Vittorio TRACANNA 2022

DISSECTING DISEASE-SUPPRESSIVE RHIZOSPHERE MICROBIOMES USING METAGENOMICS

DISSECTING DISEASE-SUPPRESSIVE RHIZOSPHERE MICROBIOMES USING METAGENOMICS

Vittorio TRACANNA

Propositions

1. All microbiome functional ecological studies require a shotgun metagenomics sequencing step.
   (This thesis)
2. To fully characterize the diversity of microbial communities, it is necessary to include single-cell metagenomic sequencing next to sequencing of bacterial isolates.
   (This thesis)
3. The difference between a good and an impactful scientific study is luck.
4. While scientists are increasingly stimulated to look for job opportunities outside academia, they are insufficiently encouraged to consider making a career in politics.
5. In the future, society will regard periodic psychologist appointments for mental hygiene as a routine process in one's life.
6. Differences in speed between human biological and societal evolution are responsible for most current global challenges.

Propositions belonging to the thesis, entitled

Dissecting disease-suppressive rhizosphere microbiomes using metagenomics

Vittorio Tracanna
Wageningen, 21 February 2022

DISSECTING DISEASE-SUPPRESSIVE RHIZOSPHERE MICROBIOMES USING METAGENOMICS

Vittorio Tracanna

Dissecting disease-suppressive rhizosphere microbiomes using metagenomics

Vittorio Tracanna

Thesis committee

Promotors
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Dr M.H. Medema
Associate professor, Bioinformatics Group
Wageningen University & Research

Other members
Prof. Dr A.H.J. Bisseling, Wageningen University & Research
Dr A. Mengoni, University of Florence, Italy
Dr I. de Bruijn, Koppert Biological Systems, Berkel en Rodenrijs
Dr R. Notebaart, Wageningen University & Research

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

Dissecting disease-suppressive rhizosphere microbiomes using metagenomics

Vittorio Tracanna

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 21 February 2022
at 4 p.m. in the Aula.

Vittorio Tracanna
Dissecting disease-suppressive rhizosphere microbiomes using metagenomics,
220 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2022)
With references, with summary in English

ISBN: 978-94-6447-074-1
DOI: 10.18174/561495

TABLE OF CONTENTS

CHAPTER 1

General Introduction

Part of this chapter was published as “Mining prokaryotes for antimicrobial compounds: from diversity to function” in *FEMS Microbiology Reviews*, 41(3):417-429 (2017).

1.1 - Plants have developed an intimate relationship with soil bacteria

Upon the arrival of the first land plants around 500 mya (Morris et al., 2018), the soil was already rich in many different forms of prokaryotic life (Battistuzzi et al., 2004). Since then, plants have formed enduring and complex relationships with microbes with benefits for both host and microbe (Rimington et al., 2019). The rhizosphere and endosphere are key areas to understand plant-microbe interaction and its evolution. The rhizosphere is no exception to one of the most famous microbial ecology mantras: “everything is everywhere but the environment selects”. Plants recruit bacteria from the surrounding soil to colonize the rhizosphere in exchange for nutrients, in the form of exudates and other plant material, to perform a plethora of functions including protection from pathogens and production of plant growth-promoting metabolites (Turner et al., 2013). The community composition of the rhizosphere depends both on the available or receptive microbes in the community and the compounds produced by the plant to initiate and maintain the symbiotic relationship. In this bipartite host-microbe interaction, plant genotypes play an important role. After all, different plants will have distinct exudate profiles, which in turn attract different members of the soil community microbiota (Zgadzaj et al., 2016).

The endosphere microbiome community is generally assumed to be a distinct and highly selected subpopulation of the rhizosphere community (Figure 1-1). Endophytes provide many beneficial functions, including nutrient acquisition, as well as conferring microbiome-associated phenotypes such as disease suppression (Carrion et al., 2019). Colonization of the endosphere from the rhizosphere generally happens at fissures in the plant root epithelium, which can be both naturally occurring or microbe-dependent. In the case of active breaching of the host by the commensal bacteria, this is done through the expression of cell wall degradation enzymes (Frank et al., 2017). Once the bacteria successfully penetrate the root system, long term colonization is strongly influenced by its interplay with the host immune system. Plant mutants lacking key immune response genes show large differences in endophytic communities (Hein et al., 2008), suggesting directional

selection on the host's microbiome recruitment abilities. At the same time, endosphere-associated bacteria have also evolved an increased adaptation to the endophytic ecological niche. Recently, studies of rhizobacteria isolates have shown specific genetic features (Levy et al., 2018) of rhizobacteria that are associated with disruption of plant immune functions to respond to microbe-associated molecular patterns or to avoid detection altogether. For commensal endo- and rhizobacteria, host health is essential for sustained carbon procurement. Hence, established microbial communities can exert pressure on invading pathogens either through competition for substrate or through the secretion of antimicrobial compounds which mediate interaction within and between microbial kingdoms.

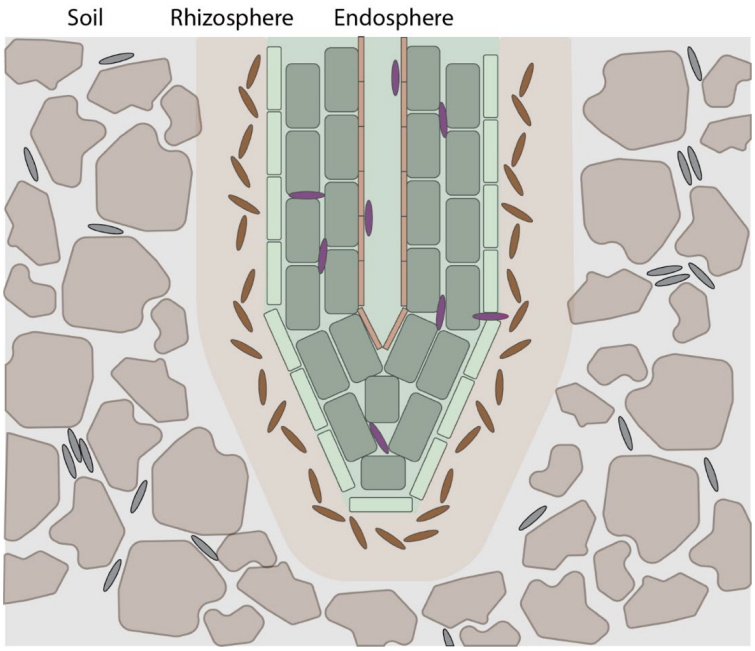


Figure 1-1. Schematic representation of soil, rhizosphere and endosphere environmental niches. Soil bacteria are shown as gray elliptic shapes, rhizosphere bacteria are shown in brown and endophytic bacteria in purple. Plant-receptive bacteria are enriched in the rhizosphere; a subset of rhizosphere bacteria is able to penetrate the host plant and adopt an endophytic lifestyle.

1.2 - Plant-associated microbial communities are an untapped resource of natural products

Microbial specialized metabolites are key mediators of interactions between microbes in the rhizosphere, as well as between microbes and their plant host. They are also the major source of antimicrobials currently used in the clinic, in agriculture and in food manufacturing. Due to the rapid development and spread of resistance against these molecules, there is an urgent need for novel compounds that can supplement the current arsenal. Since the 'golden age of antibiotics' in the sixties and seventies, there has been a steady decrease of novel antibiotic molecules entering the market. However, the recent development of computational genomic approaches to natural product discovery is replenishing hopes that this trend can be turned around: in prokaryotic genome sequences, tens of thousands of biosynthetic gene clusters (BGCs) have been identified (Cimermancic et al., 2014; Doroghazi et al., 2014; Skinnider et al., 2016). Thousands of these BGCs are likely to encode the biosynthesis of thus far unknown molecules (Dejong et al., 2016). To uncover these, many innovative approaches are being developed to link them to metabolomic data with high throughput (Kersten et al., 2011,2013; Medema et al., 2014; Mohimani et al., 2014) or to refactor synthetic versions of them for heterologous expression (Chang et al., 2013; Shao et al., 2013; Yamanaka et al., 2014; Kang, Charlop-Powers and Brady 2016; Montalban-Lopez and Kuipers 2016; van Heel et al., 2016). However, detailed biochemical characterization of biosynthetic pathways and their products is still painstakingly slow and laborious, and many BGCs encode the production of natural products without any (useful) antimicrobial activity. Therefore, targeted approaches are needed to selectively mine genomes for natural products with antimicrobial properties, and to narrow down from tens of thousands of potentially interesting BGCs to manageable numbers that can be tested in the laboratory. Several computational and experimental strategies to this end are currently emerging, based on the analysis of chemical diversity, ecology and evolution, organismal function and/or modes of action.

Genomic information has become more and more important in the process of identifying novel biosynthetic pathways for the production of antimicrobials. To

identify the potential of a bacterium to produce bioactive natural products, mining for BGCs is particularly useful. A wide range of bioinformatic tools (e.g., antiSMASH (Blin et al., 2021), BAGEL3 (van Heel et al., 2013) and PRISM (Skinnider et al., 2016)) are available to identify these, mostly based on shared properties among known classes of biosynthetic pathways (Medema and Fischbach 2015; Ziemert, Alanjary and Weber 2016). For instance, the modification enzymes for the production of lanthipeptides are well conserved (Knerr and van der Donk 2012); as such, they can be used as 'anchors' or 'signatures' for genome mining. Even though natural product BGCs can thus be accurately identified and quantified, the question still remains which of these are most likely to encode the production of potent antimicrobials. Not only the taxonomic origins, but also the chemical structures of antimicrobials are highly diverse. Indeed, known antimicrobial compounds represent a cross-section of several chemical and biosynthetic classes, such as ribosomally synthesized and post translationally modified peptides (RiPPs), nonribosomal peptides, polyketides, terpenoids and even oligosaccharides. Although the percentage of antimicrobials is probably higher among RiPPs than among saccharides, for example, each chemical class of natural products comprises both many antimicrobials and many compounds with different biological activities. To prioritize specifically for antimicrobials or other biological activities of interest, it is necessary to go beyond the genome sequences and couple the genomic information to ecological and functional data.

1.3 - Different sequencing strategies for different purposes

From soil to ocean, from plant roots to animal guts, the ecosystems in which natural products are found are highly diverse. Within these ecosystems, the diversity is enormous: a gram of soil is estimated to contain hundreds to thousands of different species (Curtis, Sloan and Scannell 2002) that form an extremely intertwined society. The metabolic potential hidden in those communities is immense, and systematic analysis of soils across the globe shows very little overlap between the secondary metabolite repertoire of similar soils (Charlop-Powers et al.2014). Potentially, understanding of ecology and microbial communities can be used to chart this variation and prioritize BGCs that are likely to encode the synthesis of molecules that function as potent antimicrobials. Metagenomics is a key technology that allows surveying BGCs and their abundances across varying communities (Wilson and Piel 2013; Charlop-Powers, Milshteyn and Brady 2014). In the hunt for novel antimicrobials from the environment, there are two important strategies: searching for novel chemical scaffolds and searching for novel congeners. Compounds with novel scaffolds are more difficult to discover (also due to rediscovery of known molecules for which no BGCs have yet been characterized), and predicting chemical structures and biological activities from BGC sequence data alone is very challenging. BGCs encoding the biosynthesis of potential congeners, variants upon an existing (and often extensively studied) molecular scaffold, however, can easily be identified based on sequence homology. While sharing the biosynthetic origin with their well-known counterparts, the small differences in the structures of some congeners can have a major effect on different characteristics of the molecule. Notably, those changes may affect the potency, toxicity profile, or the target of the compound, or even the resistance of microorganisms against it. Phylogenetic studies on novel variants of known BGCs can potentially be used to infer the substrates involved and the final products synthesized by the encoded pathway. Regardless of the product type and its discovery method, engineering expression of identified BGCs in a native or heterologous host is frequently necessary for both novel scaffolds and congeners, and is currently a rate-limiting step in antibiotic discovery. Another bottleneck lies in the charting of biosynthetic diversity, as this is often limited to organisms that are easy to culture in the

laboratory. Metagenomics can overcome culture restrictions by sampling material directly from the environment of interest. Alternatively, functional amplicon sequencing approaches can achieve high sequencing depth for BGCs by targeting the shared conserved regions of, e.g., polyketide synthases or nonribosomal peptide synthetases with degenerate primers based on previously characterized genes. The molecule type for which the biosynthesis is encoded in an underlying BGCs can be predicted by assessing the sequence similarity between the sequence of the amplicon in question and curated data aggregation platforms such as MIBiG (Medema et al., 2015) and eSNaPD (Reddy et al., 2014). Development of primer pairs based on BGC classes that specifically encode the biosynthesis of antimicrobial compounds can be used to specifically target BGCs for potential congeners of known antimicrobials. The enormous cost decrease of high throughput sequencing technologies now also allows generation of the immense quantity of data necessary to assemble BGCs directly from microbial communities. In shotgun metagenomics, environmental DNA, eDNA, is extracted from a community sample and sequenced with short read NGS technology. Sequencing platforms such as PacBio and Oxford Nanopore can be used to produce long sequence reads, which can aid in assembling contigs long enough to harbor complete BGCs, even for complex communities. At the intersection between these two solutions, approaches such as 10X Genomics can be used to reconstruct large numbers of long eDNA stretches with low error rates from complex metagenomes. eDNA is first digested in high molecular weight fragments, which are sorted and barcoded in different pools. Standard shotgun sequencing of each pool is then followed by an assembly of each high molecular weight DNA molecule. Although some species bias is introduced during library preparation, the approach has been shown to enable the assembly of synthetic long reads from relatively rare microorganisms in soil (Sharon et al., 2015). The choice of which sequencing platform to use highly depends on the question, Nanopore and PacBio are suited for contiguous assemblies of the most abundant elements of a microbiome, while short read sequencing is better used to describe less abundant members of a community. Metagenome sequencing is able to find novel BGCs regardless of their conservation or representation across known genomes. The main advantage of these techniques compared to amplicon-based approaches lies in their relatively unbiased nature (Table 1-1). The main

disadvantage of assembly-based genome mining lies in the data complexity. However, a plethora of different solutions are becoming available to tackle this. One direct solution that allows untangling of the community is represented by single-cell sequencing. Although the protocols vary and evolve over time, the overallstrategy is consistent: a single cell is isolated, its DNA extracted and amplified to undergo a PCR screen or genome sequencing step. There are examples of successful application of single cell sequencing to identify important secondary metabolite gene clusters such as the apratoxin pathway from a filamentous cyanobacterium (Grindberg et al., 2011). Important steps are being made to apply the latest chromatin interaction techniques, such as Hi-C and 3C, to improve metagenomic assemblies. As of now, those techniques were successfully used to aid the assembly of synthetic metagenomes. (Beitel et al.2014; Burton et al.2014). Potentially, chromatin interaction techniques can become powerful tools to validate metagenomics assemblies, as they provide an extra layer of information that is not used by most binning tools. In addition, information on contiguity or genomic distance between contigs will allow to reconstruct longer and more complete clusters from raw sequence data. Metagenomics-derived BGCs are not easy to revive in heterologous hosts. For instance, proper amounts of physical DNA that contains the BGC of interest is often not readily available; hence, the isolation of the natural producer or synthetic refactoring of the gene cluster is required. Also, regulation of the transcription and the precursors required for the encoded pathway are sometimes not functionally available in classic hosts. Synthetic DNA costs have decreased substantially in the last years, opening the path to high throughput expression of BGCs through synthesis and refactoring in order to match the impressive output of metagenomics analysis. Refactored BGCs are designed to achieve a better control over expression levels in the heterologous host, by replacing native regulation by synthetic promoters, ribosome binding sites and terminators (Medema et al., 2011; Smanski et al., 2016). Also, codon usage can be redesigned to match the host to increase mRNA translation speed. Further manipulation of refactored BGCs is much easier compared to the original, greatly reducing the development time for BGC derived products. Key challenges that need to be overcome here are DNA synthesis costs, tuning stoichiometry of gene expression and avoiding the introduction of unexpected functions into synthetic DNA.

1.4 - Host and environmental conditions are key factors in the expression of natural products

To understand bacteria and harness their metabolic potential, it is important to consider them within a microbial community framework. Indeed, bacteria that thrive under physicochemical conditions in which they have an elaborate interaction network with other species are excellent targets for identification of secondary metabolites. Accordingly, complex communities are reported to be more resistant to invasion from alien species (van Elsas et al., 2012). Multiple factors can influence community resilience, including the production of secondary metabolites with a negative interspecies interaction function, i.e., antibiotics (Cordero et al., 2012). Given the ubiquitous nature of potential target communities, the selection of promising candidates is vital during the experimental design phase. Biologists can use knowledge on an unexpected phenotype to deduce the presence of antimicrobial compounds. Marine sponges are a great example of prime candidates when hunting for novel and potent antibacterial compounds. Sponges have a strong endosymbiotic relation with bacteria to the point that up to 40% of their bio mass is composed by bacterial cells (Friedrich et al., 2001). The microbial community inside the sponge is radically different from the surrounding water. Lately, sponge communities were successfully targeted for identification of novel natural products with antibacterial properties, such as polytheonamides (Trindade-Silva et al., 2012), which might play a role in protecting the sponge host against predators. Suppressive soils are another great example of ecology inspired mining. These soils provide protection to plants against specific pathogens. Conducive soils, which do not restrain the development of a disease, can be gradually transformed into suppressive soils by infecting plants that grow in it in multiple cycles (Berendsen, Pieterse and Bakker 2012). Hence, by investigating the BGCs that are abundant in or expressed by a community in specific conditions (e.g., treatment of suppressive soil with a pathogen compared to control), candidates responsible for the phenotype can be prioritized: this type of strategy has led to the discovery of the lipopeptide thanamycin, which suppresses fungal root pathogens (Mendes et al., 2011; Watrous et al., 2012).

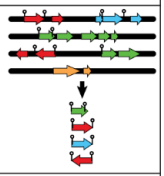
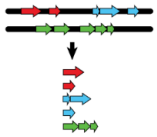
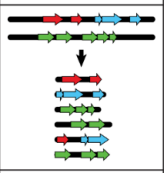
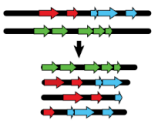
	Pros	Cons	Target	Representation
Functional amplicon sequencing	+ Inexpensive + Data has low complexity	- Incomplete coverage of biosynthetic diversity - Suitable conserved sequence regions are required - Only targets part of the BGC	Individual signature genes or domains	
Metatranscriptomics	+ Allows prioritization of BGCs based on biological responses + Allows better connection of genomic with metabolomic and phenotypic data	- Silent BGCs are not sequenced - Metagenome data required to reconstruct multi-operon BGCs	Operons	
Conventional metagenomics	+ Targets entire BGC + Unbiased	- Assemblies will be fragmented in complex communities - Complex data analysis	Complete BGCs	
HiC / TruSeq / 10X metagenomics	+ Long scaffolds: many BGCs can be completely assembled	- Complex sample preparation - Some species bias - Expensive	Complete BGCs	

Table 1-1. Strong and weak points of the different sequencing methods are described in this table. In the figures, arrows represent genes that are part of a BGC, and genes with the same colors originate from the same operon. The pins in the top right figure indicate conserved stretches targeted by custom primers.

Principled approaches to prioritize environmental BGCs-based on ecology require more than just bare metagenome sequences. Specifically, metatranscriptomics and/or environmental metadata can be used to map and understand differences in BGC abundance and expression, in order to prioritize for those that are most likely to function as antimicrobials. Also, metadata on environmental and physicochemical conditions pertaining microbial communities can potentially be used to direct the search of BGC hotspots. Localized nutritional hotspots, such as organic particles in the ocean, may create highly competitive environmental niches, where microbes utilize antimicrobials to secure resources against competitors. Therefore, metadata on, e.g., nutrient availability could potentially be exploited to identify priority targets. When metatranscriptomics data are generated for diverse samples for which metadata are also recorded, one could even identify BGCs that are expressed specifically under such conditions; these would then have an elevated probability to be involved in generating antimicrobial activity. The Tara Ocean webserver hosts metagenomic information and metadata on a wide range of marine environments.

Particular attention was directed towards the different sampling depths as they are related to the light intensity, an essential driver of community composition. For the human microbiome, which also hosts a wide range of BGCs (including antimicrobials, see Donia et al.2014), several datasets are available with rich clinical metadata, such as the Belgian Flemish Gut Flora Project (Falony et al., 2016) and the Dutch LifeLines DEEP study (Zhernakova et al., 2016). Finally, the rhizosphere is a relatively unexplored metagenome landscape which is likely to yield many novel BGCs which regulate interaction of the microbes with their host and the environment. As more and more metagenomes become available with better assemblies (and including more and more metagenome assembled genomes), richer metadata and rapidly rising amounts of comprehensive (timeseries) metatranscriptomics data, the opportunities for ecology-based antibiotic discovery are likely to increase drastically.

1.5 - Outline of this thesis

The research at the intersection between secondary metabolites discovery and metagenomics has entered a phase of incredible development. With increasing need for novel solutions against crop pathogens, researchers can peer into the untapped potential of microbial communities which naturally play a role in disease suppression for inspiration. In this thesis, I will describe the process to identify the biological mechanisms underlying rhizosphere-mediated suppression of the wheat pathogen *Fusarium culmorum*. In the **second chapter**, we perform the first large scale suppressive soil survey with a combination of phenotyping and marker gene sequencing where we identify four soils with strong *F. culmorum* suppressive characteristics. In the **third chapter**, we further characterize the nonribosomal peptide biosynthesis profiles of microbial community of soils with contrasting conducive and suppressive phenotype. To this end, we focus on co-occurrence patterns of secondary metabolism associated adenylation domains. Additionally, we develop a dedicated tool for the analysis of biosynthetic gene cluster associated domains, dom2BGC, which identifies co-occurring A-domains across samples and identifies co-occurring domains associated to known BGCs from existing repositories. Our findings are also validated by the application of 10X barcode sequencing of a rhizosphere metagenome derived from the most promising

suppressive soil. The **fourth chapter**, completes the dissection process of suppressive soil with a dilution to extinction experiment of a suppressive soil where we track changes in the microbial composition of the soils in their response to infection with the pathogen. We identify multiple MAGs with promising biosynthetic potential that mirror, in relative abundance, the decline in ability of the overall community to suppress the *F. culmorum* infection. In the **fifth chapter**, we describe a *Rhizoctonia solani* suppressive endosphere community and identify bacterial isolates with the ability to suppress fungal infection. Directed mutagenesis of an individual BGC from one of these isolates results in a loss of protection for the host, suggesting a key role of secondary metabolism in this process. Next, in the **sixth chapter** we address the challenges associated with biosynthetic gene cluster reconstruction from complex metagenomics assembly graphs with the development of biosyntheticSPAdes**;** a dedicated assembler where multiple alternative BGC conformations for reconstructed clusters are ranked based on the comparison to previously observed gene clusters. Finally, in the **discussion** I address the impact of sequencing technologies in biology and my vision for future developments in secondary metabolism related metagenomics and microbial ecology.

References

Battistuzzi, F.U., Feijao, A., Hedges, S.B., 2004. A genomic timescale of prokaryote evolution: insights into the origin of methanogenesis, phototrophy, and the colonization of land. BMC Evolutionary Biology 4, 44.

Frank, A.C., Saldierna Guzmán, J.P., Shay, J.E., 2017. Transmission of Bacterial Endophytes. Microorganisms 5, 70.

Hein, J.W., Wolfe, G.V., Blee, K.A., 2008. Comparison of Rhizosphere Bacterial Communities in Arabidopsis thaliana Mutants for Systemic Acquired Resistance. Microb Ecol 55, 333–343.

Levy, A., Salas Gonzalez, I., Mittelviefhaus, M., Clingenpeel, S., Herrera Paredes, S., Miao, J., Wang, K., Devescovi, G., Stillman, K., Monteiro, F., Rangel Alvarez, B., Lundberg, D.S., Lu, T.-Y., Lebeis, S., Jin, Z., McDonald, M., Klein, A.P., Feltcher, M.E., Rio, T.G., Grant, S.R., Doty, S.L., Ley, R.E., Zhao, B., Venturi, V., Pelletier, D.A., Vorholt, J.A., Tringe, S.G., Woyke, T., Dangl, J.L., 2018. Genomic features of bacterial adaptation to plants. Nat Genet 50, 138–150.

Morris, J.L., Puttick, M.N., Clark, J.W., Edwards, D., Kenrick, P., Pressel, S., Wellman, C.H., Yang, Z., Schneider, H., Donoghue, P.C.J., 2018. The timescale of early land plant evolution. PNAS 115, E2274–E2283.

Rimington, W.R., Pressel, S., Duckett, J.G., Field, K.J., Bidartondo, M.I., 2019. Evolution and networks in ancient and widespread symbioses between Mucoromycotina and liverworts. Mycorrhiza 29, 551–565.

Turner, T.R., James, E.K., Poole, P.S., 2013. The plant microbiome. Genome Biology 14, 209.

Zgadzaj, R., Garrido-Oter, R., Jensen, D.B., Koprivova, A., Schulze-Lefert, P., Radutoiu, S., 2016. Root nodule symbiosis in Lotus japonicus drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. PNAS 113, E7996–E8005.

Bakker, P.A.H.M., Doornbos, R.F., Zamioudis, C., Berendsen, R.L., Pieterse, C.M.J., 2013. Induced systemic resistance and the rhizosphere microbiome. Plant Pathol J 29, 136–143.

Beitel, C.W., Froenicke, L., Lang, J.M., Korf, I.F., Michelmore, R.W., Eisen, J.A., Darling, A.E., 2014. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. PeerJ 2, e415.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., Weber, T., 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research 49, W29–W35.

Burton, J.N., Liachko, I., Dunham, M.J., Shendure, J., 2014. Species-Level Deconvolution of Metagenome Assemblies with Hi-C–Based Contact Probability Maps. G3 Genes|Genomes|Genetics 4, 1339–1346.

Chang, F.-Y., Ternei, M.A., Calle, P.Y., Brady, S.F., 2013. Discovery and synthetic refactoring of tryptophan dimer gene clusters from the environment. J Am Chem Soc 135, 10.1021/ja408683p.

Charlop-Powers, Z., Owen, J.G., Reddy, B.V.B., Ternei, M.A., Brady, S.F., 2014. Chemical-biogeographic survey of secondary metabolism in soil. PNAS 111, 3757–3762.

Cimermancic, P., Medema, M.H., Claesen, J., Kurita, K., Wieland Brown, L.C., Mavrommatis, K., Pati, A., Godfrey, P.A., Koehrsen, M., Clardy, J., Birren, B.W., Takano, E., Sali, A., Linington, R.G., Fischbach, M.A., 2014. Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell 158, 412–421.

Cordero, O.X., Wildschutte, H., Kirkup, B., Proehl, S., Ngo, L., Hussain, F., Le Roux, F., Mincer, T., Polz, M.F., 2012. Ecological populations of bacteria act as socially cohesive units of antibiotic production and resistance. Science 337, 1228–1231.

Curtis, T.P., Sloan, W.T., Scannell, J.W., 2002. Estimating prokaryotic diversity and its limits. PNAS 99, 10494–10499.

Dejong, C.A., Chen, G.M., Li, H., Johnston, C.W., Edwards, M.R., Rees, P.N., Skinnider, M.A., Webster, A.L.H., Magarvey, N.A., 2016. Polyketide and nonribosomal peptide retro-biosynthesis and global gene cluster matching. Nat Chem Biol 12, 1007–1014.

Donia, M.S., Cimermancic, P., Schulze, C.J., Wieland Brown, L.C., Martin, J., Mitreva, M., Clardy, J., Linington, R.G., Fischbach, M.A., 2014. A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell 158, 1402–1414.

Doroghazi, J.R., Albright, J.C., Goering, A.W., Ju, K.-S., Haines, R.R., Tchalukov, K.A., Labeda, D.P., Kelleher, N.L., Metcalf, W.W., 2014. A roadmap for natural product discovery based on large-scale genomics and metabolomics. Nat Chem Biol 10, 963–968.

Elsas, J.D. van, Chiurazzi, M., Mallon, C.A., Elhottovā, D., Krištůfek, V., Salles, J.F., 2012. Microbial diversity determines the invasion of soil by a bacterial pathogen. PNAS 109, 1159–1164.

Falony, G., Joossens, M., Vieira-Silva, S., Wang, J., Darzi, Y., Faust, K., Kurilshikov, A., Bonder, M.J., Valles-Colomer, M., Vandeputte, D., Tito, R.Y., Chaffron, S., Rymenans, L., Verspecht, C., De Sutter, L., Lima-Mendez, G., D'hoe, K., Jonckheere, K., Homola, D., Garcia, R., Tigchelaar, E.F., Eeckhaudt, L., Fu, J., Henckaerts, L., Zhernakova, A., Wijmenga, C., Raes, J., 2016. Population-level analysis of gut microbiome variation. Science 352, 560–564.

Friedrich, A.B., Fischer, I., Proksch, P., Hacker, J., Hentschel, U., 2001. Temporal variation of the microbial community associated with the mediterranean sponge Aplysina aerophoba. FEMS Microbiology Ecology 38, 105–113.

Grindberg, R.V., Ishoey, T., Brinza, D., Esquenazi, E., Coates, R.C., Liu, W., Gerwick, L., Dorrestein, P.C., Pevzner, P., Lasken, R., Gerwick, W.H., 2011. Single Cell Genome Amplification Accelerates Identification of the Apratoxin Biosynthetic Pathway from a Complex Microbial Assemblage. PLOS ONE 6, e18565.

Kang, H.-S., Charlop-Powers, Z., Brady, S.F., 2016. Multiplexed CRISPR/Cas9- and TAR-Mediated Promoter Engineering of Natural Product Biosynthetic Gene Clusters in Yeast. ACS Synth Biol 5, 1002–1010.

Kersten, R.D., Yang, Y.-L., Xu, Y., Cimermancic, P., Nam, S.-J., Fenical, W., Fischbach, M.A., Moore, B.S., Dorrestein, P.C., 2011. A mass spectrometry–guided genome mining approach for natural product peptidogenomics. Nat Chem Biol 7, 794–802.

Kersten, R.D., Ziemert, N., Gonzalez, D.J., Duggan, B.M., Nizet, V., Dorrestein, P.C., Moore, B.S., 2013. Glycogenomics as a mass spectrometry-guided genome-mining method for microbial glycosylated molecules. Proc Natl Acad Sci U S A 110, E4407-4416.

Knerr, P.J., van der Donk, W.A., 2012. Discovery, biosynthesis, and engineering of lantipeptides. Annu Rev Biochem 81, 479–505.

Medema, M.H., Breitling, R., Bovenberg, R., Takano, E., 2011. Exploiting plug-and-play synthetic biology for drug discovery and production in microorganisms. Nat Rev Microbiol 9, 131–137.

Medema, M.H., Fischbach, M.A., 2015. Computational approaches to natural product discovery. Nat Chem Biol 11, 639–648.

Medema, M.H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J.B., Blin, K., de Bruijn, I., Chooi, Y.H., Claesen, J., Coates, R.C., Cruz-Morales, P., Duddela, S., Düsterhus, S., Edwards, D.J., Fewer, D.P., Garg, N., Geiger, C., Gomez-Escribano, J.P., Greule, A., Hadjithomas, M., Haines, A.S., Helfrich, E.J.N., Hillwig, M.L., Ishida, K., Jones, A.C., Jones, C.S., Jungmann, K., Kegler, C., Kim, H.U., Kötter, P., Krug, D., Masschelein, J., Melnik, A.V., Mantovani, S.M., Monroe, E.A., Moore, M., Moss, N., Nützmann, H.-W., Pan, G., Pati, A., Petras, D., Reen, F.J., Rosconi, F., Rui, Z., Tian, Z., Tobias, N.J., Tsunematsu, Y., Wiemann, P., Wyckoff, E., Yan, X., Yim, G., Yu, F., Xie, Y., Aigle, B., Apel, A.K., Balibar, C.J., Balskus, E.P., Barona-Gómez, F., Bechthold, A., Bode, H.B., Borriss, R., Brady, S.F., Brakhage, A.A., Caffrey, P., Cheng, Y.-Q., Clardy, J., Cox, R.J., De Mot, R., Donadio, S., Donia, M.S., van der Donk, W.A., Dorrestein,

P.C., Doyle, S., Driessen, A.J.M., Ehling-Schulz, M., Entian, K.-D., Fischbach, M.A., Gerwick, L., Gerwick, W.H., Gross, H., Gust, B., Hertweck, C., Höfte, M., Jensen, S.E., Ju, J., Katz, L., Kaysser, L., Klassen, J.L., Keller, N.P., Kormanec, J., Kuipers, O.P., Kuzuyama, T., Kyrpides, N.C., Kwon, H.-J., Lautru, S., Lavigne, R., Lee, C.Y., Linquan, B., Liu, X., Liu, W., Luzhetskyy, A., Mahmud, T., Mast, Y., Méndez, C., Metsä-Ketelä, M., Micklefield, J., Mitchell, D.A., Moore, B.S., Moreira, L.M., Müller, R., Neilan, B.A., Nett, M., Nielsen, J., O'Gara, F., Oikawa, H., Osbourn, A., Osburne, M.S., Ostash, B., Payne, S.M., Pernodet, J.-L., Petricek, M., Piel, J., Ploux, O., Raaijmakers, J.M., Salas, J.A., Schmitt, E.K., Scott, B., Seipke, R.F., Shen, B., Sherman, D.H., Sivonen, K., Smanski, M.J., Sosio, M., Stegmann, E., Süssmuth, R.D., Tahlan, K., Thomas, C.M., Tang, Y., Truman, A.W., Viaud, M., Walton, J.D., Walsh, C.T., Weber, T., van Wezel, G.P., Wilkinson, B., Willey, J.M., Wohlleben, W., Wright, G.D., Ziemert, N., Zhang, C., Zotchev, S.B., Breitling, R., Takano, E., Glöckner, F.O., 2015. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol 11, 625–631.

Medema, M.H., Paalvast, Y., Nguyen, D.D., Melnik, A., Dorrestein, P.C., Takano, E., Breitling, R., 2014. Pep2Path: Automated Mass Spectrometry-Guided Genome Mining of Peptidic Natural Products. PLOS Computational Biology 10, e1003822.

Mendes, R., Kruijt, M., de Bruijn, I., Dekkers, E., van der Voort, M., Schneider, J.H.M., Piceno, Y.M., DeSantis, T.Z., Andersen, G.L., Bakker, P.A.H.M., Raaijmakers, J.M., 2011. Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science 332, 1097–1100.

Mohimani, H., Kersten, R.D., Liu, W.-T., Wang, M., Purvine, S.O., Wu, S., Brewer, H.M., Pasa-Tolic, L., Bandeira, N., Moore, B.S., Pevzner, P.A., Dorrestein, P.C., 2014. Automated genome mining of ribosomal peptide natural products. ACS Chem Biol 9, 1545–1551.

Montalbán-López, M., van Heel, A.J., Kuipers, O.P., 2017. Employing the promiscuity of lantibiotic biosynthetic machineries to produce novel antimicrobials. FEMS Microbiology Reviews 41, 5–18.

Reddy, B.V.B., Milshteyn, A., Charlop-Powers, Z., Brady, S.F., 2014. eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. Chem Biol 21, 1023–1033.

Shao, Z., Rao, G., Li, C., Abil, Z., Luo, Y., Zhao, H., 2013. Refactoring the silent spectinabilin gene cluster using a plug-and-play scaffold. ACS Synth Biol 2, 662–669.

Sharon, I., Kertesz, M., Hug, L.A., Pushkarev, D., Blauwkamp, T.A., Castelle, C.J., Amirebrahimi, M., Thomas, B.C., Burstein, D., Tringe, S.G., Williams, K.H., Banfield, J., 2015. Accurate, multi-kb reads resolve complex populations and detect rare microorganisms. Genome Res. gr.183012.114.

Skinnider, M.A., Johnston, C.W., Edgar, R.E., Dejong, C.A., Merwin, N.J., Rees, P.N., Magarvey, N.A., 2016. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. Proc Natl Acad Sci U S A 113, E6343–E6351.

Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., Voigt, C.A., 2016. Synthetic biology to access and expand nature's chemical diversity. Nat Rev Microbiol 14, 135–149.

Trindade-Silva, A.E., Rua, C., Silva, G.G.Z., Dutilh, B.E., Moreira, A.P.B., Edwards, R.A., Hajdu, E., Lobo-Hajdu, G., Vasconcelos, A.T., Berlinck, R.G.S., Thompson, F.L., 2012. Taxonomic and functional microbial signatures of the endemic marine sponge Arenosclera brasiliensis. PLoS One 7, e39905.

van Heel, A.J., de Jong, A., Montalbán-López, M., Kok, J., Kuipers, O.P., 2013. BAGEL3: Automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. Nucleic Acids Res 41, W448-453.

Watrous, J., Roach, P., Alexandrov, T., Heath, B.S., Yang, J.Y., Kersten, R.D., van der Voort, M., Pogliano, K., Gross, H., Raaijmakers, J.M., Moore, B.S., Laskin, J., Bandeira, N., Dorrestein, P.C., 2012. Mass spectral molecular networking of living microbial colonies. Proc Natl Acad Sci U S A 109, E1743-1752.

Wilson, M.C., Mori, T., Rückert, C., Uria, A.R., Helf, M.J., Takada, K., Gernert, C., Steffens, U.A.E., Heycke, N., Schmitt, S., Rinke, C., Helfrich, E.J.N., Brachmann, A.O., Gurgui, C., Wakimoto, T., Kracht, M., Crüsemann, M., Hentschel, U., Abe, I., Matsunaga, S., Kalinowski, J., Takeyama, H., Piel, J., 2014. An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature 506, 58–62.

Yamanaka, K., Reynolds, K.A., Kersten, R.D., Ryan, K.S., Gonzalez, D.J., Nizet, V., Dorrestein, P.C., Moore, B.S., 2014. Direct cloning and refactoring of a silent lipopeptide biosynthetic gene cluster yields the antibiotic taromycin A. Proc Natl Acad Sci U S A 111, 1957–1962.

Zhernakova, A., Kurilshikov, A., Bonder, M.J., Tigchelaar, E.F., Schirmer, M., Vatanen, T., Mujagic, Z., Vila, A.V., Falony, G., Vieira-Silva, S., Wang, J., Imhann, F., Brandsma, E., Jankipersadsing, S.A., Joossens, M., Cenit, M.C., Deelen, P., Swertz, M.A., LifeLines cohort study, Weersma, R.K., Feskens, E.J.M., Netea, M.G., Gevers, D., Jonkers, D., Franke, L., Aulchenko, Y.S., Huttenhower, C., Raes, J., Hofker, M.H., Xavier, R.J., Wijmenga, C., Fu, J., 2016. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science 352, 565–569.

Ziemert, N., Alanjary, M., Weber, T., 2016. The evolution of genome mining in microbes - a review. Nat Prod Rep 33, 988–1005.

CHAPTER 2

Microbial and volatile profiling of soils suppressive to Fusarium culmorum of wheat

Authors:

Adam Ossowicki*, Vittorio Tracanna*, Marloes L. C. Petrus, Gilles van Wezel, Jos M. Raaijmakers, Marnix H. Medema# and Paolina Garbeva#

**These authors contributed equally to this work*

#: corresponding author

Published as “Microbial and volatile profiling of soils suppressive to Fusarium culmorum of wheat” in *Proceedings of the Royal Society B*, 287(1921): 20192527 (2020).

2.1 - Abstract

In disease-suppressive soils, microbiota protect plants from root infections. Bacterial members of this microbiota have been shown to produce specific molecules that mediate this phenotype. To date, however, studies have focused on individual suppressive soils and the degree of natural variability of soil suppressiveness remains unclear. Here, we screened a large collection of field soils for suppressiveness to *F. culmorum* using wheat (*Triticum aestivum*) as a model host plant. A high variation of disease suppressiveness was observed, with 14% showing a clear suppressive phenotype. The microbiological basis of suppressiveness to *F. culmorum* was confirmed by gamma-sterilization and soil transplantation. Amplicon sequencing revealed diverse bacterial taxonomic compositions and no specific taxa were found exclusively enriched in all suppressive soils. Nonetheless, co-occurrence network analysis revealed that two suppressive soils shared an overrepresented bacterial guild dominated by various *Acidobacteria*. In addition, our study revealed that volatile emission may contribute to suppression, but not for all suppressive soils. Our study raises new questions regarding the possible mechanistic variability of disease suppressive phenotypes across physicochemically different soils. Accordingly, we anticipate that larger-scale soil profiling, along with functional studies, will enable a deeper understanding of disease suppressive microbiomes.

2.2 - Introduction

The phenomenon of soil disease suppressiveness has been recognized for almost a century and was first defined by Cook and Baker (Deacon et al., 1984) as soils where a particular soil-borne disease does not develop, despite the presence of the virulent pathogen, a susceptible host and favorable conditions for disease development. Physical and chemical properties of the soil can play a role in this phenomenon, but in many cases disease suppressiveness is microbial in nature (Andrade et al., 2011, Wiseman et al., 1996, Weller et al., 2002, Cha et al., 2016).

Two types of soil suppressiveness can be distinguished, namely general and specific. General suppression is effective against a range of pathogens, whereas specific suppression operates against only one or a few of them. General soil suppressiveness is a result of the activity of the overall soil microbial community, whereas specific soil suppressiveness is due to the concerted action of specific microbial genera that interfere at some stage of the life cycle of the soil-borne pathogen. Specific soil suppressiveness can be eliminated by selective heat treatments and is transferable to a conducive soil by mixing in a small amount (1-10%) of suppressive soil (Schlatter et al., 2017, Mazzola et al., 2002).

For many suppressive soils, the microorganisms and mechanisms have not been elucidated. Some of the best-studied suppressive soils to date are the take-all-decline (TAD) soils, where root disease of wheat or barley caused by the fungal pathogen *Gaeumanomyces graminis* var. *tritici* is suppressed (Cook et al., 1976, Hornby et al., 1983, Duran et al., 2017). Suppressiveness to take-all disease is, at least in part, due to the enrichment of populations of root-associated *Pseudomonas* spp. producing the antifungal polyketide 2,4-diacetylphlorogucinol (2,4-DAPG) (Raaijmakers et al., 1999, Raaijmakers et al., 1998, Kwak et al., 2011). For *Fusarium*-wilt-suppressive soils, the microbes and mechanisms identified so far involve non-pathogenic *Fusarium oxysporum* and *Pseudomonas* species that, in a complementary manner, compete with pathogenic *F. oxysporum* for carbon and iron (Alabouvette et al., 1986, Sieger-Hertz et al., 2018). Recent studies on soils suppressive to *Rhizoctonia* damping-off disease of sugar beet revealed the involvement of multiple bacterial genera belonging to the *Pseudomonadaceae*, *Streptomycetaceae* and *Burkholderiaceae*. The suppression was linked to the production of antifungal lipopeptide thanamycin by *Pseudomonas* and to the production of volatile metabolites by *Streptomyces* spp. and *Paraburkholderia graminis* (Carrion et al., 2018, Cordovez et al., 2015). Volatiles are low-molecular-mass metabolites involved in long-distance interactions with potent antimicrobial activities (Garbeva et al., 2011, Cho et al., 2017, Ossowichi et al., 2017, Garbeva et al., 2014) For example, the effects of volatiles emitted from 50 agricultural soils on two soil-borne pathogenic fungi (*F. oxysporum, R. solani*) and a plant pathogenic oomycete (*Pythium intermedium*) have recently been studied (van Agtmaal et

al.,2018) and revealed that most soils emit volatiles that inhibit hyphal growth, but the extent of growth inhibition per soil differed strongly for the three pathogens.

To date, most studies on suppressive soils have been limited to a single field soil. Except for the case of the TAD soils (Landa et al., 2002), it is currently unclear how widespread suppressiveness to specific pathogens is as a biological phenomenon. Furthermore, it is not established yet whether suppression is mediated by one or multiple taxa and mechanisms across various suppressive soils. Finally, the role of volatiles in soil suppressiveness in the presence of both the pathogen and the host plant is yet unexplored.

In this study, we screened a large collection of field soils for suppressiveness to *Fusarium culmorum*, a ubiquitous soil-borne fungus causing foot rot, root rot and *Fusarium* head blight of different cereals, in particular wheat and barley (Wiese 1987). Using wheat (*Triticum aestivum*) as the host plant, we found *F. culmorum* suppressiveness in 4 out of 28 soils.

We hypothesised that 1) soils suppressive to *F. culmorum* have similar rhizobacterial community compositions with specific enriched taxa, 2) volatile compounds contribute to disease suppression against *F. culmorum,* and 3) disease-suppressive soils have similar volatile profiles. We anticipated that specific rhizobacterial taxa and volatiles correlate with soil suppressiveness against *F. culmorum*.

However, our comparative analysis of bacterial rhizosphere microbiome composition of suppressive and non-suppressive (conducive) soils, along with volatile profiling, indicates that the phenomenon may be mediated by different rhizobacterial genera across these soils and that volatiles may contribute to the suppressive phenotype only for a limited number of soils.

2.3 - Materials and Methods

2.3.1 - Soil collection

In order to examine soil disease suppressiveness to *F. culmorum*, 28 sites in the Netherlands and Germany were chosen based on information on soil type and crop rotation. The sites included 25 arable fields, 2 pastures and 1 forest. The arable fields included sites with wheat present in crop rotation over the last three years and sites without wheat present in the available history of the field.

Soil samples were collected in the period from January to April 2017 from 3-meter squares located in the middle of each agricultural field/pasture. In this area, top soil cores of approximately 30 cm deep were extracted. Samples were air-dried in room temperature, homogenized, sieved through 4 mm sieve and stored at 4°C. Heavy clay-type soils were additionally flaked using a jaw-crusher (Type BB-1, Retsch, Germany) after drying. Soil physical and chemical parameters were measured as described in Supplementary material part I.

2.3.2 - Wheat growth conditions and pathogen inoculation

All the greenhouse experiments were performed in growth cabinets (MC 1750 VHO-EVD, Snijders Labs) in 20°C day and night, photoperiod 12h day /12h night and 60 % relative humidity. In all the experiments, surface-sterilized and pre-germinated wheat seeds (JB Asano from Agrifirm, Netherlands) were used. Plants were watered every second day and weekly supplemented with 0.5 Hoagland solution (1ml per 80 cc of the soil, 0.5M $Ca(NO_3)_2 \cdot 4H_2O$, 1M KNO_3, 1M KH_2PO_4, 0.5M $MgSO_4 \cdot 7H_2O$ and 98.6 mM ferric EDTA).

As a standard substrate, we used a dune soil collected near Bergharen, the Netherlands (BS) (Schultz-Bohm et al., 2015). Before use, BS was air-dried, sieved and gamma-sterilized (Synergy Health Ede B.V., the Netherlands). Sub-samples of eight soils from the collection were gamma-sterilized as well in the same conditions for later use.

Prior to screening for disease suppressiveness, all soils were “activated” in order to induce microbial activity in a so-called microbiome activation step. For this, soils hosted the growth of wheat for two weeks in the conditions indicated above. After this, plants, including their whole root system, were removed and the remaining soil with activated microbiome was mixed properly.

The soil-borne pathogen *Fusarium culmorum* PV (Blom et al., 2011) was grown on 1/2 PDA media (Oxoid, the Netherlands) and maintained in continuous culture. For plant inoculation, the fungus was transferred to 1/4 PDA and grown for 2 weeks in 20°C. After the incubation, 6mm plugs were cut from the border zone of *Fusarium* hyphae and mixed with the growth substrate (1 plug per 10 cc). Controls without pathogen were inoculated with sterile 1/4 PDA plugs. At the end of the experiment, disease symptoms were assessed and plants were used for rhizosphere DNA extraction and sequencing analysis (see Supplementary material part I).

2.3.3 - Disease suppressiveness screening

Initial disease suppressiveness screening was performed in propagation trays containing 140 single pots (Teku, the Netherlands) each containing 38cc of soil per pot. In order to minimize the physicochemical differences between the 28 soils, after the activation step (described in paragraph above) the natural soil was mixed 2:1:1 in volume with sterile BS soil and sterile vermiculite (Agra-vermiculite, the Netherlands). The soil was inoculated with *Fusarium* or mock-inoculated with agar plugs, and one seedling was placed in the center of each pot. Plants were grown in ten replicates per treatment and, after three weeks, disease symptoms were assessed (see Supplementary material part I). As a control, we used BS mixed with vermiculite 3:1 in volume.

2.3.4 - Confirmation of disease suppressiveness with soils sterilization

Four suppressive and four conducive soils were selected to confirm the previously observed levels of suppressiveness. The confirmation test was performed identically

as the screening test but the system was scaled up to 380 cc soil per pot (7x7x8 cm, Teku, the Netherlands) with three plants in each. An additional treatment was included - sterilized soil inoculated with pathogen. Sterile BS, vermiculite mix was used as a control. All the treatments had an additional control without pathogen.

2.3.5 - Transplantation assay

To assess whether soil suppressiveness is transferable, we mixed natural soils in proportions 1:9 and 3:7 in volume with a conducive substrate consisting of sterile BS and vermiculite. This experiment was also performed with 380 cc soil per pot (7x7x8 cm), each containing three plants, in treatments with and without the addition of the pathogen. Sterile BS and vermiculite mixture were used as a control.

2.3.6 - Sequencing and bioinformatics analysis

Rhizosphere DNA extraction, sequencing, 16S rRNA amplicon data processing and analysis are described in Supplementary material part I.

2.3.7 - Effects of soil-emitted volatiles on fungal growth

Assays were performed as previously described by Garbeva et al., 2014 with small modifications, see Supplementary material part I.

2.3.8 - Effects of soil-emitted volatile compounds on disease suppression

The eight soils selected in the suppressiveness screening were tested for their ability to induce soil suppressiveness via volatiles. For this, the modified method described by Park et al., 2015 was used. Briefly, wheat plants growing in conducive soil mixed with *F. culmorum* plugs were exposed to volatiles emitted by soil present in the compartment below the pot. To avoid any physical contact, compartments were separated and sealed with sterile nylon mesh (Sefar, Switzerland) and paper medical tape as described above. The scheme of the system is shown on figure 2-

3B. Four pots per treatment with three plants per pot were grown for three weeks; afterwards, disease symptoms were assessed.

2.3.9 - Volatile trapping and GC-MS analysis

Volatile compounds were trapped using steel trap containing 150 mg Tenax TA and 150 mg Carbopack B (Markes International Ltd, Llantrisant, UK) and measured using GC-QTOF system. Statistical data analysis was performed using MetaboAnalyst 4.0 software (Chong et al., 2018). For details of Volatile trapping and GC-MS analysis, see Supplementary material part I.

2.4 - Results

2.4.1 - Identification of soils suppressive to *F. culmorum*

A collection of 28 soils of diverse geographical origins (Figure 2-1A), soil types and agricultural histories (table S1) was screened for disease suppressiveness against *F. culmorum* in greenhouse pot experiments. The disease symptoms of the plants grown in each of the soils were examined three weeks after pathogen inoculation. Overall, high variation in disease suppressiveness was observed between the 28 soils, while being largely consistent between replicates. The disease symptoms across the collection varied from mild or no infection to severe disease. Four soils (S01, S03, S11 and S28 - yellow color, figure 2-1) showed the lowest level of disease severity with an average score below 0.5 and were considered suppressive (figure 2-1). Four soils revealed high disease severity with average disease symptoms above 1.5 (S08, S14, S15 and S17 – black color, figure 2-1) and were considered conducive.

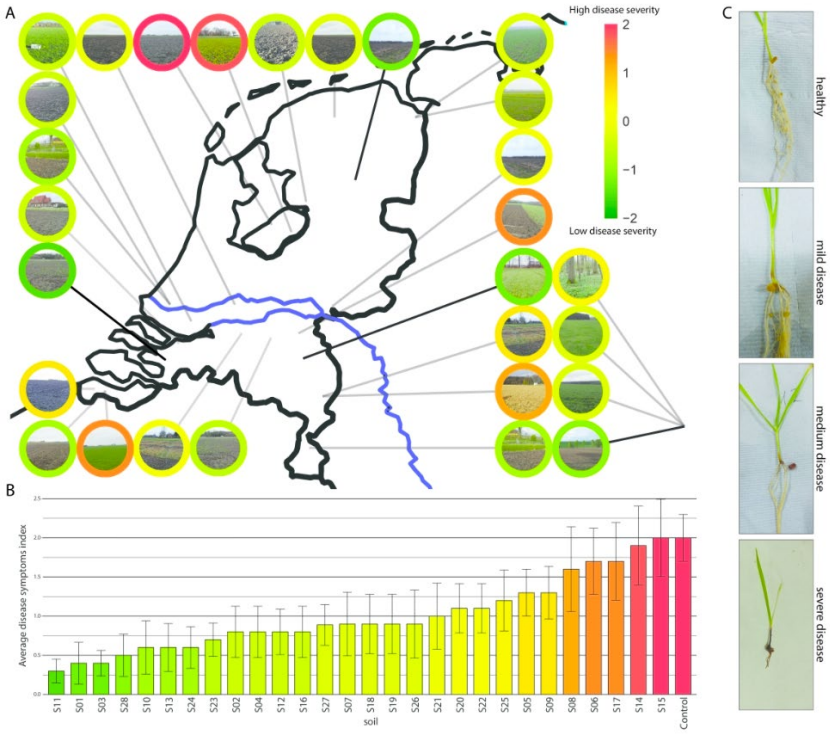


Figure 2-1. Results of the phenotypic screening towards soil suppressiveness to F. culmorum in 28 soils. (A) Map showing locations of 28 sampling sites and the results of the screening. The color of the outer circle around the photos represents the z-score transformed average disease symptoms value. Low disease severity indicates soil suppressiveness (yellow color). (B) Bar graph showing average disease symptoms indexes with standard errors.

2.4.2 - Disease suppression is independent of soil physicochemical properties

In order to investigate phenotypic variation across multiple soil types, our collection included soils from 25 arable fields, 2 pastures and 1 forest, representing different soil types, ranging from sands to heavy clays, with diverse pH (5.3 to 7.8) and C/N ratio (8.8 to 17.5). All physicochemical parameters and field history are summarized in Table S1. No clear correlations between the level of disease suppressiveness and physicochemical parameters and field history were found. Based on canonical correspondence analysis of the 28 soils, there was no separation between disease-suppressive and conducive soils (figure S1, yellow and black dots accordingly). Only weak Pearson correlations were found between physicochemical parameters and

disease severity (figure S2). Also, no relevant correlations were found between disease severity and soil pH and C/N ratio (0.24 and 0.25 respectively).

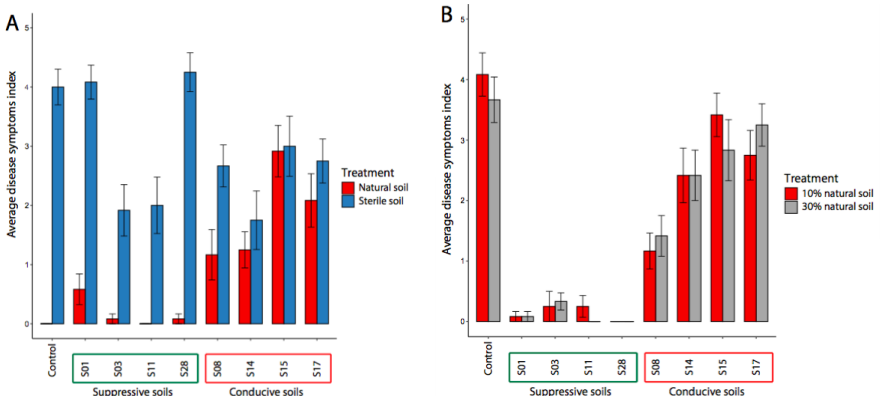


Figure 2-2. **Disease symptoms observed in wheat inoculated with *Fusarium culmorum* grown in eight prioritized soils.** (A) Natural and gamma-sterilized soil with sterile BS soil/vermiculite mix as a control (B) 10% and 30% in volume of natural soil mixed with standardized sterile substrate or with sterile BS soil as a control. The bar indicates the average of the disease symptoms index, with the error bars representing the standard error.

2.4.3 - Disease suppressiveness has a microbial basis

Based on the results from the initial screening, a set of eight soils was selected for confirmation assays and further analysis. This included the four soils with the highest level of disease suppressiveness (S01, S03, S11 and S28) and the four conducive soils showing the most contrasting phenotypes (S08, S14, S15 and S17). For the four suppressive soils, low average disease indices were again observed (figure 2-2A, red bars) but after gamma sterilization, disease indices increased significantly (figure 2-2A, blue bars). For the four conducive soils, except for S08, gamma sterilization did not substantially enhance disease severity. Furthermore, the results revealed that suppressiveness to *F. culmorum* is transferable (figure 2-2B). Mixing 10% or 30% of the suppressive soil (S01, S03, S11 and S28) in a conducive background soil transferred suppressiveness. There was no significant difference in the level of suppression between the transfer of 10% and 30% suppressive soil.

Collectively, our data suggest that the microbial community contributes to the suppressive phenotype.

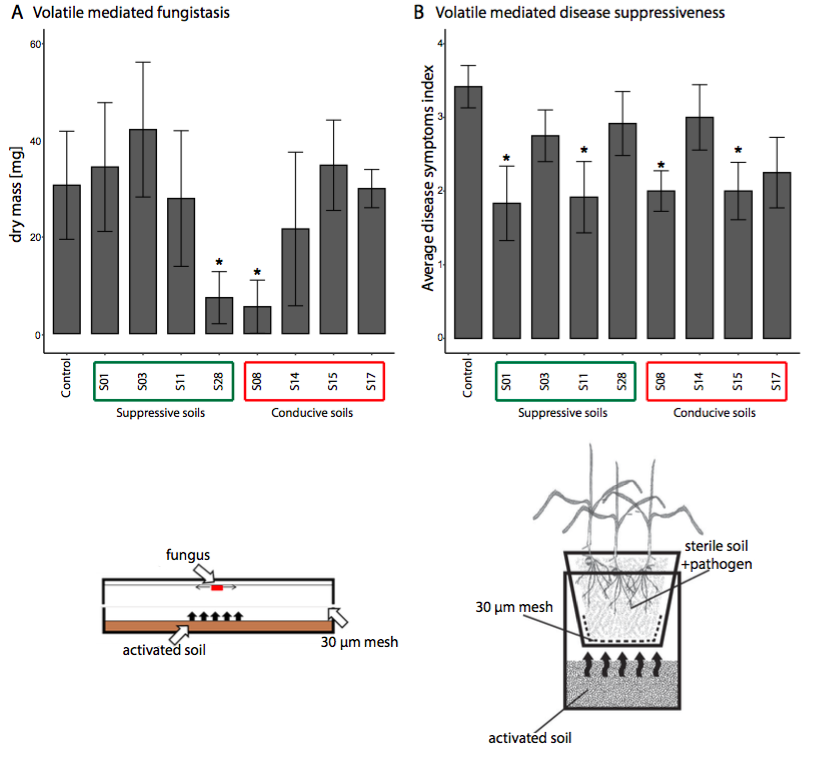


Figure 2-3. The effect of the volatiles emitted by eight prioritized soils on *Fusarium culmorum* growth (fungistasis) and disease suppression. (A) Average dry mass of the fungus with standard deviation, the statistically significant differences between treatments and control (based on ANOVA and Tuckey post-hoc test $p<0.05$) are indicated by asterisk. (B) Average symptoms index with standard error. The statistically significant differences between treatments and control (based on chi-square test $p<0.05$) are indicated by asterisk.

2.4.4 - Volatile-mediated inhibition of fungal growth and disease suppression is observed for suppressive and conducive soils

To determine the role of volatile compounds in disease suppression and antifungal activity, the eight selected soils were tested in two experimental systems: i) hyphal growth of *Fusarium* on artificial media exposed to soil volatiles, and ii) plants growing in conducive soil inoculated with the pathogen exposed to soil volatiles (figure 2-4 A, B and Supplementary material part I). The first assay revealed that only two soils (S28 suppressive and S08 conducive) emitted volatiles that significantly reduced growth of the fungus compared to the control. When plants were exposed to soil volatiles, disease suppression was observed for four out of eight soils (figure 2-4 B). Again, these included both suppressive (S01, S11) and conducive (S08, S15) soils. Subsequent analysis of volatile profiles emitted by these eight soils did not reveal clear separation between suppressive and conducive soils (figure S3). Interestingly, several suppressive and conducive soils (S03, S08, S11 and S14) revealed very similar volatile profiles. These results suggest that volatiles are not a common mechanism of soil suppressiveness to *F. culmorum*.

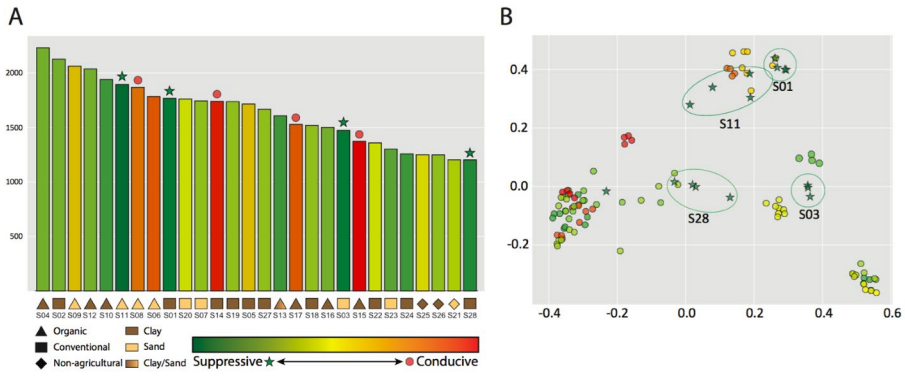


Figure 2-4. Characteristics of rhizosphere bacterial communities across all 28 soils. (A) Bar plot representing the alpha diversity using the rarefied unique ASV counts. Samples are sorted according to their alpha diversity score and color-coded according to suppressive phenotype. (B) PCoA based on Unweighted UniFrac distance between samples. Circles highlight the suppressive sample replicates clustered together.

2.4.5 - Microbial profiling of *F. culmorum*-suppressive soils

To investigate possible links between the rhizobacterial community composition and the disease suppressive phenotype, extensive 16S sequencing was performed for all 28 soils (see Supplementary material part I). The rhizobacterial communities showed large variation between samples and even between replicates of the same samples, as can be seen from the inter- and intra-sample Jaccard similarity of 0.056 (sd = 0.033) and 0.346 (sd = 0.057) respectively. Based on the alpha- and beta-diversity, there were no significant community differences between suppressive and conducive soils. Beta-diversity for all sample pairs was calculated with unweighted UniFrac, which shows consistent grouping for soil sample replicates (figure 2-4B). Subsequent PCoA analysis of the rhizobacterial community composition of the suppressive and conducive soils indicated that the different suppressive soils have diverse taxonomic compositions and did not group together (figure 2-4B). This was also the case when calculating community diversity of the samples within specific taxonomic groups (Supplementary figures S4). As for alpha diversity, Wilcoxon rank-sum test showed no significant association between observed ASVs ($p=0.74$) or Shannon diversity ($p=0.07$) and soil suppressiveness. The suppressive soils included both the most and the least diverse communities. The soils from organic farms were associated with higher bacterial diversity compared to the soils from conventional farms and to the non-agricultural soils (figure 2-4A). Again, there was no consistent correlation between soil suppressiveness and alpha- or beta-diversity.

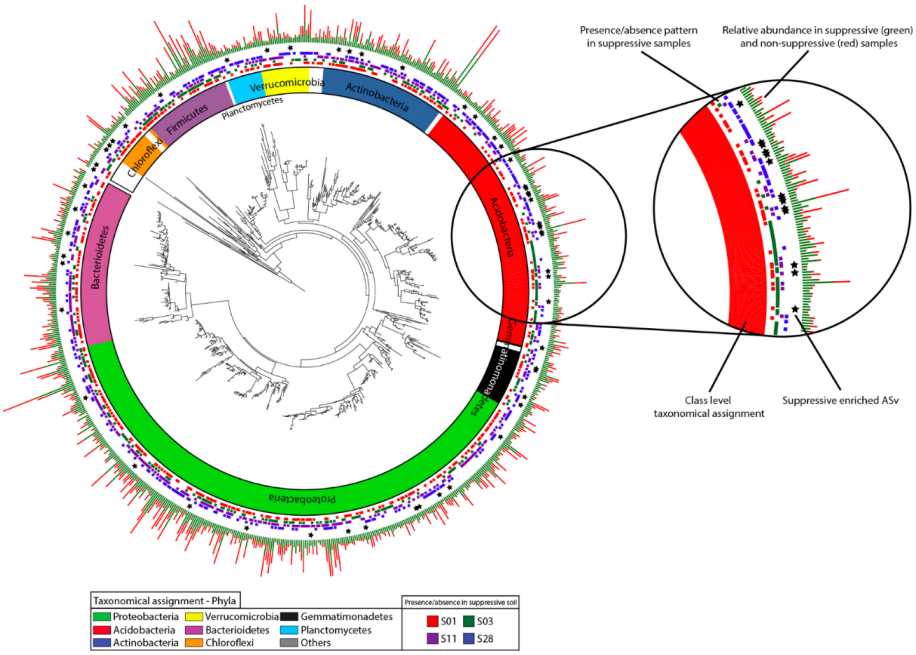


Figure 2-5. Phylogenetic tree of all ASVs consistently detected in one or more suppressive rhizosphere samples. The rings, from the inside to the outside, represent 1) the taxonomical assignment of the ASVs at the phylum level, 2) the presence/absence patterns of ASVs in the four different suppressive soils, 3) indications (with starts) of ASVs strongly enriched in suppressive samples, and 4) the average cumulative normalized abundance of each ASV in suppressive (green) versus non-suppressive (red) samples. The results show that all ASVs enriched in suppressive samples are specific to a subset of the suppressive soils, while none of the ASVs that occur throughout all suppressive soils are significantly enriched in suppressive versus non-suppressive soils.

To establish whether the disease-suppressive phenotypes were mediated by specific rhizobacterial taxon or by multiple taxa, 16S amplicon data were analyzed in more depth. To reveal if one or more individual abundant amplicon sequence variants (ASVs) could be associated with disease suppressiveness, we inspected the presence-absence patterns of all ASVs that were significantly enriched in suppressive soils (figure 2-5). No individual ASV appeared to be exclusively present in all suppressive samples compared to the conducive samples (figure S5). Random forest classifiers trained on their taxonomic profiles could not predict the suppressive soil phenotype in a leave-one-out analysis. The results were not significantly different when the same analysis was performed using OTU-level clustering of ASVs (at 97% identity).

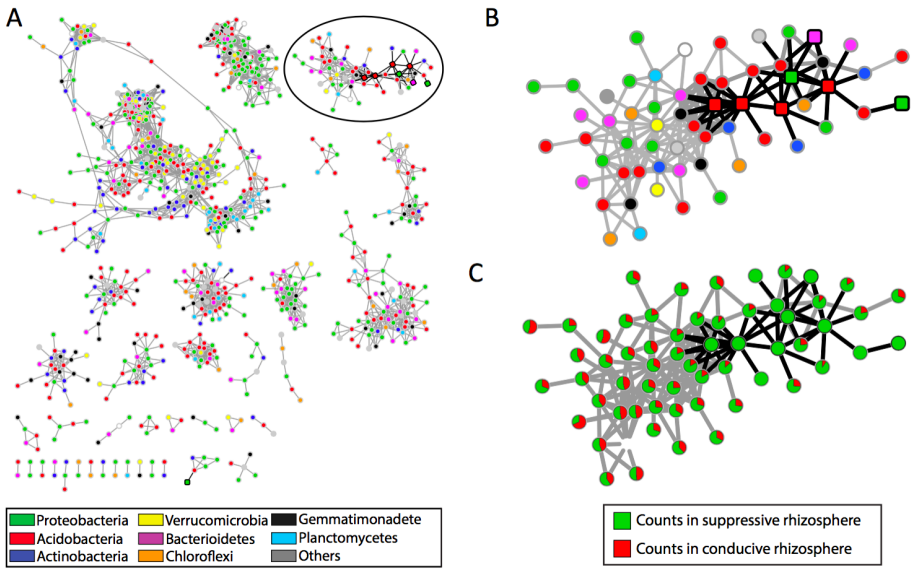


Figure 2-6. Co-occurrence network analysis of rhizosphere bacteria across samples. A) Complete network of correlated ASVs using Spearman correlation. Nodes represent individual ASVs; edges link ASVs which had a correlation score above a given threshold. The node color indicates the taxonomical annotation at the phylum level. Edges of ASVs that are significantly enriched in suppressive samples are highlighted black as the respective nodes. B) Zoomed in view of a subnetwork of interest, which shows a number of ASVs associated with the suppressive phenotype indicated by squares with black border, along with its taxonomical distribution. C) Normalized counts ratio between conducive (yellow) and suppressive (black) samples for ASVs from the network in panel B.

2.4.6 - Co-occurrence network analysis indicates bacterial communities enriched in rhizosphere microbiomes of suppressive soils

To investigate whether specific combinations of ASVs are associated with disease suppressiveness (in subsets of the four suppressive soils), we calculated Spearman correlations between ASV occurrences across all 28 soils and constructed a co-occurrence network from the most highly correlating ASVs in the correlation matrix (see Supplementary material part I). The resulting network (figure 2-6A) consists of 928 nodes (comprising 21% of all 4,322 ASVs), distributed across 37 subnetworks. The ASVs in the network that were significantly enriched in suppressive soils are indicated by squares with black border in figure 2-6B. The subnetworks are taxonomically diverse, with no network being taxonomically homogeneous, even at the phylum level. Interestingly, one of the connected components was found to be

particularly associated with the suppressive phenotype. Within this connected component, 7 out of the 163 ASVs were strongly enriched in suppressive soils and these were part of a more densely connected subcomponent comprising 60 ASVs that comprised 6 out of the 7 abovementioned ones. Taxonomic assignments of the suppression-associated subnetwork displayed an overrepresentation of ASVs belonging to different Acidobacteria, namely Blastocatellales and members of subgroup 6, compared to the overall community composition ($p=0.0003$, Fisher exact test, figure 2-6B). Moreover, several of these Acidobacteria displayed characteristics of being hub taxa in these subnetworks based on their high betweenness centrality. In addition, 11 of the 60 ASVs in the suppressive-associated network were exclusively present in two suppressive soils (figure 2-6C), 7 of which were taxonomically assigned to Acidobacteria.

2.5 - Discussion

F. culmorum is an economically important fungal plant pathogen that causes disease in many cereal and non-cereal crops. However, little is known about the occurrence and distribution of soil suppressiveness to this pathogen. To our knowledge, this study is the first to screen for disease suppressiveness in a large collection of diverse soils. Although a high variation in the level of disease suppressiveness was observed between the soils, 14% of the tested soils revealed a clear suppressive phenotype. Interestingly, no physicochemical parameter such as soil type, field history, pH, C/N ratio or content and concentration of bioavailable Fe, K, Mg, P, S correlated with the observed suppressive phenotype. In previous work, physicochemical soil parameters such as soil type, moisture, pH, organic matter and microelements content have been more commonly associated with general disease suppressiveness (Hoper et al., 1996, Janvier et al., 2007).

In our study, the microbiological basis of specific suppressiveness to *F. culmorum* was revealed by two independent approaches: gamma-sterilization and soil transplantation eliminated and conferred suppressiveness, respectively, and confirmed that suppressiveness to *F. culmorum* was not linked to soil physicochemical parameters but rather to the soil (micro)biome.

Along with the soil, the plant itself is an important determinant of the structure of soil microbial communities and disease suppressiveness. Hence, the strength of disease suppressiveness is attributed to all players in the tripartite relationship of plant-soil-microorganisms (Garbeva et al., 2004). Considering this, in the present study, all experiments included a “microbiome activation step” by growing wheat for 2 weeks prior screening for disease suppressiveness. The application of plants provides substrates for growth via root exudates and space for the soil microbial community. Several studies have associated disease-suppressive phenotypes to individual bacterial groups based on their relative enrichment compared to a conducive phenotype (Donn et al., 2014, Yin et al., 2013). However, rhizosphere microbiomes are highly diverse (Charlop-Powers et al., 2015) and multiple differences may exist even between physicochemical similar soils. In the present study, 28 different soils were examined and no individual bacterial taxon was found to be exclusively present or enriched in all suppressive soils. Bacterial taxonomic groups that were more prevalent in some of the suppressive soils were not prevalent in other suppressive soils.

Many studies have aimed to understand the relationship between microbial diversity and disease suppressiveness. However, both low (Mehrabi et al., 2016) and high (Garbeva et al., 2004) community diversity have been associated with soil suppressiveness. Here, we found suppressive samples with both high and low community richness, which suggests community diversity is not the key driver of soil suppressiveness to *Fusarium culmorum*. Community evenness is correlated with the phenotype but the strength of this association is not strong enough ($p=0.07$) to warrant speculations on its role in disease suppressiveness.

To our surprise, no ASVs or OTU-level ASV clusters were found to be shared uniquely between the microbial communities of all four suppressive soils, or to be differentially abundant across all these four soils. The fact that random forest classifiers were unable to accurately predict suppressive phenotypes in a leave-one-out analysis of the soils suggests that i) the suppressive phenotypes are mediated by different taxonomic groups across the different suppressive soils and/or ii) that rhizobacteria do not play a major role in suppressiveness to *F. culmorum*. Many

functional elements that are known to be able to drive disease suppression (such as biosynthetic gene clusters for secondary metabolites) are often strain-specific and are frequently transferred horizontally across species (Medema et al., 2014). Hence, it is still possible that the same or similar functional elements could drive the suppressiveness across all four soils, while being undetectable by 16S sequencing due to its limited resolution or due to the elements being encoded in the genomes of diverse bacterial taxa.

Representatives of a range of bacterial groups can carry out functions that result in the suppression of soil-borne diseases. For example, several *Bacillus*, *Pseudomonas*, *Streptomyces* or *Flavobacterium* species are well known to play role in suppression of various soil-borne plant pathogens (Donn et al., 2014, Garbeva et al., 2006, Chapelle et al., 2016, Kwak et al., 2018). However, various previous studies indicate that many of these bacteria reveal antimicrobial activities only as results of interspecific interaction networks (Tyc et al., 2017, Tyc et al., 2014, Traxler et al., 2013). The correlation-based network analysis performed here revealed complex inter-sample connections between individual bacterial taxa that are likely interacting either directly or indirectly, based on their observed co-occurrence. One of the network components provided insights into a bacterial guild that is potentially associated with disease suppressiveness to *Fusarium culmorum*. Multiple ASVs were found to be exclusively present in two distinct disease suppressive microbiomes, while additionally being strongly correlated with another sub-community of ASVs that also consistently occurred in a conducive microbiome. Interestingly, most of these ASVs belonged to *Acidobacteria*. We observe an over-representation of *Acidobacteria* in the suppression associated network when compared to the general community composition (Fisher exact test, p=0.0003). This phylum has previously been associated with *Rhizoctonia solani* bare patch [34] and has been shown to harbor diverse species with a large specialized metabolic potential that could be involved in interactions with fungi (Crits-Christoph et al., 2018). Based on their specific enrichment, the identified bacterial guild might represent the suppressive core at the base of the phenotype for two of the four suppressive samples (soils S01 and S11). Further shotgun sequencing efforts as well as microbiological and biochemical analysis are needed to clarify the functional

roles of these organisms. Furthermore, one should bear in mind that in addition to the bacterial community structure, the expression of functional genes conferring suppressiveness to soil-borne pathogens might be dependent on interactions with other soil microorganisms such as fungi, protists and viruses.

Frequently, the germination and growth of plant pathogenic fungi are negatively affected by direct or indirect contact with soils (without the presence of a host plant). This phenomenon is named soil fungistasis (Watson et al., 1972) and is often associated with general disease suppressiveness. Whereas abiotic soil factors can be involved, the major cause of fungistasis is biological since it is strongly reduced after soil sterilization, which is analogous to the soil disease suppressiveness. Recent studies revealed that microbial volatiles play important roles in both soil fungistasis as well as disease suppressiveness (van Agtmaal et al., 2018, Effmert et al., 2012). Our study did not reveal any congruence between volatile-mediated soil fungistasis and disease suppressiveness. Volatile-mediated soil suppressiveness was observed only with four soils, including both two suppressive and two conducive soils. This indicates that, even though volatile emission is one of the factors that may contribute to disease suppressiveness to *F. culmorum*, it is not a general one. Apparently, other (complementary) factors and mechanisms are needed to develop a full protective potential against *F. culmorum*. Alternatively, it might be that suppressiveness across the four soils is mediated by different mechanisms, involving volatiles in some cases but not in others. This could also explain the lack of commonly enriched taxa across the four suppressive soils.

In conclusion, we discovered several agricultural soils that protect wheat plants from *Fusarium culmorum* infections through a microbial component. Moreover, these suppressive soils revealed different bacterial taxonomic patterns and diversity, as well as variable degrees of antifungal volatile emissions. These observations reject the hypothesis that specific rhizobacterial taxa and specific volatiles correlate with suppressiveness of *F. culmorum.*

Co-occurrence network analysis suggested that two of the suppressive samples share a similar microbial basis of the phenotype through a uniquely overrepresented

bacterial guild dominated by *Acidobacteria*. Of course, taxonomic profiling alone cannot provide definitive answers on the actual biological mechanisms responsible for the suppressive phenotype. Accordingly, we anticipate our work to lay the foundation for a combination of functional metagenomics along with microbiological and biochemical analyses, in order to elucidate the functional mechanisms behind soil suppressiveness to *F. culmorum* in the near future.

References

Deacon, J.W. 1984 The Nature and Practice of Biological-Control of Plant-Pathogens - Cook,Rj, Baker,Kf. Nature **309**, 732-732.

Andrade, O., Campillo, R., Peyrelongue, A. & Barrientos, L. 2011 Soils suppressive against Gaeumannomyces graminis var. tritici identified under wheat crop monoculture in southern Chile. Cienc Investig Agrar **38**, 345-356.

Wiseman, B.M., Neate, S.M., Keller, K.O. & Smith, S.E. 1996 Suppression of Rhizoctonia solani anastomosis group 8 in Australia and its biological nature. Soil Biol Biochem **28**, 727-732.

Weller, D.M., Raaijmakers, J.M., Gardener, B.B.M. & Thomashow, L.S. 2002 Microbial populations responsible for specific soil suppressiveness to plant pathogens. Annual Review of Phytopathology **40**, 309.

Cha, J.Y., Han, S., Hong, H.J., Cho, H., Kim, D., Kwon, Y., Kwon, S.K., Crusemann, M., Lee, Y.B., Kim, J.F., et al., 2016 Microbial and biochemical basis of a Fusarium wilt-suppressive soil. Isme J **10**, 119-129.

Schlatter, D., Kinkel, L., Thomashow, L., Weller, D. & Paulitz, T. 2017 Disease Suppressive Soils: New Insights from the Soil Microbiome. Phytopathology **107**, 1284-1297.

Mazzola, M. 2002 Mechanisms of natural soil suppressiveness to soilborne diseases. Anton Leeuw Int J G **81**, 557-564.

Cook, R.J. & Rovira, A.D. 1976 Role of Bacteria in Biological-Control of Gaeumannomyces-Graminis by Suppressive Soils. Soil Biol Biochem **8**, 269-273.

Hornby, D. 1983 Suppressive Soils. Annual Review of Phytopathology **21**, 65-85.

Duran, P., Jorquera, M., Viscardi, S., Carrion, V.J., Mora, M.D. & Pozo, M.J. 2017 Screening and Characterization of Potentially Suppressive Soils against Gaeumannomyces graminis under Extensive Wheat Cropping by Chilean Indigenous Communities. Front Microbiol **8**.

Raaijmakers, J.M., Bonsall, R.E. & Weller, D.M. 1999 Effect of population density of Pseudomonas fluorescens on production of 2,4-diacetylphloroglucinol in the rhizosphere of wheat. Phytopathology **89**, 470-475.

Raaijmakers, J.M. & Weller, D.M. 1998 Natural plant protection by 2,4-diacetylphloroglucinol - Producing Pseudomonas spp. in take-all decline soils. Mol Plant Microbe In **11**, 144-152.

Kwak, Y.S., Han, S., Thomashow, L.S., Rice, J.T., Paulitz, T.C., Kim, D. & Weller, D.M. 2011 Saccharomyces cerevisiae Genome-Wide Mutant Screen for Sensitivity to 2,4-Diacetylphloroglucinol, an Antibiotic Produced by Pseudomonas fluorescens. Appl Environ Microb **77**, 1770-1776.

Alabouvette, C. 1986 Fusarium-Wilt Suppressive Soils from the Chateaurenard Region - Review of a 10-Year Study. Agronomie **6**, 273-284.

Siegel-Hertz, K., Edel-Hermann, V., Chapelle, E., Terrat, S., Raaijmakers, J.M. & Steinberg, C. 2018 Comparative Microbiome Analysis of a Fusarium Wilt Suppressive Soil and a Fusarium Wilt Conducive Soil From the Chateaurenard Region. Front Microbiol **9**.

Duijff, B.J., Recorbet, G., Bakker, P.A.H.M., Loper, J.E. & Lemanceau, P. 1999 Microbial antagonism at the root level is involved in the suppression of Fusarium wilt by the combination of nonpathogenic Fusarium oxysporum Fo47 and Pseudomonas putida WCS358. Phytopathology **89**, 1073-1079.

Carrion, V.J., Cordovez, V., Tyc, O., Etalo, D.W., de Bruijn, I., de Jager, V.C.L., Medema, M.H., Eberl, L. & Raaijmakers, J.M. 2018 Involvement of Burkholderiaceae and sulfurous volatiles in disease-suppressive soils. Isme J **12**, 2307-2321.

Cordovez, V., Carrion, V.J., Etalo, D.W., Mumm, R., Zhu, H., van Wezel, G.P. & Raaijmakers, J.M. 2015 Diversity and functions of volatile organic compounds produced by Streptomyces from a disease-suppressive soil. Front Microbiol **6**.

Garbeva, P., Hol, W.H.G., Termorshuizen, A.J., Kowalchuk, G.A. & de Boer, W. 2011 Fungistasis and general soil biostasis - A new synthesis. Soil Biol Biochem **43**, 469-477.

Cho, G., Kim, J., Park, C.G., Nislow, C., Weller, D.M. & Kwak, Y.S. 2017 Caryolan-1-ol, an antifungal volatile produced by Streptomyces spp., inhibits the endomembrane system of fungi. Open Biol **7**.

Ossowicki, A., Jafra, S. & Garbeva, P. 2017 The antimicrobial volatile power of the rhizospheric isolate Pseudomonas donghuensis P482. Plos One **12**.

Garbeva, P., Hordijk, C., Gerards, S. & de Boer, W. 2014 Volatiles produced by the mycophagous soil bacterium Collimonas. Fems Microbiol Ecol **87**, 639-649.

van Agtmaal, M., Straathof, A.L., Termorshuizen, A., Lievens, B., Hoffland, E. & de Boer, W. 2018 Volatile-mediated suppression of plant pathogens is related to soil properties and microbial community composition. Soil Biol Biochem **117**, 164-174.

Landa, B.B., Mavrodi, O.V., Raaijmakers, J.M., Gardener, B.B.M., Thomashow, L.S. & Weller, D.M. 2002 Differential ability of genotypes of 2,4-diacetylphloroglucinol-producing Pseudomonas fluorescens strains to colonize the roots of pea plants. Appl Environ Microb **68**, 3226-3237.

Wiese, M.V. 1987 Compendium of wheat diseases. 2nd ed. St. Paul, Minn., APS Press; viii, 112 p., 112 p. of plates p.

Schulz-Bohm, K., Zweers, H., de Boer, W. & Garbeva, P. 2015 A fragrant neighborhood: volatile mediated bacterial interactions in soil. Front Microbiol **6**.

Blom, D., Fabbri, C., Eberl, L. & Weisskopf, L. 2011 Volatile-Mediated Killing of Arabidopsis thaliana by Bacteria Is Mainly Due to Hydrogen Cyanide. Appl Environ Microb **77**, 1000-1008.

Park, Y.S., Dutta, S., Ann, M., Raaijmakers, J.M. & Park, K. 2015 Promotion of plant growth by Pseudomonas fluorescens strain SS101 via novel volatile organic compounds. Biochem Bioph Res Co **461**, 361-365.

Chong, J., Soufan, O., Li, C., Caraus, I., Li, S.Z., Bourque, G., Wishart, D.S. & Xia, J.G. 2018 MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. Nucleic Acids Res **46**, W486-W494.

Hoper, H. & Alabouvette, C. 1996 Importance of physical and chemical soil properties in the suppressiveness of soils to plant diseases. Eur J Soil Biol **32**, 41-58.

Janvier, C., Villeneuve, F., Alabouvette, C., Edel-Hermann, V., Mateille, T. & Steinberg, C. 2007 Soil health through soil disease suppression: Which strategy from descriptors to indicators? Soil Biol Biochem **39**, 1-23.

Garbeva, P., van Veen, J.A. & van Elsas, J.D. 2004 Microbial diversity in soil: Selection of microbial populations by plant and soil type and implications for disease suppressiveness. Annual Review of Phytopathology **42**, 243-270.

Donn, S., Almario, J., Mullerc, D., Moenne-Loccoz, Y., Gupta, V.V.S.R., Kirkegaard, J.A. & Richardson, A.E. 2014 Rhizosphere microbial communities associated with Rhizoctonia damage at the field and disease patch scale. Appl Soil Ecol **78**, 37-47.

Yin, C.T., Hulbert, S.H., Schroeder, K.L., Mavrodi, O., Mavrodi, D., Dhingra, A., Schillinger, W.F. & Paulitz, T.C. 2013 Role of Bacterial Communities in the Natural Suppression of Rhizoctonia solani Bare Patch Disease of Wheat (Triticum aestivum L.). Appl Environ Microb **79**, 7428-7438.

Charlop-Powers, Z., Owen, J.G., Reddy, B.V.B., Ternei, M., Guimaraes, D.O., de Frias, U.A., Pupo, M.T., Seepe, P., Feng, Z.Y. & Brady, S.F. 2015 Global Biogeographic Sampling of Bacterial Secondary Metabolism. Elife **4**.

Mehrabi, Z., McMillan, V.E., Clark, I.M., Canning, G., Hammond-Kosack, K.E., Preston, G., Hirsch, P.R. & Mauchline, T.H. 2016 Pseudomonas spp. diversity is negatively associated with suppression of the wheat take-all pathogen. Sci Rep-Uk **6**.

Medema, M.H., Cimermancic, P., Sali, A., Takano, E. & Fischbach, M.A. 2014 A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. Plos Computational Biology **10**.

Garbeva, P., Postma, J., van Veen, J.A. & van Elsas, J.D. 2006 Effect of above-ground plant species on soil microbial community structure and its impact on suppression of Rhizoctonia solani AG3. Environ Microbiol **8**, 233-246.

Chapelle, E., Mendes, R., Bakker, P.A. & Raaijmakers, J.M. 2016 Fungal invasion of the rhizosphere microbiome. Isme J **10**, 265-268.

Kwak, M.J., Kong, H.G., Choi, K., Kwon, S.K., Song, J.Y., Lee, J., Lee, P.A., Choi, S.Y., Seo, M., Lee, H.J., et al., 2018 Rhizosphere microbiome structure alters to enable wilt resistance in tomato (vol 36, pg 1100, 2018). Nat Biotechnol **36**, 1117-1117.

Tyc, O., de Jager, V.C.L., van den Berg, M., Gerards, S., Janssens, T.K.S., Zaagman, N., Kai, M., Svatos, A., Zweers, H., Hordijk, C., et al., 2017 Exploring bacterial interspecific interactions for discovery of novel antimicrobial compounds. Microb Biotechnol **10**, 910-925.

Tyc, O., van den Berg, M., Gerards, S., van Veen, J.A., Raaijmakers, J.M., de Boer, W. & Garbeva, P. 2014 Impact of interspecific interactions on antimicrobial activity among soil bacteria. Front Microbiol **5**.

Traxler, M.F., Watrous, J.D., Alexandrov, T., Dorrestein, P.C. & Kolter, R. 2013 Interspecies interactions stimulate diversification of the Streptomyces coelicolor secreted metabolome. MBio **4**.

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

Watson, A.G. & Ford, E.J. 1972 Soil Fungistasis - Reappraisal. Annual Review of Phytopathology **10**, 327.

Effmert, U., Kalderas, J., Warnke, R. & Piechulla, B. 2012 Volatile Mediated Interactions Between Bacteria and Fungi in the Soil. J Chem Ecol **38**, 665-703.

CHAPTER 3

Dissecting disease-suppressive rhizosphere microbiomes by functional amplicon sequencing and 10X metagenomics

Authors:

Vittorio Tracanna*, Adam Ossowicki*, Marloes L. C. Petrus, Sam Overduin, Barbara R. Terlouw, George Lund, Serina L. Robinson, Sven Warris, Elio G. W. M. Schijlen, Gilles P. van Wezel, Jos M. Raaijmakers, Paolina Garbeva#, Marnix H. Medema#

**These authors contributed equally to this work*

: corresponding author

Published as "Dissecting disease-suppressive rhizosphere microbiomes by functional amplicon sequencing and 10X metagenomics" in *mSystems*, 6(3):e0111620 (2021).

3.1 - Abstract

Disease-suppressive soils protect plants against soil-borne fungal pathogens that would otherwise cause root infections. Soil suppressiveness is, in most cases, mediated by the antagonistic activity of the microbial community associated with the plant roots. Considering the enormous taxonomic and functional diversity of the root-associated microbiome, identification of microbial genera and mechanisms underlying this phenotype is challenging. One approach to unravel the underlying mechanisms is to identify metabolic pathways enriched in the disease-suppressive microbial community, in particular pathways that encode natural products with antifungal properties. An important class of these natural products are peptides produced by Non-Ribosomal Peptide Synthetases (NRPSs). Here, we adopted functional amplicon sequencing of NRPS-associated Adenylation domains (A-domains) to a collection of eight soils that are suppressive or nonsuppressive (i.e., conducive) to *Fusarium culmorum*, a fungal root pathogen of wheat. To identify functional elements in the root-associated bacterial community, we developed an open-source pipeline, referred to as dom2BGC, for amplicon annotation and putative gene cluster reconstruction through analyzing A-domain co-occurrence across samples. We applied this pipeline to rhizosphere communities from four disease-suppressive and four conducive soils and found significant similarities in NRPS repertoires between suppressive soils when compared to conducive soils. Specifically, several siderophore biosynthetic gene clusters were consistently associated with suppressive soils, hinting at competition for iron as a potential mechanism of suppression. Finally, to validate dom2BGC and to allow more unbiased functional metagenomics, we performed 10X metagenomic sequencing of one suppressive soil, leading to the identification of multiple gene clusters potentially associated with the disease-suppressive phenotype.

3.1.1 - Importance

Soil-borne plant pathogenic fungi continue to be a major threat to agriculture and horticulture. The genus *Fusarium* in particular is one of the most devastating groups of soil-borne fungal pathogens for a wide range of crops. Our approach to develop

novel sustainable strategies to control this fungal root pathogen is to explore and exploit an effective, yet poorly understood naturally occurring protection, i.e. disease-suppressive soils. After screening 28 agricultural soils, we recently identified four soils that were suppressive to root disease of wheat caused by *Fusarium culmorum*. We also confirmed, via sterilization and transplantation, that the microbiomes of these soils play a significant role in the suppressive phenotype. By adopting NRPS functional amplicon screening of suppressive and conducive soils, we here show how computationally driven comparative analysis of combined functional amplicon and metagenomic data can unravel putative mechanisms underlying microbiome-associated plant phenotypes.

3.2 - Introduction

Cereals are a staple food for the human population, with wheat as the most widely consumed cereal crop worldwide. It is estimated that up to 40% of crop yields are lost due to weeds, pests and diseases (FAO 2020). Pathogenic fungi are one of the major threats to agriculture. The genus *Fusarium* in particular is one of the most devastating groups of pathogens for a wide range of crops, including wheat (Dean et al., 2012, Valverde et al., 2019). *Fusarium culmorum* causes root rot and head blights in wheat and barley. It can kill plants at early stages of development or reduce their fitness and contaminate the grain with an arsenal of mycotoxins. Intriguingly, in some agricultural soils, root rot caused by *Fusarium culmorum* does not occur or only to a small extent (Ossowicki et al., 2020). This so-called soil disease suppressiveness is a phenomenon where plants show strongly reduced disease symptoms despite the presence of a virulent pathogen and conditions favourable for disease development (Hornby et al., 1983). It is now well established that the soil and root microbiome are essential for disease suppressiveness. In recent work, we performed an extensive screening of 28 soils for their suppressiveness to *F. culmorum* (Ossowicki et al., 2020). We identified and confirmed, via sterilization and transplantation, that in four tested soils the microbiome is associated with suppressiveness to *F. culmorum*. Subsequent comparative taxonomic analysis of the root-associated bacterial communities, aimed to identify differences in abundance or absence/presence patterns of specific genera, revealed only limited

commonalities between the four suppressive soils. The overall aim of this study was to adopt a functional approach to generate hypotheses regarding putative mechanisms associated with the disease-suppressive phenotype.

Many microbe-microbe interactions are mediated by specialized metabolites with diverse functions, including inhibition of fungal growth (Raaijmakers et al., 2012). The production of these bioactive compounds is often encoded by biosynthetic gene clusters (BGCs): groups of physically clustered genes that encode molecular machineries such as Non-Ribosomal Peptide Synthetases (NRPSs) and PolyKetide Synthases (PKSs), which enzymatically assemble complex metabolites. Importantly, these BGCs are often discontinuously distributed across taxa due to high rates of horizontal gene transfer (Medema et al., 2014). Additionally, there may be functional redundancy, due to overlapping biological activities between the products of different BGCs. Therefore, looking at BGC distribution patterns may help explain microbiome-associated phenotypes for which no clear taxonomic associations are identified. PKS and NRPS enzymes are often organized in multi-domain modules, which each contain a set of enzymatic domains that extend the growing peptide or polyketide chain with a specific monomer during enzymatic assembly. Functional amplicon sequencing can target such domains using oligoprimers to amplify DNA from BGCs. Because the sequencing is highly selective, even BGCs from lowly abundant microorganisms can be detected by this technology (Hover et al., 2018, Owen et al., 2013).

Here, we use NRPS amplicon screening for comparative functional analyses of a set of four suppressive and four conducive agricultural soils in the presence and absence of the pathogen *F. culmorum*. To facilitate this analysis, we introduce dom2BGC, a pipeline for extensive annotation of BGC-related amplicons (https://git.wageningenur.nl/traca001/dom2bgc). The amplicons are annotated based on similarity to domains in MIBiG and antiSMASH-DB, two large natural product BGC databases. For NRPS adenylation (A) domains, substrate specificities are predicted based on a newly built random forest classifier trained on the amplified region of these domains. When multiple samples are available, dom2BGC creates a co-occurrence network to aid in detection of groups of amplicons that jointly

originate from known or related BGCs. We apply dom2BGC and validate the annotation and clustering results with the high-quality metagenome of a selected sample enhanced using 10X-based read clouds. The results show siderophore BGCs as key candidates associated with disease suppressiveness of the soils against *F. culmorum*. The linked-read metagenomic dataset further revealed several additional BGCs that, based on their predicted functions, may be involved in the disease-suppressive phenotype. This study exemplifies how computationally driven analysis of combined functional amplicon and metagenomic data can unravel new candidate BGCs for further investigation and help to develop new hypotheses regarding the mechanisms underlying important microbiome-associated phenotypes.

3.3 - Materials and Methods

3.3.1 - Soil collection

Eight soil samples: S01, S03, S08, S11, S14, S15, S17 and S28 were collected from 3-meter squares located at the centre of each agricultural field in January-April 2017. In this area, topsoil cores of approximately 30 cm depth were collected. Soils were air-dried at room temperature, homogenized, sieved through a 4mm mesh sieve and stored at 4°C. Soil S28 was additionally flaked after drying using a jaw-crusher (Type BB-1, Retsch, Germany). Detailed descriptions of the soil samples are included in our previous study (Ossowicki et al., 2020).

3.3.2 - Disease suppressiveness assay and A-domains amplification

Wheat growth conditions, pathogen inoculation, the suppressiveness assay, A-domains amplification and sequencing are described in detail in supplementary methods. Briefly, wheat seedlings were transferred to substrate containing one of the eight tested soils and challenged with pathogenic *Fusarium culmorum* PV using untreated plants as control, each combination had 12 replicates. After three weeks, wheat plants were inspected for disease symptoms and given a disease index

describing the severity of infection from 0 (healthy plant) to 5 (heavily diseased), like in our previous work (4). Rhizosphere DNA was isolated from 4 randomly chosen replicates per treatment. NRPS adenylation domains were amplified using A3F and A7R primers (Ayuso-Sacido et al., 2005) using Q5 polymerase.

3.3.3 - A-domain amplicon preparation

Barcoding and sequencing of the A-domain amplicons was performed at BaseClear (Leiden, the Netherlands) using Illumina MiSeq, which generated 4,181,437 paired-end reads of 250 bp in length. Sequences were de-multiplexed and adapters trimmed using Qiime2 (Bolyen et al., 2019). Quality filtering and denoising was done with Dada2 (Callahan et al., 2016). Nucleotide sequences were translated to amino acid sequences with transeq from the EMBOSS suite (Rice et al., 2000). Forward sequences were aligned with the AMP-binding domain hidden Markov model (HMM) profile PF00501 from the pfam database [version 27] (El-Gebali et al.,2019) using hmmsearch from the HMMer package [version 3.1] (Wheeler et al., 2013). The output table was parsed to retain only the conserved amino acids in the sequence corresponding to “match” states with the HMM profile. Protein sequences shorter than 66 amino acids were discarded. The resulting pre-aligned amplicon sequences from the natural source are referred to as nAMPs (natural amplicons) to distinguish them from the *in silico* amplicons used for their annotation.

3.3.4 - 10X metagenome

DNA extraction, sequencing and the assembly are described in detail in the supplementary methods. Briefly, 10X Genomics Chromium was used to generate a read cloud library from high quality rhizosphere DNA and subsequently sequenced on an Illumina NovaSeq 6000.

3.3.5 - Feature extraction from amplicons for substrate specificity prediction

1,029 experimentally validated bacterial NRPS A-domains from the MIBiG database were used as a training set. Training set sequences were aligned to the AMP-binding (PF00501) HMM and the range of 34 residues that aligned with positions 210 – 243 (PheA) were extracted. All duplicates and any sequences for which there were less than seven training examples for a given amino acid substrate were removed from the dataset, leaving a training set of 848 sequences (see data folder). Each of the 34 residues was encoded as a vector of 15 physicochemical properties including hydrophobicity, secondary structure, size, and polarity. The full vector of 510 features was used to train separate random forest models to predict amino acid monomer specificity and broad substrate groups using the SKLearn package [version 0.20.2] in Python [version 3.7.3].

Each random forest classifier was randomly split with class-specific stratification into 90% training and 10% test. Model parameters were tuned based on OOB score for the training set over 3 iterations. Overfitting was limited by pruning tree depth to a maximum of 20. The number of features randomly sampled as candidates for each split was set to default (square root of number of predictors). The random forest was grown to a size of 1,000 trees.

Final models for monomer and broad substrate group classification were used to make predictions for the 51,914 soil amplicon sequences. Sequences with a prediction probability score less than 0.5 were labeled as ‘no confident prediction.’ Approximately 65% of the broad substrate groups and 49% of the monomers could be predicted with confidence.

3.3.6 - Dom2BGC pipeline - Generation of *in silico* amplicons

To generate a reference dataset of NRPS functional amplicons, A-domain sequences were extracted from AntiSMASH-DB and MIBiG BGCs. In Dom2BGC, in silico amplicons are created by searching these sequences using hmmsearch with the A-domain hidden Markov model (HMM) profile from Pfam (PF00501). This produces reference sequences aligned to the HMM profile. In order to produce in silico amplicons comparable to the nAMPs, the alignment matching the nAMP match

coordinates is extracted. This process creates *in silico* amplicons that are pre-aligned to the nAMPs, which allows for quick matching between nAMPs and *in silico* AMPs using pairwise identity. Annotations available for *in silico* amplicons are stored to be transferred to any nAMPs matching with it. Currently supported annotations include, where available, the taxonomy of the source organism, the BGC type annotation based on antiSMASH predictions, and the name of the natural product for which the production is encoded in the BGC (for domains extracted from MIBiG entries (Kautsar et al., 2020)). Calculations for diversity measures and community composition are described in supplementary material and methods.

3.3.7 - Dom2BGC pipeline - Amplicon matching and annotation

Each nAMP is matched to an in-silico amplicon if it shares 90% or more of its amino acid sequence with the reference over the full amplicon length. For nAMPs matching to multiple in-silico AMPs within a reference database, all entries are recorded. In case of multiple nAMPs matching an individual in-silico amplicon, all matched nAMPs are grouped for evaluation of presence-absence patterns and abundance of the in-silico amplicon.

In dom2bgc, amplicons are taxonomically annotated at the lowest rank available. In case of annotation to a reference BGC with a different taxonomic annotation, dom2BGC assigns the amplicon to the lowest common ancestor of the matching references. In addition, information from the reference cluster on the gene cluster family is passed on to the matching amplicon. This annotation is based on antiSMASH classification rules for predicted gene clusters. Possible annotations include NRPS, lipopeptides, hybrid PKS and more. In case of an amplicon matching with reference clusters belonging to different gene cluster families, dom2BGC reports all matches.

3.3.8 - Dom2BGC pipeline - Co-occurrence network creation and analysis

Pairwise co-occurrence patterns of nAMPs are calculated using Spearman rank correlation of presence-absence patterns using numpy meshgrid. To filter out spurious relationships, the correlation network contains only the strongest correlations in the 99th percentile among abundant nAMPs. In the resulting network, amplicons are nodes and edges are drawn based on co-occurrence. Clustering within the network to define BGC hubs is performed with DBscan. These BGC hubs, comprising highly correlated nAMPs, are inspected for nAMP annotation enrichment. Cluster nodes and first-degree neighbors annotated to the same reference gene cluster are further selected as putative gene clusters. Networks are visualized in Cytoscape (Shannon et al., 2003) and putative clusters are reported in a separate tab separated file.

3.4 - Results and Discussion

3.4.1 - Identification of disease- suppressive agricultural soils

In our previous study (Ossowicki et al., 2020), we tested 28 diverse field soils from the Netherlands and Germany for disease suppressiveness against *F. culmorum* root rot of wheat. Based on these results we selected four disease-suppressive (S01, S03, S11 and S28) and four disease-conducive (S08, S14, S15 and S17) soils for further analysis. For the amplicon-based analyses of the rhizosphere microbiome, we again performed disease suppressiveness assays on these eight soils. We observed no disease symptoms in two inoculated suppressive soils (S11 and S28), and only low levels of disease in the other two inoculated suppressive soils (S01 and S03). This clearly contrasts with the four conducive soils, where disease levels varied from moderate (S08) to high (S14, S15 and S17, Figure 3-1). In two of the conducive soils (S14 and S17), we also identified some mild disease symptoms in treatments without addition of the pathogen, indicating the presence of indigenous populations of *F. culmorum* or of other pathogens causing similar disease symptoms (Figure 3-1, light blue bars). Altogether, these results confirm and extend the results of our previous study and show a clear distinction in phenotype between the four suppressive and the four conducive soils.

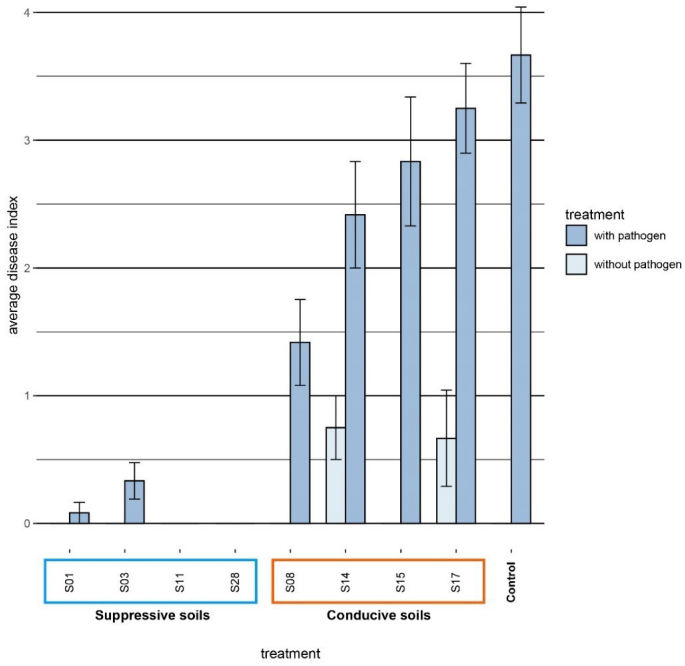


Figure 3-1. Disease index of *Fusarium* root rot disease of wheat grown in eight different agricultural soils. Four soils (S01, S03, S11, S28) were classified as disease suppressive and four soils (S08, S14, S15, S17) as disease conducive. Dark blue - inoculated with *F. culmorum*, light blue – non-inoculated, sterile BS dune soil was used as a control. The bar indicates the average disease index, with the error bars representing the standard error of the mean (n=12).

3.4.2 - Functional amplicon sequencing uncovers novel NRPS domains from low-abundant bacteria in rhizosphere microbial communities

As our previous 16S rRNA-based analysis of taxonomic similarities and differences between and across conducive and suppressive soils revealed that no taxa were unequivocally linked to disease suppression (Ossowicki et al., 2020), we turned to functional amplicon sequencing to assess whether this could point to metabolites or classes of metabolites associated with the suppressive phenotype. The selective amplification of functional domains allows capturing biosynthetic diversity found within a complex soil sample. Specifically, we used PCR amplification of A-domains of NRPSs, which are involved in the production of several types of bioactive molecules that have been previously linked to disease suppression, such as

lipopeptides and siderophores. In NRPSs, the role of A-domains is to recognise and activate amino acid substrates that are incorporated into the growing peptide (Martinez-Nunez et al 2016). Based on their sequence, it is possible to predict their amino acid specificity and match them to databases of known or predicted BGCs. Functional amplicon sequencing of adenylation domains across the four suppressive and four conducive soils produced 4,181,437 raw reads across all samples, which were used to identify association patterns of A-domains across suppressive and conducive soils. One replicate from suppressive soil S28 (FC.1, supplementary FigureS1) was removed from further analysis, because it produced significantly fewer reads compared to other samples (12,380 reads while the rest of the samples average 61,132 reads). Processing of the reads resulted in 3,396,393 reads mapping to 51,912 unique domains. Rarefaction analysis revealed that for most samples, diversity was sufficiently covered at ~30,000 reads per sample (Supplementary Table.S1).

To facilitate linking amplicon sequences to specific BGCs, we generated a high-quality shotgun metagenome assembly of one sample from the rhizosphere microbiome of plants grown in soil S11. This soil was chosen because of its strong disease suppression in this study as well as in our previous experiments. To increase assembly contiguity, we made use of 10X linked read sequencing technology, which is able to generate much more contiguous contigs compared to what is possible with conventional metagenomics with comparable coverage. We used the dedicated cloudSPAdes 10X linked reads assembler on these data, which resulted in an assembly size of 2.2Gb and an N50 of 2.8Kb for contigs above 1Kb, with the largest contig measuring 1.3 Mb. Compared to the metaSPAdes equivalent assembly, which does not make use of the linked read information, we observed a considerable improvement in the N50 and assembly size for contigs above 5Kb (7.3Kb for regular metaSPAdes assembly and 20.2Kb for cloudSPAdes) which makes the cloudSPAdes assembly more suited to obtain complete NRPS BGCs (Meleshko et al., 2019).

Functional amplicon sequencing of A-domains can achieve better coverage of domains from rare BGCs compared to metagenomics with the same sequencing

volume. This is reflected by the higher diversity of domains found in natural amplicons (nAMPs), with 40,005 unique amplicons at the protein level, compared to the shotgun assembly that yielded 8,762 unique *in silico* amplicons at the protein level. Remarkably, we observed that the number of unique sequences present in all our samples surpasses the diversity contained in antiSMASH-DB (24,085 AMPs) - the largest available annotated database for natural product-encoding BGCs that contains sequence data for 32,548 BGCs from 24,776 microbial genomes. To highlight the importance of environmental sampling efforts, we further matched the nAMP sequences to *in silico* amplicons from antiSMASH-DB. We found that most sequences could be matched at or above 70% identity. However, there are 162 instances of A-domains with < 30% amino acid sequence identity to their closest representative in the database. These domains, while still matching the Pfam domain, can potentially harbor novel functions, such as incorporation of different amino acids, or may simply belong to rare and uncharted BGCs. The percentage identity of nAMPs to the closest antiSMASH-DB AMP follows a normal distribution, with a peak to the right accounted for by (near-)perfect matches to previously sequenced clusters (Figure 3-2A).

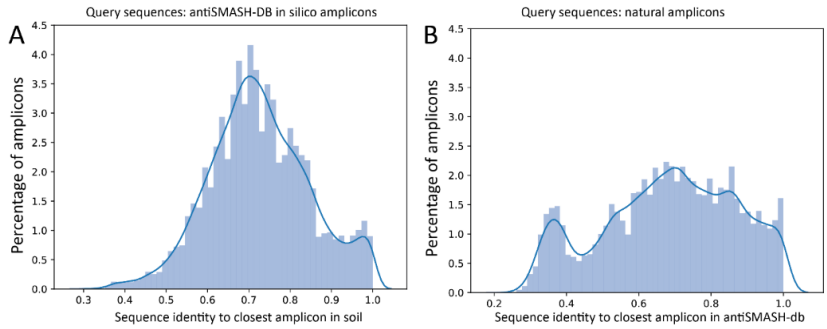


Figure 3-2. **Sequence distance between nAMPs and antismash-DB in silico amplicons.** A) Histogram showing the distribution of best matches (as in highest percentage identity at protein level) between each nAMP and the antiSMASH-DB *in silico* amplicon database. B) Histogram showing the distribution of best matches (as in highest percentage identity at protein level) between each antiSMASH-DB *in silico* amplicon and the nAMPs.

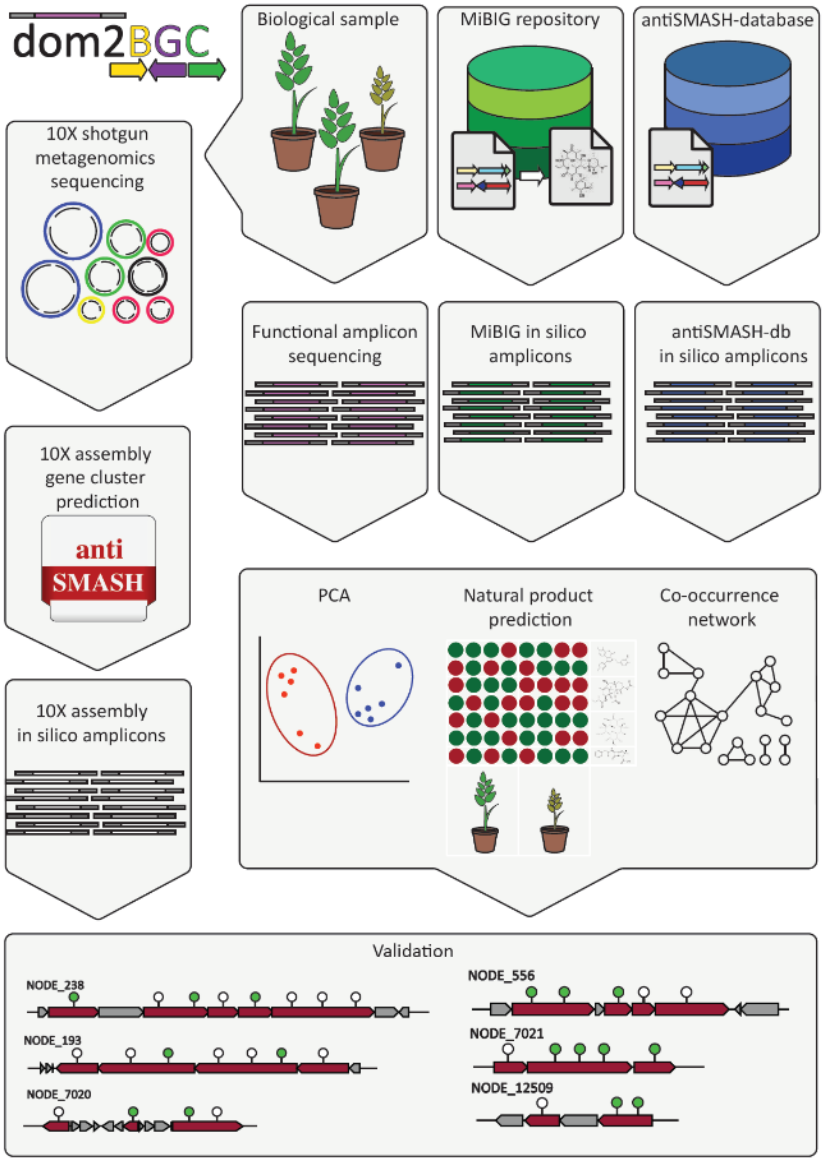


Figure 3-3. The dom2BGC annotation pipeline and validation process. Amplified sequences from the rhizosphere are translated to nAMPs as per methods section and have been annotated through comparison with *in silico* amplicons from MiBIG and antiSMASH databases. Richness and community composition measures are used to assess their associations with phenotype and treatments. Co-occurrence patterns of amplicons which share similarity to the same reference BGCs were used to predict presence of (homologues of) known BGCs. Finally, in this study, a shotgun metagenomic assembly from one of the soil samples was used to confirm the presence of these predicted gene clusters from the amplicon data.

3.4.3 - The dom2BGC pipeline facilitates automated annotation and networking of functional amplicons

To evaluate the impact of the primer bias on the observed amplicon diversity, we performed an inverse analysis by identifying the closest match of *in silico* amplicons from antiSMASH-DB to the nAMPs from the soil, as the first is not affected by primer bias. The results revealed a bimodal distribution (Figure 3-2B and supplementary table S2). The leftmost mode includes amplicons not present in the samples, as well as amplicons that might be present in the samples but absent in the nAMPs set because of their poor match to the primer sequences. Still, the majority of the *in silico* amplicons from antiSMASH-DB had a match in our sample above 60% sequence identity. This indicated that the primer bias, despite being present, does not prevent the majority of the known sequence diversity of adenylation domains to be represented in the functional amplicon data. These results confirm the high value of functional amplicon sequencing studies in charting the biosynthetic potential of environmental niches. Based on these results we see that with limited primer bias we can still get substantial coverage of nAMPs.

Current tools for the annotation of functional amplicons (eSNaPD (Reddy et al., 2014), NaPDoS) have limited applications or rely on laborious processes which require expensive laboratory automation of BAC clone library approaches (CONKAT-seq (Libis et al., 2019)). To harness the potential of A-domain functional amplicons in soils, we developed dom2BGC, a pipeline to add taxonomical, functional and product annotation to amplicon sequences and validate some of the predicted clusters using shotgun metagenomics assembly data. Within dom2BGC (Figure 3-3), amplicons are matched to antiSMASH-DB and MIBiG, two natural product BGC databases, and annotations are transferred to the query amplicons when hits are reported above a user-set threshold (default: 95% identity). Diversity measurements and community structure relationships between samples are calculated and visualized in a series of automatically generated figures (examples: Figure 3-2, Figure 3-5). Finally, a co-occurrence network of amplicons across the samples is created. Neighboring amplicons mapping onto domains of known clusters from antiSMASH-DB or MIBiG are considered as domains which potentially belong

to the same original cluster. This information can then be used in designing further experiments to validate the putative functions of the identified clusters.

To identify known natural product BGCs in the microbial communities, a total of 3,239 *in silico* amplicons were generated from MIBiG products entries (MAMPs – MIBiG amplicons). 1,312 unique nAMPs, corresponding to 8% of the total, were matched and associated to a BGC for a known natural product. Notably, the most abundant known BGC annotated encodes the biosynthesis of pyoverdine; this NRPS gene cluster is widespread among *Pseudomonas* species which are also common members of the rhizosphere. Still, even for MIBiG entries with a perfect match and consistent coverage across samples, not all the A-domains present in the reference cluster could be amplified. This illustrates how functional amplicon sequencing provides deep coverage of biosynthetic diversity across microbiome samples, but also misses certain domains because of mismatches between oligo-primers and the target sequence or other PCR biases. This is partially balanced by the fact that most NRPS gene clusters encode multiple A-domains, which increases the chance that at least one of these regions is amplified. As for database coverage, 119 out of 860 entries with an adenylation domain in MIBiG had at least one amplicon from our data mapping to one of its domains above 90% amino acid identity. This is testament to the extensive natural product potential of soil microbial communities.

To investigate the taxonomical and gene cluster class distributions of nAMPS, a total of 40,211 *in silico* amplicons were generated from antiSMASH-DB BGCs (aSAMPs) and used to annotate 5,531 nAMPs (corresponding to 29,9% of total reads), linking them to 1,443 different BGCs. This annotation rate constitutes about a 4-fold increase compared to the numbers of nAMPs that could be annotated using MIBiG as reference.

3.4.4 - Disease-suppression is not associated with increased adenylation domain diversity but shows distinct community structure

There is great need for diagnostic tools to assess the disease-suppressive potential of agricultural soils based on their microbial and functional composition. In a recently published paper, Yuan et al., (Yuan et al., 2020) explored in a meta-analysis the potential of 16S and ITS amplicons as predictors of disease occurrence. Since A-domain functional amplicon data showed more distinctive patterns than 16S data between soils with conducive and suppressive phenotypes (Hornby et al., 2012), we set out to explore if it might be feasible to use functional amplicon sequencing as a diagnostic tool of disease suppressiveness. To test the possible association of within-sample amplicon diversity measures with the suppressive phenotype, we calculated within-sample richness, evenness and phylogenetic diversity for all samples based on observed unique amplicons, Simpson-e and Faith-PD, respectively. Wilcoxon rank-sum tests showed no significant association of alpha diversity measures with the presence of the pathogen nor with the suppressive phenotype for any of these metrics (Figure 3-4A).

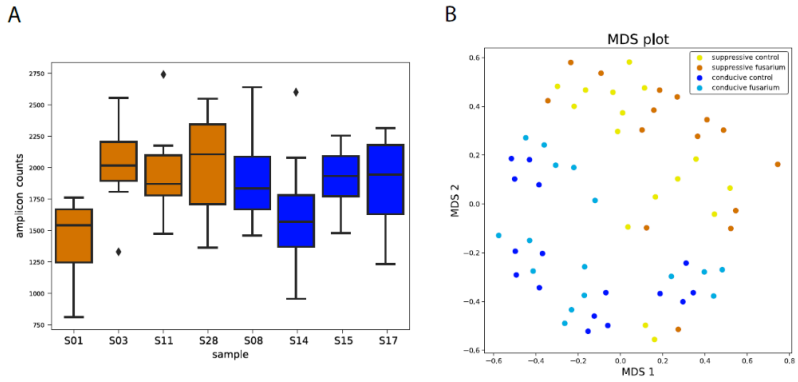


Figure 3-4. Community diversity and composition. A) Adenylation domain richness across suppressive (orange bars) and conducive (blue bars) soils, calculated as unique sequences. B) Visualization of the adenylation domain community composition with multidimensional scaling.

Several studies have associated overall microbial species richness or evenness in the soil and rhizosphere with disease suppressiveness (Garbeva et al., 2004, Janvier et al., 2007, Chaparro et al., 2012, Larkin, 2015, van Bruggen et al., 2015). In other studies, however, this was not the case, and suppressiveness was associated with the abundance/enrichment of specific genera or functions (Mendes et al., 2011, Carrion et al., 2019). Here, we note that suppressive soils were both among the most and least diverse in terms of NRPS A-domains, which highlights the importance of availability of samples from multiple sources that share the same phenotype before drawing conclusions on the role of community diversity in disease suppression.

In a multi-dimensional scaling (MDS) analysis, suppressive soils did form a distinct group based on their community profile (Figure 3-4, panel B) with significant grouping, suggesting that similar community NRPS profiles can indeed be associated with the suppressive phenotype based on unweighted Unifrac (PERMANOVA, p-value=0.010; ANOSIM, p-value=0.010). This could indicate that the observed phenotype is caused by a single or limited number of pathway(s), not detectable with overall richness or abundance measurements, that directly interfere with a pathogen's ability to colonize the rhizosphere, initiate root penetration and disease.

Thus, it appears that sequencing A-domain community composition has the potential to become a predictive tool for diagnosing soil suppressiveness. Nevertheless, we should emphasize that our study is based on only one host-pathogen system (wheat and *Fusarium culmorum*) and a collection of eight soils. Still, the fact that the production of compounds by NRPS and PKS enzymes play crucial roles in other disease-suppressive soils (Carrion et al., 2019, Duijff et al., 1999, Duijff et al 1994, Scher et al., 1982, Raaijmakers et al., 1998, Michelsen et al., 2015, Zhao et al., 2018, Hayden et al., 2018, Weller et al., 2002, Kinkel et al., 2012) supports this proposition. This method has to be further developed and validated in the future through the inclusion of more host-pathogen systems and soils suppressive to other soil-borne fungal pathogens.

3.4.5 - Suppressive soils are enriched in cyclic-peptide-associated A-domains

Adenylation domains activate and incorporate specific amino acids in the growing nonribosomal peptide during synthesis by an NRPS assembly line. The substrate specificity for different A-domains is determined by a restricted number of residues in their sequence (Stachelhaus et al., 1999). A-domains incorporate a large variety of both proteogenic and non-proteogenic amino acids, which facilitate the structural diversity of the final peptide products. We reasoned that prediction of the substrate specificities of the domain amplicons detected in suppressive and conducive rhizosphere samples could provide new insights into the abundance and diversity of different products, and trained a classifier to predict these specificities (see Methods). Intriguingly, we found predicted threonine-specific domains to be significantly more common in suppressive soils versus conducive (rank-sum test p-value<0.001, full result table in Supplementary table S3). This is particularly interesting as threonine is an amino acid commonly involved in lactone ring formation of cyclic and branched cyclic (lipo)peptides. Such peptides have a large variety of natural functions, which encompass, among others, the induction of systemic resistance in plants to fungal infection and direct antifungal activity (Raaijmakers et al., 2012, Cawoy et al., 2015, Geudens et al., 2018, Kruijt et al., 2009, Omoboye et al., 2019, Raaijmakers et al., 2010).

3.4.6 - Reconstruction of 31 gene clusters from amplicon data using domain annotation and co-occurrence pattern analysis

Co-occurrence of domains across the soil samples was used to build a pairwise co-occurrence matrix as described in methods. A strict filter was applied to remove spurious correlations. To this end, we keep only the Spearman correlations above the 99th percentile, which resulted in a co-occurrence network containing 1,618 amplicons. Associations of co-occurring amplicons into putative BGCs were predicted only for co-occurring amplicons which share annotation to one or multiple references; this resulted in the reconstruction of 31 gene clusters (supplementary

table S4). These clusters belonged to multiple taxonomical groups, namely *Pseudomonas*, *Delftia*, *Streptomyces*, *Variovorax*, *Burkholderia* and *Collimonas*. In order to validate putative network clusters, we generated 8,762 *in silico* amplicons from our 10X shotgun metagenome-assembly as described above. Two of the 31 reconstructed gene clusters could be matched to known gene cluster products predicted from the metagenome: the BGCs for nunamycin and delftibactin from *Pseudomonas* and *Delftia* respectively, as shown in Figures 3-5 and 3-6.

3.4.7 - Overview of the BGCs associated with suppressive soils

Next, we identified in more detail the BGCs detected in the wheat rhizosphere microbiome from suppressive soil S11. To this end, we used antiSMASH to identify BGCs in the 10X shotgun metagenome assembly of this soil. This resulted in 991 predicted BGCs from multiple GCFs associated with various known compounds. Notable compounds include siderophores like turnerbactin, delftibactin, fimsbactin, xanthoferrin and amonabactin, lipopeptides like nunamycin/nunapeptin and brabantamide (Michelsen et al., 2015, Han et al., 2013, Tejman-Yarden et al., 2019, Bohac et al., 2019, Pandey et al., 2019, Barghouthi et al., 1989, Schmidt et al., 2014), and known antifungal compounds like 2,4-diacetylphloroglucinol (36). This array of candidate clusters offered an initial insight into putative mechanisms associated with the disease-suppressive phenotype, in which one or multiple compounds may inhibit simultaneously or sequentially the growth of the invading pathogen and suppress root infection.

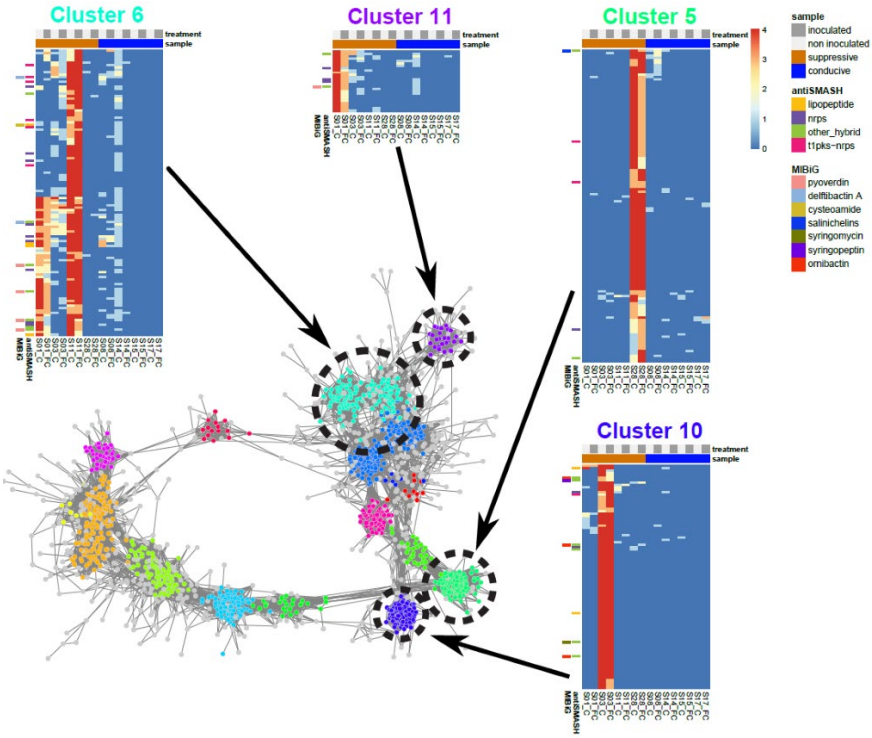


Figure 3-5. Domain co-occurrence network showing clusters associated with soil suppressiveness. For each of the four clusters (5, 6, 10 and 11), a heatmap shows distribution of A-domains across the samples. The heatmap colour scale represents the number of replicates in which the A-domain occurs (from dark blue – absent to red – present in all four replicates. Upper colour bars in the heatmaps describes samples – light grey – non-inoculated, dark grey – inoculated with pathogen and disease suppressiveness – orange – suppressive, blue - conducive. The left side of each heatmap shows which A-domains were annotated using the MiBIG or antiSMASH databases with colour bars. Colour of the bars indicate a compound or compound class shown in the legend.

3.4.8 - Analysis of siderophores and lipopeptides associated with observed phenotypes

As expected, our MIBiG-based annotations show that a considerable portion of the amplicons (955 out of 5,531) mapped to *Pseudomonas* A-domains. A-domains from this study could be mapped to BGCs belonging to 68 different genera and 208 bacterial species (Supplementary table S2). With these taxonomic annotations obtained from dom2BGC, it was possible to identify taxonomic patterns of adenylation domains associated with soil disease suppressiveness. Multiple species known for their biosynthetic potential and for involvement in disease suppressiveness in other systems were significantly enriched in suppressive soils at

high taxonomical resolution (Supplementary table S4). This suggests that these bacteria, which were previously found to exhibit antifungal activity, might also play a role in the disease suppressiveness against *F. culmorum* in wheat.

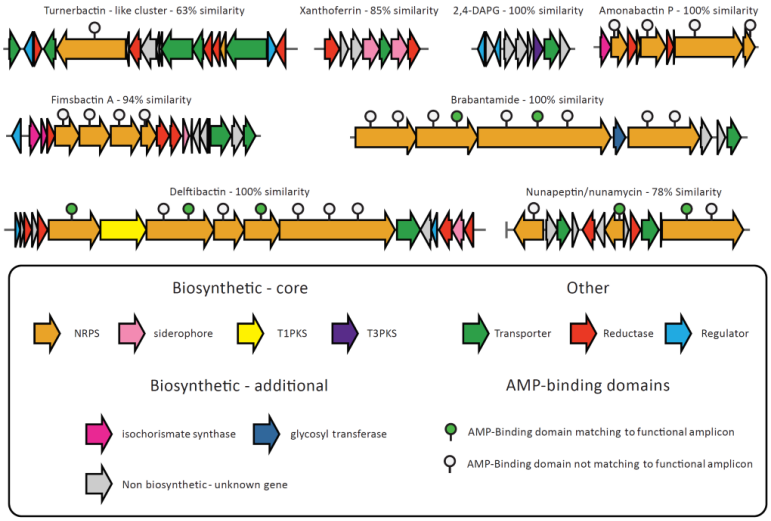


Figure 3-6. Selection of known BGCs predicted in the rhizosphere metagenome of suppressive soil S11. Arrows represent predicted genes and are colour-coded based on their annotated function. AMP-binding domains matching to functional amplicons are highlighted as described in the legend.

DBscan clustering of the A-domain co-occurrence network produced 16 clusters. Among these clusters, 4 were associated with at least one suppressive soil. The most interesting subnetwork (Figure 3-5, cluster 6) has amplicons associated with suppressive soil S11 and partially with soil S01, with some amplicons present across three suppressive soils. Three separate domain clusters were reconstructed within this subnetwork, with all three matching BGCs encoding the production of known siderophores, namely pyoverdine from *Pseudomonas*, scabichelin from *Streptomyces* and delftibactin from *Delftia*. All of these were associated with suppressive soil S11 and the last one with suppressive soil S01 as well. Siderophores are a group of secondary metabolites produced by microorganisms in iron-limited environments like soil. These metabolites form complexes with insoluble iron, facilitating the uptake of this iron by microorganisms. Often, competition for iron is a central process in soil systems with neutral to high pH (56–60). Siderophores and competition for iron were found to be involved in soil disease suppression

mechanisms against *Fusarium* wilt (Duijff et al., 1999, Duijff et al., 1994, Scher 1982, Raaijmakers 1998, Michelsen et al., 2015, Scher, 1982, Alabouvette 1986, Baker et al., 1986), *take-all* disease in wheat (Verbon et al., 2017, Lemanceau et al., 2009) and damping-off of sugar beet (Carrion et al., 2019).

The concentration of soluble iron in eight tested soils, as assayed in our previous study (4), ranged from 0.01 mg/kg in soil S17 to 0.11 mg/kg in soil S11 with the exception of soil S03, where the concentration was much higher and reached 0.45 mg/kg. The high iron concentration in soil S03 can be explained by its low pH (5.28), which increases the solubility of oxidized iron. All other soils have a neutral pH (7.13 to 7.82) or are only slightly acidic (soils S01 and S08, pH 6.22 and 6.87, respectively) (supplementary table S5 and previous work). We observe that the broad presence of siderophores is not limited to environments with a low availability of iron. Those results do not indicate a simple connection between the concentration of soluble iron and soil disease suppressiveness against *F. culmorum*. Nevertheless, the production of siderophores is so widespread among microorganisms is soil systems that we can consider it as primary process in ecosystem functioning consequently indispensable for soil disease suppressiveness.

The network hub associated with suppressive soil S03 (Figure 3-5, cluster 10) contains three predicted reconstructed gene clusters taxonomically assigned to *Burkholderia*, *Collimonas* and *Pseudomonas*. The *Burkholderia* and *Collimonas* clusters matched to multimodular NRPSs with no known associated natural product, while the reconstructed cluster from *Pseudomonas* matched to the syringafactin BGC. Finally, the pyoverdine BGC from *Pseudomonas* was recovered from a smaller amplicon subnetwork (Figure 3-5, cluster 11). While the consistent recovery of the pyoverdine BGC in multiple hubs is expected given its ubiquity in rhizosphere-associated pseudomonads, the recovery of the delftibactin and scabichelin BGCs and their association to two suppressive soils emphasize the contribution of different kinds of siderophores in disease suppression. Our results were further confirmed by the prediction of a delftibactin BGC in the associated shotgun metagenome assembly from soil S11 with antiSMASH, which has an almost perfect match with the delftibactin BGC in MiBIG (Figure 3-6). The largest suppressive-sample-

associated subnetwork by number of amplicons (Figure 3-5, cluster 5) possesses an individual cluster matching the scabichelin BGC from *Streptomyces scabies*. This siderophore has been found to be produced by previously reported *Fusarium*-suppressive strains (del Barrio-Duque et al., 2019). The reconstruction of separate instances of the same BGC suggest that the underlying amplicons belong to variants of the scabichelin cluster present in different rhizosphere communities.

All in all, the results suggest an association of siderophore BGCs with the disease-suppressive phenotype across the soils studied. They also point to a possible functional redundancy that should be validated in future work: in some soils, a suppressive function might be mediated through the production of some siderophores (e.g., delftibactin), while in other soils the same function might be mediated by other natural products (e.g., scabichelin).

Based on the MIBiG database, 15 lipopeptides were annotated in our samples. Figure S2 presents the distribution of these compounds among suppressive and conducive soils. Interestingly, most annotated lipopeptides are much more abundant in conducive soils, especially in soil S17. Many of these lipopeptides are connected to bacterial plant pathogens and act like pathogenicity factors (for example: syringafactin, tolaasin, sessilin), while others have been implicated in soil disease suppressiveness and antagonistic interactions with fungi (for example: nunamycin and thanamycin) or breaking down bacterial biofilms (for example: WLIP, entolysin, putisolvin and xantholysin A). Many of the A domains that are part of NRPS BGCs of plant pathogenic bacteria are also part of NRPS BGCs of non-pathogenic bacteria (Girard et al., 2020). Isolation of the bacteria harbouring these BGCs and subsequent genetic, genomic, transcriptomic and mutational analyses will be needed to determine the identity as well as any functional significance of these BGCs in suppressiveness.

3.5 – Conclusions

Our study provides novel insights into the NRPS AMP-binding domain diversity of agricultural rhizosphere samples. Remarkably, the set of unique amplicons from this rhizosphere collection equals the level of diversity of adenylation domains found across all publicly available genomes. Annotation rates for nAMPs were generally low, which highlights the incredible potential of plant-associated microbiomes for discovering novel natural products. We report significant community structure overlap among suppressive rhizobacterial adenylation domains profiles, and generated new hypotheses regarding possible roles for siderophores in disease suppression against *Fusarium culmorum*. We also developed a pipeline for taxonomic and functional annotation of NRPS amplicons without the requirement of a BAC-clone library. The dom2BGC pipeline can be extended to and currently supports annotation of any natural product-associated domain that occurs multiple times within a BGC, and to some extent for any BGC-associated domain. We validated the amplicon clustering results by reconstructing the delftibactin BGC, a siderophore associated with suppressive soils using a combination of amplicon sequencing and novel 10X genomics shotgun metagenomics sequencing. We conclude that combining functional amplicon sequencing and shotgun metagenomics highlighted represents a powerful approach to probe complex microbiome-associated plant phenotypes and to generate new hypotheses on the functional roles of microbial metabolites in microbe-microbe and microbe-host interactions.

3.6 - Data availability

Raw sequence data that support the findings of this study have been deposited in NCBI under project number PRJNA670155.

References

2020. Food and Agriculture Organisation of the United Nations. http://faostat.fao.org/.

Dean R, Van Kan JA, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, Rudd JJ, Dickman M, Kahmann R, Ellis J, Foster GD. 2012. The Top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol 13:414–30.

Valverde-Bogantes E, Bianchini A, Herr JR, Rose DJ, Wegulo SN, Hallen-Adams HE. 2019. Recent population changes of Fusarium head blight pathogens: drivers and implications. Canadian Journal of Plant Pathology 42:3,315-329.

Ossowicki A, Tracanna V, Petrus MLC, van Wezel G, Raaijmakers JM, Medema MH, Garbeva P. 2020. Microbial and volatile profiling of soils suppressive to Fusarium culmorum of wheat. Proceedings of the Royal Society B: Biological Sciences 287:20192527.

Hornby D. 1983. Suppressive Soils. Annu Rev Phytopathol 21:65–85.

Raaijmakers JM, Mazzola M. 2012. Diversity and Natural Functions of Antibiotics Produced by Beneficial and Plant Pathogenic Bacteria. Annual Review of Phytopathology 50:403–424.

Medema MH, Cimermancic P, Sali A, Takano E, Fischbach MA. 2014. A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. PLoS Comput Biol 10:e1004016.

Hover BM, Kim S-H, Katz M, Charlop-Powers Z, Owen JG, Ternei MA, Maniko J, Estrela AB, Molina H, Park S, Perlin DS, Brady SF. 2018. Culture-independent discovery of the malacidins as calcium-dependent antibiotics with activity against multidrug-resistant Gram-positive pathogens. 4. Nature Microbiology 3:415–422.

Owen JG, Reddy BVB, Ternei MA, Charlop-Powers Z, Calle PY, Kim JH, Brady SF. 2013. Mapping gene clusters within arrayed metagenomic libraries to expand the structural diversity of biomedically relevant natural products. Proc Natl Acad Sci U S A 110:11797–11802.

Vittorio Tracanna. 2020. Project title: dom2bgc. https://git.wageningenur.nl/traca001/dom2bgc.

Ayuso-Sacido A, Genilloud O. 2005. New PCR primers for the screening of NRPS and PKS-I systems in actinomycetes: detection and distribution of these biosynthetic gene sequences in major taxonomic groups. Microb Ecol 49:10–24.

Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet CC, Al-Ghalith GA, Alexander H, Alm EJ, Arumugam M, Asnicar F, Bai Y, Bisanz JE, Bittinger K, Brejnrod A, Brislawn CJ, Brown CT, Callahan BJ, Caraballo-Rodríguez AM, Chase J, Cope EK, Da Silva R, Diener C, Dorrestein PC, Douglas GM, Durall DM, Duvallet C, Edwardson CF, Ernst M, Estaki M, Fouquier J, Gauglitz JM, Gibbons SM, Gibson DL, Gonzalez A, Gorlick K, Guo J, Hillmann B, Holmes S, Holste H, Huttenhower C, Huttley GA, Janssen S, Jarmusch AK, Jiang L, Kaehler BD, Kang KB, Keefe CR, Keim P, Kelley ST, Knights D, Koester I, Kosciolek T, Kreps J, Langille MGI, Lee J, Ley R, Liu Y-X, Loftfield E, Lozupone C, Maher M, Marotz C, Martin BD, McDonald D, McIver LJ, Melnik AV, Metcalf JL, Morgan SC, Morton JT, Naimey AT, Navas-Molina JA, Nothias LF, Orchanian SB, Pearson T, Peoples SL, Petras D, Preuss ML, Pruesse E, Rasmussen LB, Rivers A, Robeson MS, Rosenthal P, Segata N, Shaffer M, Shiffer A, Sinha R, Song SJ, Spear JR, Swafford AD, Thompson LR, Torres PJ, Trinh P, Tripathi A, Turnbaugh PJ, Ul-Hasan S, van der Hooft JJJ, Vargas F, Vázquez-Baeza Y, Vogtmann E, von Hippel M, Walters W, Wan Y, Wang M, Warren J, Weber KC, Williamson CHD, Willis AD, Xu ZZ, Zaneveld JR, Zhang Y, Zhu Q, Knight R,

Caporaso JG. 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. 8. Nature Biotechnology 37:852–857.

Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. 2016. DADA2: High resolution sample inference from Illumina amplicon data. Nat Methods 13:581–583.

Rice P, Longden I, Bleasby A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276–277.

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432.

Wheeler TJ, Eddy SR. 2013. nhmmer: DNA homology search with profile HMMs. Bioinformatics 29:2487–2489.

Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJJ, van Santen JA, Tracanna V, Suarez Duran HG, Pascal Andreu V, Selem-Mojica N, Alanjary M, Robinson SL, Lund G, Epstein SC, Sisto AC, Charkoudian LK, Collemare J, Linington RG, Weber T, Medema MH. 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48:D454–D458.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 13:2498–2504.

Martínez-Núñez MA, López VEL y. 2016. Nonribosomal peptides synthetases and their applications in industry. Sustainable Chemical Processes 4:13.

Meleshko D, Mohimani H, Tracanna V, Hajirasouliha I, Medema MH, Korobeynikov A, Pevzner PA. 2019. BiosyntheticSPAdes: reconstructing biosynthetic gene clusters from assembly graphs. Genome Res 29:1352–1362.

Reddy BVB, Milshteyn A, Charlop-Powers Z, Sean F. Brady. 2014. eSNaPD: a versatile, web-based bioinformatics platform for surveying and mining natural product biosynthetic diversity from metagenomes. Chem Biol 21:1023–1033.

The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity.

Libis V, Antonovsky N, Zhang M, Shang Z, Montiel D, Maniko J, Ternei MA, Calle PY, Lemetre C, Owen JG, Brady SF. 2019. Uncovering the biosynthetic potential of rare metagenomic DNA using co-occurrence network analysis of targeted sequences. 1. Nature Communications 10:3848.

Yuan J, Wen T, Zhang H, Zhao M, Penton CR, Thomashow LS, Shen Q. 2020. Predicting disease occurrence with high accuracy based on soil macroecological patterns of Fusarium wilt. The ISME Journal 1–15.

Garbeva P, van Veen JA, van Elsas JD. 2004. Microbial diversity in soil: selection microbial populations by plant and soil type and implications for disease suppressiveness. Annu Rev Phytopathol 42:243–270.

Janvier C, Villeneuve F, Alabouvette C, Edel-Hermann V, Mateille T, Steinberg C. 2007. Soil health through soil disease suppression: Which strategy from descriptors to indicators? Soil Biology and Biochemistry 39:1–23.

Chaparro JM, Sheflin AM, Manter DK, Vivanco JM. 2012. Manipulating the soil microbiome to increase soil health and plant fertility. Biol Fertil Soils 48:489–499.

Larkin RP. 2015. Soil Health Paradigms and Implications for Disease Management. Annual Review of Phytopathology 53:199–221.

van Bruggen AHC, Sharma K, Kaku E, Karfopoulos S, Zelenev VV, Blok WJ. 2015. Soil health indicators and Fusarium wilt suppression in organically and conventionally managed greenhouse soils. Applied Soil Ecology 86:192–201.

Felnagle EA, Jackson EE, Chan YA, Podevels AM, Berti AD, McMahon MD, Thomas MG. 2008. Nonribosomal Peptide Synthetases Involved in the Production of Medically Relevant Natural Products. Mol Pharmaceutics 5:191–211.

Mendes R, Kruijt M, Bruijn I de, Dekkers E, Voort M van der, Schneider JHM, Piceno YM, DeSantis TZ, Andersen GL, Bakker PAHM, Raaijmakers JM. 2011. Deciphering the Rhizosphere Microbiome for Disease-Suppressive Bacteria. Science 332:1097–1100.

Carrión VJ, Perez-Jaramillo J, Cordovez V, Tracanna V, Hollander M de, Ruiz-Buck D, Mendes LW, Ijcken WFJ van, Gomez-Exposito R, Elsayed SS, Mohanraju P, Arifah A, Oost J van der, Paulson JN, Mendes R, Wezel GP van, Medema MH, Raaijmakers JM. 2019. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science 366:606–612.

Duijff BJ, Recorbet G, Bakker PAHM, Loper JE, Lemanceau P. 1999. Microbial antagonism at the root level is involved in the suppression of Fusarium wilt by the combination of nonpathogenic Fusarium oxysporum Fo47 and Pseudomonas putida WCS358. Phytopathology 89:1073–1079.

Duijff BJ, Bakker PAHM, Schippers B. 1994. Suppression of fusarium wilt of carnation by Pseudomonas putida WCS358 at different levels of disease incidence and iron availability. Biocontrol Science and Technology 4:279–288.

Scher M. BR. 1982. Effect of Pseudomonas putida and a Synthetic Iron Chelator on Induction of Soil Suppressiveness to Fusarium Wilt Pathogens. Phytopathology 72:1567.

Raaijmakers JM, Weller DM. 1998. Natural Plant Protection by 2,4-Diacetylphloroglucinol-Producing Pseudomonas spp. in Take-All Decline Soils. MPMI 11:144–152.

Michelsen CF, Watrous J, Glaring MA, Kersten R, Koyama N, Dorrestein PC, Stougaard P. 2015. Nonribosomal Peptides, Key Biocontrol Components for Pseudomonas fluorescens In5, Isolated from a Greenlandic Suppressive Soil. mBio 6.

Zhao ML, Yuan J, Zhang RF, Dong MH, Deng XH, Zhu CZ, Li R, Shen QR. 2018. Microflora that harbor the NRPS gene are responsible for Fusarium wilt disease-suppressive soil. Appl Soil Ecol 132:83–90.

Hayden HL, Savin KW, Wadeson J, Gupta V, Mele PM. 2018. Comparative Metatranscriptomics of Wheat Rhizosphere Microbiomes in Disease Suppressive and Non-suppressive Soils for Rhizoctonia solani AG8. Front Microbiol 9:859.

Weller DM, Raaijmakers JM, Gardener BBM, Thomashow LS. 2002. Microbial populations responsible for specific soil suppressiveness to plant pathogens. Annu Rev Phytopathol 40:309-+.

Kinkel LL, Schlatter DC, Bakker MG, Arenz BE. 2012. Streptomyces competition and co-evolution in relation to plant disease suppression. Res Microbiol 163:490–499.

Stachelhaus T, Mootz HD, Marahiel MA. 1999. The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chemistry & Biology 6:493–505.

Cawoy H, Debois D, Franzil L, Pauw ED, Thonart P, Ongena M. 2015. Lipopeptides as main ingredients for inhibition of fungal phytopathogens by Bacillus subtilis/amyloliquefaciens. Microbial Biotechnology 8:281–295.

Geudens N, Martins JC. 2018. Cyclic Lipodepsipeptides From Pseudomonas spp. – Biological Swiss-Army Knives. Front Microbiol 9.

Kruijt M, Tran H, Raaijmakers JM. 2009. Functional, genetic and chemical characterization of biosurfactants produced by plant growth-promoting Pseudomonas putida 267. Journal of Applied Microbiology 107:546–556.

Mm Y, Ss W, Dv M, Ov M, D von W, Ls T, Jh G, Dm W. 2014. Biological control of wheat root diseases by the CLP-producing strain Pseudomonas fluorescens HC1-07. Phytopathology 104:248–256.

Omoboye OO, Oni FE, Batool H, Yimer HZ, De Mot R, Höfte M. 2019. Pseudomonas Cyclic Lipopeptides Suppress the Rice Blast Fungus Magnaporthe oryzae by Induced Resistance and Direct Antagonism. Front Plant Sci 10:901.

Oni FE, Geudens N, Onyeka JT, Olorunleke OF, Salami AE, Omoboye OO, Arias AA, Adiobo A, Neve SD, Ongena M, Martins JC, Höfte M. 2020. Cyclic lipopeptide-producing Pseudomonas koreensis group strains dominate the cocoyam rhizosphere of a Pythium root rot suppressive soil contrasting with P. putida prominence in conducive soils. Environmental Microbiology 22:5137–5155.

Raaijmakers JM, De Bruijn I, Nybroe O, Ongena M. 2010. Natural functions of lipopeptides from Bacillus and Pseudomonas : more than surfactants and antibiotics. FEMS Microbiol Rev 34:1037–1062.

Han AW, Sandy M, Fishman B, Trindade-Silva AE, Soares CAG, Distel DL, Butler A, Haygood MG. 2013. Turnerbactin, a novel triscatecholate siderophore from the shipworm endosymbiont Teredinibacter turnerae T7901. PLoS ONE 8:e76151.

Tejman-Yarden N, Robinson A, Davidov Y, Shulman A, Varvak A, Reyes F, Rahav G, Nissan I. 2019. Delftibactin-A, a Non-ribosomal Peptide With Broad Antimicrobial Activity. Front Microbiol 10.

Bohac TJ, Fang L, Giblin DE, Wencewicz TA. 2019. Fimsbactin and Acinetobactin Compete for the Periplasmic Siderophore Binding Protein BauB in Pathogenic Acinetobacter baumannii. ACS Chem Biol 14:674–687.

Pandey SS, Patnana PK, Rai R, Chatterjee S. 2017. Xanthoferrin, the α-hydroxycarboxylate-type siderophore of Xanthomonas campestris pv. campestris, is required for optimum virulence and growth inside cabbage. Mol Plant Pathol 18:949–962.

Barghouthi S, Young R, Olson MO, Arceneaux JE, Clem LW, Byers BR. 1989. Amonabactin, a novel tryptophan- or phenylalanine-containing phenolate siderophore in Aeromonas hydrophila. J Bacteriol 171:1811–1816.

Schmidt Y, van der Voort M, Crüsemann M, Piel J, Josten M, Sahl H-G, Miess H, Raaijmakers JM, Gross H. 2014. Biosynthetic origin of the antibiotic cyclocarbamate brabantamide A (SB-253514) in plant-associated Pseudomonas. Chembiochem 15:259–266.

Kramer J, Özkaya Ö, Kümmerli R. 2020. Bacterial siderophores in community and host interactions. Nat Rev Microbiol 18:152–163.

Saha M, Sarkar S, Sarkar B, Sharma BK, Bhattacharjee S, Tribedi P. 2016. Microbial siderophores and their potential applications: a review. Environ Sci Pollut Res 23:3984–3999.

Lemanceau P, Bakker P, Jandekogel W, Alabouvette C, Schippers B. 1993. Antagonistic Effect of Nonpathogenic Fusarium-Oxysporum Fo47 and Pseudobactin 358 Upon Pathogenic Fusarium-Oxysporum F-Sp Dianthi. Appl Environ Microbiol 59:74–82.

Kloepper JW, Leong J, Teintze M, Schroth MN. 1980. Enhanced plant growth by siderophores produced by plant growth-promoting rhizobacteria. 5776. Nature 286:885–886.

Haas D, Défago G. 2005. Biological control of soil-borne pathogens by fluorescent pseudomonads. 4. Nature Reviews Microbiology 3:307–319.

Alabouvette C. 1986. Fusarium-Wilt Suppressive Soils from the Chateaurenard Region - Review of a 10-Year Study. Agronomie 6:273–284.

Baker R, Elad Y, Sneh B. 1986. Physical, Biological and Host Factors in Iron Competition in Soils, p. 77–84. In Swinburne, TR (ed.), Iron, Siderophores, and Plant Diseases. Springer US, Boston, MA.

Verbon EH, Trapet PL, Stringlis IA, Kruijs S. 2017. Iron and Immunity 24.

Lemanceau P, Expert D, Gaymard F, Bakker PAHM, Briat JF. 2009. Role of Iron in Plant-Microbe Interaction. Advances in Botanical Research 51:491–549.

del Barrio-Duque A, Ley J, Samad A, Antonielli L, Sessitsch A, Compant S. 2019. Beneficial Endophytic Bacteria-Serendipita indica Interaction for Crop Enhancement and Resistance to Phytopathogens. Front Microbiol 10.

Girard L, Höfte M, Mot RD. 2020. Lipopeptide families at the interface between pathogenic and beneficial Pseudomonas -plant interactions. Critical Reviews in Microbiology 46:397–419.

Acknowledgements
VT, AO and MP are supported by the research programme NWO-Groen, which is jointly funded by the Netherlands Organisation for Scientific Research (NWO), BASF SE and Baseclear BV, under project number ALWGR.2015.1. We thank Walter Pirovano and Danny Duijsings (Baseclear) for providing advice on library preparation and Illumina sequencing of the functional amplicons.

Funding
This work was supported by NWO grant ALWGR. 2015.1.

CHAPTER 4

Deciphering the microbiome of a disease-suppressive soil by dilution-to-extinction

Authors:

Adam Ossowicki*, Vittorio Tracanna*, Somayah S. Elsayed, Elio G. W. M. Schijlen, Mirna Baak, Walter Pirovano, Gilles P. van Wezel, Jos M. Raaijmakers, Marnix H. Medema#, Paolina Garbeva#

*: contributed equally to this work

#: corresponding author

Manuscript in preparation

4.1 – Abstract

In disease-suppressive soils, root-associated microbiomes provide plant protection against infections by specific soil-borne plant pathogens. Previously, we screened a large number of soils for suppressiveness to *Fusarium culmorum* for wheat and revealed that soil S11 showed a consistent and high level of suppressiveness that was associated with the rhizosphere microbiome. In this study, we used a dilution-to-extinction approach to investigate the underlying mechanisms of microbiome-mediated disease suppressiveness to *F. culmorum*. Our results show that disease suppressiveness can be transferred to a sterile soil by introducing the microbiome extracted from the rhizosphere of wheat grown in suppressive soil S11. Introducing different dilutions of this S11 rhizosphere microbiome into the sterile soil revealed a nonlinear relationship between the dilution factor and the level of disease suppressiveness. At low microbiome dilution, disease was significantly suppressed relative to the untreated control, but the suppressive effect was lost at higher dilutions (100X and higher). Shotgun metagenomic sequencing and reconstruction of 96 metagenome assembled genomes from the rhizosphere microbiota of the different dilutions revealed changes in microbiome composition and functions. We observed a reduction in microbial diversity along the dilution trajectory with an increase in the relative abundance of the Proteobacteria and Bdellovibrionota, and a decrease in the relative abundances of Streptomycetales, Acidobacteria and Verrucomicrobiota. Analysis of the occurrences of KEGG orthologs showed enrichment for 143 orthologs in the suppressive conditions and included terms associated with iron uptake, chitinases and components of the type 6 secretion system. Nine metagenome-assembled genomes were associated with the diminished suppressive phenotype, including a high-quality genome from the bacterial genus *Labilithrix* that contains multiple novel biosynthetic gene clusters. Collectively, this study exemplifies the high added value of dilution-to-extinction to deconstruct a naturally occurring microbiome-associated plant phenotype, i.e., disease suppressiveness. Deciphering the mechanisms and microbial consortia involved in suppressive soils is highly instrumental for the development of new disease management strategies for sustainable agriculture.

4.2 – Introduction

The thin layer of soil surrounding and influenced by plant roots - the rhizosphere – is populated by diverse microbial communities recruited from the soil microbiome by root exudates (Berg and Smalla, 2009; Morella et al., 2020). Rhizosphere microbes often confer beneficial functions to plants, such as nutrient acquisition, production of phytohormones and protection against biotic and abiotic stresses (Cordovez et al., 2019). Hence, changes in the composition of the rhizosphere microbiome can positively or negatively affect plant growth and health. For ages, people have recognized that some soils are better suited for cultivation of particular crops, but only in the last decades we started to understand how microbes present in a soil contribute to better yields and protection from diseases. In so-called disease suppressive soils, the level of disease is low to absent, despite the presence of a virulent pathogen and a susceptible host plant (Raaijmakers and Mazzola, 2011). For most disease-suppressive soils, the protection provided to the plant is mediated by activities of specific members of the soil and rhizosphere microbiota (for reviews see (Gomes Exposito et al., 2017; Schlatter et al., 2017; Weller et al., 2002)). Because of the complex diversity and dynamics of the rhizosphere microbiota (Wagg et al., 2014), unravelling the mechanisms by which specific microbial consortia provide protection against diseases is a challenge. New technologies, such as dilution-to-extinction (DTE) and genome-resolved shotgun metagenomics now provide new opportunities to disentangle the microbial traits associated with suppressiveness. DTE is a controlled perturbation method leading to a simplified derivative of the initial microbiome and has been proven effective for studying microbiome composition and function. For example, it has been used for manipulation of microbiomes of the gut (Kenters et al., 2011; Lagier et al., 2015) and water (Stingl et al., 2008; Yu et al., 2019) to study rare taxa and ecological succession. Manipulation of microbiome structure and diversity by DTE has also successfully been used in several studies of soil microbiomes in the context of the utilization of carbon sources (Garland and Lehman, 1999), fungistasis (Hol et al., 2015), chitin and cellulose degradation (Peter et al., 2011), antibiotic resistance (Chen et al., 2019, 2017), general functional potential (Yan et al., 2017)(Chen et al., 2020) and regulation of root exudation (Korenblum et al., 2020). Here, for the first

time, we employ DTE in combination with metagenomics to identify microbiome functions associated *in situ* with disease suppression. Identifying microbial consortia and microbial functions involved in suppressive soils can substantially contribute to developing new sustainable solutions for future agriculture, with reduced inputs of fertilizers and chemical pesticides.

In our previous studies, we screened 28 soils for suppressiveness to *Fusarium culmorum* of wheat and identified several suppressive soils. Amplicon sequencing showed that these suppressive soils did not share the same taxonomic patterns, but co-occurrence network analysis revealed at least one common uniquely overrepresented bacterial guild dominated by *Acidobacteria* (Ossowicki et al., 2020). In some of the suppressive soils, we also found an association between suppressiveness and the abundance of biosynthetic gene clusters (BGCs) encoding the production of siderophores based on NRPS functional amplicon sequencing (Tracanna et al., 2019). For the work presented here, we chose agricultural soil S11 because it showed the highest and most consistent level of disease suppressiveness and an association of suppressiveness with siderophore BGCs. For the experimental work, we inoculated a sterile soil with different dilutions of soil S11 and planted wheat. Then, we challenged the plants with the fungal pathogen and monitored disease severity and microbiome dynamics by in-depth metagenomic sequencing. We hypothesize that dilution of the suppressive rhizosphere microbiome i) diminishes the abundance of specific microbial taxa and biosynthetic gene clusters (BGCs) involved in disease suppressiveness, ii) diminishes microbial diversity and thereby disturbs the interactions within the microbial consortia responsible for disease suppressiveness.

4.3 - Results and discussion

4.3.1 - Impact of rhizosphere microbiome dilution on disease suppressiveness

After growing wheat for two weeks in a greenhouse pot experiment, we extracted the microbiome of disease-suppressive soil S11. When introduced into a sterile sandy soil, the extracted soil microbiome (ESM) conferred significant suppressiveness to *F. culmorum* infections of wheat in two independent bioassays with sixteen biological replicates (Figure4-1). At low dilutions (10X, 25X and 50X), the disease was significantly suppressed, but the suppressive effect was lost at higher dilutions (100X, 200X and 400X) (Figure 4-1).

4.3.2 - Dilution-to-extinction is an effective approach to dissect microbiome-associated phenotypes

Following the bioassays, DNA was extracted from the rhizosphere of wheat plants treated with four ESM dilutions that confer phenotypes ranging from disease-suppressive to conducive (non-suppressive). Based on taxonomic classification of the metagenomic reads, the results showed that the rhizosphere microbiomes maintained their overall taxonomic structure at the phylum level over the dilution series (Figure 4-2), but the relative abundance of various taxonomic groups significantly changed as a result of the dilution (presented and discussed further below).

The taxonomic distribution was consistent among all five biological replicates of the original extracted ESM and of each of the four dilutions (Figure S1). According to nonpareil estimation of coverage (Rodriguez-R et al., 2018), the in-depth sequencing effort for this experiment was able to cover even less abundant and rare members of the rhizosphere microbiota nearly completely (Figure 4-3A), providing a basis to deconstruct microbial functions associated with the disease-suppressive plant phenotype while maintaining a similar overall taxonomic composition of the rhizosphere microbiome

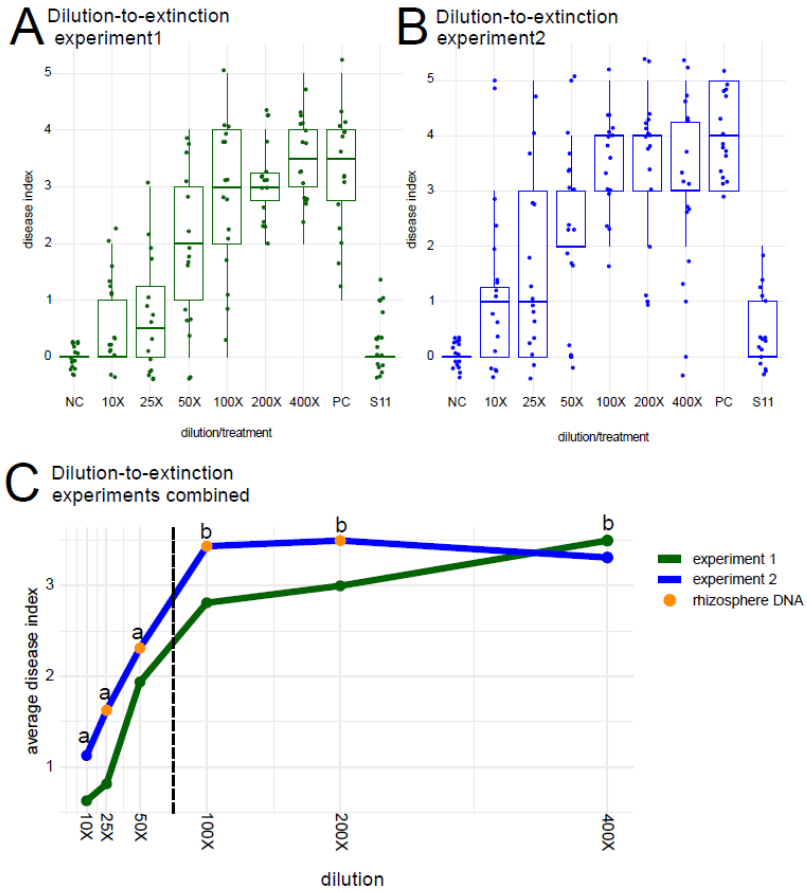


Figure 4-1. Effects of diluted soil suspensions (10X up to 400X) of disease-suppressive soil S11 on *Fusarium culmorum* disease severity of wheat plants grown in a sterile soil. Results of two independent experiments (panels A, B) are shown. Disease severity was scored with an index scale ranging from 0 (healthy) to 5 (dead plants). Average disease indices are shown for 16 replicates per treatment. NC is the 'healthy control' where wheat plants are grown in the sterile soil without the pathogen; PC is the 'diseased control' with sterile soil inoculated with the fungal pathogen; S11 is suppressive field soil S11 inoculated with the pathogen. Panel C represents the combined results of the two independent experiments shown in panels A and B. The black dashed line indicates the transition point in the plant phenotype from suppressive to conducive according to the statistical analyses of the chi-square test; different letters above the data points indicate statistically significant differences. For the metagenome analyses, DNA rhizosphere samples were collected in experiment 2 from the four dilutions indicated by orange dots on the graph.

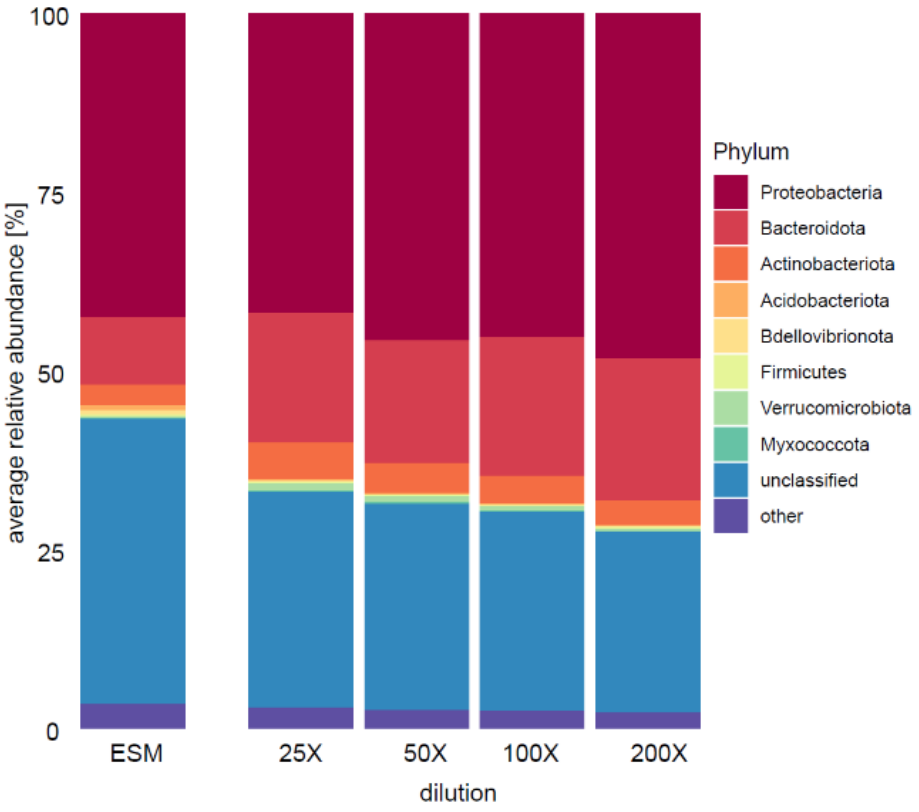


Figure 4-2. **Relative abundance of the bacterial phyla detected in different dilutions of the rhizosphere microbiome of wheat plants grown in Fusarium-suppressive soil S11.** ESM represents the extracted soil microbiome, whereas 25X – 200X represent the ESM diluted 25 – 200 times. Averages of 5 replicates are shown.

4.3.3 - Changes in microbiome composition and diversity are associated with suppressiveness

Taxonomic annotation at the read level using Kraken2 with the GTDB database (Parks et al., 2020; Wood et al., 2019) showed that the most abundant phyla were Proteobacteria and Bacteriodota. These two phyla together constituted 60 - 68 % of the bacterial community, and their relative share increased at higher ESM dilutions (Figure 4-2). Accordingly, the relative abundance of other phyla decreased, except the Bdellovibrionota. *Bdellovibrio* are considered to be potential agents for controlling phytopathogens (Yair et al., 2009; Youdkes et al., 2020), although they have been mostly shown to antagonize bacterial (and not fungal) pathogens. The

increased relative abundance of Proteobacteria was attributed to the orders of Burkholderiales, Nevskiales, Caulobacterales, Rhizobiales and Sphingomonadales (Figure 4-4). The relative abundance of Bacteriodota was stable across all the ESM dilutions. Notably, we observed a significant decrease in the relative abundances of taxa previously associated with disease suppressiveness, including the order Streptomycetales (Cordovez et al., 2015; Kinkel et al., 2012) and the phyla Acidobacteria (Ossowicki et al., 2020; Shen et al., 2015) and Verrucomicrobiota (Chapelle et al., 2016; Navarrete et al., 2015) (Figure 4-4). In our previous work on comparative taxonomic profiling of disease conducive and disease suppressive soils to *F. culmorum*, including soil S11, co-occurrence network analysis revealed similarities between the suppressive soils through an overrepresentation of Acidobacteria (Ossowicki et al., 2020). Hence, the results obtained here by metagenomic analysis supports this earlier observation and are compatible with a potential role of Acidobacteria in suppressiveness to *F. culmorum*. Validating this role will require isolation of representative genera and species of this phylum, many of which are still very difficult to culture in laboratory conditions (Costa et al., 2018).

The diversity detected in the twenty rhizosphere metagenomes and in the undiluted ESM was estimated using the database-independent Nonpareil 3 algorithm (Rodriguez-R et al., 2018). The results showed a diversity reduction with increased ESM dilutions (Figure 4-3B). The major drop in diversity between undiluted ESM and the subsequent dilutions may reflect the selective power of the rhizosphere and the inability of certain groups of microbes to establish and thrive in a sterile soil following inoculation. We noted that the diversity of the rhizosphere metagenomes significantly differed between the lower dilutions (25X and 50X) and the higher dilutions (100X and 200X) (Figure 4-3B), which correlates with their contrasting plant phenotypes. In two former studies (Berg et al., 2017; Mallon et al., 2015) it was postulated that the invasion of a pathogen in a biological system, like the rhizosphere, may lead to the extinction of a subset of the resident species, compromising community functioning. It was experimentally shown that reducing the diversity in synthetic bacterial communities made them more susceptible to invading species (Jousset et al., 2011; van Elsas et al., 2012). Our results suggest that the reduced microbial diversity negatively affects disease suppressiveness, supporting

the hypothesis that diversity of root-associated microbiomes contributes to plant resilience to biotic stress. In summary, our results showed: i) a reduction in microbial diversity along the ESM dilution trajectory with a concomitant reduction in disease suppressiveness, ii) an increase in the relative abundance of the Proteobacteria and Bdellovibrionota at higher dilutions, and iii)) a decrease in the relative abundances of Streptomycetales, Acidobacteria and Verrucomicrobiota at higher ESM dilutions.

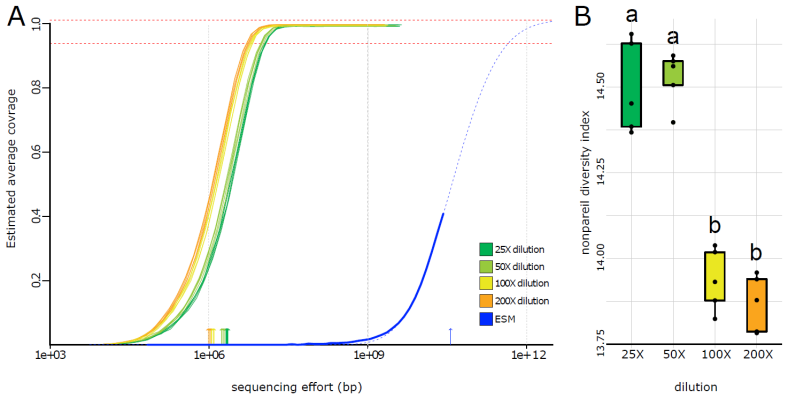


Figure 4-3. Overview of the microbial community coverage and diversity. Panel A: estimated sequencing coverage of extracted soil microbiome of suppressive soil S11 (ESM) and rhizosphere microbiomes from different ESM dilutions. ESM coverage for the original extract before inoculation is incomplete (less than 50% of the sequence space is covered) and shows much higher complexity. All the rhizosphere samples cover the genetic diversity of the microbial communities. The curves show a small but consistent separation between low (25X and 50X) and high (100X and 200X) dilutions, with lower dilutions displaying higher diversity. Panel B: Effect of diluted suspension of suppressive soil S11 introduced to a sterile soil on genetic diversity of wheat rhizosphere microbiomes expressed as a nonpareil diversity index. The disease symptoms caused by plant pathogen *F. culmorum* in two lower dilutions 25X and 50X were significantly suppressed compared to higher dilutions 100X and 200X. The significance levels according to ANOVA and Tukey post-hoc test with $p > 0.001$ are indicated by different letters. Full results are presented in Table S1.

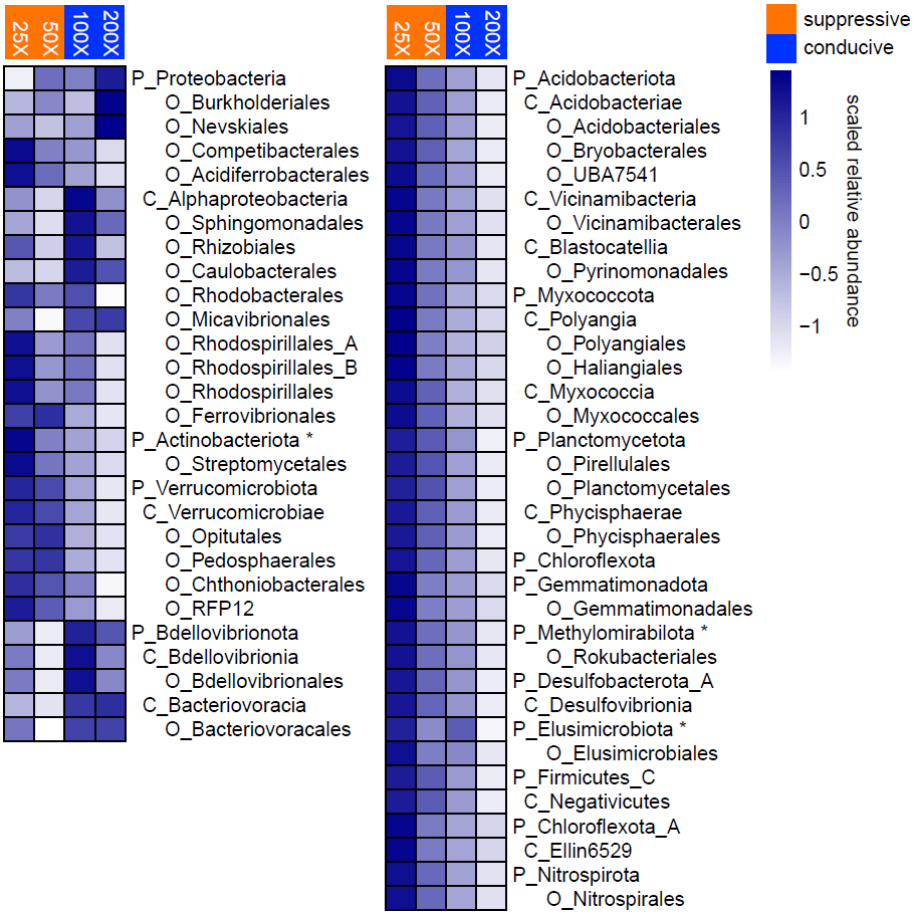


Figure 4-4. Significant changes in the relative abundance of the rhizosphere bacterial taxa between suppressive soil S11 dilutions exibiting a *Fusarium*-suppressive (25X and 50X) or a non-suppressive phenotype (100X and 200X). Relative abundance is scaled in rows (z-score), and the name of the taxonomic level is indicated by capital letters: P – Phylum, C – Class, O – Order. The significance of the change in abundance evaluated according to DESeq2 analysis with an adjusted p value < 0,05. The change in relative abundance of taxa marked with an asterisk is not statistically significant, and they are shown to indicate the phylum of lower taxonomic levels. Low abundant taxa (<1%) are not shown for clarity of presentation. All taxa significantly changing abundance are shown in Table S2.

4.3.4 - Genes associated with iron uptake, chitinase activity and T6SS secretion systems are enriched in disease-suppressive rhizosphere microbiomes

Co-assembly of all rhizosphere metagenome reads (8,214,787,741 reads, 953,849,270,414 Gb) yielded a total of 327,947 contigs above 5 Kb in length with an

N50 of 11.908 Kb and total assembly size of 3.638 Gb. The co-assembly strategy was chosen to improve the overall coverage for each organism in the niche, as significant overlap was expected for all samples derived from the same ESM. The KEGG ortholog (KO) functional annotation of diluted rhizosphere metagenomes reflected the functional potential of the microbial community. As expected in a rhizosphere metagenome, there was a large variety of KEGG orthologous groups (KO groups), with 11,459 unique entries. Investigating a general functional profile of the rhizosphere communities (top and second-level KEGG ortholog categories) across the dilution series showed no statistically significant changes (see supplementary material). To find associations between individual KEGG ortholog groups and the suppressive phenotype, we then separated metagenomic contigs in two categories based on a DESeq2 enrichment test between the lower two dilution points (25X and 50X) versus the two higher dilution points (100X and 200X). Occurrences of KO group annotations for genes of putative suppressiveness-associated contigs were grouped together and tested versus counts in the rest of the assembly. A Fisher's exact test of the occurrences of KEGG orthologs showed enrichment for 143 orthologs after multiple testing correction. Among these, we found terms associated with iron uptake like TonB [ko:K03832] and bacterioferritin [ko:K03594], chitinases [ko:K01183] and components of the Type VI Secretion System (T6SS) [ko:K11894]. Competition for iron was also identified as a potential mechanism of suppression of *F. culmorum* in our previous amplicon-based analyses of nonribosomal peptide synthetase (NRPS) genes (Tracanna et al., n.d.) and previously also for soils suppressive to *Fusarium oxysporum* (Lemanceau et al., 2009; Mazurier et al., 2009; Siegel-Hertz et al., 2018; Tamietti and Alabouvette, 1986). The enrichment of NRPS genes involved in siderophore biosynthesis discovered here with untargeted metagenome analysis corroborates our earlier results and highlights the possible contribution of competition for iron in disease suppressiveness to *F. culmorum*.

Chitinases are enzymes involved in degradation of polymeric chitin and often associated with bacteria-fungi interactions (Lacombe-Harvey et al., 2018; Veliz et al., 2017). Production of chitinases was recognized as a potential control mechanism in a soil suppressive to club-root disease of cabbage (Hjort et al., 2014). Also, in a

recent study on suppressiveness to *Rhizoctonia solani*, the expression of chitinases was found significantly upregulated in disease-suppressive consortia of root endophytic bacteria (Carrion et al., 2019). However, the enrichment of chitinases in *F.-culmorum*-suppressive rhizosphere samples may be related to saprophytic activity of bacteria as well as direct antagonism against the pathogen. Additional experiments are needed to test and validate those scenarios.

The T6SS is used by Gram-negative bacteria to deliver effectors such as toxins and other pathogenicity factors (Coulthurst, 2013; Records, 2011). T6SS is specifically known to mediate interaction between bacteria. Nevertheless, some studies have suggested a role in antagonistic activity of bacteria against fungi (Monjarás Feria and Valvano, 2020; Trunk et al., 2018). While we cannot exclude the incidental enrichment of T6SS components with taxonomic changes in the microbiome, this opens the possibility of a direct inhibition of *F. culmorum* by members of the microbial community which are diminished across dilutions.

4.3.5 - Biosynthetic gene clusters and metagenome-assembled genomes associated with suppressiveness

Secondary metabolites are often associated with direct antagonistic activity against fungal pathogens, which can result in disease suppression (Gomez-Exposito et al., 2017; Kwak and Weller, 2013; Schlatter et al., 2017). Microorganisms produce an incredible diversity of bioactive compounds, the production of which is encoded in biosynthetic gene clusters (BGCs). In this study, we identified 42,170 differentially abundant contigs using the DESeq2 package. AntiSMASH prediction of BGCs within this set of contigs yielded 504 BGCs of various classes (Figure 4-5). Statistical analysis of the distribution of BGC classes in this set versus the whole metagenome showed the enrichment of BGC classes encoding melanin, aryl polyenes and ladderanes in the lower dilutions (Table S4). Melanins in fungi are mainly associated with virulence and protection against physical and chemical stresses (Cordero and Casadevall, 2017). However, they can also be produced extracellularly by diverse bacterial taxa, including Streptomycetes, which were already found to decrease in abundance at higher ESM dilutions (Figure 4-4) (El-Naggar and El-Ewasy, 2017).

Accordingly, the melanin clusters found in this group share high similarity to other melanin BGCs found in multiple *Streptomyces* strains. Melanins appear to have a role in root colonization by streptomycetes (Chewning et al., 2019) and considering the relationship between rhizosphere population density and disease suppression (Johnson, 1994; Raaijmakers and Weller, 1998), melanins may indirectly contribute to the disease-suppressive phenotype by increasing the fitness of streptomycete strains involved in suppression of the pathogen. The enrichment of aryl polyenes and ladderanes is discussed in the Supplementary Material.

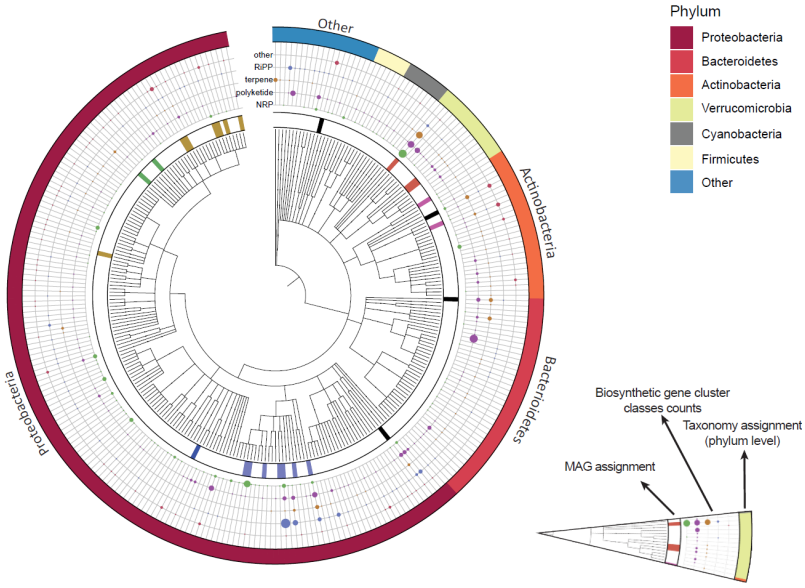


Figure 4-5. Taxonomical representation of 504 differentially abundant contigs containing biosynthetic gene clusters from the dilution metagenomes. A phylogenetic tree based on their taxonomical annotation is used to order the contigs. In this visualization, contigs assigned to the same taxa are grouped together. The innermost circle (MAG assignment) represents the taxonomic groups containing contigs assigned to a specific metagenome-assembled genome. Contigs assigned to the same MAG are marked with the same color (further description in Figure 4-7 for a color legend) or with black for MAGs containing only one BGC-harboring contig. For example, the contigs assigned to the *L. luteola* (Polyangiacee) MAG at approximately 180 degrees are shown in blue both in this inner ring and in Fig 4-7. Further, the major antiSMASH predicted BGC classes in the contigs are indicated with dots. The number of predicted BGCs is reflected by the size of the dot. Finally, the outer ring represents the contig taxonomical assignment at the phylum level.

Considering that BGC classes are generally broad and functionally heterogeneous, we focused our attention on specific gene clusters for which the abundance strongly

correlates with the changes in the suppressive phenotype. Among them, we noted BGCs with high similarity to known gene clusters involved in the biosynthesis of the siderophores delftibactin, pyoverdine, and myxochelin, as well as several BGCs putatively encoding the production of unknown siderophores, based on identification of genes involved in the biosynthesis of chelating amino acids and/or the presence of TonB-dependent siderophore receptors. We also identified multiple fragments of actinobacterial BGCs with similarity to gene clusters encoding the production of metabolites with reported antifungal properties such as those encoding the production of caniferolides / cyphomycin (Pérez-Victoria et al., 2019) and himastatin / kutznerides (Zhang et al., 2020). Full antiSMASH results can be accessed online ("AntiSMASH results," 2021). However, given such a large collection of potential targets, which in most cases cannot be associated to previously characterized gene clusters in the MIBiG database, we first prioritized those BGCs present in the metagenome-assembled genomes (MAGs), as the MAGs provide additional genomic and taxonomic context (Figure 4-5). Manual curation of individual bins resulted in a collection of 96 MAGs. Then, we analyzed the MAG collection looking for BGCs that become differentially abundant along the ESM dilution trajectory. This resulted in 9 MAGs that were strongly associated with disease suppressiveness and belong to the bacterial families *Burkholderiaceae*, *Acidimicrobiaceae*, *Verrucomicrobiaceae*, and the order Polyangiales (Figure 4-5 and Figure 4-6). Among the 9 MAGs, the ones assigned to the genera *Labilithrix* (Polyangiales) and *Prosthecobacter* (*Verrucomicrobiaceae*) belong to taxa that displayed a significant decrease in abundance at higher dilutions (table S2). We then linked the KEGG orthologues enriched genes to the MAGs and found that out of the 9 MAGs, only the *Labilithrix* bin contained genes from the previously discussed categories such as chitinases, siderophore-associated genes and core-T6SS-associated genes (Figure 4-8). Additionally, we found multiple copies of the T6SS component FHA (Forkhead-associated domain) that was enriched in the dilutions with a suppressive phenotype within the same *Labilithrix* MAG. One particular multimodular NRPS gene cluster stood out, as it is predicted to encode biosynthesis of a peptide with 11 modules, does not have a thioesterase termination domain in the BGC region and lies on a contig edge, suggesting that it is a fragment of an even larger BGC (Figure 4-7). AntiSMASH clusterblast and knownclusterblast results showed little to no similarity

to previously encountered and classified clusters, including those from its closest representative in the NCBI database. *Labilithrix* and closely related bacteria were, so far, never associated with disease suppressiveness but have been recognized as producers of antibacterial, antifungal and antiviral agents (Mulwa et al., 2018; Weissman and Müller, 2009; Yamamoto et al., 2014). Future experiments on this genus should aim at the identification of BGCs expressed on roots of wheat challenged with *F. culmorum* and the expression levels correlating with the disease suppressive phenotype. Additionally, previous studies successfully established synthetic communities from disease-suppressive endosphere isolates that were able to reproduce the suppressive phenotype (Carrión et al., 2019), a validation technique that could be repeated in this context. Here, we describe a set of 9 MAGs which constitute suitable candidates for initial leads in this direction.

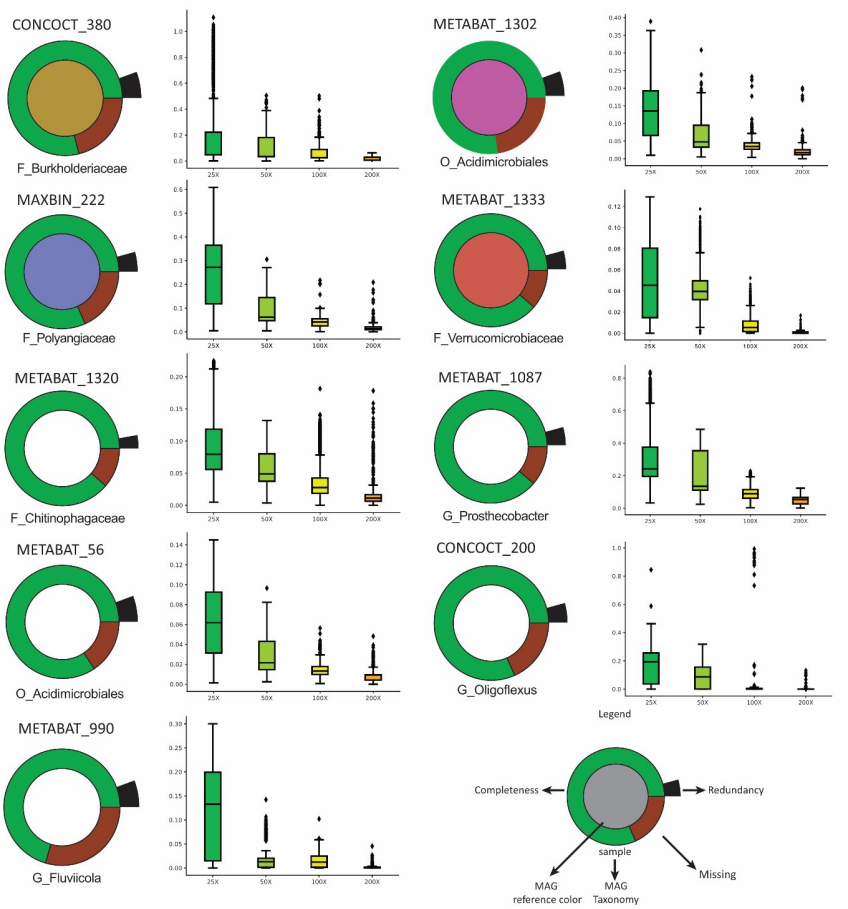


Figure 4-6. Representation of the MAGs from the metagenome showing significant association with the suppressive phenotype. The MAGs shown here are constituted for 50% or more in length by differentially abundant contigs according to DESeq2 analysis. Each ring shows an overview of a MAGs assembly completeness and redundancy. For each MAG, a boxplot shows the abundance of all its contigs across the dilutions. MAG reference colors refer to Figure 4-5. The taxonomic levels are inndicated by a capital letters: O – Order, F -family, and G – genus.

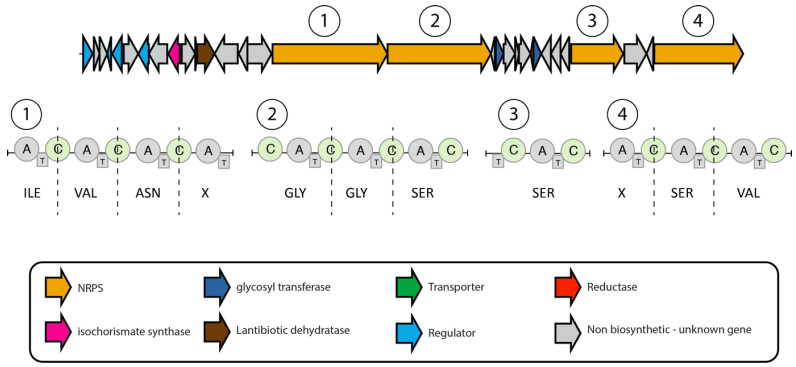


Figure 4-7. Large NRP gene cluster within *Labilithrix* MAG with predicted amino acid specificities for each module. The cluster lays on a contig edge and lacks a TE domain, and is therefore considered incomplete.

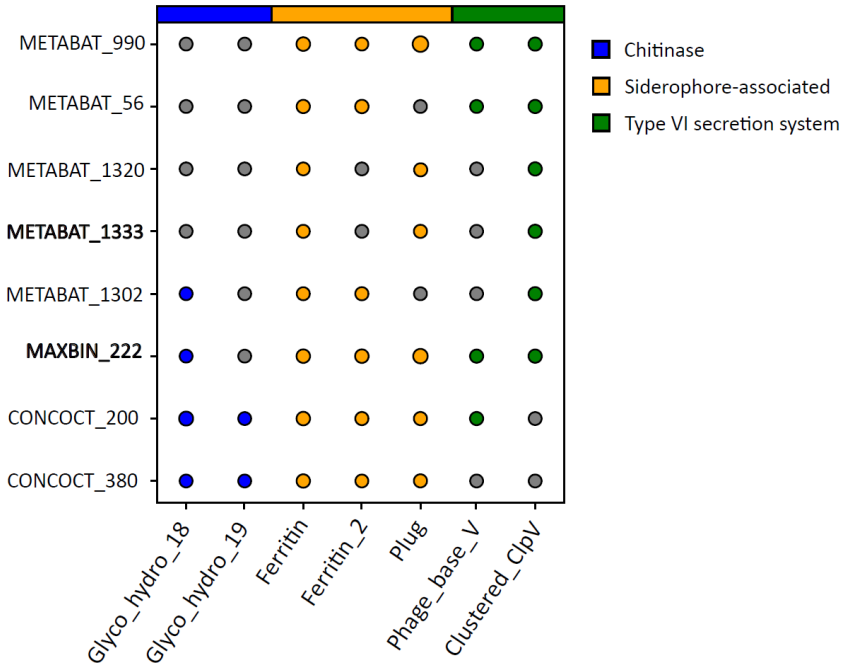


Figure 4-8. Presence (colored dot) absence (gray dot) patterns of domains contained in KEGG orthologs genes enriched in suppressive dilution points in the MAG collection. Bacterial chitinase families are shown in blue, siderophores associated domains in yellow and core-representative-components of type VI secretion system in green. Presence of both Phage_base_V and co-occurrence of all 4 ClpV subdomains (AAA, AAA_2, AAA_lid_9 and ClpB_D2-small) on the same contig are considered as minimal requirements for putative presence of type VI secretion system.

4.3.6 - Conclusions

We analyzed the microbiome of a disease suppressive soil by integrating a dilution-to-extinction approach with shotgun metagenomics. The dilution of the rhizosphere microbial community of the *Fusarium*-suppressive soil led to the decline of the plant protective effect. The dilution effect produced significant changes in the abundance of multiple bacterial taxa and in overall diversity, but these changes could not directly be matched to shifts in the microbiome functional profile. Nevertheless, key genes and gene clusters associated with chitin degradation and siderophore biosynthesis were associated with the disease suppressive phenotype. These genes were also detected in several metagenome assembled genomes (MAGs) that showed considerable yet unknown secondary metabolite production potential. Altogether, this work represents the first example of the dissection of the rhizosphere microbiome of a disease-suppressive soil using a dilution-to-extinction. This approach could be instrumental in future studies for discovering the mechanisms involved in other soil functions.

4.4 - Materials and Methods

4.4.1 - Soils used in the study

Soil S11 was collected from an agricultural field near Bergen op Zoom, the Netherlands in March 2018. This soil exhibited a high level of disease suppressiveness against *Fusarium culmorum,* which was evaluated in our previous study (Ossowicki et al., 2020). Furthermore, we used a sandy dune soil (BS) collected near Bergharen, the Netherlands (Schulz-Bohm et al., 2015). After collecting soils were air-dried in room temperature, sieved through 4 mm sieve, and stored at 4°C. BS soil was additionally gamma sterilized (Synergy Health Ede B.V., The Netherlands) before use.

4.4.2 - Seed preparation and growth conditions

JB Asano wheat seeds (Agrifirm, The Netherlands) were surface sterilized and pre-germinated in order to use in the experiment. Briefly, after surface sterilization with 70% ethanol and 10% bleach, seeds were washed with an excess of sterile water and placed on a wet 1mm paper filter (VWR, the Netherlands). Three-day-old seedlings were transferred to the soil, watered every second day and supplemented weekly with 0,5 Hoagland solution without microelements (0.5 M Ca(NO3)2·4H2O, 1 M KNO3, 1 M KH2PO4, 0.5 M MgSO4·7H2O and 98.6 mM ferric EDTA). Plants were grown in cabinets (MC 1750 VHO-EVD, Snijders Labs) at 20°C day and night, photoperiod 12 h day/12 h night with 60% relative humidity.

4.4.3 - Inoculation and disease suppressiveness evaluation

Soilborne pathogen *F. culmorum* PV (de Boer et al., 1998) was grown on ¼ strength PDA media (Oxoid, The Netherlands) for 2 weeks at 20°C in order to use in the experiment. Where indicated, the pathogen was introduced to the soil before seedling transfer as 6mm mycelium plugs (1 per 10cc of soil). After the experiment wheat plants were gently removed from the soil and cleaned. The disease symptoms on the roots were assessed according at a scale from 0- healthy to 5-severely diseased, like we described before (Ossowicki et al., 2020). Significance of differences in disease symptoms between treatments and control were assessed using a chi-square test, with a cut-off of $p<0.05$.

4.4.4 - Microbiome extraction

Wheat plants were grown in four 130 cc pots filled with S11 for two weeks. In order to perform a liquid extraction of the microbiome from the suppressive soil S11, all the glassware, materials and buffer were autoclaved. Plants were gently removed from the pots and the soil with the whole root system was transferred to a glass bottle with 10 volumes of phosphate buffer (10 mM KH_2PO_4, pH=6.5) and 5g of sterile glass beads. These bottles were shaken using an orbital shaker for 1 h at 180 rpm, sonicated for 15 min and shaken again for 1h to detach microorganisms from

soil particles. Afterwards, all the extracts were merged in one beaker and sieved through a metal sieve to separate bigger soil particles and roots. Subsequently, extract was briefly centrifuged (1000 x g, 1 min) and the supernatant was filtered again using a Buchner funnel through a filter paper (VWR Grade 415 Filter Paper, VWR the Netherlands). The extract obtained this way was considered a 10X diluted Extracted Soil Microbiome (ESM) because the soil was extracted with 10 volumes of the buffer. In parallel, as a control, the same procedure was applied with the sterile soil BS with a sterile vermiculite (Agra-vermiculite, The Netherlands) in a 1 to 1 volume ration. This way a BS control extract was obtained.

4.4.5 - Dilution to extinction - introduction to a sterile soil

In order to optimize the magnitude of the dilution we performed preliminary experiments (data not shown) investigating first, if dilution affects soil disease suppressiveness to *F. culmorum* (the phenotype), and second, at which dilution point the change in phenotype from suppressive to conducive occurs. After the preliminary experiments we chose the dilution steps in such a way that the phenotype tipping point occurs centrally. Compared to the dilution steps used in other studies (Chen et al., 2020; Hol et al., 2015; Korenblum et al., 2020; Yan et al., 2015), we applied a relatively small dilution magnitude. Nevertheless, in the experimental system we present in this work, these small dilution steps allow us to observe a gradual change of the phenotype.

The 10X dilution of the ESM was serially diluted with phosphate buffer to 25X, 50X, 100X, 200X and 400X and stored at 4°C. The dilution range was determined in a preliminary experiment (data not shown). Plastic, 100 cc pots were filled with the sterile soil BS mixed with sterile vermiculite in a volume ratio of 1 to 1 and with mycelium plugs of *F. culmorum* or sterile ¼ PDA plugs (mock inoculation) according to the treatment. Distinct dilutions of the ESM and the BS control extract were added to pots, 20 ml per day for seven days, before wheat seedlings were introduced. Plants were grown for 3 weeks before the disease assessment and a rhizosphere DNA isolation.

4.4.6 - Dilution to extinction - experimental setup

All the dilution treatments and controls consisted of sixteen biological replicates. Plants were grown in a randomized order to minimize spatial bias. A negative control and a positive control, with and without the pathogen, respectively, were used to control for the infectiousness of the pathogen and the effect of the sterile soil, buffer, and the extraction procedure. The control S11 contained natural suppressive soil S11 with the pathogen introduced to control for the soil disease suppressiveness of the soil that was used for microbiome extraction. The experiment was performed twice, and the rhizosphere samples used for the DNA extraction were collected in the second experiment from randomly selected replicates.

4.4.7 - Rhizosphere DNA extraction

Five out of sixteen replicates were randomly selected for a rhizosphere DNA extraction before the disease assessment. Plants were gently removed from the pots and the soil loosely adhering to the roots was removed by shaking. The whole root system was placed in a 50 ml falcon tube with 25 ml of sterile MQ grade water. The tubes were vortexed, sonicated, and vortexed again, each step for 1 min. Afterwards, the roots were removed with sterile forceps and disease symptoms were assessed on them. The water with extracted rhizosphere was briefly spined down (1000 x g, 1 min) and carefully decanted to a new 50 ml falcon tube. The decanted liquid was frozen at -20 °C and freeze dried overnight (Free zone 12, Labconco the USA). This resulted in forming a white powder in the tube which was collected and extracted using a DNeasy PowerSoil Kit (QIAGEN, the Netherlands) according to the manufacturer's protocol, without applying a bead beating step. Samples were subsequently purified using the DNeasy PowerClean cleanup kit (QIAGEN, the Netherlands).

4.4.8 - Data quality control, assembly, mapping and binning

Raw paired sequence reads were quality-checked with (Andrews, 2015) and trimmed using bbduk (Bushnell 2017) with a Phred score threshold of 30. Read pairs for which at least one mate was shorter than 150 bp were discarded. Reads for all dilution points were co-assembled in a single assembly using Megahit v1.2.9 (Li et al., 2016). Genes were predicted using prodigal v2.6.3 (Hyatt et al., 2010) in metagenomic mode. Reads were mapped back onto the assembly contigs using HISAT2 (Kim et al., 2019). BAM files were converted into SAM using samtools v1.7 and RPKM counts were obtained with mpileup (Li et al., 2009). Binning was performed within Anvi'o environment v6.2 (Eren et al., 2015) using DASTOOL (Sieber et al., 2018) refinement of bins obtained with MAXBIN2 v2.2.7 (Wu et al., 2016), CONCOCT v1.1.0 (Qian and Comin, 2019) and METABAT v2.15 (Kang et al., 2019). MAGs with a DASTOOL confidence score >= 0.5 were kept.

4.4.9 - Enrichment analysis, biosynthetic gene cluster annotation, reads taxonomy annotation and KEGG functional annotation

Identification of differentially abundant contigs was performed in R using DESeq v1.18.1 (Love et al., 2014, p. 2) looking for enrichment of contigs in the first two dilution points (25X and 50X) versus the higher dilution points (100X and 200X). Contigs with adjusted p-value < 0.05 were considered differentially abundant. Biosynthetic gene clusters were predicted from differentially abundant contigs with a positive fold change using antiSMASH 5.1.2 (Blin et al., 2019) using the prodigal-meta option for gene prediction, --cb-general and --cb-knownclusters for comparison of predicted clusters to the antiSMASH database (Blin et al., 2020) and MIBiG (Kautsar et al., 2020). Quality-trimmed reads were taxonomically annotated using Kraken v2.0.8 (Wood et al., 2019) against the GTDB-taxonomy database release 05-RS95 (Parks et al., 2020). KEGG annotation was performed at Baseclear (Leiden, The Netherlands) using the full Megahit assembly.

4.4.10 - Phylogeny tree construction and annotation

Differentially abundant contigs which had at least one predicted cluster were taxonomically annotated using Diamond v 0.9.21 (Buchfink et al., 2015) against the NCBI nr database Oct. 2020 (NCBI Resource Coordinators, 2016). Contigs were assigned to the taxonomical group with the highest cumulative score across the contig. Contigs annotated with the same taxonomy were grouped for tree construction. The list of taxonomic groups was used in NCBI common tree to obtain a tree in PHYLIP format representing the phylogeny relations of the taxonomies present in the differentially abundant contigs with biosynthetic gene clusters shown in Figure 4-5. Annotation of the tree was done with iTOL (Letunic and Bork, 2019); a fully annotated tree is available at

https://itol.embl.de/tree/88731562323671816060863 31#.

Presence-absence of KEGG orthologs enriched domains in MAGs was determined using hmmsearch v3.1b2 (Mistry et al., 2013) and hmm domains from pfam (El-Gebali et al., 2019) against the different MAGs using trusted cutoffs. Presence of ClpV was determined based on the co-occurrence of all 4 ClpV subdomains (AAA, AAA_2, AAA_lid_9 and ClpB_D2-small) on the same contig. ClpV and Phage_base_V are considered core components of type VI secretion system (Boyer et al., 2009).

4.5 - Supplementary material

4.5.1 - General functional profiles of the rhizosphere microbiomes are resilient to dilution

We grouped KEGG orthologs by their KEGG Orthology hierarchical classification and aggregated their RPKM counts to generate functional profiles for each sample per KO category. Interestingly, no category with a sufficient number of entries was found to be significantly different across the dilutions (Figure S3), including categories which contain differentially abundant KOs.

From these observations, we deduce that, despite the changes in the taxonomic profiles, the KO functional profiles of the communities are generally stable at the level of these broad categories across dilutions, with no statistical differences (Figure S3). This may be connected to the strong selective pressure of the environment, which restricts members of the community into a well-defined functional profile. In our experiment, every rhizosphere community occupied highly similar niche; therefore, we can conclude that the general functional profiles could also be highly similar. We can speculate that further dilutions of the ESM inoculum would have a stronger effect on the metabolic profiles, as shown in (Yan et al., 2017). Nevertheless, it would also disturb the community in considerable way and our intention in this experiment was to track the changes in microbial communities related to the change in disease suppressiveness which were shown to be significant between these dilution points. The loss of the suppressive phenotype across the dilutions without significant changes in functional profiles suggests that the origin of this phenomenon is not related to specific community-wide microbiome features but to specific “silver bullets” which cannot be tracked with large scale functional analysis.

Because of the character of KEGG database (Kanehisa et al., 2017) we can only draw conclusions about the general functional profiles of the soil microbial communities. To understand the specialized secondary metabolism of soil microbiomes, which was recognized to be crucial for soil suppressiveness in a number of studies (Gomes Exposito et al., 2017; Kwak and Weller, 2013; Schlatter et al., 2017), we need to use other approaches. To this end, we decided to expand the functional analysis with prediction and identification of biosynthetic gene clusters (BGC) using the antiSMASH pipeline (Blin et al., 2019).

4.5.2 - Additional enriched gene cluster classes in suppressive rhizosphere

In addition to melanins, two more BGC product classes were significantly more present in suppressive enriched contigs compared to the rest of the metagenome: aryl polyenes and ladderane. The functional role of aryl polyenes produced by

rhizosphere bacteria is so far unknown. This class of antioxidants is similar to carotenoids in structure and biochemical function and likely protect bacteria from reactive oxygen species produced by the plant as a response to infection (Lehmann et al., 2015). The synthesis genes of aryl polyenes are abundant among Proteobacteria but not in Alphaproteobacteria and Gammaproteobacteria (Schöner et al., 2016). Since Alphaproteobacteria are the main contributors to the increased abundance of whole phylum in high dilutions, it is likely that also the enrichment of the aryl polyene BGCs is related to the abundance of associated taxa.

Laddreanes are membrane lipids present in anammoxosomes – intracellular compartments of ammonium-oxidizing bacteria (Hancock and Brown, 2021). Since, these lipids are mostly found in Planctomycetes (Sinninghe Damsté et al., 2005) we can attribute the ladderane BGCs enrichment to the abundance of this group (Figure 4-4). Ladderanes are predominantly found in marine environments and their role in rhizosphere is unknown. Despite their enrichment in suppressive associated contigs, these biosynthetic gene cluster classes have no known association with antifungal activity nor provide fitness competitive advantage over the pathogen. Therefore, we consider this enrichment incidental to the relative decrease in abundance of the corresponding producing taxonomic groups.

4.6 – Supplementary Figures

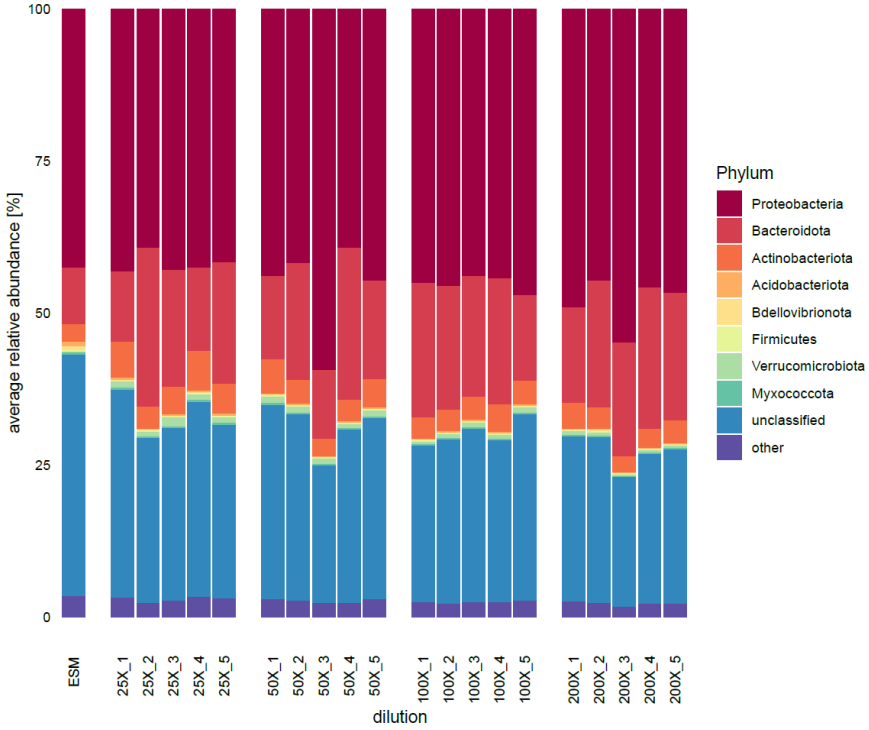


Figure S1. Barplot showing the most abundant phyla in the ESM and rhizosphere across four dilution points in five replicates.

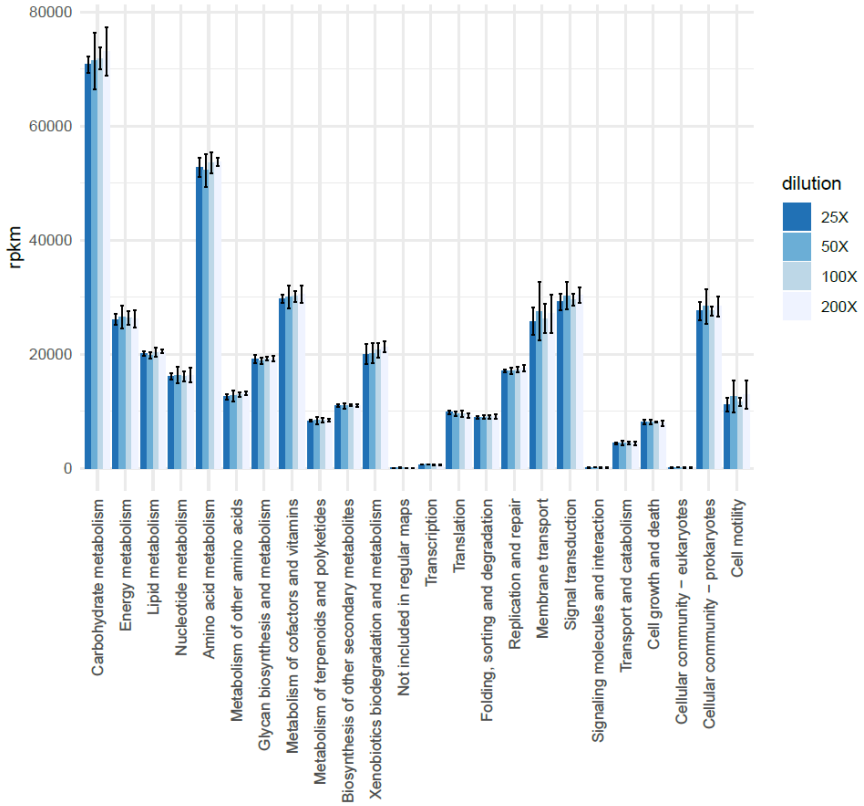


Figure S2. Graph summarising KEGG orthology annotations of rhizosphere metagenomes in four dilutions. The values on the y axis are expressed in reads per kilo base per million mapped reads (rpkm). Error bars show the standard deviation of the mean in five replicates.

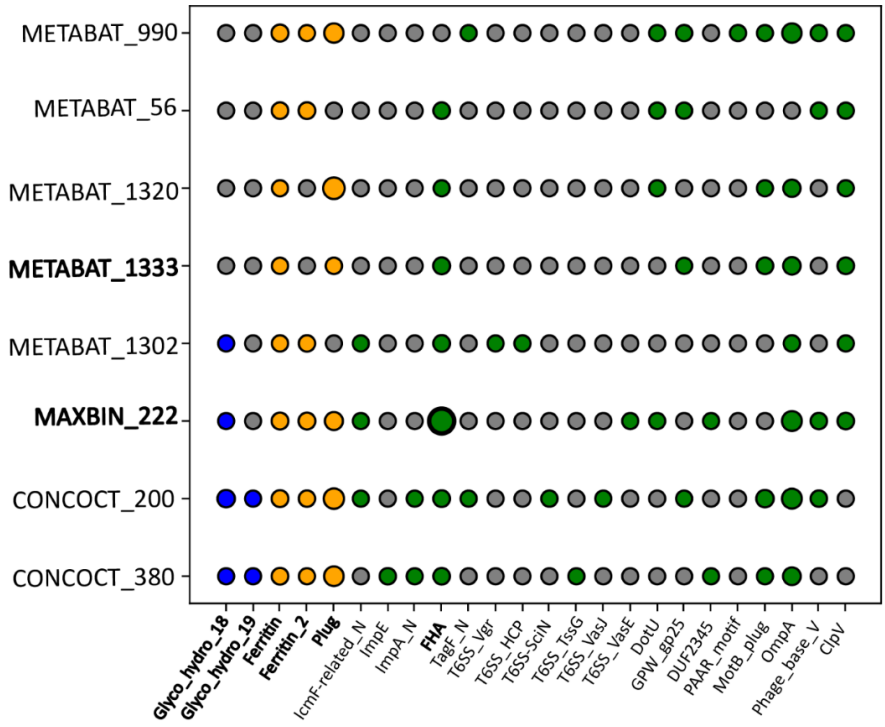


Figure S3. Presence (colored dot) absence (gray dot) patterns of domains contained in KEGG orthologues genes enriched in suppressive dilution points in the MAG collection. Bacterial chitinase families are shown in blue, siderophores associated domains in yellow and components of type VI secretion system in green. ClpV is considered present (last column) if all 4 ClpV subdomains (AAA, AAA_2, AAA_lid_9 and ClpB_D2-small) cooccur on the same contig. FHA domain counts for *Labilithrix* MAG is shown in bold. Here, multiple core and accessory components of the type VI secretion system are shown.

Supplementary Tables available on Zenodo at
https://doi.org/10.5281/zenodo.5636347

References

Andrews, S., 2015. FastQC: A Quality Control Tool for High Throughput Sequence Data [Online] [WWW Document]. URL https://qubeshub.org/resources/fastqc

AntiSMASH results [WWW Document], 2021. URL https://bioinformatics.nl/~traca001/rhizo_antismash/antismash_de_contigs/ (accessed 1.19.21).

Berg, G., Köberl, M., Rybakova, D., Müller, H., Grosch, R., Smalla, K., 2017. Plant microbial diversity is suggested as the key to future biocontrol and health trends. FEMS Microbiology Ecology 93.

Berg, G., Smalla, K., 2009. Plant species and soil type cooperatively shape the structure and function of microbial communities in the rhizosphere. FEMS Microbiol Ecol 68, 1–13.

Blin, K., Shaw, S., Kautsar, S.A., Medema, M.H., Weber, T., 2020. The antiSMASH database version 3: increased taxonomic coverage and new query features for modular enzymes. Nucleic Acids Res.

Blin, K., Shaw, S., Steinke, K., Villebro, R., Ziemert, N., Lee, S.Y., Medema, M.H., Weber, T., 2019. antiSMASH 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic Acids Res 47, W81–W87.

Boyer, F., Fichant, G., Berthod, J., Vandenbrouck, Y., Attree, I., 2009. Dissecting the bacterial type VI secretion system by a genome wide in silico analysis: what can be learned from available microbial genomic resources? BMC Genomics 10, 104.

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

Bushnell, B., n.d. BBMap [WWW Document]. URL sourceforge.net/projects/bbmap/

Carrión, V.J., Perez-Jaramillo, J., Cordovez, V., Tracanna, V., Hollander, M. de, Ruiz-Buck, D., Mendes, L.W., Ijcken, W.F.J. van, Gomez-Exposito, R., Elsayed, S.S., Mohanraju, P., Arifah, A., Oost, J. van der, Paulson, J.N., Mendes, R., Wezel, G.P. van, Medema, M.H., Raaijmakers, J.M., 2019. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science 366, 606–612.

Chapelle, E., Mendes, R., Bakker, P.A.H., Raaijmakers, J.M., 2016. Fungal invasion of the rhizosphere microbiome. The ISME Journal 10, 265–268.

Chen, Q.-L., An, X.-L., Li, H., Zhu, Y.-G., Su, J.-Q., Cui, L., 2017. Do manure-borne or indigenous soil microorganisms influence the spread of antibiotic resistance genes in manured soil? Soil Biology and Biochemistry 114, 229–237.

Chen, Q.-L., An, X.-L., Zheng, B.-X., Gillings, M., Peñuelas, J., Cui, L., Su, J.-Q., Zhu, Y.-G., 2019. Loss of soil microbial diversity exacerbates spread of antibiotic resistance. Soil Ecol. Lett. 1, 3–13.

Chen, Q.-L., Ding, J., Zhu, Y.-G., He, J.-Z., Hu, H.-W., 2020. Soil bacterial taxonomic diversity is critical to maintaining the plant productivity. Environment International 140, 105766.

Chewning, S.S., Grant, D.L., O'Banion, B.S., Gates, A.D., Kennedy, B.J., Campagna, S.R., Lebeis, S.L., 2019. Root-Associated Streptomyces Isolates Harboring melC Genes Demonstrate Enhanced Plant Colonization. Phytobiomes Journal 3, 165–176.

Cordero, R.J., Casadevall, A., 2017. Functions of fungal melanin beyond virulence. Fungal Biol Rev 31, 99–112.

Cordovez, V., Carrion, V.J., Etalo, D.W., Mumm, R., Zhu, H., van Wezel, G.P., Raaijmakers, J.M., 2015. Diversity and functions of volatile organic compounds produced by Streptomyces from a disease-suppressive soil. Front. Microbiol. 6.

Cordovez, V., Dini-Andreote, F., Carrión, V.J., Raaijmakers, J.M., 2019. Ecology and Evolution of Plant Microbiomes. Annu. Rev. Microbiol. 73, 69–88.

Costa, O.Y.A., Raaijmakers, J.M., Kuramae, E.E., 2018. Microbial Extracellular Polymeric Substances: Ecological Function and Impact on Soil Aggregation. Front. Microbiol. 9.

Coulthurst, S.J., 2013. The Type VI secretion system – a widespread and versatile cell targeting system. Research in Microbiology, Bacterial secretion systems: function and structural biology 164, 640–654.

de Boer, W., Klein Gunnewiek, P.J.A., Lafeber, P., Janse, J.D., Spit, B.E., Woldendorp, J.W., 1998. Anti-fungal properties of chitinolytic dune soil bacteria. Soil Biology and Biochemistry 30, 193–203.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A., Sonnhammer, E.L.L., Hirsh, L., Paladin, L., Piovesan, D., Tosatto, S.C.E., Finn, R.D., 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47, D427–D432.

El-Naggar, N.E.-A., El-Ewasy, S.M., 2017. Bioproduction, characterization, anticancer and antioxidant activities of extracellular melanin pigment produced by newly isolated microbial cell factories Streptomyces glaucescens NEAE-H. Sci Rep 7.

Eren, A.M., Esen, Ö.C., Quince, C., Vineis, J.H., Morrison, H.G., Sogin, M.L., Delmont, T.O., 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. PeerJ 3, e1319.

Garland, J.L., Lehman, R.M., 1999. Dilution/extinction of community phenotypic characters to estimate relative structural diversity in mixed communities. FEMS Microbiology Ecology 30, 333–343.

Gomes Exposito, R., de Bruijn, I., Postma, J., Raaijmakers, J.M., 2017. Current Insights into the Role of Rhizosphere Bacteria in Disease Suppressive Soils. Front. Microbiol. 8.

Hancock, E.N., Brown, M.K., 2021. Ladderane Natural Products: From the Ground Up. Chemistry – A European Journal 27, 565–576.

Hjort, K., Presti, I., Elväng, A., Marinelli, F., Sjöling, S., 2014. Bacterial chitinase with phytopathogen control capacity from suppressive soil revealed by functional metagenomics. Appl Microbiol Biotechnol 98, 2819–2828.

Hol, W.H.G., Garbeva, P., Hordijk, C., Hundscheid, M.P.J., Gunnewiek, P.J.A.K., Agtmaal, M. van, Kuramae, E.E., Boer, W. de, 2015. Non-random species loss in bacterial communities reduces antifungal volatile production. Ecology 96, 2042–2048.

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

Johnson, K.B., 1994. Dose-Response Relationship and Inundative Biological Control. Phytopathology.

Jousset, A., Schulz, W., Scheu, S., Eisenhauer, N., 2011. Intraspecific genotypic richness and relatedness predict the invasibility of microbial communities. The ISME Journal 5, 1108–1114.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., Morishima, K., 2017. KEGG: new perspectives on genomes, pathways, diseases and drugs. Nucleic Acids Res 45, D353–D361.

Kang, D., Li, F., Kirton, E.S., Thomas, A., Egan, R.S., An, H., Wang, Z., 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies (No. e27522v1). PeerJ Inc.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., Medema, M.H., 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Res 48, D454–D458.

Kenters, N., Henderson, G., Jeyanathan, J., Kittelmann, S., Janssen, P.H., 2011. Isolation of previously uncultured rumen bacteria by dilution to extinction using a new liquid culture medium. J Microbiol Methods 84, 52–60

Kim, D., Paggi, J.M., Park, C., Bennett, C., Salzberg, S.L., 2019. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nature Biotechnology 37, 907–915

Kinkel, L.L., Schlatter, D.C., Bakker, M.G., Arenz, B.E., 2012. Streptomyces competition and co-evolution in relation to plant disease suppression. Res. Microbiol. 163, 490–499.

Korenblum, E., Dong, Y., Szymanski, J., Panda, S., Jozwiak, A., Massalha, H., Meir, S., Rogachev, I., Aharoni, A., 2020. Rhizosphere microbiome mediates systemic root metabolite exudation by root-to-root signaling. PNAS.

Kwak, Y.S., Weller, D.M., 2013. Take-all of Wheat and Natural Disease Suppression: A Review. Plant Pathology J 29, 125–135.

Lacombe-Harvey, M.-È., Brzezinski, R., Beaulieu, C., 2018. Chitinolytic functions in actinobacteria: ecology, enzymes, and evolution. Appl Microbiol Biotechnol 102, 7219–7230.

Lagier, J.-C., Hugon, P., Khelaifia, S., Fournier, P.-E., Scola, B.L., Raoult, D., 2015. The Rebirth of Culture in Microbiology through the Example of Culturomics To Study Human Gut Microbiota. Clinical Microbiology Reviews 28, 237–264.

Lehmann, S., Serrano, M., L'Haridon, F., Tjamos, S.E., Metraux, J.-P., 2015. Reactive oxygen species and plant resistance to fungal pathogens. Phytochemistry, In Memory of G. Paul Bolwell: Plant Cell Wall Dynamics 112, 54–62.

Lemanceau, P., Expert, D., Gaymard, F., Bakker, P.A.H.M., Briat, J.F., 2009. Role of Iron in Plant-Microbe Interaction. Advances in Botanical Research 51, 491–549.

Letunic, I., Bork, P., 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research 47, W256–W259.

Li, D., Luo, R., Liu, C.-M., Leung, C.-M., Ting, H.-F., Sadakane, K., Yamashita, H., Lam, T.-W., 2016. MEGAHIT v1.0: A fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods, Pan-omics analysis of biological data 102, 3–11.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup, 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25, 2078–2079.

Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550.

Mallon, C.A., Elsas, J.D. van, Salles, J.F., 2015. Microbial Invasions: The Process, Patterns, and Mechanisms. Trends in Microbiology 23, 719–729.

Mazurier, S., Corberand, T., Lemanceau, P., Raaijmakers, J.M., 2009. Phenazine antibiotics produced by fluorescent pseudomonads contribute to natural soil suppressiveness to Fusarium wilt. ISME J 3, 977–991.

Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., Punta, M., 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Research 41, e121–e121.

Monjarás Feria, J., Valvano, M.A., 2020. An Overview of Anti-Eukaryotic T6SS Effectors. Front Cell Infect Microbiol 10, 584751.

Morella, N.M., Weng, F.C.-H., Joubert, P.M., Metcalf, C.J.E., Lindow, S., Koskella, B., 2020. Successive passaging of a plant-associated microbiome reveals robust habitat and host genotype-dependent selection. Proc Natl Acad Sci USA 117, 1148–1159.

Mulwa, L.S., Jansen, R., Praditya, D.F., Mohr, K.I., Wink, J., Steinmann, E., Stadler, M., 2018. Six Heterocyclic Metabolites from the Myxobacterium Labilithrix luteola. Molecules 23.

Navarrete, A.A., Soares, T., Rossetto, R., van Veen, J.A., Tsai, S.M., Kuramae, E.E., 2015. Verrucomicrobial community structure and abundance as indicators for changes in chemical factors linked to soil fertility. Antonie Van Leeuwenhoek 108, 741–752.

NCBI Resource Coordinators, 2016. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 44, D7-19.

Ossowicki, A., Tracanna, V., Petrus, M.L.C., van Wezel, G., Raaijmakers, J.M., Medema, M.H., Garbeva, P., 2020. Microbial and volatile profiling of soils suppressive to Fusarium culmorum of wheat. Proceedings of the Royal Society B: Biological Sciences 287, 20192527.

Parks, D.H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A.J., Hugenholtz, P., 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. Nature Biotechnology 38, 1079–1086.

Pérez-Victoria, I., Oves-Costales, D., Lacret, R., Martín, J., Sánchez-Hidalgo, M., Díaz, C., Cautain, B., Vicente, F., Genilloud, O., Reyes, F., 2019. Structure elucidation and biosynthetic gene cluster analysis of caniferolides A–D, new bioactive 36-membered macrolides from the marine-derived Streptomyces caniferus CA-271066. Org. Biomol. Chem. 17, 2954–2971.

Peter, H., Beier, S., Bertilsson, S., Lindström, E.S., Langenheder, S., Tranvik, L.J., 2011. Function-specific response to depletion of microbial diversity. ISME J 5, 351–361.

Qian, J., Comin, M., 2019. MetaCon: unsupervised clustering of metagenomic contigs with probabilistic k-mers statistics and coverage. BMC Bioinformatics 20, 367.

Raaijmakers, J.M., Mazzola, M., 2011. Diversity and Natural Functions of Antibiotics Produced by Beneficial and Plant Pathogenic Bacteria. Annu. Rev. Phytopathol. 50, 403–424.

Raaijmakers, J.M., Weller, D.M., 1998. Natural Plant Protection by 2,4-Diacetylphloroglucinol-Producing Pseudomonas spp. in Take-All Decline Soils. MPMI 11, 144–152.

Records, A.R., 2011. The type VI secretion system: a multipurpose delivery system with a phage-like machinery. Mol Plant Microbe Interact 24, 751–757.

Rodriguez-R, L.M., Gunturu, S., Tiedje, J.M., Cole, J.R., Konstantinidis, K.T., 2018. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. mSystems 3.

Schlatter, D., Kinkel, L., Thomashow, L., Weller, D., Paulitz, T., 2017. Disease Suppressive Soils: New Insights from the Soil Microbiome. Phytopathology 107, 1284–1297.

Schöner, T.A., Gassel, S., Osawa, A., Tobias, N.J., Okuno, Y., Sakakibara, Y., Shindo, K., Sandmann, G., Bode, H.B., 2016. Aryl Polyenes, a Highly Abundant Class of Bacterial Natural Products, Are Functionally Related to Antioxidative Carotenoids. ChemBioChem 17, 247–253.

Schulz-Bohm, K., Zweers, H., de Boer, W., Garbeva, P., 2015. A fragrant neighborhood: volatile mediated bacterial interactions in soil. Front. Microbiol. 6.

Shen, Z., Ruan, Y., Xue, C., Zhong, S., Li, R., Shen, Q., 2015. Soils naturally suppressive to banana Fusarium wilt disease harbor unique bacterial communities. Plant and Soil 393, 21–33.

Sieber, C.M.K., Probst, A.J., Sharrar, A., Thomas, B.C., Hess, M., Tringe, S.G., Banfield, J.F., 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nature Microbiology 3, 836–843.

Siegel-Hertz, K., Edel-Hermann, V., Chapelle, E., Terrat, S., Raaijmakers, J.M., Steinberg, C., 2018. Comparative Microbiome Analysis of a Fusarium Wilt Suppressive Soil and a Fusarium Wilt Conducive Soil From the Chateaurenard Region. Front. Microbiol. 9.

Sinninghe Damsté, J.S., Rijpstra, W.I.C., Geenevasen, J.A.J., Strous, M., Jetten, M.S.M., 2005. Structural identification of ladderane and other membrane lipids of planctomycetes capable of anaerobic ammonium oxidation (anammox). FEBS J 272, 4270–4283.

Stingl, U., Cho, J.-C., Foo, W., Vergin, K.L., Lanoil, B., Giovannoni, S.J., 2008. Dilution-to-extinction culturing of psychrotolerant planktonic bacteria from permanently ice-covered lakes in the McMurdo Dry Valleys, Antarctica. Microb Ecol 55, 395–405.

Tamietti, G., Alabouvette, C., 1986. Studies on the Disease Suppressiveness of Soil .13. Role of Nonpathogenic Fusarium-Oxysporum in the Wilt-Suppression Mechanisms of a Soil Form Noirmoutier. Agronomie 6, 541–548.

Tracanna, V., Ossowicki, A., Petrus, M.L.C., Overduin, S., Terlow, B.R., Lund, G., Robinson, S.L., Warris, S., Schijlen, E.G.W.M., van Wezel, G.P., Raaijmakers, J.M., Garbeva, P., Medema, M.H., n.d.

Dissecting disease suppressive rhizosphere microbiomes by functional amplicon sequencing and 10X metagenomics. mSystems.

Trunk, K., Peltier, J., Liu, Y.-C., Dill, B.D., Walker, L., Gow, N.A.R., Stark, M.J.R., Quinn, J., Strahl, H., Trost, M., Coulthurst, S.J., 2018. The Type VI secretion system deploys anti-fungal effectors against microbial competitors. Nat Microbiol 3, 920–931.

van Elsas, J.D., Chiurazzi, M., Mallon, C.A., Elhottovā, D., Krištůfek, V., Salles, J.F., 2012. Microbial diversity determines the invasion of soil by a bacterial pathogen. Proc Natl Acad Sci U S A 109, 1159–1164.

Veliz, E.A., Martínez-Hidalgo, P., Hirsch, A.M., 2017. Chitinase-producing bacteria and their role in biocontrol. AIMS Microbiol 3, 689–705.

Wagg, C., Bender, S.F., Widmer, F., van der Heijden, M.G.A., 2014. Soil biodiversity and soil community composition determine ecosystem multifunctionality. Proceedings of the National Academy of Sciences 111, 5266–5270.

Weissman, K.J., Müller, R., 2009. A brief tour of myxobacterial secondary metabolism. Bioorg Med Chem 17, 2121–2136.

Weller, D.M., Raaijmakers, J.M., Gardener, B.B.M., Thomashow, L.S., 2002. Microbial populations responsible for specific soil suppressiveness to plant pathogens. Annu Rev Phytopathol 40, 309-+.

Wood, D.E., Lu, J., Langmead, B., 2019. Improved metagenomic analysis with Kraken 2. Genome Biology 20, 257.

Wu, Y.-W., Simmons, B.A., Singer, S.W., 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605–607.

Yair, S., Yaacov, D., Susan, K., Jurkevitch, E., 2009. Small Eats Big: Ecology and Diversity of Bdellovibrio and Like Organisms, and their Dynamics in Predator-Prey Interactions, in: Lichtfouse, E., Navarrete, M., Debaeke, P., Véronique, S., Alberola, C. (Eds.), Sustainable Agriculture. Springer Netherlands, Dordrecht, pp. 275–284.

Yamamoto, E., Muramatsu, H., Nagai, K., 2014. Vulgatibacter incomptus gen. nov., sp. nov. and Labilithrix luteola gen. nov., sp. nov., two myxobacteria isolated from soil in Yakushima Island, and the description of Vulgatibacteraceae fam. nov., Labilitrichaceae fam. nov. and Anaeromyxobacteraceae fam. nov. International Journal of Systematic and Evolutionary Microbiology 64, 3360–3368.

Yan, Y., Kuramae, E.E., de Hollander, M., Klinkhamer, P.G.L., van Veen, J.A., 2017. Functional traits dominate the diversity-related selection of bacterial communities in the rhizosphere. The ISME Journal 11, 56–66.

Yan, Y., Kuramae, E.E., Klinkhamer, P.G.L., Veen, J.A. van, 2015. Revisiting the Dilution Procedure Used To Manipulate Microbial Biodiversity in Terrestrial Systems. Appl. Environ. Microbiol. 81, 4246–4252.

Youdkes, D., Helman, Y., Burdman, S., Matan, O., Jurkevitch, E., 2020. Potential Control of Potato Soft Rot Disease by the Obligate Predators Bdellovibrio and Like Organisms. Appl Environ Microbiol 86.

Yu, X., Polz, M.F., Alm, E.J., 2019. Interactions in self-assembled microbial communities saturate with diversity. The ISME Journal 13, 1602–1617.

Zhang, D., Lu, Y., Chen, H., Wu, C., Zhang, H., Chen, L., Chen, X., 2020. Antifungal peptides produced by actinomycetes and their biological activities against plant diseases. The Journal of Antibiotics 73, 265–282.

CHAPTER 5

Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome

Authors:

Víctor J. Carrión, Vittorio Tracanna*, Juan Perez-Jaramillo*, Viviane Cordovez*, Mattias de Hollander, Daniel Ruiz-Buck, Lucas W. Mendes, Wilfred F.J. van Ijcken, Ruth Gomez-Exposito, Somayah S. Elsayed, Prarthana Mohanraju, Adini Arifah, John van der Oost, Joseph N. Paulson, Rodrigo Mendes, Gilles P. van Wezel, Marnix H. Medema# and Jos M. Raaijmakers#

*: These authors contributed equally to this work.

#: corresponding author

Based on “Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome” published in *Science*, 366(6465):606-612 (2019).

5.1 – Abstract

Microorganisms living inside plants can promote plant growth and health, but their genomic and functional diversity remain largely elusive. Here, metagenomics and network inference showed that fungal infection of plant roots enriched for *Chitinophagaceae* and *Flavobacteriaceae* in the root endosphere and boosted expression of chitinase genes and various unknown biosynthetic gene clusters encoding the production of nonribosomal peptides (NRPS) and polyketides (PKS). Following strain-level genome reconstruction, a consortium of *Chitinophaga* and *Flavobacterium* was designed that consistently suppressed fungal root disease. Site-directed mutagenesis then revealed that a novel NRPS-PKS gene cluster from *Flavobacterium* was essential for disease suppression by the endophytic consortium. Our results highlight that endophytic root microbiomes harbor a wealth of yet unknown functional traits that, in concert, can protect the plant inside out.

5.2 – Introduction

Past and present plant microbiome studies have generated a large amount of sequence data and a wealth of (mostly) descriptive information on the diversity and relative abundance of different taxonomic groups in the rhizosphere, phyllosphere, spermosphere and endosphere of a multitude of plant species (Vorholt et al., 2017, Cordovez et al., 2019). To date, however, relatively few studies have demonstrated the functional importance of microbiomes for specific plant phenotypes, i.e., plant growth, development and health (Hacquard et al., 2015, Mendes et al., 2011, Panke-Buisse et al., 2015, Vannier et al., 2019, Oyserman et al., 2018, Duran et al., 2018). Furthermore, the molecular and chemical basis of the causal relationships between these plant phenotypes and microbiome structure and functions are in most cases still unknown. The aim of this study was to investigate the genomic diversity and functional potential of the endophytic root microbiome in protection of plants against fungal infections. To this end, we integrated multiple approaches including network inference and metagenomics to identify root endophytic bacterial consortia and functional gene clusters associated with a soil that is suppressive to disease caused

by *Rhizoctonia solani*, a fungal root pathogen of several plant species, including rice, wheat and sugar beet.

Disease-suppressive soils are exceptional ecosystems in which plants are protected from root pathogens as a result of antagonistic activities of the root-associated microbiome. Suppressive soils have been described for various soil-borne pathogens, including fungi, bacteria, oomycetes and nematodes (Kwak et al., 2018, Mendes et al., 2011, Weller et al., 2002, Chapelle et al., 2016, Berendsen et al., 2012, Mazzola et al., 2007, Mendes et al., 2013, Cha et al., 2016). Disease suppression can be eliminated by selective heat treatment and can be transplanted to non-suppressive (conducive) soils, analogous to fecal transplants in humans (Panke-Buisse et al., 2015, Voort et al., 2016). Specific suppression of soils to fungal root pathogens, such as *R. solani*, is induced in field soils by a disease outbreak during continuous cultivation of a susceptible host plant (Raaijmakers et al., 2016). Once established, the suppression can dissipate if non-host plants are grown but is regained in the presence of the host plant and the specific fungal pathogen. Therefore, the three-way interactions between the fungal pathogen, the host plant and its root microbiome are key elements of the onset and persistence of specific disease suppression. We previously showed that in a soil suppressive to the fungal root pathogen *R. solani,* several bacterial genera inhabiting the rhizosphere of sugar beet, in particular *Paraburkholderia*, *Pseudomonas* and *Streptomyces* (Mendes et al., 2011, Carrion et al., 2018, Cordovez et al., 2015), act as a first line of defense. To understand what role microorganisms that live within plant root tissues (endophytes) play in disease suppression, we conducted a metagenomic analysis of the endosphere of sugar beet seedlings grown in field soil suppressive to *R. solani* and identified the microorganisms associated with disease suppression, characterized biosynthetic gene clusters (BGCs) that were upregulated during infection, reconstructed synthetic endosphere consortia and finally made site-directed mutations to test the role of specific BGCs in disease suppression.

5.3 - Materials and Methods

5.3.1 - Soil collection, plant bioassays and endosphere samplin

The *Rhizoctonia*-suppressive soil described previously by Gómez Expósito et al., (2015) was collected from a sugar beet field in Zwaagdijk, The Netherlands. Disease suppressiveness in this field soil is typically induced by repeated cultivation of sugar beet seedlings in presence of the pathogen and an outbreak of damping-off disease. We have used composite soil samples obtained from different places in a single agricultural field site at one moment in time. This soil sampling strategy is common in plant microbiome studies as it minimizes heterogeneity in abiotic soil conditions and plant growth. Furthermore, disease suppressive soils are rare to find in the field and the disease-suppressive state of a particular field site is highly sensitive to soil management practices over time (e.g. fertilization or rotation with nonhost crops can eliminate suppressiveness). Therefore, we were restricted in the sampling date and available soil from that particular field site. For the bioassays, sugar beet (*Beta vulgaris* (cv. Rhino)) was grown in this field soil and inoculated (or not) with *R. solani* AG2-2IIIB, as described previously (*Mendes et al., 2011*). Briefly, sugar beet seedlings were grown during 42 days in square PVC pots containing 500 g of field soil, with an initial moisture content of 20% (v/w), in conducive (C) and suppressive soils (S) inoculated without (C, S) or with *R. solani* (C+R and S+R) (Figure S1). Plants were grown in a growth chamber (24 °C/24 °C day/night temperatures; 180 µmol light m^{-2} s^{-1} at plant level during 16 h/d; 70% relative humidity) and watered weekly with standard Hoagland solution (macronutrients only). The experiment was performed using four biological replicates (12 sugar beet plants per replicate) for all treatments. C, S and S+R respectively. For the C+R and S+R treatments, *R. solani* was inoculated seven days after seed germination by placing 5-mm-agar plugs (previously grown on $1/5^{th}$ strength PDA medium for 7 days at 25°C) next to the roots of the first seedling at 1 cm beneath the soil surface.

Root endosphere was harvested 34 days post-inoculation. Disease incidence was high (> 80%) for plants in the C+R treatment; hence it was not possible to collect enough root material for extraction of the endosphere microbiome for this treatment.

For the other three treatments (C, S, S+R), plants without symptoms and soil were carefully removed from the pots. Plant roots were gently shaken to remove loosely adhering soil and the shoots were discarded. Harvested roots were surface sterilized as follows: 1) 5X wash in 10 ml of 10 mM $MgSO_4$ (hereafter: buffer), 2) 1X wash in buffer containing 0.01% (vol/vol) Tween 20, 3) wash in buffer 2X, 4) 20 min immersion in 1% bleach (NaOCl) solution containing 0.01% (vol/vol) Tween 20 under slow agitation, and 5) 5X wash in buffer. To check if roots were properly surface-sterilized, roots were rolled on Luria-Bertani (LB) and $1/10^{th}$ strength TSA agar plates, and also 100 µl of the last washing step were plated on LB and $1/10^{th}$ strength TSA plates. Plates were incubated at 25°C for 7 days. Roots that showed no microbial growth were used for the isolation of endophytic bacterial community. Hence, three biological replicates remained for S and S+R whereas for the conducive soil (C) all 4 initial replicates remained. To extract the endophytic bacterial cells, root tissues were disrupted in buffer using a blender and the homogenate was filtered through 25 µm Miracloth (Merck Millipore) to remove larger plant tissue debris. The flow-through was further cleaned by centrifugation at 500 X *g* for 10 min, and the bacterial cells were collected by centrifuging the resulting supernatant at 9,500 rpm for 15 min. Cells of the endophytic bacteria were separated from the plant root cells by a Nycodenz gradient (*32-36*). Briefly, the pellet consisting of endophytic microbes and plant cells, was suspended in 3.5 ml buffer supplemented with Nycodenz® resin (PROGEN Biotechnik, Germany) to a final concentration of 50% (w/v). A Nycodenz density gradient was mounted above the sample by slowly depositing various layers of Nycodenz (3 ml of 35% Nycodenz, 2 ml of 20% Nycodenz, 2 ml of 10% Nycodenz) and the gradient was centrifuged for 45 min at 8500 rpm in a swing-out bucket rotor (Sorvall HB-6). The endophytic bacterial cells appear as an opalescent whitish band and were recovered by pipetting. The recovered cells were washed five times with buffer and centrifuged at 13,000 X *g* for 5 min to remove the Nycodenz resin. Finally, bacterial cells were suspended in 500 µl of buffer, recovered by centrifugation, (16,000 X *g*). The sample was split in two aliquots, 400 µl were used for metagenomics and 100 µl for plating on agar media to determine cell density of endophytic bacteria per gram of root tissue. For the metagenomic analysis, the samples were frozen in liquid nitrogen and stored at -80°C (Figure S1) prior to sequencing.

5.3.2 - DNA isolation and metagenome sequencing

DNA was obtained from the endophytic bacterial cell fraction using the Meta-G-Nome™ DNA Isolation Kit (Epicentre®) following the manufacturer instructions. Metagenome sequencing library preparation was performed by Erasmus MC Center for Biomics (Rotterdam, The Netherlands). Briefly, DNA was fractionated by ultrasound. Sequencing libraries were then prepared using the Nextera kit, according to Illumina's instructions, on DNA fragments in the range of 350-500 bp. Barcoded and equal molar amounts pooled libraries were paired end 300 basepairs sequenced on two rounds, an Illumina MiSeq (to check the quality of the DNA and the needed sequencing depth) followed by HiSeq2500. Finally, both datasets were merged reaching a total of 191.5 Gbp of raw metagenome sequence data. Quality filtering and removal of reads mapping to the host genome resulted in 47.2 Gbp non-plant high quality sequencing data (Figure S2 and S3A, table S1).

5.3.3 - Metagenome quality filtering and annotation

Paired end reads were trimmed with the sliding window approach used by Sickle (Ikeda et al., 2009) to keep reads with Phred score of at least 30 and 150 base pairs in length. Contamination of reads originating from the host plant were removed by mapping with Bowtie 2.2.5 (Langmead et al., 2012) in sensitive mode against the draft genome of Beta vulgaris; paired and unpaired data were stored separately. Reads of all samples were pooled together for an assembly with SPAdes 3.5 (Bankevich et al., 2012) using kmers with length 33, 55, 77, 99 and 127 and the 'careful' flag enabled. For the resulting contigs, genes were predicted with Prodigal 2.61 (Hyatt et al., 2010) in metagenomics mode and stored in General Transfer Format using Cufflinks 2.1.1 (Trapnell et al., 2012). Genes were assigned to taxonomy by running Diamond 0.7.9 (Buchfink et al., 2015) against the non-redundant Blast NCBI database from 20150311. The lowest common ancestor classification was determined using MEGAN 5.10 (Huson et al., 2007) by taking the top 50 percent hits and filtering for a minimum score of 50 and maximum expect value of 0.01 and converting the gene identifiers to taxonomy IDs using the mapping provided by MEGAN of March 4th, 2015. For functional annotation, UProc (Meinicke

et al., 2015) was used to annotate genes with KEGG release 20140317 (Kanehisa et al., 2000), COG release 2014 (Galperin et al., 2015) and Pfam 28 (Finn et al., 2016). An abundance table was created by first mapping all reads to the contigs with BamM (Imelfort et al., 2015) which uses Samtools 1.2 (Li et al., 2009) and bwa-mem 0.7.12 (Li et al., 2013) followed by counting all number of reads mapping to a contig with featureCounts (Liao et al., 2014).

5.3.4 - Taxonomic analysis of 16S rRNA reads extracted from metagenome

Ribosomal RNA reads were extracted from the quality and contamination filtered reads using a kmer strategy in BBDuk from the BBMap tool suite (Bushnell et al., 2014). These rRNA reads were aligned to SILVA 119 (Quast et al., 2013) reference set provided by Qiime (Caporaso et al., 2010) using the usearch_global algorithm implemented in VSEARCH (Rognes et al., 2016). The alignment in usearch format (uc) was converted to a BIOM file with the biom 2.1.5 package (McDonald et al., 2012) and the from-uc flag. All steps where implemented in a Snakemake workflow (Koster et al., 2012). The OTU table was filtered using QIIME (1.9.1) custom scripts (Kuczynski et al., 2012). The Bacteria domain was extracted using the command split_otu_table_by_taxonomy.py and singletons, doubletons and chloroplast sequences were discarded with the command filter_otus_from_otu_table.py, obtaining a filtered OTU table for further analysis. The alpha diversity was calculated using alpha_rarefaction.py script from QIIME. The command alpha_rarefaction.py was used to rarefy the OTU table the lowest sequencing depth obtained from a sample and therefore used as a threshold for rarefaction and alpha diversity calculations (Oksanen et al., 2007) The alpha_diversity.py command was applied to rarefied data and observed OTUs, Shannon, Chao1 and Faith's Phylogenetic Diversity metrics were obtained. One-way ANOVA and Tukey HSD were performed in R. For the Beta-diversity calculations, the entire filtered OTU table was used and normalized using the function cumNorm from the R package metagenomeSeq (v.1.12) (Paulson et al., 2013). We used a cumulative-sum scaling (CSS) method, which calculates the scaling normalization factors equal to the sum of counts up to a particular quantile to normalize the read counts, in order to avoid biases generated

with current sequencing technologies due to uneven sequencing depth (*Paulson et al., 2013*). A Bray–Curtis dissimilarity matrix was calculated and used to build Principal Coordinate Analyses and Constrained Principal Coordinate Analysis constrained by status group, that is, conducive (C), suppressive (S) and suppressive inoculated with *R. solani* (S+R), using the function capscale retrieved from Vegan package (Oksanen et al., 2007) (v.2.3-2) and implemented in the Phyloseq package (McMurdie et al., 2013) (v.1.10), both in R. The nonparametric adonis test was used to assess the percentage of variation explained by the status grouping along with its statistical significance. Permutational multivariate analyses of variance were performed to evaluate the significance of the constrained principal coordinate analyses, both retrieved from Phyloseq and Vegan packages. To compare the differences in taxonomic composition and to assess whether some bacterial taxa were differentially abundant, we conducted a three-step analysis in which we assessed separately the read counts based on Phylum, Family and OTU level. For Phylum and Family level, custom R commands were used in order to aggregate all the reads according to the level chosen. For the OTU level analysis, the function calculateEffectiveSamples from the metagenomeSeq R package was applied to the filtered OTU table and features with less than the average number of effective samples in all features were removed. For the analysis at Phylum, Family and OTU level, we used normalized tables applying the cumulative-sum scaling normalization as described above. Then, a Zero-Inflated Gaussian Distribution Mixture Model was applied using the fitZig function from metagenomeSeq. With the coefficients from the model, we applied moderated t-tests between accessions using the makeContrasts and eBayes commands retrieved from the R package Limma (v.3.22.7) (Ritchie et al., 2015). Obtained P-values were adjusted using the Benjamini–Hochberg correction method. Differences in the abundance of taxa between treatments were considered significant when adjusted P-values were lower than 0.1. Custom R scripts were used to represent the results obtained at Phylum, Class, Order, Family and Genus level. Treemap (v.3.7.3) was used to visualize the significantly abundant OTU's, the annotated taxonomy, the adjusted P-value and relative abundance, in which the size of the bubbles indicates the relative abundance of the raw read counts.

5.3.5 - Taxonomic and functional analysis of the metagenome

To compare the differences in taxonomic and functional composition, we conducted an analysis where taxonomic ranks (at phylum, class, order, family and genus level) were combined with their functions based on its annotations using COG, dbCAN, PFAM and KEGG databases. Custom R commands were used in order to aggregate all the reads according to the different levels chosen and the resulting datasets were used for the statistical analysis. In pairwise comparisons, S vs C and S+R vs S, each table was first normalized using the function cumNorm from the metagenomeSeq R package (Paulson et al., 2013). Applying the fitZig command from metagenomeSeq, as no zeros were within the count matrix, we fit a general linear model on log-transformed counts. With the coefficients from the model we performed moderated t tests between treatments using the makeContrasts and eBayes commands retrieved from the R package limma (v.3.22.7). Obtained *P*-values were adjusted using the Benjamini-Hochberg correction method. To represent the results obtained with the complete "filtered_table", a separate file including the significantly abundant taxa with their functions, the adjusted p-value and relative abundance, were used to construct bubble graphs using Treemap (v.3.7.3) in which the size of the bubbles indicates the relative abundance of the read counts.

5.3.6 - Co-occurrence networks

Network analyses were performed to assess the dynamics of the interactions in the endophytic bacterial communities of plants grown in conducive and suppressive soils. Non-random co-occurrence analyses were performed using SparCC, a tool capable of estimating correlation values from compositional data (Friedman et al., 2012). For this, we calculated SparCC correlations between microbial taxa at OTU level based on the 16S rRNA extracted from the metagenomics reads. For each network analysis, *P*-values were obtained by 99 permutations of random selections of the data table, subjected to the same analytical pipeline. SparCC correlations with a magnitude > 0.9 or < -0.9 and statistical significance ($P < 0.01$) were included into the network analyses. The nodes in the reconstructed networks represent the OTUs, whereas the edges (that is, connections) correspond to the correlation between

nodes. The topology of the network was calculated based on a set of measures, including number of nodes and edges, modularity, number of communities, average path length, network diameter, averaged degree and clustering coefficient (Newman 2003, Newman 2006). Co-occurrence analyses were carried out using the Python module 'SparCC' and network visualizations were constructed using Cytoscape (v. 3.4.0) (Shannon et al., 2003) and Gephi (Bastian et al., 2009).

5.3.7 - Metagenome annotation with dbCAN and AntiSMASH

All predicted genes were annotated using dbCAN (Yin et al., 2012). We used hmmsearch with all 308 HMMs against all the predicted genes in our metagenome. Using an E-value < $1e^{-5}$ as the cutoff, we annotated around 1,822 genes in the endophytic metagenome. The network of protein domains was generated by using HMMER v3.0 package to calculate domain-domain distances and display their relatedness. Contigs > 5 kb were processed with antiSMASH 3.0 with default parameters (Weber et al., 2015). The network of gene clusters based on shared gene content was built using BiG-SCAPE (Navarro-Munoz et al., 2013) (version of July 2016) and visualized at cutoffs 0.75, 0.80 and 0.85. The width and color intensity of the network edges was scaled with the length of the shared protein alignments, normalized to the length in base pairs of the two clusters being compared. Biosynthetic gene clusters (BGCs) were compared to experimentally characterized gene clusters previously reported in the MIBiG repository (Medema et al., 2015) using BLASTP and MultiGeneBlast (Medema et al., 2013), and the closest hits to MIBiG.

5.3.8 - Binning methods

Three separate binning tools were used in order to group contigs into genomes from the metagenome assembly (MAGs): Concoct (Alneberg et al., 2014), Maxbin 2.0 (Wu et al., 2014) and Metabat (Kang et al., 2015). These tools use tetranucleotide frequencies and differential mapping of reads across the samples onto the contigs. However, the different algorithms used to cluster the contigs into bins influence which contigs are used to assemble the final genome. The resulting genome bins were then tested to identify those that represent the same microbial entity.

5.3.9 - Draft genome selection

We defined potential genomes as any genome with an N50 above 15Kb and over 2Mb in size. For these genomes, we calculated the corresponding CC score (completeness minus contamination). Genomes with CC scores above 70 were selected and considered as sufficiently high quality (Figure S13B). Different binning tools cannot consistently assign short contigs (<5 Kb) to the bins. Also, short contigs provide only a minor contribution to the overall bin size and their coverage often do not match with the rest of the bin as shown in figure S13C and D. Therefore, these short contigs were excluded from the analysis. In total, 25 medium quality bacterial draft genomes were obtained by these binning tools.

5.3.10 – Taxonomy classification

Selected bins were classified using the pipeline described in Figure S2. Reads were mapped to the contigs of each bin with BamM extract default parameters. CheckM (Wood et al., 2014) was used to evaluate bin quality based on the identification of a specific set of single copy genes to produce completeness and contamination scores. Kraken (Li et al., 2015) is a fast and accurate tool for taxonomic classification that matches reads to a database containing unique k-mers of a microorganism. In this study, we could not reliably classify bins with Kraken, as its performance depends on the microorganism to be already present in the database, which is often not the case when accessing novel environmental niches with metagenomics. Here,

we attempted 16S rRNA taxonomic annotation of the bins with a custom pipeline. First, reads that align to a Silva (Quast et al., 2013) 16S rRNA database were extracted from the mapped reads with BBduk. Then, the subset was assembled with MegaHit (Li et al., 2014). Identification of the rRNA sequences in the MegaHit assembly was done with Barrnap, which uses HMM models derived from Rfam (Griffiths-Jones et al., 2003), Silva and RefSeq (Edgar 2004). Finally, classification of the 16S rRNA sequences was performed with Rdp-classifier. Compared to canonical tools as SortMeRNA, this pipeline could identify more 16S rRNA sequences. However, few bins contained complete rRNA sequences, which severely hinders the classification power of this method. Yet, Megan could be used to assign each mapped read to a specific taxonomic group for each taxonomic level together with a confidence score.

5.3.11 – Selection of draft genomes

We define redundant bins as genome bins created by different tools which share more than half of their sequence. Unique bins are represented by high quality bins which do not have a significant overlap with bins created by other tools. The similarity score between two bins is calculated as the ratio between the cumulative sum of shared contigs and the total bin size. When two bins have different sizes, the smaller bin is taken as reference. Contigs below 5Kb were excluded as their placement was not consistent even between highly similar redundant bins. These analyses resulted in a list that includes redundant and unique bins identified by all the tools. For each bin, we calculated a confidence score S as the difference between completeness and contamination from the CheckM output. Bins which have an S score above 70 are considered draft genomes. However, for redundant bins, the bin with the highest S score of the group was selected.

5.3.12 - Isolation and identification of culturable endophytic bacteria

To characterize the bacterial community from the endosphere of sugar beet plants grown in the three soil conditions (C, S, S+R), serial dilutions of the endophytic bacterial cell fraction obtained from the plant roots by Nycodenz separation were plated on 1/10th strength TSA (1/10 TSA) and R2A agar media. Plates were incubated at 25°C for 5 days. From each of the 10 biological replicate samples used for metagenome sequencing (C (4), S (3) and S+R (3)), approximately ninety independent colonies were randomly picked and streaked on 1/10th strength TSA plates. Colonies were re-streaked on fresh 1/10 TSA plates to ensure purity and then suspended in 96-well plates containing 40% (v/v) glycerol and stored at -80°C. For identification, the isolate collection was replicated and 16S rRNA sequencing using the primers E9F 5' GAGTTTGATCCTGGCTC 3' and E939R 5' CTTGTGCGGGCCCCCGTCAATTC 3' (V1-V4 region) was performed at BaseClear (Leiden, Netherlands). All 16S rRNA sequences were processed with custom Python scripts for trimming and quality filtering of the reads. After that, sequences were submitted to the SILVA database for taxonomic identification. The evolutionary relationship of the 16S rRNA sequences was inferred by alignment with MUSCLE (Edgar 2004) and phylogenetic analysis was performed using RAxML version 8.2.12 (Stamatakis, 2014), using -m PROTGAMMAAUTO (for amino acid sequences) and GTR (for nucleotides). Phylogenetic trees were visualized with iTOL (Letunic et al., 2011). For real-time qPCR analysis specific primers, to amplify and to correct the C_T values of the tested BGCs and chitinases genes, for the housekeeping gene serine hydroxy-methyltransferase (*glyA)* were designed for the *Chitinophaga* and *Flavobacterium* isolates (table S13, see Supplementary Material and Methods section 5.3.13). PCR reactions were performed using GoTaq® G2 Hot Start Green Master Mix. To check the generated amplicons, 5 µL of the PCR products were loaded on 1% agarose gel (1 g agarose in 100 mL TE buffer + 2.5 µL Ethidium Bromide) and run for 30 min at 100 V. The 20 µL of PCR product left of each sample were then purified with NucleoSpin® PCR Clean-up Kit. Following the manufacturer protocol, 500 µL of binding buffer were used instead of 200 µL. Finally, 15 µL of each

sample was added to two tubes, one along with the forward primer and the other with the reverse. This last step was necessary for the sequencing process carried out by Baseclear B.V. (Leiden, The Netherlands). All 16S rRNA and housekeeping genes sequences were processed by using custom python scripts for trimming and quality filtering of the reads. Taxonomic and phylogenetic analyses were performed as described above. To determine if the BGCs detected in the metagenome analysis are present in the bacterial isolates from the endosphere, a custom python script was used to automatically design primer pairs (table S8) to amplify a fragment sized between 700 and 1500 bp for each BGC. All the primers were first tested *in silico* using the online software tools Primer3Plus (Untergasser et al., 2007) and Sigma-Aldrich oligo evaluator (Sigma-Aldrich, St. Louis, Missouri, USA). Finally, each isolate of the selected subset was screened for the BGCs by PCR.

5.3.13 - Genome sequencing and comparative genomics of Bacteroidetes isolates

Following initial taxonomic and functional characterization by PCR (see point 9.), seven Bacteroidetes isolates (three *Chitinophaga* and four *Flavobacterium*) were selected for Pacbio genome sequencing. Long read assembly and comparison with draft genome bins was performed with Pacbio Sequel RSII at Keygene N.V. – The Netherlands. Raw sequences were demultiplexed with Lima 1.9.0 to produce CCS reads for all the different isolates. Reads were converted to fastq with bam2fastq 1.3.0. The assemblies were performed with Flye 2.4.2 (Kolmogorov et al., 2019). The assemblies were compared with the genomic bins from the metagenome assembly using mash 2.2.1 (Ondov et al., 2016) with kmers of length 16 and 10000 sketches. The Jaccard index was translated into a percentage using OrthoANI (Lee et al., 2016). A maximum-composite likelihood phylogeny was constructed using PhyloPhlAn and the method of Segata et al., 2013. Ortholog identification and alignment was performed in Phylophlan using the “-u” command. Functional diversity between the reference strains, bins and sequenced Bacteroidetes isolates were analyzed using the Bai et al., 2015 method with slight modifications: Briefly, for each genome in the data set, a profile of presence/absence of all COG groups between selected isolates was generated. Subsequently, a distance measure based on the

Pearson correlation of each pair of phyletic patterns was calculated, which allowed us to embed each genome as a data point in a metric space.

5.3.14 - Disease suppression by Bacteroidetes isolates and consortia

Spontaneous rifampicin-resistant mutants of the seven sequenced *Bacteroidetes* isolates were generated to allow monitoring of their population dynamics. The rifampicin-resistant Bacteroidetes isolates were grown in 10 ml of 1/10th strength TSB supplemented with 50 µg/ml rifampicin for 2 days at 25°C on a rotary shaker at 200 rpm. Liquid cultures of each isolates were centrifuged, washed 3 times and resuspended in sterile 10 mM $MgSO_4$. Cell suspensions were mixed with conducive field soil at an initial density of 10^7 cells/g soil and approximately 20% (v/w) soil hydration. Rectangular trays (19.5 × 6 × 3.5 cm) were filled with 250 g of the conducive soil and 16 sugar beet seeds were sown in a row, 1 cm apart. Non-inoculated soil was used as a control. For each isolate treatment, eight replicates were used. Trays were placed in boxes with transparent lids in a growth chamber at 24°C with a 16h photoperiod. After 5 days, seedlings were inoculated with a mycelial plug (5-mm diameter) of *R. solani* AG2-2 IIIB previously grown for 1 week on 1/5th strength PDA at 25°C. The mycelial plug was placed next to the root of the first seedling of each tray. For the inoculation of the syncom, a mixture of the isolates, each at a cell density of 10^7 cells/ml and mixed in 1:1 ratio, was used to inoculate seedlings. Instead of mixing through the soil, approximately 1 ml of the syncom mixture was inoculated close to the stem of each seedling. For three independent bioassays, disease incidence was assessed between 21-28 days after pathogen inoculation when disease incidence in the control treatment reached a level of approximately 75%. For one of these three bioassays, we monitored disease incidence at regular intervals for 2 weeks.

5.3.15 - Rhizosphere and endosphere colonization by Bacteroidetes isolates

To determine whether the selected and introduced bacterial isolates were able to colonize the endophytic compartment of the sugar beet seedlings, root colonization was assessed. For that, the rifampicin-resistant mutants were inoculated in soil. After 28 days, sugar beet plants were harvested and processed as follows: roots were suspended in 10 ml of 10 mM $MgSO_4$ buffer in triplicates, vortexed for 1 min, sonicated for 1 min and vortexed for 15 s to collect the rhizosphere. For rhizosphere, serial dilutions were plated on $1/10^{th}$ strength TSA medium supplemented with 100 µg ml^{-1} rifampicin plus other antibiotics (100 µg ml^{-1} kanamycin and 100 µg ml^{-1} trimethoprim for *Flavobacterium* spp. and 100 µg ml^{-1} kanamycin for *Chitinophaga* spp.) and incubated at 25°C for 4 days. To assess endosphere colonization, roots were surface-sterilized as described above and ground in 10 ml of 10 mM $MgSO_4$ buffer using a blender. Serial dilutions were plated on selective agar medium and incubated for 4 days at 25°C. Colony forming units (CFU) were counted for the rhizosphere and endosphere samples. The undiluted solutions with rhizosphere and endosphere samples were stored at - 80°C for qPCR.

5.3.16 - In vivo BGC expression in Bacteroidetes isolates

For the transcriptional analysis of biosynthetic gene clusters (BGCs) and chitinase genes, RNA was extracted from samples collected in the *in vivo* antagonistic assay. Samples were collected 3 weeks post-infection and for all the treatments between 3-5 biological replicates were obtained (depending on the number of remaining healthy plants). Rhizosphere and endosphere samples were obtained and processed in RNALater. Rhizosphere samples were processed according to manufacturer's instructions. The endosphere samples were washed 10 times with buffer and to extract the endophytic bacterial cells, root tissues were disrupted using a blender in a known volume of buffer. PowerSoil® RNA isolation kit (MO BIO Laboratories, Inc.) was used for RNA isolation following the manufacturer's instruction. All RNA isolated was purified following a DNA/RNA clean up kit (PowerClean® Pro DNA Clean-Up Kit) protocol and quantified by using Qubit™ 3.0

Fluorometer. Samples were standardized to the same concentration and cDNA synthesis was carried out by using RevertAidTM First Strand cDNA Synthesis Kit (Thermo Scientific™).

5.3.17- Real-Time qPCR

Expression of specific biosynthetic gene clusters (BGCs) as well as chitinase genes in the rhizosphere and endosphere of plants in the presence and absence of *R. solani* was determined by qPCR. The concentration of the primers was optimized at 400 nM, and a dissociation curve was performed to check the specificity of the primers. The primers used for the qPCR are listed in (table S14). All qPCR reactions were carried out by using iTaqTM Universal SYBR® Green Supermix and with a CFX96 Touch™ Real-Time PCR Detection System thermocycler following the program, initialization 4 min at 95°C, 40 cycles of denaturation at 95°C for 30 s, annealing at 58 (for the BGCs) or 61°C (for the chitinase genes) for 15 s and extension at 72°C for 15 s, final extension was 72°C for 5 min. To correct for small differences in template concentration, *glyA* was used as the reference housekeeping gene. The cycle in which the SYBR green fluorescence crossed a manually set cycle threshold (C_T) was used to determine transcript levels. For each gene, the threshold was fixed based on the exponential segment of the PCR curve. The C_T values of the BGCs were corrected for the housekeeping gene *glyA* as follows: $\Delta C_T = C_T$ (BGCn or chitinase) - C_T (*glyA*); the same formula was used for all the BGCs studied. The relative quantification (RQ) values were calculated by the following formula: $RQ = 2^{-}[\Delta C_T(\text{BGCn or chitinase}) - \Delta C_T(\textit{glyA})]$. qPCR analysis was performed in triplicate (technical replicates) and between three and five (depending on the amount of material available) independent RNA isolations per treatment were used (biological replicates). Statistical analyses were performed using the R packages pgirmess, car, agricolae and multcomp. Data were tested for equal variance using Levene's test and for normality using Shapiro–Wilk test at the 5% significance level. Statistical differences were determined by pairwise comparisons as compared with the controls with Student t-test (for independent samples) or One-way ANOVA followed by Tukey's honestly significant difference (HSD) post hoc test. A Generalized Linear Model (GLM, Type III Chi-square Wald test) was performed to

statistically assess differences between treatments combined with an FDR test when assumptions of normality were not met.

5.3.18 - Strains and growth media for CRISPR-Cas-based mutagenesis of BGC298 in *Flavobacterium*

Escherichia coli DH10B strain was used as a host for cloning and propagating the plasmids required for HR-Cas9 based gene editing in *Flavobacterium* sp. 98. The *E. coli* and *Flavobacterium* were routinely cultured at 37 °C and 30 °C, respectively. *E. coli* cells were grown in 10 ml LB [10 g l^{-1} peptone (Oxoid), 5 g l^{-1} yeast extract (BD), 10 g l^{-1} NaCl (Acros)] medium in 50-mL Greiner tubes or plated on LB with 15 g l^{-1} agar (Oxoid) plates. The LB culture medium was supplemented with 100 µg ml^{-1} ampicillin to select for *E. coli* harboring plasmids derived from the pCP11 *E. coli-Flavobacterium* shuttle vector. *Flavobacterium* sp. 98 was cultured in 10 mL CYE (10 g l^{-1} casitone (BD), 5 g l^{-1} yeast extract (BD), 8 mM $MgSO_4$-7 H_2O (Sigma-Aldrich) in 1 l of 10 mM Tris-buffer (Sigma-Aldrich, pH adjusted to 7.6) medium (McBride et al., 1996) in 50-ml Greiner tubes at 175 rpm or plated on CYE with 20 g l^{-1} agar (Oxoid) plates. To select for transformants, the culture medium was supplemented with 100 µg ml^{-1} erythromycin.

5.3.19 - Transformations, PCR, DNA isolation and sequencing

Flavobacterium sp. 98 cells were made electro-competent by combining previously established protocols for other *Flavobacterium* species (Chen et al., 2007, McBride et al., 1996). *Flavobacterium* sp. 98 was grown overnight in 10 ml of CYE at 30 ^oC, 175 rpm. The overnight culture was used to inoculate 100 ml CYE broth and grown until it reached an OD_{600} of 0.3. Thereafter, the cells were harvested by centrifugation at 4700 rpm for 10 minutes at 4 ^oC and washed three times with 1×volume of 10% (v/v) glycerol at 4 ^oC. The pellet was suspended using 10% glycerol to 1/100 of the initial volume. Cells were aliquoted in micro-centrifuge tubes and stored at -80 ºC. Plasmids for *Flavobacterium* sp. 98 transformations were extracted from *E. coli* via miniprep isolation (Thermo Scientific). 80-100 ng of plasmid DNA was added to 100 µL of cells. Electroporation was performed using BTX™ ECM™ 630 pulser in 1-mm

cuvette using the following settings: 1.5 kV, 200 Ω, 25μF. 900 μl of CYE medium was added immediately and the cells were incubated at 30 °C for 1.5 hours for recovery. The cells were plated on CYE agar supplemented with 100 μg ml^{-1} erythromycin and incubated at 30 °C for two days. Single colonies were randomly screened for gene deletion by colony PCR using OneTaq® 2X Master Mix with Standard Buffer (NEB) with specific primers BG15927 and BG15928. Genomic DNA of *Flavobacterium* sp. 98 colonies was isolated using the PureLink™ Genomic DNA Mini Kit (Thermo Scientific). Purification of the PCR products was performed using the Zymoclean™ Gel DNA Recovery Kit (Zymo Research). The DNA fragments were subsequently sent for Sanger sequencing (Macrogen Europe B.V) to verify the gene deletion.

5.3.20 - Plasmid construction

The oligonucleotides/primers (IDT) used for cloning and the constructed plasmids are listed in Supplementary Tables S11 and S12, respectively. The fragments for assembling the plasmids were amplified by PCR with Q5® High-Fidelity 2X Master Mix (NEB). The amplicons were resolved on an 1% agarose gel electrophoresis and purified using Zymoclean™ Gel DNA Recovery Kit (Zymo Research). Fragments were assembled using NEBuilder® HiFi DNA Assembly Master Mix (NEB). The HiFi DNA assembly mix was purified using the DNA Clean & Concentrator-5 (Zymo Research) prior to electroporation into competent *E. coli* DH10B cells. Single colonies were randomly screened for the presence of the correct plasmids by colony PCR using OneTaq® 2X Master Mix with Standard Buffer (NEB). The confirmed colonies were used to inoculate 10 ml LB medium, followed by plasmid isolation using the GeneJET Plasmid Miniprep kit (Thermo Fisher Scientific) and subsequent verification by Sanger sequencing (Macrogen Europe B.V).

All the targeting plasmids:

pSpyCas9Fb_Sp1, pSpyCas9Fb_Sp2, pSpyCas9Fb_Sp3 and the pSpyCas9Fb_NT, were created simultaneously. For the construction of the pSpyCas9Fb_Sp1 plasmid, a 5-part assembly was designed and assembled. The

parts consist of: 1) a fragment of the pCP11 backbone plasmid amplified (using BG14285 and BG14286) from the pCP11 vector; 2) *E. coli* codon optimized *spycas9* gene amplified (using BG14303 and BG14304) from pET-28b-Cas9-His (Gagnon et al., 2014) along with an *ompA* promoter (P_{ompA}) added as an overhang (Chen et al., 2007). The PCR products were resolved using agarose gel electrophoresis and isolated using Zymoclean™ Gel DNA Recovery Kit (Zymo Research) before introducing the overhangs by another PCR using BG14287 and BG14288 primers; 3) a fragment of the membrane associated zinc metalloprotease (map) terminator (Xie et al., 2004) amplified (using BG14605 and BG14290) from the synthesized G-Block (IDT) BG14419; 4) a fragment of the pCP11 backbone amplified (using BG14291 and BG14292) from pCP11 vector; 5) sgRNA containing spacer 1 amplified (using BG14305 and BG14306) from pT7_gRNA (Jao et al., 2013) and placed in between the HU promoter (P_{HU}) (Chen et al., 2007) and the *ompA* terminator (Agarwal et al., 1997) (using BG14307 and BG14308) followed by another amplification (using BG14293 and BG14294).

For the construction of the pSpyCas9Fb_Sp2, pSpyCas9Fb_Sp3 and pSpyCas9Fb_NT plasmids, a 3-part assembly was designed and assembled. The parts consist of: 1) a fragment of the pCP11 backbone plasmid and sgRNA amplified (using BG14295 and BG14296, BG14295 and BG14298, BG14295 and BG15071) from pSpyCas9Fb_Sp1 plasmid for pSpyCas9Fb_Sp2 plasmid, pSpyCas9Fb_Sp3 plasmid, and pSpyCas9Fb_NT plasmid, respectively; 2) a fragment of pCP11 backbone plasmid amplified (using BG14297 and BG14300, BG14299 and BG14300, BG15072 and BG14300) from pSpyCas9Fb_Sp1 plasmid for pSpyCas9Fb_Sp2 plasmid, pSpyCas9Fb_Sp3 plasmid, and pSpyCas9Fb_NT plasmid, respectively; 3) *E. coli* codon optimized *spycas9* gene along with the P_{ompA} promoter amplified (using BG14301 and BG14302) from pSpyCas9Fb_Sp1 plasmid.

For the construction of the pSpyCas9Fb-HR_Sp1 plasmid, a 5-part assembly was designed and assembled. The parts consist of: 1) a fragment of the pCP11 backbone plasmid amplified (using BG14295 and BG15653) from pSpyCas9Fb_Sp1 plasmid; 2) the upstream genomic region of type I PKS module of BGC298 amplified (using

BG15243 and BG15655) from the genome of *Flavobacterium* sp. 98; 3) the downstream genomic region of type I PKS module of BGC298 amplified (using BG15654 and BG15246) from the genome of *Flavobacterium* sp. 98. 4) a fragment of pCP11 backbone plasmid and sgRNA amplified (using BG15656 and BG14300) from pSpyCas9Fb_Sp1; 5) *E. coli* codon optimized *spycas9* gene along with the P_{ompA} promoter amplified (using BG14301 and BG14302) from pSpyCas9Fb_Sp1 plasmid.

Two plasmids were used as controls in the editing experiment: the plasmid containing a non-targeting spacer and HR flanks (pSpyCas9Fb-HR_NT) and the plasmid with HR flanks but without *spycas9* (pCP11_w/o-Cas9_HR). The pSpyCas9Fb_NT_HR plasmid was constructed using the same primers and design as pSpyCas9Fb-HR_Sp1 and all the respective parts were amplified from the pSpyCas9Fb_NT plasmid. For the construction of the pCP11_w/o-Cas9_HR plasmid, a 4-part assembly was designed and assembled. The parts consist of: 1) a fragment of the pCP11 backbone plasmid amplified (using BG15564 and BG15653) from pSpyCas9Fb_NT; 2) the upstream genomic region of type I PKS module of BGC298 amplified (using BG15243 and BG15655) from the genome of *Flavobacterium* sp. 98; 3) the downstream genomic region of type I PKS module of BGC298 amplified (using BG15654 and BG15246) from the genome of *Flavobacterium* sp. 98; 4) a fragment of the pCP11 backbone plasmid amplified (using BG15656 and BG15565) from pSpyCas9Fb_NT plasmid.

5.3 – Results and Discussion

5.4.1 - Taxonomic diversity and network inference of the endophytic microbiome

Sugar beet plants were grown in disease-conducive (C) and disease-suppressive (S) soils inoculated (or not) with the root pathogen *R. solani* (Figure S1). Disease incidence in the pathogen-inoculated suppressive soil (S+R) was 15-30%, whereas disease incidence in the pathogen-inoculated conducive soil (C+R) exceeded 80% (Figure S1A), typical of our previous studies (Mendes et al., 2011, van der Voort et al., 2016). Given the high disease incidence in C+R, there was not enough root material left for in-depth microbiome analysis of this condition. The taxonomic diversity and functional potential of the root endophytic microbiome of plants grown in the remaining three soil conditions (C, S, S+R) was investigated after 4 weeks of plant growth. Following metagenome sequencing and bioinformatic analyses (Figure S2, table S1 and S2), taxonomic assignment of the microbial cell fraction from the sugar beet endosphere showed that 76.1%, 10.5% and 0.0065% of the sequence reads corresponded to the domains Bacteria, Eukaryotes and Archaea, respectively (Figure S3, A and B). For the Eukaryotic reads, Constrained Analysis of Principal Coordinates (CAP) showed significant differences (PERMANOVA, $P < 0.05$) between the endophytic fungal community composition in C, S and S+R (Figure S4A). This was largely due to a significant increase in *Rhizoctonia*-related sequence reads in the suppressive soil inoculated with *R. solani* (S+R) (Figure S4, B and C). Most of the other sequence reads could not be reliably assigned to specific fungal phyla. Collectively, these results indicate that after inoculation into the disease-suppressive soil, *R. solani* colonized and penetrated the plant roots but caused little disease.

16S rRNA data from the metagenome sequences (Figure S2) showed that Proteobacteria and Bacteroidetes dominated the endophytic bacterial community with ten OTUs spanning *Pseudomonadaceae* (2), *Xanthomonadaceae* (4), *Chitinophagaceae* (1), *Flavobacteriaceae* (2) and *Veillonellaceae* (1) (Figure S5), all of which became enriched in the S+R condition compared with the S condition

(Figure 5-1A). Co-occurrence network analysis revealed increased complexity in the S+R condition (Table S3, Figure S6, A, B and C) compared with C and S conditions (Table S3). Highly connected networks, like those in the S+R samples, can occur when microbiota face environmental perturbation, such as pathogen invasion (*20*). Interestingly, 80% of the interacting nodes in the S+R network belonged to *Chitinophaga*, *Flavobacterium* and *Pseudomonas* spp. (Table S4). When sequence reads from the Bacteroidetes were removed from the datasets, the endophytic signal from the C and S soils were indistinguishable (Figure S7, A and B), once again indicating an association of the Bacteroidetes genera *Chitinophaga* and *Flavobacterium* with the disease-suppressive phenotype.

5.4.2 - Functional diversity of the endophytic microbiome

Fifty to seventy percent of the genes retrieved from the metagenome data were assigned to a known function (Figure S3, C, D and E). For the other genes, grouping annotations indicated 56,175 taxa-associated functions of which 402 functions were significantly enriched in the endophytic bacterial community of plants grown in the S soil compared with plants in the C soil (FDR < 0.1; Figure S8, B and C). In the S+R condition, this proportion of functional enrichment increased over ten-fold (4,443) (FDR < 0.1; Figure 5-1B). These genes belonged mainly to pathways classified as 'carbohydrate transport and metabolism' and 'signal transduction mechanisms'. Several endophytic bacterial families, including *Chitinophagaceae* and *Flavobacteriaceae* (Bacteroidetes), *Pseudomonadaceae* and *Xanthomonadaceae* (Gammaproteobacteria), *Hyphomicrobiaceae* and *Rhizobiaceae* (Alphaproteobacteria), and *Burkholderiaceae* (Betaproteobacteria) were specifically associated with the functional enrichment we observed (Figure 5-1B, C and Figure S9A). The majority of the overrepresented genes in S+R (3,138 genes of 4,443) were associated with *Chitinophagaceae* and *Flavobacteriaceae* (Figure 5-1B, Figure S9A). When we used a more stringent significance level of $P<0.05$, 2,063 of 56,175 taxa-associated functions were overrepresented with 461 functions associated mainly with *Chitinophagaceae* and *Flavobacteriaceae*. Cumulative differential abundance analyses of all Bacteroidetes' genes between samples highlighted that genes from COG category Q (secondary metabolites biosynthesis, transport, and

catabolism) were among the most differentially abundant between S+R and S, while genes from category G (carbohydrate transport and metabolism) were among the most differentially abundant between S and C (Figure 5-1C).

For more detailed resolution of the specific functions associated with COGs G and Q, we searched for carbohydrate-active enzymes and secondary metabolite biosynthetic gene clusters within the metagenome sequences using dbCAN (Yin et al., 2012, Lombard et al., 2014) and antiSMASH (Weber et al., 2015), respectively. Using dbCAN we were able to annotate 1,822 genes in the endophytic metagenome with glycoside hydrolase (GH), glycosyltransferase (GT), polysaccharide lyase (PL), and carbohydrate esterase (CE) domains as well as non-catalytic carbohydrate-binding modules (CBMs). Because many of these domains are evolutionary related and have related functionalities, we mapped the domain diversity in a protein family similarity network using the hhsearch algorithm (*24*). Glycoside hydrolases and glycosyltransferases were more abundant in the S+R endophytic microbiome and correlated with disease suppression (Figure 5-2A, Figure S9, B and C). Three endophyte families (*Chitinophagaceae*, *Burkholderiaceae*, *Xanthomonadaceae*) showed statistically significant differences in CAzyme composition between S+R and S (FDR < 0.1, Figure 5-2A, Figure S9, A and B). Further, we found that *Chitinophagaceae* harbored several enzymes with domains associated with fungal cell wall degradation, such as chitinases, beta-glucanases, endoglucanases (Figure 5-2A), and also possessed de-branching enzymes, including α-1,6-mannanase and α-L-rhamnosidase. *Burkholderiaceae* and *Xanthomonadaceae* families (Figure S9B, C) also contributed two chitinase domains and three other enzymes involved in chitin degradation, including chitin deacetylase and chitosanase. Only five domains were shared between *Chitinophagaceae, Burkholderiaceae* and *Xanthomonadaceae* (Figure 5-2B), indicating limited functional redundancy among these endophytes for this trait. The enrichment of genes encoding chitin-degrading enzymes points to a role in disease suppression for these endophytes (Bowman et al., 2006).

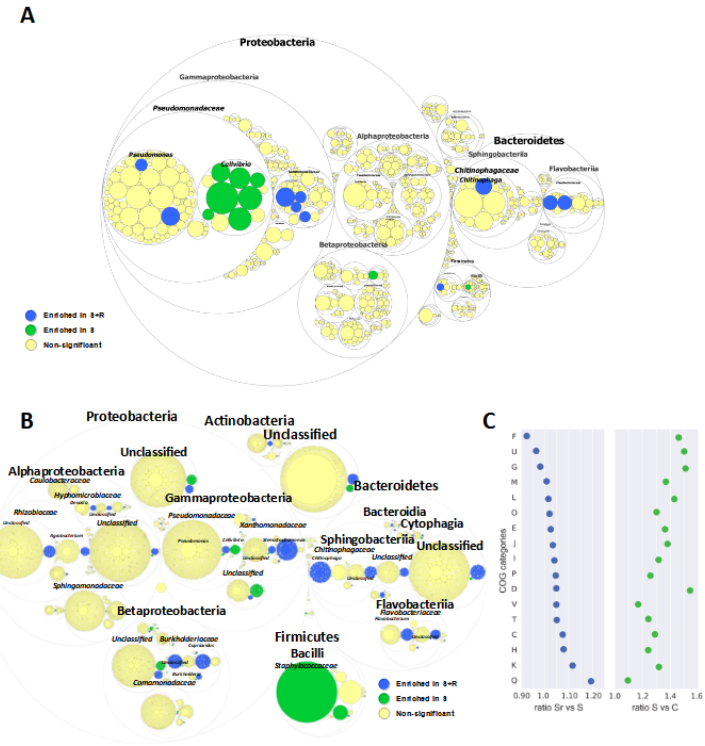


Figure 5-1. Pathogen-induced changes in endophytic microbiome diversity and functions. Differential abundance of endophytic bacterial communities from plants grown in suppressive soil (S) or in suppressive soil inoculated with the fungal root pathogen *R. solani* (S+R). (**A**) Taxonomic differences are based on 16S rRNA sequences extracted from the metagenome. The largest circles represent Phylum level, the inner circles represent Class, Family and Genus. (**B**) Functional differences are based on the metagenome sequence data and assigned to taxonomic groups. The smallest circles represent the COG categories groups. The circle sizes represent the mean read relative abundance of the differentially abundant taxa and functions. Bacterial taxa or functions that are significantly enriched (FDR<0.1) in the comparison between S and S+R are indicated in green for S and in blue for S+R; non-significant taxa and functions are indicated in yellow. (**C**) Strip plot depicting the average abundance ratios of all genes from Bacteroidetes belonging to core COG functional categories that contain significantly enriched genes in S+R compared with S, and in S compared with C. Categories are sorted by S+R/S and S/C ratios. Each COG type has been abbreviated as follows: C: energy production and conversion, D: cell cycle control, cell division, and chromosome partitioning, E: amino acid transport and metabolism, F: nucleotide transport and metabolism, G: carbohydrate transport and metabolism, H: coenzyme transport and metabolism, I: lipid transport and metabolism, J: translation, ribosomal structure, and biogenesis, K: Transcription, L: replication, recombination, and repair, M: cell wall/membrane/envelope biogenesis, O: post-translational modification, protein turnover, and chaperones, P: inorganic ion transport and metabolism, Q: secondary metabolites biosynthesis, transport, and catabolism, T: signal transduction mechanisms, U: intracellular trafficking, secretion, and vesicular transport, V: defense mechanisms, W: extracellular structures.

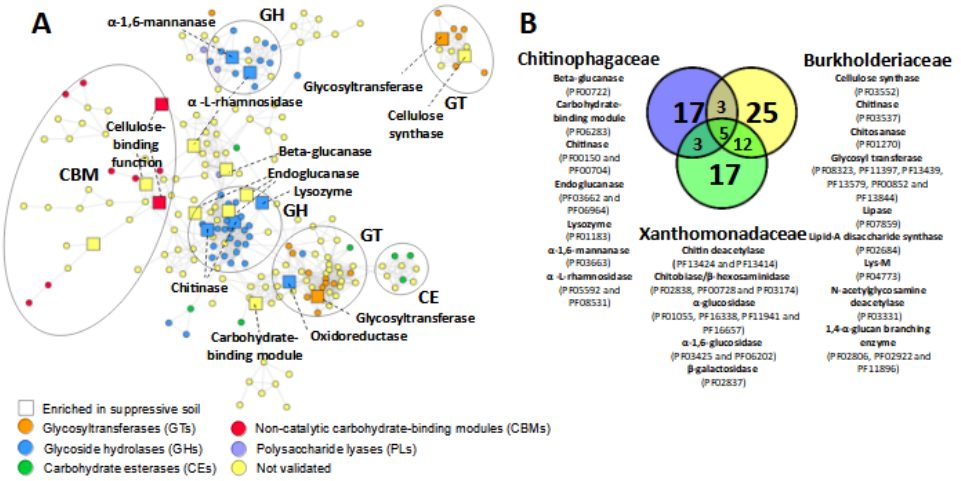


Figure 5-2. Diversity and distribution of carbohydrate-active enzymes in the endophytic microbiome (**A**) similarity network of known and putative HMM domains of enzymes involved in carbohydrate metabolism (CAzymes). From the endophytic metagenome of plants grown in suppressive soil (S) or in suppressive soil inoculated with the fungal root pathogen *R. solani* (S+R), a total of 1,822 genes were annotated as CAzymes. Domain-domain distances and their relatedness are shown in the network. Nodes were grouped into five functional classes: glycoside hydrolases (GH, blue), glycosyltransferases (GTs, orange), polysaccharide lyases (PL, purple), carbohydrate esterases (CE, green) and the non-catalytic carbohydrate-binding modules (CBM, red). Unknown domains or domains for which the function has not been experimentally validated are shown in yellow. Squared nodes represent enzymes that are significantly overrepresented (FDR < 0.1) in S+R compared with S, and taxonomically assigned to the *Chitinophagaceae*. Enzymes significantly overrepresented in S+R and taxonomically classified as *Burkholderiaceae* and *Xanthomonadaceae* are shown in Figure S9B and C, respectively. (**B**) Venn diagram with different CAzymes annotated for three endophytic bacterial families enriched in S+R, i.e. *Burkholderiaceae*, *Chitinophagaceae* and *Xanthomonadaceae*. For each of the CAzymes, the Pfam number is shown between brackets. The Venn diagram shows the number of domains detected exclusively for each bacterial family and the domains shared by these endophytic bacterial families.

Bacterial genomes contain a large diversity of biosynthetic gene clusters (BGCs), the vast majority of which have not yet been linked to specific molecules or functions (Mendes et al., 2011, Watrous et al., 2012, Medema et al., 2015, Cimermancic et al., 2014). Our antiSMASH analysis for secondary metabolites revealed a total of 730 BGCs associated with the biosynthesis of nonribosomal peptides, polyketides, terpenes, aryl polyenes, ribosomally synthesized and post-translationally modified peptides (RiPPs), phosphonates, phenazines, and siderophores (Figure 5-3A, Figures S10, 11, 12). Of these 730 BGCs, only 12 BGCs have previously been described and the chemical structure of their products elucidated (Table S5, Figure S11). Among these were the BGCs for thanamycin and brabantamide, which are two NRPS-derived products previously detected in the rhizosphere microbiome of

plants grown in *Rhizoctonia*-suppressive soil (Mendes et al., 2011, Watrous et al., 2012, Schmidt et al., 2014). For the other 718 BGCs, no near or exact matches were found for their genetic architecture and predicted products in the MIBiG repository (Medema et al., 2015). Of the BGCs detected, several proteobacterial RiPPs and NRPSs were noted (Figure 5-3C), as well as NRPS and aryl polyene clusters originating from Bacteroidetes (mainly *Flavobacterium* and *Chitinophaga* [Figure 5-3D]) and a larger proportion of NRPS clusters from a group of unclassified phyla (Figure 5-3E). Altogether, 117 BGCs were significantly overrepresented (two-tailed Welch's T-test, P <0.1) in the endosphere under the S+R conditions with 34 BGCs belonging to Bacteroidetes (Figure 5-3A-F, Figure S10, S11and S12). Notably, these did not include the thanamycin and brabantamide BGCs identified previously for the disease-suppressive *Pseudomonas* spp. from the rhizosphere (Mendes et al., 2011, Schmidt et al., 2014). For the Bacteroidetes species, 10 NRPS gene clusters out of the 117 were overrepresented under S+R conditions and none of these had a match in antiSMASH with gene clusters from MIBiG.

5.4.3 - De novo assembly of endophytic bacterial genomes

From the 730 BGCs identified in the metagenome by antiSMASH, 157 were found in a set of 25 metagenome-assembled genomes (MAGs) we reconstructed (Table S6, Figure S13 and S14). The MAGs, housekeeping genes and identified BGCs were subsequently used to generate specific primer sets for transcriptome analyses and to associate the BGCs to isolates in the bacterial endophyte collection.

The initial collection of 935 bacterial endophyte isolates (Figure S1) were taxonomically characterized by 16S rRNA sequencing (Figure S15, A and B, Table S7), revealing eight different genera, mostly represented by Bacteroidetes and Gammaproteobacteria. Although no BGCs associated with *Chitinophaga* or *Pseudomonas* spp. (Table S8) were detected, four BCGs (298, 396, 471 and 592) were found in the endophytic *Flavobacterium* isolates obtained from the S+R condition. Three of these encoded an NRPS (BCGs 396, 471, 592) and the fourth a hybrid NRPS-PKS gene cluster (BGC298, Figure 5-4A). A similar approach confirmed the presence of glycosyl hydrolase (GH18) genes in three endophytic

Chitinophaga isolates obtained from the S+R condition (Figure 5-2A). Subsequent *in vitro* assays with the bacterial isolate collection showed that the three *Chitinophaga* isolates also had extracellular chitinolytic active

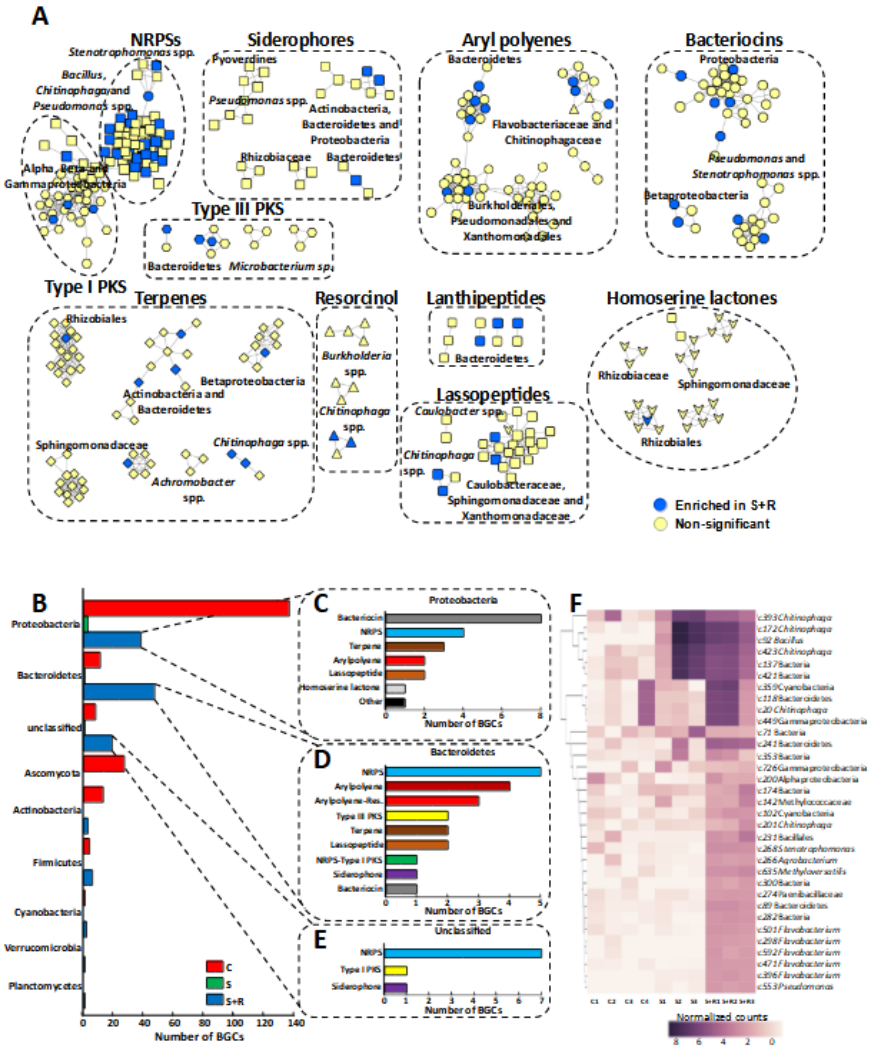


Figure 5-3. Diversity and distribution of biosynthetic gene clusters (BGCs) in the endophytic microbiome. (**A**) Sequence similarity network (constructed with BiG-SCAPE (Kautsar et al., 2019), threshold: 0.8) of the different classes of BGCs detected in the endophytic microbiome. Taxonomic assignment and BGC class annotation of the nodes are shown. Nodes with less than three connections were removed - original network with all nodes can be found in Figure S10. Node colors represent statistical significance based on a Welch test (FDR < 0.1): yellow nodes are non-significant and blue nodes are significantly overrepresented in S+R condition. (**B**) Number of overrepresented BGCs (two-tailed Welch's T-test, *P* <0.1) detected by the antiSMASH and Clusterfinder algorithms for the different bacterial phyla in the endophytic root microbiome of plants grown in conducive (C), suppressive (S) and

suppressive soils challenged with the fungal pathogen *R. solani* (S+R). (**C**, **D** and **E**) Number and type of BGCs assigned to Proteobacteria, Bacteroidetes and unclassified bacterial phyla, respectively, that were significantly (two-tailed Welch's T-test, P <0.1) more enriched in S+R. (**F**) Clustered heat map of relative abundances (CSS-normalized RPKM values) of the 33 NRPS gene clusters that were significantly overrepresented in the different replicate samples of S or S+R versus C. The NRPS cluster number and the corresponding taxonomic assignment are shown on the right side of the panel.

Subsequent genome sequencing of the three *Chitinophaga* and four *Flavobacterium* isolates showed >99% similarity among the isolates within each genus (table S9, Figure S15C and S16A). The isolate genomes also clustered with MAGs assigned to each of these genera (Figure S15B and C), confirming that they correspond to taxa abundant in the microbiome. For the key BGCs, no signs of metagenome mis-assemblies were identified based on comparisons with the complete genome sequences of the Bacteroidetes isolates (Figure S16B and C, and S17A, B and C).

5.4.4 - Reconstruction and functional analysis of disease-suppressive consortia

We selected the seven sequenced Bacteroidetes isolates for root colonization assays and BGC-transcript analysis. All isolates colonized the rhizosphere and the root endosphere of sugar beet seedlings (Figure S18 and S19). Transcriptional analysis showed that chitinase expression was significantly ($P < 0.05$) higher in the consortium colonizing the rhizosphere and endosphere compartments inoculated with the fungal pathogen (Figure 5-4B and C, Figure S20). Of the four *Flavobacterium* gene clusters, BGC298 was expressed at significantly ($P < 0.05$) higher levels in the endosphere than in the rhizosphere when the plant roots were challenged with the fungal pathogen *R. solani* (Figure 5-4C). This BGC was consistently assembled in all four *Flavobacterium* genomes and in a MAG (Figure S16B) and showed no match with known BGCs in MIBiG (Figure S17).

The central place of *Flavobacterium* and *Chitinophaga* in the functional network of plants grown in the disease suppressive soil, their ability to colonize the endosphere and the fact that expression of BGC298 and chitinase genes in the synthetic consortium are induced by the fungal pathogen indicate a role in *R. solani*-disease suppression. To test this hypothesis, three independent bioassays showed that the consortium of *Chitinophaga* and *Flavobacterium* conferred more significant and

more consistent protection against fungal root infection than the individual consortium members (Figure 5-4D, E and F, and Figure S21A, B and C). Even when single isolates showed little benefit against disease, consortia always showed a greater degree of protection (Figure 5-4D-F, Figure S21A-C). The apparent 'minimal' consortium to reconstitute the plant phenotype consisted of one *Chitinophaga* isolate and one *Flavobacterium* isolate (syncom-2) as this consortium showed the same level of disease control observed for the seven-member consortium (syncom-7; Figure 5-4F).

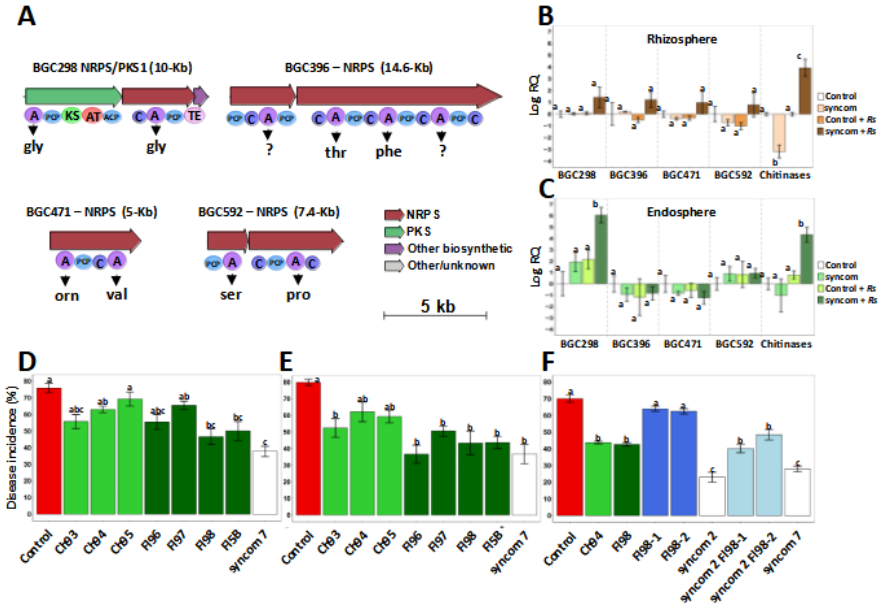


Figure 5-4. Transcriptional and functional analyses of disease-suppressive consortia. (**A**) Genetic organization of BGCs 298, 396, 471 and 592 identified in both the *Flavobacterium* MAG nbed44b64, and in the genome sequences of the four endophytic *Flavobacterium* isolates. Shown below the nonribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) genes are the module and domain organizations of the encoded proteins. The domains are labeled as: C, condensation; A, adenylation; KS, ketosynthase; AT, acyltransferase; and TE, thioesterase. Predicted substrates of the NRPS and PKS modules in BGC298 are glycine, malonyl-CoA and again glycine. (**B** and **C**) qPCR-based analysis of the expression of BGC298, BGC396, BGC471, BGC592, and chitinase genes (GH18) in the rhizosphere and endosphere of sugar beet seedlings treated with the synthetic endophytic consortium of *Chitinophaga* and *Flavobacterium* isolates (syncom). LogRQ represents the gene expression levels by relative quantification scores: values below 0 indicate lower expression of the BGC relative to that of the housekeeping gene (*glyA*) used for data normalization. Bars represent the average of 3-5 biological replicates per treatment and error bars indicate the standard error of the mean. Different letters indicate statistically significant differences between treatments as determined by one-way ANOVA with post-hoc Tukey HSD ($P < 0.05$). **(D)** and **(E)** *Rhizoctonia* damping-off disease incidence of sugar beet seedlings

treated with single *Chitinophaga* (Ch93, 94, 95) and *Flavobacterium* (Fl96, 97,98 and 5B) isolates and with a consortium of all seven endophytic isolates (synthetic community, syncom 7), and **(F)** single *Chitinophaga* (Ch94), *Flavobacterium* (Fl98) isolates, two independent Fl98-mutants (Fl98-1 and Fl98-2) with a deletion in BGC298, syncom 7 and the consortium of *Chitinophaga* sp. 94 and *Flavobacterium* sp. 98 (syncom 2). Single isolates and the two syncoms were applied at an initial density of 10^7 CFU g^{-1} of *Rhizoctonia*-conducive field soil. Bars represent the average of 4-8 biological replicates per treatment and error bars represent the standard error of the mean. Disease incidence was scored 21- 28 days after *R. solani* inoculation. Different letters indicate statistically significant differences between treatments as determined by one-way ANOVA with post-hoc Tukey HSD ($P < 0.05$). Note: for Figure 5-4B-F, box plots with the individual data of each replicate are provided in Figures S20 and S21.

To confirm the role of the *Flavobacterium* BGC298 in the disease-suppressive activity, we developed a SpyCas9-mediated system for introduction of double-stranded DNA breaks in *Flavobacterium* sp. 98. We obtained two independent BGC298 mutants (table S10, 11 and 12, and Figure S22A, B, C and D), for which the PKS gene deletion was verified by Sanger sequencing with specific primers (Figure S22D). The two mutants colonized the rhizosphere and endosphere to the same extent as wild type *Flavobacterium* sp. 98 when introduced alone or with *Chitinophaga* sp. 94 (Table S13). When the two independent BGC298 mutants were tested in the disease bioassay, the mutation reduced disease-suppressive activity of Fl98 alone and when paired with the *Chitinophaga* isolate (Figure 5-4F).

5.4 - Conclusions

In our previous studies on soils suppressive to fungal root diseases, we have shown that rhizosphere bacteria act as first line of defense (Mendes et al., 2011, Panke-Buisse et al., 2015, Vannier et al., 2019, Duran et al., 2018). If the pathogen breaks through this first line of defense, it will encounter the basal and induced defense mechanisms of the plant (Jones et al., 2006). Here, we show that in this second stage of pathogen invasion of the plant roots, the endophytic microbiome can provide an additional layer of protection. Our experiments showed that on pathogen invasion, members of the *Chitinophagaceae* and *Flavobacteriaceae* became enriched within the plant endosphere and showed enhanced enzymatic activities associated with fungal cell wall degradation, as well as secondary metabolite biosynthesis encoded by NRPSs and PKSs. Following *de novo* assembly of 25 bacterial genomes from metagenome sequences we were able to reconstruct a synthetic community (syncom) of *Flavobacterium* and *Chitinophaga* that provided disease protection. Site-directed mutagenesis further confirmed the contribution of

BGC298 in *Flavobacterium* to this phenotype. Where these two bacterial genera are localized inside the root tissue and how they interact at the molecular level in the endosphere is not yet known. Possibly, chitinase-generated chitooligosaccharides induce expression of the *Flavobacterium* BGC298. Whether BCG298 encodes a metabolite that exerts direct antifungal activity or acts as a regulator of other protective traits is not yet known. Another consideration is that the consortium may have indirect effects via induction of local or systemic resistance in the roots. The results of this study highlight the wealth of yet unknown microbial genera and functional traits in the endophytic root microbiome. Adopting metagenome-guided analyses and network inference was successful in pinpointing taxa and functions for targeted design of microbial consortia to attain a specific microbiome-associated plant phenotype.

5.6 - Supplementary Figures

Selection of supplementary figures that the thesis author contributed to.

The remaining supplementary figures are available online.

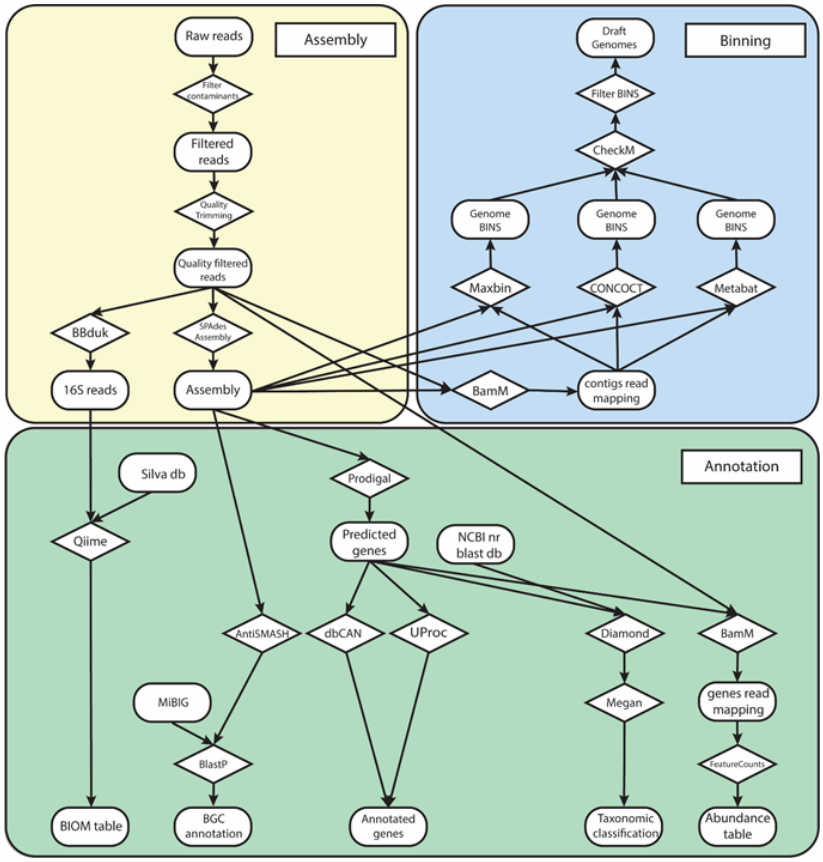


Figure S2. Schematic representation of the bioinformatic tools used for the metagenomics analysis of the endophytic microbiome.

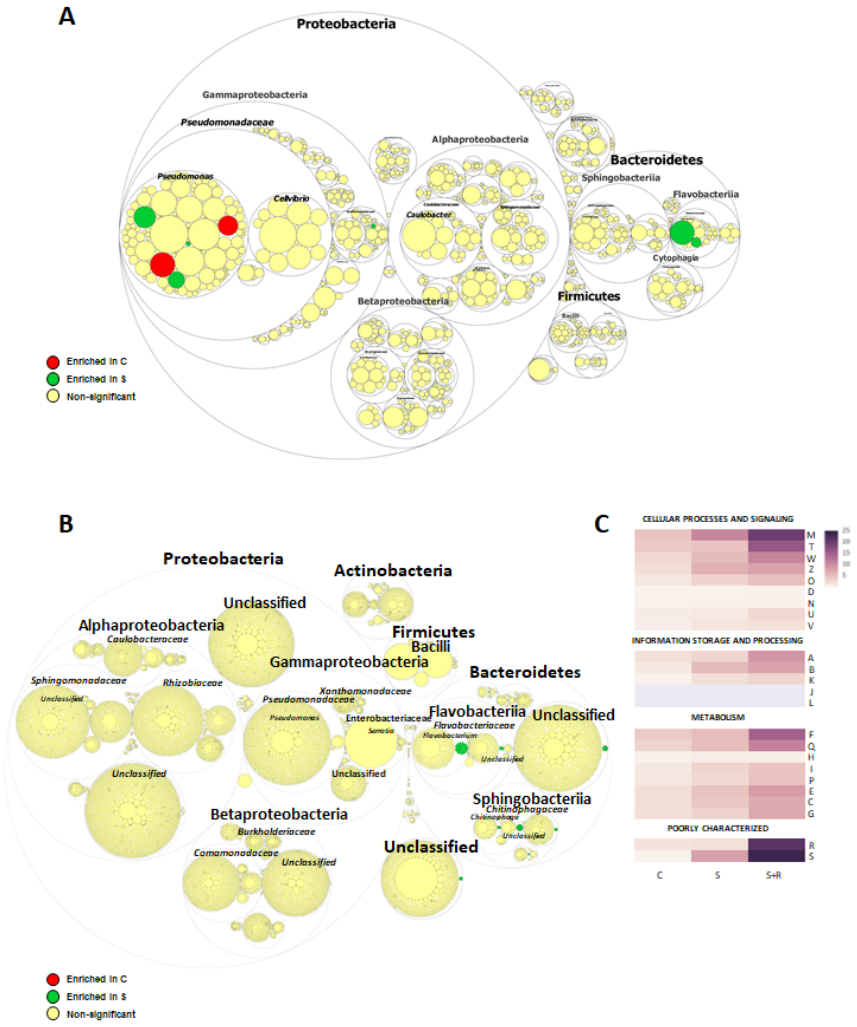


Figure S8. **Differences in endophytic microbiome diversity and functions between *Rhizoctonia*-conducive and suppressive soils**. Differential abundance of endophytic bacterial communities from plants grown in conducive soil (C) or suppressive soil (S). Taxonomic differences (panel **A**) are based on 16S rRNA sequences extracted from the metagenome. Functional differences (panel **B**) are based on the metagenome sequence data and assigned to taxonomic groups. The largest circles represent Phylum level, the inner circles represent Class, Family and Genus for panel (**A**) and for panel (**B**) the smallest circles represent the groups of COG categories. The circle sizes represent the mean read relative abundance of the differentially abundant taxa and functions. Bacterial taxa or functions that are significantly enriched (FDR<0.1) in the comparison between C and S are indicated in red for C and in green for S; non-significant taxa and functions are indicated in yellow. **C**, heat maps depicting the relative abundance of the significantly enriched COG functional categories in S as compared to C (FDR<0.1). Each COG type has been abbreviated D: cell cycle control, cell division, and chromosome partitioning, M: cell wall/membrane/envelope biogenesis, N: cell motility, O: post-translational modification, protein turnover, and chaperones, T: signal transduction mechanisms, U: intracellular trafficking, secretion, and vesicular transport, V: defense mechanisms, W: extracellular structures, Y: nuclear structure, Z: cytoskeleton, A: RNA processing and modification, B: chromatin structure and dynamics, J: translation,

ribosomal structure, and biogenesis, K: Transcription, L: replication, recombination, and repair, C: energy production and conversion, E: amino acid transport and metabolism, F: nucleotide transport and metabolism, G: carbohydrate transport and metabolism, H: coenzyme transport and metabolism, I: lipid transport and metabolism, P: inorganic ion transport and metabolism, Q: secondary metabolites biosynthesis, transport, and catabolism, R: general function prediction only and S: function unknown.

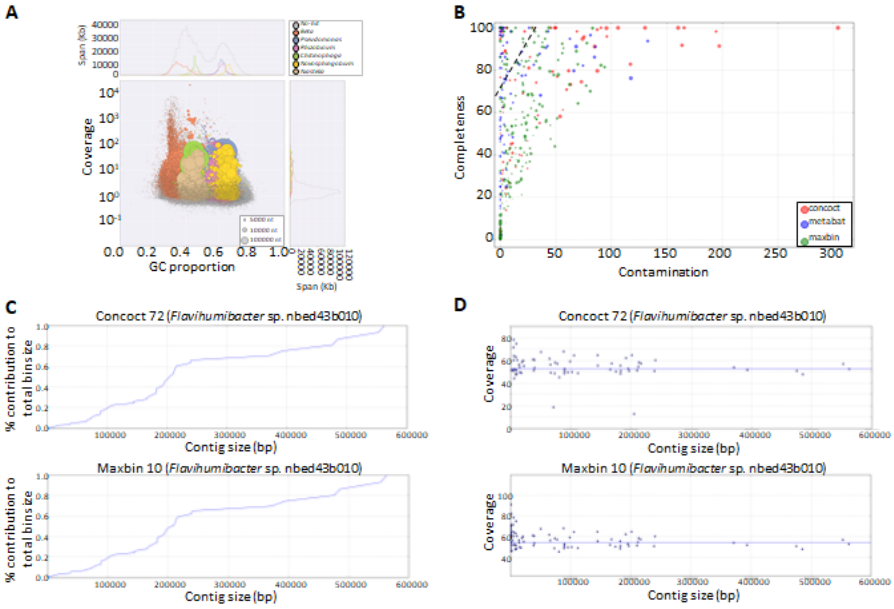
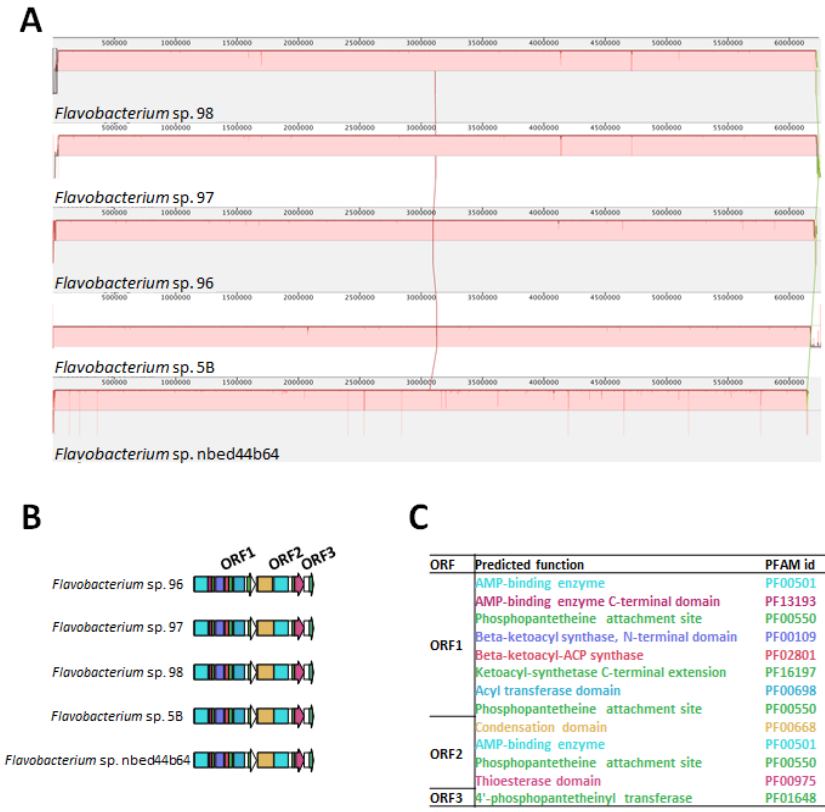


Figure S13. **Binning of the metagenome sequences from the endophytic microbiome.** (**A**) Scatterplot representing the distribution of the assembled contigs based on their GC content and coverage. The dot size is related to the contig size while the color is based on their annotation. (**B**) Representation of the metagenome bins generated by the different binning tools by CheckM completeness and contamination scores. Here, presence/absence and duplication of conserved single copy genes were used to evaluate the bin's quality. The blue line (completeness >70%, contamination < 30%) separates the portions of bins that were considered medium quality draft genomes. (**C**) Example of the impact short and long contigs to the total bin size. Large contigs (above 5kb) contribute to the majority of the bin. **(D),** Dot plot highlighting the coverage and size patterns of contigs (blue dots) assigned to redundant bins. Longer contigs from the redundant bins are consistently assigned to the same bin (dots on the right side) while the shorter ones (dots on the left) do not.



C

ORF	Predicted function	PFAM id
ORF1	AMP-binding enzyme	PF00501
	AMP-binding enzyme C-terminal domain	PF13193
	Phosphopantetheine attachment site	PF00550
	Beta-ketoacyl synthase, N-terminal domain	PF00109
	Beta-ketoacyl-ACP synthase	PF02801
	Ketoacyl-synthetase C-terminal extension	PF16197
	Acyl transferase domain	PF00698
	Phosphopantetheine attachment site	PF00550
ORF2	Condensation domain	PF00668
	AMP-binding enzyme	PF00501
	Phosphopantetheine attachment site	PF00550
	Thioesterase domain	PF00975
ORF3	4'-phosphopantetheinyl transferase	PF01648

Figure S16. Comparative genomics of the sequenced *Flavobacterium* genomes and the metagenome assembled genomes (MAGs). (A) Progressive Mauve alignment of the *Flavobacterium* genomes (strains 5B, 96, 97 and 98) and the MAG 'bin nbed44b64'. Contigs are reordered using the genome of *Flavobacterium* sp. 98 as the reference **(B)** genetic architecture of BGC 298 in 4 *Flavobacterium* genomes and MAG 'bin nbed44b64'. Colors indicate the domains predicted for each gene. **(C)** Table showing the predicted functions of all the domains in each gene belonging to the BGC 298.

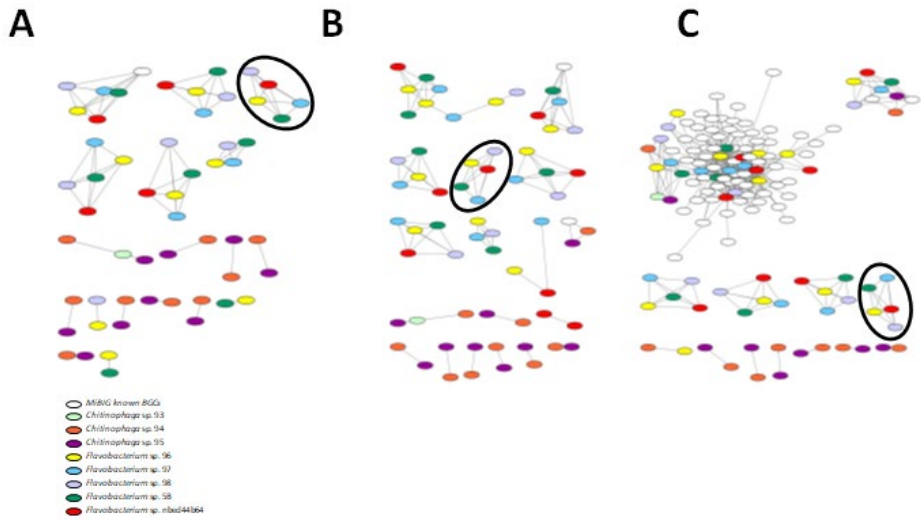


Figure S17. BiG-SCAPE network containing NRPS-containing BGCs related to the known gene clusters in the MiBIG database. (A), **(B)** and **(C)** show a BiG-SCAPE network of the BGCs from the 7 sequenced isolates (*Chitinophaga* sp. 93, 94 and 95, and *Flavobacterium* sp. 5B, 96, 97 and 98) and the binned genome (*Flavobacterium* sp. nbed44b64) at 0.3, 0.5 and 0.7 identity threshold respectively. White nodes are representing the known BGCs in the MiBIG database, nodes with different colors are the different BGCs detected with Antismash in the 7 sequenced genomes and in the binned genome. The black circle in A, B and C is highlighting the gene cluster family containing BGC 298 in the 5 *Flavobacterium* genomes.

References

J. A. Vorholt, C. Vogel, C. I. Carlström, D. B. Mueller, Establishing causality: opportunities of synthetic communities for plant microbiome research. Cell Host Microbe **22**, 142-155 (2017).

V. Cordovez, F. Dini-Andreote, V. J. Carrión, J. M. Raaijmakers, Ecology and evolution of plant microbiomes. Annu. Rev. Microbiol. **73**, 69-88 (2019).

M.-J. Kwak et al., Rhizosphere microbiome structure alters to enable wilt resistance in tomato. Nat. Biotech. **36**, 1100-1109 (2018).

S. Hacquard et al., Microbiota and host nutrition across plant and animal kingdoms. Cell Host Microbe **17**, 603-616 (2015).

R. Mendes et al., Deciphering the rhizosphere microbiome for disease-suppressive bacteria. Science **332**, 1097-1100 (2011).

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

N. Vannier, M. Agler, S. Hacquard, Microbiota-mediated disease resistance in plants. PLoS Path. **15**, e1007740 (2019).

B. O. Oyserman, M. H. Medema, J. M. Raaijmakers, Road MAPs to engineer host microbiomes. Curr. Opin. in Microbiol. **43**, 46-54 (2018).

P. Durán et al., Microbial interkingdom interactions in roots promote Arabidopsis survival. Cell **175**, 973-983. e914 (2018).

D. M. Weller, J. M. Raaijmakers, B. B. M. Gardener, L. S. Thomashow, Microbial populations responsible for specific soil souppressiveness to plant pathogens. Annu. Rev. Phytopathol. **40**, 309-348 (2002).

E. Chapelle, R. Mendes, P. A. H. M. Bakker, J. M. Raaijmakers, Fungal invasion of the rhizosphere microbiome. ISME J **10**, 265-268 (2016).

R. L. Berendsen, C. M. J. Pieterse, P. A. H. M. Bakker, The rhizosphere microbiome and plant health. Trends Plant Sci. **17**, 478-486.

M. Mazzola, Manipulation of rhizosphere bacterial communities to induce suppressive soils. J. Nematol. **39**, 213-220 (2007).

R. Mendes, P. Garbeva, J. M. Raaijmakers, The rhizosphere microbiome: significance of plant beneficial, plant pathogenic, and human pathogenic microorganisms. FEMS Microbiol. Rev. **37**, 634-663 (2013).

J.-Y. Cha et al., Microbial and biochemical basis of a Fusarium wilt-suppressive soil. ISME J **10**, 119-129 (2016).

M. Voort, M. Kempenaar, M. Driel, J. M. Raaijmakers, R. Mendes, Impact of soil heat on reassembly of bacterial communities in the rhizosphere microbiome and plant disease suppression. Ecol. Lett. **19**, 375-382 (2016).

J. M. Raaijmakers, M. Mazzola, Soil immune responses. Science **352**, 1392-1393 (2016).

V. J. Carrión et al., Involvement of Burkholderiaceae and sulfurous volatiles in disease-suppressive soils. ISME J **12**, 2307-2321 (2018).

V. Cordovez et al., Diversity and functions of volatile organic compounds produced by Streptomyces from a disease-suppressive soil. Front. Microbiol. **6**, 1081 (2015).

K. Faust, J. Raes, Microbial interactions: from networks to models. Nat. Rev. Microbiol. **10**, 538-550 (2012).

Y. Yin et al., dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. **40**, W445-W451 (2012).

V. Lombard, H. G. Ramulu, E. Drula, P. M. Coutinho, B. Henrissat, The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. **42**, D490-D495 (2014).

T. Weber et al., antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. **43**, W237-243 (2015).

M. Remmert, A. Biegert, A. Hauser, J. Soding, HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Meth. **9**, 173-175 (2012).

S. M. Bowman, S. J. Free, The structure and synthesis of the fungal cell wall. BioEssays **28**, 799-808 (2006).

J. Watrous et al., Mass spectral molecular networking of living microbial colonies. Proc. Natl. Acad. Sci. U.S.A. **109**, E1743–E1752 (2012).

M. H. Medema et al., Minimum Information about a Biosynthetic Gene cluster. Nat. Chem. Biol. **11**, 625-631 (2015).

P. Cimermancic et al., Insights into secondary metabolism from a global analysis of prokaryotic biosynthetic gene clusters. Cell **158**, 412-421 (2014).

Y. Schmidt et al., Biosynthetic origin of the antibiotic cyclocarbamate brabantamide A (SB-253514) in plant-associated Pseudomonas. ChemBioChem **15**, 259-266 (2014).

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

R. Gómez Expósito, J. Postma, J. M. Raaijmakers, I. de Bruijn, Diversity and activity of Lysobacter species from disease suppressive soils. Front. Microbiol. **6**, 1243 (2015).

E. Chapelle et al., A straightforward and reliable method for bacterial in planta transcriptomics: application to the Dickeya dadantii/Arabidopsis thaliana pathosystem. The Plant J. **82**, 352-362 (2015).

S. Courtois et al., Quantification of bacterial subgroups in soil: comparison of DNA extracted directly from soil or from cells previously released by density gradient centrifugation. Environ. Microbiol. **3**, 431-439 (2001).

P. N. Holmsgaard et al., Bias in bacterial diversity as a result of Nycodenz extraction from bulk soil. Soil Biol. Biochem. **43**, 2152-2159 (2011).

A. Hevia, S. Delgado, A. Margolles, B. Sánchez, Application of density gradient for the isolation of the fecal microbial stool component and the potential use thereof. Sci. Rep. **5**, 16807 (2015).

S. Ikeda et al., Development of a bacterial cell enrichment method and its application to the community analysis in soybean stems. Microb. Ecol. **58**, 703-714 (2009).

N. A. Joshi, J. N. Fass, Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files. (Version 1.33 [Software]. Available at https://github.com/najoshi/sickle., 2011).

B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat. Meth. **9**, 357-359 (2012).

A. Bankevich et al., SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. **19**, 455-477 (2012).

D. Hyatt et al., Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics **11**, 119 (2010).

C. Trapnell et al., Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7**, 562-578 (2012).

B. Buchfink, C. Xie, D. H. Huson, Fast and sensitive protein alignment using DIAMOND. Nat. Meth. **12**, 59-60 (2015).

D. H. Huson, A. F. Auch, J. Qi, S. C. Schuster, MEGAN analysis of metagenomic data. Genome Res. **17**, 377-386 (2007).

P. Meinicke, UProC: tools for ultra-fast protein domain classification. Bioinformatics **31**, 1382-1388 (2015).

M. Kanehisa, S. Goto, KEGG: Kyoto Encyclopedia of Genes and Genomes. Nucleic Acids Res. **28**, 27-30 (2000).

M. Y. Galperin, K. S. Makarova, Y. I. Wolf, E. V. Koonin, Expanded microbial genome coverage and improved protein family annotation in the COG database. Nucleic Acids Res. **43**, D261-269 (2015).

R. D. Finn et al., The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. **44**, D279-D285 (2016).

M. Imelfort, B. Woodcroft, D. Parks, Ecogenomics/BamM [WWWDocument] (Available at: https://github.com/Ecogenomics/BamM, 2015).

H. Li et al., The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**, 2078-2079 (2009).

H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv:1303.3997 [q-bio], (2013).

Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**, 923-930 (2014).

B. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner. Lawrence Berkeley National Laboratory. LBNL Report #: LBNL-7065E. Retrieved from https://escholarship.org/uc/item/1h3515gn. (2014).

C. Quast et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Res. **41**, D590-D596 (2013).

J. G. Caporaso et al., QIIME allows analysis of high-throughput community sequencing data. Nat. Meth. **7**, 335-336 (2010).

T. Rognes, T. Flouri, B. Nichols, C. Quince, F. Mahé, VSEARCH: Versatile open-source tool for metagenomics. PeerJ 4:e2584 (2016).

D. McDonald et al., The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. GigaScience **1**, 7 (2012).

J. Köster, S. Rahmann, Snakemake—a scalable bioinformatics workflow engine. Bioinformatics **28**, 2520-2522 (2012).

J. Kuczynski et al., Using QIIME to analyze 16S rRNA gene sequences from microbial communities. Curr. Protoc. Microbiol. **27**, 1E. 5.1-1E. 5.20 (2012).

J. Oksanen et al., The vegan package. Community ecology package **10**, 631-637 (2007).

J. N. Paulson, O. C. Stine, H. C. Bravo, M. Pop, Differential abundance analysis for microbial marker-gene surveys. Nat. Meth. **10**, 1200-1202 (2013).

P. J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PloS one **8**, e61217 (2013).

M. E. Ritchie et al., limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. **43**, e47-e47 (2015).

J. Friedman, E. J. Alm, Inferring correlation networks from genomic survey data. PLoS Comput. Biol. **8**, e1002687 (2012).

M. E. Newman, The structure and function of complex networks. SIAM review **45**, 167-256 (2003).

M. E. Newman, Modularity and community structure in networks. Proc. Natl. Acad. Sci. U.S.A. **103**, 8577-8582 (2006).

P. Shannon et al., Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**, 2498-2504 (2003).

M. Bastian, S. Heymann, M. Jacomy, in Third international AAAI conference on weblogs and social media. (2009).

T. Weber et al., antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Res. **43**, W237-W243 (2015).

J. Navarro-Muñoz et al., A computational framework for systematic exploration of biosynthetic diversity from large-scale genomic data. BioRxiv, 445270 (2018).

M. H. Medema, E. Takano, R. Breitling, Detecting sequence homology at the gene cluster level with MultiGeneBlast. Mol. Biol. Evol. **30**, 1218-1223 (2013).

J. Alneberg et al., Binning metagenomic contigs by coverage and composition. Nat. Meth. **11**, 1144 (2014).

Y.-W. Wu, Y.-H. Tang, S. G. Tringe, B. A. Simmons, S. W. Singer, MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. Microbiome **2**, 1-18 (2014).

D. D. Kang, J. Froula, R. Egan, Z. Wang, MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ **3**, e1165 (2015).

D. H. Parks, M. Imelfort, C. T. Skennerton, P. Hugenholtz, G. W. Tyson, CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. Genome Res. **25**, 1043-1055 (2015).

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

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

S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, S. R. J. Eddy, Rfam: an RNA family database. Nucleic Acids Res. **31**, 439-441 (2003).

K. D. Pruitt, T. Tatusova, D. R. J. N. a. r. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. **35**, D61-D65 (2006).

R. C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. **32**, 1792-1797 (2004).

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

I. Letunic, P. Bork, Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. Nucleic Acids Res. **39**, W475-W478 (2011).

A. Untergasser et al., Primer3Plus, an enhanced web interface to Primer3. Nucleic Acids Res. **35**, W71-W74 (2007).

M. Kolmogorov, J. Yuan, Y. Lin, P. A. Pevzner, Assembly of long, error-prone reads using repeat graphs. Nat. Biotech. **37**, 540-546 (2019).

B. D. Ondov et al., Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol. **17**, 132 (2016).

I. Lee, Y. O. Kim, S.-C. Park, J. Chun, OrthoANI: an improved algorithm and software for calculating average nucleotide identity. Int. J. Syst. Evol. Micr. **66**, 1100-1103 (2016).

N. Segata, D. Börnigen, X. C. Morgan, C. Huttenhower, PhyloPhlAn is a new method for improved phylogenetic and taxonomic placement of microbes. Nat. Commun. **4**, 2304 (2013).

Y. Bai et al., Functional overlap of the Arabidopsis leaf and root microbiota. Nature **528**, 364-369 (2015).

M. McBride, M. Kempf, Development of techniques for the genetic manipulation of the gliding bacterium Cytophaga johnsonae. J. Bacteriol. **178**, 583-590 (1996).

S. Chen, M. Bagdasarian, M. Kaufman, E. Walker, Characterization of strong promoters from an environmental Flavobacterium hibernum strain by using a green fluorescent protein-based reporter system. Appl. Environ. Microbiol. **73**, 1089-1100 (2007).

M. J. McBride, S. A. Baker, Development of techniques to genetically manipulate members of the genera Cytophaga, Flavobacterium, Flexibacter, and Sporocytophaga. Appl. Environ. Microbiol. **62**, 3017-3022 (1996).

J. A. Gagnon et al., Efficient mutagenesis by Cas9 protein-mediated oligonucleotide insertion and large-scale assessment of single-guide RNAs. PloS one **9**, e98186 (2014).

S. Chen, M. Bagdasarian, M. G. Kaufman, A. K. Bates, E. D. Walker, Mutational analysis of the ompA promoter from Flavobacterium johnsoniae. J. Bacteriol. **189**, 5108-5118 (2007).

H. Xie, P. Nie, B. Sun, Characterization of two membrane-associated protease genes obtained from screening out-membrane protein genes of Flavobacterium columnare G4. J. Fish Dis. **27**, 719-729 (2004).

L.-E. Jao, S. R. Wente, W. Chen, Efficient multiplex biallelic zebrafish genome editing using a CRISPR nuclease system. Proc. Natl. Acad. Sci. U.S.A. **110**, 13904-13909 (2013).

S. Agarwal, D. W. Hunnicutt, M. J. McBride, Cloning and characterization of the Flavobacterium johnsoniae (Cytophaga johnsonae) gliding motility gene, gldA. Proc. Natl. Acad. Sci. U.S.A. **94**, 12139-12144 (1997).

Data, scripts and code used for statistical and bioinformatic analyses are available at:.

Acknowledgments: We thank Irene de Bruijn, Silvia P. Vega-Hernández, Victor de Jager and Roos Keijzer for their valuable advice for genomic and metagenomic DNA extractions. We also would like to thank the BSc and MSc students Cristina Rotoni, Ryan Hijkoop, Hannah McDermott, Azkia Nurfikari, Jelle Spooren and Rik Peters for their valuable help in the *Flavobacterium* and *Chitinophaga* isolation and phenotyping. We thank Mark J McBride at the University of Wisconsin for providing the plasmids for *Flavobacterium* transformation. This is publication number --- of the NIOO-KNAW. **Data and materials availability:** All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. Additional data related to this paper may be requested from the authors. The sequencing data are available under EBI submission number PRJEB8920. Data, scripts and code used for statistical and bioinformatic analyses are available at: http://doi.org/10.5281/zenodo.3405564 (http://doi.org/10.5281/zenodo.3405564).

CHAPTER 6

BiosyntheticSPAdes: Reconstructing biosynthetic gene clusters from assembly graphs

Authors:

Dmitry Meleshko, Hosein Mohimani, Vittorio Tracanna, Iman Hajirasouliha, Marnix H. Medema, Anton Korobeynikov, Pavel A. Pevzner[#]

#: Corresponding author

Based on “BiosyntheticSPAdes: Reconstructing biosynthetic gene clusters from assembly graphs” published in *Genome Research*, 29(8):1352-1362 (2019).

6.1 - Abstract

Predicting Biosynthetic Gene Clusters (BGCs) is critically important for discovery of antibiotics and other natural products. While BGC prediction from complete genomes is a well-studied problem, predicting BGC in fragmented genomic assemblies remains challenging. The existing BGC prediction tools often assume that each BGC is encoded within a *single* contig in the genome assembly, a condition that is violated for most sequenced microbial genomes where BGCs are often scattered through several contigs, making it difficult to reconstruct them. The situation is even more severe in shotgun metagenomics, where the contigs are often short, and the existing tools fail to predict a large fraction of long BGCs. While it is difficult to assemble BGCs in a single contig, the structure of the genome assembly graph often provides clues on how to combine multiple contigs into segments encoding long BGCs. We describe biosyntheticSPAdes, a tool for predicting BGCs in assembly graphs and demonstrate that it greatly improves the reconstruction of BGCs from genomic and metagenomics datasets.

6.2 – Introduction

Although there exist many tools for assembling microbial genomes or metagenomes (Simpson et al., 2009, Li et al., 2015, Nurk et al., 2017), they all have limitations with respect to assembling contigs that contain long genes encoding proteins with repetitive domains. Since long genes are often scattered between multiple contigs in fragmented assemblies, the existing gene prediction tools (Besemer et al., 2005, Delcher et al., 2007, Pati et al., 2010, Hyatt et al., 2010) cannot predict them. The challenge of assembling long genes in a single contig is illustrated by genes encoding *Nonribosomal Peptides Synthetases* (*NRPSs*), *Polyketide Synthases (PKSs),* and other genes that are parts of *biosynthetic gene clusters (BGCs)* encoding the production of antibiotics and other natural products. BGCs usually include multiple consecutive genes that participate in a single metabolic pathway responsible for synthesizing a natural product. NRPS BGCs encode *Nonribosomal Peptides* (*NRPs*) built from amino acids and PKS BGCs encode *polyketides* (*PSs*) built from keto groups. Mixed NRPS/PKS BGCs contain both NRPS-specific and

PKS-specific domains and their natural products represent fusions of peptides and polyketides (Cane et al., 1999). Klassen and Currie, 2012 showed that long and repetitive NRPSs and PKSs are responsible for a large fraction of fragmentation in microbial assemblies.

This paper focuses on NRPSs because NRPs represent an important class of natural product drugs (Newman and Cragg, 2016) that is most amenable to downstream peptidogenomics analysis as compared to other classes of natural products (Kersten et al., 2011, Mohimani et al., 2014, Medema et al., 2014).

NRPS BGCs constitute 34% of all BGCs in publicly available genomes, as found in the antiSMASH database (https://antismash-db.secondarymetabolites.org/#!/stats).

Since NRPSs are very common (albeit elusive) in diverse bacterial datasets (Mukherjee et al., 2017) and since the downstream peptidogenomics analysis of NRPs is greatly impaired by fragmented assemblies, most examples in this paper refers to NRPs. In addition to NRPS BGCs, biosynteticSPAdes is also applicable to PKS BGCs and mixed NRPS-PKS BGCs (NRPS, PKS, and mixed NRPS-PKS BGCs constitute the majority of BGCs in the MIBiG database). Klassen and Currie, 2012 have shown that fragmented ORFs in genome assemblies are highly enriched in NRPSs and PKSs, which thus constitute a prominent source of breakpoint in (meta)genome assemblies. The fact that the vast majority of genomes contain either an NRPS or a PKS, or a mixed NRPS-PKS BGC (for some species, over 30% of the genome is allocated to these BGCs) and direct interest to a large research community is a good reason to provide a specialized assembler for these BGCs.

NRPSs are large modular protein complexes containing multiple highly similar *adenylation domains* (*A-domains)* responsible for recruiting amino acids that form NRPs according to the substrate specificity of each A-domain (Stachelhaus and Marahiel 1999). NRPSs are often accompanied by other adjacently located genes that together form NRP BGCs and contribute to NRP synthesis, transport, and regulation. NRP BGCs are typically long with an average length of ~60 kb and some exceeding 100 kb in length. Assembling NRP BGCs into single contigs is a crucial

step in natural product discovery by genome mining (Weber et al., 2015) and peptidogenomics (Mohimani et al., 2014, Mohimani and Pevzner, 2016, Mohimani et al., 2017, Gurevich et al., 2018).

The recent *Genomic Encyclopedia of Bacteria and Archaea (GEBA)* study of over 1000 bacterial genomes revealed over 23,000 BGCs (Mukherjee et al., 2017). An average GEBA genome devotes nearly 10% of its genome to BGCs (some genomes devote >30%). However, the vast majority of predicted BCG products remain unknown, in part due to difficulties in predicting long BGCs (Hadjithomas et al., 2015).

The recently proposed genome mining and peptidogenomic approaches elucidate the amino acid sequences of NRPs by matching tandem mass spectra against predicted NRP synthetases in the assembled genomes (Mohimani et al., 2014, Medema et al., 2014, Mohimani et al., 2017). The success of these approaches depends on accurate prediction of genes encoding NRP synthetases followed by machine-learning algorithms to predict their substrate specificities, and matching mass spectral datasets against the predicted NRP amino acid sequences. This is a challenging task requiring the recovery of the *complete* NRPS genes and the corresponding NRP BGCs in a single contig.

This challenge is further amplified in metagenomics assemblies, because NRP synthetases from different species within a microbial community often share similar domains. This makes it difficult to assemble them in a single contig in cases when multiple domains are collapsed into a single edge in the assembly graph (Coates et al., 2014). Therefore, while metagenomes represent a gold mine for antibiotics discovery, a limited number of antibiotics have been discovered from metagenomics datasets so far (Freeman et al., 2012, Donia et. al., 2014, Donia and Fischbach, 2015).

Despite the fact that it is difficult to reconstruct long NRPS BGCs from individual contigs, the structure of the assembly graph often provides clues on how to combine various contigs into intact BGCs. We describe the biosyntheticSPAdes tool for

assembling NRPS BGCs in assembly graphs constructed by SPAdes (Bankevich et al., 2012) and metaSPAdes (Nurk et al., 2017) assemblers. Below we show how biosyntheticSPAdes contributes to the discovery of NRPS BGCs in various genomes and metagenomes.

6.3 – Results

6.3.1 - The challenge of assembling BGCs

Contrary to the standard practice in existing gene prediction tools that attempt to reconstruct genes from *individual* contigs/scaffolds, biosyntheticSPAdes analyzes the assembly graph to join fragments of long BGCs (scattered over multiple contigs) into a single contig. Below, we describe the biosyntheticSPAdes algorithm and illustrate how it works using the genome of *Streptomyces coelicolor A3(2)* (referred to as *S. coelicolor* for brevity), a well-studied antibiotics-producing bacterium, which encodes four NRP BGCs (Bentley et al., 2002), including *calcium-dependent antibiotic (CALC)*.

We illustrate the challenge of assembling long repetitive genes using a subgraph of the *S. coelicolor* assembly graph encoding the CALC BGC (Figure 6-1). To generate this graph, we simulated error-free short paired-end reads (Huang et al., 2012) from the *S. coelicolor* genome using the ART read simulator. The reads from the resulting dataset with coverage 180× (referred to as the STREP dataset and containing paired reads of length 150 bp with mean insert size 300 bp) were assembled using the SPAdes assembler (Bankevich et al., 2012). The assembly graph constructed from these simulated reads contains 626 vertices and 697 edges (484 of them are longer than 1000 bp). The total edge length in the assembly graph is 8,598,860 with N50=41 kb. SPAdes uses paired reads to resolve repeats in the genome and combines some edges in the assembly graphs into contigs/scaffolds using exSPAnder (Prjibelsky et al., 2014). exSPAnder constructed 145 scaffolds longer than 1000 bp with N50=135 kb after the repeat resolution step.

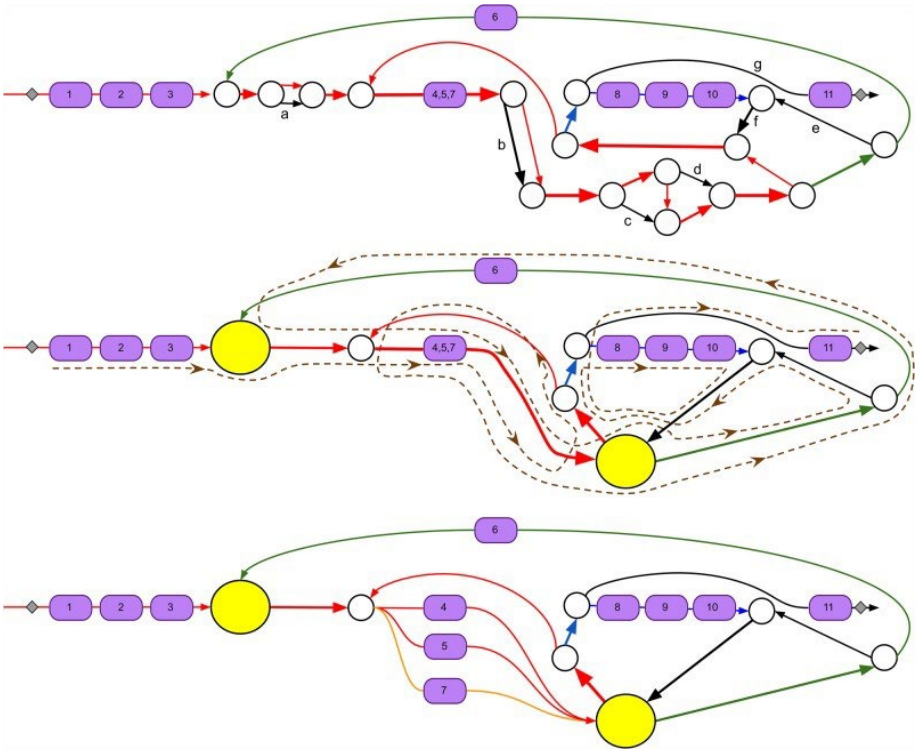


Figure 6-1. Subgraph of the assembly graph of *S. coelicolor* corresponding to the CALC NRP BGC. (Top) Edges of the assembly graph traversed by the CALC BGC. Nodes of the assembly graph are shown as white circles. After applying exSPAnder, the CALC BGC remains scattered over ten scaffolds. Three of them are shown as red, blue, and green paths through the assembly graph, the remaining seven consist of a single edge each (shown in black and marked with letters a through g). The positions of eleven A-domains (with their indices) along the CALC BGC are shown by violet boxes. Edges with low and high coverage by reads are shown as thin and thick edges, respectively. The edge harboring three A-domains 4, 5, and 7 has approximately triple coverage by reads as compared to other domain-harboring edges. The 11 A-domains in CALC are split over three NRP synthetases with 6, 3, and 2 A-domains, respectively. (Middle) A simplified representation of the graph with all short edges (shorter than 300 bp) contracted into single vertices. The two contracted subgraphs of the assembly graph (formed by short edges) are represented by yellow vertices. The brown dashed path illustrates how the CALC NRP synthetase traverses the contracted assembly graph. (Bottom) The bubble restoration procedure described below transforms the collapsed edge harboring three A-domains (A-domains 4, 5, and 7) into three edges, each of them harboring a single A-domain. Applying exSPAnder to the modified assembly graph results in seven scaffolds that differ from scaffolds before bubble restoration (shown as red, blue, green, and orange paths as well as three black edges). Grey squares show the starting and ending positions of the CALC BGC.

AntiSMASH (Weber et al., 2015) is a popular genome mining tool for detecting and annotating BGCs. AntiSMASH revealed 29 BGCs in the *S. coelicolor* genome, including four NRP BGCs. The CALC BGC with eleven A-domains traverses 25

edges in the assembly graph. exSPAnder (Prjibelsky et al., 2014) combined some of these edges into single contigs, but even after applying exSPAnder, CALC was split into 7 scaffolds (Figure 6-1). This illustrates the challenge of reconstructing long genes even for isolated bacteria, let alone metagenomes. Note that 11 A-domains in CALC are represented by only 9 A-domains in Figure 6-1 because 3 out of 11 A-domains got collapsed into a single edge in the assembly graph.

The CALC BGC illustrates just one example of the difficulties with assembling long and repetitive genes in genomic and metagenomic datasets. Table 6.2 illustrates that 285 out of 7,910 genes (≈ 3%) in the *S. coelicolor* genome are split over multiple edges in the assembly graph. The fraction of split genes further increases when we consider long genes: 11 out of the 100 longest genes (length > 3200 bp) traverse multiple edges and 17 out of these 100 longest genes corresponds to BGCs (Table 6.2). While the repeat resolution step in SPAdes (Prjibelsky et al., 2014) captures some of the split genes in a single contig/scaffold, many long genes remain split even after repeat resolution and three of them correspond to BGC genes (Supplementary Table S3). The fraction of such split genes further increases in metagenomics assemblies.

6.3.2 - BiosyntheticSPAdes outline

The biosyntheticSPAdes pipeline includes six steps (Figure 6-2) that are described in the Methods section:

- assembling genomic/metagenomic reads with SPAdes/metaSPAdes
- identifying domain-edges in the assembly graph,
- extracting BGC subgraphs from the assembly graph
- restoring collapsed domains in the assembly graph,
- constructing the scaffolding graph
- constructing putative BGCs by solving the Rural Postman Problem in the scaffolding graph.

6.3.3 - Benchmarking design

To benchmark biosyntheticSPAdes, we compared its output (a single or multiple contigs) against the reference genome(s). Since the downstream applications, such as NRPquest (Mohimani et al., 2014), do not require a single contig output and work equally well when a small set of output contigs contain a correct one, we classify the biosyntheticSPAdes output as correct if at least one of the reported contigs is contained in one of the reference genomes (with percent identity exceeding 95%).

In the case when the reference genomes are not available, we check whether a BGC subgraph contains a rural postman path. If it is the case, it is likely that one of the reported contigs is contained in an unknown reference genome.

6.3.4 – Datasets

We analyzed the following datasets assembled using SPAdes or metaSPAdes with *k*-mer sizes varying from 21 to 55 nucleotides during the iterative assembly.

Pseudomonas datasets (PSEUDO). The PSEUDO dataset (accession number ERR1890333) contains ≈4.5 million paired reads from the isolate of *Pseudomonas protegens (fluorescens) pf-5* (read length 100 bp, a mean insert size 440 bp, and a standard deviation of the insert size 140 bp). The genome sequence was finished using a combination of primer walking, generation and sequencing of transposon-tagged libraries, and multiplex PCR (Paulsen et al., 2005).

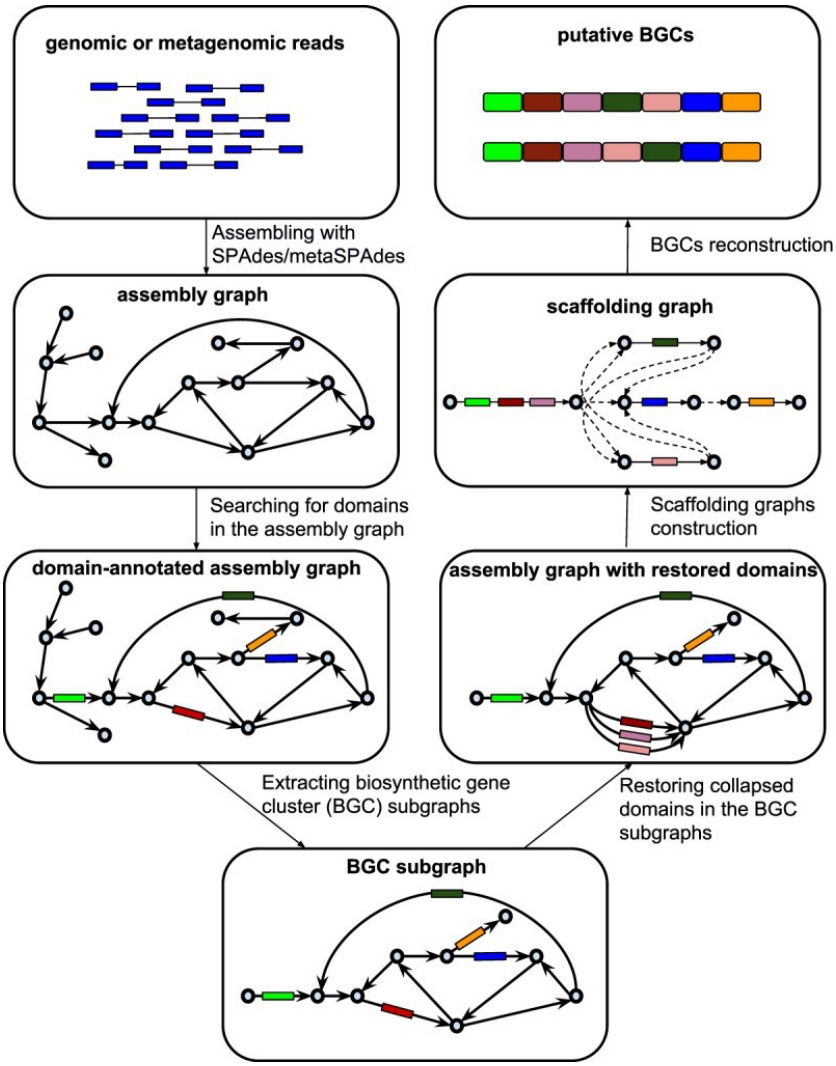


Figure 6-2. The biosyntheticSPAdes pipeline. Six steps of the biosyntheticSPAdes pipeline: (i) assembling genomic/metagenomic reads with SPAdes/metaSPAdes, (ii) searching for edges harboring biosynthetic domains in the assembly graph, (iii) extracting biosynthetic gene cluster subgraphs from the assembly graph, (iv) restoring the collapsed domains in the BGC-subgraphs, (v) constructing the scaffolding graph, and (vi) generating putative BGC by solving the Rural Postman Problem in the scaffolding graph.

Cyanobacteria dataset (CYANO). The CYANO dataset contains genomic reads from cultured marine bacteria *Moorea producens JHB* (referred as JHB below) described

in Kleigrewe et al., 2015. The sample is contaminated with heterotrophic bacteria and thus represents a low-complexity metagenome. The JHB strain encodes various NRPs, PKs and mixed NRP-PKs, including *hectochlorin* (Marquez et al., 2002) and *jamaicamides* (Edwards et al., 2004). The JHB dataset contains ≈6 million paired reads (length 150 bp, a mean insert size 292 bp, and a standard deviation of the insert size 74 bp).

MIBiG datasets (MIBIG). The Minimum Information about a Biosynthetic Gene Cluster (MIBiG) database contains information about BGCs and their products (Medema et al., 2015). Each entry in the MIBiG database contains the nucleotide sequence of a BGC, the natural product type (NRPs, PKs, and other types), and its annotation. In order to benchmark biosyntheticSPAdes on a wide range of BGCs, we extracted all MIBiG entries describing NRPSs and PKSs with complete BGC sequences (665 entries) and used the ART read simulator (Huang et al., 2012) to simulate reads from BGC sequences with the default MiSeq parameters. Admittedly, generating reads from BGCs results in a simpler problem than simulating reads from the entire genome. However, since entire genomes are not available for many MIBiG entries, we simulated reads from BGCs only. We define the *complexity* of a BGC as the total number of A-domains and AT-domains in this BGC. Note that this is a very naïve definition of complexity (e.g., trans-AT PKSs have few AT domains). 139 out of 665 BGCs in the MIBiG dataset have complexity 10 and larger.

HMP datasets (HMP). The HMP dataset consists of 20 metagenomic sub-datasets from seven parts of human body that included keratinized gingiva, buccal mucosa, stool, gingivival plaque, subpravingal plaque, tongue dorsum, and throat (Table S4). The description of these datasets is given in Methé et al., 2012.

6.3.5 - Analyzing the PSEUDO dataset

AntiSMASH (Weber et al., 2015) identified 12 BGCs in the *Pseudomonas protegens pf-5* genome, including seven NRP and PK BGCs. SPAdes assembled each of them into a single contig with the exception of the pyoverdine NRP BGC (with eight A-domains), which was assembled into four contigs that revealed only seven A-domains (Figure 6-3, top left). In contrast, the domain restoration procedure in biosyntheticSPAdes succeeded in reconstructing two A-domains that were collapsed on a single edge by SPAdes (Figure 6-3, top right). The resulting scaffolding graph contains a single rural postman route that revealed the correct arrangement of A-domains (Figure 6-3, bottom). The reconstructed pyoverdine NRP BGC aligns to the *Pseudomonas protegens pf-5* genome with 99.9% identity.

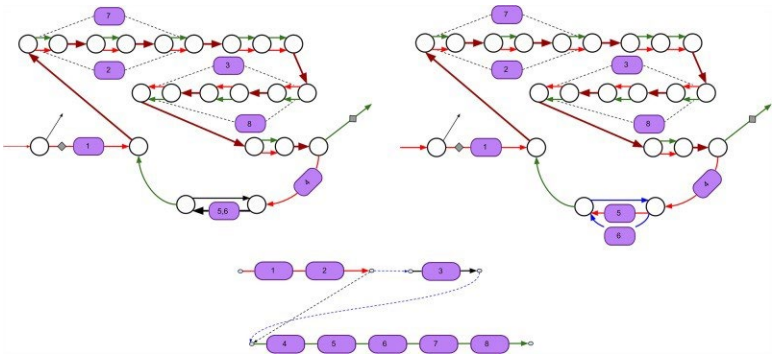


Figure 6-3. Subgraph of the assembly graph of *Pseudomonas protegens* Pf-5 corresponding to the pyoverdine NRP BGC. (Top left) The pyoverdine BGC is scattered over four scaffolds in the SPAdes assembly. Two scaffolds traversing single edges are shown by black color and two scaffolds traversing multiple edges are shown by red and green colors. The repeat edges traversed by both red and green scaffolds are shown by brown color. Edges with low and high depth of coverage by reads are shown as thin and thick edges, respectively. Some A-domains span multiple edges (starting and ending positions of such domains are shown with dashed lines). (Top right) The domain restoration procedure restored two A-domains (5 and 6) in the assembly (SPAdes collapsed these domains into a single edge). Four scaffolds in the assembly graph are shown by red, green, blue and black colors. (Bottom) The scaffolding graph of the pyoverdine BGC with a single rural postman route (dashed edges in this route are shown in blue).

6.3.6 - Analyzing the CYANO dataset

Kleigrewe et al., 2015 assembled the CYANO dataset using SPAdes. metaSPAdes assembled the CYANO dataset into the assembly graph with 217,826 vertices and 116,066 edges (8454 of them are longer than 1 kb). metaSPAdes assembled the jamaicamide BGC with complexity 9 into a single contig but failed to assemble the hectochlorin BGC with complexity 5 into a single contig.

biosyntheticSPAdes extracted 781 BGC subgraphs, including 12 non-trivial BGC subgraphs with complexities 21, 20, 11, 9, 6, 6, 5, 5, 5, 5, 4, and 4. The hectochlorin BGC contains 22 domains (four A-domains, one AT-domain, four C-domains, one KS-domain, three KR domains and several others, one of them was also identified by HMMER as A-domain). biosyntheticSPAdes assembled the hectochlorin BGCs into a single contig (Figure 6-4) that aligns with the *Moorea producens* JHB genome with 99.9% identity. The jamaicamide BGC contains 42 domains (three A-domains, six AT-domains, four KR-domains, seven KS-domains, two C-domains, one TE-domain and several others). The jamaicamide scaffolding graph contains a single solid edge (usually, it means that the entire BGC was recovered after the repeat resolution step with exSPAnder).

Besides reconstructing the hectochlorin and the jamaicamides BGCs, biosyntheticSPAdes recovered sequences for 5 more putative NRP BGCs that were missed in previous studies (see Appendix: “Putative NRP BGCs in the CYANO dataset).

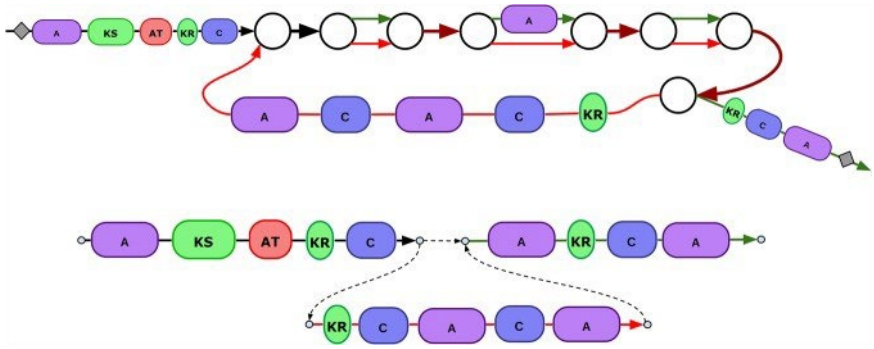


Figure 6-4. biosyntheticSPAdes assembly of the hectochlorin BGCs (the CYANO dataset). (Top) The subgraph of the assembly graph corresponding to the hectochlorin BGC. metaSPAdes assembly results in 4 scaffolds shown by a red path, a green path, and two black edges. The repeat edges traversed by both red and green scaffolds are shown by the brown color. The domain restoration procedure had no effect on this graph. (Bottom) The scaffolding graph of the hectochlorin BGC has only one rural postman route that revealed the correct domain order.

6.3.7 - Analyzing the MIBiG datasets

For each of 665 MIBiG datasets corresponding to a single known NRP or PK, we launched biosyntheticSPAdes on the SPAdes assembly graph. We also compared them with the other popular assemblers: MEGAHIT v.1.1.3 (Li et al., 2015) and ABySS assembler v.2.1.0 (Simpson et al., 2009). For each assembler and each MiBiG dataset. the assembly was classified as successful if it meets the following criteria: (i) one of the contigs in the assembly covers more than 95% of the BGC and has at least 95% identity with the BGC being assembled and (ii) this contig has no misassemblies as identified by QUAST (Gurevich et al., 2013). biosyntheticSPAdes failed to successfully assemble only 11% of BGCs versus 22% for SPAdes, 35% for MEGAHIT and 34% for ABySS (Table 1). For 139 out of 665 BGCs with complexity >10, biosyntheticSPAdes failed to successfully assemble 22% of BGCs versus 58% for SPAdes, 79% for MEGAHIT and 83% for ABySS.

Assembler	Failed to assemble				
	all BGCs	**BGCs with complexity 1-3**	**BGCs with complexity 4-6**	**BGCs with complexity 7-9**	**BGCs with complexity 10**
SPAdes	149 (22%)	6 (2%)	23 (18%)	39 (37%)	81 (58%)
SPAdes + domain restoration	121 (18%)	9 (3%)	24 (19%)	30 (28%)	58 (41%)
biosyntheticSPAdes	69 (11%)	7 (2%)	16 (13%)	16 (16%)	30 (22%)
MEGAHIT	235 (35%)	`28 (10%)	34 (27%)	60 (58%)	111 (79%)
ABySS	227 (34%)	15 (5%)	35 (27%)	58 (56%)	117 (83%)

Table 6-1. Results of Spades, biosyntheticSPAdes, MEGAHIT, and ABySS assemblies on 665 MIBiG data sets

6.3.8 - Analyzing the HMP datasets

To reconstruct BGCs in the human microbiome, we assembled each HMP dataset with biosyntheticSPAdes. We define the *biosynthetic capacity* of an assembly as the number of A and AT domains identified in this corresponding assembly. The biosynthetic capacity of the HMP datasets varies from 60 to over 400 across various human body sites (see Appendix: “Biosynthetic capacity of the HMP datasets”), suggesting that many HMP samples may encode over a dozen of NRP and PK BGCs. However, the amount of high-complexity BGC subgraphs suggests that sequencing depth in some datasets from the HMP project may be insufficient to capture the diversity of BGCs.

Below, we focus on analyzing the subpravingal plaque metagenome (dataset SRS013723) with large biosynthetic capacity. The assembly graph of this dataset contains 1540 BGC subgraphs, including seven non-trivial BGC subgraphs. We analyzed one of the complex BGC subgraphs with six predicted A-domains, five C-domains and two TE-domains that was assembled into six contigs. Figure 6-5 shows the BGC subgraph and two rural postman routes in the scaffolding graph generated by biosyntheticSPAdes. A nucleotide BLAST search of two predicted BGCs against the nt/rt database revealed only the short regions of similarity (less than 200 bp) with various *Pseudomonas* species, suggesting that Figure 6-5 represents a still

unknown BGC. See Appendices “Biosynthetic capacity of the HMP datasets” and “Putative NRP BGCs in subpravingal plaque datasets” for detailed analysis of the subpravingal plaque datasets.

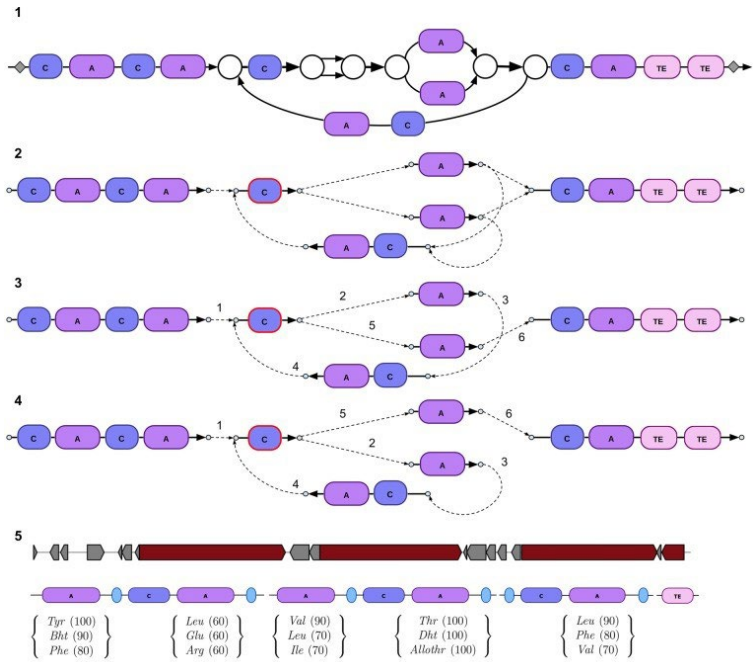


Figure 6-5. The BGC subgraph and the scaffolding graph for the subpravingal plaque metagenome (SRS013723) in the HMP dataset. (1,2) The BGC subgraph and the scaffolding graph, (3,4) Two rural postman routes in the scaffolding graph. The duplicated C-domain is highlighted with red border and is traversed twice in the rural postman routes. The numbers labeling the dashed edges indicate their order in the resulting tour. (5) Since biosyntheticSPAdes and antiSMASH use different thresholds and filtering options antiSMASH identified only 5 (rather than 6) A-domains in the NRP BGC predicted by biosyntheticSPAdes. Three most likely amino acids for each A-domain are shown along with their NRPSpredictor2 (Röttig et al., 2011) scores for the first of two rural postman routes.

6.4 – Discussion

While the human microbiome encodes natural products with great biomedical potential, little is known about these abundant small molecules, despite the fact that the human host is chronically exposed to them (Donia et al., 2014). One of the bottlenecks in discovering natural products from human and other metagenomes is

deriving full-length BGCs from short metagenomics reads (Donia and Fischbach, 2015). This bottleneck negatively affects various genome mining efforts. Indeed, although the discovery of coelichelin (Challis et al., 2005) was one of the first successes of genome mining that was followed by the characterization of many NRPs from sequenced genomes, genome mining in fragmented assemblies remains challenging.

The discovery of the bioactive peptides teixobactin (Wilson et al., 2014) and polytheonamides (Freeman et al., 2012) marks a new era of natural product discovery from uncultivated bacteria. However, while various metagenomes serve as a rich source of natural products (Cragg et al., 2013, Katz et at. 2016), reconstructing complex BGCs from metagenomic assemblies is nearly impossible with short read sequencing technologies. Since gene prediction of BGC scattered between multiple contigs is challenging, the full-length BGC reconstruction is usually difficult without additional biological experiments and extensive manual analysis (Kleigrewe et al., 2015).

biosyntheticSPAdes is a step toward enabling high-throughput natural product discovery by coupling metagenomics and mass spectrometry projects using tools such as NRPquest (Mohimani et al., 2014). It represents the first automated pipeline for BGC reconstruction from genomic and metagenomic sequencing datasets that takes advantage of the assembly graph rather than individual contigs. While we demonstrated that biosyntheticSPAdes is able to recover long BGCs, it can also be extended to other types of long and highly repetitive genes, such as 16S rRNA genes or insecticide toxins (Palma et al., 2014). Although biosyntheticSPAdes currently has the predefined options only for the most important classes of BGCs (NRPS, PKSs, and their fusions), we plan to create presets for other it can be extended for other BGCs with different domain compositions. A user can replace the default HMM-profiles with any profiles of interest, such as TPR-proteins, mucus-binding proteins, etc.

We emphasize that, similarly to all gene prediction tools, a putative BGC predicted by biosyntheticSPAdes may be incorrect and should be used with caution. In

particular, the homology-based mode of biosyntheticSPAdes is most useful when one or more closely related reference genomes are available that have well-annotated BGCs. In the case when multiple feasible paths exist in the assembly graph, we recommend to experimentally verify biosyntheticSPAdes predictions, e.g., using targeted PCR amplification or matching against mass-spectrometry data. Also, peptidogenomics tools (Mohimani et al., 2014) can be applied to all feasible paths in the assembly graph rather than to a single highest-scoring path.

Third generation sequencing technologies have greatly improved isolate bacterial sequencing, thus turning BGC assembly into a relatively simple task. However, they have not yet had a large impact on metagenomic sequencing due to relatively high cost of long-read technologies and difficulties in assembly (no specialized long read metagenomic assembler has been released yet). Since most new natural products are analyzed through metagenomics (or mini-metagenomics) rather than isolate datasets, short reads remain the workhorse of genome mining for natural products.

Some researchers use hybrid approaches for metagenomics assemblies by combining short and long reads (Frank et al., 2016, Tsai et al., 2016). biosyntheticSPAdes is implemented in a manner that allows one to use new sequencing technologies as long as they are supported by the SPAdes pipeline. Since both SPAdes and metaSPAdes support hybrid datasets (Illumina + Pacific Bioscience/Oxford Nanopores), biosyntheticSPAdes can also assemble BGCs in hybrid datasets.

6.5 - Methods

Below we describe the six steps of the biosyntheticSPAdes pipeline (Figure 6-2) and illustrate them using reconstruction of the CALC BGC (Figure 6-1).

6.5.1 – biosyntheticSPAdes pipeline

Step 1: Assembling genomic/metagenomic reads with SPAdes/metaSPAdes

BiosyntheticSPAdes starts with launching SPAdes (Bankevich et al., 2012) or metaSPAdes (Nurk et al., 2017) assemblers. SPAdes and metaSPAdes first construct a *de Bruijn graph* (Compeau et al., 2011) of all reads and subsequently perform various graph simplification procedures (e.g., *bubble collapsing* and *tip removal*) to transform it into an *assembly graph*. Both SPAdes and metaSPAdes use exSPAnder (Prjibelsky et al., 2014) to utilize the read-pair information for repeat resolution and scaffolding in the assembly graph.

Step 2: Identifying domain-edges in the assembly graph

The first step towards reconstructing the nucleotide sequence of a BGC is reconstruction of the arrangement of its biosynthetic domains. In many cases, this arrangement alone provides sufficient information for predicting the structure of the core scaffold of a natural product encoded by the BGC.

To identify edges harboring biosynthetic domains in the assembly graph, contigs generated by SPAdes/metaSPAdes are searched for the domain motifs using HMMER (Zhang et al., 2003, Eddy, 2011). For illustration purposes, here we analyze only A-domains. After mapping contigs back to the assembly graph, biosyntheticSPAdes identifies the positions of all detected domains in the assembly graph (Figure 6-1, top). Mapping the A-domains from the CALC BGC back to the assembly graph reveals that three A-domains (4, 5, and 7) map to the same positions on a single edge of the assembly graph. The edge harboring these positions has

approximately three times higher coverage than the average coverage of edges that contain only a single copy of an A-domain. Supplementary Figure S1 illustrates that these three domains are similar to each other, and share identical repeats of length ≈100 bp and longer. Sequences of these domains are collapsed during assembly, because the assembly graph was constructed from *k*-mers that are shorter than 100 nucleotides.

Step 3: Extracting BGC subgraphs from the assembly graph

BGCs contain various domains and multiple biosynthetic genes in close proximity to each other. Analysis of all complete NRP BGCs from the MIBiG repository of BGCs (Medema et al., 2015) revealed that the distances between consecutive NRPS- or PKS-related domains do not exceed 20 kb in 95% of the cases (Supplementary Figure S2).

Hence, we consider all edges in the assembly graph within 10 kb from the positions of domains on the domain edges identified in the previous step to capture all consecutive domains separated by at most 20 kb. The subgraph of the assembly graph formed by these edges, referred to as the *BGC assembly graph*, usually consists of multiple connected components, where each component, referred to as a *BGC subgraph*, usually corresponds to a single BGC. For example, four NRP BGCs in *S. coelicolor* genome are represented by four different connected components of the BGC assembly graph. However, in some cases a single component of the BGC assembly graph may combine multiple BGCs, particularly when these BGCs share very similar domains with identical sequences exceeding the maximum default *k*-mers size in SPAdes. The *complexity of the BGC subgraph* is defined as the total number of A-domains and AT-domains in this subgraph. We define *non-trivial BGC subgraphs* as BGC subgraphs of complexity at least 3.

The BGC assembly graph for *S. coelicolor* consists of 24 BGC subgraphs. Three of them are non-trivial BGC subgraphs with complexities 9 (for the CALC BGC), 4, and 3. The BGC subgraph corresponding to the CALC BGC with 11 A-domains revealed only 9 A-domains, since three A-domains were collapsed into a single edge.

Step 4: Restoring collapsed domains in the assembly graph

Figure 6-1 reveals a limitation of existing assemblers (*repeat collapsing*) that negatively affects gene prediction tools: three A-domains sharing long identical segments are collapsed into a single edge in the assembly graph. As a result, valuable information about the differences between these A-domains is lost (Supplementary Figure S1). This effect is amplified in metagenomics assemblies since they aggressively collapse bubbles to improve contiguity of the assembly (Nurk et al., 2017), particularly in the case of metagenomes containing similar strains. A side effect of the bubble collapsing procedure is collapsing similar domains, which leads to a high number of mismatches and indels in reconstructed BGC sequences (referred to as an “assembly deterioration”).

This limitation of the existing assemblers can be remedied by restoring subtle variations in the collapsed repeats to enable better repeat resolution. Since SPAdes and metaSPAdes provide map each read to the assembly graph, we consider all reads mapped to edges of all BGC subgraphs and compute the median depth of coverage of each edge. Given an edge with coverage *cov* in a BGC subgraph, we extract all *k*-mers from the reads mapped to this edge. A *k*-mer is defined as *solid* if it does not belong to the edge but appears in at least $\alpha^{*}cov$ reads mapped to this edge (the default value α=0.2). Solid *k*-mers reveal variations in repeats (rather than sequencing errors), as the expected frequency of erroneous *k*-mers is typically below $\alpha^{*}cov$. We define a path formed by solid *k*-mers as a *solid bubble* if it forms an alternative path in a BGC subgraph. We restore all such solid bubbles in a BGC subgraph and rerun the exSPAnder repeat resolution on the modified BGC subgraphs with restored solid bubbles. We emphasize that we applied the domain restoration step to the domain edges in the BGC subgraphs only since applying it to the entire assembly graph leads to deterioration of the assembly and reduced N50 statistics.

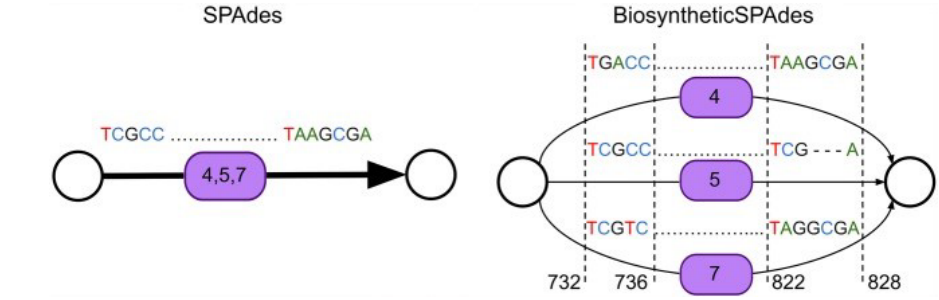


Figure 6-6. Effect of bubble restoration on the reconstruction of the CALC BGC. Schematic representation of repeat collapsing and consensus deterioration in the case of the CALC BGC assembly. While SPAdes outputs a single (and incorrect) consensus sequences of all three collapsed A-domains, these three sequences are not identical. In contrast, biosyntheticSPAdes utilized restored domains and reconstructed their distinct sequences with 100% accuracy (as compared to 99.6% accuracy for SPAdes). Numbers near dashed vertical lines represent the column numbers in the multiple alignment of three A-domain.

Note that the consensus sequence of the edge harboring three similar but not identical A-domains in the CALC assembly (Figure 6-6) differs from the sequences of each of these A-domains. Therefore, it provides slightly inaccurate sequences for each of these three domains. However, after the domain restoration procedure, these three A-domains correspond to three different and 100% accurate consensus sequences. In some cases, the domain restoration procedure even enables exSPAnder to utilize the restored variations between domains for further repeat resolution by utilizing variations between long imperfect repeats.

We note that although the described bubble restoration procedure has a potential to resolve close strains in metagenomics assemblies, it has not been implemented in metaSPAdes yet.

Running exSPAnder on the modified BGC subgraph with restored bubbles often results in a more accurate estimate of the total number of domains (Figure 6-6). In contrast to the initial BGC subgraph with only 9 identified A-domains for the CALC BGC, all 11-A-domains are now captured in 5 resulting contigs in the modified BGC subgraph.

Step 5: Constructing the scaffolding graph

We represent each domain-containing contig as an isolated solid edge in the scaffolding graph (Figure 6-7). Given solid edges e and e', we connect the ending vertex of e with the starting vertex of e' by a dashed edge if the last domain on e and the first domain of e' are close in the BGC assembly graph, i.e., the distance between them is below 10 kb. Given a directed graph with solid and dashed edges, the Rural Postman Problem is to find a rural postman route, i.e., a path visiting all solid edges of the graph (Orloff, 1974).

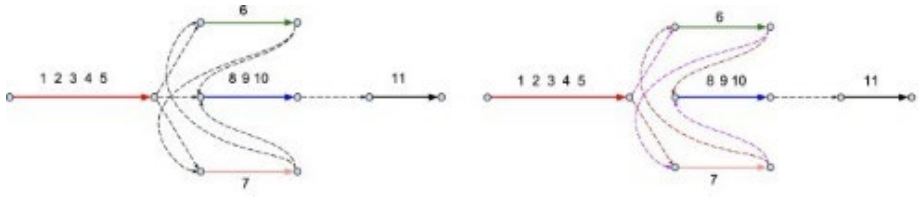


Figure 6-7. The scaffolding graph of the CALC BGC. (Left) Five solid edges in the scaffolding graph correspond to 5 contigs shown in Figure 6-4 (bottom) that contain A-domains. These contigs are shown as a red edge (A-domains 1, 2, 3, 4, and 5), a green edge (A-domain 6), a pink edge (A-domain 7), a blue edge (A-domains 8, 9, and 10), and a black edge (A-domain 11). Eight dashed edges in the scaffolding graph connect solid edges that contain closely located domains in the BGC subgraph. (Right) Two rural postman routes in the CALC scaffolding graph. First tour contains all violet dashed edges and results in the (1, 2, 3, 4, 5, **6, 7**, 8, 9, 10, 11) arrangement of A-domains while the second tour contains all brown dashed edges and results in the (1, 2, 3, 4, 5, **7, 6**, 8, 9, 10, 11) arrangement of A-domains.

Step 6: Constructing putative BGCs by solving the Rural Postman Problem

Inferring the arrangement of domains in an NRP BGC is crucial for identifying the NRP encoded by this NRPS. Since each NRP synthetase corresponds to a rural postman routes in the scaffolding graph, biosyntheticSPAdes searches for all rural postman routes in the scaffolding graph using a brute force algorithm (most scaffolding subgraphs have less than 20 vertices). Figure 6-7 shows two rural postman routes in the CALC scaffolding graph.

Some bacterial genomes contain 100% identical domains that are collapsed into a single edge even after domain restoration. As the result, a rural postman route may

visit the collapsed solid edges in some scaffolding graphs multiple times. For each solid edge in the scaffolding graph, the approximate number of times it should be traversed is defined by the ratio of the coverage of the corresponding domain-edge in the BGC subgraph to the median coverage across all edges of the BGC subgraph.

As Figure 6-7 illustrates, biosyntheticSPAdes may output multiple arrangements of A-domains, each arrangement corresponding to a rural postman route. For each rural postman route, biosyntheticSPAdes reconstructs a path in the BGC assembly graph corresponding to this route and its nucleotide sequence. Dashed edges in a rural postman route may correspond to multiple paths in the BGC assembly graph, and we report the path with the length closest to any of distances from the set of 550, 1500 and 2400 bp, the values of the three pronounced peaks in the distribution of the distances between consecutive domains in known NRPSs (Supplementary Figure S2).

6.5.2 - biosyntheticSPAdes and NRPquest for PNP reconstruction

Even when biosyntheticSPAdes fails to assemble a BGC into a single contig, it typically reduces the number of contigs as compared to SPAdes, e.g., outputs two contigs A and B without providing one of two possible orders to concatenate these contigs (B after A or A after B). This feature is important for natural product researchers since they often perform additional experiments to reconstruct the correct order of contigs (Kleigreve, 2015). For example, in the case of NRP BGC, one can generate all possible concatenates, predict putative NRPs for each concatenate, and match a spectral dataset against all putative NRPs to find a concatenate with the best match. Appendix: “Output format of biosyntheticSPAdes” specifies the details of the biosyntheticSPAdes output. Appendix: “Coupling biosyntheticSPAdes and NRPquest for PNP reconstruction” presents an example of combining genomic and mass spectrometry data to infer the correct arrangement of A-domains.

6.5.3 - Extending biosyntheticSPAdes from NRP BGCs to other BGCs

In addition to the A-domains, biosyntheticSPAdes analyzes other domains in NRP BGCs such as *C–condensation* domains (*C-domains*) and *thioesterase* domains (*TE-domains*), among others. Moreover, biosyntheticSPAdes is not limited to NRP BGCs and also works with BGCs encoding PKS BGCs (Robinson, 1991). PKSs are built from various domains including *acyltransferase* domains (AT-domains), *keto-synthase* domains (KS-domains), *keto-reductase* domains (KR-domains) and *acyl carrier protein domains (ACP-domains).*

6.5.4 - Reference-based BGC ranking algorithm

When a database of reference genomes is available, it can help to predict the correct order of contigs by identifying a genome with a similar BGC. This is especially relevant when assembling genomes that are related to an already sequenced species, or during studies of microbial communities from which individual strains have been isolated and sequenced. biosyntheticSPAdes includes a pipeline that matches all possible orders of multiple putative BGC sequences to gene clusters in antiSMASH-DB (Blin et al., 2016) and ranks them based on how well the order of the matching domains corresponds to the domain order in the most similar BGC.

Note that the reference-based BGC ranking algorithm is an optional module in biosyntheticSPAdes that should be called only in cases when there is more than one plausible path in the assembly graph. In most of our test cases, biosyntheticSPAdes leads to a single plausible path through the assembly graph, and thus a single BGC architecture. In all such cases, reference genomes are not required to infer the correct assembly.

If biosyntheticSPAdes outputs several putative BGCs (pBGCs) for a single BGC gene cluster, it is not clear which of them is correct. In such cases, biosyntheticSPAdes uses a BGC ranking algorithm to compare each putative BGC against all reference BGCs (rBGCs) from a database of all BGCs from the reference

genome sequences, and report the pair of pBGC and rBGC that are most similar to each other.

First, the order and positions of all domains in a pBGCs and all reference rBGCs are predicted with antiSMASH. For each pBGC-rBGC pair, biosynthetiSPAdes constructs a bipartite graph, where nodes are domains and edges connect a domain in pBGC with a domain rBGC if both these domains have the same type, e.g., A-domains. The edge weight is defined as the amino acid sequence similarity for the corresponding domain pair. biosynthetiSPAdes further computes the maximum-weight matching in the constructed bipartite graph using the Hungarian algorithm (Kuhn et al., 1955) (Figure 6-8).

The matching nodes in the maximum-weight matching are referred to as the domain twins.

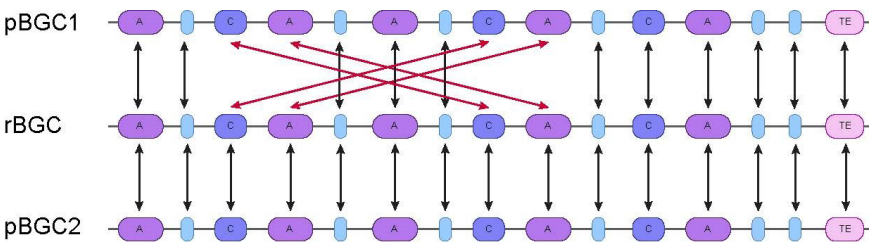


Figure 6-8. Reference-based ranking of two fictional putative BGCs (pBGC1 and pBGC2) according to their similarity to an rBGC in the antiSMASH-DB database. To find which of two pBGC has a better match with the rBGC, the Hungarian algorithm determines domain twins between each pBGCs and the rBGC. Black and red arrows connect twin domains, red arrows further connect twin domains which will lower the score between rBGC and pBCG1 as the domain order in pBGC1 does not match the reference.

The closest rBCG from the database is taken based on the Domain Sequence Similarity (DSS) score described below.

The similarity score between two BGC clusters should consider the sequence similarity, the domain composition, and the ordering of the domains. We also use a concept of highly similar domains – domain twins to find sequence similarity only between relevant domains of BGCs. We find a set of domain twins of a pBGC and rBGC as follows:

1. Construct a bipartite graph G = (U, V, E), where U is the set of nodes that correspond to the domains of the first BGC, V is the set of nodes that correspond to the domains of the second BGC, and E is the set of edges that connecting pairs of domains from U and V of the same type (e.g. A-domains, C-domains, etc.)

2. Compute the similarity score between all pairs of domains of the same type as the amino acid sequence identity of their alignment. The weight of the edge between two domains in the bipartite graph is defined as the similarity score between these domains.

3. Find the maximum weight matching in the bipartite bipartite graph using the Hungarian algorithm (Kuhn, H. W., 1955). Pairs of domains connected by an edge from the maximum weight matching are called the domain twins.

To find a best matching pBGC-rBGC pair, we define the Domain Sequence Similarity (DSS) score. The DSS score is a measure of similarity between the amino acid sequences of twin domains between two BGCs. DSS also penalizes for domains that have no twin or different ordering of twin domains.

Let M be the subset of edges in the maximum weight matching for anrBGC-pBGC Pair, and DT be a set of domain types (e.g. A-domains, C-domains, etc.). Given a BGC, we refer to the number of domains of the specific type in this BGC Ntype(BGC). Given the order of the twin domains in an rBGC (r+, r,, ... , r|/|) and a pBGC (p+, p,, ... , p|/|), we analyze all domain twins (r1, p2) and (r3, p4) and classify a pair as an inversion if k > i and j > l. We define the inversion index $I(rBGC, pBGC)$ as the total number of inversions between an rBGC and a pBGC divided by the $\binom{M}{2}$, the maximum possible number of inversions between two permuatations of length |M|. Given an rBGC-pBGC Pair (rBGC, pBGC) we define its Domain Sequence Similarity score DSS(rBGC, pBGC) as follows:

$$DSS(rBGC, pBGC) = \sum_{type \in DT} \frac{\sum_{e \in M_{type}} weight(e)}{max(N^{type}(rBGC), N^{type}(pBGC))} (1 - I(rBGC, pBGC))$$

where MTUVJ is the subset of edges of the given type in the maximum weight matching and $weight(e)$ is the weight of an edge e in the bipartite graph. Note that the DSS score penalizes domains that do not participate in twin pairs.

Given a set of putative BGCs and a set of reference BGCs, biosyntheticSPAdes selects an rBGC-pBGC Pair with the maximum DSS score and outputs the pBGC from this pair as the most likely solution.

6.5.5 - Ranking putative BGCs from *Streptomyces coelicolor* A3(2) and *Streptomyces avermitilis* MA-4680

biosyntheticSPAdes assembly of the calcium dependent antibiotic (CALC) NRPS in *S. coelicolor* produced two putative BGCs that we refer to as CALC_1 and CALC_2. These two putative BGCs were scored against all BGCs in antiSMASH-DB (excluding CALC itself) to identify which putative BGC is the most similar to known BGCs. To illustrate our approach, we analyzed an rBGC with the highest DSS scores against both CALC_1 and CALC_2: calcium dependent antibiotic BGC from *Streptomyces lividans* TK24. The rBGC chosen using the DSS score belong to the same genus suggesting that the concept of the DSS score helps to identify the correct domain order. Since the MiBIG database contains the CALC BGC from S. Coelicolor A3(2) database, it was possible to also compare the two putative BGCs to their annotated version in MiBIG. Table 6-2 and Figure 6-8 illustrate that CALC_2 has higher domain order consistency and achieves higher DSS score with both the rBGC from antiSMASH-DB and MiBIG making it the best candidate for the biosyntheticSPAdes assembly.

	S. lividans TK24 CALC	S. coelicolor CALC
	DSS	DSS
CALC_1	0.250	0.276
CALC_2	0.253	0.307

Table 6-2. Comparing the DSSs between the two putative BGCs and the two reference BGC from antiSMASH-DB and the reference BGC from MiBIG.

The domain twins generated by the Hungarian algorithm reveal significant differences between the domain structures produced by the rural postman algorithm for the two putative CALC BGCs which affect the order of entire genes within the gene cluster (Figure 6-9).

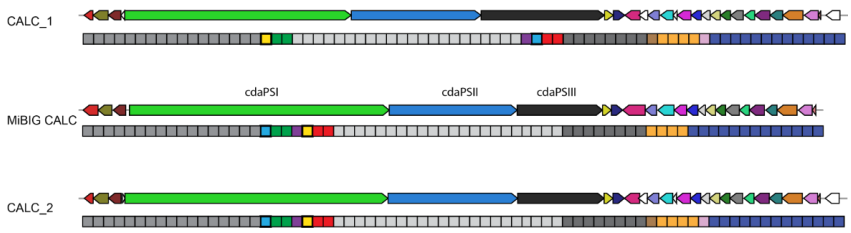


Figure 6-9. The domain orders of CALC_1, CALC_2 and reference CALC from MiBIG. The cdaPSI and cdaPSIII genes from the reference were matched with the green and black labeled genes in CALC_1 and CALC_2. However, the cdaPSI gene in CALC_1 is shorter than the corresponding gene in the reference and in CALC_2, while the cdaPSIII gene (in black) is longer in CALC_1 compared to the reference and CALC_2. These differences are due to an incorrect assembly in CALC_1. This indicates that CALC_2 is the better candidate among the two.

We also assembled the genome of *Streptomyces avermitilis* MA-4680 (Ikegami et al., 2015), which contains a complex repeat-rich gene cluster that produced 6 candidate BGCs from the assembly graph. The ranking algorithm compared the pBGC structures with the filipin BGC, a polyketide synthase BGC, which is present in both antiSMASH-DB and MIBiG (accession: BGC0000059). Table 6-3 illustrates that two out of six candidate BGCs (FILIPIN_2 and FILIPIN_6) produced an identical domain arrangement and the highest-ranking candidate was chosen based on small differences in amino acid sequence.

Putative BGC	Correctly ordered domain twins
FILIPIN_2	102/125
FILIPIN_6	102/125
FILIPIN_3	100/125
FILIPIN_1	100/125
FILIPIN_5	100/125
FILIPIN_4	99/125

Table 6-3. Number of domain twins which had the same order between the putative BGC structure and the reference FILIPIN from antismash-db. The highest-ranking putative structures FILIPIN_2 and FILIPIN_6 have identical domain order. The tie is broken by the DSS score, which indicated that FILIPIN 2 putative BGC had higher sequence similarity to the reference

Figure 6-10 illustrates that the domain architecture for candidates 2 and 6 is more similar to the reference BGC domain architecture compared to lower-ranking pBGCs candidate FILIPIN_5.

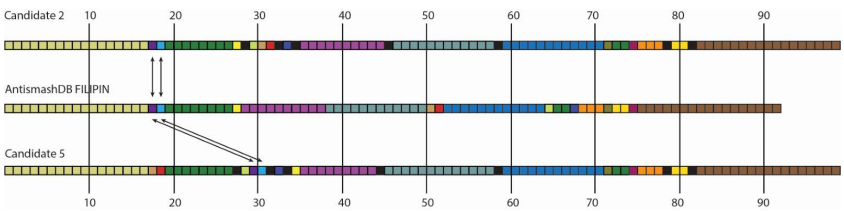


Figure 6-10. The domain orders of two of the FILIPIN putative BGCs and reference FILIPIN from AntismashDB. The domains are color coded to represent blocks with conserved order in the three BGCs even when considering twin domains. The black arrows highlight an example of relocation of two domains for which the reference agrees on the placement for only one of the putative BGCs, notably the highest scoring putative FILIPIN.

As for other reference-based methods, the ranking is affected by database completeness and correctness. Also, the top-ranking pBGC is not necessarily 100% correct, as complex BGCs with high repeat content can result in misassemblies, even with biosyntheticSPAdes. Therefore, results from the ranking algorithm will give insight on which structure better matches the reference BGC but do not guarantee that the highest-ranking structure is also the actual sequence in the assembled genome. In the case of the filipin BGC, even the top-ranking pBGC has small

differences with the reference, indicating that further analysis (e.g., by PCR) would be necessary to confirm the actual structure. We provide this example as a case in point to not blindly trust the results of biosyntheticSPAdes and instead verify them whenever possible.

6.6 - Software Availability

biosyntheticSPAdes will be included in the next version of the SPAdes toolkit available from

http://cab.spbu.ru/software/spades starting from version 3.14.

The pre-release version, that was used for benchmarking in this paper and the biosyntheticSPAdes ranking pipeline, is available in Supplemental material. BiosyntheticSPAdes source code is alternatively available from

http://dx.doi.org/10.6084/m9.figshare.6948260.v1

and the biosyntheticSPAdes ranking pipeline is alternatively available from

https://git.wur.nl/medema-group/biosyntheticSpadesRankingPipeline.

6.7 - Acknowledgements

We are grateful to Alexey Gurevich, Sergey Nurk, Bahar Behsaz, and Jeremy Owen for useful discussions and assistance with data analysis. A.K. was supported by the Russian Science Foundation (grant 19-14-00172). V.T. is supported by the research program NWO-Groen, which is jointly funded by the Netherlands Organization for Scientific Research (NWO), BASF SE and Baseclear BV (project ALWGR.2015.1). M.H.M. is supported by VENI grant 863.15.002 from The Netherlands Organization for Scientific Research (NWO). DM was supported by the Tri-Institutional Training Program in Computational Biology and Medicine (NIH grant 1T32GM083937).

References

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M., Challis, G. L., Thomson, N. R., James, K. D., & Bateman, A. (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3 (2). Nature, 417(6885), 141-147.

Besemer, J., & Borodovsky, M. (2005). GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Research, 33 (suppl. 2), W451-W454.

Blin, K., Medema, M. H., Kottmann, R., Lee, S. Y., & Weber, T. (2016). The antiSMASH database, a comprehensive database of microbial secondary metabolite biosynthetic gene clusters. Nucleic Acids Research, 45, D555-D559.

Blin, K., Medema, M. H., Kazempour, D., Fischbach, M. A., Breitling, R., Takano, E., & Weber, T. (2013). antiSMASH 2.0—a versatile platform for genome mining of secondary metabolite producers. Nucleic Acids Research, 41(W1), W204-W212.

Cane, D. E., & Walsh, C. T. (1999). The parallel and convergent universes of polyketide synthases and nonribosomal peptide synthetases. Chemistry & Biology, 6(12), R319-R325.

Challis, G. L., & Ravel, J. (2000). Coelichelin, a new peptide siderophore encoded by the Streptomyces coelicolor genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. FEMS Microbiology Letters, 187(2), 111-114.

Chu, J., Vila-Farres, X., Inoyama, D., Ternei, M., Cohen, L. J., Gordon, E. A., ... & Jaskowski, M. (2016). Discovery of MRSA active antibiotics using primary sequence from the human microbiome. Nature Chemical Biology, 12(12), 1004.

Coates RC, Podell S, Korobeynikov A, Lapidus A, Pevzner P, Sherman DH, Allen EE, Gerwick L., Gerwick WH. (2014) Characterization of cyanobacterial hydrocarbon composition and distribution of biosynthetic pathways. PLoS One. 9(1):e85140.

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

Cragg, G. M., & Newman, D. J. (2013). Natural products: a continuing source of novel drug leads. Biochimica et Biophysica Acta, 1830(6), 3670-3695.

Delcher, A. L., Bratke, K. A., Powers, E. C., & Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics, 23(6), 673-679.

Donia, M. S., Cimermancic, P., Schulze, C. J., Brown, L. C. W., Martin, J., Mitreva, M., Clardy J.,

Linington R.G., and Fischbach, M. A. (2014). A systematic analysis of biosynthetic gene clusters in the human microbiome reveals a common family of antibiotics. Cell, 158(6), 1402-1414

Donia, M. S., and Fischbach, M. A. (2015). Small molecules from the human microbiota. Science, 349:1254766.

Eddy, S. R. Accelerated profile HMM searches. PLoS computational biology 7.10 (2011): e1002195.

Edwards, D. J., Marquez, B. L., Nogle, L. M., McPhail, K., Goeger, D. E., Roberts, M. A., & Gerwick, W. H. (2004). Structure and biosynthesis of the jamaicamides, new mixed polyketide-peptide neurotoxins from the marine cyanobacterium Lyngbya majuscula. Chemistry & Biology, 11(6), 817-833.

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

Freeman, M. F., Gurgui, C., Helf, M. J., Morinaka, B. I., Uria, A. R., Oldham, N. J., ... & Piel, J. (2012). Metagenome mining reveals polytheonamides as posttranslationally modified ribosomal peptides. Science, 338(6105), 387-390.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. Bioinformatics, 29(8), 1072-1075.

Gurevich, A., Mikheenko, A., Shlemov, A., Korobeynikov, A., Mohimani, H., & Pevzner, P. A. (2018). Increased diversity of peptidic natural products revealed by modification-tolerant database search of mass spectra. Nature microbiology, 3(3), 319.

Hadjithomas, M., Chen, I. M. A., Chu, K., Ratner, A., Palaniappan, K., Szeto, E., ... & Ivanova, N. N. (2015). IMG-ABC: a knowledge base to fuel discovery of biosynthetic gene clusters and novel secondary metabolites. MBio, 6(4), e00932-15.

Hobbs, G. A. I. C., Obanye, A. I., Petty, J., Mason, J. C., Barratt, E., Gardner, D. C., ... & Oliver, S. G. (1992). An integrated approach to studying regulation of production of the antibiotic methylenomycin by Streptomyces coelicolor A3 (2). Journal of Bacteriology, 174(5), 1487-1494.

Huang, W., Li, L., Myers, J. R., & Marth, G. T. (2012). ART: a next-generation sequencing read simulator. Bioinformatics, 28(4), 593-594.

Hyatt, D., Chen, G. L., LoCascio, P. F., Land, M. L., Larimer, F. W., & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics, 11(1), 119.

Ikegami, T., Inatsugi, T., Kojima, I., Umemura, M., Hagiwara, H., Machida, M., Asai, K. (2015 Hybrid De Novo Genome Assembly Using MiSeq and SOLiD Short Read Data PLoS ONE 10(4): e0126289.

Katz, M., Hover, B. M., & Brady, S. F. (2016). Culture-independent discovery of natural products from soil metagenomes. Journal of Industrial Microbiology & Biotechnology, 43(2-3), 129-141.

Kersten, R. D., Yang, Y.-L., Xu, Y., Cimermancic, P. and Nam, S.J., Fenical, W., Fischbach, M.A., Moore, B.S. Dorrestein, P.C. (2011) A mass spectrometry-guided genome mining approach for natural product peptidogenomics. Nat. Chem. Biol. 7, 794–802.

Klassen, J. L., & Currie, C. R. (2012). Gene fragmentation in bacterial draft genomes: extent, consequences and mitigation. BMC genomics, 13(1), 14.

Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., ... & Gerwick, L. (2015). Combining mass spectrometric metabolic profiling with genomic analysis: A powerful approach for discovering natural products from cyanobacteria. Journal of Natural Products, 78(7), 1671-1682.

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.

Liu, C. M., McDaniel, L. E., & Schaffner, C. P. (1972). Fungimicyn biogenesis of its aromatic moiety. The Journal of Antibiotics, 25(3), 187-188.

Magnolo, S. K., Leenutaphong, D. L., DeModena, J. A., Curtis, J. E., Bailey, J. E ., Galazzo, J. L., & Hughes, D. E. (1991). Actinorhodin production by Streptomyces coelicolor and growth of Streptomyces lividans are improved by the expression of a bacterial hemoglobin. Biotechnology, 9(5), 473-476.

Marquez, B. L., Watts, K. S., Yokochi, A., Roberts, M. A., Verdier-Pinard, P., Jimenez, J. I., ... & Gerwick, W. H. (2002). Structure and absolute stereochemistry of hectochlorin, a potent stimulator of actin assembly. Journal of Natural Products, 65(6), 866-871.

Medema, M. H., Blin, K., Cimermancic, P., de Jager, V., Zakrzewski, P., Fischbach, M. A., ... & Breitling, R. (2011). antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. Nucleic Acids Research, 39 (suppl. 2), W339-W346.

Medema, M.H., Cimermancic P., Sali A., Takano E., Fischbach, M.A. (2014) A Systematic Computational Analysis of Biosynthetic Gene Cluster Evolution: Lessons for Engineering Biosynthesis. PLOS Computational Biology 10, e1004016

Medema, M. H., Kottmann, R., Yilmaz, P., Cummings, M., Biggins, J. B., Blin, K., ... & Cruz-Morales, P. (2015). Minimum information about a biosynthetic gene cluster. Nature Chemical Biology, 11(9), 625-631.

Methé, B. A., Nelson, K. E., Pop, M., Creasy, H. H., Giglio, M. G., Huttenhower, C., ... & Chinwalla, A. T. (2012). A framework for human microbiome research. Nature, 486(7402), 215.

Medema MH, Paalvast Y, Nguyen DD, Melnik A, Dorrestein PC, Takano E, et al., (2014) Pep2Path: automated mass spectrometry-guided genome mining of peptidic natural products. PLoS Comput Biol 10: e1003822.

Mohimani, H., Liu, W. T., Kersten, R. D., Moore, B. S., Dorrestein, P. C., & Pevzner, P. A. (2014). NRPquest: coupling mass spectrometry and genome mining for nonribosomal peptide discovery. Journal of Natural Products, 77(8), 1902-1909.

Mohimani, H., Kersten, R. D., Liu, W. T., Wang, M., Purvine, S. O., Wu, S., ... & Pevzner, P. A. (2014). Automated genome mining of ribosomal peptide natural products. ACS Chemical Biology, 9(7), 1545-1551.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L. F., Ninomiya, A., ... & Pevzner, P. A. (2017). Dereplication of peptidic natural products through database search of mass spectra. Nature Chemical Biology, 13(1), 30-37.

Mukherjee, S., Seshadri, R., Varghese, N. J., Eloe-Fadrosh, E. A., Meier-Kolthoff, J. P., Göker, M., ... & Yoshikuni, Y. (2017). 1,003 reference genomes of bacterial and archaeal isolates expand coverage of the tree of life. Nature Biotechnology. 35(7):676-683.

Newman, D.J., Cragg, G.M. Natural Products as Sources of New Drugs from 1981 to 2014 (2016) J. Natural Products, 79, 629–661

Nurk, S., Meleshko, D., Korobeynikov, A., & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. Genome Research, 27(5), 824-834.

Orloff, C. S. (1974). A fundamental problem in vehicle routing. Networks, 4(1), 35-64.

L. Palma, D. Muñoz, C. Berry, J. Murillo, P. Caballero (2014) Bacillus thuringiensis toxins: an overview of their biocidal activity. Toxins, 6, 3296-3325.

Pati, A., Ivanova, N. N., Mikhailova, N., Ovchinnikova, G., Hooper, S. D., Lykidis, A., & Kyrpides, N. C. (2010). GenePRIMP: a gene prediction improvement pipeline for prokaryotic genomes. Nature Methods, 7(6), 455-457.

Paulsen, I. T., Press, C. M., Ravel, J., Kobayashi, D. Y., Myers, G. S., Mavrodi, D. V., ... & Dodson, R. J. (2005). Complete genome sequence of the plant commensal Pseudomonas fluorescens Pf-5. Nature Biotechnology, 23(7), 873-878.

Robinson, J. A. (1991). Polyketide synthase complexes: their structure and function in antibiotic biosynthesis. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 332(1263), 107-114.

Röttig, M., Medema, M. H., Blin, K., Weber, T., Rausch, C., & Kohlbacher, O. (2011). NRPSpredictor2—a web server for predicting NRPS adenylation domain specificity. Nucleic Acids Research, 39 (suppl.2), W362-W367.

Prjibelski, A. D., Vasilinetc, I., Bankevich, A., Gurevich, A., Krivosheeva, T., Nurk, S., ... & Pevzner, P. A. (2014). ExSPAnder: a universal repeat resolver for DNA fragment assembly. Bioinformatics, 30(12), i293-i301.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. Genome Research, 19(6), 1117-1123.

Stachelhaus, T., Mootz, H. D., & Marahiel, M. A. (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. Chemistry & Biology, 6(8), 493-505.

Strieker, M., Tanović, A., & Marahiel, M. A. (2010). Nonribosomal peptide synthetases: structures and dynamics. Current opinion in structural biology, 20(2), 234-240.

Takano, E., Gramajo, H. C., Strauch, E., Andres, N., White, J., & Bibb, M. J. (1992). Transcriptional regulation of the redD transcriptional activator gene accounts for growth-phase-dependent production of the antibiotic undecylprodigiosin in Streptomyces coelicolor A3 (2). Molecular Microbiology, 6(19), 2797-2804.

Tsai, Y.C., Conlan, S., Deming, C., Segre, J.A., Kong, H.H., Korlach, J., Oh, J. and NISC Comparative Sequencing Program, 2016. Resolving the complexity of human skin metagenomes using single-molecule sequencing. MBio, 7(1),.e01948-15.

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C., Knight, R., & Gordon, J. I. (2007). The human microbiome project: exploring the microbial part of ourselves in a changing world. Nature, 449(7164), 804.

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., ... & Breitling, R. (2015). antiSMASH 3.0—a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, 43(W1), W237-W243.

Wilson, M. C., Mori, T., Rückert, C., Uria, A. R., Helf, M. J., Takada, K., ... & Rinke, C. (2014). An environmental bacterial taxon with a large and distinct metabolic repertoire. Nature, 506(7486), 58-62.

Zhang, Z., & Wood, W. I. (2003). A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics, 19(2), 307-308.

Cummings, S. L., Barbé, D., Leao, T. F., Korobeynikov, A., Engene, N., Glukhov, E., ... & Gerwick, L. (2016). A novel uncultured heterotrophic bacterial associate of the cyanobacterium Moorea producens JHB. BMC Microbiology, 16(1), 198.

Kleigrewe, K., Almaliti, J., Tian, I. Y., Kinnel, R. B., Korobeynikov, A., Monroe, E. A., ... & Gerwick, L. (2015). Combining mass spectrometric metabolic profiling with genomic analysis: A powerful approach for discovering natural products from cyanobacteria. Journal of Natural Products, 78(7), 1671-1682.

Kuhn, H. W. (1955). The Hungarian method for the assignment problem. Naval Research Logistics (NRL), 2(1-2), 83-97.

Mohimani, H., Gurevich, A., Mikheenko, A., Garg, N., Nothias, L. F., Ninomiya, A., ... & Pevzner, P. A. (2017). Dereplication of peptidic natural products through database search of mass spectra. Nature Chemical Biology, 13(1), 30-37.

CHAPTER 7

Discussion

7.1 - Influence of high-throughput sequencing

All scientific discoveries in the field of microbiology over the last decades are inevitably tied to the development of sequencing technologies. It is my opinion that the advances of the last 40 years are as important as the wealth of knowledge that was built in the three centuries from Leeuwenhoek's 1677 '*Letter on the Protozoa*' to the development of Sanger sequencing in 1977. The advent of DNA sequencing brought about a radical change in tools and ideas that revolutionized both our understanding of microbial ecology and our approach to studying it.

Beyond the disruptive effect of this technology on the field, the relentless speed at which novel sequencing platforms reduce sequencing costs (while still further improving in quality) promises a future where the only limit is the ability of the researcher to generate and test new hypotheses. No chapter of this thesis could be written or executed without the explosive change that the field has seen. Within the span of the work described in this thesis, we saw the rise and fall of novel sequencing platforms and development of technologies that enabled us to analyze rhizosphere microbial communities at a resolution and precision that was not possible when this project started.

The design of the dilution to extinction experiment, which was made possible by the expertise and understanding developed in our previous work described in chapters 2 and 3), allowed for a simplification of the complex rhizosphere environment in a controlled system where specific parameters could be altered and we could directly observe changes in the phenotype. Still, even in this relatively simplified metagenome, we can observe complex interactions that required complete information on the genetic makeup of the community to draw conclusions on the mechanisms associated with the suppression of the fungal pathogen. However, while sequencing costs continuously decrease over time and yields increase, computational limitations become more and more apparent with even major studies having to compromise on key data analysis steps to deal with such an overwhelming amount of information (Pasolli et al., 2019, Sunagawa et al., 2020). Unsurprisingly,

the most significant hardware limitation is the memory footprint stemming from large assembly graphs that are typical for highly diverse metagenomic samples.

Various assembly tools resort to heuristic solutions to significantly reduce the memory requirements of large assemblies. However, those tradeoffs may result in errors or loss of information, which in turn can affect the final quality of the analysis. There are solutions to the increasing complexity of the microbial communities under study, but these usually come at a cost. For example, it is possible to bypass physical memory shortages with virtual memory, where you can preemptively assign a physical non-RAM disk to serve as additional space should the physical memory run out. The main disadvantage of this approach is the significantly increased running time of the assembly when accessing the non-RAM disk. Alternatively, it is possible to reduce the overall complexity of the sample by filtering out excessively redundant and unique reads before the assembly begins. Filtering reads that do not share similarity to other reads in the dataset will consequently decrease the computational time, but most importantly lessens the memory requirements, as this affects the size of the assembly graph which is directly correlated to the number of unique *k*-mers found in the reads. Removing overly redundant reads originating from sequences that have been already extensively covered will result in a net decrease of the running time especially for assemblers (such as SPAdes) which include a read correction step. This last approach was successfully applied to the complex 10X metagenome described in chapter 3, where reads composed of either unique or overrepresented kmers were filtered before the assembly.

Parallel to increasing sequencing yields, we are also witnessing a race of sequencing technologies to produce increasingly longer reads. Microbiology in general and metagenomics in particular are not deaf to developments that promise improved contiguity information, which is often the principal challenge in microbial ecology studies of complex communities. Accordingly, there is increasing attention in microbiological and metagenomic projects towards the use of long-read technologies. The two main alternatives at this moment in time are represented by Pacific Bioscience single-molecule real time sequencing (SMRT) and Oxford Nanopore sequencing. Long-read technologies are characterized by significantly

increased contiguity, with sequence lengths limited by the extracted DNA fragment sizes, but sport significantly lower accuracy for each individual position compared to Illumina-generated reads. Both traits pose unique challenges in metagenomics contexts. Namely, the higher error rates of long-read technologies are generally corrected by overlapping multiple reads spanning the same region, as these errors are random. However, in metagenomics contexts this problem is exacerbated by the presence of multiple strain variations, with sequence variants indistinguishable from sequencing errors at lower coverages. Additionally, environmental DNA extraction methods can be particularly harsh, all the more so in the rhizosphere or soil samples, which results in greatly reduced DNA fragment size far from the read lengths these technologies can potentially deliver. Finally, hybrid approaches with short and long reads, which are commonplace in complex eukaryotic genome assembly projects, are strongly affected by the DNA extraction methods required to collect environmental DNA, which are known to bias the genetic material towards different organisms (Sui et al., 2020, Wesolowska-Andersen et al., 2014, Vesty et al., 2017) .

Despite some initial efforts that make good use of parallel short- and long-read solutions (Stewart et al., 2019), most non-human-gut-microbiota-centered projects in the near future will still rely on short-length, high-accuracy, high-coverage datasets. To overcome these limitations, large research consortia that have the budget and manpower to devote part of their resources to high-risk initiatives should develop generally applicable protocols and methods to extract material from a diverse range of environmental niches that can produce comparable eDNA samples to be sequenced on multiple platforms. Additionally, extensive studies are needed that compare the effects of the different sequencing platforms on the resulting DNA sequence composition compared to the original DNA material (Xie et al., 2020), which could inform the sequencing platform choice based on previous knowledge on an environmental niche. This is especially relevant in case of proven extensive bias for the combination between expected microbial composition and sequencing method. For example, it may be important to avoid PCR-reliant methods when targeting particularly GC-rich BGCs to avoid underrepresentation in the sequencing data (Browne et al., 2020). For every shotgun metagenomic project, there are many

more culturing-based studies that attempt characterization of bacterial isolates from the environment. However, despite the improvements to large-scale culturomics methods to increase the yield of these processes (Lagier et al., 2018), the gap between the number of metagenome-assembled genomes (MAGs) and that of isolate genomes is bound to grow. Culturomics efforts are useful to answer key questions on many bacterial traits such as their potential to express specific compounds for characterization (Carrion et al., 2019) or verify the ability of bacteria to metabolize specific substrates (Ma et al., 2021). However, culturomics results are likely not representative of the original community composition (Bai et al., 2015), nor of the bacterial metabolic behavior in situ (Amagai et al., 2017). Most studies that associate MAGs with specific phenotypes or functions make use of the guilt-by-association principle. Moreover, MAGs generally represent but a fraction of the assembly which in turn is a fraction of the total sequenced material. To bypass culturability limitations, single-amplified genomes (SAGs) represent a promising alternative strategy that can complement MAGs and isolate studies. In SAG sequencing, individual prokaryotic cells from a specific environment are sorted and lysed before sequencing (Alneberg et al., 2018). I believe that future research on complex microbial environments attempting to elucidate specific phenotypes will benefit greatly from pairing shotgun metagenome and metatranscriptome data from eDNA with exhaustive SAG collections. This way, it would be possible to map reads from the metatranscriptomes onto MAGs and SAGs to identify genes and bacteria associated with the phenomenon under study. In turn, identified MAGs and SAGs can be used to carry out taxonomically targeted isolation efforts that focus on the members of interest of a community, where the information gathered on the predicted metabolic pathways of the uncultured bacteria can guide isolation efforts in the design of the most appropriate isolation medium.

Finally, within this thesis, suppressive soils were dissected with a large variety of sequencing techniques, giving me the opportunity to delve into their nuances. At every step of the process, we reduced the genetic space under study. The second chapter attempted to characterize a very large cohort of complex environments. In such cases, 16S gene-based surveys are commonplace to assess the taxonomical composition and diversity of multiple samples. However, I find the choice of 16S as

reporter gene for these studies to be a relic of the past; the choice seems based on preexisting expertise and excellently detailed and maintained tools for downstream analysis (e.g., qiime (Bolyen et al., 2019)). Multiple studies (Soriano-Lerma et al., 2020, Walker et al., 2015, Tremblay et al., 2015) reported major effects on the predicted community composition based on the primer set used to amplify the target gene, ranging from inability to amplify certain variants of the target sequence to significant changes in amplification efficiencies. This phenomenon derives from the effects of different annealing temperatures for the different version of the oligoprimers and potential mismatches between the primer sequences and the target. Hence, comparison of different studies should be limited to those that share both identical primer sets and amplicon creation protocols. Besides, the variable number of gene copies even between closely related organisms can have a tremendous impact by misrepresenting the community composition, as it influences comparisons of the relative abundances of different taxa. Additionally, the presence of an individual 16S gene within a sample does not provide much information beyond taxonomy, given the diversity of genotypes and functional gene repertoires that can be associated with the same marker. In hindsight, I believe other approaches would have allowed to better compare the large soil collection from chapter 2. For example, shallow sequencing using Oxford Nanopore or Illumina HiSeq would result in a more faithful representation of the taxonomical diversity of the samples even at coverages that do not allow assembly of the raw reads.

To tackle some of these issues, in the third chapter we approach the complexity reduction problem by looking at genes functionally associated with antifungal activity. While this process does not lessen the drawbacks linked to primer bias and comparison between studies, it results in an increased resolution for the target genes that would be hard, if not impossible, to achieve with other methods. Still, non-16S-based marker studies require previous understanding and knowledge of the possible mechanisms associated with a given phenotype. In our case, we relied on the extensive wealth of knowledge associated with disease suppression and the roles of nonribosomal peptides therein to select the marker. Hence, this solution is not applicable in many cases and the focus should still be on nontargeted metagenome approaches when possible.

7.2 - Microbial ecology

Everywhere we look, microorganisms wage a continuous struggle for survival. The main weapon to ensure their success is their genome. The objective of this weapon is to produce an efficient machinery that can transform available nutrients from the outside environment into energy to pass on their genetic information to the next generation. This cycle has been successful for billions of years and has allowed bacteria to colonize virtually every ecological niche on the planet. From underwater hot vents to termite guts to acid drainage, bacteria have shown the ability to strive and adapt to ever-changing conditions. Still, no organism can efficiently utilize all available nutrients in a natural environment. Hence, organisms do not live in a vacuum and have to share their environment with others trying to achieve the same goal: to pass their genetic information to the next generation. The constant competition and collaboration with other organisms represents the most complex factor that drives evolution and population equilibrium within an ecological niche. Along the way, mistakes are made (e.g., genetic mutations) which negatively or positively affect the fitness of the next generation. Given the microbial biological life cycle speed and the significant amount of time that passed since their appearance we can assume that most, if not all, the natural genetic space that results in a phenotypically distinct microorganism metabolically fit for survival in stable environmental niches has been explored at some point. Additionally, beyond sudden changes resulting from evolutionary events, driven by manmade or external factors, most bacterial organisms we can observe today have been genetically stable throughout history.

The appearance of genetically related bacteria in very different environments with very different biotic and abiotic stressors suggests the existence of ‘archetypal’ organisms, which consist of a set of genetic characteristics, morphology and metabolic abilities that distinguishes them from others and allow them to occupy a specific type of niche. They do not necessarily coalesce with an existing or real organism with actual properties or roles within an environment, but they do take shape in the different bacterial species and variants that are observable within and between different ecological niches. The actual microorganisms that can be

observed in the different habitats are but temporary niche-enabled fitness optima selected from the larger pool of available archetypal organisms, which would promptly disappear once the conditions change, but can later be replaced by different species that have similar functional characteristics but may have evolved these independently. This is not an attempt to redefine what is a species based on functional rather than taxonomic data, but a warning against excessive stress put on abundant or enriched bacterial species in many microbial ecology papers, especially when such conclusions are uniquely driven by marker gene studies. After all, many species share significant functional overlap and different combinations of species may fulfill the same 'archetypal' roles within ecosystems; which combination is successful in a specific situation may be largely determined by chance effects.

Any multicellular organism necessarily evolved in the presence of a myriad of unicellular organisms. Eukaryotic organisms that are unable to physically relocate, such as plants, gain a significant evolutionary advantage in forming a mutually beneficial collaboration with the microbiome that inhabits the same environment. As detailed multiple times across the different chapters, plants exert a great influence on the rhizosphere community and selectively recruit microorganisms which carry desirable traits. The ability to select (and be selected as) so-called beneficial commensals carries an evolutionary advantage for both uni- and multi-cellular organisms and is likely to have played a large role in shaping the current viable soil genotypes. As shown in the fourth chapter, preferential selection of microorganisms from the available pool results in a consistent metagenome-wide genetic functional profile. Despite changes of the initial inoculum composition, the population stabilizes towards functionally (rather than taxonomically) fixed profiles.

Ideally, bacteria recruited to the rhizosphere should bring some benefits to the host in the form of expanded specialized metabolism. Rather than being constituted by fast-growing microbes, which devote a larger portion of their energy to explosive growth, rhizosphere communities can be rather diverse and host a plethora of slower-growing but biosynthetically rich taxa, such as actinobacteria. Accordingly, rhizobacteria are known to display overall larger cell size and to grow in higher density clusters compared to the surrounding soil. We can conclude that host-

microbiome interactions in the rhizosphere reduce fitness requirements, enable strain divergence and encompass a diverse functional repertoire.

Since functional profiles in the rhizosphere seem dictated by environmental pressures, rather than following the growing trend of functional prediction of metagenomes from taxonomical gene marker data, I propose a metadata-driven approach to predict metagenome functions based on biome data such as host genotype, soil characteristics and environmentally available microbes. These relationships can also be drawn in the opposite direction, where host genotype can be assumed (predicted) based on the metagenomic composition of its rhizosphere.

With metagenomics, we aim to understand this complex and layered network of interaction by observing the genetic makeup of an ecological niche. This highly reductive approach can be enriched by adding gene expression information. Expression data can be particularly useful in detecting naturally occurring behavior, such as activation of a specific biosynthetic cluster which is otherwise silent when the microbe is grown in isolation. However, our understanding of rhizosphere dynamics is limited and the information from RNAseq data has to be analyzed while dealing with additional unknown factors like relations between phenotype and time, while maintaining data volume and experimental work at reasonable scales. For example, time-series expression data for a dilution to extinction experiment, in which the aim is to track expression changes before, during and after pathogen inoculation, would require an experimental scale which is unattainable in current scientific settings without the formation of large research consortia. Each time point in this possible experiment linearly increases the number of samples needed as the rhizosphere extraction process requires the sacrifice of the sample. Additionally, the time scale and signal duration on which this phenotype operates is completely unknown. Missing the right expression peak generates misleading conclusions for phenotypes which may not have strong ties to expression in the first place. This last point is of particular relevance when trying to identify direct links between secondary metabolite abundance, regardless of the metric used to describe it, and phenotype. Natural products such as antibiotics can have a great effect on their targets even when present in small concentrations, but direct correlations are still regularly

employed when the results are analyzed with conventional enrichment-based approaches.

7.3 - Microbially derived natural products

The diversity of microbial life is staggering. Diversity is primarily evaluated from a taxonomical point of view, as in numbers of different bacteria living in a given ecological niche. However, diversity in bacterial species becomes relevant only in view of the associated diversity in genetic content. Generally speaking, a more taxonomically diverse community is considered to be more genetically rich. However, while it is still possible to observe variations in primary metabolic capabilities of a community, the secondary metabolism is where this uniqueness stands out. In the fourth chapter we observe significant resilience of the community to the dilution treatment. Regardless of the initial inoculum composition, the stability of the functional profiles is remarkable. The selective pressure exerted by the ecological niche results in communities with taxonomically distinct profiles and highly similar genetic composition. However, secondary metabolic capabilities cannot be tracked by measuring the abundance of individual genes within a sample and are tied to the fate of individual strains carrying biosynthetic gene clusters. The best tool currently available to detect and analyze secondary metabolism biosynthetic gene clusters in microbial ecology is antiSMASH (Blin et al., 2021).

Despite antiSMASH's remarkable ability to predict novel gene clusters from metagenomics data, leveraging the formulaic patterns of biosynthetic gene cluster classes, the link between a gene cluster and its product is very faint in absence of concrete previous evidence from single organism studies. Such links are easily drawn for complete metagenome-derived clusters when a perfect match is found with gene clusters in MIBiG (Kautsar et al., 2020), which accounts for only a minority of clusters. For the majority of gene clusters, no exact match exists to known or previously identified BGCs. Tools like BiG-SCAPE (Navarro-Muñoz et al., 2020) calculate distances between BGCs and can be used to identify the closest representative. The BiG-SCAPE-calculated distance is a great starting point for implying relatedness of BGCs, but falls short when it comes to drawing conclusions

about the degree of structural relatedness of their products. After all, small changes in BGC domain order architecture or substrate specificity can result in major differences in the final product. To this end, I recommend the development of an additional natural product prediction tool, that focuses on the key players in BGC product biosynthesis. Such a tool should not be based on overall sequence similarity but on substrate specificity, module order and presence-absence patterns of tailoring enzymes. This would create an abstract representation of a product encoded by the BGC that better reflects the biosynthesis aspects of a BGC rather than the evolutionary relation to other BGCs.

Assembly of metagenomes derived from complex microbial communities can be a challenging task and the quality of the final assembly has a significant influence on the number and completeness of predicted BGCs. This produces a number of different artifacts, which make the assessment of biosynthetic diversity and capabilities hard to correctly estimate. For example, BGC length is highly variable, but it usually surpasses the average metagenomic contig N50 (Dittmann et al., 2015). Therefore, a single BGC will frequently be fragmented across multiple contigs and will thus appear as separate occurrences of a cluster in the antiSMASH output, obscuring the actual number of gene clusters in a sample and inflating the counts for longer BGC classes such as NRPS BGCs. Furthermore, a significant number of sequenced reads in metagenomics assemblies are not assembled, due to lack of coverage or region complexity which cannot be resolved. Opposite to the fragmented clusters problem, this leads to an underestimation of the actual complexity and potential of the community secondary metabolism. There are multiple possible solutions to alleviate this problem. From a data analysis point of view, simple awareness of these downsides can lead to a fairer representation of the sample by manual curation of the data, where highly fragmented clusters are culled from the downstream analysis and the assembly results are critically scrutinized. Software solutions to this problem are also appealing: metagenomics assemblies can be highly complex and, as we saw in chapter 3 and chapter 4, they can produce multiple thousands of putative BGCs, which makes manual curation of individual BGCs hard. To overcome assembly-related biases in metagenomics analysis of secondary metabolism of complex communities, it is vital to develop

assembly-free methods for classification and identification of BGCs. Initially, assembly-free tools to estimate and identify clusters can leverage the existing and growing repositories of gene clusters such as the antiSMASH database by mapping the raw reads against these databases. This process allows for sample diversity estimation regardless of assembly quality. Still, this estimate would initially be biased towards clusters already in already sampled organisms. However, clusters detected in the follow-up assembly of the sample that were not included in the database could be integrated within the growing BGC catalogue, which eventually could become an excellent source of information for estimating biosynthetic potential and novelty. Once integrated in the database, different samples can be compared to each other based on their biosynthetic composition just as we compare samples' taxonomical composition. This is not really possible at the moment with available tools and could change the way we look at biosynthetic characteristics of a community. This could be achieved by clustering BGCs into BGC families (Kautsar et al., 2021) and aggregating the counts for a BGC family, which in turn are used as functional units for calculating the beta diversity between samples and the individual sample richness and evenness.

For most metagenomics natural product studies, the aim is to associate an observed phenotype with one or multiple compounds. Ultimately, this cannot be achieved exclusively through computational methods, which at best can provide evidence supporting one or more hypotheses given the currently limited number of direct links between BGCs and their product. For future studies, priority should go towards creation of open-source isolate collections that cover as much as possible of the natural community. This ensures that hypotheses generated by the analysis of the biosynthetic characteristics of a community can lead to conclusive evidence regarding the causal agents behind a phenotype even for novel natural products.

The field of metagenomics witnessed an explosive growth in the latest years and a large number of studies attempt to describe highly complex phenomena connected to microbial communities from a variety of environments. Like the microorganisms that fascinate us so much, we as scientists must evolve and adapt to tackle the new

questions that arise as we develop new technologies and face new challenges. What is impossible today will be possible tomorrow.

References

Alneberg, J., Karlsson, C.M.G., Divne, A.-M., Bergin, C., Homa, F., Lindh, M.V., Hugerth, L.W., Ettema, T.J.G., Bertilsson, S., Andersson, A.F., Pinhassi, J., 2018. Genomes from uncultivated prokaryotes: a comparison of metagenome-assembled and single-amplified genomes. Microbiome 6, 173.

Amagai, K., Ikeda, H., Hashimoto, J., Kozone, I., Izumikawa, M., Kudo, F., Eguchi, T., Nakamura, T., Osada, H., Takahashi, S., Shin-ya, K., 2017. Identification of a gene cluster for telomestatin biosynthesis and heterologous expression using a specific promoter in a clean host. Sci Rep 7, 3382.

Bai, Y., Müller, D.B., Srinivas, G., Garrido-Oter, R., Potthoff, E., Rott, M., Dombrowski, N., Münch, P.C., Spaepen, S., Remus-Emsermann, M., Hüttel, B., McHardy, A.C., Vorholt, J.A., Schulze-Lefert, P., 2015. Functional overlap of the Arabidopsis leaf and root microbiota. Nature 528, 364–369.

Blin, K., Shaw, S., Kloosterman, A.M., Charlop-Powers, Z., van Wezel, G.P., Medema, M.H., Weber, T., 2021. antiSMASH 6.0: improving cluster detection and comparison capabilities. Nucleic Acids Research 49, W29–W35.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., Alexander, H., Alm, E.J., Arumugam, M., Asnicar, F., Bai, Y., Bisanz, J.E., Bittinger, K., Brejnrod, A., Brislawn, C.J., Brown, C.T., Callahan, B.J., Caraballo-Rodríguez, A.M., Chase, J., Cope, E.K., Da Silva, R., Diener, C., Dorrestein, P.C., Douglas, G.M., Durall, D.M., Duvallet, C., Edwardson, C.F., Ernst, M., Estaki, M., Fouquier, J., Gauglitz, J.M., Gibbons, S.M., Gibson, D.L., Gonzalez, A., Gorlick, K., Guo, J., Hillmann, B., Holmes, S., Holste, H., Huttenhower, C., Huttley, G.A., Janssen, S., Jarmusch, A.K., Jiang, L., Kaehler, B.D., Kang, K.B., Keefe, C.R., Keim, P., Kelley, S.T., Knights, D., Koester, I., Kosciolek, T., Kreps, J., Langille, M.G.I., Lee, J., Ley, R., Liu, Y.-X., Loftfield, E., Lozupone, C., Maher, M., Marotz, C., Martin, B.D., McDonald, D., McIver, L.J., Melnik, A.V., Metcalf, J.L., Morgan, S.C., Morton, J.T., Naimey, A.T., Navas-Molina, J.A., Nothias, L.F., Orchanian, S.B., Pearson, T., Peoples, S.L., Petras, D., Preuss, M.L., Pruesse, E., Rasmussen, L.B., Rivers, A., Robeson, M.S., Rosenthal, P., Segata, N., Shaffer, M., Shiffer, A., Sinha, R., Song, S.J., Spear, J.R., Swafford, A.D., Thompson, L.R., Torres, P.J., Trinh, P., Tripathi, A., Turnbaugh, P.J., Ul-Hasan, S., van der Hooft, J.J.J., Vargas, F., Vázquez-Baeza, Y., Vogtmann, E., von Hippel, M., Walters, W., Wan, Y., Wang, M., Warren, J., Weber, K.C., Williamson, C.H.D., Willis, A.D., Xu, Z.Z., Zaneveld, J.R., Zhang, Y., Zhu, Q., Knight, R., Caporaso, J.G., 2019. Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. Nat Biotechnol 37, 852–857.

Carrión, V.J., Perez-Jaramillo, J., Cordovez, V., Tracanna, V., de Hollander, M., Ruiz-Buck, D., Mendes, L.W., van Ijcken, W.F.J., Gomez-Exposito, R., Elsayed, S.S., Mohanraju, P., Arifah, A., van der Oost, J., Paulson, J.N., Mendes, R., van Wezel, G.P., Medema, M.H., Raaijmakers, J.M., 2019. Pathogen-induced activation of disease-suppressive functions in the endophytic root microbiome. Science 366, 606–612.

Dittmann, E., Gugger, M., Sivonen, K., Fewer, D.P., 2015. Natural Product Biosynthetic Diversity and Comparative Genomics of the Cyanobacteria. Trends in Microbiology 23, 642–652.

Kautsar, S.A., Blin, K., Shaw, S., Navarro-Muñoz, J.C., Terlouw, B.R., van der Hooft, J.J.J., van Santen, J.A., Tracanna, V., Suarez Duran, H.G., Pascal Andreu, V., Selem-Mojica, N., Alanjary, M., Robinson, S.L., Lund, G., Epstein, S.C., Sisto, A.C., Charkoudian, L.K., Collemare, J., Linington, R.G., Weber, T., Medema, M.H., 2020. MIBiG 2.0: a repository for biosynthetic gene clusters of known function. Nucleic Acids Research 48, D454–D458.

Kautsar, S.A., van der Hooft, J.J.J., de Ridder, D., Medema, M.H., 2021. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. GigaScience 10, giaa154.

Lagier, J.-C., Dubourg, G., Million, M., Cadoret, F., Bilen, M., Fenollar, F., Levasseur, A., Rolain, J.-M., Fournier, P.-E., Raoult, D., 2018. Culturing the human microbiota and culturomics. Nat Rev Microbiol 16, 540–550

Ma, M., Zheng, L., Yin, X., Gao, W., Han, B., Li, Q., Zhu, A., Chen, H., Yang, H., 2021. Reconstruction and evaluation of oil-degrading consortia isolated from sediments of hydrothermal vents in the South Mid-Atlantic Ridge. Sci Rep 11, 1456.

Pasolli, E., Asnicar, F., Manara, S., Zolfo, M., Karcher, N., Armanini, F., Beghini, F., Manghi, P., Tett, A., Ghensi, P., Collado, M.C., Rice, B.L., DuLong, C., Morgan, X.C., Golden, C.D., Quince, C., Huttenhower, C., Segata, N., 2019. Extensive Unexplored Human Microbiome Diversity Revealed by Over 150,000 Genomes from Metagenomes Spanning Age, Geography, and Lifestyle. Cell 176, 649-662.e20.

Stewart, R.D., Auffret, M.D., Warr, A., Walker, A.W., Roehe, R., Watson, M., 2019. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. Nat Biotechnol 37, 953–961.

Sui, H., Weil, A.A., Nuwagira, E., Qadri, F., Ryan, E.T., Mezzari, M.P., Phipatanakul, W., Lai, P.S., 2020. Impact of DNA Extraction Method on Variation in Human and Built Environment Microbial Community and Functional Profiles Assessed by Shotgun Metagenomics Sequencing. Frontiers in Microbiology 11, 953.

Sunagawa, S., Acinas, S.G., Bork, P., Bowler, C., Eveillard, D., Gorsky, G., Guidi, L., Iudicone, D., Karsenti, E., Lombard, F., Ogata, H., Pesant, S., Sullivan, M.B., Wincker, P., de Vargas, C., 2020. Tara Oceans: towards global ocean ecosystems biology. Nat Rev Microbiol 18, 428–445.

Vesty, A., Biswas, K., Taylor, M.W., Gear, K., Douglas, R.G., 2017. Evaluating the Impact of DNA Extraction Method on the Representation of Human Oral Bacterial and Fungal Communities. PLOS ONE 12, e0169877.

Wesolowska-Andersen, A., Bahl, M.I., Carvalho, V., Kristiansen, K., Sicheritz-Pontén, T., Gupta, R., Licht, T.R., 2014. Choice of bacterial DNA extraction method from fecal material influences community structure as evaluated by metagenomic analysis. Microbiome 2, 19.

Xie, H., Yang, C., Sun, Y., Igarashi, Y., Jin, T., Luo, F., 2020. PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning. Frontiers in Genetics 11, 1077.

Browne, P.D., Nielsen, T.K., Kot, W., Aggerholm, A., Gilbert, M.T.P., Puetz, L., Rasmussen, M., Zervas, A., Hansen, L.H., 2020. GC bias affects genomic and metagenomic reconstructions, underrepresenting GC-poor organisms. GigaScience 9, giaa008.

Soriano-Lerma, A., Pérez-Carrasco, V., Sánchez-Marañón, M., Ortiz-González, M., Sánchez-Martín, V., Gijón, J., Navarro-Mari, J.M., García-Salcedo, J.A., Soriano, M., 2020. Influence of 16S rRNA target region on the outcome of microbiome studies in soil and saliva samples. Sci Rep 10, 13637.

Tremblay, J., Singh, K., Fern, A., Kirton, E., He, S., Woyke, T., Lee, J., Chen, F., Dangl, J., Tringe, S., 2015. Primer and platform effects on 16S rRNA tag sequencing. Frontiers in Microbiology 6, 771.

Walker, A.W., Martin, J.C., Scott, P., Parkhill, J., Flint, H.J., Scott, K.P., 2015. 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. Microbiome 3, 26.

Summary

Plants and microbes have coexisted for hundreds of millions of years and have developed deeply intertwined mutually beneficial relations. Among the many benefits of a selected microbial community, pathogen suppression is a particularly desirable trait for both the plant and agronomy industry. Suppressive soils have been described for many years but technology to develop a deep understanding of this phenomenon was only recently introduced. In this thesis, I applied different metagenomic sequencing approaches to study the biological basis of suppressive soils with particular interest toward the biosynthetic potential of the suppressive rhizosphere community. In the introduction, I describe ecology-inspired solutions for biosynthetic gene cluster (BGC) mining and the existing tools and sequencing technologies that can be used to this end.

We take a structured approach to the dissection of the suppressive-associated rhizosphere communities. In the first part of the work, we perform the first large-scale soil survey aiming at establishing a soil collection to identify unique characteristics of suppressive soils. Through a combination of phenotyping and marker gene sequencing, we identify four soils with strong *Fusarium culmorum*-suppressive characteristics. We compare taxonomy composition and volatile profile of both suppressive and non-suppressive soils to identify features that distinguish suppressive soils. The suppressive soils found in this collection do not share physiochemical or categorical characteristics. In addition, diversity and community structure do not separate suppressive and non-suppressive soils. However, network-based analysis shows a group of acidobacteria which co-occur in a suppressive-soil-associated hub.

Then, to better understand the secondary metabolite diversity of the suppressive samples we characterize of nonribosomal peptide synthetase diversity in suppressive and conducive soils with the use of functional amplicon sequencing of NRPS adenylation domains. To this end, we developed dom2BGC, a pipeline for the annotation of domains associated with BGCs. We also perform cooccurrence-based clustering of the sequenced domains to restore, through guilt by association, the

physical clustering of the different domains annotated to the same (BGC). We identified multiple NRPSes with potentially antifungal activity that occur exclusively in suppressive soils. Furthermore, we sequenced one suppressive sample with 10X metagenomic sequencing, which was used to confirm the presence of dom2BGC reconstructed clusters.

After extensive study of suppressive rhizosphere communities, we zoom in and perform a dilution to extinction experiment with a microbial extract from a suppressive soil which progressively loses its phenotype in accordance with the dilution. We evaluate the effect of dilution on the microbial composition and functional profile of the community. Genetic characteristics and taxonomic groups that correlate with the dilution and phenotype loss are investigated for links that can shed light on the key players of suppressive soils. We found multiple metagenome-assembled genomes rich in BGCs and chitin-degrading ability that closely correlate with the loss in pathogen suppression.

This work then proceeds to describe the characteristics of a suppressive endosphere. We compare *Rhizoctonia solani* suppressive and conducive endosphere communities of sugar beet. Shotgun metagenome sequencing showed significant differences in taxonomic abundance and genes associated with *Chitinophaga* and *Flavobacterium* bacteria. Additionally, we compose a synthetic community from endosphere isolates that provides disease suppression against *Rhizoctoria solani* infection. Finally, disease suppression of the synthetic community is lost upon site-directed mutagenesis of a candidate suppressive NRPS in a flavobacterial isolate, providing a model to explain the phenotype.

Finally, we detail the development of an assembly tool that aims to improve the assembly of complex BGCs. BGCs often contain repetitive domains that are hard to assemble, but are still very informative as they strongly influence the predicted natural product. Such repetitive domains are sometimes inadvertedly collapsed during the assembly graph formation, which inevitably leads to an erroneous or incomplete cluster. These problems are exacerbated in complex metagenomics assemblies. BiosyntheticSPAdes is designed to identify and isolate BGC-harboring

neighborhoods in the assembly graph by finding multiple adjacent BGC-associated domains. Once identified, the BGC subgraph is extracted and collapsed domains are restored based on local coverage. Depending on the subgraph, multiple paths can be traversed to produce a BGC sequence. When multiple putative BGCs are produced, a ranking pipeline shows which candidate BGC is structurally similar to previously assembled BGCs based on sequence similarity and domain order.

To conclude, in the discussion I offer some considerations on the effects of sequencing technologies on microbiology and microbial ecology, to then propose experimental and computational strategies that are best fit to identify microbial natural products from complex ecosystems.

Acknowledgements

“E quindi uscimmo a riveder le stelle” is the last sentence of Dante’s Inferno which can be translated as “Thence we came forth to rebehold the stars” (Longfellow 1867). It details the moment in which the writer and his guide, Virgil, exit hell and are finally able to see the stars once more. However, the stars that welcome the duo are not the ones Dante is used to but the Southern hemisphere. Likewise, it is now my moment to finalize my doctorate and see new stars.

I find many similarities between Dante’s journey, the challenges and stepping stones in a doctorate degree. While the poet references important people in his life, I would like to use a similar stratagem and thank the people who made this journey possible by assigning them to characters in the books. Don’t worry if you find yourself in Hell, even Dante placed people he appreciated there.

One of the main factors in a doctorate degree success is guidance. I could not have asked for a better Virgil than Marnix. Described as maestro, prophet and father figure, it is easy to see how these titles apply to you. Your guidance and support were second to none, you are the scientist I wish one day to become. You lead by example by practicing what you preach. You expect the most from yourself and your heart is full of empathy. Thankfully, unlike Dante, I don’t have to visit hell, but Bennekom, to see you. Which is not as bad.

My favorite character from the Comedy, is Ulysses. Found in hell, it represents the restless and never-ending search for knowledge. Just as Dante meets this figure early in its travels, I met Dick before the beginning of my doctorate. From the very beginning, I was impressed by your ability to immediately understand problems and identify solutions. Challenges took me a week to overcome were solved in a matter of seconds during our meeting. Coming from a highly hierarchical educational system, I was not expecting to connect so easily with a supervisor. Your humor and friendly demeanor made it possible to me to realize that a different style of management is not only possible but successful.

When it comes to my colleagues, I have an endless list of people that I must thank. Being a “people person” by ability to thrive in the social environment is just as important as the supervision. In these years I made life-long friendship bonds with many of you. Instead of finding individual characters from the books for each of you, I will refer to you as the “stilnovisti” colleagues of Dante himself when he was writing his book. I leave it to you to decide who is Petrarca, Boccaccio, Cavalcanti, Frescobaldi or Cino de' Sigilbuldi da Pistoia. Thanks to Raul, you were destined to be the organizational center of all our activities, you hold the sky so that it does not fall on our heads. You were my Dedalus, continuously sending me, your Icarus, flying during our Tuesday football matches. Thanks Carlos, you are Ulysses’s second flame, Diomedes, embarking on ambitious projects with the recklessness of those that don’t put a limit to their potential. You inspire me with your drive and dedication. Talking about audacity, Satria, you are my Garibaldi, hero of two worlds. Fighting for your place in the pantheon of science. Your ability to get up every time you are struck, just like Garibaldi, puts into perspective my struggles. I have nothing but admiration for you. Miguel, you are a Cyrano that conquered his Rossana. You always speak your mind and are not afraid to challenge power, whatever it takes. I would like to believe that you transmitted some of this characteristic to me. Wherever you are, know that I will always be here for you to mock 16S papers.

Now, it is time I address the elephant in the room. Thank you, Mehmet and Janani, I do not have words to express my gratitude to you. I will instead praise your incredible bond that makes you more than the sum of its parts. You are Paolo e Francesca, the lovers of Dante’s inferno. Both of you are incredibly talented people that share a unique link that flourished in a stunning union. Your future is bright and I am looking forward to seeing its fruits.

During my journey, I was never alone. I shared every single step with a truly remarkable companion. Adam, you were Dante’s moral compass. Always directing our efforts toward the finish line. I will not forget the dystopian apparition of an ostrich just as it started snowing in some forgotten potato field in east Germany. Thank you for being my rock in this time.

Mamma e Papà, il mio tragitto verso questo traguardo parte da molto lontano. Voi siete l'arco ed io la freccia. Devo la traiettoria della mia vita a voi. Ogni volta che mi avete incoraggiato a seguire la mia curiosità, ogni volta che avete dedicato tempo ad ascoltare le mie infinite storie su dinosauri, coltivato il mio interesse per la storia, avete contribuito a questo successo. Spero che questo risultato sia un motivo di orgoglio tanto mio quanto vostro. Grazie.

Grazie Chiara, senza di te non avrei potuto imbarcarmi in questa avventura. La tua presenza a casa mi porta serenità perche so che c'è *sempre qualcuno su cui posso contare quando necessario. Tendiamo a dimenticarlo, ma noi siamo molto simili. Entrambi abbiamo voglia di migliorarci, ambizione e una* tendenza alla socialità*. I nostri tragitti nella vita sono diversi ma le caratteristiche che condividiamo sono un testamento del nostro legame.*

"Io son Beatrice, che ti faccio andare
Vegno di loco, ove tornar disio
Amor mi mosse, che mi fa parlare."
Inferno. Canto II, verso 70-72.

My acknowledgements cannot but end with the reason I made it so far. Mela, you are my Beatrice, Dante's inspiration, muse and guide. In the Comedy, Virgilio has to leave when Dante reaches heaven and Beatrice becomes the poet guide. Consequently, this means that the Netherlands are hell and heaven is in Germany. Life is full of surprises. We speak so many languages and yet, words fail to describe our bond. You are the best companion I could have asked for. We complement each other in all things. Where I lack, you excel. During my doctorate, we were separated by physical distance but you never felt absent. I knew I could count on your support whenever I needed it. Now that this step of my life is completed, I am not afraid of what comes next because I know you will be by my side. Thanks for being the single best thing in my life.

This research described in this thesis was financially supported by the Graduate School Experimental Plant Sciences (EPS), Wageningen University & Research

Cover design by Maria Vittoria Sorgi (mvsorgi@gmail.com)

Printed by Digiforce - ProefschriftMaken