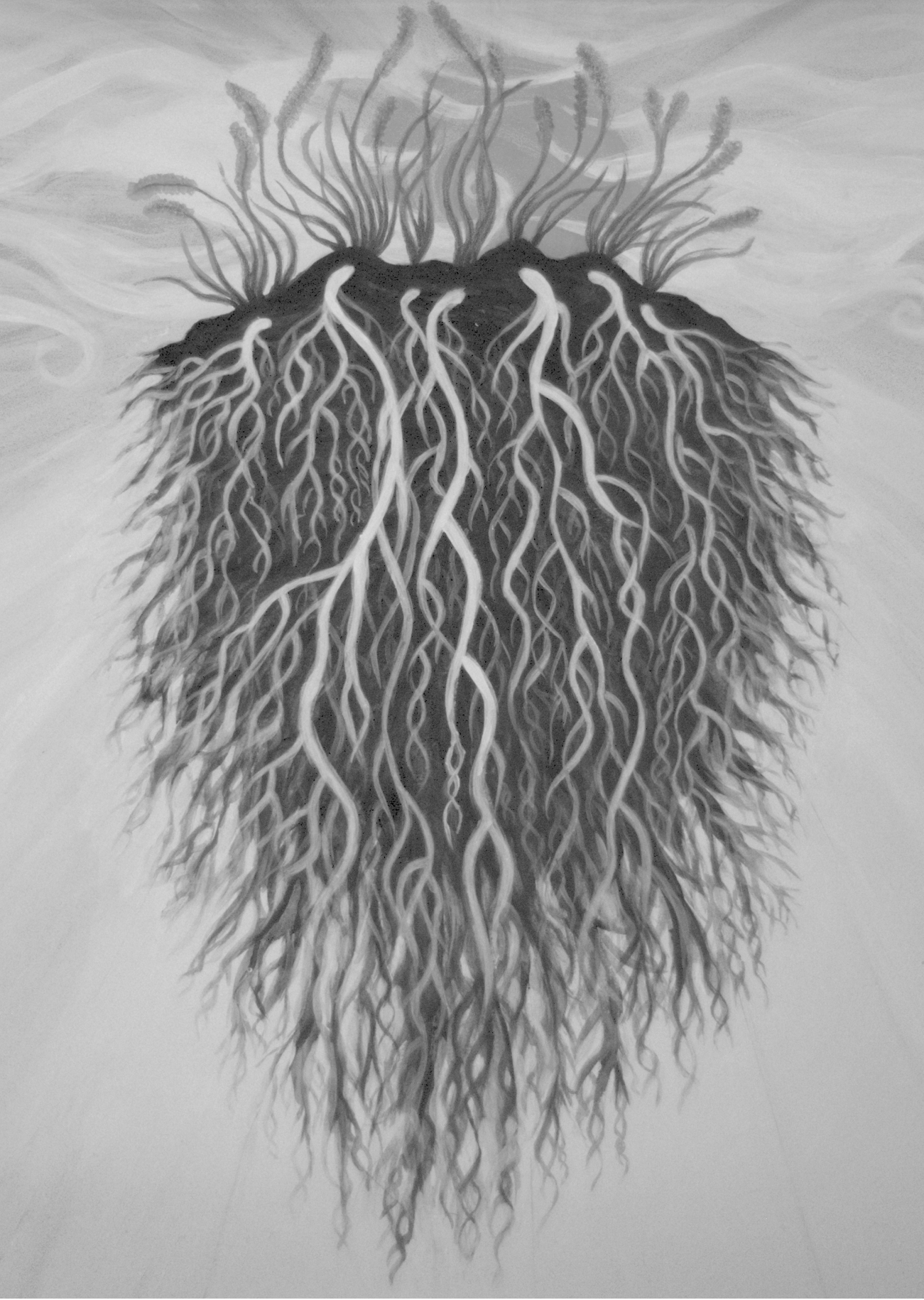
Understanding disease suppressive soils: molecular and chemical identification of microorganisms and mechanisms involved in soil suppressiveness to Fusarium culmorum of wheat

Ossowicki, A.S.



Chapter 3

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

Vittorio Tracanna[1]*, Adam Ossowicki[2]*, Marloes L. C. Petrus[3], Sam Overduin[1], Barbara R. Terlouw[1], George Lund[4], Serina L. Robinson[5], Sven Warris[6], Elio G. W. M. Schijlen[6], Gilles P. van Wezel[2,3], Jos M. Raaijmakers[2,3], Paolina Garbeva[2], Marnix H. Medema[1]

1. Bioinformatics Group, Wageningen University and Research, Wageningen, The Netherlands
2. Microbial Ecology, Netherlands Institute of Ecology (NIOO-KNAW), Wageningen, The Netherlands
3. Microbial Biotechnology, Leiden Institute of Biology, Leiden, The Netherlands
4. Biointeractions and Crop Protection, Rothamsted Research, Harpenden, United Kingdom
5. BioTechnology Institute, University of Minnesota-Twin Cities, MN, USA
6. Bioscience, Wageningen University and Research, Wageningen, The Netherlands
*: contributed equally to this work

A revised version was accepted for publication in mSystems

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.

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.

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 ("Food and Agriculture Organisation of the United Nations," 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-Bogantes 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 little (5). 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, 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 and Mazzola, 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 (2020). 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.

Materials and Methods

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 deep 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).

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. After three weeks, wheat plants were

inspected for disease symptoms together with rhizosphere DNA isolation. NRPS adenylation domains were amplified using 3F and A7R primers (Ayuso-Sacido and Genilloud, 2005).

A-domain amplicon preparation

Barcoding and sequencing of the A-domain amplicons was performed at BaseClear (Leiden, the Netherlands) using Illumina MiSeq, which generated 843,536 paired-end reads of 250 bp in length per read. Sequences were de-multiplexed and adapters trimmed using Qiime2 (Bolyen et al., 2019, p. 2). Quality filtering and denoising was done with Dada2 (Benjamin J 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 and Eddy, 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.

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.

Feature extraction from amplicons for substrate specificity prediction

These methods, including substrate specificity prediction model training and tuning, are described in detail in the supplementary methods.

Dom2BGC pipeline - Generation of *in silico* amplicons

To generate a reference dataset of NRPS functional amplicons, A-domain sequences were extracted from BGCs from a variety of reference databases. Hmmsearch with default parameters was used to detect the A-domains, and the domain regions targeted by the oligoprimers were extracted from the matched sequences using HMM profile match coordinates. This process creates *in silico* amplicons which 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)).

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. 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.

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 99^{th} 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.

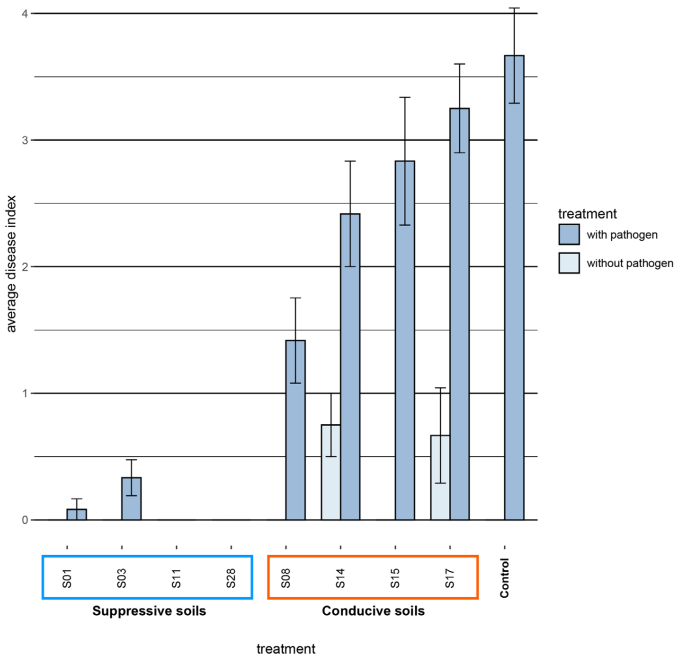


Figure 1. Disease index of Fusarium root rot disease of wheat grown in eight different agricultural soils. Four soils (S01, S03, S11, S28) were classified as disease suppressive and four soils (S08, S14, S15, S17) as disease conducive. Dark blue - inoculated with F. culmorum, light blue – non-inoculated, 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).

Results and Discussion

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, Fig.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 (Fig. 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.

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 (Martínez-Núñez and López, 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 Fig.S1) 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 (Ossowicki et al., 2020). 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. 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 (Fig. 2A).

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 (Fig. 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.

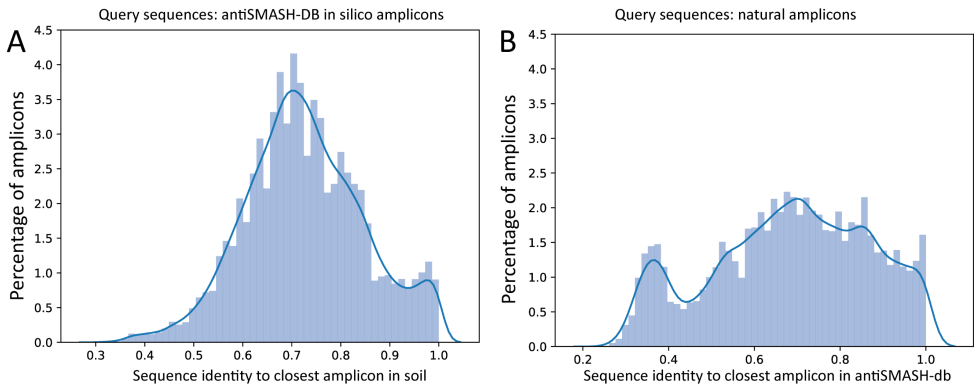


Figure 2. Sequence distance between nAMPs and antismash-DB in silico amplicons. A) Histogram showing the distribution of best matches (as in highest percentage identity at protein level) between each nAMP and the antiSMASH-DB in silico amplicon database. B) Histogram showing the distribution of best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

The dom2BGC pipeline facilitates automated annotation and networking of functional amplicons

Current tools for the annotation of functional amplicons (eSNaPD (Reddy et al., 2014), NaPDoS ("The Natural Product Domain Seeker NaPDoS: A Phylogeny Based Bioinformatic Tool to Classify Secondary Metabolite Gene Diversity," n.d.)) 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 (Fig. 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: Fig.2, Fig.5 - network). 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.

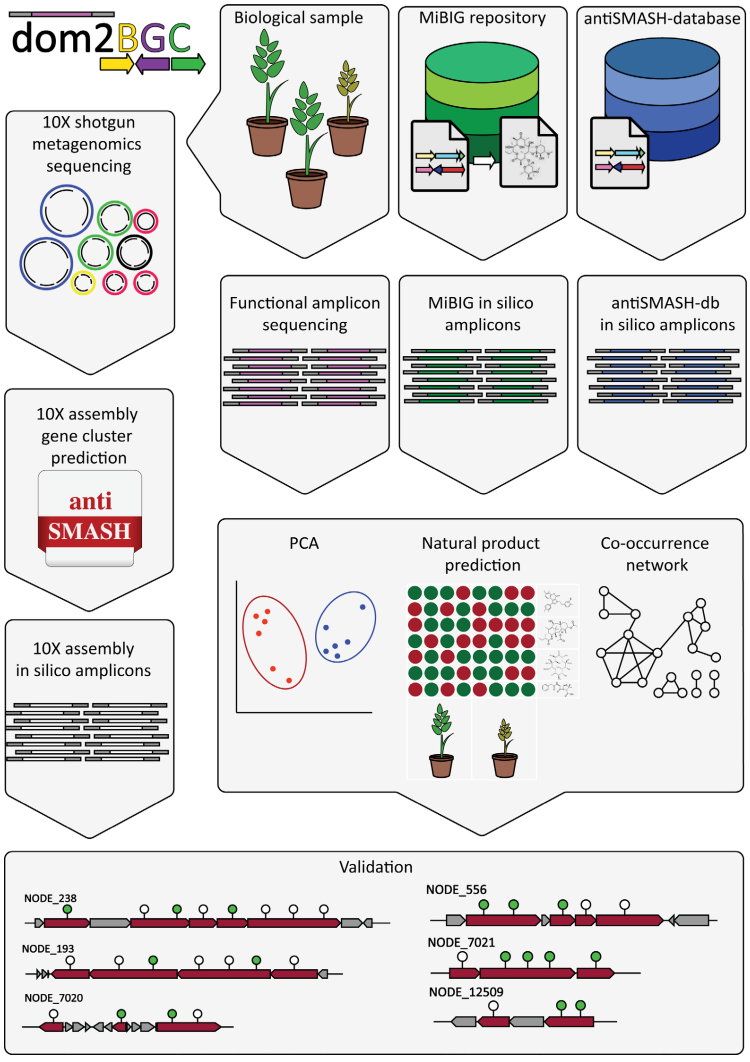


Figure 3. The dom2BGC annotation pipeline and validation process. Amplified sequences from the rhizosphere are translated to nAMPs as per methods section and have been annotated through comparison with in silico amplicons from MiBIG and antiSMASH databases. Richness and community composition measures 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.

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.

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 (5), 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 (Fig.4A).

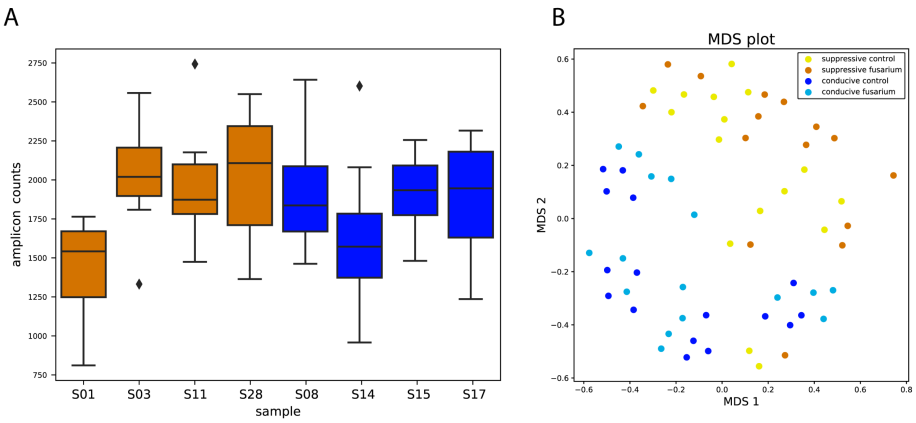


Figure 4. Community diversity and composition. A) Adenylation domain richness across suppressive (orange bars) and conducive (blue bars) soils, calculated as unique sequences. B) Visualization of the adenylation domain community composition with multidimensional scaling.

Several studies have associated overall microbial species richness or evenness in the soil and rhizosphere with disease suppressiveness (Chaparro et al., 2012; Garbeva et al., 2004; Janvier et al., 2007; 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 (Carrión et al., 2019; Mendes et al., 2011). 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 (Fig.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 (Carrión et al., 2019; Duijff et al., 1999, 1994; Hayden et al., 2018; Kinkel et al., 2012; Michelsen et al., 2015; Raaijmakers and Weller, 1998; Scher M., 1982; Weller et al., 2002a; Zhao et al., 2018) 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.

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 (Cawoy et al., 2015; Geudens and Martins, 2018; Kruijt et al., 2009; Mm et al., 2014; Omoboye et al., 2019b, 2019a, p. 3; Oni et al., n.d.; Raaijmakers et al., 2010; Raaijmakers and Mazzola, 2012).

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, creating 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 5 and 6.

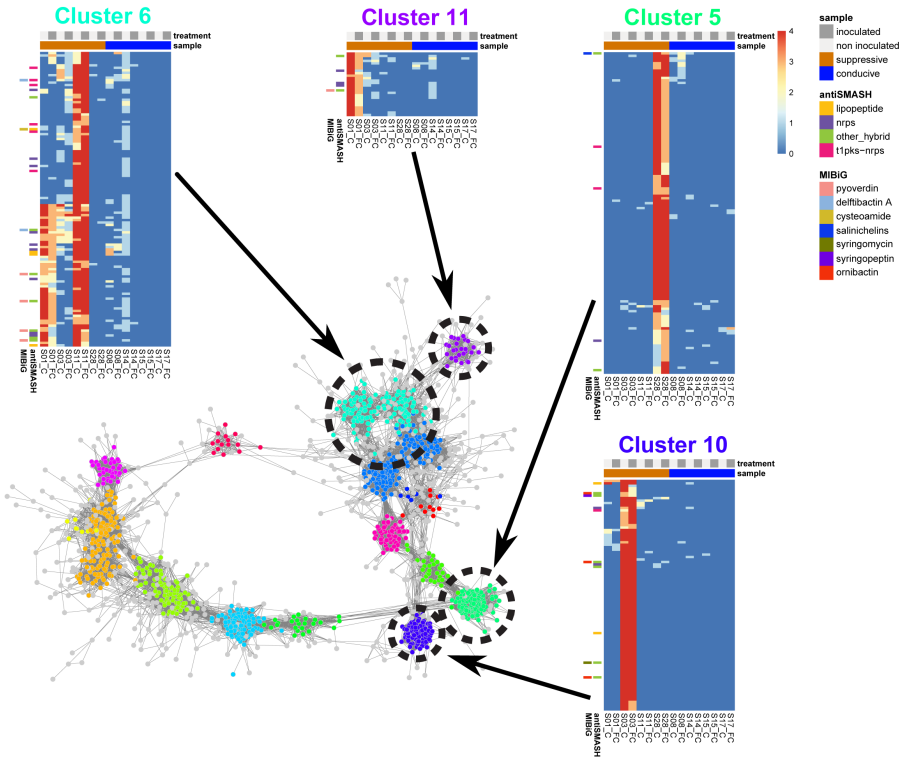


Figure 5. Domain co-occurrence network showing clusters associated with soil suppressiveness. For each of the four clusters (5, 6, 10 and 11), a heatmap shows distribution of A-domains across the samples. The heatmap colour scale represents the number of replicates in which the A-domain occurs (from 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.

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 (Barghouthi et al., 1989; Bohac et al., 2019; Han et al., 2013; Michelsen et al., 2015, 2015; Pandey et al., 2017; Schmidt et al., 2014; Tejman-Yarden et al., 2019), and known antifungal compounds like 2,4-diacetylphloroglucinol (Raaijmakers and Weller, 1998). 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.

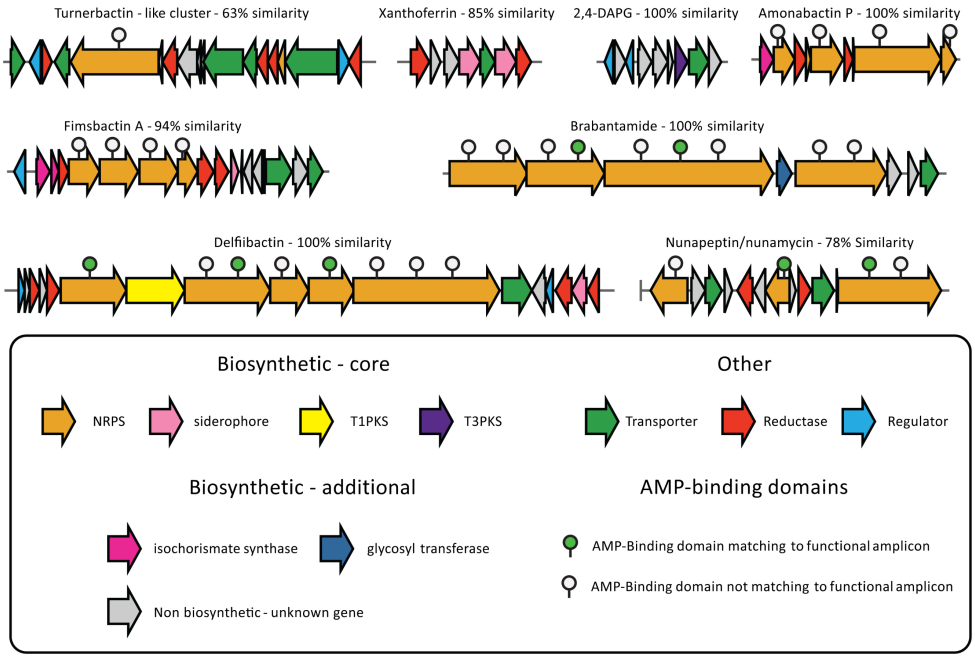


Figure 6. Selection of known BGCs predicted in the rhizosphere metagenome of suppressive soil S11. Arrows represent predicted genes and are colour-coded based on their annotated function. AMP-binding domains matching to functional amplicons are highlighted as described in the legend.

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.

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 (Fig. 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 (Haas and Défago, 2005; Kloepper et al., 1980; Kramer et al., 2020; Lemanceau et al., 1993; Saha et al., 2016). Siderophores and competition for iron were found to be involved in soil disease suppression mechanisms against Fusarium wilt (Alabouvette, 1986; Baker et al., 1986; Duijff et al., 1999, 1994; Scher M., 1982) , *take-all* disease in wheat (Lemanceau et al., 2009; Verbon et al., 2017) and damping-off of sugar beet (Carrión et al., 2019).

The concentration of soluble iron in eight tested soils, as assayed in our previous study (Ossowicki et al., 2020), 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 (Ossowicki et al., 2020)). 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 (Fig. 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 (Fig. 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 (Fig.6). The largest suppressive-sample-associated subnetwork by number of amplicons (Fig. 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.

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.

Data availability

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

Supplementary material

Table S1. Read counts per sample table pre and post filtering.

Sample-ID	Raw counts	Filtered counts
S01-30-C-1_38060-A01	61522	56356
S01-30-C-2_38060-B01	69249	63627
S01-30-C-3_38060-C01	93329	86633
S01-30-C-4_38060-D01	88579	81528
S01-30-FC-1_38060-E01	73573	68386
S01-30-FC-3_38060-G01	59679	54557
S01-30-FC-4_38060-H01	54754	50564
S03-30-C-1_38060-A02	49444	45627
S03-30-C-2_38060-B02	59744	55266
S03-30-C-3_38060-C02	45294	42639
S03-30-C-4_38060-D02	55966	51235
S03-30-FC-1_38060-E02	69574	66177
S03-30-FC-2_38060-F02	72277	66783
S03-30-FC-3_38060-G02	56285	51549
S03-30-FC-4_38060-H02	54935	49470
S08-30-C-1_38060-A05	57233	53380
S08-30-C-2_38060-B05	66362	61590
S08-30-C-3_38060-C05	64633	60426
S08-30-C-4_38060-D05	82370	77089
S08-30-FC-1_38060-E05	68239	64447
S08-30-FC-2_38060-F05	89136	82245
S08-30-FC-3_38060-G05	69234	63569
S08-30-FC-4_38060-H05	56346	51398
S11-30-C-1_38060-A03	56932	53328
S11-30-C-2_38060-B03	72236	67006
S11-30-C-3_38060-C03	60323	56541
S11-30-C-4_38060-D03	76486	72089
S11-30-FC-1_38060-E03	65674	62059
S11-30-FC-2_38060-F03	69165	63976
S11-30-FC-3_38060-G03	61222	57484
S11-30-FC-4_38060-H03	62399	58216

S14-30-C-1_38060-A06	65009	60765
S14-30-C-2_38060-B06	70802	65514
S14-30-C-3_38060-C06	64070	59514
S14-30-C-4_38060-D06	54727	50559
S14-30-FC-1_38060-E06	72434	66960
S14-30-FC-2_38060-F06	81112	71312
S14-30-FC-3_38060-G06	59644	55225
S14-30-FC-4_38060-H06	64895	58831
S15-30-C-1_38060-A07	59465	53857
S15-30-C-2_38060-B07	68717	62513
S15-30-C-3_38060-C07	75261	69264
S15-30-C-4_38060-D07	74238	68226
S15-30-FC-1_38060-E07	83840	77529
S15-30-FC-2_38060-F07	85701	77805
S15-30-FC-3_38060-G07	60580	56114
S15-30-FC-4_38060-H07	69506	63197
S17-30-C-1_38060-A08	56302	51984
S17-30-C-2_38060-B08	78187	67847
S17-30-C-3_38060-C08	85159	76649
S17-30-C-4_38060-D08	80787	74714
S17-30-FC-1_38060-E08	73688	69596
S17-30-FC-2_38060-F08	76891	70879
S17-30-FC-3_38060-G08	63541	59126
S17-30-FC-4_38060-H08	61208	56702
S28-30-C-1_38060-A04	54744	49232
S28-30-C-2_38060-B04	72221	65227
S28-30-C-3_38060-C04	48818	43892
S28-30-C-4_38060-D04	69285	63359
S28-30-FC-1_38060-E04	14606	12380
S28-30-FC-2_38060-F04	85430	78041
S28-30-FC-3_38060-G04	54205	50309
S28-30-FC-4_38060-H04	54170	49003
Total:	3459604	3396393

Table S2. Statistic tests for Figure 2B

Phenotype, permanova	
method name	PERMANOVA
test statistic name	pseudo-F
sample size	63
number of groups	2
test statistic	4.9766
p-value	0.0010
number of permutations	999
Name: PERMANOVA results, dtype: object	
Treatment, permanova	
method name	PERMANOVA
test statistic name	pseudo-F
sample size	63
number of groups	2
test statistic	0.8660
p-value	0.6890
number of permutations	999
Name: PERMANOVA results, dtype: object	
Phenotype, anosim	
method name	ANOSIM
test statistic name	R
sample size	63
number of groups	2
test statistic	0.3529
p-value	0.0010
number of permutations	999
Name: ANOSIM results, dtype: object	
Treatment, anosim	
method name	ANOSIM
test statistic name	R
sample size	63
number of groups	2
test statistic	-0.0008
p-value	0.4030
number of permutations	999
Name: ANOSIM results, dtype: object	

Table S3. Cumulative relative abundance of predicted amino acid substrates in suppressive and conducive rhizosphere soils. Rank sum statistics is calculated using relative counts of appearances of amplicons annotated to the monomer for calculating ranks.

Substrate specificity	Cumulative relative abundance suppressive	Cumulative relative abundance conducive	Rank sums statistic	Rank sums	p-value
ala	0,1136	0,1355	0,8383	-2,3834	0,0203
asn	0,0077	0,0064	1,2166	-0,5024	0,6172
asp	0,0013	0,0004	3,5484	2,2128	0,0307
cys	0,0810	0,0864	0,9370	-3,3174	0,0015
dab	0,0006	0,0005	1,0138	2,5104	0,0148
gln	0,0341	0,0373	0,9149	2,5214	0,0144
glu	0,0009	0,0011	0,8449	-2,9170	0,0050
gly	0,2195	0,2234	0,9825	-0,8890	0,3776
leu	0,0122	0,0065	1,8587	3,4547	0,0010
phe	0,0015	0,0015	1,0138	0,1616	0,8722
pro	0,0271	0,0238	1,1377	-0,1885	0,8512
ser	0,2781	0,2843	0,9782	-1,2785	0,2060
thr	**0,2103**	**0,1817**	**1,1569**	**4,8296**	**9,83E-06**
tyr	0,0122	0,0109	1,1152	0,6348	0,5280

Table S4 Annotated MIBig clusters with counts in all suppressive and all conducive soils

cluster	suppressive counts	conducive counts
BGC0000438_AAF99707.2	5499	58
BGC0000438_AAO72425.1	4836	47
BGC0000437_AAY37655.1	4834	47
BGC0001416_KPN93063.1	4367	86
BGC0001416_KPN90376.1	4417	234
BGC0001721_AUD11994.1	3824	0
BGC0000984_ABX37383.1	2021	167
BGC0000984_ABX37382.1	1987	137
BGC0001346_AOA33122.1	1503	72
BGC0000438_AAO72424.1	1493	78
BGC0000437_AAY37654.1	1482	78
BGC0000437_AAY37653.1	1443	55
BGC0001842_ABA73955.1	1668	409
BGC0001980_QDF82255.1	1698	474

BGC0001509_WP	1987	800
BGC0000435_AAO56329.1	925	0
BGC0001567_WP	1142	473
BGC0001312_CAY48788.1	418	208
BGC0001416_KPN93064.1	276	76
BGC0000305_BAC67535.1	376	263
BGC0001752_ctg1	409	308
BGC0001984_QED55423.1	76	3
BGC0001519_WP	105	35
BGC0000305_BAC67536.1	142	77
BGC0001331_WP	37	0
BGC0001332_WP	31	6
BGC0000443_AED90003.1	26	3
BGC0000429_AEA30273.1	27	5
BGC0001999_WP	42	23
BGC0000425_AFH75321.1	96	78
BGC0001806_APU91750.1	96	78
BGC0001330_WP	24	6
BGC0000425_AFH75320.1	98	81
BGC0000429_AEA30272.1	18	1
BGC0000447_CCJ67639.1	100	83
BGC0000447_CCJ67638.1	100	83
BGC0001389_AHZ34242.1	96	80
BGC0001389_AHZ34241.1	96	80
BGC0000371_BAH33409.1	34	20
BGC0000431_EFE73312.1	14	0
BGC0000315_CAB38518.1	13	0
BGC0001370_ALV82356.1	13	0
BGC0000437_AAY37647.1	14	2
BGC0001192_AJM89738.1	19	8
BGC0001153_AEZ51520.1	19	8
BGC0001192_AJM89735.1	19	8
BGC0000408_ACA97580.1	19	8
BGC0000403_AFJ14794.1	19	8
BGC0000408_ACA97576.1	19	8

BGC0001153_AEZ51516.1	19	8
BGC0002014_WP	13	2
BGC0001715_CUX79061.1	16	6
BGC0001715_CUX79060.1	16	6
BGC0000429_AEA30274.1	13	4
BGC0001657_BAV56270.1	9	0
BGC0000461_AEP18655.1	5	0
BGC0000461_AEP18656.1	4	0
BGC0001042_ACY06285.1	20	16
BGC0001370_ALV82384.1	3	0
BGC0000315_CAB38517.1	3	0
BGC0001763_BBA20967.1	5	2
BGC0001620_ASX95241.1	5	2
BGC0001214_AJV88375.1	2	0
BGC0001984_QED55421.1	2	0
BGC0000418_AIE77059.1	1	0
BGC0002001_WP	5	5
BGC0000311_CAC48361.1	4	5
BGC0000955_CCA29203.1	0	1
BGC0000341_ABD65958.1	0	2
BGC0000359_AAZ55900.1	0	2
BGC0000385_AEH59099.1	0	2
BGC0000385_AEH59100.1	0	2
BGC0001462_OLZ52457.1	0	2
BGC0000419_AIG79241.1	0	2
BGC0001416_KPN90369.1	0	2
BGC0001459_OKA09423.1	0	2
BGC0001460_EME52990.1	0	2
BGC0000311_CAC48360.1	0	2
BGC0001460_EME52989.1	0	3
BGC0000326_AAK81826.1	0	3
BGC0001975_QBG38784.1	2	5
BGC0001459_OKA09424.1	1	5
BGC0000354_CAM56770.1	31	35
BGC0000393_ABF87031.1	0	4

BGC0000455_AEI58866.1	0	5
BGC0001792_WP	0	8
BGC0000463_AGM14934.1	11	28
BGC0000333_ABW00331.1	0	22
BGC0000447_CCJ67637.1	131	156
BGC0001368_WP	1	26
BGC0000289_CAD91212.1	3	31
BGC0000440_CAE53352.1	2	31
BGC0000441_CAG15011.1	2	31
BGC0001178_AGS77309.1	2	31
BGC0001806_APU91751.1	22	52
BGC0000425_AFH75322.1	15	50
BGC0000413_AAY93356.2	5	41
BGC0000413_AAY93354.1	5	41
BGC0001389_AHZ34243.1	31	73
BGC0000344_CAK15814.1	41	89
BGC0000463_AGM14933.1	41	89
BGC0000344_CAK15815.1	41	89
BGC0001838_AFJ23826.1	38	89
BGC0000439_AHH53506.1	0	66
BGC0001838_AFJ23825.1	15	82
BGC0000411_ABW17377.1	51	119
BGC0000411_ABW17376.1	47	139
BGC0001842_ABA73956.1	279	417
BGC0001980_QDF82259.1	325	477
BGC0001767_ctg1	2438	2651
BGC0000325_CAB53322.1	100	342
BGC0001312_CAY48789.1	124	400
BGC0000389_ABH06369.2	159	444
BGC0000423_CBG75492.1	2445	2741
BGC0000349_CAM02313.1	2291	2589
BGC0000389_ABH06368.2	268	611
BGC0001211_WP	2448	2820
BGC0000413_AAY93445.1	52510	73141

Table S5. Chemical properties of tested soils

	S01	S03	S11	S28
pH	6,22 SD 0,17	5,28 SD 0,27	7,28 SD 0,19	7,13 SD 0,07
OM [%]	3,46 SD 0,35	5,49 SD 0,68	3,48 SD 0,47	3,29 SD 0,12
Fe [mg/kg]	0,09 SD 0	0,45 SD 0,06	0,11 SD 0	0,02 SD 0
K [mg/kg]	34 SD 0,92	59,95 SD 2,29	68,77 SD 1,1	71,36 SD 0,44
Mg [mg/kg]	163,19 SD 1,75	56,74 SD 2,11	56,43 SD 0,58	160,26 SD 2,7
P [mg/kg]	1,4 SD 0,01	2,69 SD 0,08	5,43 SD 0,04	0,79 SD 0,02
S [mg/kg]	1,29 SD 0,08	2,6 SD 0,16	1,17 SD 0,15	1,74 SD 0,24
C [%]	2,31 SD 0,18	3,77 SD 1,13	1,99 SD 0,88	1,52 SD 0,01
N [%]	0,14 SD 0,01	0,26 SD 0,1	0,16 SD 0,07	0,17 SD 0
C:N	16,5	14,5	12,44	8,94

	S08	S14	S15	S17
pH	6,87 SD 0,06	7,61 SD 0,05	7,82 SD 0,05	7,75 SD 0,06
OM [%]	2,81 SD 0,51	5,04 SD 0,35	3,77 SD 0,63	4,23 SD 0,79
Fe [mg/kg]	0,08 SD 0	0,04 SD 0	0,02 SD 0	0,01 SD 0
K [mg/kg]	181,54 SD 3,76	82,49 SD 4,08	45,83 SD 0,53	87,17 SD 2,45
Mg [mg/kg]	109,68 SD 1,7	98,91 SD 4,04	60,6 SD 0,52	74,54 SD 0,8
P [mg/kg]	16,23 SD 0,05	0,7 SD 0,06	0,9 SD 0,05	1,03 SD 0,04
S [mg/kg]	1,14 SD 0,05	8,61 SD 0,08	5,95 SD 0,06	2,27 SD 0,06
C [%]	1,94 SD 0,24	2,35 SD 0,14	1,91 SD 0,08	2,63 SD 0,05
N [%]	0,16 SD 0,01	0,15 SD 0,01	0,11 SD 0,01	0,15 SD 0
C:N	12,13	15,67	17,36	17,53

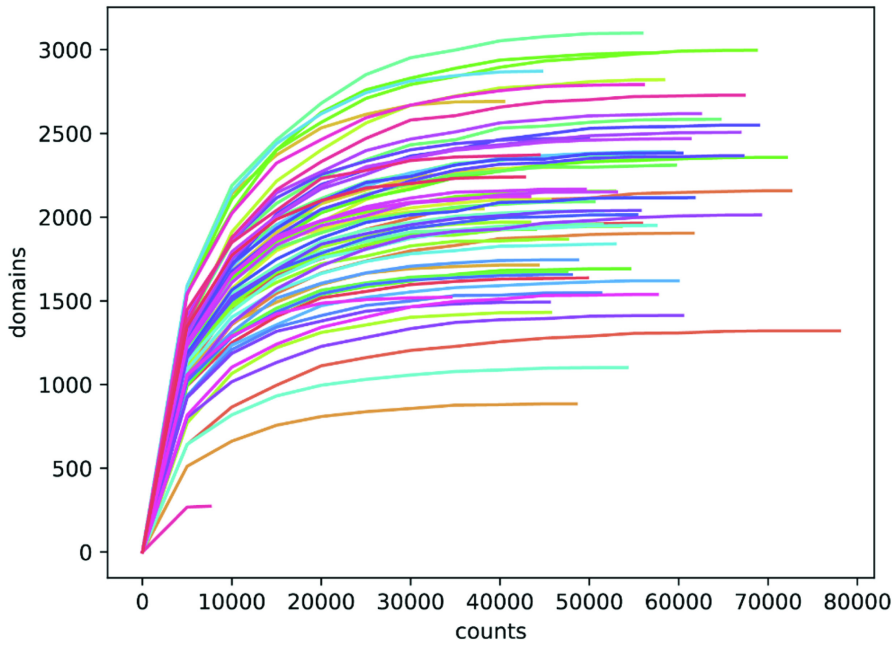


Figure S1. Rarefaction analysis of rhizosphere A-domain amplicons. Samples rarefaction analysis showing number of unique amp-binding domains (ordinate) at different rarefaction points (abscissa). One sample was clearly an outlier (pink arrow on the bottom) and was removed from the analysis.

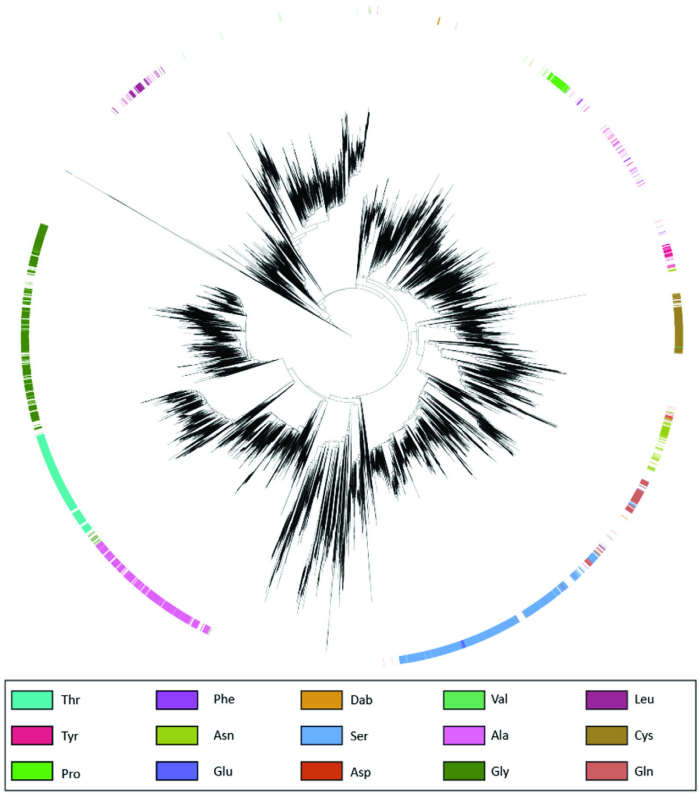


Figure S2. – Phylogeny tree of the natural amplicons analyzed in this study. On the outer ring, colors represent predicted amino acid specificity and amino acid group specificity as detailed in supplementary material.