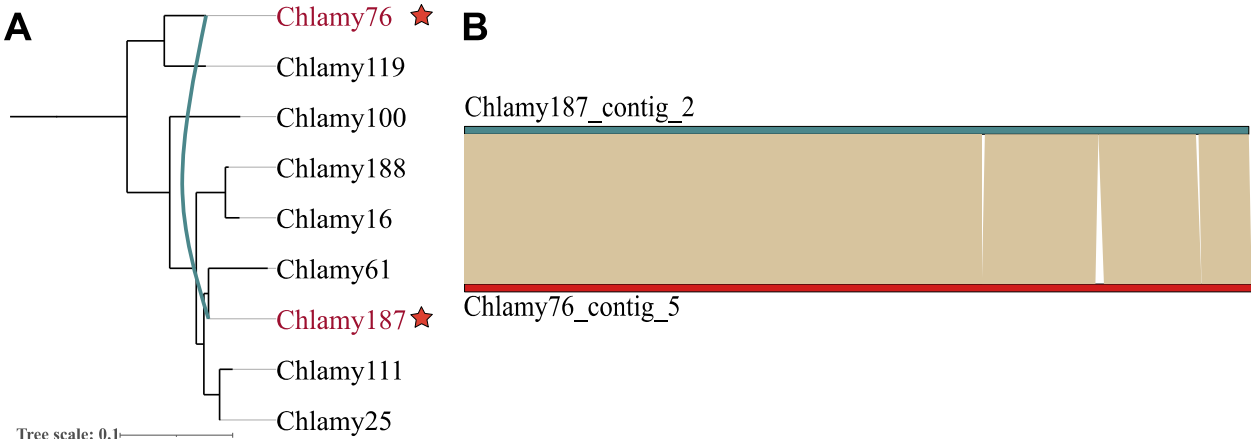


Figure 4.4 The evolution of transferred mega-plasmids. (A) The phylogenetic tree of Chlamy76 and Chlamy187 together with their close relatives. The cognate mega-plasmid was only found in Chlamy76 and Chlamy187 as shown in red stars, suggesting a plasmid transfer happening between the two strains. (B) The synteny plot between the two mega-plasmids shows high similarity except some minor genetic variations.

4.3.3 Taxonomic distribution of HGT events in the plant microbiota

Next, we set out to identify the active agents of HGT in the plant-associated microbiota. In general, we observed remarkable differences in HGT frequencies in distinct taxa (Figure 4.5A). A small portion of families showed patterns of high HGT in multiple culture collections, while others

showed culture collection-specific patterns (Figure 4.5B). In line with the above-mentioned finding that abundant taxa exhibited significantly greater HGT incidences, we found that *Burkholderiaceae*, which is most abundant in rhizosphere and phycosphere samples (Durán et al. 2022; Ling et al. 2022), exhibited the highest HGT frequency and accounted for 46.0% of all HGT events, with clear signatures observed in genomes from all microhabitats and host species (Figure 4.5C). *Burkholderiaceae* contributed 94.2%, 69.0%, 34.3%, 13.5%, 11.8% and 0.7% of HGT in *CrG*-sphere, *LjGr*-sphere, *AtGr*-sphere, *AtIr*-sphere, *AtUr*-sphere and *AtSl*-sphere, respectively. In the root culture collections, *Xanthomonadaceae* that is second most abundant bacteria in rhizosphere was active in HGT in *AtIr*-sphere and *LjGr*-sphere, accounting for 46.3% and 19.1% of HGT in respective culture collection, but was devoid in *AtGr*-sphere and *AtUr*-sphere. *Rhizobiaceae*, whose members are responsible for nodule formation in legumes, contributed more HGT in *LjGr*-sphere compared to other non-legume hosts and *LjGr*-sphere accounted for 56.6% of *Rhizobiaceae*-involved HGT across the whole dataset. In the leaf culture collection, *Sphingomonadaceae* and *Beijerinckiaceae* that are abundant in phyllosphere (Almario et al. 2022) contributed the majority (86.9%) of HGTs in leaf microbiota. Despite the dominance of HGTs within the same family, we still found prevalent HGT cases between phylogenetically distant taxa. For instance, phylogenetically distant transfer was prevalent between *Burkholderiaceae* and other bacterial taxa in *LjGr*-sphere. By comparing the taxonomic distribution of HGT events in microbiota from the same host grown in different sites (*AtGr*-sphere, *AtIr*-sphere and *AtUr*-sphere) or from different hosts grown in the same soil (*AtGr*-sphere, *LjGr*-sphere and *CrG*-sphere, Figure 4.5C), the unique pattern of active HGT agents in each culture collection suggested that HGT initiators can be determined by both the hosts and environmental factors.

4.3.4 The HGT frequency depends on the functional category

To elucidate the potential impact of HGT in the recipient organisms, we performed an analysis of functional capacity of transferred genes. Annotation of the transferred genes reflected that their majority were assigned an unknown function, followed by genes related to transcription, inorganic ion transport and metabolism, replication, recombination and repair.

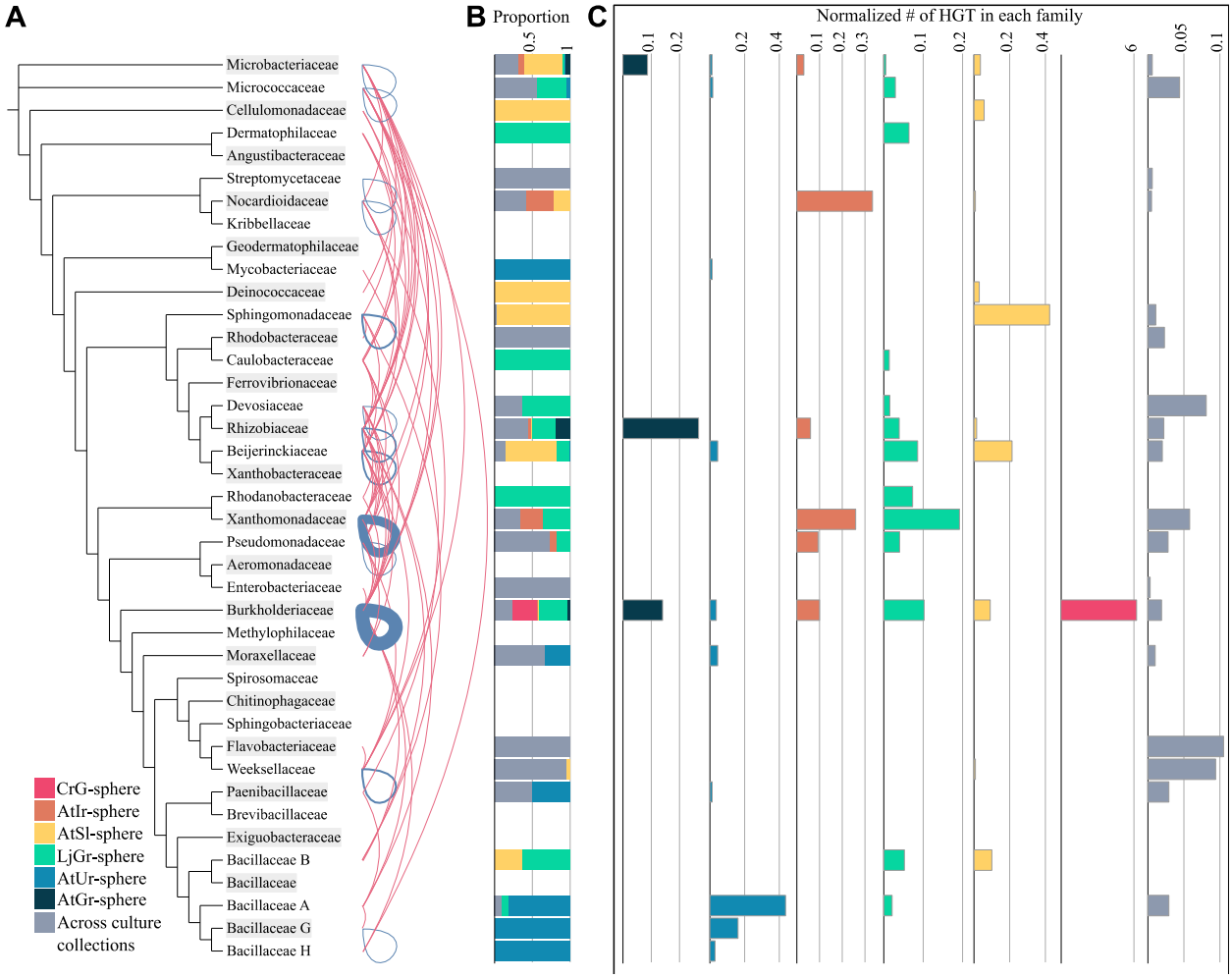


Figure 4.5 The taxa actively involved in HGT in plant-associated microbiota. (A) Mapping HGT onto the bacterial phylogenetic tree across plant-associated microbiota. Blue curves indicate the HGT within the strains from the same family and red curves indicate the HGT across different families. Width of the curves represents the HGT frequency. (B) and (C) Active taxa involved in HGT in each culture collection for a given family. The barplot in panel **B** represents the proportion of number of HGT events contributed by each culture collection for a given taxa. The normalized HGT frequency by total number of available genome pairs in each culture collection is shown in **C**.

To better understand which functions were specifically selected to be transferred rather than by chance, we identified categoriesthat were enriched/depleted in HGT compared with the background whole-genome level. Interestingly, we found general or culture collection-specific enriched functions (Figure 4.6A). For example, 'replication, recombination and repair' was found to be enriched in *CrG*-sphere, *AtGr*-sphere, *LjGr*-sphere and *AtIr*-sphere. 'Intracellular trafficking, secretion, and vesicular transport' was enriched in *CrG*-sphere, *AtGr*-sphere and *LjGr*-sphere, 'inorganic ion transport and metabolism' and 'transcription' were enriched specifically in two culture collections, i.e., *AtSl*-sphere and *AtIr*-sphere, *AtSl*-sphere and *CrG*-sphere, respectively,

and ‘amino acid transport and metabolism’ was exclusively enriched in *CrG*-sphere but depleted in 3 other culture collections. As expected, vertically transferred functions were commonly depleted in HGT events in multiple culture collections, e.g., ‘Coenzyme transport and metabolism’, ‘Nucleotide transport and metabolism’ and ‘Translation, ribosomal structure and biogenesis’ (Figure 4.6A). The highly dynamic functional patterns of HGT in different microhabitats suggest that the selective force of HGT in plant-associated microbiota corresponds to the physicochemical properties of each microhabitat.

Since many transferred genes had unknown function, we clustered genes into orthologues, which allowed us to identify which functional groups were frequently transferred in plant-associated microbiota at high resolution. 97 orthologues were found to include at least 10 transferred genes in our dataset and all of them were transferred at significantly higher frequencies than expected by chance (Figure 4.6B). The frequently transferred orthologues could be separated into two groups based on their prevalence across microhabitats. For example, orthologues from ‘transcription’ and ‘inorganic ion transport and metabolism’ were frequently transferred in multiple microhabitats, but orthologues belonging to ‘energy production and conversion’ and ‘intracellular trafficking, secretion, and vesicular transport’ were mainly found in *CrG*-sphere and *LjGr*-sphere, respectively. Those frequently transferred orthologues may suggest that they can play important roles in mediating microbial adaptation to these microhabitats.

Next, we assessed the underlying factors shaping the dynamic functional patterns of transferred genes in different microhabitats. Given that our culture collections originated from the same site (*AtGr*-sphere, *LjGr*-sphere and *CrG*-sphere) or same host (*AtGr*-sphere, *AtIr*-sphere and *AtUr*-sphere), it was possible to estimate the extent to which the two factors contributed to the selection of transferred functionalities. Interestingly, we found a greater overlap of transferred orthologues among culture collections derived from the same sites but different species than that from same species but different sites (Figure 4.7). This result suggests that the environment (geographical location or soil type in this case) plays a more important role in driving HGT than host species.

4.3.5 Frequently transferred functions vary across taxa

Given the dynamic functional patterns observed at the microhabitat level, we questioned whether impact of environments and host on patterns HGT for single taxa was still present. We focused on

Figure 4.6 Distinct functions are positively selected to be transferred in different culture collections. (A) Frequently transferred functional categories in each culture collection. The x axis represents the proportion of HGT events contributed by each functional category in each culture collection. The figures right next to bars represent the absolute number of HGT events related to the functional categories in a given culture collection. Red arrows indicate the enrichment of the functional categories in HGT events compared to the whole-genome background level in the studied culture collection, and green arrows indicate the depletion of the functional categories in HGT. (B) The frequently transferred orthologues in each culture collection ordered by the functional categories. The enrichment index highlights how the transfer frequency of a given orthologue deviates from the whole-genome level, which is calculated as $\frac{The\ percentage\ of\ a\ given\ orthologue\ in\ transferred\ genes\ in\ a\ culture\ collection}{The\ percentage\ of\ a\ given\ orthologue\ across\ genomes\ in\ a\ culture\ collection}$.

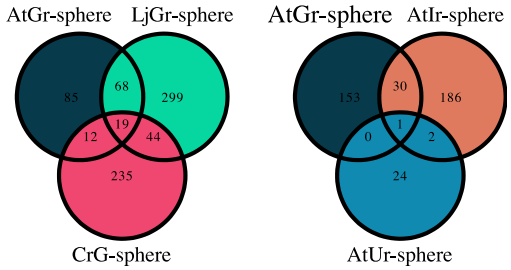


Figure 4.7 The shared transferred orthologues among microbiota from different habitats. The left and right venn plot represents the shared orthologues among microbiota from the same site but different hosts and from the same host but different sites, respectively.

the functions of frequently transferred orthologues in three abundant bacterial taxa in rhizosphere and phycosphere: *Burkholderiaceae*, *Rhizobiaceae* and *Xanthomonadaceae* (Durán et al. 2022b; Ling et al. 2022), and two abundant taxa in phyllosphere: *Sphingomonadaceae* and *Beijerinckiaceae* (Almario et al. 2022). *AtUr*-sphere was excluded in the analysis because none of HGT was found in the above-mentioned families.

Burkholderiaceae accounted for almost half of the HGT events in plant-associated microbiota. In *LjGr*-sphere, we found orthologues belonging to mobile genetic elements were frequently transferred in *Burkholderiaceae*, including integrase (OG0000593), and conjugation T4SS-related proteins (OG0001491, OG0002010 and OG0000321), which provided an evidence of flourishing HGT within *Burkholderiaceae* in *LjGr*-sphere (Figure 4.8). In *CrG*-sphere, orthologues annotated as enoyl-CoA hydratase isomerase (OG0000018), CoA-transferase family (OG0000051), lactoylglutathione lyase (OG0019548), and glycolate oxidase (OG0001137), acyl-CoA dehydrogenase (OG0000319) and aldehyde dehydrogenase family (OG0000015) were most frequently transferred in *Burkholderiaceae*, many of which belong to carbon metabolism. In the two *Arabidopsis thaliana* root culture collections, oxidoreductase molybdopterin binding domain (OG0000186), lactonase (OG0004162), and response regulator receiver (OG0000001) were most frequently transferred in *Burkholderiaceae* in *AtGr*-sphere. However, for *AtIr*-sphere, HAMP domain (OG0000005), and outer membrane efflux protein (OG0003368) were most frequently transferred. In the top 50 frequently transferred orthologues contributed by *Burkholderiaceae* in each culture collection, only 3 orthologues were shared in 4 culture collections, 0~4 shared orthologues in 2 or 3 culture collections (Figure 4.9A).

HGT events in *Rhizobiaceae* were only found in *AtGr*-sphere, *AtIr*-sphere and *LjGr*-sphere culture collections. Similar to *Burkholderiaceae*, only 1 out of top 50 frequently transferred orthologues was found to be shared between *AtGr*-sphere and *AtIr*-sphere, and 3 orthologues were found to be shared between *AtGr*-sphere and *LjGr*-sphere (Figure 4.9B). For all the *Rhizobiaceae*-involved HGTs, short-chain alcohol dehydrogenases (OG0000000), and response regulator receiver (OG0000001) were most frequently transferred in *LjGr*-sphere; transcriptional regulators (OG0000002), lysR family (OG0000003), and conjugation proteins (OG0001974) were found to be most frequently transferred in *AtGr*-sphere; permease (OG0000103), periplasmic binding protein (OG0002353), and carbon–phosphorus lyase (OG0000012) were most frequently

transferred in *AtIr*-sphere (Figrue 4.10), which could be a consequence of the fact that soluble phosphate is limiting for the growth of both plants and bacteria in the calcareous Italian soil.

Xanthomonadaceae only appeared to contain HGTs in *AtIr*-sphere and *LjGr*-sphere culture collections. Contrarily to *Burkholderiaceae* and *Rhizobiaceae*, 16 out of 50 culture collection-specific most frequently transferred orthologues were found to be shared between two culture collections, which was higher than the other two families (Figure 4.9C). Surprisingly, in the 16 commonly transferred orthologues, we found many genes related to copper resistance (Figure 4.11), including heavy metal translocating P-type ATPase (OG0000122), CopB (OG0002700), multicopper oxidase (OG0000396), RND family (OG0000484), Co Zn Cd efflux system component (OG0002241). Given that *Xanthomonadaceae* strains from *AtIr*-sphere and *LjGr*-sphere were isolated from roots of two different plant species in Italy and Germany respectively, this conserved pattern may imply that *Xanthomonadaceae* in root-associated environments is sensitive to copper concentrations.

By comparing the top 50 most frequently transferred orthologues among *Burkholderiaceae*, *Rhizobiaceae* and *Xanthomonadaceae* in each culture collection separately, large taxa-specific transferred orthologues were identified with only a small overlap (Figure 4.9D, 4.9E and 4.9F). Taken together, these findings suggest that diversity of physicochemical properties in plant-associated microhabitats triggers the dynamic patterns of gene gains in a given taxa, which contributes to genomic diversity and the expansion of the accessory pan-genome. However, the observation that different taxa transfer different genes in the same microhabitat can be potentially attributed to their variation in functional capacity.

For the phyllosphere, the two abundant families *Sphingomonadaceae* and *Beijerinckiaceae* shared a high portion of orthologues regarding the top 20 most frequently transferred orthologues in each family (Figure 4.12), which includes 14 orthologues, e.g., oxidoreductase molybdopterin binding domain (OG0000186), lysR family (OG0000003), lactonase (OG0004162), transmembrane transcriptional regulator (OG0001431), voltage-dependent anion channel (OG0003339), polymerase sigma factor (OG0000025), membrane fusion protein (OG0000717), RND family (OG9999484) and heavy metal translocating P-type ATPase (0G9999122). Except for this relatively conserved pattern, taxa-specific preference for HGT was also observed. Copper resistance-related genes, including multicopper oxidase (OG0000396) and CopB (OG0002700),

were found to be transferred exclusively in *Sphingomonadaceae*. However, arsenical resistance-related genes, including arsenical pump membrane protein (OG0001303), and arsenic resistance protein ArsH (OG0001966), were found to be transferred only in *Beijerinckiaceae*. 2-nitropropane dioxygenase was also found to be involved in HGT in *Sphingomonadaceae*, which has been reported to be related with detoxification, virulence and colonization in nodules and leaves (Koch et al. 2010; Sexton et al. 2006; Zhang et al. 2017).

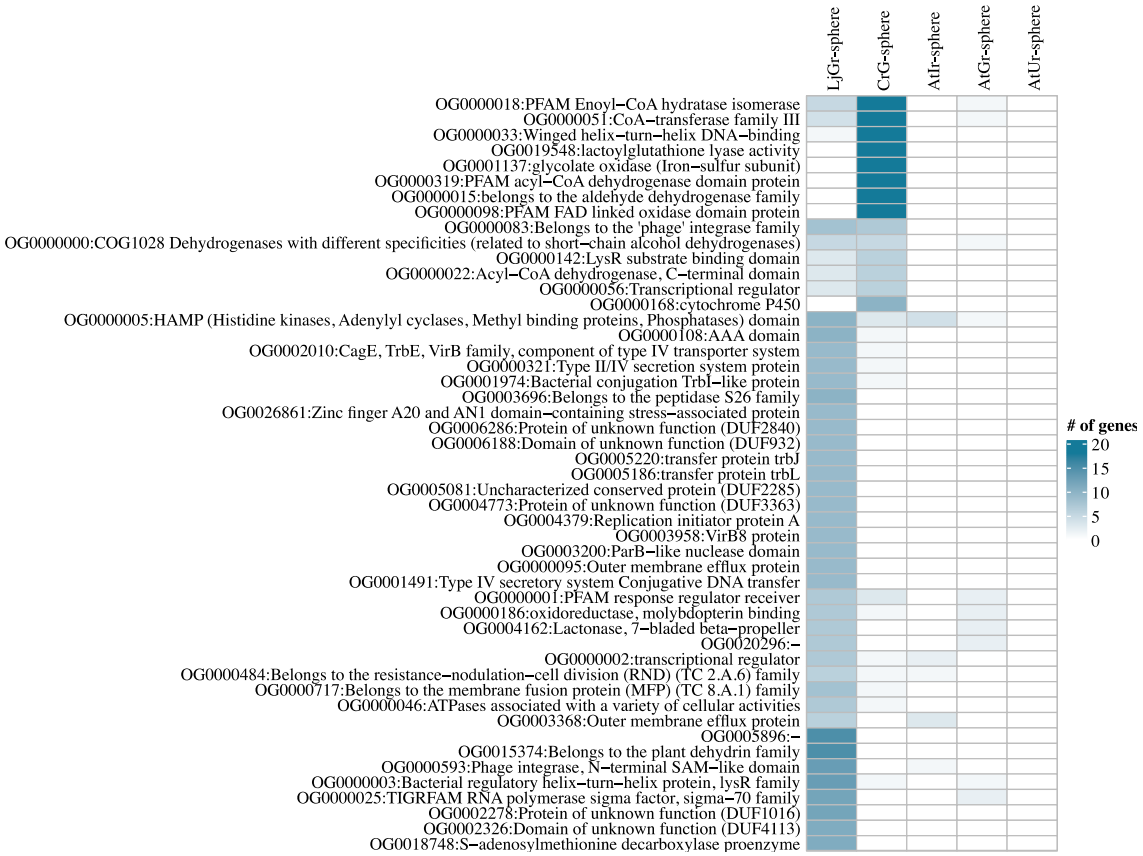
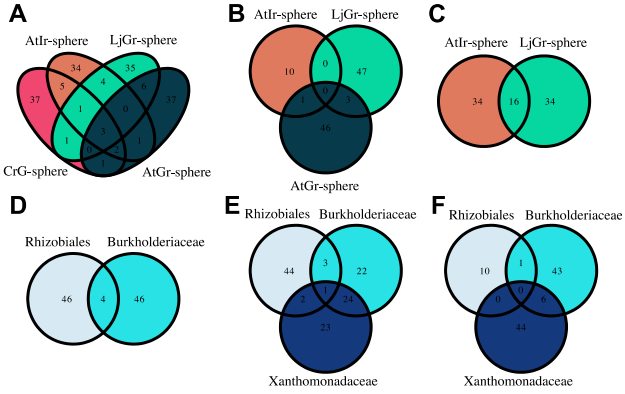


Figure 4.8 The top 50 most frequently transferred orthologues within *Burkholderiaceae*. The color scale represents the number of transferred genes for a specific orthologue in a given culture collection.



(Captions on the next page)

Figure 4.9. The shared transferred orthologues for the same taxa or among different taxa. The shared orthologues among top 50 most frequently transferred orthologues in each culture collection transferred in *Burkholderiaceae* (A), *Rhizobialeceae* (B) and *Xanthomonadaceae* (C). The shared most frequently transferred orthologues between different taxa from the same culture collection, i.e. *AtGr*-sphere (D), *AtIr*-sphere (E) and *LjGr*-sphere (F).

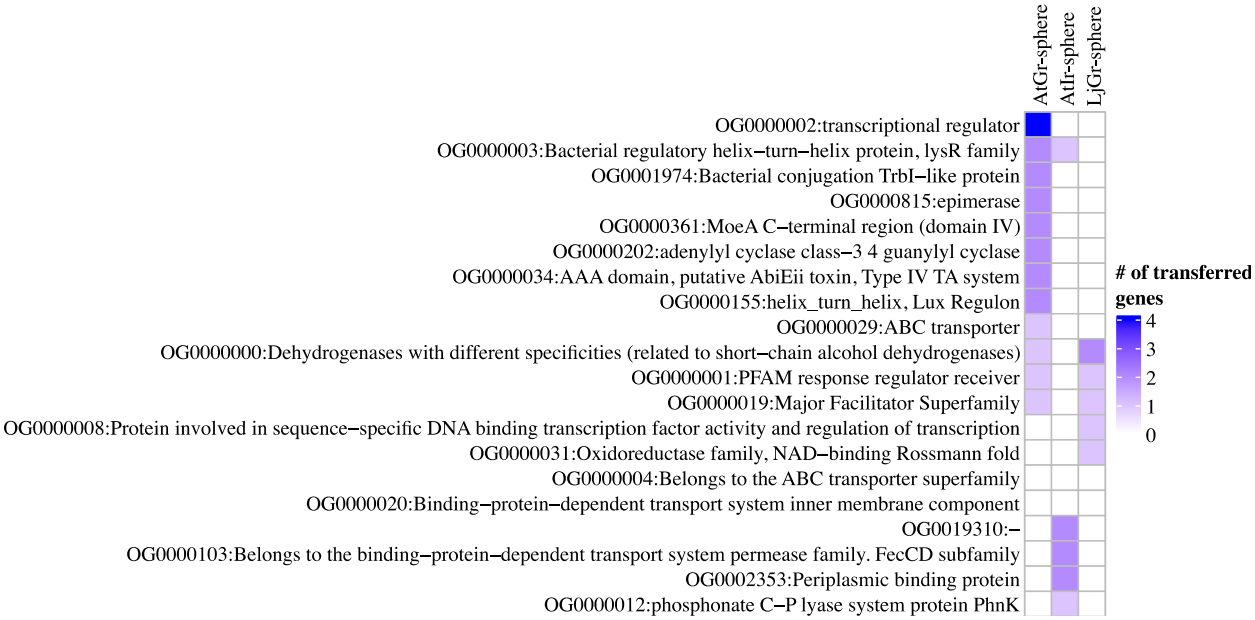


Figure 4.10 The top 50 most frequently transferred genes within *Rhizobialeceae* within culture collections. The color scale represents the number of transferred genes for a specific orthologue in a given culture collection.

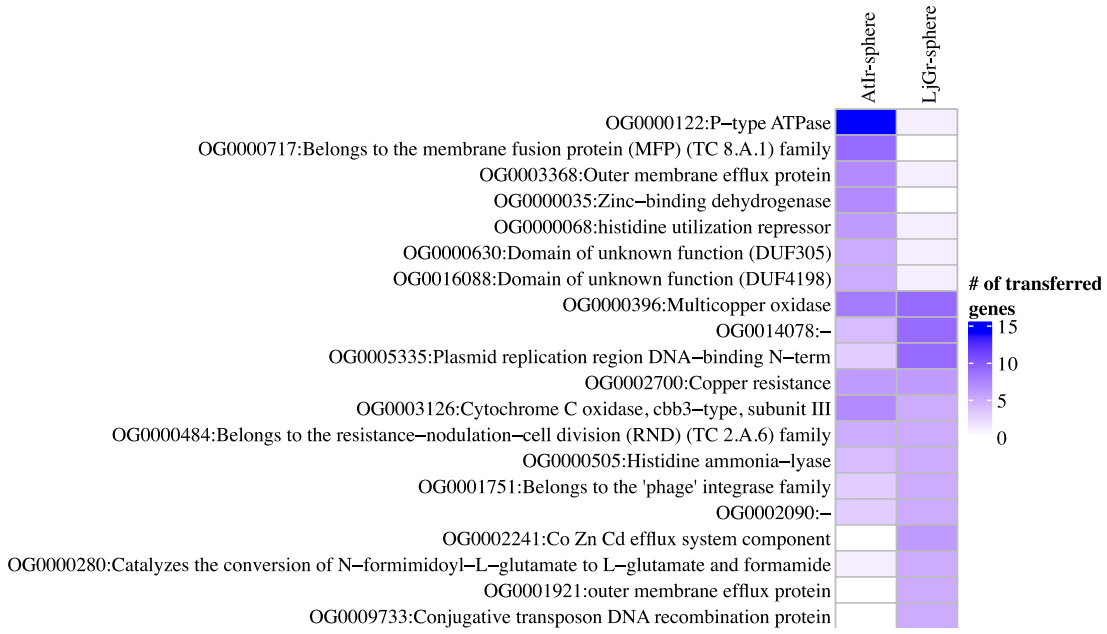


Figure 4.11 The top 50 most frequently transferred genes within *Xanthomonadaceae* within culture collections. The color scale represents the number of transferred genes for a specific orthologue in a given culture collection.

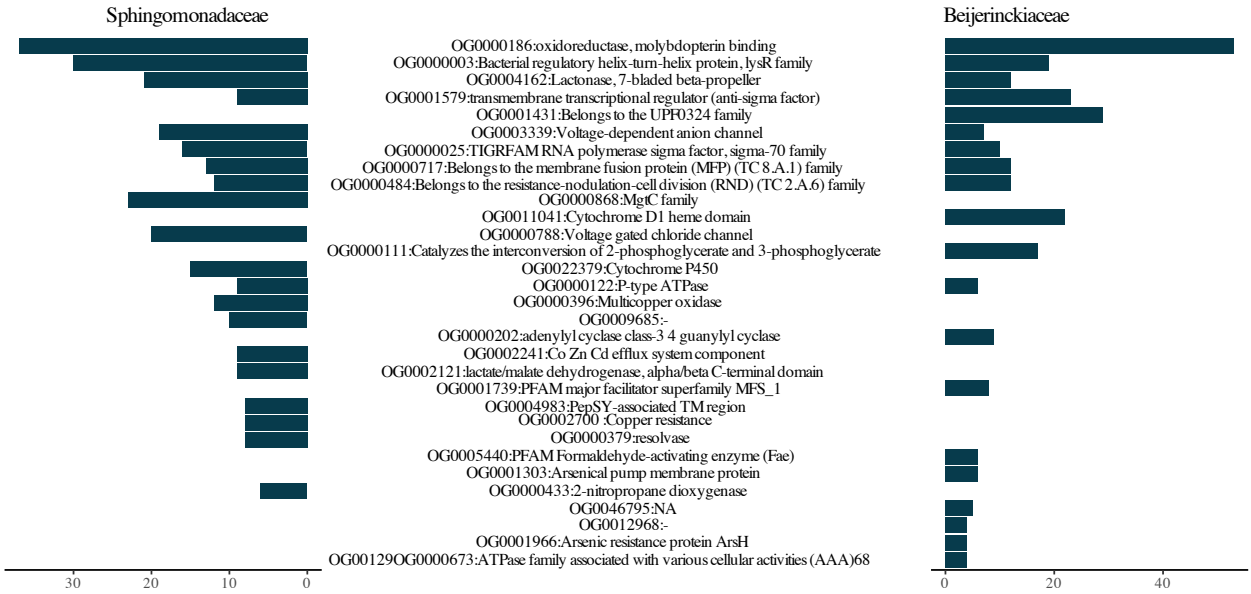


Figure 4.12 The overlaps of frequently transferred orthologues between taxa in the phyllosphere microbiota. The x axis represents the number of transferred genes from each orthologue. Only the top 20 most frequently transferred orthologues in *Sphingomonadaceae* and *Beijerinckiaceae* respectively are shown here.

4.3.6 Transfer of molybdopterin-binding sulfite dehydrogenase encoding genes in the phyllosphere microbiota

The most frequently transferred orthologue OG0000186 in phyllosphere belonged to oxidoreductase molybdopterin binding domain, whichhas been reported to be induced during phyllosphere colonization of *Methylobacterium extorquens* (Müller et al. 2016a). The molybdopterin binding domain is a large and heterogeneous superfamily that binds molybdopterin as a cofactor and is usually found in a variety of oxidoreductases involved in different reactions (Kisker et al. 1997). Interestingly, on the tree of OG0000186 spanning all the non-redundant genomes, we found that 84 out of 86 genes involved in HGT belonging to OG0000186 came from a monophyletic clade (**Figure 4.13**), which was predominate by the genes from phyllosphere and mainly originated from *Proteobacteria* and *Bacteroidota*. Although several transfer cases regarding OG0000186 were also found in *AtGr*-sphere, *CrG*-sphere and *LjGr*-sphere culture collections, those genes form a separate clade than those in phyllosphere. The conservation of transferred genes prompted us to investigate whether the 84 transferred genes actually work with functionally similar oxidoreductases. To address this question, we searched for the neighboring genes surrounding the transferred genes. 41 out of the 84 genes (48.8%) were accompanied with sulfite dehydrogenase, but they exclusively came from a subclade inside the monophyletic clade.

The sulfite dehydrogenases next to the molybdopterin binding domain also showed 100% sequence similarities between all pairs of genomes exchanging the binding domain, which indicated that sulfite dehydrogenase was concurrently transferred with the binding domain but the transfer was undetectable in our analysis due to the short length of this gene whose maximum length was 483bp. This finding supports the notion that sulfur is limiting bacterial growth on leaves (Müller et al. 2016a).

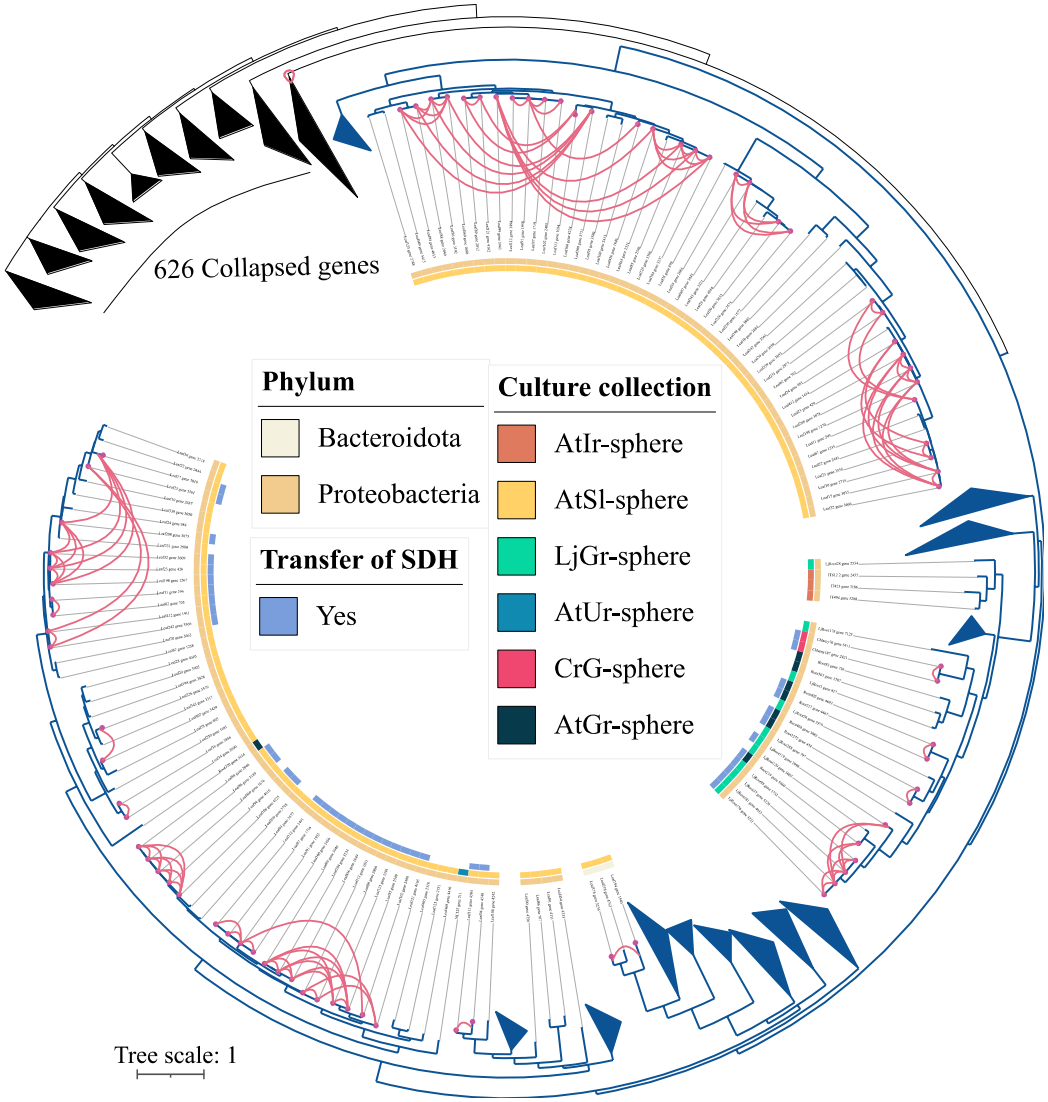


Figure 4.13 The phylogenetic tree of the molybdopterin binding domain from OG0000186. All the genes belonging to OG0000186 are extracted from all the non-redundant genomes and are subject to multiple alignment for tree construction. Any two strains shareing the binding domain are connected by a curve. The innermost ring indicates whether the domain is accompanied by SDH in the genome. The middle ring represents the culture collection where they come from and the outermost ring represents the taxonomic information of each strain.

4.3.7 Gain of novel functions positively associated with bacterial abundance in rhizosphere

Grouping genes into orthologues allowed us to identify the transfer of novel functions into the recipient organisms. For this analysis, we first identified orthologues involved in HGT and found to be present with single copy in at least one genome in which HGT of the orthologue was detected. We refer to these orthologs as 'orphans'. By incorporating the presence/absence of orphan orthologues and microbial composition data, we tried to identify the HGT events potentially transferring beneficial novel functions (Figure 4.14A). Though we could not identify exactly which genome in a pair is the recipient, we assumed the transfer direction was from genomes with multiple paralogue copies to those with single-copy orthologue based on the assumption that the transferred genes in the recipient genome did not have enough time to duplicate and diverge upon recent HGT. Based on this notion, we found that 45.8% (2,378/5,193) of HGT in our dataset carry orphan orthologues (Figure 4.14B), which suggests the transfer of potential novel functions in the plant-associated microbiota.

Next, we identified the orphan orthologues positively associated with bacteria abundance by correlating the presence of the orphan orthologues across all the strains with the abundance of corresponding strains in the root compartment. We found 1, 18 and 14 "orphan" orthologues positively associated with strain abundance in *AtGr*-sphere, *LjGr*-sphere and *AtIr*-sphere respectively. In *AtGr*-sphere, the orthologue belonged to periplasmic binding protein (OG0003572) showing a fold change of 2.8 in terms of the relative abundance between strains with or without this orthologue (Figure 4.14C). In *LjGr*-sphere, ammonia monooxygenase (OG0001853) displayed the strongest signal, potentially supporting bacteria colonization with a fold change of 3.3. Ammonia monooxygenase oxidizes ammonia to hydroxylamine and is the rate-limiting step in nitrification. Nitrification can produce nitrate that can be assimilated by bacteria as a nitrogen source (Merrick and Edwards 1995). This implies that bacteria may compete with plants for nitrogen resources in the root of legumes. Some T4SS genes were also found to positively correlate with strain abundance in *LjGr*-sphere, which implies that either abundant bacteria tend to possess conjugative T4SS in line with the elevated HGT frequencies in abundant bacteria or the transferred T4SS contributes to bacterial virulence and host-cell binding (Frank et al. 2005). In *AtIr*-sphere, we found the presence of an iron-cheating domain (OG0009487), potentially contributing to the

strain abundance. Since the Italian soil is alkaline and the iron is usually preent in an insoluble form that cannot be directly taken in by both microbes and plants (Harbort et al. 2020), this highlighted that bacteria can cooperate with each other *via* HGT during iron scavenging.

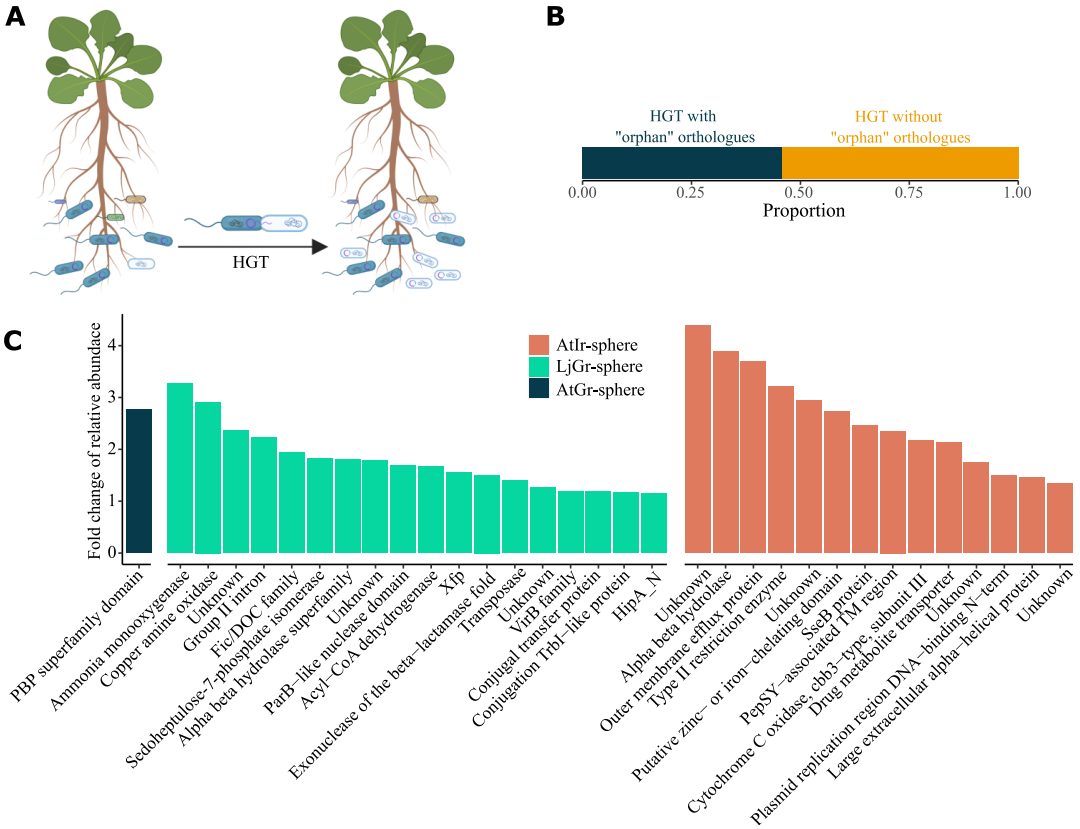


Figure 4.14 The transferred 'orphan' orthologues potentially help bacterial root colonization. (A) A conceptual diagram shows the gain of novel functions contributes to bacterial fitness in the root compartment of plants. 'Orphan' orthologues refer to the orthologues absent in recipient genomes before the acquisition of the functional orthologues via HGT, which is equivalent to gain of novel functions. (B) The proportion of HGT events transferring 'orphan' orthologues/novel functions in plant-associated microbiota. (C) 'Orphan' orthologues potentially provide beneficial outcomes to the recipients after HGT. The transferred 'orphan' orthologues whose presence in the genomes positively correlate with the relative abundance of strains in different root compartments and their corresponding fold changes of the relative abundance between strains with or without the orthologues are shown here. The correlation analysis was performed on *AtGr*-sphere, *AtIr*-sphere and *LjGr*-sphere whose microbial compositions are available.

4.3.8 Extensive dissemination of antimicrobial resistance genes in microbial communities via HTG

Plants and soil-dwelling microbes can release varieties of antibiotics to suppress the colonization of pathogens and compete for resources (Savoia 2012; Schäfer et al. 2022; Zhou et al. 2021). Under

such strong antibiotics pressure, microbes in the plant-associated environments may gradually acquire antimicrobial resistance genes (ARGs) to defend against antibiotics and cooperate with other members to stabilize the microbial communities even without additional manure treatment with antibiotics. This process is unlike the human gut microbiota, in which the dissemination of ARGs are strongly selected by antibiotic usage (Forster et al. 2022; Groussin et al. 2021). The plant-associated microbiota represents a relevant system to study the intrinsic horizontal transfer of ARGs in host-associated microbiota without external antibiotic intervention. In our dataset, we found massive ARGs transfer accounting for 8.7% (454/5,193) of total HGT, which supported our hypothesis that plant-associated microbes actively disseminated ARGs to defend against the pressure from hosts and other microbial members. ARGs were mainly transferred within *Burkholderiaceae*, *Sphingomonadaceae*, *Beijerinckiaceae*, *Xanthomonadaceae* and *Rhizobiaceae* (Figure 4.15A). In *Burkholderiaceae*, a variety of ARG families including RND_mfp, ACR_tran, MFS_1, drrA, vanS and vanR were frequently transferred. For *Sphingomonadaceae*, *Beijerinckiaceae* and *Xanthomonadaceae*, RND_mfp and ACR_tran were frequently transferred. In *Rhizobiaceae*, drrA was most frequently transferred. Across all culture collections, we found a conserved pattern of ARG transfer in rhizosphere and phyllosphere that RND_mfp and ACR_tran were most frequently transferred, but in phycosphere, MFS_1 was most frequently transferred (Figure 4.15B). This result suggests that either *Chlamydomonas* and / or phycosphere microbiota possess an antibiotics reservoir distinct from land plants and / or the rhizosphere and phyllosphere microbiota.

4.3.9 HGT of glycolate oxidase is linked to bacterial colonization of the phycosphere

Microalgae release diverse chemicals, including photoassimilates and other carbon sources, to phycosphere to attract the colonization of heterotrophic bacteria and in return the bacteria provide beneficial functions to algae (Seymour et al. 2017). From the list of most frequently transferred orthologues in the *CrG*-sphere, we found a carbohydrate metabolism-related gene – glycolate oxidase (Figure 4.8), including three subunits: *glcD*, *glcE* and *glcF*. Glycolate is reported to be produced and secreted by *Chlamydomonas reinhardtii* through photorespiration that occurs at low CO_2 and high O_2 concentrations (Bauwe et al. 2010). Conversely, bacteria can assimilate glycolate as a carbon source to support its growth via glycolate oxidase (Schada von Borzyskowski et al.

2019). There are 22 (46.8%) phycosphere strains possessing glycolate oxidase genes (*glcDEF*) (Figure 4.16A) and 8 strains might involve in horizontal transfer of *glcDEF*, including Chlamy35, Chlamy40, Chlamy76, Chlamy95, Chlamy109, Chlamy119, Chlamy183 and Chlamy187. All the 10 potentially transferred genes (2 HGT events in Chlamy76 and Chlamy187) were located on 10

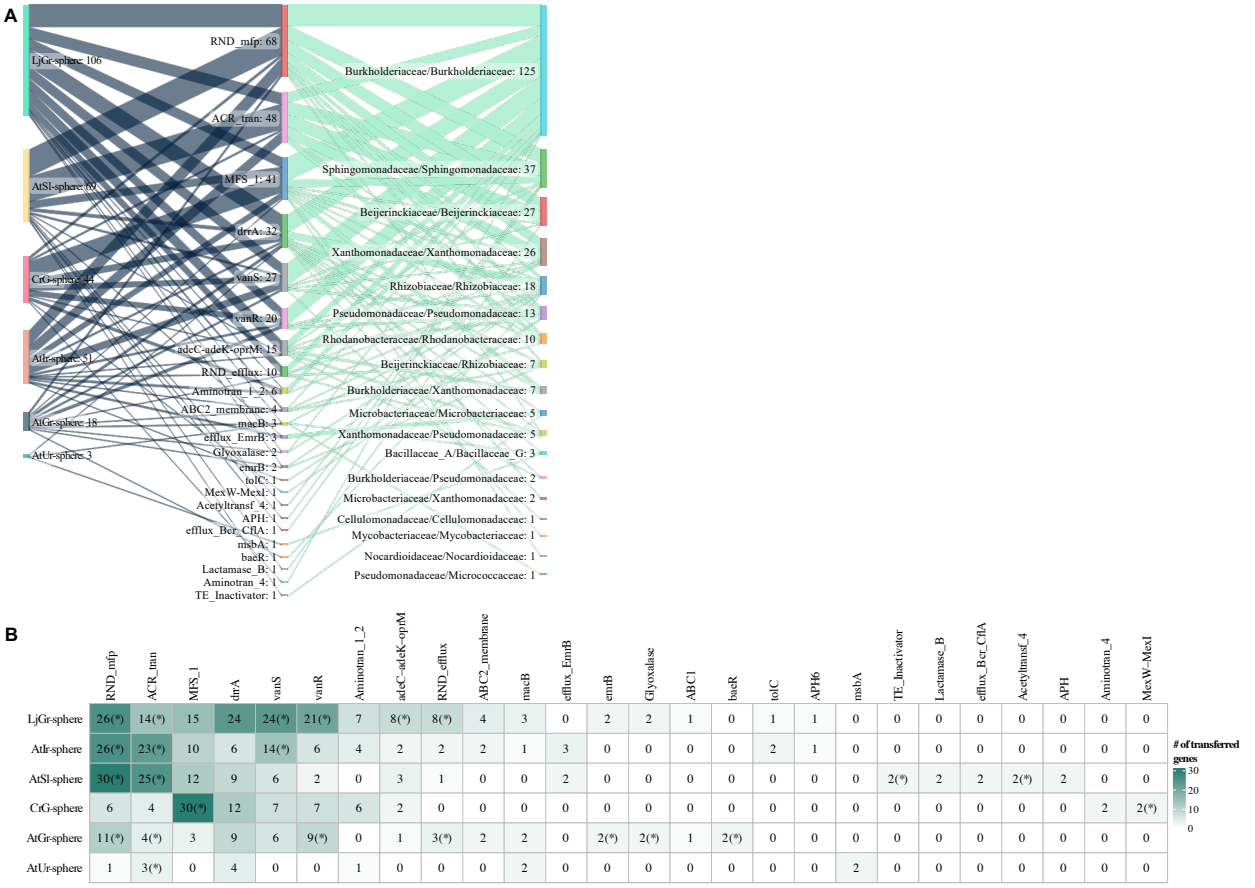


Figure 4.15 The landscape of ARGs transfer in the plant-associated microbiota. (A) The sankey plot connects the transferred ARG families and the taxonomy pairs in which the transfer is found. (B) The overall display of the ARG families transferred in each culture collection. The asterisk sign (*) indicates the enrichment of the ARG family in HGT compared to the whole-genome level.

independent transferrable plasmids, including the two mega-plasmids mentioned above. From the *glcDEF* gene tree, we observed that there were 8 transferred identical genes (Figure 4.16A), which complicated the analysis because it is not possible to determine the direction of transfer. Based on the pairwise alignment across 8 plasmid contigs, we verified 2 transfer events between Chlamy76 and Chlamy187, and transfer among Chlamy35, Chlamy119 and Chlamy183, in which they showed highest similarity to each other compared to the other contigs. Those strains were well separated in the species tree in marked contrast to the gene tree (Figure 4.16A). A clear separation

between plasmid-based and chromosome-based *glcDEF* in terms of their phylogenetic relatedness was observed (Figure 4.16A). The phylogenetic topology of chromosome-based *glcDEF* matched quite well with the phylogenetic relatedness of species, except 4 glycolate oxidase genes from *Alphaproteobacteria* and *Bacilli* that fell into the chromosome-based and plasmid-based clade of *Gammaproteobacteria* in the gene tree, respectively (Figure 4.16A), which implies that the *Alphaproteobacteria* and *Bacilli* acquired the *glcDEF* from *Gammaproteobacteria* through ancient HGT. This is most likely for *Bacilli*, which probably acquired *glcDEF* via plasmid transfer followed by the integration into their chromosomes.

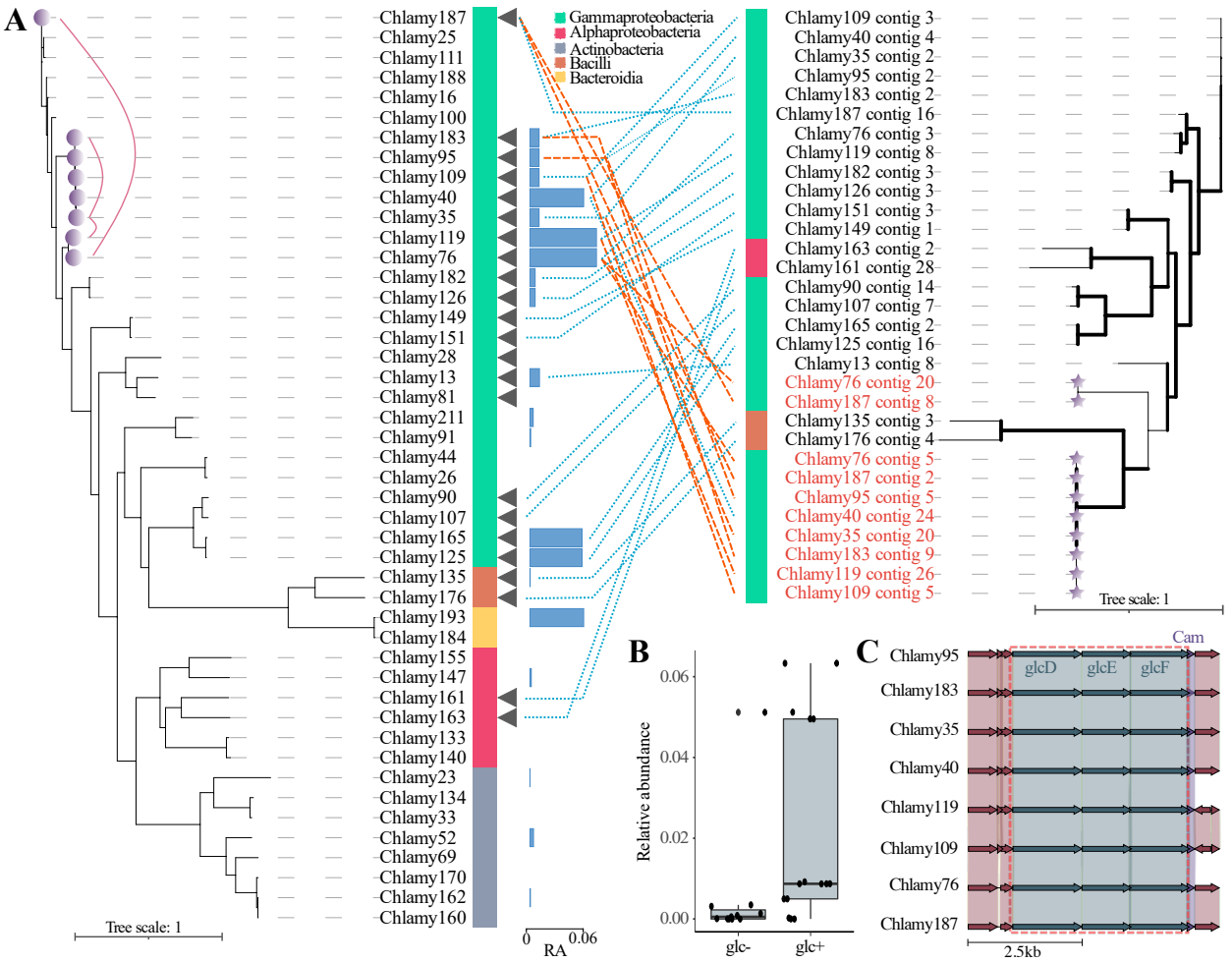


Figure 4.16 The transfer of glycolate oxidase contribute to bacterial colonization in the phycopshere. (A) The discrepancy of the species tree (left) and the *glcDEF* tree (right). On the species tree, the triangle and barplot indicate the presence of *glcDEF* in genomes and relative abundance of the strains in phycopshere respectively. The points on the tips indicate the identified transfer of *glcDEF* in the genome and curves connect two strains with confident transfer of *glcDEF* supported by the whole-contig alignment. On the gene tree, the stars on the tips of correspond to the plasmid-locating *glcDEF*. Contig names are used in the gene tree to show the origin of the *glcDEF*. The lines between

the two trees connect *glcDEF* and their genome sources. Brown lines connect plasmid-locating genes and blue lines connect chromosome-locating genes. (B) The relative abundance of the strains with or without *glcDEF* in the phycosphere microbiota (Wilcox test, *P* value < 0.05). (C) The synteny block of the plasmid-locating *glcDEF*. Cam is the abbreviation form for the gamma carbonic anhydrase.

Given the wide spread and frequent transfer of *glcDEF* in phycosphere strains, we assumed that *glcDEF* could help bacteria assimilate glycolates secreted by *Chlamydomonas* and support their growth in phycosphere. To test this hypothesis, we compared the relative abundance of strains with or without *glcDEF* in phycosphere. As expected, we found a significant increase in the relative abundance of strains with *glcDEF* (Wilcoxon test $P < 0.05$, Figure 4.16A and 4.16B). Next, we further scrutinized the functions of other genes on the 10 transferrable plasmids besides *glcDEF*. Gamma carbonic anhydrase was found to locate next the *glcDEF* (Figure 4.16C), which catalyzes the reversible hydration of CO_2 to the HCO_3^- such that may regulate the CO_2 concentration in phycosphere to mediate the production rate of glycolate by host. A large synteny block containing *glcDEF* was found to be conserved across the 10 plasmids (Figure 4.17). The synteny block was enriched in the genes related to carbon metabolism except for *glcDEF* (Figure 4.18). For instance, we found fatty acid metabolism related genes: enoyl-CoA hydratase and acyl-CoA dehydrogenase, genes producing intermediates in the TAC cycle: pyruvate dehydrogenase complex, followed by cytochrome P450 and ferredoxin that convert pyruvate into acetyl-CoA, succinate-semialdehyde (SSA) dehydrogenase converting SSA into succinate. This result suggests that these plasmids may play symbiotic roles during bacterial adaptation to the phycosphere environment.

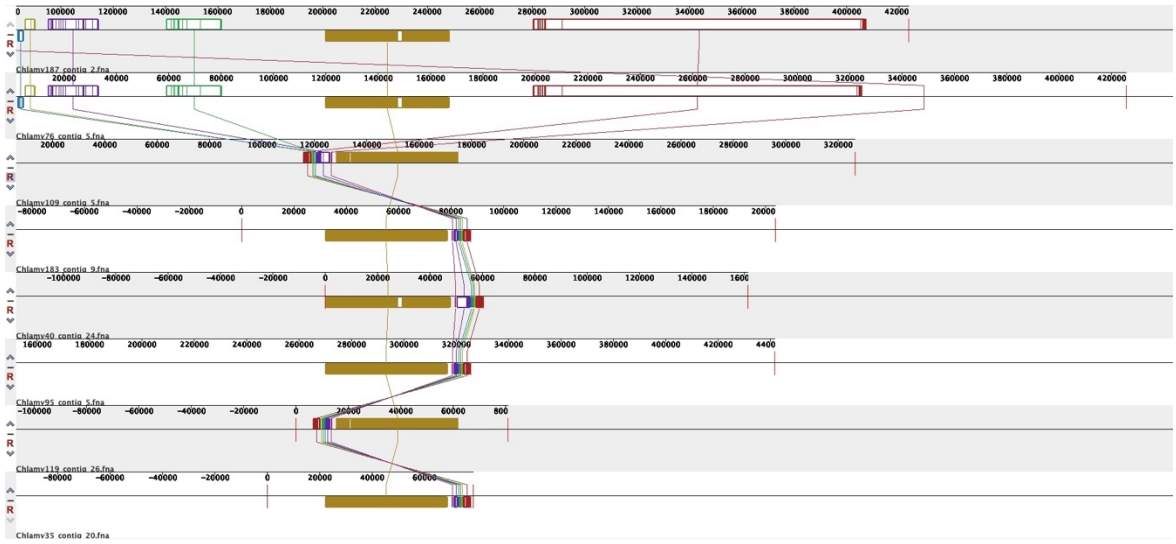


Figure 4.17 The synteny plot of the 8 plasmid contigs showing identical *glcDEF*. The colored blocks highlight the conserved backbone in the plasmid. The large synteny block conserved across the 8 plasmids are depicted in brown.

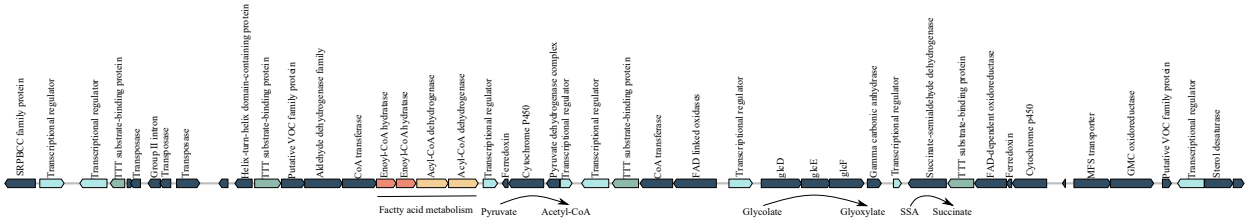


Figure 4.18 The functional annotation of the genes within the large conserved backbone shown in Figure 4.17. The synteny blocks shows an enrichment in carbohydrate metabolism-related genes.

4.4 Discussion

Plants release a variety of chemicals to the surrounding environments thus forming specific niches for microbial colonization. The shifts of lifestyle from the free-living state to plant-associated commensals is a process that determines the fate of the soil-borne bacteria in plant-associated niches. The evolution of free-living bacteria into symbionts can be facilitated by HGT contributed by other microbial members in the community, which can be driven by plant derivatives (Ling et al. 2016; Mølbak et al. 2007). The microbial symbiosis islands tend to be located on plasmids or close to mobile elements, such as ICEs and insertion sequences (Dobrindt et al. 2004), and this special feature invests them with the mobility across species. In this chapter, we systematically investigated the recent HGT in plant-associated microbiota by taking advantage of different genome-indexed bacterial culture collections we established in the past decade spanning 3 different photosynthetic host-associated niches. Unlike the soil conditions where restricted nutrients are available for dwelling bacteria, plant-associated niches are rich in varieties of secreted carbohydrates that can greatly support bacterial proliferation such that increased cell densities promote HGT incidence (Heuer and Smalla 2007), in line with the observation of higher abundance of genes affiliated to HGT processes in rhizosphere microbiota compared to bulk soil (Lopes et al. 2016). As expected, 42% of non-redundant bacterial genomes in plant-associated microbiota have shown to be involved in recent HGT. We found three factors: shared environments, phylogenetic relatedness and bacterial abundance in natural environments, can limit the incidence of HGT in plant-associated niches (Figure 4.1B, 4.1C and 4.1D). HGT between close relatives has been reported in gut and cheese microbiota (Handley et al. 2017; Smillie et al. 2011), but the transfer between phylogenetically distant bacteria is also found in our study, in which the fixation of transferred genes in recipient organisms can probably be maintained by error-prone DNA polymerases (Remigi et al. 2014). Plasmids potentially benefit the host via the coding of degradative, antibiotic resistance and metal tolerance genes (Heuer and Smalla 2012). Carriage of

plasmids is a common phenomenon in bacteria isolated from plants (Powell et al. 1993). Transferred gene clusters pinpoint the importance of plasmids in transferring gene modules in plant-associated microbiota via conjugative T4SS (Figure 4.3) and transposons can participate in the rapid evolution of transferred plasmids in recipient genomes (Figure 4.4).

From the perspective of taxa, we found a consistent pattern that abundant taxa are preferentially involved in HGT across all the niches (Figure 4.1D and 4.5). However, from the functional perspective, neither the same taxon across rhizospheric and phycospheric microhabitats, nor different taxa within the same microhabitat show a conserved pattern of the frequently transferred functional orthologues (Figure 4.6B, 4.8, 4.9, 4.10 and 4.11), which facilitate the expansion of intra-species genomic diversity and pangenomes of a given taxa and potentially provide adaptation to different environments. Previous studies highlight that soil properties explain a large portion of the variation found in root microbial communities (Schlaeppi et al. 2014), and our findings further show that the environmental setting is a major force in the selection of transferred genes. We found that the phyllosphere microbiota is relatively consistent in terms of HGT. Many frequently transferred orthologues were found in abundant taxa (Figure 4.12), and their in-depth analysis led us to the finding that the sulfite dehydrogenase and its molybdopterin binding domain are often transferred between the genomes obtained from this environment (Figure 4.13). Cumulative evidence supports the induction of sulfur metabolism related genes during bacteria colonizing plant leaves and consolidated the hypothesis of sulfur limitation on leaves (Müller et al. 2016a). Our finding further supports this hypothesis and suggests that bacteria can interact with each other to scavenge sulfur by HGT.

It has long been known that administration of antibiotics can accelerate the dissemination of ARGs in microbes, especially in the human gut microbiota, but there is a paucity of knowledge in how ARGs are transferred in natural host-associated microbiomes without anthropogenic antibiotics intervention. We found 8.7% of HGT to be ARGs in plant-associated environments, however, in a recent study of HGT in the human gut microbiota (Forster et al. 2022), the authors identified 16.5% of HGT to be ARGs. This supports that additional antibiotics usage indeed accelerates dissemination of ARGs in host-associated environments but it also highlights the intrinsic prevalence of ARGs transfer in natural microbial communities, which is far overlooked in current studies, and prompts that intrinsic agents of dissemination of ARGs should also be considered during the assessment of the risk of ARGs transfer.

Nevertheless, bacteria do not live alone. Instead, they interact with other counterparts in plant-associated niches, including fungi, plants and animals. A recent study reported the bidirectional HGT between plant-associated microbiota and *Arabidopsis thaliana*, in which the authors reported bacterial *DET2* transferred from the host plant could replace plant homologue in brassinosteroid production (Haimlich et al. 2022), therefore it is important to elucidate cross-kingdom HGT for better understanding the co-evolution of microbiota and its hosts. Taken together, identification of recent HGT in plant-associated microbiota allow us to identify potential mechanisms of interaction among bacterial members, adaptation to a host, to trace back patterns of microbial evolution, and to expand our knowledge of plant-associated microbiota assembly.

4.5 Materials and methods

Bacterial culture collections from plant-associated environments

Based on the great efforts on bacterial isolation from plant-associated environments, including rhizosphere, phyllosphere and phycosphere, made by our former colleagues, we established different bacterial culture collections, which encompass 194 and 236 root isolates from vascular model plant *Arabidopsis thaliana* grown in Cologne CAS soil and Italy (Bai et al. 2015; Harbort et al. 2020), respectively, 292 root isolates from model legume *Lotus Japonicus* grown in Cologne CAS soil (Wippel et al. 2021), 185 phycosphere isolates from *Chlamydomonas reinhardtii* grown in Cologne CAS soil (Durán et al. 2022b), which serves as a unicellular model organism for photosynthesis study, and 206 leaf isolates from *Arabidopsis thaliana* grown in Switzerland and Germany (Bai et al. 2015). All of the bacterial isolates were whole-genome sequenced and assembled, which provides an opportunity to have an insight into the microbial functional genes responsible for bacterial adaptation to host plants and mediating the community assemblage. Besides the 5 above-mentioned culture collections, we also included another culture collection assembled from *Arabidopsis thaliana* in the USA by Jeff Dangle's lab (Levy et al. 2018), which contains 156 isolates. Given the fact that each culture collection contains bacteria isolated from the same plant species grown in the same field and they formed robust microbial community structures distinguishable from that of surrounding bulk soil without the growth of plants, those bacteria can form physical contact with others and potentially transfer genetic elements horizontally to each other. It has long been known that HGT can help the microbe recipients to adapt to new changing environments and several studies have elucidated the extensive HGT in

human microbiota (Groussin et al. 2021; Smillie et al. 2011). By taking advantage of the sequence-indexed culture collections we have, it allows us to identify the HGT events in plant-associated environments and search for the microbial genes essential for bacterial host adaptation.

Preprocess of the bacterial genomes

To eliminate the contaminations in the genome assemblies, all the 1,269 bacterial genomes underwent quality check by CheckM (Parks et al. 2015) with the threshold of 90% completeness and 5% contamination and all of the genomes passed the quality check. Stepwise dereplication of genomes were conducted. Firstly, the genomes were dereplicated by super high similarities (-pa 0.999 --SkipSecondary) to agglomerate the duplicate genomes from the same strain with dRep (Olm et al. 2017), here we call the retained 868 genomes as non-redundant genomes as shown in Figure 4.1A. In the second step, we clustered the non-replicated genomes into species clusters based on 95% identity and 90% covergae (-pa 0.9 -sa 0.95 -nc 0.30 -cm larger) (Almeida et al. 2021a).

Detection of HGT

PHIX phage contamination is widely spread in the genome assembly cause this phage works as a spike-in sequence in Illumina sequencing (Mukherjee et al. 2015). Moreover, to avoid the viral genes in bacteria infected by the same bacteriophage, we removed the viral genes found in our database by mapping against the viral Refseq database (O'Leary et al. 2016) by Diamond (Buchfink et al. 2014) with the threshold of 80% identity and 75% coverage. This step removed the potential false positives of HGT events caused by the prevalent diffusion of identical viral genes in our assemblies and only focused on bacterial-origin HGT. To note, some viral genes can facilitate the horizontal transfer of host genes when they loop out of the host genome and carry some neighboring host genes. By conducting the viral decontamination, we're still capable of detecting the transfer of bacterial genes initiated by viral integration because we only abandon the viral parts.

We used MMseqs2 (Mirdita et al. 2019) to cluster the genes (≥500bp) from non-redundant bacteria based on 99.9% identity. Only a pair of genes from the same cluster but originate from two bacterial species were identified as a potential horizontal gene transfer event. Here, we only focused on genes with a minimum length of 500bp, which took up 76.1±0.1% of the genomes on

average, so we only excluded a small portion of shorter genes in our analysis. In this study, a HGT event refers to a pair of potentially transferred genes showing high similarities.

To test whether the identity threshold is robust in identifying recent HGT and incompetent in falsely calling strictly vertically transferred genes, we performed a pairwise comparison of the ribosomal protein genes between strains from different species since they are all strictly vertically transferred. The ribosomal protein-coding genes were identified by blasting protein sequences against the RiboDB database (Jauffrit et al. 2016) to search for the 61 universal single-copy ribosomal protein families by using diamond (with parameters --query-cover 75 --id 70) (Buchfink et al. 2014). Ribosomal genes bTHX, cs23, el30, el43, el8, es27 and es30 were excluded because they were not broadly distributed in our dataset. Any pair of genes belonging to the same ribosomal protein family in two strains from different species were aligned to obtain the sequence similarity. Within our expectation, none of the sequence similarity of ribosomal genes exceeds the 99.9% percent identity (Figure 4.19), which supports that the threshold used for HGT detection is sufficient to identify the recent HGT events.

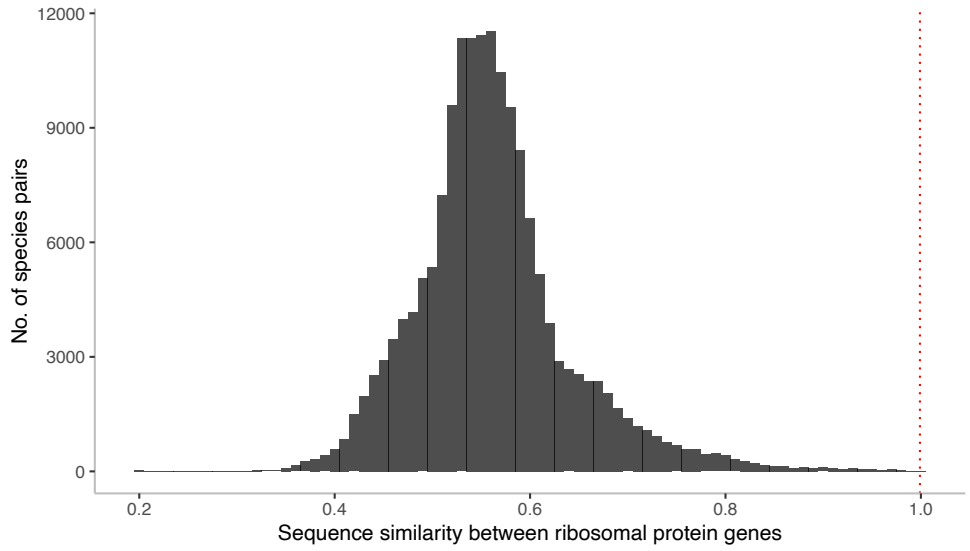


Figure 4.19 The pairwise comparison of vertically transferred ribosomal protein genes. The ribosomal protein genes are extracted as described above (See *Materials and methods*). Pairwise sequence comparison of the ribosomal protein genes from the same protein family is performed between any pair of genomes from different species clusters.

Functional annotation of genomes

Functional capacity of the genomes was identified with EGGNOG mapper (Cantalapiedra et al. 2021). The complete conjugative T4SS in each genome was predicted by m(Abby et al. 2014)acsyfinder2.0 and the TXSScan database. Macsyfinder2 searches for the essential genes for different secretion systems and their genomic proximity to identify the complete secretion system. Antimicrobial resistance genes were annotated by hmmer3 v3.1b2 (Finn et al. 2011) was used with the Resfam hmm database (Gibson et al. 2015) with a cutoff e-value of 1e-5 and score of 80. To better understand the functional relationships of the transferred genes, we clustered genes across all the genomes into orthogroups by Orthofinder (Emms and Kelly 2019). The genes from the same orthogroups are thought to execute similar functions and make it possible to identify which functional orthologues were frequently transferred in plant-associated microbiota.

Identification of transferred gene clusters

The transferred genes were clustered into gene clusters based on genomic proximity if they met the following two requirements: 1) The genes were located on the same contig; 2) Any two neighboring genes in a cluster were separated by at most 10 genes. This results in 553 singleton genes that were not clustered with any other genes and 545 gene clusters (≥2 transferred genes) containing 5546 genes. To probe whether the transferred gene clusters were located on plasmids, we combined the predicted plasmid contigs from two software platon and plasflow (Krawczyk et al. 2018). Platon uses plasmid-specific features and proteins to identify plasmid contigs and plasflow uses neural network models to predict plasmid contigs that can complement the results from platon.

Phylogenetic tree construction

Non-redundant genomes were fed into Phylophlan (Asnicar et al. 2020) to construct the phylogenetic tree according to the multiple-sequence alignment of 400 universal single-copy marker genes in each genome, which were internally inferred by Phylophlan. The glcDEF genes were extracted from phycosphere genomes and aligned with mafft (Katoh et al.), followed by tree construction with Fasttree (Price et al. 2010). The nucleotide sequences of OG0000186 were extracted from all the non-redundant genomes and the tree was constructed in the same way as glcDEF.

Assigning relative abundance to strains

To evaluate the relationship between HGT and strain abundance, we retrieved the microbial community structures of *Arabidopsis thaliana* root samples collected from Cologne CAS soil and Italy, *Lotus japonicus* root samples from Cologne CAS soil and phycosphere samples from flask experiments where the strains were originally isolated (Bai et al. 2015; Durán et al. 2022b; Harbort et al. 2020; Wippel et al. 2021). The V5-V7 regions of the strains were mapped to the ASV representative sequences in the corresponding samples based on 100% identity with Usearch (Edgar 2010) and the average relative abundance of the mapped ASV was assigned to the strain to represent its abundance in the relevant environment. To compare the HGT network with the co-occurrence network, SparCC (Friedman and Alm 2012) was used to construct the co-occurrence network among the mapped ASVs.

Statistical analysis

To test which functional categories and ARGs families were enriched in HGT rather than by random transfer, we first calculated the percentage of genes belonging to each category across all of the non-redundant genomes in each culture collection. Then a *binomial* distribution was used to calculate the significance of enrichment of genes from each category in HGT compared the genome background by the following formula:

$$p\{X \geq N\} = 1 - \sum_{0}^{N-1} \binom{n}{r} p^n (1-p)^{r-n}$$

where *N* represents the number of genes from a category involved in HGT in a culture collection, *r* represents the total number of genes involved in HGT in a culture collection, *p* is the percentage of genes belonging to a specific category across all of the non-redundant genomes in a culture collection. The same method was used to evaluate the enrichment of positive/depletion of negative interactions in strain pairs exchanging genes in each culture collection by comparing to the whole co-occurrence network of the corresponding environments. Wilcoxon text was used to compare the relative abundance of phycosphere strains with glcDEF and without glcDFG. PhyloGLM (Ives and Garland 2010) was used to find the “orphan” orthologues whose presence and frequencies positively correlated with the abundance of the strains in each culture collections to alleviate the effect of phylogenetic relatedness of strains because close strains often resemble each other in terms of physiology and phenotypes.

4.6 Acknowledgements

I would like to thank Yang Bai and Daniel Müller for establishing the *AtGr*-sphere culture collection; thank Kathrin Wippel and Ke Tao for establishing the *LjGr*-sphere culture collection; thank CJ Harbort and Masayoshi Hashimoto for establishing the *AtIr*-sphere culture collection; thank Paloma Duran and José Flores-Uribe for establishing the *CrG*-sphere culture collection.

Outlook

Microbial compositions of the plant-associated microbiota and the potential driving factors of community assemblages have been extensively studied in nature with amplicon sequencing and metagenomic sequencing in the past decade. Now researchers are moving from basic understanding of the microbiota structure to the reveal of the molecular mechanisms underlying the assemblage process in nature, which can further guide the manipulation of the microbiota to achieve the desired phenotypes. Accumulative studies have utilized simplified SynComs to study microbial ecologies by adding another layer of information from microbe-microbe interactions. One dilemma frequently argued about SynCom experiments is that they cannot represent the complexity of microbes in natural environments such that we are taking the risks of missing some interactions, though we have set out to strive for a holistic overview in the way of transferring from the study of mono-association to the multitudinous associations. Nevertheless, the increase of complexity of SynComs is greatly attributed to the taxonomic diversity of strains in culture collections. With the inclusion of close strains or sub-strains that are insufficient to be differentiated by 16S rRNA gene or ITS into SynComs, another intractable issue is how to distinguish one from the others easily in a community. The SynCom experimentation set-up in gnotobiotic systems requires strict quality controls to exclude any unexpected biological contaminants that potentially dramatically influence the robustness and readouts of experiments. Rbec, presented in *Chapter 2*, provides a novel function to detect contaminants in SynCom samples based on the high proportion of uncorrected reads by any of the reference sequences.

Greater efforts should be put into the establishment of culture collections from diverse environments and hosts given local adaptation to specific environments and hosts as discussed in *Chapter 3.1*, which alerts us that case-specific culture collections would be required to study microbial adaptation. Due to the high proportions of hard-to-culture bacteria in soil (Steen et al. 2019), the taxa in established culture collections bias towards easy-to-culture and fast-growing bacteria, which underestimates the taxonomic and functional diversities of the plant microbiota by using the reductionist approach for microbiome studies. Culture-independent metagenomic sequencing offers the opportunity to recover microbial genomes in a taxonomically unbiased way, termed metagenome-assembled genomes (MAGs), which has been extensively used in the human microbiomes (Almeida et al. 2021b; Almeida et al. 2019; Nayfach et al. 2019; Pasolli et al. 2019). Considering the far greater diversity of plant-associated microbiomes, high-cost deep metagenomic sequencing is required to recover MAGs from those samples. Fully understanding the functional capacities of the plant

microbiota necessitates the inputs from a multitude of research communities interested in this area. An integrated genome database for the plant microbiota with well-documented information, e.g. isolation sources or sampling sites, host plants and soil conditions, further facilitates the study of genomic contexts mediating microbial adaptations to plants, evolutionary trajectories, population dynamics within a given taxon and the usability of metagenomic reads.

The aim of the plant microbiota research is to translate the theoretical findings into practical applications. Current studies mainly rely on mono-association between a single microbe and its host or multi-association in a reduced synthetic microbial gnotobiotic system, one difficulty in translating is that how these findings can be applied to natural environments/fields where the system is far more complex than the laboratory condition. Targeted genome editing in microbiome *in situ* is a new technique (Rubin et al. 2021), which shows a promising and strong application to manipulate the microbial communities in nature. For instance, researchers can expect to genetically perturb some strains to allow the colonization and persistence of other beneficial commensals in the plant-associated niches. That being said, this CRISPR-Cas system still needs further validation and evaluation before it can be fully applied to natural environments. Continuous innovations in techniques, algorithmic breakthrough in data analysis of omics and explosion of relevant genomic data will tremendously aid in the molecular study of the plant microbiota and contribute to the sustainable agriculture.

References

Abby, S.S., Néron, B., Ménager, H., Touchon, M., Rocha, E.P.C. (2014). MacSyFinder: A program to mine genomes for molecular systems with an application to CRISPR-Cas systems Public Library of Science), 9.

Ailloud, F., Lowe, T., Cellier, G., Roche, D., Allen, C., Prior, P. (2015). Comparative genomic analysis of Ralstonia solanacearum reveals candidate genes for host specificity 16, 270.

Almario, J., Mahmoudi, M., Kroll, S., Agler, M., Placzek, A., Mari, A., et al. (2022). The Leaf Microbiome of Arabidopsis Displays Reproducible Dynamics and Patterns throughout the Growing Season American Society for Microbiology), 13.

Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., et al. (2019). A new genomic blueprint of the human gut microbiota 568, 499–504.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., et al. (2021a). A unified catalog of 204,938 reference genomes from the human gut microbiome Nature Research), 39, 105–114.

Almeida, A., Nayfach, S., Boland, M., Strozzi, F., Beracochea, M., Shi, Z.J., et al. (2021b). A unified catalog of 204,938 reference genomes from the human gut microbiome 39, 105–114.

Amin, S.A., Hmelo, L.R., van Tol, H.M., Durham, B.P., Carlson, L.T., Heal, K.R., et al. (2015). Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria 522, 98–101.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J.T., Zech Xu, Z., et al. (2017). Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns 2.

Armanhi, J.S.L., de Souza, R.S.C., Damasceno, N. de B., de Araújo, L.M., Imperial, J., Arruda, P. (2018). A Community-Based Culture Collection for Targeting Novel Plant Growth-Promoting Bacteria from the Sugarcane Microbiome 8.

Asnicar, F., Thomas, A.M., Beghini, F., Mengoni, C., Manara, S., Manghi, P., et al. (2020). Precise phylogenetic analysis of microbial isolates and genomes from metagenomes using PhyloPhlAn 3.0 Nature Research), 11.

Bai, Y., Müller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., et al. (2015). Functional overlap of the Arabidopsis leaf and root microbiota Nature Publishing Group), 528, 364–369.

Batstone, R.T., O'Brien, A.M., Harrison, T.L., Frederickson, M.E. (2020a). Experimental evolution makes microbes more cooperative with their local host genotype 370, 476–478.

Batstone, R.T., O'brien, A.M., Harrison, T.L., Frederickson, M.E. (2020b). Experimental evolution makes microbes more cooperative with their local host genotype 370, 476–478.

Bauwe, H., Hagemann, M., Fernie, A.R. (2010). Photorespiration: players, partners and origin pp. 330–336.

Beilsmith, K., Thoen, M.P.M., Brachi, B., Gloss, A.D., Khan, M.H., Bergelson, J. (2019). Genome-wide association studies on the phyllosphere microbiome: Embracing complexity in host-microbe interactions 97, 164–181.

Berendsen, R.L., Pieterse, C.M.J., Bakker, P.A.H.M. (2012). The rhizosphere microbiome and plant health 17, 478–486.

Berendsen, R.L., Vismans, G., Yu, K., Song, Y., de Jonge, R., Burgman, W.P., et al. (2018). Disease-induced assemblage of a plant-beneficial bacterial consortium 12, 1496–1507.

Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M., Vorholt, J.A. (2014). A Synthetic Community Approach Reveals Plant Genotypes Affecting the Phyllosphere Microbiota 10, e1004283.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., et al. (2019). Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2 37, 852–857.

Bouffaud, M., KYSELKOVÁ, M., GOUESNARD, B., GRUNDMANN, G., MULLER, D., MOËNNE-LOCCOZ, Y. (2012). Is diversification history of maize influencing selection of soil bacteria by roots? 21, 195–206.

Brito, I.L. (2021). Examining horizontal gene transfer in microbial communities 19, 442–453.

Buchfink, B., Xie, C., Huson, D.H. (2014). Fast and sensitive protein alignment using DIAMOND Nature Publishing Group), pp. 59–60.

Bulgarelli, D., Rott, M., Schlaeppi, K., ver Loren van Themaat, E., Ahmadinejad, N., Assenza, F., et al. (2012). Revealing structure and assembly cues for Arabidopsis root-inhabiting bacterial microbiota pp. 91–95.

Bulgarelli, D., Schlaeppi, K., Spaepen, S., van Themaat, E.V.L., Schulze-Lefert, P. (2013). Structure and functions of the bacterial microbiota of plants pp. 807–838.

Butler, M.I., Stockwell, P.A., Black, M.A., Day, R.C., Lamont, I.L., Poulter, R.T.M. (2013). Pseudomonas syringae pv. actinidiae from Recent Outbreaks of Kiwifruit Bacterial Canker Belong to Different Clones That Originated in China 8, e57464.

Cairns, J., Jokela, R., Becks, L., Mustonen, V., Hiltunen, T. (2020). Repeatable ecological dynamics govern the response of experimental communities to antibiotic pulse perturbation 4, 1385–1394.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P. (2016). DADA2: High-resolution sample inference from Illumina amplicon data 13, 581–583.

Cantalapiedra, C.P., Herņandez-Plaza, A., Letunic, I., Bork, P., Huerta-Cepas, J. (2021). eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale Oxford University Press), 38, 5825–5829.

Carlström, C.I., Field, C.M., Bortfeld-Miller, M., Müller, B., Sunagawa, S., Vorholt, J.A. (2019). Synthetic microbiota reveal priority effects and keystone strains in the Arabidopsis phyllosphere 3, 1445–1454.

Carvalho, F.M., Souza, R.C., Barcellos, F.G., Hungria, M., Tereza, A., Vasconcelos, R. (2010). Genomic and evolutionary comparisons of diazotrophic and pathogenic bacteria of the order Rhizobiales [Internet]

Castrillo, G., Teixeira, P.J.P.L., Paredes, S.H., Law, T.F., de Lorenzo, L., Feltcher, M.E., et al. (2017). Root microbiota drive direct integration of phosphate stress and immunity 543, 513–518.

Ciancio, A., Pieterse, C.M.J., Mercado-Blanco, J. (2019). Editorial: Harnessing Useful Rhizosphere Microorganisms for Pathogen and Pest Biocontrol - Second Edition 10.

Crits-Christoph, A., Diamond, S., Butterfield, C.N., Thomas, B.C., Banfield, J.F. (2018). Novel soil bacteria possess diverse genes for secondary metabolite biosynthesis 558, 440–444.

Deng, S., Caddell, D.F., Xu, G., Dahlen, L., Washington, L., Yang, J., et al. (2021). Genome wide association study reveals plant loci controlling heritability of the rhizosphere microbiome 15, 3181–3194.

Dobrindt, U., Hochhut, B., Hentschel, U., Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms pp. 414–424.

Durán, P., Flores-Uribe, J., Wippel, K., Zhang, P., Guan, R., Melkonian, B., et al. (2022a). Shared features and reciprocal complementation of the Chlamydomonas and Arabidopsis microbiota 13, 406.

Durán, P., Flores-Uribe, J., Wippel, K., Zhang, P., Guan, R., Melkonian, B., et al. (2022b). Shared features and reciprocal complementation of the Chlamydomonas and Arabidopsis microbiota Nature Research), 13.

Durán, P., Thiergart, T., Garrido-Oter, R., Agler, M., Kemen, E., Schulze-Lefert, P., et al. (2018). Microbial Interkingdom Interactions in Roots Promote Arabidopsis Survival 175, 973-983.e14.

Durrant, M.G., Li, M.M., Siranosian, B.A., Montgomery, S.B., Bhatt, A.S. (2020). A Bioinformatic Analysis of Integrative Mobile Genetic Elements Highlights Their Role in Bacterial Adaptation 27, 140-153.e9.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST 26, 2460–2461.

Edgar, R.C., Flyvbjerg, H. (2015). Error filtering, pair assembly and error correction for next-generation sequencing reads 31, 3476–3482.

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

Emms, D.M., Kelly, S. (2019). OrthoFinder: Phylogenetic orthology inference for comparative genomics BioMed Central Ltd.), 20.

Escudero-Martinez, C., Coulter, M., Alegria Terrazas, R., Foito, A., Kapadia, R., Pietrangelo, L., et al. (2022). Identifying plant genes shaping microbiota composition in the barley rhizosphere 13, 3443.

Finn, R.D., Clements, J., Eddy, S.R. (2011). HMMER web server: Interactive sequence similarity searching 39.

Forster, S.C., Liu, J., Kumar, N., Gulliver, E.L., Gould, J.A., Escobar-Zepeda, A., et al. (2022). Strain-level characterization of broad host range mobile genetic elements transferring antibiotic resistance from the human microbiome Nature Research), 13.

Frank, A.C., Alsmark, C.M., Thollesson, M., Andersson, S.G.E. (2005). Functional divergence and horizontal transfer of type IV secretion systems 22, 1325–1336.

Friedman, J., Alm, E.J. (2012). Inferring Correlation Networks from Genomic Survey Data Public Library of Science), 8.

Fu, H., Uchimiya, M., Gore, J., Moran, M.A. (2020). Ecological drivers of bacterial community assembly in synthetic phycospheres 117, 3656–3662.

Garrido-Oter, R., Nakano, R.T., Dombrowski, N., Ma, K.W., McHardy, A.C., Schulze-Lefert, P. (2018). Modular Traits of the Rhizobiales Root Microbiota and Their Evolutionary Relationship with Symbiotic Rhizobia Cell Press), 24, 155-167.e5.

Gibson, M.K., Forsberg, K.J., Dantas, G. (2015). Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology Nature Publishing Group), 9, 207–216.

Groussin, M., Poyet, M., Sistiaga, A., Kearney, S.M., Moniz, K., Noel, M., et al. (2021). Elevated rates of horizontal gene transfer in the industrialized human microbiome Elsevier B.V.), 184, 2053-2067.e18.

Guo, X., Zhang, X., Qin, Y., Liu, Y.-X., Zhang, J., Zhang, N., et al. (2020). Host-Associated Quantitative Abundance Profiling Reveals the Microbial Load Variation of Root Microbiome 1, 100003.

Haimlich, S., Fridman, Y., Khandal, H., Savaldi-Goldstein, S., Levy, A. (2022). Widespread horizontal gene transfer between plants and their microbiota

Handley, K., Bonham, K.S., Wolfe, B.E., Dutton, R.J. (2017). Extensive horizontal gene transfer in cheese-associated bacteria

Harbort, C.J., Hashimoto, M., Inoue, H., Niu, Y., Guan, R., Rombolà, A.D., et al. (2020). Root-Secreted Coumarins and the Microbiota Interact to Improve Iron Nutrition in Arabidopsis Cell Press), 28, 825-837.e6.

Hartman, K., van der Heijden, M.G., Roussely-Provent, V., Walser, J.-C., Schlaeppi, K. (2017). Deciphering composition and function of the root microbiome of a legume plant 5, 2.

Heuer, H., Smalla, K. (2007). Horizontal gene transfer between bacteria pp. 3–13.

Heuer, H., Smalla, K. (2012). Plasmids foster diversification and adaptation of bacterial populations in soil pp. 1083–1104.

Horňák, K., Kasalický, V., Šimek, K., Grossart, H.-P. (2017). Strain-specific consumption and transformation of alga-derived dissolved organic matter by members of the *Limnohabitans* -C and *Polynucleobacter* -B clusters of *Betaproteobacteria* 19, 4519–4535.

Hou, S., Thiergart, T., Vannier, N., Mesny, F., Ziegler, J., Pickel, B., et al. (2021). A microbiota–root–shoot circuit favours Arabidopsis growth over defence under suboptimal light 7, 1078–1092.

Hu, L., Robert, C.A.M., Cadot, S., Zhang, X., Ye, M., Li, B., et al. (2018). Root exudate metabolites drive plant-soil feedbacks on growth and defense by shaping the rhizosphere microbiota 9, 2738.

Ives, A.R., Garland, T. (2010). Phylogenetic logistic regression for binary dependent variables 59, 9–26.

Jauffrit, F., Penel, S., Delmotte, S., Rey, C., de Vienne, D.M., Gouy, M., et al. (2016). RiboDB Database: A Comprehensive Resource for Prokaryotic Systematics Oxford University Press), 33, 2170–2172.

Karasov, T.L., Almario, J., Friedemann, C., Ding, W., Giolai, M., Heavens, D., et al. (2018). Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales 24, 168-179.e4.

Katoh, K., Misawa, K., Kuma, K.-I., Miyata, T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform

Kent, A.G., Vill, A.C., Shi, Q., Satlin, M.J., Brito, I.L. (2020). Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C 11, 4379.

Khan Chowdhury, MD.E., Jeon, J., Ok Rim, S., Park, Y.-H., Kyu Lee, S., Bae, H. (2017). Composition, diversity and bioactivity of culturable bacterial endophytes in mountain-cultivated ginseng in Korea 7, 10098.

Kim, B.-H., Ramanan, R., Cho, D.-H., Oh, H.-M., Kim, H.-S. (2014). Role of Rhizobium, a plant growth promoting bacterium, in enhancing algal biomass through mutualistic interaction 69, 95–105.

Kisker, C., Schindelin, H., Pacheco, A., Wehbi, W.A., Garrett, R.M., Rajagopalan, K. v, et al. (1997). Molecular Basis of Sulfite Oxidase Deficiency from the Structure of Sulfite Oxidase enzymes have been sequenced. The enzyme is located in the mitochondrial intermembrane space and was found to be a homodimer with a molecular mass be

Koch, M., Delmotte, N., Rehrauer, H., Vorholt, J.A., Pessi, G., Hennecke, H. (2010). Rhizobial Adaptation to Hosts, a New Facet in the Legume Root-Nodule Symbiosis 23, 784–790.

Kolmer, J.A. (2005). Tracking wheat rust on a continental scale 8, 441–449.

Krawczyk, P.S., Lipinski, L., Dziembowski, A. (2018). PlasFlow: predicting plasmid sequences in metagenomic data using genome signatures Oxford University Press), 46, E35.

Kudjordjie, E.N., Sapkota, R., Steffensen, S.K., Fomsgaard, I.S., Nicolaisen, M. (2019). Maize synthesized benzoxazinoids affect the host associated microbiome 7, 59.

Kwak, M.-J., Kong, H.G., Choi, K., Kwon, S.-K., Song, J.Y., Lee, J., et al. (2018). Rhizosphere microbiome structure alters to enable wilt resistance in tomato 36, 1100–1109.

Lambers, H., Mougel, C., Jaillard, B., Hinsinger, P. (2009). Plant-microbe-soil interactions in the rhizosphere: an evolutionary perspective 321, 83–115.

Langridge, G.C., Fookes, M., Connor, T.R., Feltwell, T., Feasey, N., Parsons, B.N., et al. (2015). Patterns of genome evolution that have accompanied host adaptation in *Salmonella* 112, 863–868.

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., et al. (2018). Genomic features of bacterial adaptation to plants Nature Publishing Group), 50, 138–150.

Li, E., de Jonge, R., Liu, C., Jiang, H., Friman, V.-P., Pieterse, C.M.J., et al. (2021). Rapid evolution of bacterial mutualism in the plant rhizosphere 12, 3829.

Ling, J., Wang, H., Wu, P., Li, T., Tang, Y., Naseer, N., et al. (2016). Plant nodulation inducers enhance horizontal gene transfer of Azorhizobium caulinodans symbiosis island National Academy of Sciences), 113, 13875–13880.

Ling, N., Wang, T., Kuzyakov, Y. (2022). Rhizosphere bacteriome structure and functions Nature Research), 13.

Lopes, L.D., Pereira e Silva, M. de C., Andreote, F.D. (2016). Bacterial abilities and adaptation toward the rhizosphere colonization Frontiers Media S.A.), 7.

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

Ma, K.-W., Niu, Y., Jia, Y., Ordon, J., Copeland, C., Emonet, A., et al. (2021). Coordination of microbe–host homeostasis by crosstalk with plant innate immunity 7, 814–825.

Magoc, T., Salzberg, S.L. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies 27, 2957–2963.

Mallott, E.K., Amato, K.R. (2021). Host specificity of the gut microbiome 19, 639–653.

Mathesius, U. (2018). Flavonoid Functions in Plants and Their Interactions with Other Organisms 7, 30.

McCarty, N.S., Ledesma-Amaro, R. (2019). Synthetic Biology Tools to Engineer Microbial Communities for Biotechnology 37, 181–197.

Merrick, M.J., Edwards, R.A. (1995). Nitrogen Control in Bacteria [Internet]

Mesny, F., Miyauchi, S., Thiergart, T., Pickel, B., Atanasova, L., Karlsson, M., et al. (2021). Genetic determinants of endophytism in the Arabidopsis root mycobiome 12, 7227.

Mirdita, M., Steinegger, M., Söding, J. (2019). MMseqs2 desktop and local web server app for fast, interactive sequence searches Oxford University Press), 35, 2856–2858.

Miyauchi, S., Kiss, E., Kuo, A., Drula, E., Kohler, A., Sánchez-García, M., et al. (2020). Large-scale genome sequencing of mycorrhizal fungi provides insights into the early evolution of symbiotic traits 11, 5125.

Mølbak, L., Molin, S., Kroer, N. (2007). Root growth and exudate production define the frequency of horizontal plasmid transfer in the Rhizosphere 59, 167–176.

Mukherjee, S., Huntemann, M., Ivanova, N., Kyrpides, N.C., Pati, A. (2015). Large-scale contamination of microbial isolate genomes by illumina Phix control BioMed Central Ltd), 10.

Müller, D.B., Schubert, O.T., Röst, H., Aebersold, R., Vorholt, J.A. (2016a). Systems-level Proteomics of Two Ubiquitous Leaf Commensals Reveals Complementary Adaptive Traits for Phyllosphere Colonization 15, 3256–3269.

Müller, D.B., Vogel, C., Bai, Y., Vorholt, J.A. (2016b). The Plant Microbiota: Systems-Level Insights and Perspectives 50, 211–234.

Nayfach, S., Shi, Z.J., Seshadri, R., Pollard, K.S., Kyrpides, N.C. (2019). New insights from uncultivated genomes of the global human gut microbiome 568, 505–510.

Oksanen, A.J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., Mcglinn, D., et al. (2020). vegan: Community Ecology Package

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016). Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation Oxford University Press), 44, D733–D745.

Olm, M.R., Brown, C.T., Brooks, B., Banfield, J.F. (2017). DRep: A tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication Nature Publishing Group), 11, 2864–2868.

Palmer, J.D., Foster, K.R. (2022). Bacterial species rarely work together 376, 581–582.

Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W. (2015). CheckM: Assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes Cold Spring Harbor Laboratory Press), 25, 1043–1055.

Pascale, A., Proietti, S., Pantelides, I.S., Stringlis, I.A. (2020). Modulation of the Root Microbiome by Plant Molecules: The Basis for Targeted Disease Suppression and Plant Growth Promotion 10.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., et al. (2019). Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle 176, 649-662.e20.

Peiffer, J.A., Spor, A., Koren, O., Jin, Z., Tringe, S.G., Dangl, J.L., et al. (2013). Diversity and heritability of the maize rhizosphere microbiome under field conditions 110, 6548–6553.

Peng, X., Dorman, K.S. (2021). AmpliCI: a high-resolution model-based approach for denoising Illumina amplicon data 36, 5151–5158.

Pérez-Carrascal, O.M., Choi, R., Massot, M., Pees, B., Narayan, V., Shapira, M. (2022). Host Preference of Beneficial Commensals in a Microbially-Diverse Environment 12.

Perisin, M., Vetter, M., Gilbert, J.A., Bergelson, J. (2016). 16Stimator: statistical estimation of ribosomal gene copy numbers from draft genome assemblies 10, 1020–1024.

Pop, M. (2009). Genome assembly reborn: recent computational challenges 10, 354–366.

Powell, B.J., Purdy, K.J., Thompson, I.P., Bailey, M.J. (1993). Demonstration of tra+ plasmid activity in bacteria indigenous to the phyllosphere of sugar beet; gene transfer to a recombinant pseudomonad 12, 195–206.

Price, M.N., Dehal, P.S., Arkin, A.P. (2010). FastTree 2 - Approximately maximum-likelihood trees for large alignments 5.

Remigi, P., Capela, D., Clerissi, C., Tasse, L., Torchet, R., Bouchez, O., et al. (2014). Transient Hypermutagenesis Accelerates the Evolution of Legume Endosymbionts following Horizontal Gene Transfer Public Library of Science), 12.

Rubin, B.E., Diamond, S., Cress, B.F., Crits-Christoph, A., Lou, Y.C., Borges, A.L., et al. (2021). Species- and site-specific genome editing in complex bacterial communities 7, 34–47.

Savoia, D. (2012). Plant-derived antimicrobial compounds: Alternatives to antibiotics pp. 979–990.

Schada von Borzyskowski, L., Severi, F., Krüger, K., Hermann, L., Gilardet, A., Sippel, F., et al. (2019). Marine Proteobacteria metabolize glycolate via the β-hydroxyaspartate cycle Nature Research), 575, 500–504.

Schäfer, M., Vogel, C.M., Bortfeld-Miller, M., Mittelviefhaus, M., Vorholt, J.A. (2022). Mapping phyllosphere microbiota interactions in planta to establish genotype–phenotype relationships Nature Research), 7, 856–867.

Schlaeppi, K., Dombrowski, N., Oter, R.G., ver Loren Van Themaat, E., Schulze-Lefert, P. (2014). Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives 111, 585–592.

Schmidt, S.K., Lipson, D.A., Raab, T.K. (2000). Effects of Willows (Salix brachycarpa) on Populations of Salicylate-Mineralizing Microorganisms in Alpine Soils 26, 2049–2057.

Sexton, A.C., Cozijnsen, A.J., Keniry, A., Jewell, E., Love, C.G., Batley, J., et al. (2006). Comparison of transcription of multiple genes at three developmental stages of the plant pathogen Sclerotinia sclerotiorum 258, 150–160.

Seymour, J.R., Amin, S.A., Raina, J.B., Stocker, R. (2017). Zooming in on the phycosphere: The ecological interface for phytoplankton-bacteria relationships Nature Publishing Group),.

Shen, W., Le, S., Li, Y., Hu, F. (2016). SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation 11, e0163962.

Smillie, C.S., Smith, M.B., Friedman, J., Cordero, O.X., David, L.A., Alm, E.J. (2011). Ecology drives a global network of gene exchange connecting the human microbiome 480, 241–244.

Song, W., Wemheuer, B., Zhang, S., Steensen, K., Thomas, T. (2019). MetaCHIP: community-level horizontal gene transfer identification through the combination of best-match and phylogenetic approaches 7, 36.

Steen, A.D., Crits-Christoph, A., Carini, P., DeAngelis, K.M., Fierer, N., Lloyd, K.G., et al. (2019). High proportions of bacteria and archaea across most biomes remain uncultured 13, 3126–3130.

Sun, Z., Huang, S., Zhang, M., Zhu, Q., Haiminen, N., Carrieri, A.P., et al. (2021). Challenges in benchmarking metagenomic profilers 18, 618–626.

Thiergart, T., Durán, P., Ellis, T., Vannier, N., Garrido-Oter, R., Kemen, E., et al. (2019). Root microbiota assembly and adaptive differentiation among European Arabidopsis populations 4, 122–131.

Thomas, C.M., Taib, N., Gribaldo, S., Borrel, G. (2021). Comparative genomic analysis of Methanimicrococcus blatticola provides insights into host adaptation in archaea and the evolution of methanogenesis 1, 47.

Thomson, N.R., Clayton, D.J., Windhorst, D., Vernikos, G., Davidson, S., Churcher, C., et al. (2008). Comparative genome analysis of *Salmonella* Enteritidis PT4 and *Salmonella* Gallinarum 287/91 provides insights into evolutionary and host adaptation pathways 18, 1624–1637.

Tkacz, A., Hortala, M., Poole, P.S. (2018). Absolute quantitation of microbiota abundance in environmental samples 6, 110.

Trivedi, P., Leach, J.E., Tringe, S.G., Sa, T., Singh, B.K. (2020). Plant–microbiome interactions: from community assembly to plant health 18, 607–621.

Vogel, C.M., Potthoff, D.B., Schäfer, M., Barandun, N., Vorholt, J.A. (2021). Protective role of the Arabidopsis leaf microbiota against a bacterial pathogen 6, 1537–1548.

Vorholt, J.A., Vogel, C., Carlström, C.I., Müller, D.B. (2017). Establishing Causality: Opportunities of Synthetic Communities for Plant Microbiome Research 22, 142–155.

Vrancken, G., Gregory, A.C., Huys, G.R.B., Faust, K., Raes, J. (2019). Synthetic ecology of the human gut microbiota 17, 754–763.

Wang, Y., Wang, X., Sun, S., Jin, C., Su, J., Wei, J., et al. (2022). GWAS, MWAS and mGWAS provide insights into precision agriculture based on genotype-dependent microbial effects in foxtail millet 13, 5913.

Wardell, G.E., Hynes, M.F., Young, P.J., Harrison, E. (2022). Why are rhizobial symbiosis genes mobile? Royal Society Publishing),.

Wheatley, R.M., Ford, B.L., Li, L., N Aroney, S.T., Knights, H.E., Ledermann, R., et al. (2020). Lifestyle adaptations of Rhizobium from rhizosphere to symbiosis 117, 23823–23834.

Wink, M. (2016). Secondary Metabolites: Deterring Herbivores In eLS, Wiley), pp. 1–10.

Wippel, K., Tao, K., Niu, Y., Zgadzaj, R., Kiel, N., Guan, R., et al. (2021). Host preference and invasiveness of commensal bacteria in the Lotus and Arabidopsis root microbiota Nature Research), 6, 1150–1162.

Xu, J., Zhang, Y., Zhang, P., Trivedi, P., Riera, N., Wang, Y., et al. (2018). The structure and function of the global citrus rhizosphere microbiome Nature Publishing Group), 9.

Yeoh, Y.K., Dennis, P.G., Paungfoo-Lonhienne, C., Weber, L., Brackin, R., Ragan, M.A., et al. (2017). Evolutionary conservation of a core root microbiome across plant phyla along a tropical soil chronosequence 8, 215.

Yoshida, K., Schuenemann, V.J., Cano, L.M., Pais, M., Mishra, B., Sharma, R., et al. (2013). The rise and fall of the Phytophthora infestans lineage that triggered the Irish potato famine 2.

Zhang, J., Liu, Y.-X., Guo, X., Qin, Y., Garrido-Oter, R., Schulze-Lefert, P., et al. (2021). High-throughput cultivation and identification of bacteria from the plant root microbiota 16, 988–1012.

Zhang, J., Liu, Y.-X., Zhang, N., Hu, B., Jin, T., Xu, H., et al. (2019). NRT1.1B is associated with root microbiota composition and nitrogen use in field-grown rice 37, 676–684.

Zhang, J., Zhang, N., Liu, Y.-X., Zhang, X., Hu, B., Qin, Y., et al. (2018). Root microbiota shift in rice correlates with resident time in the field and developmental stage 61, 613–621.

Zhang, W., Li, J., Tang, Y., Chen, K., Shi, X., Ohnishi, K., et al. (2017). Involvement of NpdA, a putative 2-nitropropane dioxygenase, in the T3SS expression and full virulence in Ralstonia solanacearum OE1-1 Frontiers Media S.A.), 8.

Zhou, L., Song, C., Li, Z., Kuipers, O.P. (2021). Antimicrobial activity screening of rhizosphere soil bacteria from tomato and genome-based analysis of their antimicrobial biosynthetic potential BioMed Central Ltd), 22.

Curriculum Vitae

Pengfan Zhang

Email: pzhang@mpipz.mpg.de | Phone: +49 15776854513;
Address: Max Planck Institute for Plant Breeding Research,
Carl-von-Linne-Weg 10, 50829, Cologne, Germany
Date of Birth: November/09/1994 Gender: Male

PROFILE

I'm a final-year bioinformatic PhD student affiliated to the department of plant-microbe interactions directed by Prof. Dr. Paul Schulze-Lefert at the Max Planck Institute for Plant Breeding Research. I obtained my master degree in Genomics at the University of Chinese Academy of Sciences and BGI-Shenzhen. I started to learn microbiome data analysis since 11/2015 at BGI-Shenzhen and have equipped myself with almost 7-year experience in this area so far. I am always keeping myself updated with the state-of-the-art studies and open-minded to all kinds of novel things and ideas. I also work out often and am into making new friends.

AREAS OF INTEREST

- The mechanisms underlying the assembly of microbial communities, including metabolic modelling
- Microbial ecology and host-microbiota co-evolution
- Dark matter mining from metagenomic data, e.g. novel species and biosynthetic gene cluster investigation

EDUCATION

Doctor of Philosophy, Ruben Garrido-Oter's lab, June/2019-Present,
Botany, Max Planck Institute for Plant Breeding Research, Cologne, Germany

Master of Science, September/2016 - May/2019
Metagenomics and Genome Mining, University of Chinese Academy of Sciences, Beijing, China
BGI-research, Shenzhen, China

Visiting Scholar, June/2018 - September/2018
Citrus Research and Education Center, University of Florida, Florida, USA

Bachelor of Science, September/2012 - May/2016
Biological Science, Zhejiang Normal University, Jinhua, Zhejiang, China

SKILLS

- **Experimental Skills**: Bacterial culturing and isolation, PCR, vector construction, plasmid transformation and western blot.
- **Computational Programming**: Proficient in Shell, Perl, and R; basic knowledge of Python, C++, Matlab.

• **Bioinformatics**: Proficient with conventional bioinformatic tools and metagenomic/metatranscriptomic/amplicon sequencing analysis pipelines, and keep up with the latest software.

LANGUAGES

English, fluent
Chinese, native

EXPERIENCES

PhD, MPIPZ, June/2019 to present

Cologne, Germany

• Developed an R package called ***Rbec***, which can be used to accurately profile the microbial compositions in synthetic microbiota by conducting reference-guided error correction in amplicon sequencing data and identifying variable paralogues within the same strain. The package can also detect the contaminated synthetic microbial communities and output sequences from contaminants. This package is archived on both Github (https://github.com/PengfanZhang/Rbec) and Bioconductor. The relevant work has been accepted for publication in ***ISME communications***.

• Established a computational pipeline for designing universal primers for amplicon sequencing of single-copy marker genes from a specific bacterial phylogenetic clade of interest, e.g. *Rhizobiales* and *Pseudomonas*. Those amplicons provide 3-4X higher resolution than 16S sequences and circumvent the bias in profiling bacterial communities caused by copy number variation.

• Analyzing the horizontal gene transfer events in plant-associated microbiome obtained from different host plants and different countries to shed light on bacterial adaptation to plants and community assemblage.

• Integrating the plant-associated bacterial genomes by both retrieving the publicly available genomes and metagenomic binning of novel genomes from plant-associated metagenome samples. 10335 non-redundant prokaryotic genomes were found and they can be grouped into 4255 species clusters.

Master, BGI-research, March/2017 to May/2019

Shenzhen, Guangdong, China

• Analyzed the taxonomic and functional structure of Foxtail Millet root microbiome by amplicon and metagenomic sequencing. During this period, I proposed and verified a new and ultra-fast strategy to construct the non-redundant gene catalogue for large-scale metagenomic samples and evaluated the required metagenomic sequencing depth for rhizosphere samples. I also raised a potential mechanism for the co-existence of rhizoplane microbiome. One of the relevant results was published in ***Gigascience***.

• Analyzed the functional compositions of global citrus root microbiome across six continents. In this project, I constructed the first and largest reference gene catalogue for root microbiome. This work is accepted by ***Nature Communications (Co-first author)***.

• Used random forest model to predict the shifted root-associated bacteria during different development stages of rice. This work was published in ***Science China Life Sciences*** (**Cover paper**).

• Large-scale genomic analysis of Trp-dependent IAA synthesis pathways in bacteria. This work unveils the distribution of IAA synthesis pathways across different bacterial phylum and was published in ***Molecules***.

Visiting Scholar, University of Florida, June/2018 to September/2018
Florida, USA
Worked as a visiting scholar in the Citrus Research and Education Center in USA.

Master, University of Chinese Academy of Sciences, September/2016 to January/2017
Beijing, China
I organized and founded a group called Academic Sharing Seminar and invited lots of classmates to give a speech about his/her research areas or some basic software usages every two weeks. These seminars attracted lots of attendees from other faculties and were highly praised and spread.

Undergraduate, Zhejiang Normal University, September/2012 to May/2016
Jinhua, Zhejiang, China
Conducted site-directed mutagenesis of *HIF-2alpha* from blind mole rats, which exerts as a transcription factor and could activate the hypoxic adaptation, and tested whether this mutation affected the phosphorylation level of this protein.

AWARDS

- Outstanding graduate of Beijing. 2019
- Outstanding graduate in the university of Chinese Academy of Sciences. 2019
- The national scholarship for masters. 2017-2018
- The first prize of BGI Education Center's scholarship. 2016-2017
- The second prize of undergraduate scholarship. 2015
- The second prize of the 21st biology contest for undergraduates in Zhejiang Province. 2014

PUBLICATIONS

1. Jin, T., Wang, Y., Huang, Y., Xu, J., ***Zhang, P.***, & Wang, N., et al. (2017). Taxonomic structure and functional association of foxtail millet root microbiome. ***Gigascience***, 6(10), 1-12.
2. Zhang, J., Zhang, N., Liu, Y. X., Zhang, X., Hu, B., & Qin, Y., et al. (2018). Root microbiota shift in rice correlates with resident time in the field and developmental stage. ***Science China Life Sciences***, 61(6), 1-9.
3. Xu, J., Zhang, Y., ***Zhang, P.***, Trivedi, P., Riera, N., Wang, Y., ... & Wang, N. (2018). The structure and function of the global citrus rhizosphere microbiome. ***Nature communications***, 9(1), 1-10. (Co-first author)
4. ***Zhang, P.,*** Jin, T., Kumar Sahu, S., Xu, J., Shi, Q., Liu, H., & Wang, Y. (2019). The distribution of tryptophan-dependent indole-3-acetic acid synthesis pathways in bacteria unraveled by large-scale genomic analysis. ***Molecules***, 24(7), 1411.
5. Gan, S. H., Yang, F., Sahu, S. K., Luo, R. Y., Liao, S. L., Wang, H. Y., ... & Liu, H. (2019). Deciphering the composition and functional profile of the microbial communities in Chinese Moutai liquor starters. ***Frontiers in microbiology***, 10, 1540.
6. Liao, S., Wang, Y., Liu, H., Fan, G., Sahu, S. K., Jin, T., ... & Liu, X. (2020). Deciphering the microbial taxonomy and functionality of two diverse mangrove ecosystems and their potential abilities to produce bioactive compounds. ***Msystems***, 5(5), e00851-19.

7. Schütz, V., Frindte, K., Cui, J., ***Zhang, P.***, Hacquard, S., Schulze-Lefert, P., ... & Dörmann, P. (2021). Differential impact of plant secondary metabolites on the soil microbiota. ***Frontiers in microbiology***, 12, 1267.
8. Wippel, K., Tao, K., Niu, Y., Zgadzaj, R., Kiel, N., Guan, R., Dahms E., ***Zhang P.***, ... & Garrido-Oter, R. (2021). Host preference and invasiveness of commensal bacteria in the Lotus and Arabidopsis root microbiota. ***Nature Microbiology***, 1-13.
9. ***Zhang, P.***, Spaepen, S., Bai, Y., Hacquard, S., & Garrido-Oter, R. (2021). Rbec: a tool for analysis of amplicon sequencing data from synthetic microbial communities. ***ISME Communications***.
10. Duran, P., Flores-Uribe, J., Wippel, K., ***Zhang, P.***, Guan, R., & Garrido-Oter, R. (2021). Characterization of the Chlamydomonas reinhardtii phycosphere reveals conserved features of the plant microbiota. ***Nature communications***.
11. Getzke, F., Amine, H. M., Crüsemann, M., Malisic, M., ***Zhang, P***, Ishigaki, Y., … & Schulze-Lefert, P. Co-functioning of bacterial exometabolites drives root microbiota establishment. Under review in ***Cell host & microbes***.
12. ***Zhang, P***, Garrido-Oter, R. Horizontal gene transfer in plant-associated microbiota shed light on bacterial host adaptation and microbial community assemblage. ***In preparation***.

Journal version of published papers