Use of Whole Genome Shotgun Sequencing for the Analysis of Microbial Communities in *Arabidopsis thaliana* Leaves

Dissertation
der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard Karls Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Julián Regalado Pérez
aus Ciudad de México, México

Tübingen
2019

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	04.07.2019
Dekan:	Prof. Dr. Wolfgang Rosenstiel
1. Berichterstatter:	Prof. Dr. Detlef Weigel
2. Berichterstatter:	Prof. Dr. Daniel Huson

Acknowledgments

To Detlef Weigel, who I sometimes refer as: The BAUS, The D, or El Jefe (but never in front of him), thank you for your constant support, valuable feedback, joyful enthusiasm, and who's love for science is absolutely inspiring. It has been an opportunity, only few people enjoy, to have you as a mentor. I just have to confess, sometimes I don't understand your jokes :).

Just as parenting is easier when it's done in a team. I would like to thank Daniel Huson who, like the teenager I can sometimes be, I would go to when I couldn't get what I wanted from The BAUS. It is incredible to have one great supervisor, it is beyond my greatest expectations to have two of them. Your support along the way will forever be treasured.

There is a crucial third person who I can't emphasize enough how important he has been in my PhD. Derek Lundberg (aka the baus, not to be confused with The Baus), has been in the trenches with me since the very beginning. You produced all the data I used for my project together with Sonja Kersten. Your input in experimental design and general microbiology knowledge is an integral part of this work. You are an incredible scientist, colleague, friend, and I really look up to you.

Special thanks to Kay Nieselt for being part of my defense committee. Great thanks to Anders J. Hansen for supporting me in the last part of my PhD and in my next step as postdoc.

I believe science to be the epitome of collaborative work, and I am anything but part of one of the greatest teams of scientists. I am greatly thankful to all members of department six for their support and great feedback after presentations or simply during coffee breaks. All of you make going to work something to full of joy.

I am greatly thankful to each and every single teacher who have tirelessly paved the way I have taken in my education. Special thanks to Federico Sanchez who taught me the value of passion in what one studies. Paty Torres who always pushed me to go beyond the requirements and to value knowledge for its own sake. Anne Kane, not only were you my first science teacher, without you this text would be unreadable. Jim Valero, you transmitted the power of language and showed me the joy it can bring to the reader. Maru Serratos, you taught me how the arts complement our interaction with the world, and consequently how it influences scientific thinking.

To my mother Guadalupe, who knows love above everything else. Thank you for your support. I love you and I have missed you every single instant during all these years.

To my father Justino, my first source of admiration, infinite support, and never ending love. If I grow up to be just a fraction of what you are, it will be more than what I have ever dreamed of.

To my siblings, Lulu, Santiago, and Lucia, you are my best friends, but I really expect you will refer to me as Dr. Regalado from now on. You know …. I am the favourite.

A mis abuelos Fernando y Lupe, que me han compartido su sabiduría, su amistad y su amor. Los quiero mucho, y admiro más.

A mis abuelos Henoc y Lourdes, que me han apoyado siempre en todo lo que hago.

To my uncle Pantxo, who has always taught me what can not be explored through the scientific method.

To my uncle Fernando, we share the burden of being the favourite.

To all my family Ana, Alexandra, Alonso, La Pitoja, Toya, Lupita, Jose Carlos, Pablo, Diego, Paquirri, Fabienne, Thabea, Noeli, Peter y Anita who have tolerated me discussing science time and time again, who has seen be destroy appliances, get suspended from school, and been in all sorts of trouble for the sake of scientific exploration.

The universe might be big, scary and with no purpose, but it can also be the finest of all places when you are surrounded by the friends you love. I love all my friends and a great part of who

I am is because of all of you. I have Luis Medina who is my friend since before we were born and whom, together with Lalo Legarreta, enjoy the greatest game there is to play. Luis Miguel Cervantes, who even though has seen me dedicate my life to study, knows I am not that smart. Dario Urbina, Carlos Flores, and Rodrigo Muños, my travelling friends who I have spent way too much time. Jorge Saviñon who always swims faster than me, but has shared his video games since forever. Roberto Lopez who I don't see much but know he is always there for me. Mariana Werner (La Babosa) for still being my friend even though I can still be a bit of a Baboso.

The following persons are probably the OKAYEST group of people I have ever met. Clemens Weiß (aka Clementino, aka FFFFF), an incredible friend and colleague who tolerated me sitting next to him for what I am sure he thinks was enough time. On exceptionally rare occasions we went for beers though. Sergio Latorre (aka SirGay), you irradiate joy wherever you go, thank you for your never ending support. Moises Exposito (aka Moi, aka Mua, aka Moisex) I'm lucky to have you as a friend, sorry if I bully you sometimes. Cristina Barragan (aka Barraganja), even though you bully me a lot I know it is only tough love. Efthymia Symeonidi (aka Effie) thank you for being there for me no matter the circumstances. Best roommate ever!! Sebastian Petersen, you and Teresa have always been a big part of my PhD life also, you make Mondays awesome, Tuesdays …. not so much. Jorge Kageyama, you are the one who told me to apply to Detlef's lab, so any complains will be forwarded to you. Patricia Lang (aka Paty Larga), thank you for all the support and great times, specially when the rest of my friends were being useless!!. Max Collenberg (aka Cafe Max) You are a great friend and we always had the greatest breakfasts. Talia Karasov, Michael Werner, and Laika (aka Like a Dog) who not only are scientist I greatly admire (Talia and Mike), but also incredible friends (all three of them). Hernan Burbano, you are among the few who truly understands my love for Barcelona, throughout these years we've been full of joy and drowned in sorrow. It is what it is. Kelly Swarts and Laura Morales for adopting me during my homeless period.

I would also like to thank Hugo Samano (aka Drugo) and Juan Escalona (aka Juanal) collectively known as "Mis Champs", who have never believed in me and attribute any success I have had to my "güero" phenotype, but still remain friends for some reason beyond my comprehension. You will forever be my friends.

Finally, I would like to thank Tatiana Feuerborn for always having my back. You are the most amazing person in my life and only with your support could I have achieved everything that has happened to me. You have tolerated my bad mood, constant traveling, and you have a titanic patience with me. I love you so much. I would also like to thank all of the Feuerborn Richtman family, specially Jacquie, for their continued support throughout these years.

To Xolo, the best BOI.

After years of hard work it is difficult to to summarize in an acknowledgments section all those who have helped me, not only during the years of my PhD studies, but throughout my entire life. I consider myself an extremely privileged person, thus my merits are in part a reflection of the actions done by multitudes. I am, for the most part, the lucky one through which such efforts are expressed. It is difficult to put my gratitude down in writing, yet I am incredibly thankful to each and every single individual who has shaped me along the way.

Zusammenfassung

Mikroorganismen, wie alle Bakterien, Archäen und einige Eukaryoten, bewohnen alle erdenklichen Lebensräume auf dem Planeten, von den Wasserschlitzen im tiefen Ozean bis hin zu extremen Umgebungen mit hoher Temperatur und Salzgehalt. Mikroben stellen auch die vielfältigste Gruppe von Organismen dar, wenn es um genetische Information, Stoffwechselfunktion und Taxonomie geht. Darüber hinaus bauen viele dieser Mikroben komplexe Wechselwirkungen untereinander und mit vielen anderen multizellulären Organismen auf. Die Sammlung von Mikroben, die sich einen Körperraum mit einer Pflanze oder einem Tier teilen, wird Mikrobiota genannt, und ihre genetische Information wird Mikrobiom genannt.

Die Mikrobiota hat sich als entscheidende Determinante für die allgemeine Gesundheit eines Wirtes herausgestellt und das Verständnis dafür ist in vielen biologischen Bereichen entscheidend geworden. Bei Säugetieren wurde die Darm-Mikrobiota mit wichtigen Krankheiten wie Diabetes, entzündlichen Darmerkrankungen und Demenz in Verbindung gebracht. In Pflanzen kann die Mikrobiota Schutz vor bestimmten Krankheitserregern bieten oder Resistenz gegen raue Umgebungsbedingungen wie Trockenheit vermitteln. Darüber hinaus stellen die Blätter von Pflanzen eine der größten Oberflächen dar, die möglicherweise von Mikroben besiedelt werden können.

Das Aufkommen der Sequenzierungstechnologien hat es den Forschern ermöglicht, mikrobielle Gemeinschaften in bisher unerreichter Auflösung und Größe zu untersuchen. Durch die Ausrichtung auf einzelne Loci wie den 16S rDNA-Locus in Bakterien können viele Arten gleichzeitig untersucht werden, ebenso wie ihre Eigenschaften wie die relative Häufigkeit, ohne dass eine individuelle Isolierung der Zieltaxa erforderlich ist. Sinkende Kosten der DNA-Sequenzierung haben auch zu einer vollständigen Schrotflinten-Sequenzierung geführt,

bei der anstelle der Ausrichtung auf einen einzelnen oder eine Reihe von Loci zufällige DNA-Fragmente sequenziert werden. Dadurch wird das gesamte Mikrobiom effektiv für die Untersuchung zugänglich, die so genannte Metagenomik. Infolgedessen stehen viel mehr Untersuchungsgebiete zur Verfügung, wie z.B. die Erforschung der genetischen Vielfalt innerhalb des Wirtes, die Funktionsanalyse oder die Zusammenstellung einzelner Genome aus Metagenomen.

In dieser Studie habe ich die Analyse von metagenomischen Sequenzierungsdaten aus mikrobiellen Gemeinschaften in Blättern von wilden Arabidopsis thaliana Individuen aus dem Südwesten Deutschlands beschrieben. Als Modellorganismen ist A. thaliana nicht nur in der Wildnis zugänglich, sondern verfügt auch über einen reichen Bestand an früheren Forschungen zur Wechselwirkung zwischen Pflanze und Mikrobe. Im ersten Abschnitt beschreibe ich, wie die gesamte Schrotflinten-Sequenzierung von Blatt DNA-Extrakten verwendet werden kann, um die taxonomische Zusammensetzung der mikrobiellen Gemeinschaft einzelner Wirte genau zu beschreiben. Die Art der Sequenzierung von ganzen Schrotflinten wird verwendet, um echte mikrobielle Häufigkeiten zu schätzen, die mit der Sequenzierung von Amplikonen nicht erreicht werden können. Ich zeige, wie unterschiedlich diese Gemeinschaft zwischen den Wirten ist, aber es werden einige Trends beobachtet, wie z.B. die Dominanz der Bakteriengattungen Pseudomonas und Sphingomonas. Außerdem, obwohl es Unterschiede zwischen den Individuen gibt, untersuche ich den Einfluss von Ursprungsort und Wirtsgenotyp. Schließlich wird die metagenomische Zusammensetzung auf einzelne Proben angewendet, was die Grenzen von WGS in Pflanzenblättern zeigt.

Im zweiten Abschnitt untersuche ich die genomische Vielfalt der am häufigsten vorkommenden Gattungen: Pseudomonas und Sphingomonas. Ich verwende einen Kerngenomansatz, bei dem ein Satz gängiger Gene aus zuvor sequenzierten und montierten

Genomen gewonnen wird. Danach werden die Gensequenzen des Kerngenoms als Referenz für eine kurze Genomkartierung verwendet. Basierend auf diesen Mappings werden aus der Häufigkeitsverteilung von Nicht-Referenzbasen bei jedem detektierten Single Nucleotide Polymorphism (SNP) individuelle Stammgemische abgeleitet. Schließlich werden SNP's verwendet, um die Populationsstruktur von Stammmischungen über Proben und mit bekannten Referenzgenomen hinweg abzuleiten.

Zusammenfassend lässt sich sagen, dass diese Arbeit Erkenntnisse über die Verwendung von metagenomischer Sequenzierung zur Untersuchung mikrobieller Populationen in Wildpflanzen liefert. Ich identifiziere die Stärken und Schwächen der Verwendung der gesamten Genomsequenzierung für diesen Zweck. Sowie eine Möglichkeit, die Dynamik des Stammniveaus von vorherrschenden Taxa innerhalb eines einzigen Wirtes zu untersuchen.

Abstract

Microorganisms, such as all Bacteria, Archaeae, and some Eukaryotes, inhabit all imaginable habitats in the planet, from water vents in the deep ocean to extreme environments of high temperature and salinity. Microbes also constitute the most diverse group of organisms in terms if genetic information, metabolic function, and taxonomy. Furthermore, many of these microbes establish complex interactions with each others and with many other multicellular organisms. The collection of microbes that share a body space with a plant or animal is called the microbiota, and their genetic information is called the microbiome.

The microbiota has emerged as a crucial determinant of a host's overall health and understanding it has become crucial in many biological fields. In mammals, the gut microbiota has been linked to important diseases such as diabetes, inflammatory bowel disease, and dementia. In plants, the microbiota can provide protection against certain pathogens or confer resistance against harsh environmental conditions such as drought. Furthermore, the leaves of plants represent one of the largest surface areas that can potentially be colonized by microbes.

The advent of sequencing technologies has let researchers to study microbial communities at unprecedented resolution and scale. By targeting individual loci such as the 16S rDNA locus in bacteria, many species can be studied simultaneously, as well as their properties such as relative abundance without the need of individual isolation of target taxa. Decreasing costs of DNA sequencing has also led to whole shotgun sequencing where instead of targeting a single or a number of loci, random fragments of DNA are sequenced. This effectively renders the entire microbiome accessible to study, referred to as metagenomics. Consequently many more areas of investigation are open, such as the exploration of within host genetic diversity, functional analysis, or assembly of individual genomes from metagenomes.

In this study, I described the analysis of metagenomic sequencing data from microbial

communities in leaves of wild *Arabidopsis thaliana* individuals from southwest Germany. As a model organisms, *A. thaliana* not only is accessible in the wild but also has a rich body of previous research in plant-microbe interactions. In the first section, I describe how whole shotgun sequencing of leaf DNA extracts can be used to accurately describe the taxonomic composition of the microbial community of individual hosts. The nature of whole shotgun sequencing is used to estimate true microbial abundances which can not be done with amplicons sequencing. I show how this community varies across hosts, but some trends are seen, such as the dominance of the bacterial genera *Pseudomonas* and *Sphingomonas*. Moreover, even though there is variation between individuals, I explore the influence of site of origin and host genotype. Finally, metagenomic assembly is applied to individual samples, showing the limitations of WGS in plant leaves.

In the second section, I explore the genomic diversity of the most abundant genera: *Pseudomonas* and *Sphingomonas*. I use a core genome approach where a set of common genes is obtained from previously sequenced and assembled genomes. Thereafter, the gene sequences of the core genome is used as a reference for short genome mapping. Based on these mappings, individual strain mixtures are inferred based on the frequency distribution of non reference bases at each detected single nucleotide polymorphism (SNP). Finally, SNP's are then used to derive population structure of strain mixtures across samples and with known reference genomes.

In conclusion, this thesis provides insights into the use of metagenomic sequencing to study microbial populations in wild plants. I identify the strengths and weaknesses of using whole genome sequencing for this purpose. As well as a way to study strain level dynamics of prevalent taxa within a single host.

Table of Contents

Chapter 1: Introduction

It is a surprise to no one that what matters the most are the little things in life. A good espresso in the morning, the cheerful chirp of birds, the soothing sensation that petrichor brings after a stormy night. It is in the minute where one can find that which develops into the significant. It is the stories of the little ones overcoming great odds that allure us the most.

Of all those many stories, I would like to tell you a story involving a man named William Wood. William Wood was English mintmaster who stroke a dubious contract to produce 360 tons of halfence for Ireland in 1722. Naturally the Irish were not happy, they wanted to have their own monetary policy, but the English empire was one big beast to defeat. The Irish, not letting themselves be bullied, quickly mobilized. Among a boycott and many other actions to challenge the Wood contract, a series of pamphlets was published in 1724 to increase public awareness about the dangers of a privately minted coin. The author was M. B. Drapier and his pamphlets were called *Drapier's Letters*. The campaign proved successful and the already minted coins were recalled, only to be used with other little ones, the British American colonies.

The reader might think this to be just a small story where the little ones win, but it's more than that. M. B. Drapier turns out to be a pseudonym used by the real author, Jonathan Swift. Drapier for draper a common person, someone everybody could relate to, some that could unite. A draper who, from small thin fibers, creates comfortable clothes. A cloth maker who dresses people for a cause.

The next story also started with a draper. Curiously, this draper died only one year before the publication of *Drapier's Letters*. His name was Antonie Philips van Leeuwenhoek, and with his lenses, was the first one to show how it's the little things that can influence us the most.

The Ubiquity and Utility of Microbial Life

Microorganisms, referring to bacteria, archaea, and some Eukaryotes, occupy every ecological niche (Fontaneto, 2011), and their ubiquity makes them an extremely important subject of study in almost all areas of biology. Furthermore, inhabiting a countless number of environmental niches, microbes have been able to develop all sorts of metabolisms suitable for the variety of living conditions they encounter. Not only are microbes located in association with other microbes as well as multicellular hosts, but in every imaginable location on this planet including all sorts of environments, ranging from extreme conditions such as high temperature and pressure like the ones found in deep water vents (Dick et al., 2013), high salinity bodies of water (Antunes, Ngugi, & Stingl, 2011), severe radiation (Battista, Earl, & Park, 1999) or even the vacuum of outer space (Sancho et al., 2007), to every imaginable surface and even clouds (Joly et al., 2013)**.** The magnitude and scope of their presence and the impact they have on the planet has made microbiology one of the cornerstones of the biological sciences. Even before the discovery of microbes in the 17th century some premodern societies discussed their existence (Scarborough, 1970), and more formally since the establishment of the germ theory of disease. Bacteria and other microorganisms have also been consequential outside the direct study of their biology as many of the fundamental advances in our understanding of molecular biology were made with microorganisms, including the identification of DNA as genetic material (Griffith, 1928), the semiconservative replication of DNA (Meselson & Stahl, 1958), the discovery and deciphering of the genetic code (Jones & Nirenberg, 1966; Matthaei, Jones, Martin, & Nirenberg, 1962), and many others (Fry, 2016). More recently, the CRISPR-Cas system, one of the most transformative techniques in genetic engineering, was developed as a direct application of a fundamentally microbial process, the adaptatie immunity of bacteria to phages (Jinek et al., 2012). Additionally many modern techniques used in laboratories around the world rely in part on the use of some microbe; these include yeast two-hybrid (Fields & Song, 1989), agroinfiltration (Chilton et al., 1977), or heterologous protein expression (J. Lee & Ramirez, 1994), among others.

Microbes as Subject of Study

An important aspect of microbiology has been the attempt to understand the ecological processes governing the dynamics that control the interactions among microbial groups, their relationship with abiotic factors, and their evolution. For example, different phototrophic sulfur bacterial groups occupy distinct functional features in the chemocline of lakes (Čanković, Petrić, Marguš, & Ciglenečki, 2017). Pachiadaki and colleagues showed that nitrite-oxidizing bacteria constitute one of the most important bacterial groups responsible for carbon cycling in the mesopelagic ocean, a major part of the biosphere by volume where most of the exchange of this element is believed to happen (Pachiadaki et al., 2017). Another crucial biogeochemical process is the fixation of atmospheric nitrogen, where free living and in particular legume-associated bacteria are responsible for the vast majority of nitrogen assimilation (Stein & Klotz, 2016). The availability of nitrogen is a critical requirement for any ecosystem as different compounds based on this element are centerpiece in essential biological processes. Other major elementary cycles such as the cycles of sulfur and phosphorous, depend, at least in part, on a variety of microorganisms (Anantharaman et al., 2018). In recent years, research has revealed many more geochemical processes that involve microbes (Majumder & Wall, 2017), and it has become apparent that almost every process in the planet's crust involves some sort of biological process (Colman, Poudel, Stamps, Boyd, & Spear, 2017).

The distinct biological characteristics of microorganisms have made the study of the processes driving their evolution and relationship with the environment a fascination for scientists since those first observations by Leeuwenhoek (van Leeuwenhoek, 1800). Evidence of life on Earth as unicellular organisms dates back to approximately 3.5 billion years before present (Schopf, Kitajima, Spicuzza, Kudryavtsev, & Valley, 2018), these first life-forms were most likely of bacterial nature. Subsequently some of these early unicellular lineages repeatedly and independently went through the process of acquiring multicellularity (Knoll, 2011). This gave rise to the most common multicellular microbes: fungi, with other important groups of multicellular microbes including the slime molds and the oomycetes. The first multicellular animals did not arrive until much later, 600 million years before present (Brunet & King, 2017), which means life on Earth emerged just a billion years after the formation of the planet, but non-microbial entities took three times as much time to develop. Naturally, multicellularity can be viewed as a more complex system requiring more time to develop, but if planetary origin of life is assumed to be the most parsimonious hypothesis (Sutherland, 2016), this means the emergence of life from pure abiotic processes took much less time to originate than the time it took for more complex life-forms to evolve from single cells. This has given bacteria, archaea and a select group of Eukaryotes an immense evolutionary history, a surface that is just beginning to be scratched. As all major metabolic pathways are present in one form or another in microorganisms (Schuetz, Zamboni, Zampieri, Heinemann, & Sauer, 2012), laying the evolutionary basis for all other life forms, a deep understanding of these foundations is necessary for the understanding of life itself. The evolutionary history of microorganisms is an active field of research concerned with a number of aspects including the Tree Of Life (TOL) initiative (Parks et al., 2018), which tries to systematically organize microbial groups according to their genetic relatedness and not based on purely metabolic similarities. The delineation of bacterial species continues to be challenging (Maderankova, Jugas, Sedlar, Vitek, & Skutkova, 2019), as debate still exists whether microbes, more specifically bacterial species, can be defined as discrete entities, or if a continuum of genetic similarity best describes their phylogenetic properties. With the development of the endosymbiotic theory of Eukaryotic origin (Sagan, 1967), the evolution of basic cell structures has become tightly associated with bacterial evolution.

This has lead to a spillover in research, where certain microbial properties are further investigated for other applications. This ranges from production of specific metabolites to inhibitory dynamics of pathogens. In terms of applied research, several active fields deal with a variety of topics in microbial evolution such as antibiotic resistance of human and non human pathogens (Blair, Webber, Baylay, Ogbolu, & Piddock, 2015), or the evolutionary history and surge of several plant pathogens. With the increased understanding of microbial biology and more specifically with the development of recombinant DNA, industrial-scale processes have made continuous use of microorganisms and their distinct metabolic properties. A combination of metabolic engineering and artificial evolution has enabled the production of countless number of compounds produced in bacteria or yeast for purposes as varied as medically relevant substances such as insulin (Crea, Kraszewski, Hirose, & Itakura, 1978), naringenin (Raman, Rogers, Taylor, & Church, 2014), human growth hormone (Patra et al., 2000), and interferon (Wang et al., 2002), among others. Many commercially important amino acids are synthesised via fermentation techniques, vitamin synthesis has proven quite successful with modified strains able to yield up to 10,000 times more biotin than wild type counterparts in *E. coli* (Adrio & Demain, 2010) . More recently cannabinoid compounds have been synthesised in yeast (Luo et al., 2019). These are all examples of microbes used as a monoculture, but more complex systems of microorganisms are also being used in many areas of industry and research such as in energy production and waste management where microbial communities tailored for maximum efficiency in biodigesters are being develop to achieve cost effective biofuel manufacturing (Parisutham, Kim, & Lee, 2014).

Recent examples of important plant diseases caused by microbes are oleander leaf scorch (Temsah, Hanna, & Saad, 2015), wheat rust (Kiran et al., 2016) and bacterial canker in kiwifruit (Morán et al., 2018), which have been responsible for major outbreaks resulting in significant decrease in crop yield. As concerns about food security increase, more research is focusing on understanding these plant-microbe associations. On the other hand, improvements in agricultural performance are also tied to microbial communities of the plant and the environment around it, where the interactions between plants and the microbes in soil, water, and the microbiota are at the center of attention in the development of sustainable agronomy (Tikhonovich & Provorov, 2011). These are just some selected examples where the integral understanding of microbial biology has fueled developments in many other areas of science and industry.

Host Associated Microbes

Even though free living microorganisms constitute a fundamental aspect in biology, the vast majority of microbial phenomena happen in association with other organisms. Bacterial mats constitute an entirely self contained system of intrinsically cooperating members often of different species and sometimes even different domains (Visscher & Stolz, 2005). These associations can be so strong that they can appear to behave as a single organisms such as in the case of lichens. Here, the symbiotic interaction between an alga or cyanobacterium and a fungus, has properties not found in its constituent parts (Koch et al., 2019). Corals provide another example of an entity whose association with a microbe, in this case dinoflagellates, is the basis for an entire ecosystem. Close knit interactions may also happen in animals; for example, some marine slugs in the *Elysia* genus undergo the process of kleptoplasty, where chloroplasts of the red/green algae they consume are not digested and instead held in their digestive tract in order to take advantage of its photosynthetic products (Christa et al., 2015). At least one threeway association has been studied in which the symbiotic association between the panicgrass, *Dichanthelium lanuginosum*, and a fungus, *Curvularia protuberata*, only results in heat resistance for both organisms if it happens in conjunction with an infection with a mycovirus (Márquez, Redman, Rodriguez, & Roossinck, 2007). These are just a few examples of systems involving an association with a microorganism, but generally these tight bonds are usually static and serve a well defined biological process. Nevertheless, as a consequence of the extreme versatility and capacity to colonize every available niche, microbes have been able to establish multispecies communities with higher organisms, namely animals and plants, the nature of which has slowly surfaced to be of fundamental importance to the host.

The examples discussed so far mostly involve a single microbe interacting with its host. In reality, organisms interact with a much more diverse mixture of organisms, usually from all three domains of life. In addition, these associations are often dynamic in space and time. This collection of microorganisms, which share a body space with a host, is referred as the microbiota and the collection of their genomes as the microbiome, which together with the host constitute the holobiome (Mitter, Pfaffenbichler, & Sessitsch, 2016). Importantly, the individual relationships of these microbes and their host can have different properties. Broadly, they can be classified as symbiotic, commensal, or pathogenic. In a symbiotic interaction, both partners benefit from the relation, such as with the microbes of the mammalian gut. Commensalism happens when one member of the interaction benefits at no cost or little harm to the other, typically the host. Nasal colonization by *Staphylococcus aureus* in humans is a prime example of a commensal relation. Finally, pathogenic interactions occur when an organism benefits at the expense of and great harm to its host.

As previously mentioned, these communities are also dynamic in terms of their composition and the roles they play in their host. For example, the human gut microbiota is different from the human skin microbiota (Byrd, Belkaid, & Segre, 2018); furthermore, the skin of different body parts has been shown to vary in microbial composition (Lloyd-Price et al., 2017), and body sites in different individuals may also vary in taxonomic composition. The microbiome of an individual, regardless of body part of origin, may also vary in time (Gilbert et al., 2018), or based on a myriad of other factors such as diet, health status, ethnicity, or geographic location (Gupta, Paul, & Dutta, 2017). A large body of research has revealed that the microbiota of an organisms has significant effects on the well-being of the host (Berendsen, Pieterse, & Bakker, 2012), and that it can be of central importance in many diseases. For example, in the case of humans, research has demonstrated the impact of the microbiome on inflammatory bowel disease, diabetes, skin conditions (Cho & Blaser, 2012), and more recently its involvement in several mental health illnesses such as depression and Alzheimer's disease (Kaplan, Rucklidge, Romijn, & McLeod, 2015). All these are examples of differential composition of microbial communities relative to their environment and the correlated effects this might have in the host.

The individual organisms in the community can directly interact with each other and the host. For example, several species of Gram-negative bacteria, directly inject proteins into host cells. These proteins are known as effectors and have a wide range of functions, usually to suppress the host immune system or facilitate the colonization by the microbe overall (Stavrinides, McCann, & Guttman, 2007). Another example of direct interaction comes from mycorrhizal fungi and some plant associated oomycetes, which develop structures called haustoria, which form in the intracellular space and eventually make their way through the plant cell wall to exchange or extract via the plant cell membrane water, metabolites and proteins, in either a pathogenic or symbiotic fashion. In addition to these one-to-one interactions, there are community-wide interactions between the members of a microbiome. Usually these are mediated by the production of secondary metabolites that induce specific effects in other microbes in the community or in the host. For example, in the mammalian gut, the synthesis of small molecules such as short chain fatty acids, niacin, polysaccharides, among others, help modulate the hosts to mount adaptive immune response (Kau, Ahern, Griffin, Goodman, & Gordon, 2011).

Even though substantial research has concentrated on humans and other vertebrates, substantial efforts in the field of plant-microbe interactions have revealed the great importance of the root and leaf associated microbial community for several aspects of plant health such as the priming of the plant immune system (Selosse, Bessis, & Pozo, 2014), protection against pathogens (Lugtenberg & Kamilova, 2009), yield stability in crops (Chaparro, Sheflin, Manter, & Vivanco, 2012), modifying tolerance to abiotic stress (Rolli et al., 2015), promoting plant growth, and overall plant fitness. All these properties have sparked the development of new processes and applications that aim to take advantage of plant-microbe interactions in order to achieve biocontrol of pests (Ciancio, Pieterse, & Mercado-Blanco, 2016) or selectively modify certain aspects of plant development (Panke-Buisse, Lee, & Kao-Kniffin, 2017). With the increasing demand of sustainable agriculture, exploiting the plant microbiota will be of critical importance.

Methodologies for Studying Microbial Communities: From the Bench to the Computer

Biochemical Classification

When studying microbial communities, host associated or not, key properties are often prioritized. These include, but are not limited to, taxonomic composition, functional composition, variation within an individual, variation across individuals, and influence of host genetics. Since the beginnings of microbiology, the primary way to elucidate which organisms were present in a sample was until very recently the isolation of individual colonies followed by phenotypic characterization of observable properties such as color, smell, shape, and eventually biochemical properties (Zillig, 1991). Depending on the similarity or difference in the observed phenotypes, isolates could be placed in groups of shared traits, the most common example being Gram staining (Gram & C, 1884), where bacteria were classified as Gram positive, stained by the dye Crystal Violet, to detect the presence of peptidoglycan in the cell wall, and Gram negative, which are cells without peptidoglycan that remained unstained. Isolates could also be grouped based on their origin, for example, the family name “Enterobacteriaceae” is derived from “enteric”, pertaining to the intestines, and as such many taxa in this family were isolated from the guts of mammals. With the advent of more elaborate biochemical assays it became possible to divide microorganisms based on the effects of their metabolisms. For example, microorganisms could be classified based on their nutrient source, such as photosynthetic, chemotrophic, or lithotrophic, among others (Druschel & Kappler, 2015), or based on their ability to metabolise certain compounds, as used in the beta-Glucuronidase test (Rice, Allen,

& Edberg, 1990), urea test (Graham et al., 1987), or the oxidase test for detecting cytochrome C activity (Tarrand & Gröschel, 1982). A multitude of these techniques exist usually for the identification of only a small group microorganisms, usually bacteria. Naturally it was impossible to address the greater community in multiple individuals and in a quantitative way.

DNA as an Identifying Agent

The advent of DNA sequencing brought a revolution to all fields of biology. In the specific case of microbiology it was now possible to start tracking the phylogenetic relationship of microbial isolates by comparing specific common loci. This saw its first successful applications when used with ribosomal genes, specifically with the 16S rDNA locus in bacteria (Woese et al., 1975). In this technique, the ribosomal genes are targeted for sequencing and due to its low mutation rate, the sequence can then be used to infer phylogenetic relationships of distant organisms. Due to its role as an essential gene, it is always present in the organism of interest. This resulted, among many things, in the division of all lifeforms into the three domains we know today: Bacteria, Archaea, and Eukaryota (Woese & Fox, 1977). Quickly, large scale bacterial groups could be defined from their genetic similarity instead of common biophysical properties. In addition, the relationship of these groups and which host organisms or niches they were colonizing started to be explored.

With the decreasing costs in DNA sequencing, it became possible to study an ever greater number of microbial strains. Critically, the process of colony isolation could be entirely skipped and applied much more easily to DNA extracts of microbial mixtures. This was achieved by exploiting the highly conserved regions of the 16S rDNA locus for selective amplification with PCR. Following ligation to a plasmid and transformation into *E. coli*, DNA containing the 16S rDNA locus could be sequenced and the variable regions of the gene used for phylogenetic comparison (Weisburg, Barns, Pelletier, & Lane, 1991). As the collection of 16S rDNA reference sequences started to grow, it became possible to make comparisons with previously sequenced organisms for classification purposes (Stackebrandt & Goebel, 1994). This quickly spilled over onto other fields of research, especially in clinical microbiology, where early detection of microbial species is paramount (Woo, Lau, Teng, Tse, & Yuen, 2008). Yet, techniques were still confined to analyse bacterial communities or, later on, fungal communities via the sequencing of the inter-transcribed spacer located between the genes for the small and large ribosomal subunits (Baldwin et al., 1995).

Microbiology in the Era of High-throughput Sequencing

High-throughput sequencing technologies brought yet another revolution in biological sciences as a much larger number of sequences could be obtained from DNA extracts (Kircher & Kelso, 2010) at the expense of a much shorter sequencing read length, initially <100 bp. Compared to earlier work based on Sanger sequencing, this substantially decreased the resolution at which any two microbial groups could be distinguished. Additionally the error rate in high-throughput sequencing, further decreased resolution especially in closely related taxa. Nevertheless, the large amounts of data that could be obtained by high-throughput sequencing meant that large quantities of strains could be measured. Microbial communities were able to be studied studied at much higher sequencing depth giving access to low abundance microbes. Researchers moved quickly to adapt this technology to the study of microbial communities, primarily by targeting the 16S and ITS rDNA loci. In addition to the ability of directly studying taxa for which no known culturing method existed. This brought great insight into the biological processes shaping all kinds of microbial communities in the gut (Yatsunenko et al., 2012), ocean and lake water (Sunagawa et al., 2015), soil (Roesch et al., 2007), extreme environment like acid lakes or sewage (Zhang, Shao, & Ye, 2012), and basically any medium from which DNA could be extracted. To date, amplicon sequencing remains one of the most important tools of any microbiology laboratory and will be used in years to come.

Despite its usefulness, amplicon sequencing has some pitfalls such as the exclusion of some organisms due to primer bias. Even though a conserved region is targeted for amplification, such a region is usually determined based on existing sequences. If a microbe in the target comunity happens to vary in such a region, the amplification step may not work for the particular microbe. In addition, as amplicon sequencing is a compositional technique, estimates of taxon abundances are only relative to each other and real abundances cannot be directly obtained. This property can further confound downstream analyses if a high abundance taxon is not detected on its own. Depending on the targeted locus there will be a limit to the resolution at which any two taxa will be able to be distinguished, for example in the case of 16S rDNA locus, species of the same genus are usually not possible to recognize separately (Kisand & Wikner, 2003). Finally, functionally, not much can be inferred from just a single genomic location, giving restricted insights in this regard on par with biochemical techniques to elucidate functional properties of microbes.

Sequencing costs have decreased enormously over the years (Wetterstrand, 2018), which has made amplicon sequencing increasingly available, at reduced price and consequently at an increased scale. It is now common to apply this technique to hundreds of samples (Tourlousse, Ohashi, & Sekiguchi, 2018) and with the development of better automation the time needed for sample preparation is greatly reduced. A combination of all these factors, on the other hand, makes it possible to address the metagenome, the entire collection of genomes, in a sample via the sequencing of random DNA fragments. Instead of focusing on a single locus, DNA extracts are fragmented, and individual stretches of DNA sequence are obtained. This makes for unbiased identification of any taxon represented in the DNA obtained from a microbial mixture (i.e., there are remaining biases in the form of differences in DNA extractability). Now the entire genome of organisms can potentially produce sequencing reads and many new kinds of questions can be addressed. Because no targeted amplification of an individual locus is performed, it is less likely that individual taxa will be missed, greatly reducing the chance of excluding microbes from downstream analyses. Functional analysis is more accurate, as individual genes and their functions can be identified (Campanaro et al., 2016). With the use of appropriate controls such as spiked in DNA, real abundances can be estimated with much higher accuracy (Lu, Breitwieser, Thielen, & Salzberg, 2017). Insight into the evolutionary dynamics of specific groups in and across samples is possible by having genome wide access (Garud, Good, Hallatschek, & Pollard, 2019). This increase in resolution makes it easy to discover new taxa that otherwise might be missed due to similarities in a targeted amplicon. Finally, entire genomes can be reconstructed by assembling sequenced reads into longer contigs and scaffolds (Narasingarao et al., 2012).

Metagenomics as a Data Analysis Challenge

Unfortunately, certain circumstances may reduce the usefulness of whole genome shotgun sequencing of complex microbial mixtures. For example, significant depth (amount of sequenced fragments) must be obtained in order to have reliable estimates of any property of a microbiome (Zaheer et al., 2018); otherwise, microorganisms at low abundances will not be represented in the pool of sequenced reads and be completely missed. This can be particularly important in extremely diverse microbiomes, such as the ones found in soil, or where the amount of extracted DNA is limited. Even though there is no simple, universal answer as to how much sequence must be obtained to accurately study a sample, current research suggests 0.5 million short read sequences per sample obtains taxonomic information on par with 16S rDNA amplicon sequencing (Hillmann et al., 2018). Other caveats of metagenomics arise when studying host associated microbiomes where it is not possible to efficiently separate host material from the targeted microbial mixture; this greatly decreases sequencing depth of the microbiome by overrepresentation of the host genome. In the case of metagenomic assembly both depth and diversity play an important role, the efficiency of current algorithms is negatively affected when applied to very complex mixture where many species or strains are present (Scholz, Lo, & Chain, 2012). This is due to a combination of algorithmic and technical challenges.

On the algorithmic side, most metagenome assemblers begin by creating a de Bruijn graph whose nodes correspond to kmers (subsequences of length k) extracted from sequenced reads (Zerbino & Birney, 2008). The assembled contigs are generated by traversing this graph and concatenating the sequences of the kmers. When a partition in the path is encountered, meaning more than one path can be followed in the graph, kmer concatenation stops to start a new contig. As sequence diversity in a sample increases, the complexity of this graph increases proportionately. For example, given two sequences of the same length, the number of kmers differing between the two is a linear function of the number of bases that are different between the two sequences and how far apart each mismatch is from each (Compeau, Pevzner, & Tesler, 2011). A consequence of increasing the number of kmers is an increase in the number of paths by which a graph can be traversed. Shorter contigs are generated as a consequence. On the technical side, increasing diversity means less sequencing depth per taxon, so less information is available about the genome for assembly purposes.

Depth of sequencing plays a major role in the ability of any assembler to generate contigs from low abundance genomes. Nevertheless, successful reconstruction of multiple draft quality genomes from metagenomic data has been achieved in high depth samples. For example, close to 1,000 draft genomes could be assembled from cow rumen metagenomes revealing the high content of carbohydrate metabolism associated genes expected from such environment (Stewart et al., 2018). Metagenomically assembled genomes have also been obtained from high diversity environments such as ocean water (Tully, Graham, & Heidelberg, 2018). Finally, the ability to produce metagenomically assembled genomes has demonstrated to give key insights into previously unknown diversity in the tree of life (Parks et al., 2017). Whole genome shotgun sequencing of microbial communities also requires the existence of extensive well curated databases of reference organisms to which metagenomic sequences or assemblies can be compared in order to identify known species, genes, and genetic variation previously reported.

Currently there are many such options (Dunivin, Choi, Howe, & Shade, 2019; Meyer et al., 2008), but regardless of the source database, there will always be a risk of ignoring sequences without known references, as well as the fact that each of these databases may introduce biases because the way they were compiled.

Nevertheless, whole genome shotgun sequencing of the microbiome has made great strides and contributed to diverse areas of microbiology. Metagenomics of sewage water, a naturally rich microbial habitat, not only has revealed an incredible diversity of bacteria and archaea, but also has identified diverse viral populations that continually interact with their bacterial hosts (Klausa, Piešiniene, Staniulis, & Nivinskas, 2003); additionally, pathogenic strains of multiple bacterial species are continuously detected in this environment (García-Aljaro, Blanch, Campos, Jofre, & Lucena, 2019). Antibiotic resistance in water treatment plants is an important topic in this regard due to the extreme influence of human activity; research has revealed sewage as a hotspot for the development of resistomes (the collection of genes associated with antibiotic resistance) (Su et al., 2017). Ocean water is another source where metagenomic sequencing has been used intensely, showing, as expected, a great microbial diversity found all over the world (Bork et al., 2015). In an applied setting, Appolinario and colleagues observed an enrichment of hydrocarbon degrading bacteria in oil contaminated water, which was supported by the reconstruction of 12 genomes with clear signatures of genes involved in hydrocarbon metabolism (Appolinario et al., 2019). In another study, a taxonomic and genomic description of communities involved in nutrient cycling in the western subarctic ocean was obtained (Y. Li et al., 2018). Similarly, strain level diversity of sulfur oxidising symbionts has been amply described with metagenomic and metatranscriptomic sequencing of gill pieces of mussels of the genus *Bathymodiolus* in deep sea water vents (Ansorge et al., 2019). In studies of the gut, metagenomic sequencing has been extensively used to study the prevalence and diversity of sequences associated with the crAssphage, the most abundant virus of the human gut (Yutin et al., 2018). In another gut study, taxonomic profiling of human, dog, mice, and pigs revealed remarkable similarity between canine and human gut profiles relative to the other two taxa, highlighting the impact of diet in shaping the gut microbiome (Coelho et al., 2018). In the field of metagenomic assembly, a single study was able to recover almost 9,000 microbial genomes, most of them of high quality in terms of completeness, from different metagenomic sequencing projects of mainly environmental settings and non-human guts (Parks et al., 2017)).

The development of analysis methods and algorithms, capable of efficiently processing the enormous volumes of data typical of high throughput experiments, has unfolded on par with the growth of sequencing capacity. Particularly at this moment in time where it has become almost trivial for the average researcher to generate tremendous amounts of data. BLAST has always been the goto tool when comparing sequences (Altschul, Gish, Miller, Myers, & Lipman, 1990). Even though it has been used since the early days of random fragment sequencing of microbial mixtures, it quickly became unsuitable for finding alignments to reference databases in terms of time and resources needed for such computations. This forced researchers to innovate new data processing algorithms; one of the first methods was BLAT (Kent, 2002), which is up to 100 times faster than BLAST in identifying protein sequences of high similarity but fails to find a significant fraction of more distant alignments. MetaPhlAn classifier (Truong et al., 2015), where a set of clade specific genes is preselected as reference markers to which metagenomic reads could be aligned, brought a great decrease in analysis time with very little loss in accuracy. Its main disadvantage is the use of preselected markes, which reduces significantly the number of targets a read could map to. Additionally, only bacterial reference markers are used, which may not be suitable for many microbiome studies. RAPSearch2 (Zhao, Tang, & Ye, 2012) was one of the first algorithms that could incorporate entire databases of protein sequences and find significant alignments at different levels of identity at accuracies equivalent to BLAST. More recently, most methods for alignment in protein sequence space have fallen in disuse and have become superseded by DIAMOND (Buchfink, Xie, & Huson, 2015), which uses a combination of spaced seeds (Burkhardt & Kärkkäinen, 2003; B. Ma, Tromp, & Li, 2002), reduced alphabet (Murphy, Wallqvist, & Levy, 2000), and indexing of query, as well as reference sequences to obtain speedups of as much as 20,000 compared to BLASTX.

Significant amounts of research have focused on developing methods that can accurately find alignments to protein sequences, which works well for almost all prokaryotic genomes because they are dominated by protein coding sequences, but which does not work so well with microbiomes rich in Eukaryotes such as fungi, oomycetes or protozoa, all of which have much more complex genomes with vast amount of non-coding sequence not represented in a protein database. MALT (Megan ALignment Tool) (Herbig et al., 2016) is a tool similar to DIAMOND, but it can find alignements in nucleotide space; unfortunately its implementation requires computer resources not available to most researchers. Minimap2 (H. Li, 2018), a nucleotide pairwise sequence aligner, has proven successful in analyzing metagenomic datasets. It specializes in long read data such as the ones generated by Oxford Nanopore's MinION or PacBio SMRT technologies. This type of data has proven to be extremely valuable in many genomic applications. In the case of microbial communities, it has been used to better estimate genetic diversity of human gut and aquifer sediment microbial communities (Bankevich & Pevzner, 2018).

Sequence alignment, and by extension sequence comparison, can be considered a pillar of bioinformatic analysis. The methods mentioned so far are grounded in the computation of these sequence alignments. Presently, this may be affected by several problems when studying metagenomic data from microbial communities. Such problems include the long processing times needed to evaluate vast amounts of data typical of high throughput sequencing. The growth of reference databases, to which continually new sequences from new as well as known species are added, introduces additional bottlenecks in mapping algorithms. Conversely, these databases may also suffer from lacking representative genomes or having significant biases towards oversampled taxa. These problems can be avoided by alignment-free pipelines, where sequence matches between queries and reference sequences is avoided. These algorithms usually rely on some data transformation of metagenomic and reference sequences into a form that can be processed much more quickly and with less computer resources. Two well known algorithms are Kraken (Wood & Salzberg, 2014) and Centrifuge (D. Kim, Song, Breitwieser, & Salzberg, 2016), which leverage the kmer content of reference and query sequences. In Kraken, a reference database is first built by extracting kmers from reference sequences and assigning them to the lowest common taxon in the tree of all taxa where the kmer was found to occur. Later, during the classification phase, kmers are extracted from query sequences and matched to kmers in the prebuilt database, and the distribution of assignments of all kmers in the query sequence is then used for classification. Centrifuge uses a similar algorithm, but additionally takes advantage of the Burrows-Wheeler transform (M. Burrows, 1994) and the Ferragina-Manzini index (Ferragina & Manzini, 2000), as well as variable length kmer size in order to achieve a significantly reduced memory footprint as well as increased performance.

The limitations of databases can in principle be overcome with alignment free methods. The use of machine learning has recently started to be applied in the areas of metagenomic classification and binning of metagenomic contigs in an alignment free way. Short kmers in the form of tetra- or pentanucleotide frequencies are extensively used in contig binning (Sangwan, Xia, & Gilbert, 2016), and hexa- and heptamer kmer frequency vectors of reference sequences are used to train non-negative least-square models to assign query sequences (Silva, Cuevas, Dutilh, & Edwards, 2014). Cui and Zhang train a support vector machine of short kmers in order to separate metagenomes from healthy and inflammatory bowel disease patients (Cui & Zhang, 2013). More recently, ideas of image processing have been translated to the world of contig binning where “textures” in DNA sequences can be identified and used as features (Kouchaki, Tapinos, & Robertson, 2019). Briefly, sequences are coded in local binary patterns according to some per-base predefined value, which results in each sequence being converted into a matrix to which singular value decomposition can be applied for reduced representation. The obtained features are used in stochastic neighbor embedding (Maaten & Hinton, 2008) in order to generate clusters of binned contigs. Finally, deep learning has shown utility in metagenomic classification: Rojas-Carulla and colleagues trained a deep convolutional network to label sequences at a desired taxonomic level (Rojas-Carulla et al., 2019). This was done via a transformation of query sequences into matrices using a typical machine learning technique called one-hot encoding. In this technique a sequence is represented by a matrix with as many rows as the length of the sequence. Each row has one column for every nucleotide and only the nucleotide present in the sequence receives a value of ‘1’, leaving the other columns at ‘0’. After this transformation, a convolutional network is applied for taxonomic prediction of the sequence.

Even though alignment-free classification methods and especially those relying on modern machine learning frameworks are continuously being developed, the fundamental lack of an alignment gives them limited utility in terms of the questions that can be addressed. For example, details regarding genetic variation, amino acid substitution, selection or recombination require the existence of alignment data, which makes traditional mapping algorithms still very necessary in today’s research.

Metagenomics of the Plant and Its Microbiota

It is evident that metagenomics is a very powerful tool in the study of microbial communities and, by extension, the host-microbiome system in which such communities interact with other organisms. The body parts of animals and certain environments have extensively been studied in this way, but much less research has used metagenomics to study the community of microorganisms living in and on leaves, roots, flowers and fruits of plants. It is known that many important plant traits are affected by the microorganisms living in leaves and roots (Müller, Vogel, Bai, & Vorholt, 2016). Opportunistic microbes can slow down growth or kill a plant, while beneficial ones can prime the plant immune system (Berendsen et al., 2012) or antagonize pathogens directly or indirectly through contributing to a suppressive environment (Mendes, Garbeva, & Raaijmakers, 2013). Microbes may adjust plant hormone levels (Glick, 1995) and participate in nutrient acquisition (Lareen, Burton, & Schäfer, 2016), among other mechanisms. Research in this area has revealed that most organisms on and in healthy plant leaves are bacteria, and 16S rDNA sequencing and culturing approaches have discovered many properties of the leaf microbiome (Müller et al., 2016). Other groups of organisms have also been found to play an important role in the phyllosphere such as fungi and oomycetes, which have remarkably different biology and in general more complex genomes (Sapkota, Knorr, Jørgensen, O'Hanlon, & Nicolaisen, 2015).

This lack of metagenomic studies has mainly been due to the difficulty of obtaining deeply sequenced plant microbiomes. In general there is substantial amount of host contamination in plant metagenomic samples (Müller et al., 2016), which makes it difficult to obtain sufficient data for downstream analysis. Nevertheless metagenomics has been successfully applied in plant microbial communities in a number of studies. The higher concentration of microbes in plant roots has lead more successful uses of this technology. For example, a combination of 16S rDNA and metagenomic sequencing was used to detect variation in bacterial root microbiota between wild and domesticated barley (Bulgarelli et al., 2015). In another case, metagenomic sequencing of root microbes was used to systematically address functional aspects of root colonization in wheat and cucumber (Ofek-Lalzar et al., 2014). In rice, metagenomics has revealed important aspects of the root endophytic compartment. Signs of different metabolic pathways such as polymer degradation, metal transportation, reactive oxygen species sequestration, and even nitrogen fixation were inferred from bacterial gene sequences (Sessitsch et al., 2012). In the case of phyllosphere communities, metagenomic sequencing followed by assembly together with metaproteomic profiling found signatures of methylotrophy in rice (Knief et al., 2012). Finally, another area largely unexplored is the study of viral communities in plants, where a substantial amount of diversity is believed to exist (Roossinck, 2012).

Results Summary

Whole metagenome shotgun sequencing of DNA extracts is an attractive tool for dissecting complex microbial communities such as those found on and in leaves of plants (Breitwieser, Lu, & Salzberg, 2017). Unfortunately, due to the difficulty of adequately capturing microbial complexity and diversity, we still lack a good understanding of the composition and dynamics of leaf microbial communities, and how they relate to other aspects of the host biology such as genotype, location, or environmental conditions. Because metagenomic sequencing supplies information on the total DNA content of microorganisms as opposed to just revealing what taxa are present, metagenome analysis enables us to ask many different types of questions not only about community composition but also bacterial strain dynamics in leaf microbial communities.

In this thesis I first present the use of metagenome sequencing to characterize the leaf associated microbiome (phyllosphere) of 275 wild *Arabidopsis thaliana* individuals from around Tübingen in southwest Germany, at 4 different timepoints between 2014 and 2016. Of these, 176 were also subjected to 16S rDNA and ITS1 sequencing. Because *A. thaliana* reproduces predominantly in a self-fertilizing fashion, this creates relatively homogenous subpopulations, which differ between local stands (Bomblies et al., 2010), and site of origin and host genotype were thus confounded. Microbial load varied widely, ranging from less than 1% of reads to up to 65% of sequenced reads, with low (< 100 Mb) to high (> 1 Gb) depths of microbe-associated sequences per sample obtained. As expected, the wild *A. thaliana* microbiota was highly variable between individuals; however, there were some clear patterns in the dominant microbes, in particular with *Pseudomonas* dominating in some sites and *Sphingomonas* in others. Overall, *Pseudomonas* reached the highest abundances, comprising >90% of all microbe associated reads in some samples, a phenomenon that greatly impacts compositional data such as 16S rDNA, because the increase in relative abundance of any one taxa reduces the relative abundance of other microbes.

However, when looking at actual taxonomic abundances in metagenome data, or using metagenome data to correct 16S rDNA, the effect of high abundance taxa such as *Pseudomonas* on other microbes was much more modest than relative abundance alone would have suggested. Relative abundance of eukaryotic and archaeal microbes is also analyzed, revealing a small but noteworthy prevalence of fungi and oomycetes.

In a second section, the high relative abundance of *Pseudomonas* and *Sphingomonas* was leveraged to study strain level variation in the core genomes of the two genera. For samples where sufficient coverage of the core genome was obtained, strain mixture and relative abundance was inferred by computing the relative abundance of non-reference bases at each single nucleotide polymorphism position, revealing distinct colonization patterns between the two taxa. Across-sample variation of metagenomic strains and comparison with existing reference sequences of different species in each respective genus allowed for the conclusion that *Pseudomonas* in metagenomes is mostly from the syringae/viridiflava complex and most strongly allied with a previously reported *Pseudomonas* strain referred to as OTU5, while *Sphingomonas* in metagenomes showed signatures of being composed of more complex mixtures of genetically distinct strains.

Methods

Profiling Leaf Associated Microbiomes of Wild *Arabidopsis thaliana* Populations with Metagenomic Short Read Sequencing

Sample Collection and Library Preparation

Wild *A. thaliana* individuals were sampled and processed by Derek Lundberg, Sonja Kersten, Dino Jolic, and Gautam Shirsekar in four distinct batches representing the evolution of our approach in four collection seasons spanning winter of 2014, spring of 2015, winter of 2015 and spring of 2016 from three sites in southwest Germany in the vicinity of of Tübingen, corresponding to Pfrondorf, Eyach, and Dettenhausen. Summary of collection numbers can be found in (Table 1).

Table 1. Samples by site

Season	Winter 2014	Spring 2015	Winter 2015	Spring 2016	**Total**
Site					
Eyach	33	21	40	44	138
Pfrondorf	25	23	–	49	96
Dettenhausen	–	–	18	22	40
Total	57	44	58	115	275

Batch 0: Pilot testing of two plant rosettes

In fall of 2014 a first plant visibly infected with what appeared to be *Albugo spp.* and *Hyaloperonospora arabidopsidis* was collected from Gniebel (48° 34' 34.10" North Lat., 9° 10' 55.42" East Long.) using sterile equipment and returned to the laboratory for storage at -80°C and further processing. A second plant in apparently healthy conditions was collected in Eyach. To remove loosely bound surface microbes, it was washed three times with sterile water. Both frozen samples were ground in liquid nitrogen and approximately 250 mg of material was used for DNA extraction using a custom protocol described in (Karasov et al., 2018). In short, samples were bead beaten in 1.5% sodium dodecyl sulfate (SDS) with 1 mm garnet rocks followed by SDS cleanup with 5 M potassium acetate and SPRI beads. Illumina short read libraries were prepared using TrueSeq® Nano kit with DNA shearing performed with a Covaris® S2 instrument. Instead of using the kit's Illumina adaptors, custom oligos were ligated following (Rowan, Patel, Weigel, & Schneeberger, 2015). Libraries were sequenced in one lane per sample of a Illumina HiSeq 2000 instrument using single-end 100 bp reads.

Batch 1: Shearing testing with 9 plants

In late December 2014, 9 plants were sampled in Eyach, collected with sterile instruments and brought back to the laboratory. Rosettes were then divided in three groups of three individuals and left unwashed, washed with sterile water, or washed with a Silwet L-77® solution at 0.02%, respectively, followed by quick freezing and grinding as described before. DNA extraction was performed as described for batch-0 plants. Wach plant was subjected to two different library preparation protocols that differed in the fragmentation method used: Covaris® S2 mechanical shearing and Shearase® enzymatic fragmentation (Table 2).

For mechanical shearing, 100 ng of DNA eluted in 130 µL of buffer was processed in the Covaris S2 instrument for 65 seconds at intensity = 4 and Duty Cycle = 10% at a frequency of 200 cycles per burst to obtain an approximate fragment size of 350 bp. SPRI bead cleanup in a 0.8:1 ratio was used for purification and finally eluted in 15 µL EB. Adapter ligation, A-tailing, and end-repair were performed as described in (Quail, Swerdlow, & Turner, 2009) following "Alternative Protocol 2" with the exception of using SPRI beads in place of AMPPure® XP beads together with custom adapters, as previously mentioned. For Shearase fragmentation a mixture of 100ng of DNA eluted in 20 µL of EB buffer with 9.5 µL of 3X reaction buffer and 0.5 µL of dsDNA Shearase Plus® was incubated for half an hour at 37° to provide a fragment size between 200 and 1000 bp according to the manufacturer. 3µL of EDTA was used to stop the reaction. Again, fragmented DNA was cleaned using SPRI beads following elution in 17 µL EB. Final steps of library preparation were identical to the Covaris approach. PicoGreen® was used to quantify 1 µL of DNA in 100 µL reactions for both methods. Afterwards, all reactions were pooled and size selected for fragments between 350 and 700 bp using the BluePippin® instrument. A single lane of a Illumina HiSeq3000 instrument was used to sequence all samples in 2x150 paired end mode.

Table 2. Fragmentation scheme by sample

Fragmentation	Covaris	Sherase	**Total**
Washing method			
Water	3	3	6
None	3	3	6
Silwet	3	3	6
Total	57	44	18

Batch 2: Set of 90 plants

In the fall of 2014 and spring of 2015, 90 plant rosettes were collected in Eyach and Pfrondorf, samples were brought to the lab in 50 mL tubes and washed as described for batch-0 plant-2. After quick freezing and storage at -80°C, leaf material was ground, followed by DNA extraction and sequencing library preparation using the same protocol as described in the Covaris methodology of batch-1 plants. All samples were sequenced in 2x150 paired end mode using one lane of a Illumina HiSeq3000 instrument. A subset of 59 samples was resequenced in 3 more lanes of the same instrument.

Batch 3: Set of 176 plants

A final batch of 176 plants was obtained from Eyach, Pfrondorf and Dettenhausen, all sampled in both December 2015 and March 2016. As previously described, whole rosettes were collected using sterile instruments and washed to remove dust and soil particles. During sample processing, two leaves were removed in order to isolate bacteria as described in (Karasov et al., 2018) following flash freezing, and sequencing library preparation. For this batch of samples, a modified Nextera® protocol was used in order to handle smaller volumes (Baym et al., 2015). During sample processing, 12 samples with large rosettes were split in two tubes each, which resulted in 12 samples prepared in duplicate, leaving a final sample size of 188. Following the same procedure as for batch-1 and batch-2, all 196 samples were pooled and size selected for 350 - 700 bp fragments. Paired end sequencing was performed on an Illumina HiSeq3000 instrument over multiple lanes.

Metagenomic Read Library Pre-processing

As a result of the multiple sample preparation methodologies implemented in the generation of metagenomic libraries, individual data batches were processed with slightly different pipelines. Broadly, sequencing libraries were first subject to demultiplexing based on adapter sequences and trimmed to desired phred quality. Batch-0 required no demultiplexing, as individual samples were sequenced in single lanes, while individual sequencing files for samples in batch-1 and batch-2 were obtained with a custom script (https://github.com/jregalad-o/plexSeq) based on custom barcodes in the adapter sequences. Finally, for batch-3 demultiplexed samples were obtained via the normal Illumina pipeline. Reads were trimmed at their 3’ end until an average phred score across the whole read of ≥20 was obtained using skewer. Reads shorter than 30 bp were discarded and in the case of paired end sequencing, only pairs where both reads passed filtering criteria were kept. After sample preprocessing a total of 1.5 Tb of reads were kept for downstream analysis (Table 3).

Table 3: Bases sequenced by batch

batch	1	2	3	4	**Total**
bases	0.23 Tb	0.04 Tb	0.39 Tb	0.92 Tb	1.56 Tb

Leaf Metagenome Pipeline

After preprocessing of samples, the following pipeline (Figure M1) was used to:

- derive microbial profiles of each metagenome,
- obtain host genotype,
- evaluate strain level variation of abundant microbes,
- generate metagenomic assemblies of non-host sequences.

In brief, sequencing reads were processed to separate host from microbial sequences, followed by taxonomic binning of samples, and SNP calling in host genomes. Read data determined to be not of plant origin was used for metagenome assembly of individual samples.

Due to the fact that leaf tissue almost always contains overwhelmingly more host DNA than microbial DNA, which leads to the majority of sequencing reads to be of plant origin, the first step of the analysis pipeline consisted in separating host associated data from sequences of possible microbial origin. In order to separate most of these plant sequences, samples were aligned to the TAIR10 *A. thaliana* reference genome using bwa as the short read mapper.

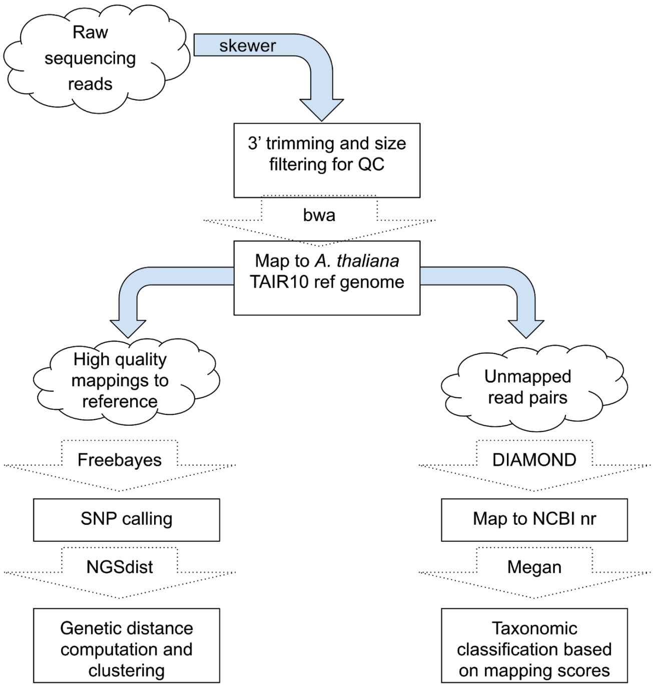
Raw sequencing reads
skewer
3' trimming and size filtering for QC
bwa
Map to *A. thaliana* TAIR10 ref genome
High quality mappings to reference
Unmapped read pairs
Freebayes
DIAMOND
SNP calling
Map to NCBI nr
NGSdist
Megan
Genetic distance computation and clustering
Taxonomic classification based on mapping scores

Figure M1. Schematic representation of primary data analysis pipeline

Mapping used a single reference genome as opposed to using a diverse set of *A. thaliana* genomes to keep analysis time short and minimise the possibility of randomly misassigning microbial sequences as host (because of the increased search space) at the expense of not being able to identify all possible plant derived data. This resulted in, on average, more than 90% of reads to be cataloged as belonging to the host, leaving all the sequences of possible microbial origin in the fraction of unaligned reads. In order to increase the certainty that a read was of microbial origin, only read pairs where neither mate mapped to the reference genome were kept for metagenomic processing and microbial profiling. Reads that did align to the plant reference genome were used for genotyping of the host, as described in a later section.

Metagenomic Profiling

Deriving microbial profiles was done in a two-step process: first, in order to fairly compare sequenced data against a representative set of reference sequences spanning a diverse set of organisms in a reasonable amount of time, sequences of possible microbial origin were mapped against NCBI nr protein (March 2018) database using DIAMOND (Buchfink et al., 2015). This was followed by taxonomic binning with Megan (Huson et al., 2016) using the weighted LCA (last common ancestor) (Bender & Farach-Colton, 2000) algorithm, described in detail later. To keep file sizes and processing time reasonable without compromising profiling accuracy, a maximum of 25 matches per read was allowed, which was more than sufficient for the LCA settings used as an average of 12 alignments per read were used for binning. The remaining alignments were discarded. The final step for taxonomic profile generation was individual read binning, consisting in placing each read that had mappings in the reference database above an alignment score of 50 in the taxonomic tree. In summary, for every read in the dataset, all the reference sequences that were matched to that read with a score that did not differ by more than 10% from the best score were used to place the read in a node of the taxonomic tree. After every single read was placed in the taxonomic tree, taxa count tables were obtained based on the number of reads assigned to each taxon, and in the case of higher taxonomic ranks, all taxa bellow any given node of the tree.

As a final step, count data and therefore abundance at any taxonomic rank were adjusted to sequencing depth and host DNA content by linear normalization. Final count tables were generated by adjusting individual sample counts based on the fraction of total sequences assigned to host chromosome sequences in that sample, multiplied by a common factor across the entire dataset. This can also be represented with the following formula:

$$\overline{X}_i = \overline{P} \cdot \frac{X_i}{P_i}$$

Where $\overline{X_i}$ is the normalized count vector for sample *i*, $\overline{P}$ is the mean number of plant chromosome read count, X_i is the raw count vector for sample i and P_i is the number of host chromosomal read counts for sample *i*.

Plant Genotyping

The final step of data processing consisted in the SNP (single nucleotide polymorphism) calling of host samples. For this, the original mappings to the host reference genome (TAIR 10) were used. After removal of PCR duplicates with Picard Tools (https://github.com/broadinstitute/picard), only alignments with mapping quality above 20 were used for genotyping, this yielded a median genome coverage of 20x. Afterwards, freebayes (Garrison & Marth, 2012) was used to generate VCF (variant call format) files of called genotypes. SNPs were filtered using custom scripts (https://github.com/jregalad-o/RegaladoLundberg2019) in order to retain only biallelic variants with a minimum alternate count of 3 and minimum read depth of 6. SNPs with more than 5% of total samples having no information were removed from the analysis. This yielded a total of ~1 million variants.

The raw SNP data were further processed to generate a full *n*X*n* genetic distance matrix, *n* corresponding to the number of samples. Raw SNP calls were transformed to an alternate allele count *m*x*n* matrix (*m* SNPs by *n* samples), with entries having values equal to 0, 1 or 2 corresponding to the number of alternate alleles at any given locus. A distance matrix was then generated with ngsDist (Vieira, Lassalle, Korneliussen, & Fumagalli, 2016) using standard parameters. Finally, genotype groups were derived from the x,y ordination coordinates with t-SNE (Maaten & Hinton, 2008).

Strain Level Diversity in Wild *Arabidopsis thaliana* Microbiomes

Sphingomonas Isolation and Library Preparation

Individual endophytic *Sphingomonas* spp. Isolates were obtained by plating homogenized, surface sterilized leaves from wild *A. thaliana* plants on culture plates containing solid LB with the antibiotic streptomycin to enrich for selective growth of *Sphingomonas* spp. (Vanbroekhoven K 2004). Plates were incubated at room temperature until visible colonies had formed (Figure M2). Subsequently, colonies that were deemed to likely be *Sphingomonas* (with diagnostic bright orange or yellow color) were inoculated into liquid LB and grown overnight. In parallel, colony PCR was performed in order to amplify 16S rDNA sequences for Sanger sequencing, to confirm each isolate's taxonomic classification. Sanger sequencing of PCR products was performed followed by taxonomic assignment, with which 20 of the isolates could be assigned to *Sphingomonas* spp. Genomic DNA was extracted from these isolates and Illumina sequencing libraries prepared using the modified Nextera protocol (Baym et al., 2015) followed by 2x250 paired end sequencing from ~650 bp fragments on the Illumina MiSeq instrument, yielding 6.7 Gb of data.

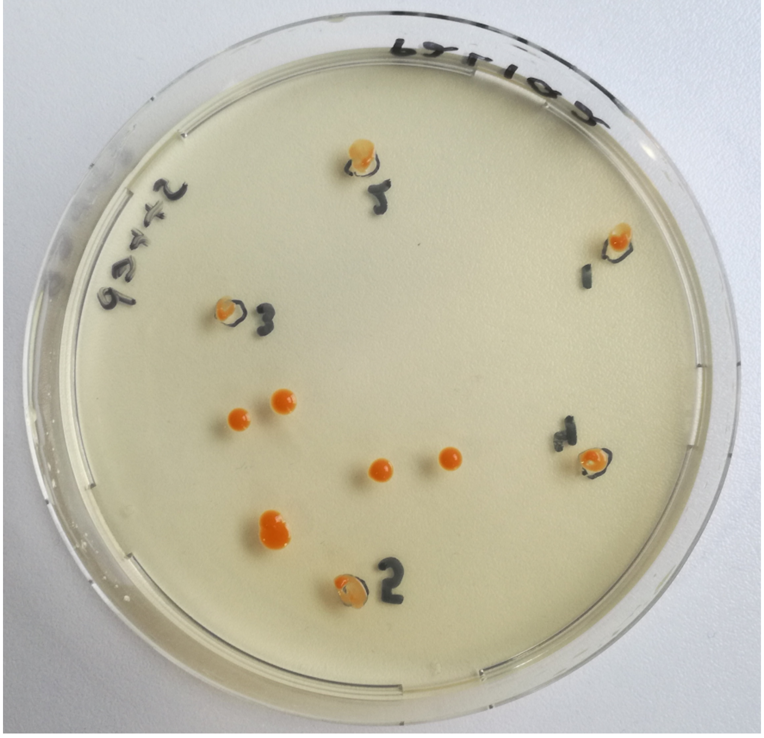

Figure M2. Representative plate of *Sphingomonas* spp. *Colonies*

Data Preprocessing

Before downstream analysis, individual sequencing libraries were QC filtered by trimming the 3’ end of individual reads until an average read phred quality ≥20 and minimum read length of 30 was obtained. Read pairs with any one read not passing these thresholds were discarded. To evaluate library quality, kmer frequency plots were generated with jellyfish (Marcais & Kingsford, 2012) and visually inspected to assess whether sequencing depth was sufficient for genome assembly. We obtained at least 25x coverage per sample.

Genome Assembly

QC filtered sequencing libraries were subjected to denovo genome assembly with Spades (Bankevich et al., 2012), and the resulting output contigs were polished using Pilon (Walker et al., 2014). Polished genomes were annotated using prokka 1.12 (Seemann, 2014) and annotations were analyzed with BUSCO (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) to measure completeness of assemblies based on single copy orthologs. The resulting annotation files were used for downstream analysis.

Reference Core Genome Setup

In the case of *Pseudomonas*, the reference sequence used was derived from a single isolate belonging to the OTU5 clade (Karasov et al., 2018) that is dominant in this population of host plants. There were 805 core genes, defined as being present in all isolates, for a total core genome sequence of 600 kb (Karasov et al., 2018). In the case of *Sphingomonas*, an independent core genome was determined using PanX (Ding, Baumdicker, & Neher, 2018) based on genomes produced in this work together with publicly available genomes of *Sphingomonas* isolates from *A. thaliana* (Bai et al., 2015). In total, >2,000 genes were identified as core, corresponding to a total of 1.3 Mb of sequence. Again, a hard core genome was used and the gene sequences of a single isolate utilized as mapping reference.

Full metagenomes were then mapped against these references using bwa in single-end mode and alignment files were processed with samtools to discard reads with alignment quality <20. PCR duplicates were removed with PICARD tools and final alignment files were processed for for strain level identification.

Variant Calling and Strain Differentiation

SNPs were identified using read support per position in the core genome computed with nQuire (Weiß, Pais, Cano, Kamoun, & Burbano, 2018). In brief, to determine variation across the reference, coverage per base was measured as the number of reads mapping over each position and observing how many bases in each read matched or did not match the reference sequence. A SNP was determined as present if it had a minimum coverage of five reads and was supported by at least three alternate base calls. Subsequently, strain relative abundance was measured by computing the fraction of reference calls against alternate calls. As a final step to compute sample wide genotypes, an alternate allele count matrix was constructed based on the variant detecting method previously described. For each sample, a count of the number of alternate allele calls per variant site was computed, resulting in a genotype matrix. This matrix was subjected to different dimensionality reduction techniques such as principal component analysis (PCA).

Chapter 2: Results

Whole Genome Shotgun Sequencing to Describe the Taxonomic Composition of Wild *Arabidopsis thaliana* Leaf Microbiomes

To evaluate to what extent shotgun sequencing can be used to assess microbial taxonomic diversity of leaves from wild *A. thaliana* plants, two test individuals from previously known populations in southwest Germany were sampled, taking care that both were still vegetative and had not yet begun to flower. The only distinguishable difference between samples was an apparently heavy white rust (*Albugo* spp.) infection co-occurring with downy mildew (most likely *Hyaloperonospora arabidopsidis*), two well known pathogens of wild *A. thaliana* plants (Coates & Beynon, 2010; Cooper et al., 2008). The first, diseased plant was processed without any treatment, while the second plant was subject to light washing with sterile water. This was performed in order to remove dust and soil particles that could harbor microbes that are not part of the leaf microbiome. Sterilization of the leaf surface was not performed, so that the analysis could capture not only microorganisms present in the internal (endophytic) compartment, but also from the external (epiphytic) surfaces. Many epiphytic microbes, even though they populate the external part of the leaves, are true plant colonizers and not stochastically present due to rain, wind, or other factors (Rout, 2014). Metagenomic shotgun libraries were constructed from total DNA extracts as indicated in methods (see Batch-1 plants). Next, samples were subject to high throughput sequencing without further treatment of libraries in order to increase the proportion of microbial sequences present. This yielded approximately 20 Gb of sequencing data per plant. In order to determine the amount of host derived DNA content, sequencing reads passing quality filters (see methods) were mapped to the Col-0 TAIR10 reference genome (Lamesch et al., 2012) with bwa mem (H. Li, 2013) using standard parameters (see Methods). All leftover reads, >60% of all reads for the infected plant and ~35% for the non infected plant, consistent with the majority of these reads being microbial. These reads were then subject to metagenome alignment using DIAMOND (Buchfink et al., 2015) again, with standard parameters, where individual reads are translated to each of the six reading frames and mapped to the NCBI nr (March 2018) multi-organism protein sequence database. Alignments resulting from this step were subsequently processed with MEGAN (Huson et al., 2016). In both samples, around 40% of non-host derived reads had at least one high scoring alignment to a sequence in the reference database. That the remaining fraction of reads without matches is of similar proportion, suggests that it most likely corresponds not just to host DNA sequences not included in the reference genome and thus impossible to capture via alignment, but very likely mostly to microbial sequences not included in the protein database, such as intergenic and ribosomal sequences, or microbial reads without close matches in known genomes.

In both cases, the majority of microbially assigned reads, >90%, corresponded to bacteria, with the remainder being classified as either archaea, oomycetes, fungi or other eukaryotic taxa. Nevertheless, clear differences could be detected between the microbial profiles of each sample.

The visually infected plant, plant 1, was colonized primarily by proteobacteria of all three major classes: alpha, beta, and gamma proteobacteria. Additionally, an important proportion of reads were binned as Oomycete with *Albugo* as the main genus within this group, consistent with the initial diagnosis of infection (Figure 1).

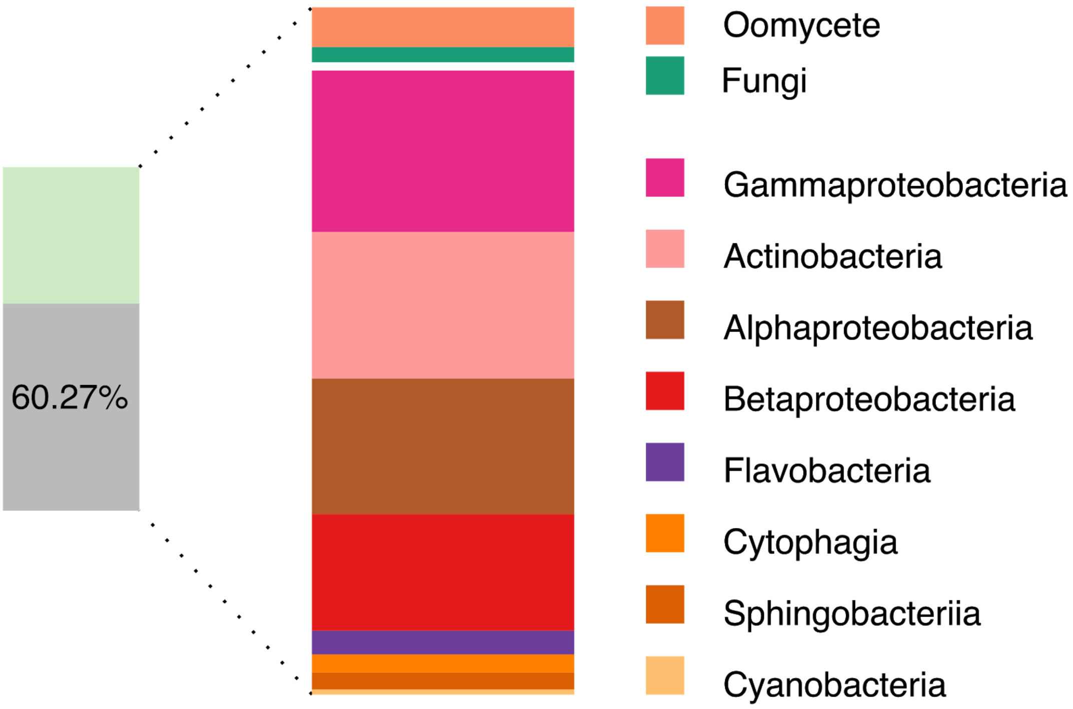

Figure 1. Batch-0, plant 1, microbial fraction and taxonomic composition

Left - Fraction of total reads sequenced that could not be assigned to the host genome (autosomes + chloroplast + mitochondria). Right - Microbial community composition at class level. Taxa are listed in sorted relative abundance of Eukaryotes followed by Bacteria.

In contrast, the not obviously infected plant, plant 2, presented a very different taxonomic profile, where Cyanobacteria represented almost half of microbial sequences detected, with drastically fewer eukaryotically assigned reads, showing that two *A. thaliana* rosettes can have widely different taxonomic profiles (Figure 2).

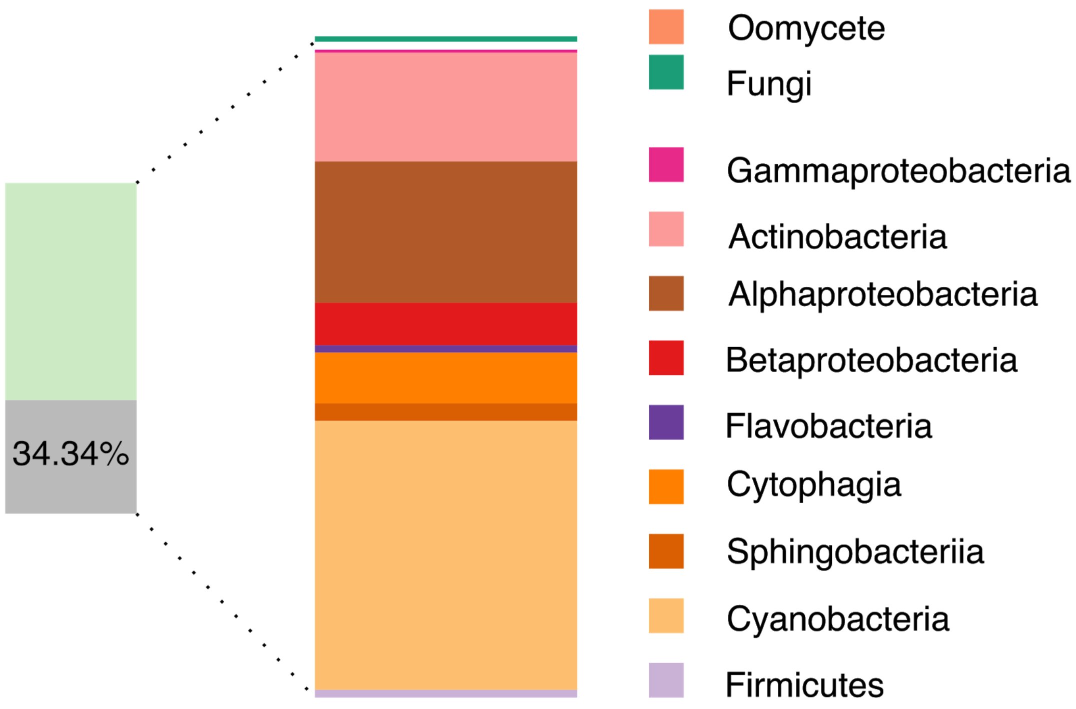

Figure 2. Batch-1, plant 2, microbial fraction and taxonomic composition

Left - Fraction of total reads sequenced that could not be assigned to the host genome (autosomes + chloroplast + mitochondria). Right - Microbial community composition at class level. Taxa are sorted in same order as in Figure 1.

To better understand the variation in microbial load as well as to better characterize the taxonomic composition of leaf phyllosphere communities, a dataset of three batches, with 275 wild *A. thaliana* plants, was assembled from four collection trips, spanning winter 2014/2015 and spring 2015/2016. These not only captured different ecological conditions, but also non-overlapping populations due to the annual life cycle of the host. All samples were washed with sterile water to remove soil particles as well as microbes loosely bound to the leaf surface. This step was considered particularly important based on the results of the pilot experiment with the visually uninfected plants, on which only a light wash had been performed, which had still resulted in a large fraction of microbial reads, many of which almost certainly originated from microbes weakly attached to the leaf surfaces.

With a larger collection of samples came the challenge of an accurate comparison of absolute microbial loads and therefore taxonomic counts between microbial profiles. Thus, the data for each sample were scaled by the number of mapped reads to the five *A. thaliana* chromosomes based on the correlation between sequences obtained from the nuclear genome and cell size (Beaulieu, Leitch, Patel, Pendharkar, & Knight, 2008). It was decided to exclude plastid reads from normalization, as plastid number per cell is known to vary based on developmental stage, genotype and environment (Mackenzie & McIntosh, 1999; Possingham, 1980).

This method of scaling microbial reads by host chromosomal reads is analogous to ‘spike in’ controls used to calibrate sample weights or volumes (Smets et al., 2016; Stämmler et al., 2016; Tourlousse et al., 2017). The result is a constant number of host reads while microbial counts are adjusted by host sequencing depth. Finally, a subset of 12 plants was processed in duplicate where ground plant material was split into two for independent DNA extraction and sequencing library preparation. In another set of 9 plants, the same DNA extraction method was used to test two different DNA fragmentation methods. This gave the opportunity to test the reproducibility of sample preparation as well as the analysis pipeline.

In summary, the proportion of total sequenced reads that could be assigned to microbial taxa, that is: bacteria, archaeae, oomycete, or fungi ranged from 1% to 45%. As with the pilot experiment, the vast majority of sequences were assigned as bacteria, and on average included 47 families. Across all samples Sphingomonadaceae and Pseudomonadaceae were consistently present as the most abundant bacterial taxa. Both taxa had remarkable differences in their distributions. Sphingomonadaceae were relatively constant in abundance, accounting for approximately 10% or reads in most samples, while Pseudomonadaceae varied substantially more, in some cases making up as much as 90% of all bacterial reads (Figure 3).

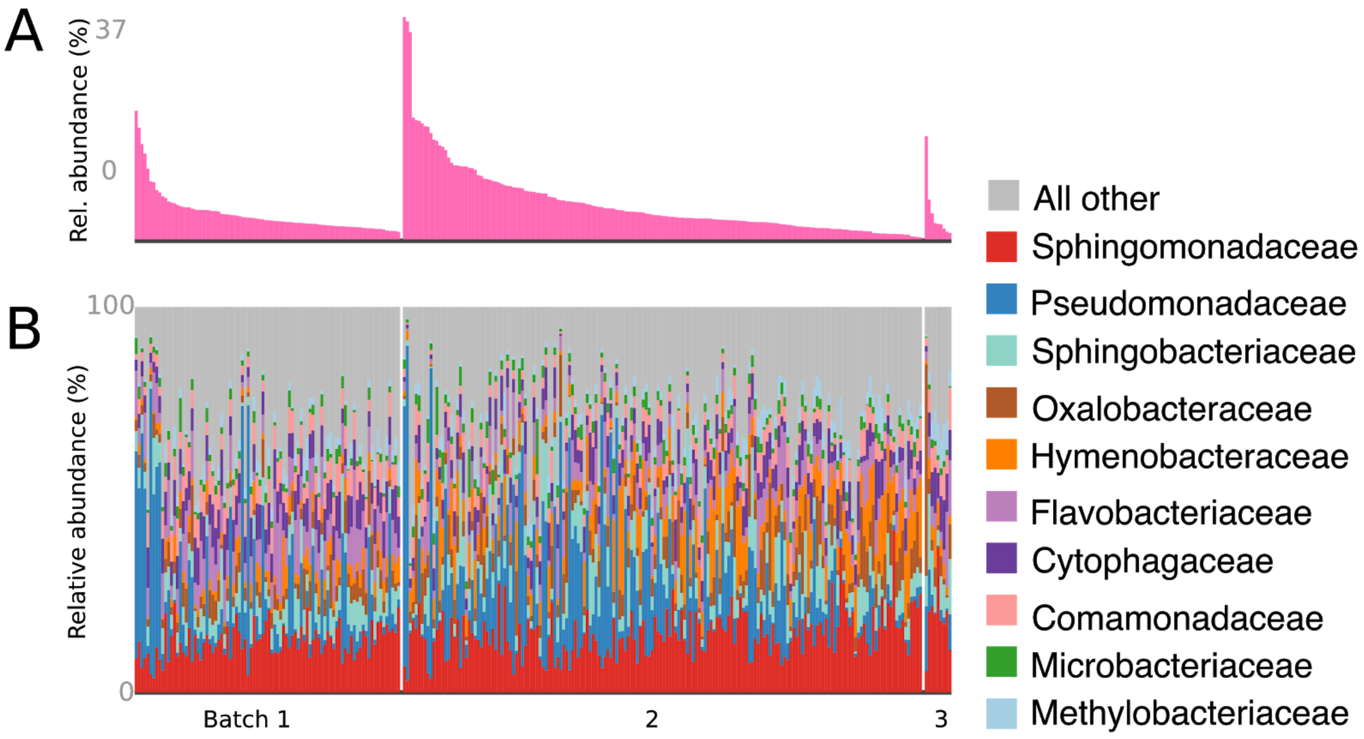

Figure 3. Bacterial taxonomic profiles of 275 plant samples

A - Scaled bacterial load as fraction of sequenced reads. **B -** Relative abundance of 10 most abundant bacterial families. Samples are organized by sample preparation batch and ordered from highest to lowest load. Bacterial families are stacked by overall abundance in the dataset.

The reproducibility of sample processing and library preparation as well as the analysis pipeline was tested with independent sequencing libraries generated from the same plant sample. Firstly, consistency in the measured relative abundance of microbes at family level was observed when subsampling as far down as 200,000 reads in two plants (Figure 4). For one, ground material was split, and for the other, the extracted DNA was split.

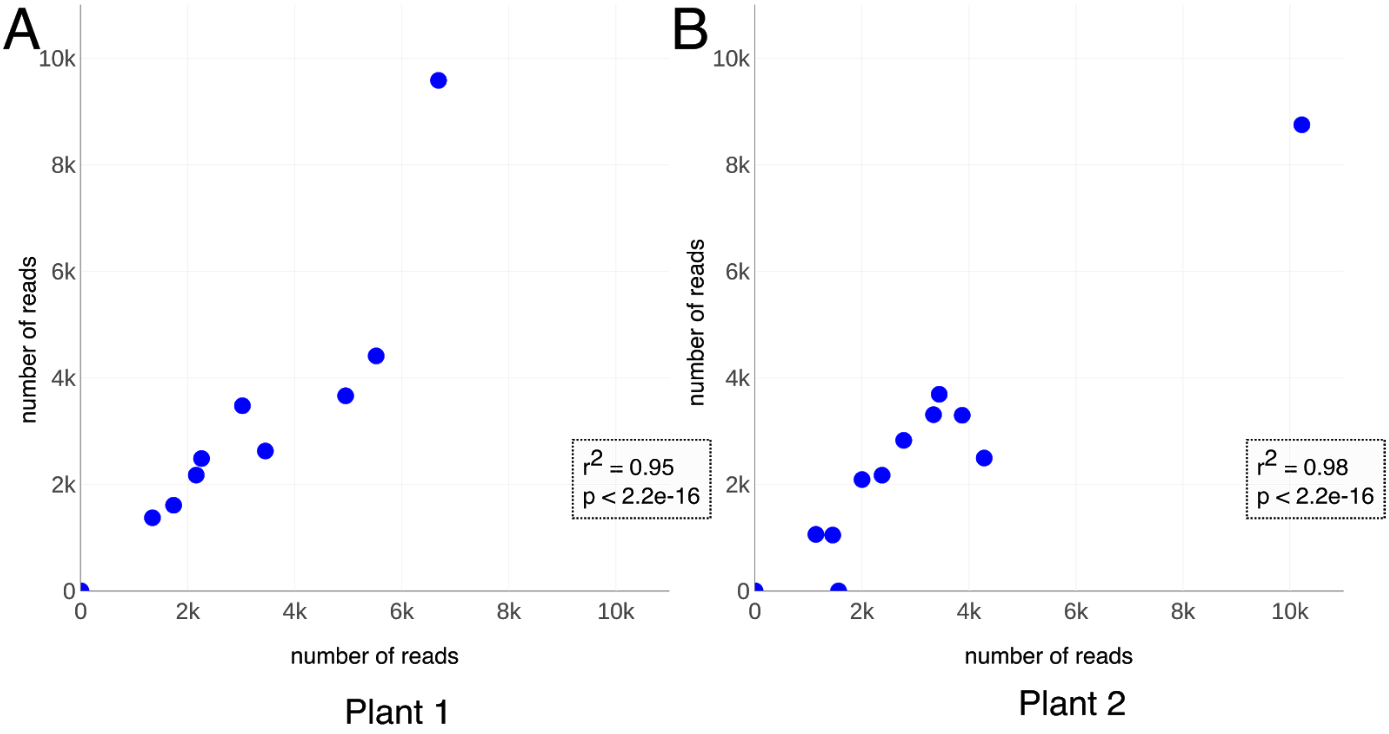

Figure 4. Family assignment correlation in replicated samples

Correlation of subsampled count data. Each plot corresponds to a single plant from which two samples were generated, either **A -** after grinding plant material, or **B -** after DNA extraction, and total microbial reads were downsampled to 200,000 reads for each library. Points correspond to individual taxa measured in read counts.

Additionally, any two microbiomes can be tested for similarity by using a distance measure that takes taxonomic composition as input and outputs a single number. For this case and throughout this analysis, the Euclidean distance was used to measure similarity between microbiomes. More specifically the square difference between all taxa in a pair of samples was added, with the square root of the sum representing the distance between the two microbiomes. Hierarchical clustering by microbial distances revealed that data from the same input DNA or from different DNA extractions were always the closest to each other (Figure 5a). Downsampling the number of non-host reads and repeating the analysis pipeline in replicated samples, a lower bound of 200,000 reads could be established as faithfully recovering plant of origin based on taxonomic profiles (Figure 5b). This high correlation between samples derived from the same plant was mainly influenced by high abundant taxa, which is due to abundant taxa being more easily detected after downsampling. In other words, taxa of low abundance are more likely to be missed or not be highly correlated after subsampling.

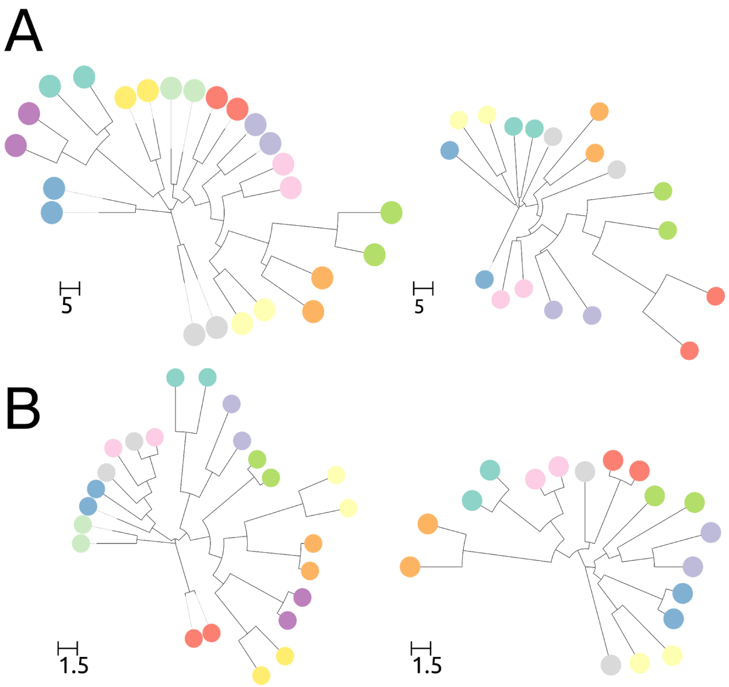

Figure 5. Hierarchical clustering on microbial distance

Dendrograms representing Euclidean pairwise microbial distances of samples processed in replicate. **A** - Distances computed on all taxa in entire dataset where plant material was split before DNA extraction (left) and where independent sequencing libraries were prepared from the same DNA extract (right). **B** - Same as in ‘A’ except distances were computed on a subsample of 200,000 non-plant reads subject to the same pipeline. In both cases, replicates are able to recover plant of origin. Individual plants are color-coded. Grey samples in right side trees (independent sequencing libraries) did not recover plant of origin due to initial low sequencing depth.

Influence of Site, Season and Host Genetics on the *A. thaliana* Microbiome

To understand the dynamics of leaf microbial communities, one needs to know how microbiomes fluctuate depending on variables such as host genetics, site of origin, or environmental conditions. Most commonly a similarity/dissimilarity index such as the Bray-Curtis index is calculated for each pair of samples studied, followed by ordination techniques such as principal component analysis (PCA). Unfortunately, two main problems arise with the most popular ecological indexes used: First, bias introduced by high abundant taxa can skew relatedness measurements to be overly determined by only the most abundant taxa, hiding potential sources of similarity or difference between samples. On the other hand, indexes such as the Bray-Curtis index do not satisfy the triangle inequality (Orlóci, 1974). This is a common mathematical criterium necessary for true distance metrics. It is established that given non overlapping three points A, B, and C, the sum of the distances AB + AC should always be greater than the distance BC. In other words, the sum of two sides of a triangle should be greater than the remaining side. If this criterion is not satisfied, the metric in question may be ill suited for downstream analyses. Therefore, to quantify similarities between microbiomes, pairwise Euclidean distances of double-square root transformed data were used. In this technique, the fourth root of individual count data is computed for all Bacteria, Fungi, and Oomycete taxa at any given taxonomic level. This transformation has a two fold purpose: It corrects for positive skewness in the distribution of data common when measuring species richness, as well as to mitigate the bias of highly abundant taxa due to the stabilizing nature of the square root function where larger values are affected more than lower ones. The distribution of Euclidean distances resulted in a typical bell-shaped curve (Figure 6) showing that, at first instance, there are no sizable clusters of samples with an above average microbial similarity.

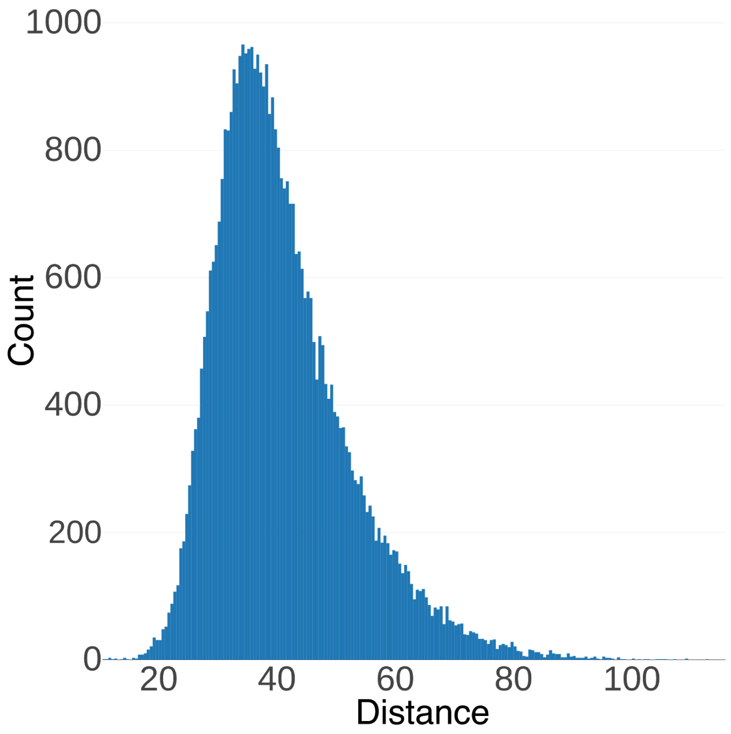

Figure 6. Distribution of pairwise Euclidean microbial distances

Shown is an all-against-all comparison of 275 samples (37,401 pairwise distances)

To investigate the effects of sampling site, PCA was computed directly on fourth-root transformed data and the first three principal components were visualized as a scatterplot. This showed a weak separation of samples by location, consistent with a slightly lower mean microbial distance within the two sites sampled, Jugendhaus and Eyach, compared to between-site distances (Figure 7).

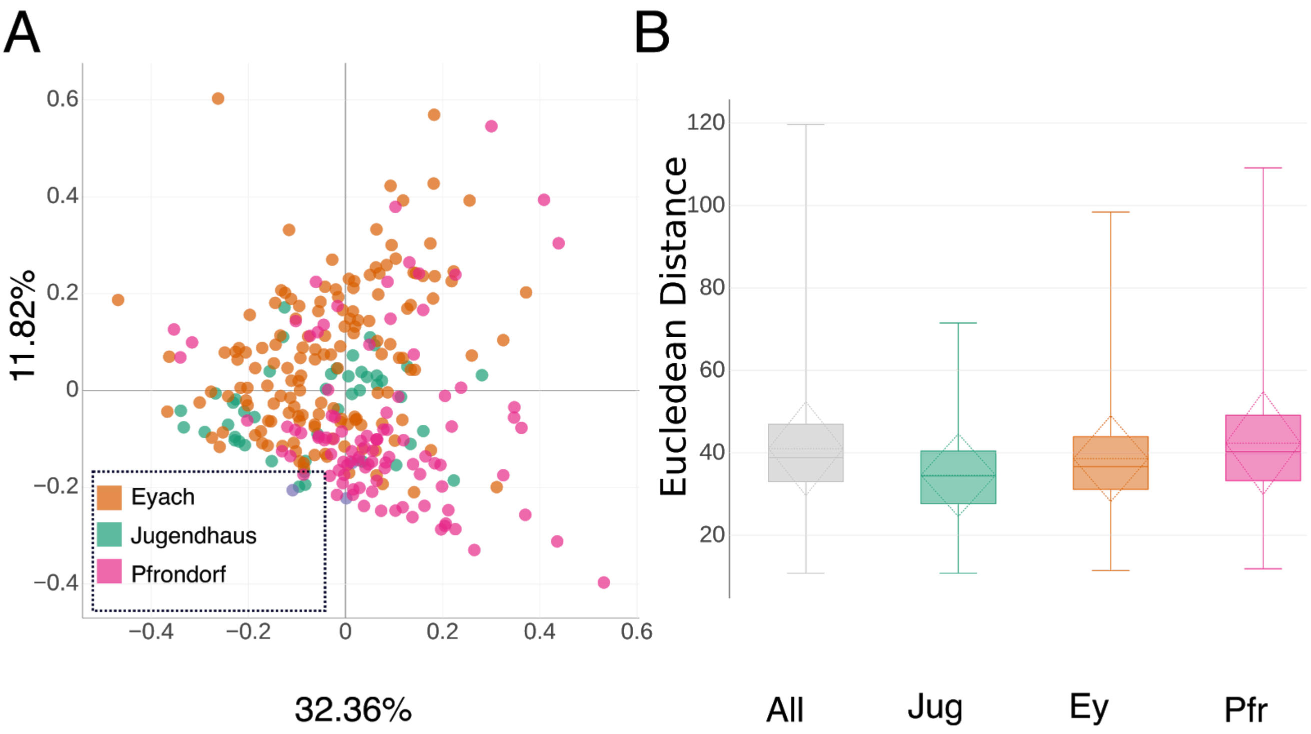

Figure 7. PCA transformed microbial counts and distribution of microbial distances per site of collection

A - Principal component analysis on scaled fourth-root transformed microbial counts; samples are colored by site of collection. **B -** Histograms of microbial distances grouped by site of origin. Colors are the same as in A.

Upon further inspection of variable loadings on principal component vectors, an effect of individual taxa could be observed, with the most abundant taxon in each sample correlating best with separation of samples in the largest principal component. Samples that either had Pseudomonadaceae or Sphingomonadaceae as the most abundant taxon were separated by the second principal component, while a combination of component 2 and 3 further separated individual samples based on most prevalent shared microbe (Figure 8).

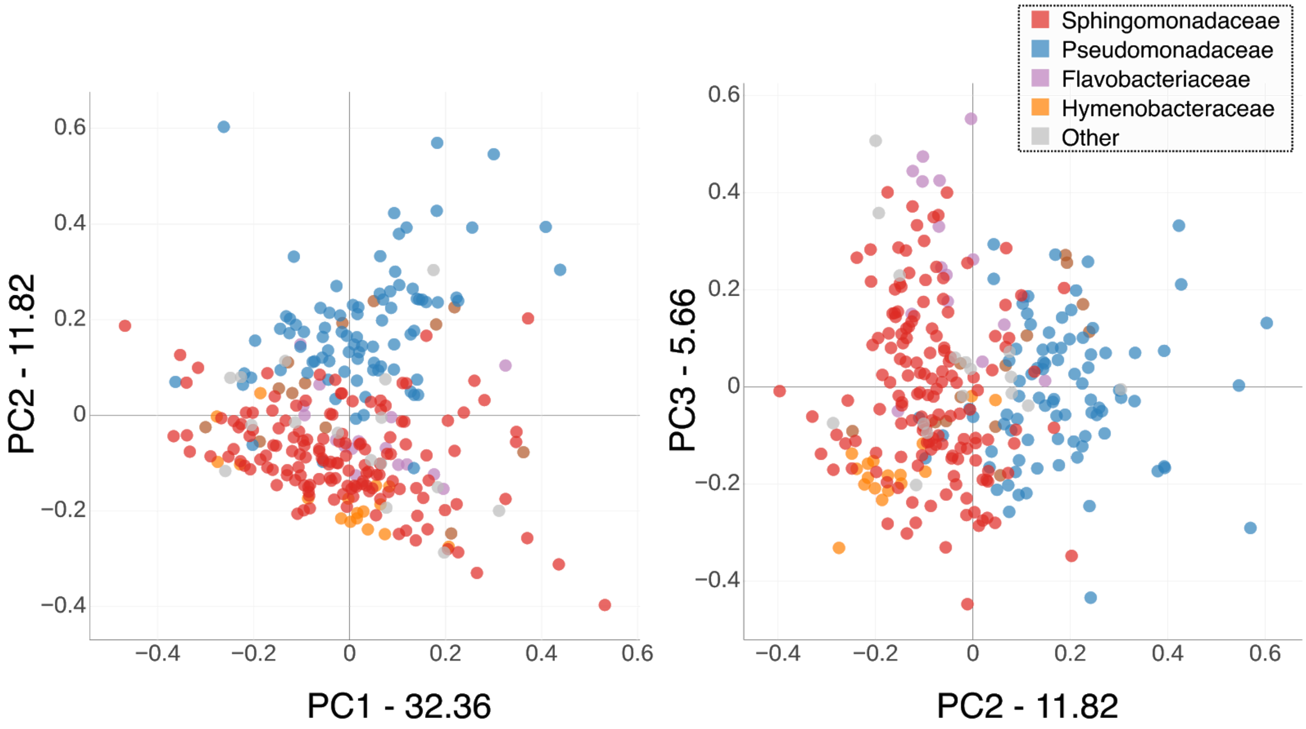

Figure 8. PCA of fourth-root transformed, scaled microbial counts, with individual data points colored by most prevalent taxa per sample

PCA as shown in figure 7. Samples are colored by most abundant taxa, colors follow figure 3. PC2 vs PC3 is also shown.

Finally, the excess of host derived sequencing reads was used to determine genotypic differences between individual hosts. Using freebayes (Garrison & Marth, 2012) with standard parameters, over 1 million high-confidence SNPs were obtained from quality filtered alignments to the TAIR10 reference genome. To measure relatedness between host genotypes, two approaches were taken. First, genetic similarity was measured by computing pairwise distances based on alternate allele count using NGS-dist (Vieira et al., 2016). The distribution of these distances revealed three distinct peaks (Figure 9), such a signal is indicative of well defined genetic structure within host genotypes. The most likely interpretation of the pattern observed in pairwise distances is the presence of samples that are very closely related, or identical to each other (first peak). The middle peak corresponds to the average distance between most samples, hence the tallest and broadest of the three. The last peak comprises the comparison between a small group of genetically distinct plants and the rest.

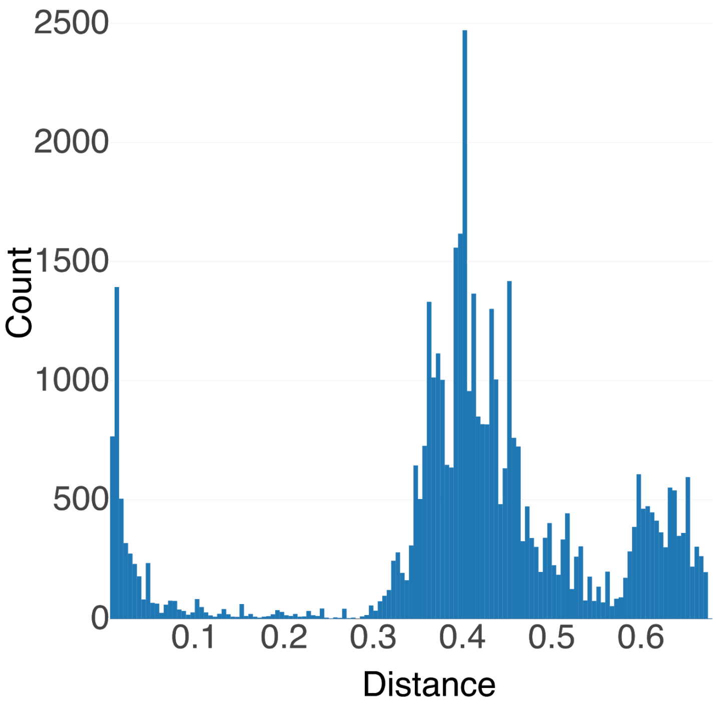

Figure 9. Distribution of pairwise genetic distances between host plants

Count frequency of pairwise genetic distance between host plants.

Additionally, stochastic neighbor embedding (t-SNE) (Maaten & Hinton, 2008) was used to visualize sample clustering based on the same alternate allele count data used to compute distances. This revealed clear clusters of samples not readily apparent in the genetic distance histogram, but supporting the presence of groups of plants with very high genetic similarity (or identity) (Figure 10).

Genetic clusters were correlated with sample collection site due to the reproductive nature of *A. thaliana*, which as a self fertilizer usually grows in local stands with a few groups of identical genotypes (Bomblies et al., 2010). An added advantage of knowing host genotypes is the potential to correlate microbial composition with host genetics (Bodenhausen, Bortfeld-Miller, Ackermann, & Vorholt, 2014; Wagner et al., 2016). Unfortunately, as stated before, genotype is strongly correlated with site, which makes both confounded and difficult to separate given the sampling setup described here.

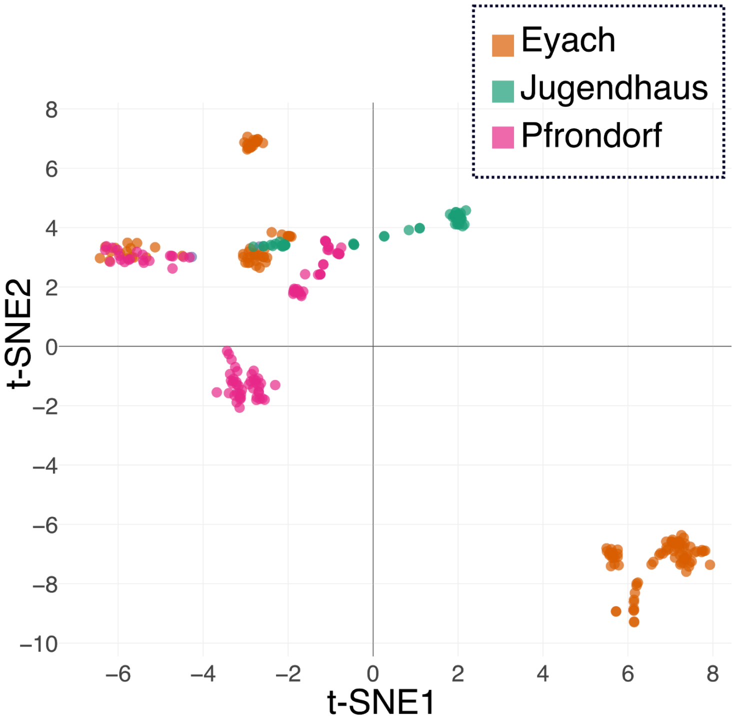

Figure 10. t-SNE plot on genotype calls colored by site of collection

Stochastic neighbor embedding computed on genotype count matrix for all 1 million SNPs. Samples were colored by location as in Figure 7. Pfrondorf appears to host three groups of identical or nearly identical genotypes. Jugendhaus likely host two groups of identical or nearly identical genotypes, as well as some individuals that are likely the result of crosses between distinct parents and that fall onto a line connecting the two main clusters. Finally, Eych hosts four groups of identical or nearly identical genotypes, with one being very different from the other genotypes, both at Eyach and at Pfrondorf and Jugendhaus.

Inferring Potential Microbial Interactions from Abundance Correlations

A key advantage of using whole genome shotgun sequencing data of the microbiome is the estimation of true microbial abundances, which permits the exploration of possible correlations between taxa. To investigate these interactions, taxa abundances were analyzed based on host chromosome count scaled data, fourth-root corrected counts, and finally relative abundance across all samples.

Firstly, all pairwise linear correlations were computed as Pearson’s product moment between all families having at least 1,000 assigned reads in at least 10 samples, in order to minimize spurious correlations of low prevalence taxa. Taxa pairs with absolute R^2 values smaller than 0.2 and with a p-value lower than 0.05 after Student’s t-test were ignored, to focus on strong correlations. Cutoffs were set in order to keep a sufficient number of nodes without cluttering the graph for visualization purposes (Faust & Raes, 2012; Zhou et al., 2010). This resulted in every taxon being correlated with on average 13 other taxa, mostly between high abundance microbes present in many samples, which are more likely to yield stronger correlation values at the desired significance threshold (Figure 11). All taxa pairs were positively correlated due to the dependence between overall load and individual taxa abundance. In other words, because the total amount of microbes in a sample is the result of adding individual taxa abundance values, overall load becomes a function of individual taxa load. Nevertheless, the amount that each taxon contributes to load varies across microbes, with families such as Pseudomonadaceae contributing substantially more than most other families. The same trend is observed after fourth-root transformation of counts, the only difference being increased correlation values and more correlated pairs due to the normalizing effect of the transformation.

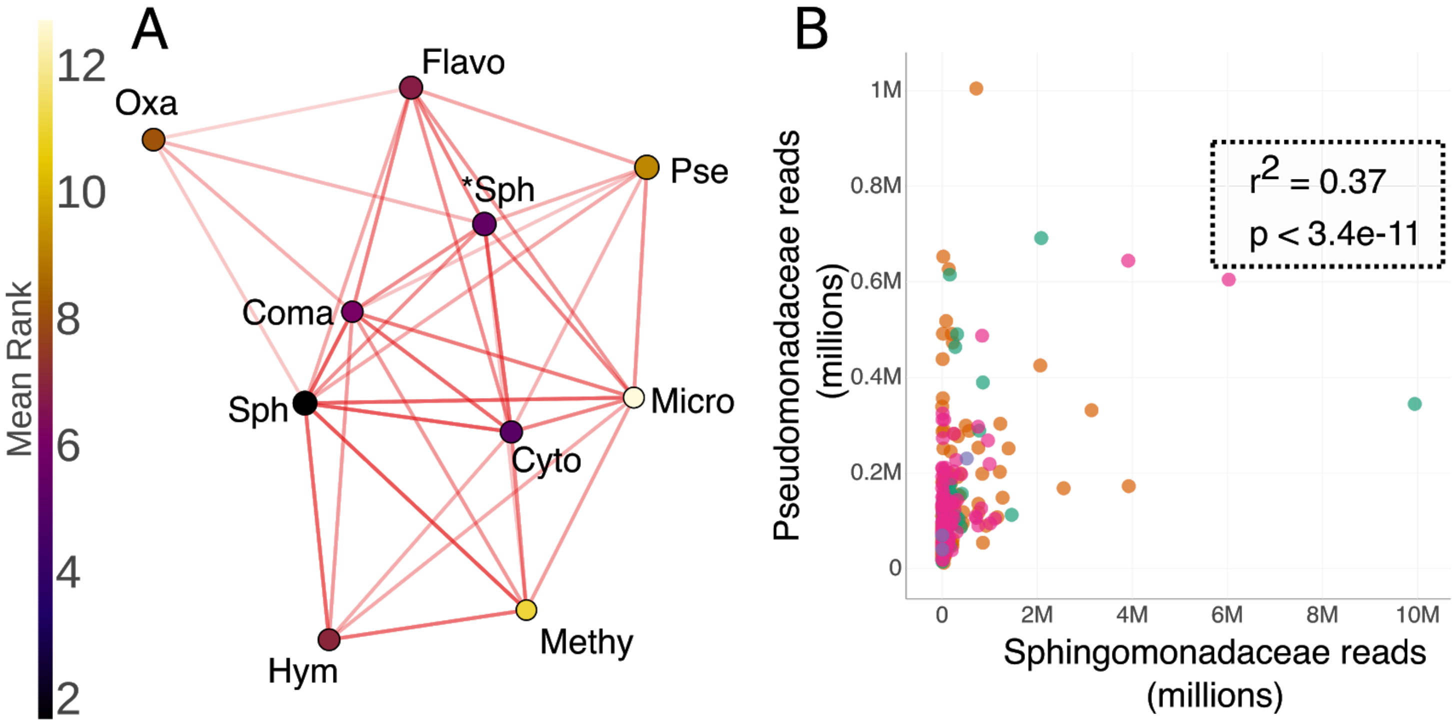

Figure 11. Load corrected correlation network of high abundant taxa and Sphingomonadaceae - Pseudomonadaceae correlation in individual samples

A - Pearson's product moment correlation network of microbial count abundance between highly abundant pairs. Nodes represent individual taxa colored by their mean rank across all samples in the dataset, edges are drawn based on linear correlation values between taxa across all samples. Pse - Pseudomonadaceae, Sph - Sphingomonadaceae, Hym - Hymenobateriaceae, Methy - Methylobacteriaceae, Cyto - Cytophagaceae, Micro - Micrococcaceae, Coma - Comamonadaceae, *Sph - Sphingobacteriaceae, Flavo - Flavobacteriaceae, Oxa - Oxalobacteraceae. **B -** Scatterplot of load corrected counts for Sphingomonadaceae and Pseudomonadaceae. Colors indicates sampling sites as in figure 7.

Count data can also be converted to relative abundance before computation of correlations, which is automatically the case for compositional measurements such as amplicon data, where differences in sequencing depth contribute no information other than higher confidence in estimates of low abundance taxa. In this case, measurements by definition become constrained as relative abundance transformation makes the sum of all taxa constant. In other terms, the relative increase in abundance of one microbe will be correlated with a decrease in all other microorganisms detected in the sample. This phenomenon, while not completely undesirable, can greatly influence any inferred interaction. For example, when metagenomic count data were transformed to relative abundance in this dataset, a number of correlations between taxa no longer passed magnitude or significance thresholds, while in other cases, interactions flipped from positive to negative. For instance, with compositional data, Pseudomonadaceae negatively correlates with all other taxa it previously correlated positively with (Figure 12).

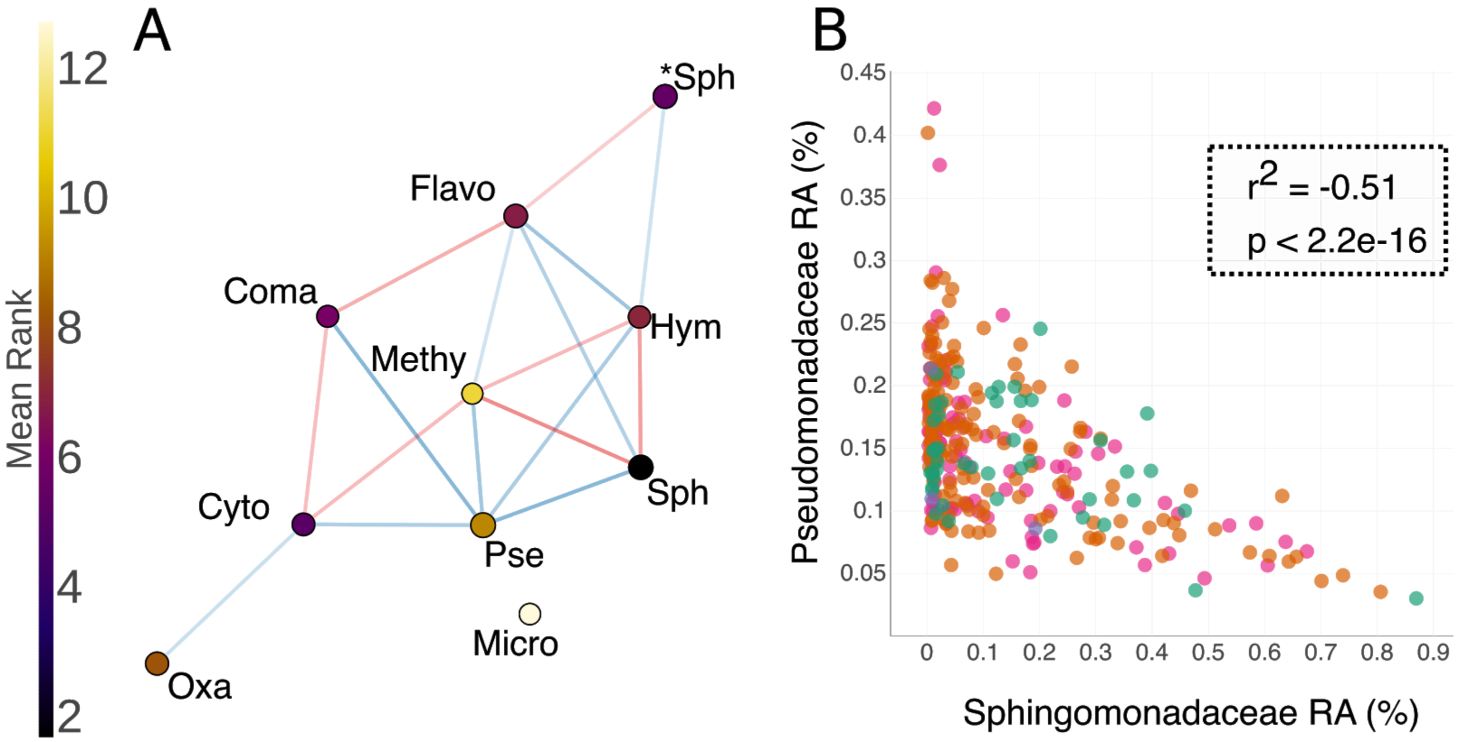

Figure 12. Relative abundance correlation network of high abundant taxa and Sphingomonadaceae - Pseudomonadaceae relative abundance scatterplot

A - Pearson's product moment correlation network of microbial relative abundance between high abundant taxa pairs. Nodes and edges are colored as in Figure 11 with the addition of negative correlations depicted in blue. **B -** Scatterplot of Pseudomonadaceae and Sphingomonadaceae relative abundance values. Colors indicates sampling sites as in figure 7..

Had load data not been taken into account, it would have been easy to interpret these results as evidence of antagonism between taxa pairs such as Pseudomonadaceae and Sphingomonadaceae. This has indeed been shown to occur between specific strains of these families when they were grown on an *A. thaliana* strain in gnotobiotic laboratory conditions (Innerebner, Knief, & Vorholt, 2011). Even though some taxa might be anticorrelated across samples, metagenomic count data, fourth-root transformed or not, did not provide any evidence for this.

The issues with compositional data just discussed have been thoroughly studied and methods to overcome them have been developed (Friedman & Alm, 2012; Gloor, Macklaim, Pawlowsky-Glahn, & Egozcue, 2017; Kurtz et al., 2015; Silverman, Washburne, Mukherjee, & David, 2017; Tsilimigras & Fodor, 2016), usually based on assumptions that may not necessarily hold true for all use cases. Having unbiased estimates of microbial abundance such as in this datasets provides an opportunity to use both approaches.

Metagenome Assembly of Leaf Phyllosphere Sequencing Reads

As an alternative to short read mapping, metagenome assembly of sequencing was performed, to overcome some of the disadvantages present when working with shorter sequences. Assemblies were done with MEGAHIT (D. Li, Liu, Luo, Sadakane, & Lam, 2015) using default parameters of the "meta-sensitive" preset. All resulting contigs were then size filtered to exclude all sequences shorter than 200 bp. This cutoff value was based on the 150 bp read length of the original reads, which implies that at least some overlap between two reads was required to reach this minimum contig length. It is important to note that average insert size of the sequencing libraries was 650 bp and paired end information was taken into account during contig computation, meaning that 200 bp constitutes a very lenient threshold.

To asses assembly quality, a number of metrics were assessed: N50 contig length, mean contig length, and total assembly size. Additionally, to evaluate to what extent reads were incorporated into contigs, input short reads were mapped back to their corresponding assembly.

Overall, this approach resulted in very poor results in all metrics being considered (Figure 13). On average, each sample resulted in ~30 Mb of assembled contigs, which is at most six medium-size bacterial genomes. Assemblies were also extremely fragmented, having on average

50,000 contigs after filtering by contig length. The average contig length across all samples was around 700 bp, a disappointing result considering this is approximately double the mean insert size of sequencing libraries. Finally, assembly N50 (the length of the shortest contig among all contigs that make up 50% of the assembly when ranking contigs from longest to shortest) was 650 bp.

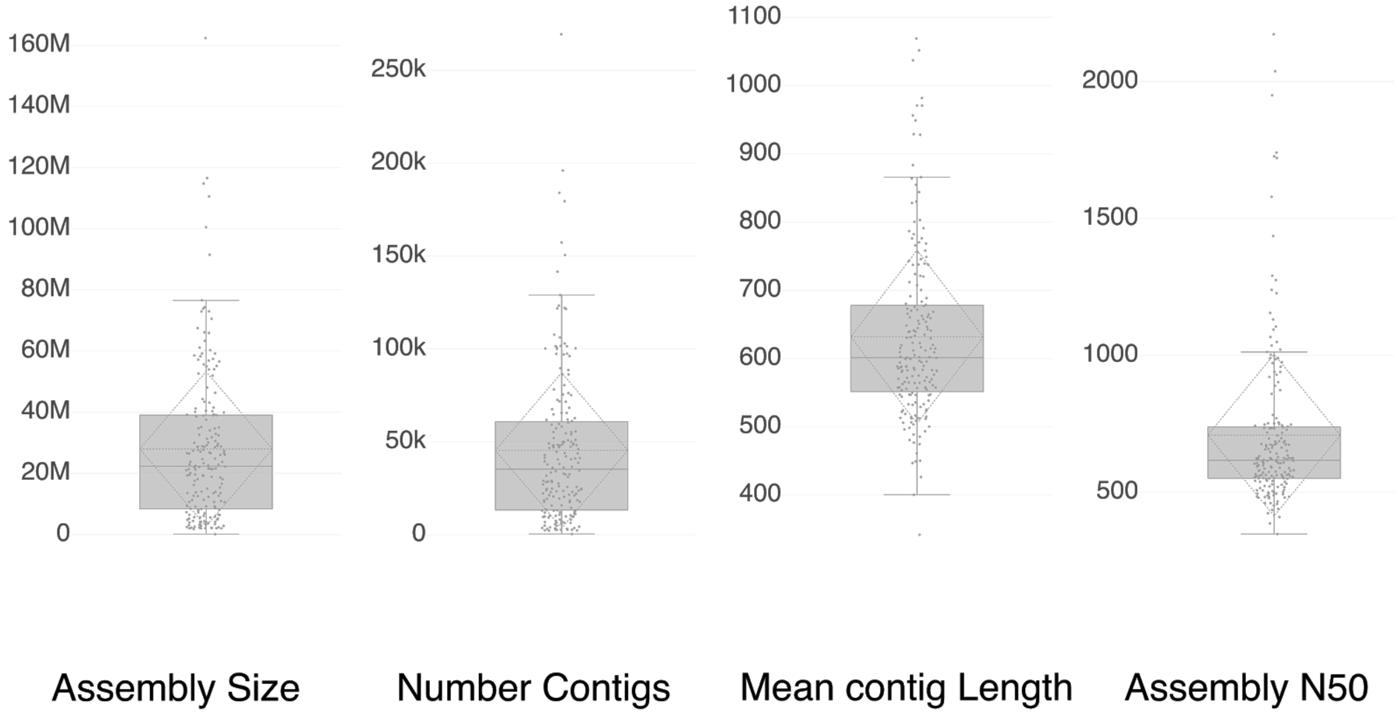

Figure 13. Metagenome assembly metrics

Boxplots of individual metagenome assembly metrics. All non host reads were used to generate assemblies. Points indicate individual samples, solid line indicates median, horizontal dotted line corresponds to mean and diamond dotted line corresponds to standard deviation.

This outcome is a consequence of several factors that are not mutually exclusive but inherent to phyllosphere microbial communities. Shallow sequencing depth due to the approach used resulted in overall low coverage of microbial genomes, which makes assembly difficult. Second, and most importantly, the leaf microbiome is a very diverse environment not only in terms of the number of taxa present, but also with respect to sequence diversity of closely related organisms.

When these two factors occur together, de Bruijn graph based assemblers such as MEGAHIT perform particularly poorly. Unfortunately, again, because shallow depth of sequencing, other assembly approaches like read overlap are not possible with this dataset. Ultimately, these outcomes demonstrate how challenging it can be to obtain full genomes from shotgun sequences in the presence of complex microbial communities in combination with shallow sequencing depth. For example, the average number of putatively microbial sequences in this dataset was 3.3 million reads. If a taxon is at 5% relative abundance,, assuming a genome size of 6 Mb, fewer than 200,000 reads could have been assigned to it. This would have corresponded to an average genome coverage of only 4.5x. Assembling such a genome would be challenging even from pure samples. This does not consider strain diversity, sequence homology of distantly related taxa, or uneven coverage, all factors that negatively impact genome assembly. Hence, only the most abundant, least diverse microbes are likely to produce informative contigs.

Correlation of Metagenomic Microbial Profiles with Amplicon Profiles

An important question when using whole-metagenome shotgun sequencing data to infer microbial profiles is how similar the inferences about relative composition are when comparing to amplicon based inferences. To this end, the V4 region of 16S rDNA of bacteria and the ITS rDNA region of fungi were amplified and sequenced, for samples from all 176 samples in batch-3. Taxonomic profiles were obtained and relative abundances of individual taxa per sample were compared to assess the congruence between amplicon derived and shotgun derived estimates.

In the case of bacterial taxa, there was high correlation (PCC $r^2 = 0.94$) (Figure 14) between the two approaches, with very few instances of taxa being included in only metagenome or amplicon derived profiles. Cyanobacteria were excluded from this comparison as the 16S rDNA sequences of this family are indistinguishable from those of host chloroplasts.

There were, however, some taxa that consistently deviated in one of the methods. For example, Pseudomonadaceae were almost always estimated at higher levels in shotgun data relative to amplicon data, while Sphingomonadaceae, Sphingobacteriaceae and Oxalobacteraceae all had lower relative abundance values in shotgun data. A number of reasons could explain the differences between the two techniques. For instance, smaller genome sizes will inevitably lead to fewer reads and thus underestimates of true abundance in shotgun data. In the case of 16S rDNA amplicon data, 16S rDNA loci can be repeated, and additional copies can lead to an overestimation of a given taxon.

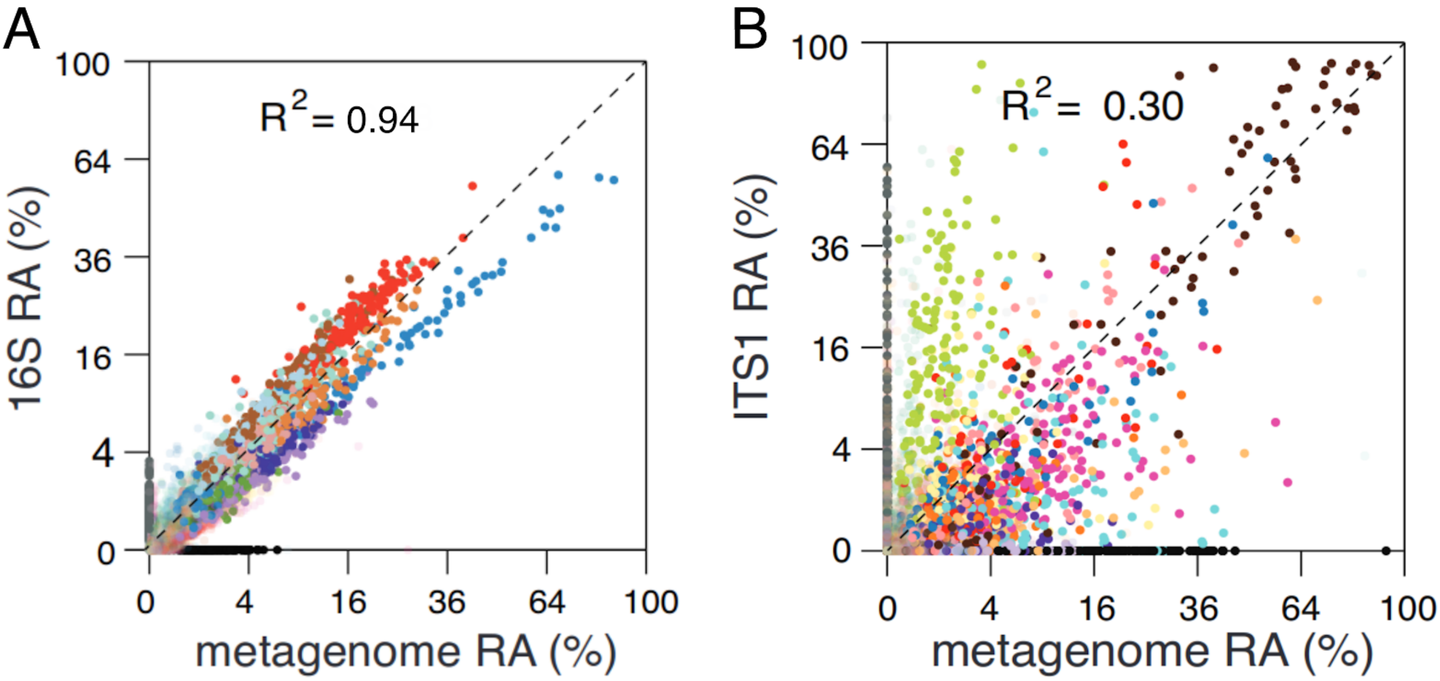

Figure 14. Correlation of taxa relative abundance values between metagenome and amplicon abundance estimation

Scatterplot of relative abundance values per sample per taxon in **A -** shotgun and 16S rDNA amplicon data and **B -** shotgun and ITS1 amplicon data. Each point corresponds to a taxonomic family in an individual sample. Colors indicate taxa as in Figure 3.

ITS1 amplicon data were much less correlated, (PCC $r^2 = 0.46$) with shotgun data (Figure 14). Several factors could explain the greater discrepancy for fungal taxa: First, fungi are present in much lower abundance, which in turns makes quantification from fewer sequences noisier. Indeed, if only the most abundant taxon is considered, Ceratobasidiaceae, r^2 increases to 0.96. Perhaps more importantly, genome size varies much more in fungi than in bacteria, which will skew abundance estimates compared to amplicon data, as rDNA copy number is not correlated with genome size. Similarly, since shotgun data are classified with a protein database, noncoding sequences, which compromise a larger fraction of the total genome in fungi, will remain unclassified. This is somewhat mitigated by the fact that genome size variation overwhelmingly affects noncoding, especially repetitive, sequences, with the protein coding content of genomes being much more constant. Lastly, primer amplification biases can lead to over- or underestimation of taxa with ITS1; this is likely the case for the family Helotiaceae. Similarly, on the analysis side, because of how the LCA binning algorithm works, many sequences of a particular taxon in the shotgun data may be placed at higher taxonomic levels due to higher similarity within fungal families, decreasing the number of sequences available for abundance estimates.

Strain Level Analysis of *Pseudomonas* and *Sphingomonas* in High-depth Leaf Metagenomes

Another key advantage of deep shotgun sequencing of microbiomes is the ability to analyze specific taxa of interest for strain level diversity in an unbiased manner, as no bacterial culturing has to be performed. This allows for the rapid and comprehensive detection of genetic diversity both within individual metagenomes as well as the overall population for focal taxa where sufficient sequencing reads can be obtained. In this particular case, *Sphingomonas* and *Pseudomonas* were the two most abundant genera in this dataset (Figure 15) and the most likely candidates to have sufficient genome coverage for the discovery of genetic variants. Additionally, these taxa have different colonization patterns, which should produce contrasting signals in metagenomic data, as some *Pseudomonas* strains behave as pathogens, while *Sphingomonas* is usually a plant comensal in nature (Innerebner et al., 2011).

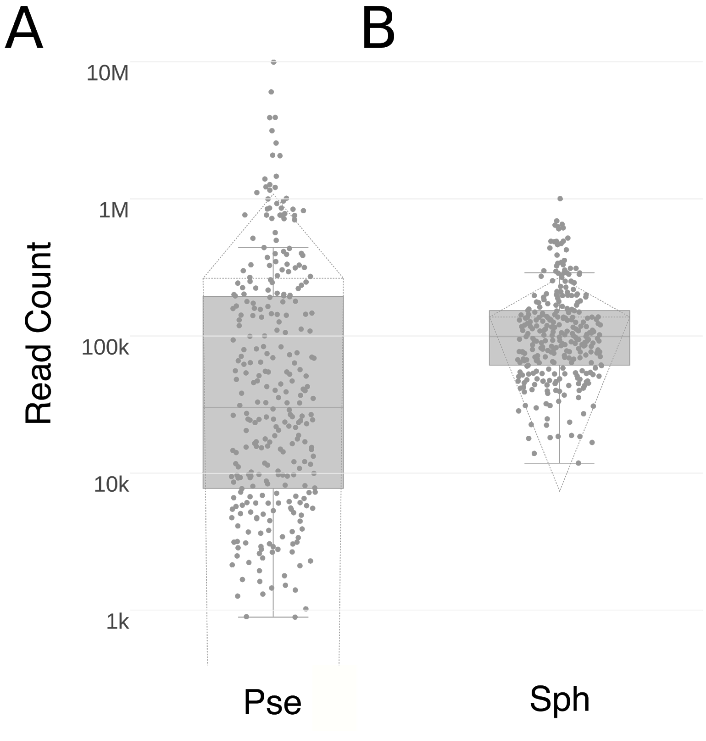

Figure 15. Normalized read count for Pseudomonadaceae and Sphingomonadaceae in *A. thaliana* metagenomes

Boxplot of distribution of read counts (in logarithmic scale) in all metagenomes for **A -** Pseudomonadaceae and **B -** Sphingomonadaceae. Solid line inside box corresponds to median read count, dotted line to mean and dotted diamond to standard deviation.

An important aspect of metagenomics to consider is that metagenomic data come from a mixture of organisms and by extension a mixture of strains/species of any genus of interest. These strains will have different levels of sequence variation and will not share 100% of their genes. This has two consequences: First, if the reference sequence used to call genetic polymorphisms is divergent from the strains in the sample, some sequencing reads from the taxa of interest may not have high quality mappings to such a reference. Second, sequences from a gene not represented in the mapping reference will be missed in the analysis, potentially introducing further biases. To address the second issue, a core genome of the species of interest can be computed in order to produce a set of reference genes more likely to be present in metagenomic samples (Medini, Donati, Tettelin, Masignani, & Rappuoli, 2005). In the case of *Pseudomonas* investigated here, a hard core genome was computed from the 1,524 genomes reported in (Karasov et al., 2018) using panX (Ding et al., 2018), with hard core genome defined by genes present in 100% of genomes. The genes of a single strain were picked as reference for the core genome. This strain was a representative of OTU5, the OTU that is the most abundant in *A. thaliana* leaves in this local population (Karasov et al., 2018). The reference core genome had 805 genes, comprising a little more than 600 kb of sequence. In the case of *Sphingomonas*, the core genome was produced de novo (Figure 16). In brief, 20 individual *Sphingomonas* ssp. isolates were obtained from leaves of wild *A. thaliana* collected in the same region as the metagenome samples, with genus membership confirmed by 16S rDNA analysis. Genomes were assembled from short read sequencing data with Spades (Bankevich et al., 2012) and annotated with Prokka (Seemann, 2014). The hard core genome was computed with panX (Ding et al., 2018), resulting in 1,955 genes (21% of all genes) for a total of ~1.9 Mb of reference sequence. The genes of a randomly selected strain were chosen as the core genome reference. Many more genes were present in the *Sphingomonas* core genome as a result of the much smaller number of strains used, 20 in this case.

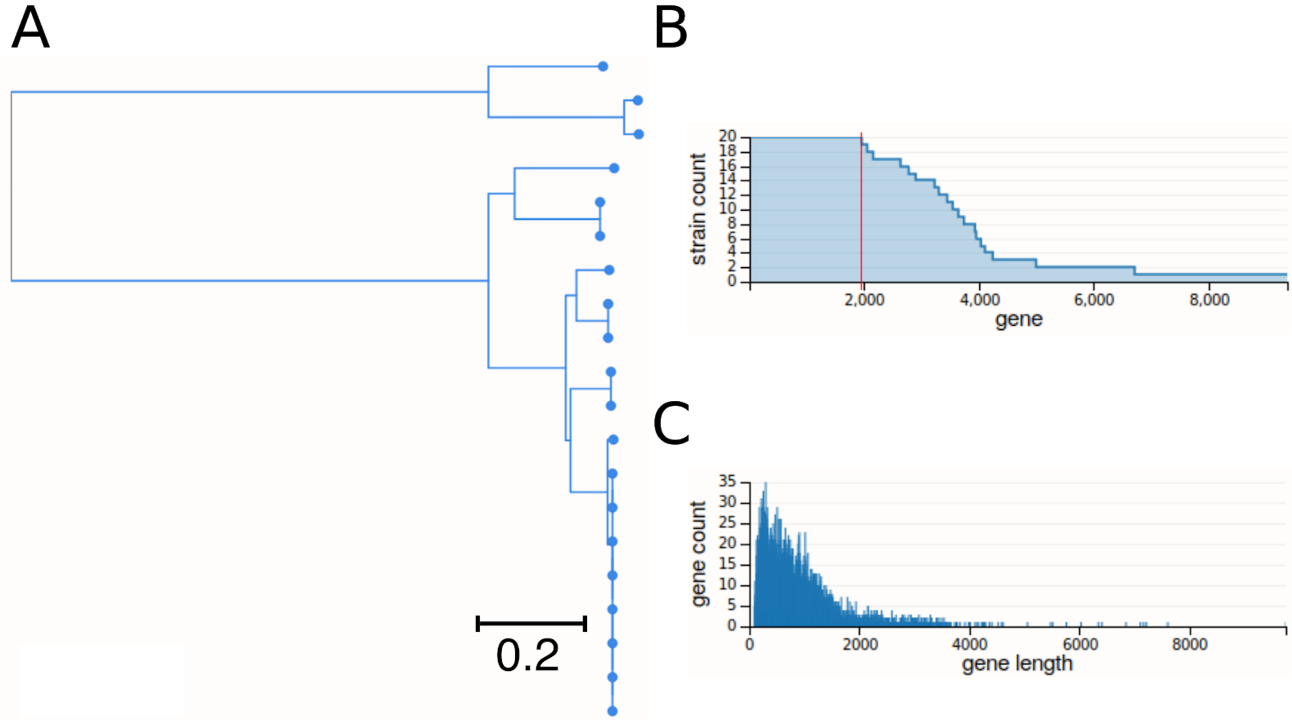

Figure 16. *Sphingomonas* core genome tree and pangenome analysis metadata

A - Core genome tree computed from concatenated core genes from *A. thaliana Sphingomonas* isolates, scale represents SNP rate. **B -** Strain count per gene in total pangenome. Red line indicates core genome cutoff, in this case 100% of strains. **C -** Distribution of gene length (in base pairs).

Using a 100% cutoff for both core genomes ensured the highest likelihood of these genes to also be present in the strains found the metagenomes.

Strain Identification by Core Genome Genotyping

With the core genomes in hand, SNPs were detected by mapping all non plant reads to each of the reference core genomes. All reads were used instead of only mapping reads already binned as either *Pseudomonas* or *Sphingomonas* in order to be as inclusive as possible, since the metagenomic classification pipeline places a significant amount of reads in taxonomic nodes higher than the genus level. This comes at little extra computation time, but with the disadvantage that sequences not belonging to the taxa of interest might have positive matches with genes present in the core genome, but because *Pseudomonas* and *Sphingomonas* were generally the most abundant taxa, this was not considered a critical issue. After mapping, PCR duplicates were removed and core genome coverage computed for all samples (Figure 17).

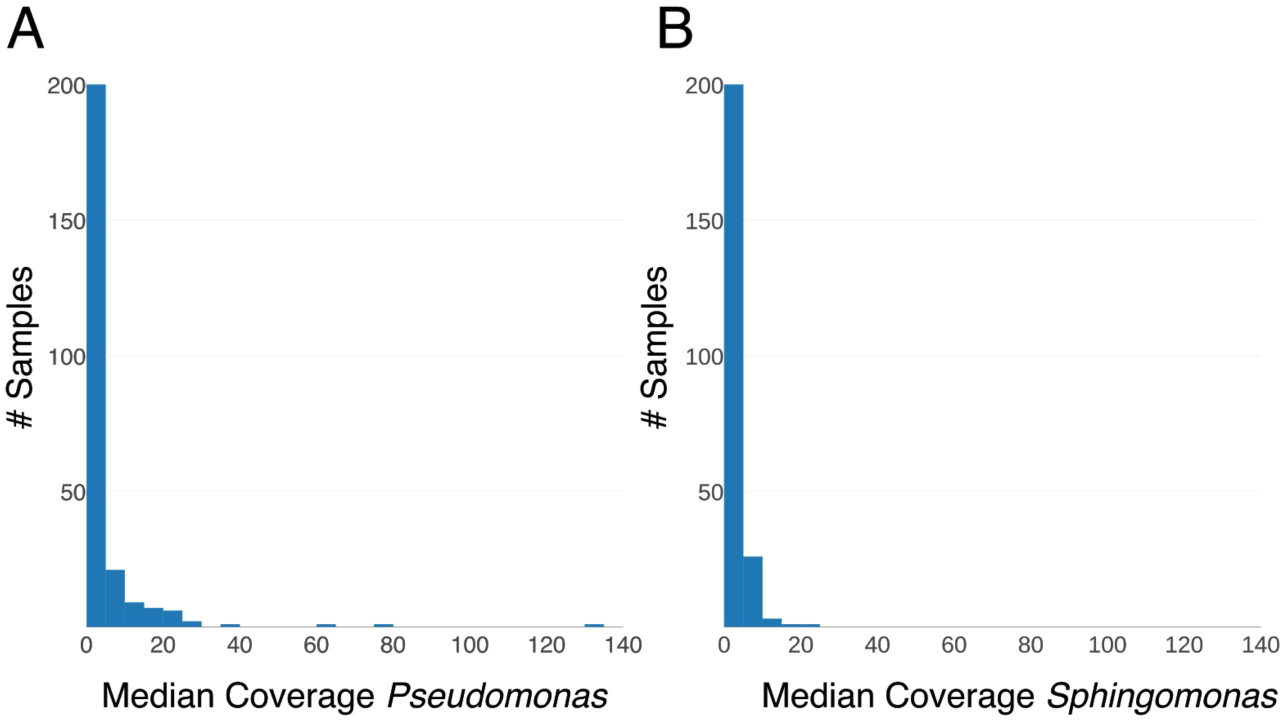

Figure 17. Median core genome coverage for *Pseudomonas* and *Sphingomonas*

Distribution of median core genome coverage for **A -** *Pseudomonas* and **B -** *Sphingomonas*. Median coverage is shown instead of mean due to the long right side tail in mean coverage distribution.

Some individual *Pseudomonas* samples had the greatest core genome coverage, with up to 130X sequencing depth, but only up to 25X for individual *Sphingomonas* with the highest coverage. Even though in the overall dataset *Sphingomonas* was more abundant than *Pseudomonas*, its core genome was present at a lower depth because it accounts for only ~20% in individual samples across the entire dataset (Figure 3). In the case of *Pseudomonas* most samples were sequenced at less than 3X coverage, but because some microbiomes contained an extremely high proportion of

Pseudomonas sequences, up to 90% (Figure 3), the core genome coverage in these samples is much higher.

For downstream analysis of SNPs, samples were selected where at least 50% of the core genome reference sequence was covered at a minimum depth of 5X, to mitigate noise introduced by sequencing errors. There were 49 samples with sufficient data for *Pseudomonas* and 31 samples for *Sphingomonas*.

Per base depth distribution indicated a largely even coverage of the reference core genome. In other words, the depth distribution had a single peak which shows that most positions are covered at a single depth with a long right sided tail of few bases with unusually high coverage, most likely the result of small repeats. This translates to the reference core genome capturing read mappings from a single source organism or collection of closely related ones (Figure 18).

An important issue to point out is that even with the use of a hard core genome, there is still a substantial risk of excluding samples that might have enough depth for SNP analysis. That is, because they lack genes considered to be core, some microbiomes are excluded. The thresholds of 50% reference coverage at at least 5X depth aim to compile a set of samples that can be fairly compared to each other. In other words, the samples chosen will have enough shared sequence. Excluding samples with insufficient coverage is important because positions in the core genome without information are in many analysis pipelines considered as reference, which will introduce problematic biases (Nielsen, Paul, Albrechtsen, & Song, 2011). In addition, if many variable positions are informative only for partially overlapping sets of samples, different parts of the core genome might lead to contradictory conclusions. In such cases, metagenomic assembly could be better suited to extract genes common in all samples.

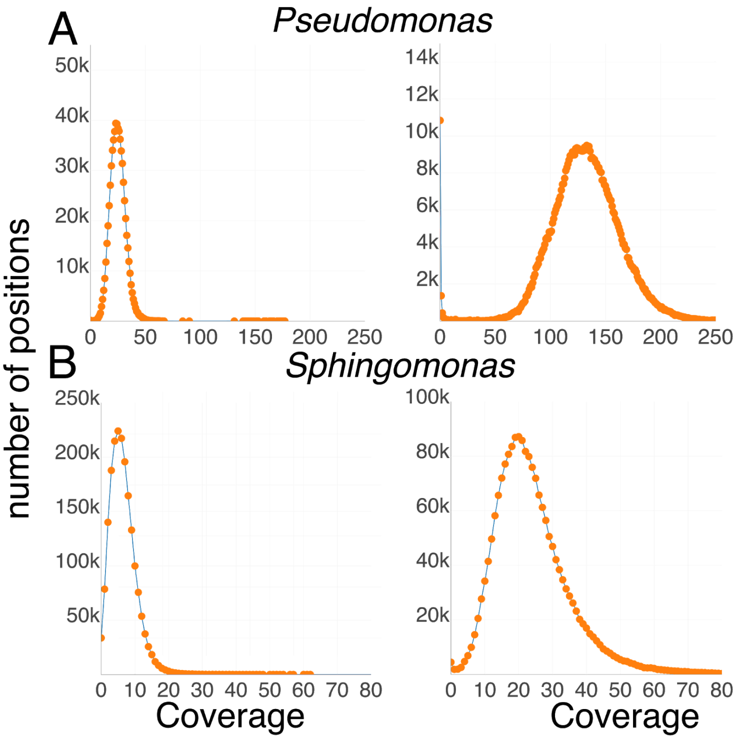

Figure 18. Per-base coverage distribution in individual high-depth and low-depth samples for *Pseudomonas* and *Sphingomonas*

Per-base core genome coverage distribution for four selected samples in **A -** *Pseudomonas* and **B -** *Sphingomonas.* Left, samples with low sequencing depth. Right, samples with high sequencing depth.

When microbiomes contain related strains of the same taxon, these will normally occur at different relative abundances. For example, two strains might be at relative proportions of 2:1 in one plant, while in another host they might exist together with a third strain in a 4:1:1 proportion. To decompose variation in such a mixture, the relative abundance of mapped bases per position in the reference genome can be used. By computing the distribution of frequencies across the entire core reference genome, it is possible to infer the presence of mixed populations, and, if coverage is sufficiently high and abundances are sufficiently different, assign different variants to different strains. Briefly, each position of the core reference genome is scanned, and if the position is covered by at least five bases, the relative abundance of all non reference nucleotides is computed. An alternate base (mismatch) is considered to be present if it is observed in at least 3 reads, if all conditions are met, the position is considered a SNP (Figure 19).

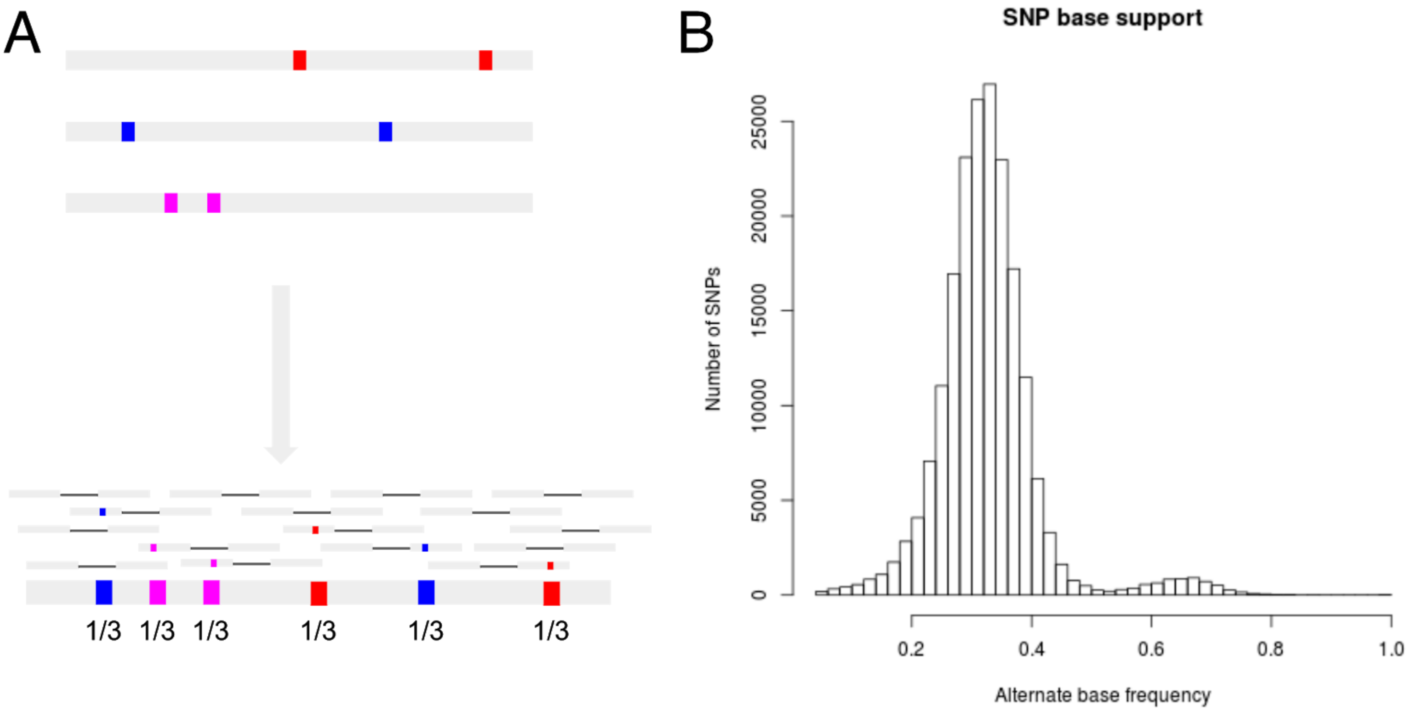

Figure 19. Schematic representation of a 3-way strain mixture at equal proportions and simulated distribution of alternate allele support.

Schematic representation of strain mixture detection based on distribution of non reference base frequency. **A -** top: Diagram of genomes from three strains present in equal proportions. Bottom: simulated reads and detection of SNPs and their proportions. **B** - Distribution of non-reference base frequency counts. Data is based on a simulation of a mixture of three genomes that have independently diverged. As expected from a three-way mixture of strains, a major peak of non-reference SNPs at ⅓ is observed, accompanied by a minor peak at ⅔, corresponding to non-reference SNPs shared in two of the three strains.

The rather high alternate base threshold assures true segregating sites are recovered, avoiding the inclusion of mismatches introduced by taxa other that *Pseudomonas* and *Sphingomonas*, but at the expense of not being able to detect SNPs from strains at very low abundance. In total, 20,437 SNPs were obtained for *Pseudomonas*, corresponding to 3 SNPs per 100 bp of core genome for an average fraction of 0.03 of polymorphic sites, and 147,681 for *Sphingomonas* or 7 SNPs per 100 bp of core genome. Most variants corresponded to diallelic sites where only a single alternate allele was observed (Figure 20).

Multiallelic sites were more prevalent in *Sphingomonas*, indicative of a greater diversity of this taxon in leaf metagenomes, not only in terms of core genome diversity, but also with respect to the number of segregating alleles, with a handful of sites even having four alternate states (any one of 3 non reference nucleotides and deleted).

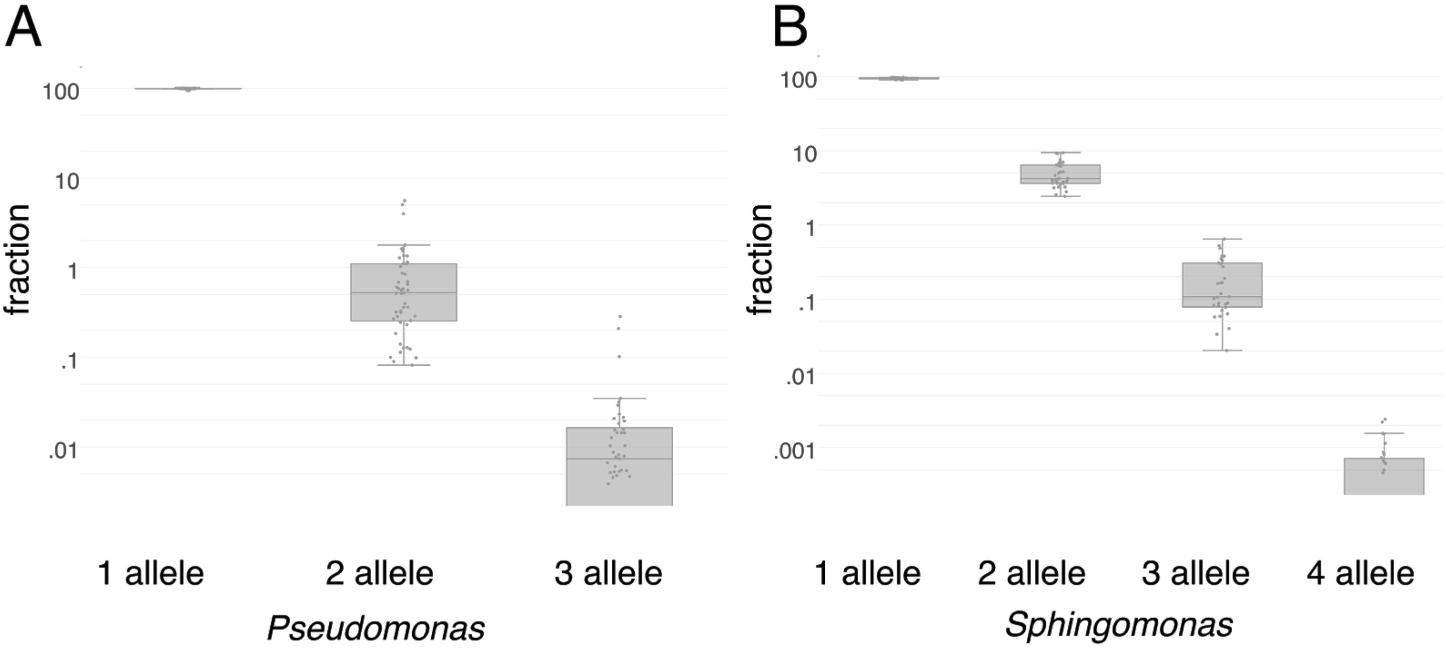

Figure 20. Fractions of different classes of non-reference variants

For each sample SNPs are grouped based on the number of alternate bases at each position. **A -** *Pseudomonas* and **B -** *Sphingomonas*. Fractions are plotted on logarithmic scales. For any given position, four alternate alleles are possible if deletions are counted.

To ascertain that the pipeline used largely ignores sequencing errors, the focus on coding sequences in the core genome was leveraged to detect whether variants were enriched for synonymous changes – as expected under neutral evolution –, which preferentially occur at third codon positions (Figure 21). In both *Pseudomonas* and *Sphingomonas* genomes, the vast majority of polymorphisms occured in third-codon positions, as expected.

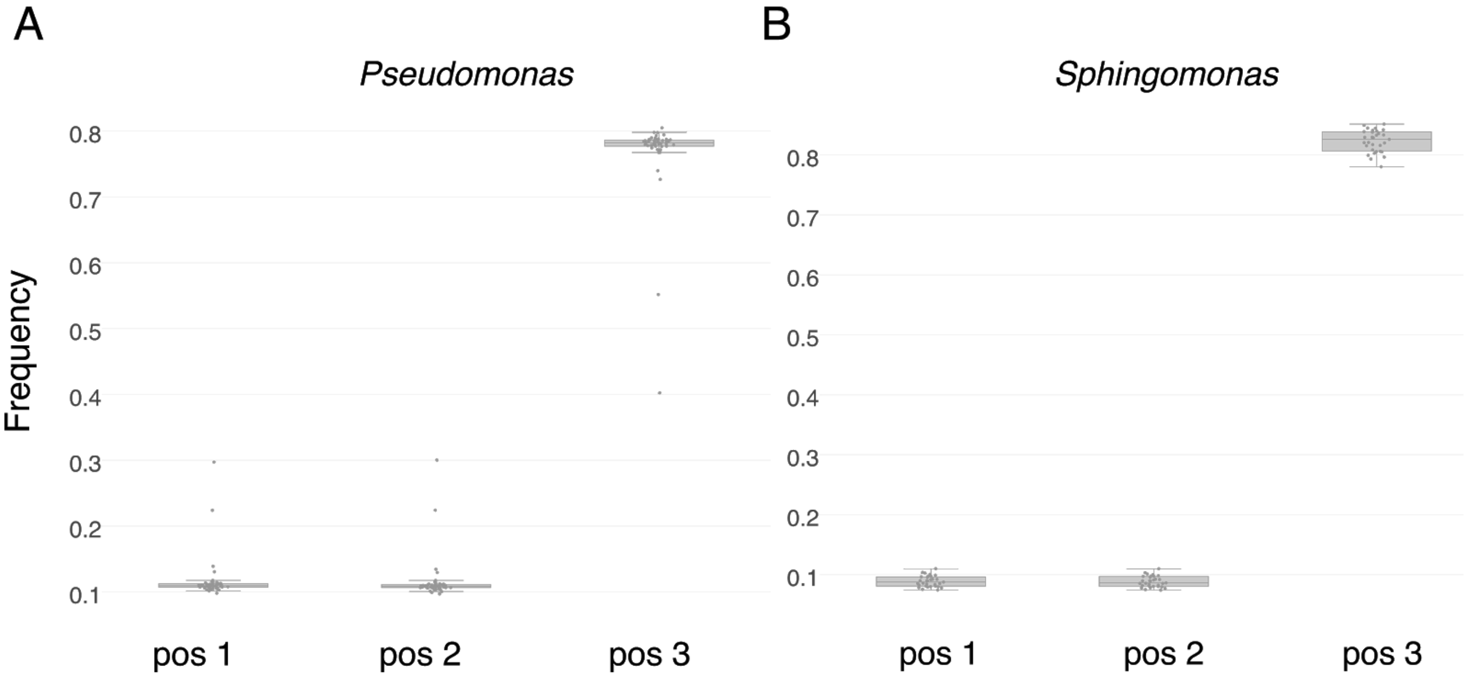

Figure 21 Within-codon position of SNPs

A - *Pseudomonas*, and **B -** *Sphingomonas*.

Distinct Colonization Patterns of *Pseudomonas* and *Sphingomonas* Revealed by Metagenomic Analysis

Inferences about strain dynamics can be made by comparing the frequency at which a nonreference base is present at each segregating site against the number of times that frequency is observed. For example, in a simulation of a three-way mixture of equally abundant strains, a large peak at alternate base frequency of 1/3 is clearly visible. An additional smaller peak at frequency of 2/3 is also seen. This corresponds to positions in the core genomes of these strains where two strains share the same alternate base at the segregating site. It is worth mentioning that such a result would only be observed in simple strain mixtures with enough coverage and number of SNPs. Metagenomic data will probably be much noisier, as samples are likely to comprise more complex mixtures. In addition, low sequencing depths will decrease the power to compute and separate frequency estimates.

In the case of *Pseudomonas,* nonreference base frequencies almost always had a bimodal distribution, with the alternate bases being either at low frequency, ~10%, or at the other end of the spectrum, ~90% (Figure 22), with only a few samples having major peaks of alternate bases at intermediate frequencies.

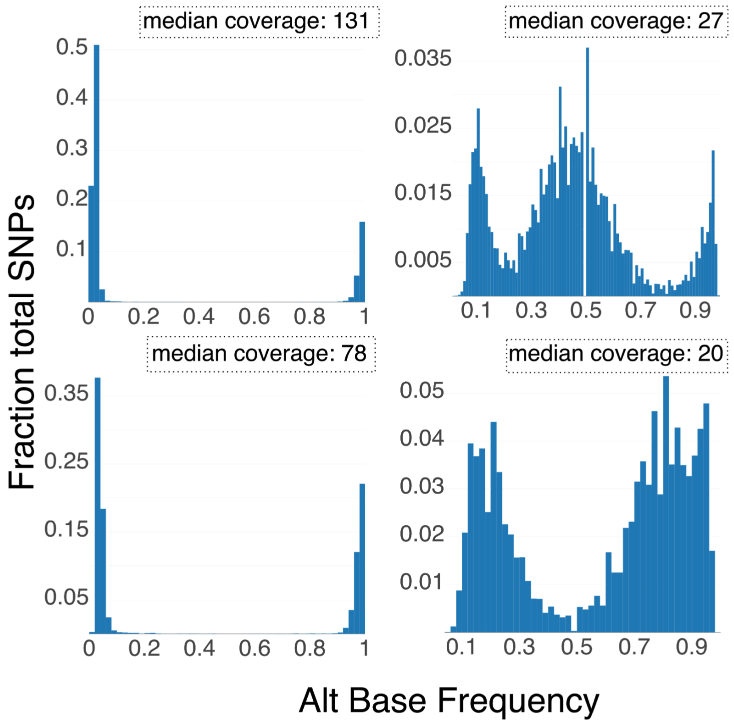

Figure 22. Alternate base frequency distribution in *Pseudomonas* core genomes

The distribution of alternate, nonreference base frequencies in four individual samples. Frequency is defined as the number of reads with alternate base calls divided by the total coverage at that position. Median coverage refers to core genome coverage.

This pattern suggests a scenario in which the vast majority of *Pseudomonas* sequencing reads in each sample is originating from a single strain or from very closely related strains. This is not surprising for a bacterium such as *Pseudomonas*, which is known to be a common pathogen of plants, with pathogens in turn known to often expand to high loads in a host (Sarkar & Guttman, 2004). However, this observation was not only made in samples with high core genome coverage, but also in samples with comparatively low coverage, albeit not as obvious as in high abundance samples (Figure 22, bottom right). An example of a more complex nonreference base frequency pattern is shown in Figure 22, top right, where two peaks at ~90% and ~10% plus an additional peak at ~45% are seen, a distribution best explained by the coexistence of two distinct lineages in near-equal proportions. The ~90% peak corresponds to sites shared by the two strains, while the 45% peak corresponds to SNPs differentiating the two strains.

In the case of *Sphingomonas,* the patterns are very different (Figure 23). None of the samples that passes any of the filtering thresholds had a distribution that can be explained by a simple colonization pattern as seen with *Pseudomonas.* Instead, a more even distribution of alternate base frequencies was observed for all samples, with a tendency to a skew toward low frequencies. There are two main factors that distinguish the *Sphingomonas* and *Pseudomonas* data. First, due to the overall lower relative abundance of *Sphingomonas*, core genome coverage was also lower, decreasing the power to detect discrete peaks in alternate allele frequency distribution, although this is somewhat mitigated by the larger reference core genome for *Sphingomonas*. Effectively, this means there is a reduction in power to distinguish variation in the relative abundance of strains, but more confidence in the overall distribution of abundances. A possible interpretation of the patterns observed in this genus is to think of *Sphingomonas* as being present as a mixture of genetically diverse lineages, consistent with its comensal nature, where a single strain rarely dominates and outcompetes all other microbes (H. Kim et al., 1998).

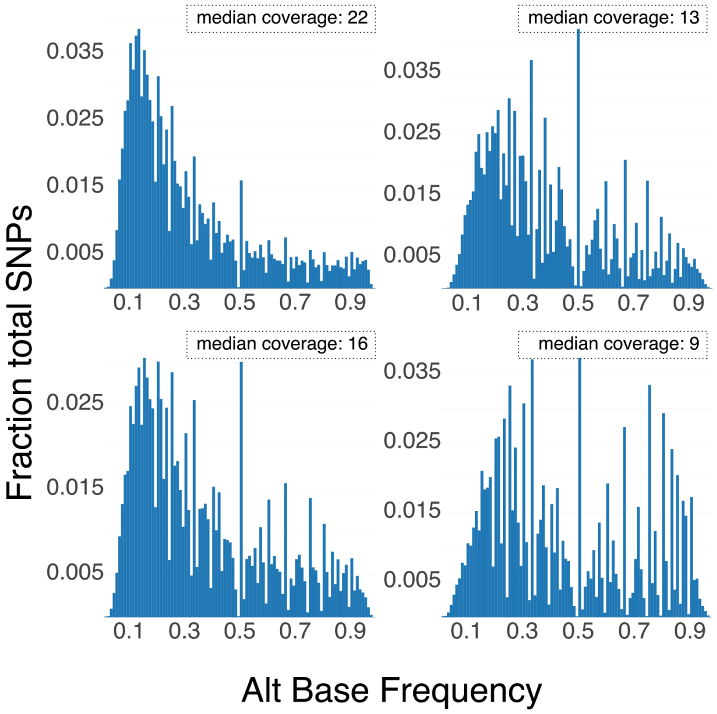

Figure 23. Alternate base frequency distribution in *Sphingomonas* core genomes

The distribution of alternate, nonreference base frequencies in four individual samples. Frequency is defined as the number of reads with alternate base calls divided by the total coverage at that position. Median coverage refers to core genome coverage.

Across-host Genetic Diversity

Variation in the core genome informs on how strains of the genus of interest relate to each other and to known reference strains in this genus. SNP information from the previous step was therefore used to assess genetic diversity across samples for both *Pseudomonas* and *Sphingomonas* strains and to compare the observed diversity with that present in reference genomes, which were downloaded from RefSeq. In the case of *Pseudomonas*, 7 genomes were used, including *P. stutzeri, P. syringae, P. viridiflava, P. stutzeri, P. aeruginosa, and P. putida* as well as the local OTU5 reference. For *Sphingomonas,* a more diverse panel of 146 genomes was selected, including RefSeq reference genomes for *S. meloni, S. adhaesiva,* and *S. koreensis,* plus all genomes reported in (Bai et al., 2015) of *Sphingomonas* spp. Isolated from *A. thaliana* leaves. A wider set of *Sphingomonas* references was used because its taxonomic diversity among strains isolated from *A. thaliana* has been less explored. In *Pseudomonas*, several studies (Karasov et al., 2018; Katagiri, Thilmony, & He, 2002) have identified species in the *syringae*/*viridiflava* complex as the most common *A. thaliana* colonizers. By including more reference sequences, more diversity is incorporated in the reference panel and a greater chance of associating any of the metagenomic strains with a known species.

Before analyzing genetic diversity, a set of genes common to all samples and reference genomes must be selected in order to eliminate potential biases introduced by including too many genes unique to just a handful of samples. That is, samples that do not have a certain gene are treated as having missing information. This will introduce biases for example, such missing information is treated as being reference. BLAST (Altschul et al., 1990) was used to align all downloaded genomes against the previously defined core genomes of *Pseudomonas* and *Sphingomonas*. Core genes with unique matches, an e-value < 0.005 and being present in 100% of the reference sequences were used. This resulted in 12 genes for *Pseudomonas*, and 185 for *Sphingomonas*. The stark difference in the number of genes is due to the larger diversity in the selected *Pseudomonas* references, even though only seven genomes were used. In addition, the core genome used for *Pseudomonas* was much smaller than that of *Sphingomonas*.

Next, an allele count matrix was constructed as input for principal component analysis. This matrix took the form *n*x*m* (*n* columns by *m* rows), where columns corresponded to positions in the reference core genome and rows to each genome analyzed, in this case one for each metagenomic sample and reference genome. This matrix was then filled with the number of alternate alleles present at each position per sample. In the case of other reference genomes, this value was always either 0 or 1. Finally, this matrix was pruned to remove all invariant sites, usually where all samples have the value 0, corresponding to non segregating positions relative to the reference genome. With such a matrix, the number of alleles per site determined similarity between samples, and not the exact state of the allele(s). In other words, given two samples and the segregating position n_1, if in one sample the nonreference base A was found and in the other the nonreference base C, both samples would have had the value 1 at that position, indicating the presence of only one nonreference allele in each sample, even though the alternate allele was different between the two samples.

In the case of *Pseudomonas*, this yielded a total of 6,512 SNPs, and the first two components of PCA showed general clustering of the metagenomic samples (Figure 24).

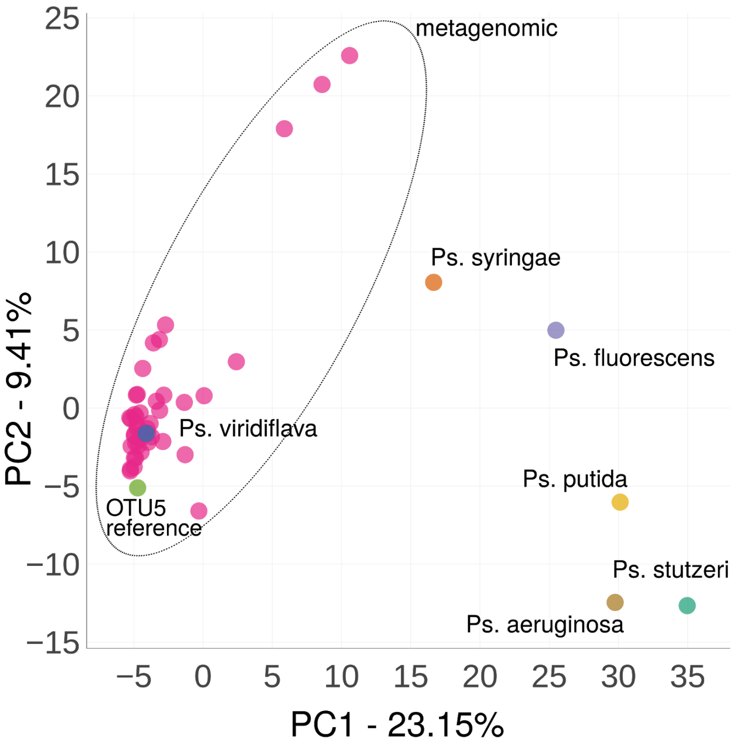

Figure 24. PCA of *Pseudomonas* allele count matrix

Principal component analysis of the allele count matrix from 6,512 SNPs in the *Pseudomonas* core genome. Metagenomes in pink, reference genomes in other colors. OTU5 was the reference genome for alignment.

As expected, of the reference genomes used, metagenomic samples were most closely associated with *Pseudomonas viridiflava* and OTU5, the locally dominant strain (Karasov et al., 2018). *Ps. syringae* was the next most closely associated reference genome, indicating that it is likely that some of the metagenomes are composed of a mixture of *viridiflava* and *syringae* strains.

The same type of analysis for *Sphingomonas* strains yielded a total of 153,283 SNPs, and PCA provided a much more complex picture (Figure 25). First, reference genomes from the same species formed distinct clusters, often including reference genomes from *A. thaliana* leaves. Surprisingly, most metagenomic samples formed a distinct cluster close to some of the reference genomes. This is might be due to their mixed nature, where computed genotypes are derived from alleles in different strains, potentially making metagenomes appear more similar to each other than expected.

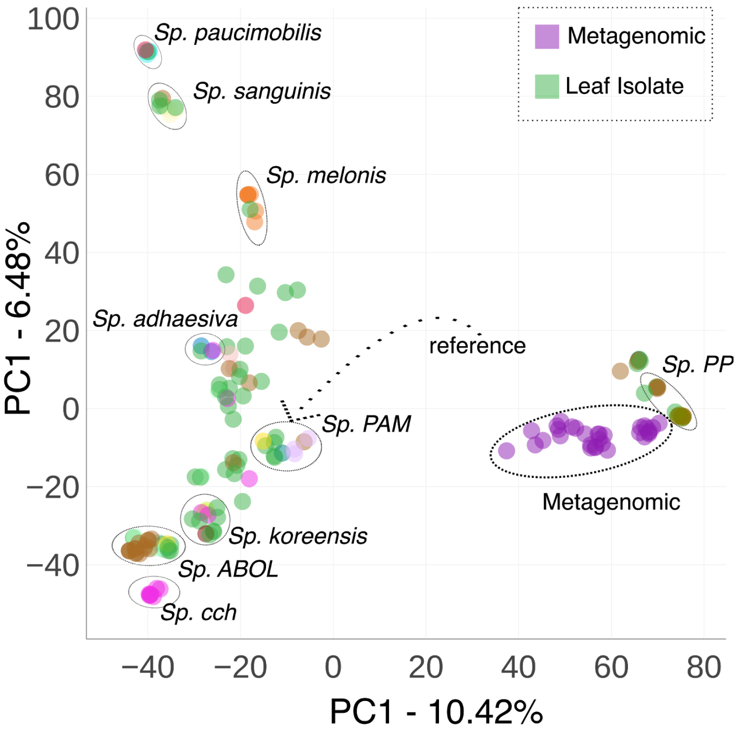

Figure 25. PCA of *Sphingomonas* allele count matrix

Principal component analysis of the allele count matrix from 153,283 SNPs in the *Sphingomonas* core genome. Reference genome for alignment indicated by dotted arrow.

Chapter 3: Discussion

Whole Genome Shotgun Sequencing can Describe the Taxonomic Composition of Wild *Arabidopsis thaliana* Leaf Microbiomes

Having an integral understanding of the leaf microbiome regarding composition, colonization patterns, and its dynamics in a natural setting is of critical importance to understand many phenomena central to plant biology (Hassani, Durán, & Hacquard, 2018). Accurate knowledge of host - microbe interactions will also be a prerequisite for the development of methodologies that can address a number of key issues pertaining to plants in wild and agricultural settings. Some of these include conservation (Bahrndorff, Alemu, Alemneh, & Lund Nielsen, 2016), since plant microbiota can help or interfere with adaptation to new environments, which is especially in a changing global climate. Agricultural improvement is another area where many efforts are being directed toward understanding how plant-microbe associations can be exploited in order to increase crop productivity (Chaparro et al., 2012)**.** And with direct impact on human health, the focus on foodborne illness (Nyachuba, 2010) has led to questions how enteric bacteria can establish themselves as plant colonizers (Brandl, 2006).

For a long time, a main approach to address how plants interact with microorganisms was to isolate and culture of specific microbes, as in many other fields of microbiology (Dickinson, Austin, & Goodfellow, 1975). Most efforts have gone into dissecting the composition and interactions of root associated microbial communities (Berendsen et al., 2012), since the soil in which roots grow is a particularly rich source of microorganisms. As DNA sequencing technologies continue to improve, getting a grasp on the larger community of microbes has become easier to achieve. With individual loci such as 16S rDNA or by multilocus sequence typing (MLST), a large number of taxa can be studied simultaneously in a complex sample (Hayashi, Sakamoto, & Benno, 2002). Moreover, decreasing sequencing costs enable the study of entire microbial communities in many more samples, which provides more statistical power to detect changes in community composition (Rodriguez-R & Konstantinidis, 2014), or to detect rare microbes, which can act as keystone species (Knight et al., 2012). Additionally all three domains of life can be studied simultaneously (Somboonna, Assawamakin, Wilantho, Tangphatsornruang, & Tongsima, 2012). In the case of whole genome shotgun sequencing, genetic diversity can be explored in an extremely granular fashion, in the best cases by comparing strains of single species (Garud et al., 2019).

In the first section of this work I developed a pipeline for estimating taxonomic composition of the leaf community of *A. thaliana*, the phyllosphere, from wild populations in southwest germany. I used whole genome shotgun sequencing data and microbial genome reference databases to bin sequencing reads according to their taxa of origin. Additionally, a method was introduced to estimate microbial load by taking advantage of host derived reads. Specifically, reads attributed to the host were used as proxy of total plant material, which was then used to scale microbial abundances. The relative abundances thus inferred correlated highly with estimates from 16S rDNA amplicon sequencing. Finally, I highlighted some of the limitations of this type of data for generating assembled genomes from metagenomes.

The results presented in this work clearly demonstrate that microbial taxonomic profiles can be derived from shotgun data. These profiles are not restricted to bacteria, as the presence of oomycetes and fungi, known plant colonizers (Kemen, 2014; Porras-Alfaro & Bayman, 2011), could be robustly detected, despite their low abundances (Figure 1, 2), which speaks to the usefulness of the methods developed here for the unbiased study of microbial communities in plants. The plant samples investigated here confirm that bacteria constitute the majority of the typical leaf microbiota (Redford, Bowers, Knight, Linhart, & Fierer, 2010). Alpha- and betaproteobacteria were the most prevalent bacterial classes in the sampled plants, mainly due to the presence of *Pseudomonas and Sphingomonas,* two well known *A. thaliana* colonizers (Innerebner et al., 2011; Katagiri et al., 2002), but other, less abundant taxa such as Hymenobacteriaceae and Flavobacteriaceae were also consistently detected (Figure 3). The whole genome shotgun data were also used to assess variation in bacterial abundances, with plant chromosomal DNA as an internal standard to derive microbial load estimates that can be compared across samples. The method is in its approach reminiscent of spiking in a known amount of DNA in order to control for differences in load (Stämmler et al., 2016), and indeed revealed substantial variability of total microbial (figure 3). How much of this reflects stochasticity in microbial colonization due to the open environment of the leaf, and how much reflects the fact that some apparently healthy plants can support more microbes than others remains to be investigated (Hirano & Upper, 2000).

The replicability of this pipeline was demonstrated with plants that were either split and processed individually, or repeatedly treated with slight modifications in the DNA shearing step, seen not only in terms of Bray-Curtis dissimilarity, but also in terms of read counts of individual taxa (Figure 4). Information on the plant of origin can be recovered from microbial community distances computed with either the original data or by subsampling the number of reads in each sampling and repeating the analysis, providing guidance for efficient application of this method in future (Figure 5). This is in the range of 200,000 Illumina short reads, in agreement with suggestions for other systems (Hillmann et al., 2018). In terms of plant microbiomes, this is important knowledge to have because of the difficulty of obtaining deeply sequenced metagenomes due to the overrepresentation of host DNA content. Thus, shallow sequencing of a relatively limited set of plants should be able to reveal large-scale properties of the microbiome, from which informed decisions can be made in order to sequence targeted samples at a higher depth. A similar approach is followed for ancient DNA samples, where content of endogenous DNA in ancient specimens is often first assayed by shallow sequencing of many samples, before choosing those that will be most informative for deep sequencing (Hagelberg, Hofreiter, & Keyser, 2015).

Differences in community composition were analyzed by deriving microbial distances based on double square root transformed taxonomic counts, which reduces the influence of abundant taxa, as the square root operation has a greater effect on larger numbers. The end result is a reduction in the long right tail of the distribution of taxonomic counts, in other words, it shift this distribution to become more normal (Xie, Goh, & Tang, 2000). Choosing the right dissimilarity measure is a contentious topic in microbial ecology, specially because different similarity indexes are useful for different situations and data types, with each having different advantages and shortcomings (Hill, Walsh, Harris, & Moffett, 2003). For example, Bray-Curtis, one of the most used metrics, provides a compromise between weighing presence/absence of taxa and overall abundance differences between samples (Ricotta & Podani, 2017). Because of this, Bray-Curtis is well suited for datasets where not only differential abundance is important, but also the number of private taxa. In the case of the leaf microbiomes investigated here, taxa exclusive to only a handful of samples were rare and these were usually low abundance microbes, which contribute relatively little to overall species richness. An alternative to Bray-Curtis dissimilarity is Euclidean distance, which was chosen here to measure differences between samples because of its simplicity, mathematical interpretability, and type of data used here (Silverman et al., 2017). Between-sample distances computed in this manner had a near-normal distribution (Figure 6) indicating that samples did not fall into distinct clusters and an overall lack of discrete community types. This has also been observed in the microbiome within different body parts in healthy humans (Koren et al., 2013). In such case, the lack of distinct taxonomic diversity, based on different methods has been interpreted as evidence for the lack of different enterotypes. Microbial communities well adapted to different body parts such as the gut.

Host microbiome variation was assessed by computing principal component analysis on transformed count data and explore to what extent individual microbiomes are similar or different from each other and how these similarities correlate with variables such as site of origin or sampling season (Figure 7). It did not appear that any environmental variable strongly affected microbiome community structure nor did host genotype, as host genetic clusters (Figure 9, 10) were not reflected in microbial distances. Nevertheless, a combination of site and genotype likely plays a role in how microbial communities assemble in the phyllosphere (Bodenhausen et al., 2014; Hacquard, 2016). Unfortunately, population structure of wild *A. thaliana* is highly confounded by location because local stands often feature multiple individuals with identical or closely related genotypes (Bomblies et al., 2010), which makes separating the effects of sampling site and host genetics in our collection difficult, if not impossible.

It also needs to be considered that the nature of the data transformation contributes to the low correlation between site/host genotype and microbiome composition, if there were individual high-abundance taxa that distinguished sites. For example in this case, an abundance of *Pseudomonas* can be seen in samples mainly originating from Eyach. Additionally, due to the open nature of the leaf ecosystem, individual host microbiomes may be affected by transient colonization events due to environmental heterogeneity (Kristin & Miranda, 2013). These may be caused by a number of factors including, differences in soil ph, UV exposure, or water availability, which may all obscure the effects of host genotype. Finally, a major source of community structure was found to be the most abundant taxa in individual samples. When most abundant taxon is used to label samples, principal components one and two divide samples that have either *Pseudomonas* or *Sphingomonas* as top taxon (Figure 8). Projecting the data to components two and three, reveal further clustering of samples based on shared most abundant taxon (Figure 8). There is a body of work suggesting that specific microbes can have a large role in shaping community structure (Banerjee, Schlaeppi, & van der Heijden, 2018; Trosvik & de Muinck, 2015). Additionally, in the specific case of the *A. thaliana* phyllosphere in populations in southwest Germany, it has been observed that hub taxa can have a dramatic influence in shaping the microbiome (Agler et al., 2016). Therefore, a combination of host genetics, site of origin and intermicrobial interactions contribute to community patterns observed in this collection, although stochastic effects cannot be discounted.

Microbe-microbe interactions in multispecies community are central to microbial ecology. For example, a metabolite produced by one microbe may inhibit growth of another microbe, reflecting antagonism between the first two taxa. Now if this second microbe was inhibiting a third species that in turn was promoting growth of the first species, the relative strengths of individual interactions as well as starting inocula, i.e., historical contingency, would determine the static outcome observed in the sort of data analyzed here.

Because a number of such complex interactions may exist in a community such as the leaf phyllosphere, network analysis is an important tool to study microbial communities, where a microbiome is viewed as a collection of nodes (the microbes) connected by edges (microbe-microbe interaction) (Layeghifard, Hwang, & Guttman, 2017). One way to infer interactions between two taxa is by measuring their co-occurrence. For example, linear correlations of abundance between taxa pairs is often used as a metric to build such networks (Friedman & Alm, 2012), although other measurements can also be used, such as mutual information (M. S. Lee, Oh, & Tang, 2014). A key component that greatly influences these co-occurrence metrics is the type of data used as input. For example, when the abundances of microbial taxa are only relative to each other, the real absolute prevalence of each microbe in any given sample can not be determined; this has been shown to be a problem when estimating species diversity (Kemp & Aller, 2004). Additionally taxa correlations measured this way may not reflect species interactions, instead, the increase in the relative abundance of a taxon leads to a decrease of other taxa in the dataset (Aitchison, 1981).

In the data analyzed here, microbe-microbe correlations had noticeably different and often even opposite patterns when comparing compositional data (relative abundance as computed by 16S rDNA sequencing) to load corrected estimates (based on whole-genome shotgun sequencing) (Figure 11,12). Specifically, for the two most abundant microbial families (Pseudomonadaceae and Sphingomonadaceae) across all samples and also per individual leaf microbiome, antagonism between the two would be inferred from the clear anticorrelation observed in relative abundance data (Figure 12). An inhibitory relationship would be consistent with the literature, since this has been shown to occur between certain strains of these families (Innerebner et al., 2011; Vogel, Innerebner, Zingg, Guder, & Vorholt, 2012). However, when load corrected data is incorporated into the analysis, this correlation became instead positive. Such a switch from a potentially negative to a potentially positive interaction was also observed in other taxa pairs, showing that the confounding effect of relative abundance has a community-wide effect. It is important to mention that in load corrected abundance estimates, these correlations were always positive (Figure 11), which could be interpreted as less complex stable relationships among different community members. In such a situation, the *A. thaliana* leaf microbiome can be seen as a community that altogether successfully colonizes the phyllosphere, with the occasional infection of particular pathogens.

A bonus of using whole genome shotgun data is the ability to assemble sequenced reads into longer contigs, in the best case resulting in complete genomes, which can give new insights into community properties (Breitwieser et al., 2017). For example, instead of relying on reference databases, taxonomic assignment can be performed de novo based on entire genes present in the assembled sequences (Alneberg et al., 2014; Boisvert, Raymond, Godzaridis, Laviolette, & Corbeil, 2012). Metagenomic assembly of the phyllosphere microbiome remains a challenging task; in the case of the data used here, with disappointing results in terms assembly size and contig length (Figure 13). A number of reasons are likely to have contributed to this, sequencing depth and genetic diversity being the main ones. High coverage of target genomes is usually required in order to assemble longer contigs (Zerbino & Birney, 2008). Due to the overwhelming amount of host DNA content, it was difficult to obtain high depth for any but the most abundant microbes. Most of the algorithms for metagenome assembly use a data structure called the de Bruijn Graph in which input sequences are cut into relatively short subsequences. Afterwards, a graph of overlapping is constructed and traversing this graph is what generates assembled sequences (Compeau et al., 2011). When a sample contains high sequence diversity, it generates an increasingly complex de Bruijn Graph, which is then more difficult to traverse, resulting in shorter contigs.

As sequencing costs decrease, more and more studies are moving towards using whole genome shotgun sequencing to study microbial communities(Cameron et al., 2016; Figueroa et al., 2018; Kose, Grice, Orsi, Ballal, & Coolen, 2018). Nevertheless, it is still critical to identify to what extent metagenomic and amplicon sequencing inferences are comparable, especially if existing knowledge in microbiome science is to be reinterpreted through the lens of this new technology. Metagenomic sequencing has been discussed in terms of the advantages it provides when compared to 16S rDNA profiling in bacterial communities (Ranjan, Rani, Metwally, McGee, & Perkins, 2016), but whole genome sequencing can also lead to underestimates of taxonomic diversity compared to amplicon based approaches (Shah, Tang, Doak, & Ye, 2010; Tessler et al., 2017). Here, I showed that these two methods agreed well when comparing bacterial families (Figure 14), but less so for fungi, with most of the discrepancies originating from the analysis pipeline rather than the limits of whole genome shotgun sequencing. It must be noted that the use of a protein database greatly decreases the power to accurately estimate the abundance of organisms with a high proportion of non coding sequences in the genome. Clearly, more work needs to be done to better understand how different analysis methods can exacerbate or mitigate biases in inferring relationships between microbes from either amplicon or metagenomic sequencing.

Strain Level Analysis of *Pseudomonas* and *Sphingomonas* in High Depth Leaf Metagenomes

While having information on the large-scale taxonomic composition of a bacterial community (e.g., family or genus) can reveal important aspects of a host’s microbiota, it is also clear that we are often missing out on understanding the relationship between closely related organisms. For example, two strains that belong to the same genus and are subsumed under the same taxonomic label, even though they have different impacts on the host (Freschi et al., 2019) (Heintz-Buschart & Wilmes, 2018). In animal systems, there are multiple examples of variation between members of the same species having revealed important dynamics between microbes and their host (Garud et al., 2019; Mizrahi & Jami, 2018). Nevertheless, the accurate naming and distinction between strains based on genetic variation among microbial groups remains subject to a great amount of debate. Confusion regarding strain classification remains rampant in modern microbiology (Baltrus, 2016). While taxonomists have traditionally maintained order in the nomenclature of microbial species, the advent of new sequencing technologies has produced an avalanche of new strains at an unprecedented pace. In addition, advancements in metagenomic sequencing and assembly will certainly only contribute to this situation.

Learning the genetic diversity of individual taxa from metagenomic data represents a great opportunity to study such strain level variation at potentially larger scales. In the case of the leaf microbiota, many aspects of leaf colonization can be addressed by exploring the genetic makeup of individual colonizers. For example, if in a group of plants, the microbiome is formed mainly by a single taxon, one may assume a single colonization event followed by clonal expansion such as in a pathogenic infection (Straub et al., 2018). On the other hand, nucleotide diversity in sequences of a specific taxon will point to near-simultaneous colonization by multiple strains, with all of them being similarly competitive. Ultimately, revealing differences between communities in turn can lead to different interpretations of many aspects of a microbial community (Metwaly & Haller, 2019).

Inferring strains from metagenomes comes with the added advantage of not having to perform colony isolation, which not only translates into less laboratory work, but also reduces the dangers of biases that come from selective culturing (Davis, Joseph, & Janssen, 2005). It comes, however at the expense of having to invest more resources in sequencing, but as costs continue to fall (“DNA Sequencing Costs: Data,” n.d.), obtaining sufficient depth for bacterial genotyping becomes a viable option. All of this further justifies the use of metagenomic sequencing in phyllosphere microbial communities as a method for deriving strain genotypes, as has been used in human microbiomes (Garud et al., 2019).

Whole genome shotgun sequencing of wild *A. thaliana* leaf microbiomes in southwest Germany revealed *Pseudomonas* and *Sphingomonas* to be the most prevalent genera in the phyllosphere microbiota. Both taxa achieve high relative abundance across all samples (Figure 3, Figure 16), which in turn allowed for strain level population analysis. *Pseudomonas* strains are one of the most important plant disease causing agents, with dozens of pathogenic varieties described (Bull et al., 2010), infecting a wide range of species including tomato (Preston, 2000), kiwifruit (Straub et al., 2018), and *A. thaliana* (Karasov et al., 2014). *Sphingomonas* strains are also an important component of the plant microbiome (Chen et al., 2018; Compant, Samad, Faist, & Sessitsch, 2019; Kecskeméti, Berkelmann-Löhnertz, & Reineke, 2016; Purahong et al., 2018), with some strains providing protection against pathogens (Adhikari, Joseph, Yang, Phillips, & Nelson, 2001; Enya et al., 2007; Innerebner et al., 2011).

There are different strategies for the strain-level analysis of microbial communities (Schürch, Arredondo-Alonso, Willems, & Goering, 2018). A challenge for this type of analyses is the establishment of a reference data sets against which metagenomic reads could be mapped. Compared to many eukaryotes, microbial genomes are extremely plastic. This is especially true in bacteria where horizontal gene transfer is rampant (Snel, Bork, & Huynen, 2002). In addition, although mutation rates in most bacteria are very low, this is not true for all species, plus the large population sizes allow nevertheless for rapid changes in the genetic makeup of bacterial strains (Denamur & Matic, 2006), with even closely related strains often differing substantially in gene content (Rouli, Merhej, Fournier, & Raoult, 2015). Pertinent to this, progress has been made with detection of copy number variation in gut microbiomes (Greenblum, Carr, & Borenstein, 2015). In the case of *Pseudomonas*, a significant amount of gene content variability has been found, with only the most essential housekeeping genes being common to all known species (Hesse et al., 2018).

Sphingomonas on the other hand has been less well studied, but great functional diversity has been observed in several species of this genus (M. K. Kim et al., 2007; Miyauchi, Adachi, Nagata, & Takagi, 1999; Nagata, Miyauchi, & Takagi, 1999). A common method to circumvent this challenge is by using core genomes as a reference with which different strains can be compared. A core genome represents a set of genes common to all or most organisms in question. While this can drastically decrease the number of genes that can be used to detect genetic variants, it makes it more likely to have a reference sequence that can be found other isolates of interest or in this case within metagenomically obtained sequences. In this study, I used genes present in 100% of individually isolated bacterial colonies. For *Pseudomonas* this was previously computed in (Karasov et al., 2018) while I obtained *Sphingomonas* colonies whose assembled genomes were used to compute the core genome.

One of the main limitations of strain typing from metagenomic data is the high sequencing depth required to achieve informative coverage of the core genomes (Scholz et al., 2012). In this dataset, only a few samples had sufficient depth to provide 5X coverage at each position of the core genome (Figure 18). Nevertheless, coverage distribution of these samples demonstrated no significant interference from sequencing errors or reads originating from other taxa (Figure 19). This is particularly important as there is the risk of confounding genetic variation with sequences from other microbes that happen to have some genes with high nucleotide similarity (Townsend, Bøhn, & Nielsen, 2012). The preference for third codon mutations (Figure 22) was consistent with the expectation of faster rate of evolution at such sites (Felsenstein, 1978).

A key challenge of interpreting strain level variation in microbial mixtures is the ability to distinguish from which strain a genetic variant originates. This is akin to phasing of genetic variants in heterozygous genomes (Browning & Browning, 2011). In the case of microbial genetics, each strain’s chromosome can be viewed as a homolog and a metagenomic mixture of strains would be analogous to an extremely polyploid and heterozygous individual. Furthermore, not only is the number of strains not known, but they also occur in unknown proportions. Coming back to the homologous chromosome analogy, a challenge related to accurately describing a strain mixture corresponds to inferring ploidy of an individual from proportions of alternative bases in reads covering heterozygous positions (Weiß et al., 2018). In microbial mixtures, the distribution of proportions of alternate base calls can be similarly exploited to infer the composition of such mixtures (Figure 20).

Pseudomonas showed a situation where alternate bases were present at either low frequency of below ~10% or near unity (Figure 23). In *Sphingomonas* these patterns of alternate base frequency were very different. With no clear signal of any frequency dominating the distribution of alternative bases. In across-sample comparisons, analyses were restricted to sequences found in all metagenomes and reference genomes. This was done mainly to circumvent having to deal with missing data, which can produce erroneous results if not controlled for. In conventional population genetics, this is overcome by imputing missing genotypes (Marchini & Howie, 2010). In the metagenome data analyzed here, no imputation was attempted, because too little is known about linkage among variants. The nature of genetic diversity in strain mixtures means multiple alleles will typically be present within a sample (Smith, Smith, & O’Rourke, 1993) and considering only biallelic positions would be a great oversimplification of diversity. I therefore used an allele count matrix, where the number of alleles per segregating position is used to infer population structure (Novembre & Stephens, 2008; Reich, Price, & Patterson, 2008), without taking the specific nature of variants into account. A possible caveat is that in highly diverse organisms, identical changes could occur at the same position (Duchêne et al., 2016). At the opposite end of the spectrum are closely related strains that may appear identical if not a sufficient fraction of the core genome is used to identify variants (Casali et al., 2016).

The different patterns in alternative base frequency observed between *Pseudomonas* and *Sphingomonas* can be explained by two models of leaf colonization. For example, in the case of

Pseudomonas strains, alternative bases being present only at either very low or high frequency can be a consequence of the presence of a single strain at high abundance, with one or more genetically distinct lineages at much lower frequency. Such a model would be consistent with *Pseudomonas* strains acting as a pathogen, and a single successful colonization event leading to one strain largely taking over the phyllosphere, with the low-frequency strain being outcompeted commensals. One could also think of these patterns as resulting from continuous colonization by commensals until a well adapted strain is able to outcompete all other strains (Karasov et al., 2018). The specific distribution of alternate base frequencies could also reflect the presence of many strains co-occurring as commensals followed by the emergence of a pathogenic variety that take over the microbial community in fashion similar to antibiotic resistance (Levy & Marshall, 2004). This scenario is, however, less likely, as *A. thaliana* is a relatively short-lived species.

In many cases, a bimodal pattern of alternate base frequency could be observed. This was independent of core genome coverage and relative abundance of the taxon, which speaks to the sensitivity of this method. There were also cases where there was a distinct third peak of alternate bases at intermediate frequency (Figure 23), consistent with two lineages being much more common that all other strains. Such a situation indicates that leaf colonization by *Pseudomonas* is not necessarily a zero sum game and that multiple lineages may co-occur (Karasov et al., 2018; Kniskern, Barrett, & Bergelson, 2011).

On the other hand, models that incorporate a simple mixture of two or three strains cannot explain the apparent distribution of alternate base frequencies seen in *Sphingomonas* (figure 24). Low coverage of the core genome might contribute to this unclear signal, but this should be at least partially mitigated by the many more positions with alternate bases. *Sphingomonas* strains in the *A. thaliana* phyllospheres seem to be often very diverse (Bai et al., 2015) and the observed patterns were likely to arise primarily from complex mixtures of relatively distantly related isolates. That often two or three, or even four, alternate bases are found at the same position was also consistent with greater genetic distance among *Sphingomonas* strains (Figure 21). Based on the lower coverage of the *Sphingomonas* core genome, sequencing errors were more likely to confound these frequency distributions, especially in terms of alternate bases at low abundance. Nevertheless, as with *Pseudomonas*, third codon positions constituted the vast majority of alternate base calls (Figure 22). Altogether, the results demonstrate how metagenomic sequencing can be a useful tool for exploring taxonomic genetic diversity among taxa with remarkably different lifestyles in the phyllosphere (HU, Jie,HE Xiaohong,LI Daping & LIU Qiang, 2007; Morán et al., 2018).

As with alternate allele frequency distribution, relatedness between samples and reference genomes based on principal component analysis revealed different pictures for *Pseudomonas* and *Sphingomonas*. In terms of relatedness between metagenomically derived strains and known reference genomes within the *Pseudomonas* genus, there was overall agreement with previous findings. The expectation being for metagenomic samples to cluster with reference sequences of representative genomes of the syringae/viridiflava complex (Jakob et al., 2002; Karasov et al., 2018; Kniskern et al., 2011), which was indeed largely the case (Figure 25). The results, together with the patterns of alternate base frequency (Figure 23), support that in most samples a single *Pseudomonas* strain dominates*.* An important next step will be to also study presence/absence patterns of entire genes, a well know feature of *Pseudomonas* genomes (Silby, Winstanley, Godfrey, Levy, & Jackson, 2011).A more complex situation was observed with *Sphingomonas*. As with *Pseudomonas*, metagenomes clustered with reference sequences, but with multiple reference strains in such clusters. The most abundant reference taxa designated as “PP” from (Kyrpides et al., 2014) a cryptic name that speaks to the problems of strain nomenclature in microbial taxonomy. The strains represented by these reference genomes had been found in association with plant leaves and soil. It is unclear whether this also represents a case of relatedness between strains*.* Based on the known diversity of this genus in leaf microbiomes (Lebeis, 2014; Lebeis, Rott, Dangl, & Schulze-Lefert, 2012) as well as the observed patterns of alternate base frequency (Figure 24), it appears that multiple species are associated with each metagenomic sample. The presence of multiple species on single *A. thaliana* individuals was also observed when culturing *Sphingomonas* isolates (Bai et al., 2015). Two main factors likely explain the differences observed between *Pseudomonas* and *Sphingomonas*. As discussed, the input data used to compute relatedness is an allele count matrix that ignores the nature of genetic variants, with similarity between samples being derived from the covariances in this matrix (Patterson, Price, & Reich, 2006; Reich et al., 2008). The observed clustering could reflect the presence of multiple alleles in the first component caused by high covariance within metagenomes only. Alternatively, this pattern could be explained by the mixed nature of *Sphingomonas* genotypes derived from metagenomic data. Such samples, when projected along the first axis, can be interpreted as having shared relatedness between two major groups of this taxa, as has been shown in admixture analysis (Jeong et al., 2014; J. Ma & Amos, 2012). This results represents a scenario expected from the mixture of a diverse set of strains, although it remains unknown how many individual strains there were. Nevertheless, even though the results regarding *Sphingomonas* are less conclusive than the ones for *Pseudomonas*, both reflect the different biological properties of these two genera.

My results also highlight that whole genome shotgun sequencing can be used to detect strain mixtures in phyllospheres, which can be of particular interest for commercially important crops, where strain mixtures may be superior for pest biocontrol compared to single strains (Stockwell, Johnson, Sugar, & Loper, 2011). Moreover, if there are specific biological functions of interest, detecting different genotypes in genes related to that function is an option, in fashion similar to phyloFlash, where 16S rDNA sequences are directly mined from shotgun data. A method that can easily be extended to any gene of interest (Gruber-Vodicka, Seah, & Pruesse, 2019). Finally, strain detection from metagenomic data also represents an important opportunity for environmental microbiome studies (A. C. Howe et al., 2014).

Outlook

I have demonstrated the effective use of metagenome sequencing for studying microbial communities of wild *Arabidopsis thaliana* plants. The study of leaf microbiomes is an active area of research (Compant et al., 2019), addressing questions about community composition (M. Kim et al., 2012), influence of host genetics (Wagner et al., 2016), long term evolutionary associations (Karasov et al., 2018), microbe-microbe interactions and community dynamics (Agler et al., 2016), and many others (Vorholt, 2012; Wallace, Kremling, Kovar, & Buckler, 2018). Additionally, microbial communities can affect how other organisms interact with plants (Ramírez-Puebla et al., 2013). I have used metagenomic sequencing to build not only informative taxonomic profiles from wild *A. thaliana* plants, but to also analyze within taxon variation. At the sequencing depth employed here, metagenome assembly worked only poorly. If sequencing costs continue to drop, along with improvements of algorithm development (A. Howe & Chain, 2015), gene centric analyses might come within reach. It might be sensible to first search for functional microbial signatures in plants that produce specialized structures as the venus flytrap (Sickel, Van de Weyer, Bemm, Schultz, & Keller, 2019). genomes. For example, in the rumen microbiota of bovines, clear signatures of carbohydrate metabolism have been discovered through metagenome assembly (Brulc et al., 2009).

Estimates put the global leaf surface area at 6.4 × 10^8 km^2 (Lindow & Brandl, 2003), which is one third larger than the surface area of the entire planet including its oceans (Pidwirny, 2010). Most of the taxonomic diversity of this vast habitat remains, however, to be explored. Consequently, increasing the number of plant species subject to the type of analyses performed in this thesis together with sampling plants exposed to a greater range of environmental factors represents a critical area to further our understanding of interactions between plant hosts, their microbes and their surroundings. Investigating plant microbiomes at higher resolution is fundamental if the properties of less abundant taxa are to be understood, since microbes at low relative abundance may constitute important indicators of plant health or be active at physiological relevant levels, as has been observed in other microbial communities (Campbell, Yu, Heidelberg, & Kirchman, 2011). To cope with the increasing amount of data, improved algorithms are needed, especially since these will soon contain more and more long-read data. Long reads can be binned directly (Huson et al., 2018), they can be used for strain level variation analysis in analogy with haplotype phasing (Huddleston et al., 2017), and they can dramatically improve the capacity to obtain high quality genomes from metagenomes (Frank et al., 2016; Olson et al., 2017). Ultrafast sequence comparison (Buchfink et al., 2015; H. Li, 2018), alignment free pipelines (D. Kim et al., 2016; Ondov et al., 2016; Wood & Salzberg, 2014), and machine learning frameworks that can incorporate different types of data (Fiannaca et al., 2018; Rojas-Carulla et al., 2019) are also all areas of active algorithm development.

Finally, while bacteria are generally the most common microorganisms associated with plants, eukaryotic microbes and archaea constitute an important fraction of the microbial community and can be significant drivers of the microbiota (Agler et al., 2016). Increasing the repertoire of reference sequences for plant associated microbes from these taxa will be necessary to understand exactly how these two groups impact the overall microbial community and its consequences for the plant host. Finally, work with tomato plants has shown that transplantation of viral extracts from field grown plants can affect bacterial colonization of plants in the greenhouse (Morella, Gomez, Wang, Leung, & Koskella, 2018), yet we still know little about the diversity and functional impacts of natural plant-associated viromes.

References

Adhikari, T. B., Joseph, C. M., Yang, G., Phillips, D. A., & Nelson, L. M. (2001). Evaluation of bacteria isolated from rice for plant growth promotion and biological control of seedling disease of rice. *Canadian Journal of Microbiology*, *47*(10), 916–924.

Adrio, J.-L., & Demain, A. L. (2010). Recombinant organisms for production of industrial products. *Bioengineered Bugs*, *1*(2), 116–131.

Agler, M. T., Ruhe, J., Kroll, S., Morhenn, C., Kim, S.-T., Weigel, D., & Kemen, E. M. (2016). Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation. *PLoS Biology*, *14*(1), e1002352.

Aitchison, J. (1981). A new approach to null correlations of proportions. *Journal of the International Association for Mathematical Geology*, *13*(2), 175–189.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., … Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, *11*(11), 1144–1146.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410.

Anantharaman, K., Hausmann, B., Jungbluth, S. P., Kantor, R. S., Lavy, A., Warren, L. A., … Banfield, J. F. (2018). Expanded diversity of microbial groups that shape the dissimilatory sulfur cycle. *The ISME Journal*, *12*(7), 1715–1728.

Ansorge, R., Romano, S., Sayavedra, L., Kupczok, A., Tegetmeyer, H. E., Dubilier, N., & Petersen, J. (2019). *Diversity matters: Deep-sea mussels harbor multiple symbiont strains*. *bioRxiv*. https://doi.org/10.1101/531459

Antunes, A., Ngugi, D. K., & Stingl, U. (2011). Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environmental Microbiology Reports*, *3*(4), 416–433.

Appolinario, L. R., Tschoeke, D., Paixão, R. V. S., Venas, T., Calegario, G., Leomil, L., … Thompson, F. L. (2019). Metagenomics sheds light on the metabolic repertoire of oil-biodegrading microbes of the South Atlantic Ocean. *Environmental Pollution* . https://doi.org/10.1016/j.envpol.2019.03.007

Bahrndorff, S., Alemu, T., Alemneh, T., & Lund Nielsen, J. (2016). The Microbiome of Animals: Implications for Conservation Biology. *International Journal of Genomics and Proteomics*, *2016*, 5304028.

Bai, Y., Müller, D. B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., … Schulze-Lefert, P. (2015). Functional overlap of the Arabidopsis leaf and root microbiota. *Nature*, *528*(7582), 364–369.

Baldwin, B. G., Sanderson, M. J., Porter, J. M., Wojciechowski, M. F., Campbell, C. S., & Donoghue, M. J. (1995). The its Region of Nuclear Ribosomal DNA: A Valuable Source of Evidence on Angiosperm Phylogeny. *Annals of the Missouri Botanical Garden. Missouri Botanical Garden*, *82*(2), 247–277.

Baltrus, D. A. (2016). Divorcing Strain Classification from Species Names. *Trends in Microbiology*, *24*(6), 431–439.

Banerjee, S., Schlaeppi, K., & van der Heijden, M. G. A. (2018). Keystone taxa as drivers of microbiome structure and functioning. *Nature Reviews. Microbiology*, *16*(9), 567–576.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., … Pevzner, P. A. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, *19*(5), 455–477.

Bankevich, A., & Pevzner, P. A. (2018). Joint Analysis of Long and Short Reads Enables Accurate Estimates of Microbiome Complexity. *Cell Systems*, *7*(2), 192–200.e3.

Battista, J. R., Earl, A. M., & Park, M. J. (1999). Why is Deinococcus radiodurans so resistant to ionizing radiation? *Trends in Microbiology*, *7*(9), 362–365.

Baym, M., Kryazhimskiy, S., Lieberman, T. D., Chung, H., Desai, M. M., & Kishony, R. (2015). Inexpensive multiplexed library preparation for megabase-sized genomes. *PloS One*, *10*(5), e0128036.

Beaulieu, J. M., Leitch, I. J., Patel, S., Pendharkar, A., & Knight, C. A. (2008). Genome size is a strong predictor of cell size and stomatal density in angiosperms. *The New Phytologist*, *179*(4), 975–986.

Bender, M. A., & Farach-Colton, M. (2000). The LCA Problem Revisited. In *LATIN 2000: Theoretical*

Informatics (pp. 88–94). Springer Berlin Heidelberg.

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

Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O., & Piddock, L. J. V. (2015). Molecular mechanisms of antibiotic resistance. *Nature Reviews. Microbiology*, *13*(1), 42–51.

Bodenhausen, N., Bortfeld-Miller, M., Ackermann, M., & Vorholt, J. A. (2014). A synthetic community approach reveals plant genotypes affecting the phyllosphere microbiota. *PLoS Genetics*, *10*(4), e1004283.

Boisvert, S., Raymond, F., Godzaridis, E., Laviolette, F., & Corbeil, J. (2012). Ray Meta: scalable de novo metagenome assembly and profiling. *Genome Biology*, *13*(12), R122.

Bomblies, K., Yant, L., Laitinen, R. A., Kim, S.-T., Hollister, J. D., Warthmann, N., … Weigel, D. (2010). Local-scale patterns of genetic variability, outcrossing, and spatial structure in natural stands of Arabidopsis thaliana. *PLoS Genetics*, *6*(3), e1000890.

Bork, P., Bowler, C., de Vargas, C., Gorsky, G., Karsenti, E., & Wincker, P. (2015). Tara Oceans. Tara Oceans studies plankton at planetary scale. Introduction. *Science*, *348*(6237), 873.

Brandl, M. T. (2006). Fitness of human enteric pathogens on plants and implications for food safety. *Annual Review of Phytopathology*, *44*, 367–392.

Breitwieser, F. P., Lu, J., & Salzberg, S. L. (2017). A review of methods and databases for metagenomic classification and assembly. *Briefings in Bioinformatics*. Retrieved from https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbx120/4210288

Browning, S. R., & Browning, B. L. (2011). Haplotype phasing: existing methods and new developments. *Nature Reviews. Genetics*, *12*(10), 703–714.

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

Brunet, T., & King, N. (2017). The Origin of Animal Multicellularity and Cell Differentiation. *Developmental Cell*, *43*(2), 124–140.

Buchfink, B., Xie, C., & Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, *12*(1), 59–60.

Bulgarelli, D., Garrido-Oter, R., Münch, P. C., Weiman, A., Dröge, J., Pan, Y., … Schulze-Lefert, P. (2015). Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host & Microbe*, *17*(3), 392–403.

Bull, C. T., De Boer, S. H., Denny, T. P., Firrao, G., Saux, M. F.-L., Saddler, G. S., … Takikawa, Y. (2010). COMPREHENSIVE LIST OF NAMES OF PLANT PATHOGENIC BACTERIA, 1980-2007. *Journal of Plant Pathology: An International Journal of the Italian Phytopathological Society*, *92*(3), 551–592.

Burkhardt, S., & Kärkkäinen, J. (2003). Better Filtering with Gapped Õ-Grams. *Fundamenta Informaticae*, *23*, 1001–1018.

Byrd, A. L., Belkaid, Y., & Segre, J. A. (2018). The human skin microbiome. *Nature Reviews. Microbiology*, *16*(3), 143–155.

Cameron, S. J. S., Lewis, K. E., Huws, S. A., Lin, W., Hegarty, M. J., Lewis, P. D., … Pachebat, J. A. (2016). Metagenomic Sequencing of the Chronic Obstructive Pulmonary Disease Upper Bronchial Tract Microbiome Reveals Functional Changes Associated with Disease Severity. *PloS One*, *11*(2), e0149095.

Campanaro, S., Treu, L., Kougias, P. G., De Francisci, D., Valle, G., & Angelidaki, I. (2016). Metagenomic analysis and functional characterization of the biogas microbiome using high throughput shotgun sequencing and a novel binning strategy. *Biotechnology for Biofuels*, *9*, 26.

Campbell, B. J., Yu, L., Heidelberg, J. F., & Kirchman, D. L. (2011). Activity of abundant and rare bacteria in a coastal ocean. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(31), 12776–12781.

Čanković, M., Petrić, I., Marguš, M., & Ciglenečki, I. (2017). Spatio-temporal dynamics of sulfate-reducing bacteria in extreme environment of Rogoznica Lake revealed by 16S rRNA analysis. *Journal of Marine Systems*, *172*, 14–23.

Casali, N., Broda, A., Harris, S. R., Parkhill, J., Brown, T., & Drobniewski, F. (2016). Whole Genome

Sequence Analysis of a Large Isoniazid-Resistant Tuberculosis Outbreak in London: A Retrospective Observational Study. *PLoS Medicine*, *13*(10), e1002137.

Chaparro, J. M., Sheflin, A. M., Manter, D. K., & Vivanco, J. M. (2012). Manipulating the soil microbiome to increase soil health and plant fertility. *Biology and Fertility of Soils*, *48*(5), 489–499.

Chen, H., Wu, H., Yan, B., Zhao, H., Liu, F., Zhang, H., … Liang, Z. (2018). Core Microbiome of Medicinal Plant Salvia miltiorrhiza Seed: A Rich Reservoir of Beneficial Microbes for Secondary Metabolism? *International Journal of Molecular Sciences*, *19*(3). https://doi.org/10.3390/ijms19030672

Chilton, M. D., Drummond, M. H., Merio, D. J., Sciaky, D., Montoya, A. L., Gordon, M. P., & Nester, E. W. (1977). Stable incorporation of plasmid DNA into higher plant cells: the molecular basis of crown gall tumorigenesis. *Cell*, *11*(2), 263–271.

Cho, I., & Blaser, M. J. (2012). The human microbiome: at the interface of health and disease. *Nature Reviews. Genetics*, *13*(4), 260–270.

Christa, G., Händeler, K., Kück, P., Vleugels, M., Franken, J., Karmeinski, D., & Wägele, H. (2015). Phylogenetic evidence for multiple independent origins of functional kleptoplasty in Sacoglossa (Heterobranchia, Gastropoda). *Organisms, Diversity & Evolution*, *15*(1), 23–36.

Ciancio, A., Pieterse, C. M. J., & Mercado-Blanco, J. (2016). Editorial: Harnessing Useful Rhizosphere Microorganisms for Pathogen and Pest Biocontrol. *Frontiers in Microbiology*, *7*, 1620.

Coates, M. E., & Beynon, J. L. (2010). Hyaloperonospora Arabidopsidis as a pathogen model. *Annual Review of Phytopathology*, *48*, 329–345.

Coelho, L. P., Kultima, J. R., Costea, P. I., Fournier, C., Pan, Y., Czarnecki-Maulden, G., … Bork, P. (2018). Similarity of the dog and human gut microbiomes in gene content and response to diet. *Microbiome*, *6*(1), 72.

Colman, D. R., Poudel, S., Stamps, B. W., Boyd, E. S., & Spear, J. R. (2017). The deep, hot biosphere: Twenty-five years of retrospection. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(27), 6895–6903.

Compant, S., Samad, A., Faist, H., & Sessitsch, A. (2019). A review on the plant microbiome: Ecology, functions, and emerging trends in microbial application. *Journal of Advanced Research*. https://doi.org/10.1016/j.jare.2019.03.004

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991.

Cooper, A. J., Latunde-Dada, A. O., Woods-Tör, A., Lynn, J., Lucas, J. A., Crute, I. R., & Holub, E. B. (2008). Basic compatibility of Albugo candida in Arabidopsis thaliana and Brassica juncea causes broad-spectrum suppression of innate immunity. *Molecular Plant-Microbe Interactions: MPMI*, *21*(6), 745–756.

Crea, R., Kraszewski, A., Hirose, T., & Itakura, K. (1978). Chemical synthesis of genes for human insulin. *Proceedings of the National Academy of Sciences of the United States of America*, *75*(12), 5765–5769.

Cui, H., & Zhang, X. (2013). Alignment-free supervised classification of metagenomes by recursive SVM. *BMC Genomics*, *14*, 641.

Davis, K. E. R., Joseph, S. J., & Janssen, P. H. (2005). Effects of growth medium, inoculum size, and incubation time on culturability and isolation of soil bacteria. *Applied and Environmental Microbiology*, *71*(2), 826–834.

Denamur, E., & Matic, I. (2006). Evolution of mutation rates in bacteria. *Molecular Microbiology*, *60*(4), 820–827.

Dick, G. J., Anantharaman, K., Baker, B. J., Li, M., Reed, D. C., & Sheik, C. S. (2013). The microbiology of deep-sea hydrothermal vent plumes: ecological and biogeographic linkages to seafloor and water column habitats. *Frontiers in Microbiology*, *4*, 124.

Dickinson, C. H., Austin, B., & Goodfellow, M. (1975). Quantitative and Qualitative Studies of Phylloplane Bacteria from Lolium perenne. *Microbiology*, *91*(1), 157–166.

Ding, W., Baumdicker, F., & Neher, R. A. (2018). panX: pan-genome analysis and exploration. *Nucleic Acids Research*, *46*(1), e5.

DNA Sequencing Costs: Data. (n.d.). Retrieved March 15, 2019, from https://www.genome.gov/27541954/dna-sequencing-costs-data/

Druschel, G. K., & Kappler, A. (2015). Geomicrobiology and Microbial Geochemistry. *Elements* , *11*(6), 389–394.

Duchêne, S., Holt, K. E., Weill, F.-X., Le Hello, S., Hawkey, J., Edwards, D. J., … Holmes, E. C. (2016). Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics*, *2*(11), e000094.

Dunivin, T. K., Choi, J., Howe, A., & Shade, A. (2019). RefSoil+: a Reference Database for Genes and Traits of Soil Plasmids. *mSystems*, *4*(1). https://doi.org/10.1128/mSystems.00349-18

Enya, J., Shinohara, H., Yoshida, S., Tsukiboshi, T., Negishi, H., Suyama, K., & Tsushima, S. (2007). Culturable leaf-associated bacteria on tomato plants and their potential as biological control agents. *Microbial Ecology*, *53*(4), 524–536.

Faust, K., & Raes, J. (2012). Microbial interactions: from networks to models. *Nature Reviews. Microbiology*, *10*(8), 538–550.

Felsenstein, J. (1978). Cases in which Parsimony or Compatibility Methods will be Positively Misleading. *Systematic Biology*, *27*(4), 401–410.

Ferragina, P., & Manzini, G. (2000). Opportunistic data structures with applications. In *Proceedings 41st Annual Symposium on Foundations of Computer Science* (pp. 390–398). ieeexplore.ieee.org.

Fiannaca, A., La Paglia, L., La Rosa, M., Lo Bosco, G., Renda, G., Rizzo, R., … Urso, A. (2018). Deep learning models for bacteria taxonomic classification of metagenomic data. *BMC Bioinformatics*, *19*(Suppl 7), 198.

Fields, S., & Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, *340*(6230), 245–246.

Figueroa, I. A., Barnum, T. P., Somasekhar, P. Y., Carlström, C. I., Engelbrektson, A. L., & Coates, J. D. (2018). Metagenomics-guided analysis of microbial chemolithoautotrophic phosphite oxidation yields evidence of a seventh natural CO2 fixation pathway. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(1), E92–E101.

Fontaneto, D. (2011). *Biogeography of Microscopic Organisms: Is Everything Small Everywhere?* Cambridge University Press.

Frank, J. A., Pan, Y., Tooming-Klunderud, A., Eijsink, V. G. H., McHardy, A. C., Nederbragt, A. J., & Pope, P. B. (2016). Improved metagenome assemblies and taxonomic binning using long-read circular consensus sequence data. *Scientific Reports*, *6*, 25373.

Freschi, L., Vincent, A. T., Jeukens, J., Emond-Rheault, J.-G., Kukavica-Ibrulj, I., Dupont, M.-J., … Levesque, R. C. (2019). The Pseudomonas aeruginosa Pan-Genome Provides New Insights on Its Population Structure, Horizontal Gene Transfer, and Pathogenicity. *Genome Biology and Evolution*, *11*(1), 109–120.

Friedman, J., & Alm, E. J. (2012). Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, *8*(9), e1002687.

Fry, M. (2016). *Landmark Experiments in Molecular Biology*. Academic Press.

García-Aljaro, C., Blanch, A. R., Campos, C., Jofre, J., & Lucena, F. (2019). Pathogens, faecal indicators and human-specific microbial source-tracking markers in sewage. *Journal of Applied Microbiology*, *126*(3), 701–717.

Garrison, E., & Marth, G. (2012). *Haplotype-based variant detection from short-read sequencing*. *arXiv [q-bio.GN]*. Retrieved from http://arxiv.org/abs/1207.3907

Garud, N. R., Good, B. H., Hallatschek, O., & Pollard, K. S. (2019). Evolutionary dynamics of bacteria in the gut microbiome within and across hosts. *PLoS Biology*, *17*(1), e3000102.

Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392–400.

Glick, B. R. (1995). The enhancement of plant growth by free-living bacteria. *Canadian Journal of Microbiology*, *41*(2), 109–117.

Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V., & Egozcue, J. J. (2017). Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology*, *8*, 2224.

Graham, D. Y., Klein, P. D., Evans, D. J., Jr, Evans, D. G., Alpert, L. C., Opekun, A. R., & Boutton, T. W. (1987). Campylobacter pylori detected noninvasively by the 13C-urea breath test. *The Lancet*, *1*(8543), 1174–1177.

Gram, & C, H. (1884). Ueber die isolirte Färbung der Schizomyceten in Schnitt- und Trockenpräparaten. *Fortschritte Der Medicin*, *2*, 185–189.

Greenblum, S., Carr, R., & Borenstein, E. (2015). Extensive strain-level copy-number variation across human gut microbiome species. *Cell*, *160*(4), 583–594.

Griffith, F. (1928). The Significance of Pneumococcal Types. *The Journal of Hygiene*, *27*(2), 113–159.

Gruber-Vodicka, H. R., Seah, B. K. B., & Pruesse, E. (2019). *phyloFlash — Rapid SSU rRNA profiling and targeted assembly from metagenomes*. *bioRxiv*. https://doi.org/10.1101/521922

Gupta, V. K., Paul, S., & Dutta, C. (2017). Geography, Ethnicity or Subsistence-Specific Variations in Human Microbiome Composition and Diversity. *Frontiers in Microbiology*, *8*, 1162.

Hacquard, S. (2016). Disentangling the factors shaping microbiota composition across the plant holobiont. *The New Phytologist*, *209*(2), 454–457.

Hagelberg, E., Hofreiter, M., & Keyser, C. (2015). Introduction. Ancient DNA: the first three decades. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *370*(1660), 20130371.

Hassani, M. A., Durán, P., & Hacquard, S. (2018). Microbial interactions within the plant holobiont. *Microbiome*, *6*(1), 58.

Hayashi, H., Sakamoto, M., & Benno, Y. (2002). Phylogenetic analysis of the human gut microbiota using 16S rDNA clone libraries and strictly anaerobic culture-based methods. *Microbiology and Immunology*, *46*(8), 535–548.

Heintz-Buschart, A., & Wilmes, P. (2018). Human Gut Microbiome: Function Matters. *Trends in Microbiology*, *26*(7), 563–574.

Herbig, A., Maixner, F., Bos, K. I., Zink, A., Krause, J., & Huson, D. H. (2016). *MALT: Fast alignment and analysis of metagenomic DNA sequence data applied to the Tyrolean Iceman*. *bioRxiv*. https://doi.org/10.1101/050559

Hesse, C., Schulz, F., Bull, C. T., Shaffer, B. T., Yan, Q., Shapiro, N., … Loper, J. E. (2018). Genome-based evolutionary history of Pseudomonas spp. *Environmental Microbiology*, *20*(6), 2142–2159.

Hillmann, B., Al-Ghalith, G. A., Shields-Cutler, R., Zhu, Q., Gohl, D., Beckman, K. B., … Knights, D. (2018). *Evaluating the information content of shallow shotgun metagenomics*. *bioRxiv*. https://doi.org/10.1101/320986

Hill, T. C. J., Walsh, K. A., Harris, J. A., & Moffett, B. F. (2003). Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology*, *43*(1), 1–11.

Hirano, S. S., & Upper, C. D. (2000). Bacteria in the Leaf Ecosystem with Emphasis onPseudomonas syringae—a Pathogen, Ice Nucleus, and Epiphyte. *Microbiology and Molecular Biology Reviews: MMBR*. Retrieved from https://mmbr.asm.org/content/64/3/624.short

Howe, A., & Chain, P. S. G. (2015). Challenges and opportunities in understanding microbial communities with metagenome assembly (accompanied by IPython Notebook tutorial). *Frontiers in Microbiology*, *6*, 678.

Howe, A. C., Jansson, J. K., Malfatti, S. A., Tringe, S. G., Tiedje, J. M., & Brown, C. T. (2014). Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(13), 4904–4909.

Huddleston, J., Chaisson, M. J. P., Steinberg, K. M., Warren, W., Hoekzema, K., Gordon, D., … Eichler, E. E. (2017). Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Research*, *27*(5), 677–685.

HU, Jie,HE Xiaohong,LI Daping & LIU Qiang. (2007). Progress in Research of Sphingomonas. *Ying Yong Yu Huan Jing Sheng Wu Xue Bao = Chinese Journal of Applied and Environmental Biology / Zhongguo Ke Xue Yuan Chengdu Sheng Wu Yan Jiu Suo Zhu Ban*. Retrieved from http://en.cnki.com.cn/Article_en/CJFDTotal-YYHS200703029.htm

Huson, D. H., Albrecht, B., Bağcı, C., Bessarab, I., Górska, A., Jolic, D., & Williams, R. B. H. (2018). MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biology Direct*, *13*(1), 6.

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., … Tappu, R. (2016). MEGAN Community Edition - Interactive Exploration and Analysis of Large-Scale Microbiome Sequencing Data. *PLoS Computational Biology*, *12*(6), e1004957.

Innerebner, G., Knief, C., & Vorholt, J. A. (2011). Protection of Arabidopsis thaliana against leaf-pathogenic Pseudomonas syringae by Sphingomonas strains in a controlled model system. *Applied and Environmental Microbiology*, *77*(10), 3202–3210.

Jakob, K., Goss, E. M., Araki, H., Van, T., Kreitman, M., & Bergelson, J. (2002). Pseudomonas viridiflava and P. syringae--natural pathogens of Arabidopsis thaliana. *Molecular Plant-Microbe Interactions: MPMI*, *15*(12), 1195–1203.

Jeong, C., Alkorta-Aranburu, G., Basnyat, B., Neupane, M., Witonsky, D. B., Pritchard, J. K., … Di Rienzo, A. (2014). Admixture facilitates genetic adaptations to high altitude in Tibet. *Nature Communications*, *5*, 3281.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J. A., & Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science*, *337*(6096), 816–821.

Joly, M., Attard, E., Sancelme, M., Deguillaume, L., Guilbaud, C., Morris, C. E., … Delort, A.-M. (2013). Ice nucleation activity of bacteria isolated from cloud water. *Atmospheric Environment*, *70*, 392–400.

Jones, O. W., Jr, & Nirenberg, M. W. (1966). Degeneracy in the amino acid code. *Biochimica et Biophysica Acta*, *119*(2), 400–406.

Kaplan, B. J., Rucklidge, J. J., Romijn, A., & McLeod, K. (2015). The Emerging Field of Nutritional Mental Health: Inflammation, the Microbiome, Oxidative Stress, and Mitochondrial Function. *Clinical Psychological Science*, *3*(6), 964–980.

Karasov, T. L., Almario, J., Friedemann, C., Ding, W., Giolai, M., Heavens, D., … Weigel, D. (2018). Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host & Microbe*, *24*(1), 168–179.e4.

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

Katagiri, F., Thilmony, R., & He, S. Y. (2002). The Arabidopsis thaliana-pseudomonas syringae interaction. *The Arabidopsis Book / American Society of Plant Biologists*, *1*, e0039.

Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L., & Gordon, J. I. (2011). Human nutrition, the gut microbiome and the immune system. *Nature*, *474*(7351), 327–336.

Kecskeméti, E., Berkelmann-Löhnertz, B., & Reineke, A. (2016). Are Epiphytic Microbial Communities in the Carposphere of Ripening Grape Clusters (Vitis vinifera L.) Different between Conventional, Organic, and Biodynamic Grapes? *PloS One*, *11*(8), e0160852.

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

Kemp, P. F., & Aller, J. Y. (2004). Estimating prokaryotic diversity: when are 16S rDNA libraries large enough? *Limnology and Oceanography, Methods / ASLO*, *2*(4), 114–125.

Kent, W. J. (2002). BLAT—the BLAST-like alignment tool. *Genome Research*. Retrieved from http://genome.cshlp.org/content/12/4/656.short

Kim, D., Song, L., Breitwieser, F. P., & Salzberg, S. L. (2016). Centrifuge: rapid and sensitive classification of metagenomic sequences. *Genome Research*, *26*(12), 1721–1729.

Kim, H., Nishiyama, M., Kunito, T., Senoo, K., Kawahara, K., Murakami, K., & Oyaizu, H. (1998). High population of Sphingomonas species on plant surface. *Journal of Applied Microbiology*, *85*(4), 731–736.

Kim, M. K., Schubert, K., Im, W.-T., Kim, K.-H., Lee, S.-T., & Overmann, J. (2007). Sphingomonas kaistensis sp. nov., a novel alphaproteobacterium containing pufLM genes. *International Journal of Systematic and Evolutionary Microbiology*, *57*(Pt 7), 1527–1534.

Kim, M., Singh, D., Lai-Hoe, A., Go, R., Abdul Rahim, R., Ainuddin, A. N., … Adams, J. M. (2012). Distinctive phyllosphere bacterial communities in tropical trees. *Microbial Ecology*, *63*(3), 674–681.

Kiran, K., Rawal, H. C., Dubey, H., Jaswal, R., Devanna, B. N., Gupta, D. K., … Sharma, T. R. (2016). Draft Genome of the Wheat Rust Pathogen (Puccinia triticina) Unravels Genome-Wide Structural Variations during Evolution. *Genome Biology and Evolution*, *8*(9), 2702–2721.

Kircher, M., & Kelso, J. (2010). High-throughput DNA sequencing--concepts and limitations. *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology*, *32*(6), 524–536.

Kisand, V., & Wikner, J. (2003). Limited resolution of 16S rDNA DGGE caused by melting properties and closely related DNA sequences. *Journal of Microbiological Methods*, *54*(2), 183–191.

Klausa, V., Piešiniene, L., Staniulis, J., & Nivinskas, R. (2003). Abundance of T4-type bacteriophages in municipal wastewater and sewage. *Ekologija* , *1*, 47–50.

Knief, C., Delmotte, N., Chaffron, S., Stark, M., Innerebner, G., Wassmann, R., … Vorholt, J. A. (2012).

Metaproteogenomic analysis of microbial communities in the phyllosphere and rhizosphere of rice. *The ISME Journal*, *6*(7), 1378–1390.

Knight, R., Jansson, J., Field, D., Fierer, N., Desai, N., Fuhrman, J. A., … Gilbert, J. A. (2012). Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology*, *30*(6), 513–520.

Kniskern, J. M., Barrett, L. G., & Bergelson, J. (2011). Maladaptation in wild populations of the generalist plant pathogen Pseudomonas syringae. *Evolution; International Journal of Organic Evolution*, *65*(3), 818–830.

Knoll, A. H. (2011). The Multiple Origins of Complex Multicellularity. *Annual Review of Earth and Planetary Sciences*, *39*(1), 217–239.

Koch, N. M., Matos, P., Branquinho, C., Pinho, P., Lucheta, F., Martins, S. M. de A., & Vargas, V. M. F. (2019). Selecting lichen functional traits as ecological indicators of the effects of urban environment. *The Science of the Total Environment*, *654*, 705–713.

Koren, O., Knights, D., Gonzalez, A., Waldron, L., Segata, N., Knight, R., … Ley, R. E. (2013). A guide to enterotypes across the human body: meta-analysis of microbial community structures in human microbiome datasets. *PLoS Computational Biology*, *9*(1), e1002863.

Kose, S. H., Grice, K., Orsi, W. D., Ballal, M., & Coolen, M. J. L. (2018). Metagenomics of pigmented and cholesterol gallstones: the putative role of bacteria. *Scientific Reports*, *8*(1), 11218.

Kouchaki, S., Tapinos, A., & Robertson, D. L. (2019). A signal processing method for alignment-free metagenomic binning: multi-resolution genomic binary patterns. *Scientific Reports*, *9*(1), 2159.

Kristin, A., & Miranda, H. (2013). The root microbiota—a fingerprint in the soil? *Plant and Soil*, *370*(1), 671–686.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, *11*(5), e1004226.

Kyrpides, N. C., Hugenholtz, P., Eisen, J. A., Woyke, T., Göker, M., Parker, C. T., … Others. (2014). Genomic encyclopedia of bacteria and archaea: sequencing a myriad of type strains. *PLoS Biology*, *12*(8), e1001920.

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., … Huala, E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Research*, *40*(Database issue), D1202–D1210.

Lareen, A., Burton, F., & Schäfer, P. (2016). Plant root-microbe communication in shaping root microbiomes. *Plant Molecular Biology*, *90*(6), 575–587.

Layeghifard, M., Hwang, D. M., & Guttman, D. S. (2017). Disentangling Interactions in the Microbiome: A Network Perspective. *Trends in Microbiology*, *25*(3), 217–228.

Lebeis, S. L. (2014). The potential for give and take in plant--microbiome relationships. *Frontiers in Plant Science*, *5*, 287.

Lebeis, S. L., Rott, M., Dangl, J. L., & Schulze-Lefert, P. (2012). Culturing a plant microbiome community at the cross-Rhodes. *The New Phytologist*, *196*(2), 341–344.

Lee, J., & Ramirez, W. F. (1994). Optimal fed-batch control of induced foreign protein production by recombinant bacteria. *AIChE Journal. American Institute of Chemical Engineers*, *40*(5), 899–907.

Lee, M. S., Oh, S., & Tang, H. (2014). Characterization of microbial associations in human oral microbiome. *Bio-Medical Materials and Engineering*, *24*(6), 3737–3744.

Levy, S. B., & Marshall, B. (2004). Antibacterial resistance worldwide: causes, challenges and responses. *Nature Medicine*, *10*(12 Suppl), S122–S129.

Li, D., Liu, C.-M., Luo, R., Sadakane, K., & Lam, T.-W. (2015). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* , *31*(10), 1674–1676.

Li, H. (2013). *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM*. *arXiv [q-bio.GN]*. Retrieved from http://arxiv.org/abs/1303.3997

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* , *34*(18), 3094–3100.

Lindow, S. E., & Brandl, M. T. (2003). Microbiology of the phyllosphere. *Applied and Environmental Microbiology*, *69*(4), 1875–1883.

Li, Y., Jing, H., Xia, X., Cheung, S., Suzuki, K., & Liu, H. (2018). Metagenomic Insights Into the Microbial Community and Nutrient Cycling in the Western Subarctic Pacific Ocean. *Frontiers in Microbiology*, *9*, 623.

Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A. B., … Huttenhower, C. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. *Nature*, *550*(7674), 61–66.

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

Lu, J., Breitwieser, F. P., Thielen, P., & Salzberg, S. L. (2017). Bracken: estimating species abundance in metagenomics data. *PeerJ Computer Science*, *3*, e104.

Luo, X., Reiter, M. A., d'Espaux, L., Wong, J., Denby, C. M., Lechner, A., … Keasling, J. D. (2019). Complete biosynthesis of cannabinoids and their unnatural analogues in yeast. *Nature*, *567*(7746), 123–126.

Maaten, L. van der, & Hinton, G. (2008). Visualizing Data using t-SNE. *Journal of Machine Learning Research: JMLR*, *9*(Nov), 2579–2605.

Ma, B., Tromp, J., & Li, M. (2002). PatternHunter: faster and more sensitive homology search. *Bioinformatics* , *18*(3), 440–445.

Mackenzie, S., & McIntosh, L. (1999). Higher plant mitochondria. *The Plant Cell*, *11*(4), 571–586.

Maderankova, D., Jugas, R., Sedlar, K., Vitek, M., & Skutkova, H. (2019). Rapid Bacterial Species Delineation Based on Parameters Derived From Genome Numerical Representations. *Computational and Structural Biotechnology Journal*, *17*, 118–126.

Ma, J., & Amos, C. I. (2012). Principal components analysis of population admixture. *PloS One*, *7*(7), e40115.

Majumder, E. L.-W., & Wall, J. D. (2017). Uranium Bio-Transformations: Chemical or Biological Processes? *Open Journal of Inorganic Chemistry*, *07*(02), 28–60.

Marcais, G., & Kingsford, C. (2012). Jellyfish: A fast k-mer counter. eagle.fish.washington.edu. Retrieved from http://eagle.fish.washington.edu/whale/fish546/Trinity_r2013-08-14_analysis1-2014-02-08-20-44-13.233/bin/trinityrnaseq_r2013_08_14/trinity-plugins/jellyfish-1.1.6/doc/jellyfish.pdf

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews. Genetics*, *11*(7), 499–511.

Márquez, L. M., Redman, R. S., Rodriguez, R. J., & Roossinck, M. J. (2007). A virus in a fungus in a plant: three-way symbiosis required for thermal tolerance. *Science*, *315*(5811), 513–515.

Matthaei, J. H., Jones, O. W., Martin, R. G., & Nirenberg, M. W. (1962). Characteristics and composition of RNA coding units. *Proceedings of the National Academy of Sciences of the United States of America*, *48*, 666–677.

M. Burrows, D. J. W. (1994). A block-sorting lossless data compression algorithm. Retrieved from http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.3.8069

Medini, D., Donati, C., Tettelin, H., Masignani, V., & Rappuoli, R. (2005). The microbial pan-genome. *Current Opinion in Genetics & Development*, *15*(6), 589–594.

Mendes, R., Garbeva, P., & Raaijmakers, J. M. (2013). The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiology Reviews*, *37*(5), 634–663.

Meselson, M., & Stahl, F. W. (1958). THE REPLICATION OF DNA IN ESCHERICHIA COLI. *Proceedings of the National Academy of Sciences of the United States of America*, *44*(7), 671–682.

Metwaly, A., & Haller, D. (2019). Strain-Level Diversity in the Gut: The P. copri Case. *Cell Host & Microbe*, *25*(3), 349–350.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., … Edwards, R. A. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, *9*, 386.

Mitter, B., Pfaffenbichler, N., & Sessitsch, A. (2016). Plant-microbe partnerships in 2020. *Microbial Biotechnology*, *9*(5), 635–640.

Miyauchi, K., Adachi, Y., Nagata, Y., & Takagi, M. (1999). Cloning and sequencing of a novel meta-cleavage dioxygenase gene whose product is involved in degradation of gamma-hexachlorocyclohexane in Sphingomonas paucimobilis. *Journal of Bacteriology*, *181*(21), 6712–6719.

Mizrahi, I., & Jami, E. (2018). Review: The compositional variation of the rumen microbiome and its effect on host performance and methane emission. *Animal: An International Journal of Animal Bioscience*, *12*(s2), s220–s232.

Morán, F., Marco-Noales, E., Escrich, A., Barbé, S., López, M. M., & Others. (2018). Biodiversity and Biogeography of Three Pseudomonas syringae Pathovars which Affect Kiwi Fruit Cultivation. *Biodiversity Online Journal*, *1*(1), 1–3.

Morella, N. M., Gomez, A. L., Wang, G., Leung, M. S., & Koskella, B. (2018). The impact of bacteriophages on phyllosphere bacterial abundance and composition. *Molecular Ecology*, *27*(8), 2025–2038.

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

Murphy, L. R., Wallqvist, A., & Levy, R. M. (2000). Simplified amino acid alphabets for protein fold recognition and implications for folding. *Protein Engineering*, *13*(3), 149–152.

Nagata, Y., Miyauchi, K., & Takagi, M. (1999). Complete analysis of genes and enzymes for γ-hexachlorocyclohexane degradation in Sphingomonas paucimobilis UT26. *Journal of Industrial Microbiology & Biotechnology*, *23*(4), 380–390.

Narasingarao, P., Podell, S., Ugalde, J. A., Brochier-Armanet, C., Emerson, J. B., Brocks, J. J., … Allen, E. E. (2012). De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME Journal*, *6*(1), 81–93.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. *Nature Reviews. Genetics*, *12*, 443.

Novembre, J., & Stephens, M. (2008). Interpreting principal component analyses of spatial population genetic variation. *Nature Genetics*, *40*(5), 646–649.

Nyachuba, D. G. (2010). Foodborne illness: is it on the rise? *Nutrition Reviews*, *68*(5), 257–269.

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

Olson, N. D., Treangen, T. J., Hill, C. M., Cepeda-Espinoza, V., Ghurye, J., Koren, S., & Pop, M. (2017). Metagenomic assembly through the lens of validation: recent advances in assessing and improving the quality of genomes assembled from metagenomes. *Briefings in Bioinformatics*. https://doi.org/10.1093/bib/bbx098

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, *17*(1), 132.

Orlóci, L. (1974). Revisions for the Bray and Curtis ordination. *Canadian Journal of Botany. Journal Canadien de Botanique*, *52*(7), 1773–1776.

Pachiadaki, M. G., Sintes, E., Bergauer, K., Brown, J. M., Record, N. R., Swan, B. K., … Stepanauskas, R. (2017). Major role of nitrite-oxidizing bacteria in dark ocean carbon fixation. *Science*, *358*(6366), 1046–1051.

Panke-Buisse, K., Lee, S., & Kao-Kniffin, J. (2017). Cultivated Sub-Populations of Soil Microbiomes Retain Early Flowering Plant Trait. *Microbial Ecology*, *73*(2), 394–403.

Parisutham, V., Kim, T. H., & Lee, S. K. (2014). Feasibilities of consolidated bioprocessing microbes: from pretreatment to biofuel production. *Bioresource Technology*, *161*, 431–440.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A., & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, *36*(10), 996–1004.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., … Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, *2*(11), 1533–1542.

Patra, A. K., Mukhopadhyay, R., Mukhija, R., Krishnan, A., Garg, L. C., & Panda, A. K. (2000). Optimization of inclusion body solubilization and renaturation of recombinant human growth hormone from Escherichia coli. *Protein Expression and Purification*, *18*(2), 182–192.

Patterson, N., Price, A. L., & Reich, D. (2006). Population structure and eigenanalysis. *PLoS Genetics*, *2*(12), e190.

Pidwirny, M. (2010, April 6). Introduction to the Oceans. Retrieved April 28, 2019, from http://www.physicalgeography.net/fundamentals/8o.html

Porras-Alfaro, A., & Bayman, P. (2011). Hidden fungi, emergent properties: endophytes and microbiomes. *Annual Review of Phytopathology*, *49*, 291–315.

Possingham, J. V. (1980). Plastid Replication and Development in the Life Cycle of Higher Plants. *Annual Review of Plant Physiology*, *31*(1), 113–129.

Preston, G. M. (2000). Pseudomonas syringae pv. tomato: the right pathogen, of the right plant, at the right time. *Molecular Plant Pathology*, *1*(5), 263–275.

Purahong, W., Orrù, L., Donati, I., Perpetuini, G., Cellini, A., Lamontanara, A., … Spinelli, F. (2018). Plant Microbiome and Its Link to Plant Health: Host Species, Organs and Pseudomonas syringae pv. actinidiae Infection Shaping Bacterial Phyllosphere Communities of Kiwifruit Plants. *Frontiers in Plant Science*, *9*, 1563.

Quail, M. A., Swerdlow, H., & Turner, D. J. (2009). Improved protocols for the illumina genome analyzer sequencing system. *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]*, *Chapter 18*, Unit 18.2.

Raman, S., Rogers, J. K., Taylor, N. D., & Church, G. M. (2014). Evolution-guided optimization of biosynthetic pathways. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(50), 17803–17808.

Ramírez-Puebla, S. T., Servín-Garcidueñas, L. E., Jiménez-Marín, B., Bolaños, L. M., Rosenblueth, M., Martínez, J., … Martínez-Romero, E. (2013). Gut and root microbiota commonalities. *Applied and Environmental Microbiology*, *79*(1), 2–9.

Ranjan, R., Rani, A., Metwally, A., McGee, H. S., & Perkins, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochemical and Biophysical Research Communications*, *469*(4), 967–977.

Redford, A. J., Bowers, R. M., Knight, R., Linhart, Y., & Fierer, N. (2010). The ecology of the phyllosphere: geographic and phylogenetic variability in the distribution of bacteria on tree leaves. *Environmental Microbiology*, *12*(11), 2885–2893.

Reich, D., Price, A. L., & Patterson, N. (2008). Principal component analysis of genetic data. *Nature Genetics*, *40*(5), 491–492.

Rice, E. W., Allen, M. J., & Edberg, S. C. (1990). Efficacy of beta-glucuronidase assay for identification of Escherichia coli by the defined-substrate technology. *Applied and Environmental Microbiology*, *56*(5), 1203–1205.

Ricotta, C., & Podani, J. (2017). On some properties of the Bray-Curtis dissimilarity and their ecological meaning. *Ecological Complexity*, *31*, 201–205.

Rodriguez-R, L. M., & Konstantinidis, K. T. (2014). Estimating coverage in metagenomic data sets and why it matters. *The ISME Journal*, *8*(11), 2349–2351.

Roesch, L. F. W., Fulthorpe, R. R., Riva, A., Casella, G., Hadwin, A. K. M., Kent, A. D., … Triplett, E. W. (2007). Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal*, *1*(4), 283–290.

Rojas-Carulla, M., Tolstikhin, I., Luque, G., Youngblut, N., Ley, R., & Schölkopf, B. (2019). *GeNet: Deep Representations for Metagenomics*. *arXiv [q-bio.GN]*. Retrieved from http://arxiv.org/abs/1901.11015

Rolli, E., Marasco, R., Vigani, G., Ettoumi, B., Mapelli, F., Deangelis, M. L., … Daffonchio, D. (2015). Improved plant resistance to drought is promoted by the root-associated microbiome as a water stress-dependent trait. *Environmental Microbiology*, *17*(2), 316–331.

Roossinck, M. J. (2012). Plant virus metagenomics: biodiversity and ecology. *Annual Review of Genetics*, *46*, 359–369.

Rouli, L., Merhej, V., Fournier, P.-E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85.

Rout, M. E. (2014). The plant microbiome. In *Advances in Botanical Research* (Vol. 69, pp. 279–309). Elsevier.

Rowan, B. A., Patel, V., Weigel, D., & Schneeberger, K. (2015). Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3* , *5*(3), 385–398.

Sagan, L. (1967). On the origin of mitosing cells. *Journal of Theoretical Biology*, *14*(3), 255–274.

Sancho, L. G., de la Torre, R., Horneck, G., Ascaso, C., de Los Rios, A., Pintado, A., … Schuster, M. (2007). Lichens survive in space: results from the 2005 LICHENS experiment. *Astrobiology*, *7*(3), 443–454.

Sangwan, N., Xia, F., & Gilbert, J. A. (2016). Recovering complete and draft population genomes from metagenome datasets. *Microbiome*, *4*, 8.

Sapkota, R., Knorr, K., Jørgensen, L. N., O'Hanlon, K. A., & Nicolaisen, M. (2015). Host genotype is an important determinant of the cereal phyllosphere mycobiome. *The New Phytologist*, *207*(4), 1134–1144.

Sarkar, S. F., & Guttman, D. S. (2004). Evolution of the core genome of Pseudomonas syringae, a highly clonal, endemic plant pathogen. *Applied and Environmental Microbiology*, *70*(4), 1999–2012.

Scarborough, J. (1970). Romans and physicians. *The Classical Journal*, *65*, 296–306.

Scholz, M. B., Lo, C.-C., & Chain, P. S. G. (2012). Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. *Current Opinion in Biotechnology*, *23*(1), 9–15.

Schopf, J. W., Kitajima, K., Spicuzza, M. J., Kudryavtsev, A. B., & Valley, J. W. (2018). SIMS analyses of the oldest known assemblage of microfossils document their taxon-correlated carbon isotope compositions. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(1), 53–58.

Schuetz, R., Zamboni, N., Zampieri, M., Heinemann, M., & Sauer, U. (2012). Multidimensional optimality of microbial metabolism. *Science*, *336*(6081), 601–604.

Schürch, A. C., Arredondo-Alonso, S., Willems, R. J. L., & Goering, R. V. (2018). Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *24*(4), 350–354.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics* , *30*(14), 2068–2069.

Selosse, M.-A., Bessis, A., & Pozo, M. J. (2014). Microbial priming of plant and animal immunity: symbionts as developmental signals. *Trends in Microbiology*, *22*(11), 607–613.

Sessitsch, A., Hardoim, P., Döring, J., Weilharter, A., Krause, A., Woyke, T., … Reinhold-Hurek, B. (2012). Functional characteristics of an endophyte community colonizing rice roots as revealed by metagenomic analysis. *Molecular Plant-Microbe Interactions: MPMI*, *25*(1), 28–36.

Shah, N., Tang, H., Doak, T. G., & Ye, Y. (2010). COMPARING BACTERIAL COMMUNITIES INFERRED FROM 16S rRNA GENE SEQUENCING AND SHOTGUN METAGENOMICS. In *Biocomputing 2011* (pp. 165–176). WORLD SCIENTIFIC.

Sickel, W., Van de Weyer, A.-L., Bemm, F., Schultz, J., & Keller, A. (2019). Venus flytrap microbiotas withstand harsh conditions during prey digestion. *FEMS Microbiology Ecology*, *95*(3). https://doi.org/10.1093/femsec/fiz010

Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B., & Jackson, R. W. (2011). Pseudomonas genomes: diverse and adaptable. *FEMS Microbiology Reviews*, *35*(4), 652–680.

Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E., & Edwards, R. A. (2014). FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ*, *2*, e425.

Silverman, J. D., Washburne, A. D., Mukherjee, S., & David, L. A. (2017). A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, *6*. https://doi.org/10.7554/eLife.21887

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* , *31*(19), 3210–3212.

Smets, W., Leff, J. W., Bradford, M. A., McCulley, R. L., Lebeer, S., & Fierer, N. (2016). A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biology & Biochemistry*, *96*, 145–151.

Smith, J. M., Smith, N. H., & O'Rourke, M. (1993). How clonal are bacteria? *Proceedings of the*. Retrieved from https://www.pnas.org/content/90/10/4384.short

Snel, B., Bork, P., & Huynen, M. A. (2002). Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Research*, *12*(1), 17–25.

Somboonna, N., Assawamakin, A., Wilantho, A., Tangphatsornruang, S., & Tongsima, S. (2012). Metagenomic profiles of free-living archaea, bacteria and small eukaryotes in coastal areas of Sichang island, Thailand. *BMC Genomics*, *13 Suppl 7*, S29.

Stackebrandt, E., & Goebel, B. M. (1994). Taxonomic Note: A Place for DNA-DNA Reassociation and 16S rRNA Sequence Analysis in the Present Species Definition in Bacteriology. *International Journal of Systematic and Evolutionary Microbiology*, *44*(4), 846–849.

Stämmler, F., Gläsner, J., Hiergeist, A., Holler, E., Weber, D., Oefner, P. J., … Spang, R. (2016). Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome*, *4*(1), 28.

Stavrinides, J., McCann, H. C., & Guttman, D. S. (2007). Host–pathogen interplay and the evolution of bacterial effectors. *Cellular Microbiology*, *0*(0), 071127144819001 – ???

Stein, L. Y., & Klotz, M. G. (2016). The nitrogen cycle. *Current Biology: CB*, *26*(3), R94–R98.

Stewart, R. D., Auffret, M. D., Warr, A., Wiser, A. H., Press, M. O., Langford, K. W., … Watson, M. (2018). Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nature Communications*, *9*(1), 870.

Stockwell, V. O., Johnson, K. B., Sugar, D., & Loper, J. E. (2011). Mechanistically compatible mixtures of bacterial antagonists improve biological control of fire blight of pear. *Phytopathology*, *101*(1), 113–123.

Straub, C., Colombi, E., Li, L., Huang, H., Templeton, M. D., McCann, H. C., & Rainey, P. B. (2018). The ecological genetics of Pseudomonas syringae from kiwifruit leaves. *Environmental Microbiology*, *20*(6), 2066–2084.

Su, J.-Q., An, X.-L., Li, B., Chen, Q.-L., Gillings, M. R., Chen, H., … Zhu, Y.-G. (2017). Metagenomics of urban sewage identifies an extensively shared antibiotic resistome in China. *Microbiome*, *5*(1), 84.

Sunagawa, S., Coelho, L. P., Chaffron, S., Kultima, J. R., Labadie, K., Salazar, G., … Bork, P. (2015). Structure and function of the global ocean microbiome. *Science*, *348*(6237), 1261359.

Sutherland, J. D. (2016). The Origin of Life--Out of the Blue. *Angewandte Chemie* , *55*(1), 104–121.

Tarrand, J. J., & Gröschel, D. H. (1982). Rapid, modified oxidase test for oxidase-variable bacterial isolates. *Journal of Clinical Microbiology*, *16*(4), 772–774.

Temsah, M., Hanna, L., & Saad, A. (2015). First Report of Xylella fastidiosa associated with oleander leaf scorch in Lebanon. *Journal of Crop Protection*, *4*(1), 131–137.

Tessler, M., Neumann, J. S., Afshinnekoo, E., Pineda, M., Hersch, R., Velho, L. F. M., … Brugler, M. R. (2017). Large-scale differences in microbial biodiversity discovery between 16S amplicon and shotgun sequencing. *Scientific Reports*, *7*(1), 6589.

Tikhonovich, I. A., & Provorov, N. A. (2011). Microbiology is the basis of sustainable agriculture: an opinion. *The Annals of Applied Biology*, *159*(2), 155–168.

Tourlousse, D. M., Ohashi, A., & Sekiguchi, Y. (2018). Sample tracking in microbiome community profiling assays using synthetic 16S rRNA gene spike-in controls. *Scientific Reports*, *8*(1), 9095.

Tourlousse, D. M., Yoshiike, S., Ohashi, A., Matsukura, S., Noda, N., & Sekiguchi, Y. (2017). Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Research*, *45*(4), e23.

Townsend, J. P., Bøhn, T., & Nielsen, K. M. (2012). Assessing the probability of detection of horizontal gene transfer events in bacterial populations. *Frontiers in Microbiology*, *3*, 27.

Trosvik, P., & de Muinck, E. J. (2015). Ecology of bacteria in the human gastrointestinal tract--identification of keystone and foundation taxa. *Microbiome*, *3*, 44.

Truong, D. T., Franzosa, E. A., Tickle, T. L., Scholz, M., Weingart, G., Pasolli, E., … Segata, N. (2015). MetaPhlAn2 for enhanced metagenomic taxonomic profiling. *Nature Methods*, *12*(10), 902–903.

Tsilimigras, M. C. B., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Annals of Epidemiology*, *26*(5), 330–335.

Tully, B. J., Graham, E. D., & Heidelberg, J. F. (2018). The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Scientific Data*, *5*, 170203.

van Leeuwenhoek, A. (1800). *The Select Works of Anthony Van Leeuwenhoek: Containing His Microscopical Discoveries in Many of the Works of Nature*. translator.

Vieira, F. G., Lassalle, F., Korneliussen, T. S., & Fumagalli, M. (2016). Improving the estimation of genetic distances from Next-Generation Sequencing data. *Biological Journal of the Linnean Society. Linnean*

Society of London, *117*(1), 139–149.

Visscher, P. T., & Stolz, J. F. (2005). Microbial mats as bioreactors: populations, processes, and products. In N. Noffke (Ed.), *Geobiology: Objectives, Concepts, Perspectives* (pp. 87–100). Amsterdam: Elsevier.

Vogel, C., Innerebner, G., Zingg, J., Guder, J., & Vorholt, J. A. (2012). Forward genetic in planta screen for identification of plant-protective traits of Sphingomonas sp. strain Fr1 against Pseudomonas syringae DC3000. *Applied and Environmental Microbiology*, *78*(16), 5529–5535.

Vorholt, J. A. (2012). Microbial life in the phyllosphere. *Nature Reviews. Microbiology*, *10*(12), 828–840.

Wagner, M. R., Lundberg, D. S., Del Rio, T. G., Tringe, S. G., Dangl, J. L., & Mitchell-Olds, T. (2016). Host genotype and age shape the leaf and root microbiomes of a wild perennial plant. *Nature Communications*, *7*, 12151.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One*, *9*(11), e112963.

Wallace, J. G., Kremling, K. A., Kovar, L. L., & Buckler, E. S. (2018). Quantitative Genetics of the Maize Leaf Microbiome. *Phytobiomes Journal*, *2*(4), 208–224.

Wang, Y.-S., Youngster, S., Grace, M., Bausch, J., Bordens, R., & Wyss, D. F. (2002). Structural and biological characterization of pegylated recombinant interferon alpha-2b and its therapeutic implications. *Advanced Drug Delivery Reviews*, *54*(4), 547–570.

Weisburg, W. G., Barns, S. M., Pelletier, D. A., & Lane, D. J. (1991). 16S ribosomal DNA amplification for phylogenetic study. *Journal of Bacteriology*, *173*(2), 697–703.

Weiß, C. L., Pais, M., Cano, L. M., Kamoun, S., & Burbano, H. A. (2018). nQuire: a statistical framework for ploidy estimation using next generation sequencing. *BMC Bioinformatics*, *19*(1), 122.

Wetterstrand, K. A. (2018, April 25). DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Retrieved April 29, 2019, from https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

Woese, C. R., & Fox, G. E. (1977). Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(11), 5088–5090.

Woese, C. R., Fox, G. E., Zablen, L., Uchida, T., Bonen, L., Pechman, K., … Stahl, D. (1975). Conservation of primary structure in 16S ribosomal RNA. *Nature*, *254*(5495), 83–86.

Wood, D. E., & Salzberg, S. L. (2014). Kraken: ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46.

Woo, P. C. Y., Lau, S. K. P., Teng, J. L. L., Tse, H., & Yuen, K.-Y. (2008). Then and now: use of 16S rDNA gene sequencing for bacterial identification and discovery of novel bacteria in clinical microbiology laboratories. *Clinical Microbiology and Infection: The Official Publication of the European Society of Clinical Microbiology and Infectious Diseases*, *14*(10), 908–934.

Xie, M., Goh, T. N., & Tang, X. Y. (2000). Data transformation for geometrically distributed quality characteristics. *Quality and Reliability Engineering International*, *16*(1), 9–15.

Yatsunenko, T., Rey, F. E., Manary, M. J., Trehan, I., Dominguez-Bello, M. G., Contreras, M., … Gordon, J. I. (2012). Human gut microbiome viewed across age and geography. *Nature*, *486*(7402), 222–227.

Yutin, N., Makarova, K. S., Gussow, A. B., Krupovic, M., Segall, A., Edwards, R. A., & Koonin, E. V. (2018). Discovery of an expansive bacteriophage family that includes the most abundant viruses from the human gut. *Nature Microbiology*, *3*(1), 38–46.

Zaheer, R., Noyes, N., Ortega Polo, R., Cook, S. R., Marinier, E., Van Domselaar, G., … McAllister, T. A. (2018). Impact of sequencing depth on the characterization of the microbiome and resistome. *Scientific Reports*, *8*(1), 5890.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zhang, T., Shao, M.-F., & Ye, L. (2012). 454 pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *The ISME Journal*, *6*(6), 1137–1147.

Zhao, Y., Tang, H., & Ye, Y. (2012). RAPSearch2: a fast and memory-efficient protein similarity search tool for next-generation sequencing data. *Bioinformatics* , *28*(1), 125–126.

Zhou, J., Deng, Y., Luo, F., He, Z., Tu, Q., & Zhi, X. (2010). Functional molecular ecological networks.

mBio, *1*(4). https://doi.org/10.1128/mBio.00169-10

Zillig, W. (1991). Comparative biochemistry of Archaea and Bacteria. *Current Opinion in Genetics & Development*, *1*(4), 544–551.