Elucidating the diversity and potential function of nonribosomal peptide and polyketide biosynthetic gene clusters in the root microbiome

Barak Dror[a,b], Zongqiang Wang[c], Sean F. כBrady[c], Edouard Jurkevitch[b], Eddie Cytryn[a]#

[a] Institute of Soil, Water and Environmental Sciences, Agricultural Research Organization, Rishon-Lezion, Israel.

[b] Department of Plant Pathology and Microbiology, The Robert H. Smith Faculty of Agriculture, Food and Environment, The Hebrew University of Jerusalem, Rehovot, Israel.

[c] Laboratory of Genetically Encoded Small Molecules, The Rockefeller University, New York, New York, USA.

#Address correspondence to Eddie Cytryn, eddie@volcani.agri.gov.il

Abstract

Polyketides (PKs) and nonribosomal peptides (NRPs) are two microbial secondary metabolite (SM) families known for their variety of functions, including antimicrobials, siderophores and others. Despite their involvement in bacteria-bacteria and bacteria-plant interactions, root-associated SMs are largely unexplored due to the limited cultivability of bacteria. Here, we analyzed the diversity and expression of SM-encoding biosynthetic gene clusters (BGCs) in root microbiomes by culture-independent amplicon sequencing, shotgun metagenomics and metatranscriptomics. Roots (tomato and lettuce) harbored distinct compositions of nonribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) relative to the adjacent bulk soil, and specific BGC markers were both enriched and highly expressed in the root microbiomes. While several of the highly abundant and expressed sequences were remotely associated with known BGCs, the low similarity to characterized genes suggests their potential novelty. Low similarity genes were screened against a large set of soil-derived cosmid libraries, from which five whole BGCs of unknown function were retrieved. Three clusters were taxonomically affiliated with Actinobacteria, while the remaining were not associated with known bacteria. One Streptomyces-derived BGC was predicted to encode for a polyene with potential antifungal activity, while the others were too novel to predict chemical structure. Screening against a suite of metagenomic datasets revealed a higher abundance of retrieved clusters in roots and soil samples. In contrast, they were almost completely absent in aquatic and gut environments, supporting the notion that they might play an important role in root ecosystems. Overall, our results indicate that root microbiomes harbor a specific assemblage of undiscovered SMs.

Importance

We identified distinct secondary metabolite (polyketide and nonribosomal peptide) encoding genes that are enriched (relative to adjacent bulk soil) and expressed in root ecosystems, yet almost completely absent in human gut and aquatic environments. Several of the genes were distantly related to genes encoding for antimicrobials and siderophores, and their high sequence variability relative to known sequences suggests that they may encode for novel metabolites and may have unique ecological functions. This study demonstrates that plant roots harbor a diverse array of unique secondary metabolite encoding genes that are highly enriched and expressed in the root ecosystem. The secondary metabolites encoded by these genes might assist the bacteria that produce them in colonization and persistence in the root environment. To explore this hypothesis, future investigations should assess their potential role in inter-bacterial and bacterial-plant interactions.

Keywords: Soil microbiome, root microbiome, polyketides, nonribosomal peptides, secondary metabolites, plant-microbe interactions

Introduction

Soil is an extremely diverse ecosystem that contains a myriad of micro- and macro-organisms, including nematodes, arthropods, fungi and bacteria. The rhizosphere is a narrow region of soil directly influenced by root exudates and mucilage (1, 2) . This "hot spot" of organic matter and nutrients "enriches" a specific fraction of the soil microbial community known as the root microbiome, which is significantly different than the surrounding soil microbiome (3). Over the past two decades, several studies have linked specific constituents of the root microbiome to enhanced plant growth and development and inhibition of soilborne plant pathogens (4), either by direct antagonism, and/or induced systemic resistance (5). These functions are often facilitated by the vast array of secondary metabolites (SMs) produced by root-associated bacteria, which play a key role in inter- and intra-species interactions (6, 7).

Many important soil and root-associated bacterial SMs are nonribosomal peptides (NRPs), or polyketides (PKs), produced by nonribosomal peptide synthetases (NRPSs) or polyketide synthases (PKSs), respectively. These are encoded on large biosynthetic gene clusters (BGCs) that often exceed 50,000 bp (8). Enzymatic complexes in these families follow a similar biosynthetic logic wherein molecules are assembled in an iterative building process using conserved domains that are organized in modules (9, 10). NRPSs and PKSs are responsible for the synthesis of a wide array of siderophores, toxins, pigments and antimicrobial compounds (11) that are believed to play a pivotal role in bacterial adaptation to soil and rhizosphere ecosystems, and in plant health and development (12). Despite their ecological (rhizosphere competence) and translational (biocontrol agents and novel antimicrobial compounds for plant protection) importance, little is known about the occurrence, diversity and dynamics of NRPSs and PKSs in root ecosystems, or their role in intra- and inter- microbial and plant-bacterial interactions.

A major challenge in exploring the role and function of SMs in soil stems from the fact that the majority of bacteria cannot be cultivated using conventional methods, making it difficult to study these bacteria and the diversity, expression and function of the metabolites they produce (13). Despite progress made in culturing techniques, our capacity to isolate soil and root-associated bacteria is highly constrained, primarily because it is challenging to mimic the natural conditions required for growing these bacteria (14). Furthermore, many bacterial BGCs are silent under laboratory conditions and therefore the metabolites that they encode are extremely challenging to isolate (15).

To circumvent the above mentioned barriers, a myriad of culture-independent sequencing-based and omics tools have been developed to reveal the scope and composition of soil-derived BGCs encoding NRPSs and PKSs (16, 17), and to infer the chemical composition and structure of the metabolites produced by these synthases (18, 19). For instance, amplicon sequencing-based approaches have been developed to target short fragments within adenylation (AD- in NRPS) and ketosynthase (KS- in PKS) domains. These amplicons can be used to ascertain the diversity and abundance of bacterial BGCs in complex environments, as both AD and KS domains are important (in concert with other components) for the assembly, and thus identity and activity, of the synthesized metabolites (20, 21). To date, a handful of studies have explored the diversity and composition of bacterial SM-encoding BGCs in soil, demonstrating the vast genetic diversity and novelty of NRPSs and PKSs genes (22, 23). However, little is known regarding the distribution and of these gene families in the root microbiome, and their functional role in this complex community remains an enigma (24).

This study proposes a unique approach to analyze the diversity and potential functions of NRPSs and PKSs in the root, specifically focused on elucidating: (i) the composition and diversity of

NRPSs and PKSs encoding genes in the root environment relative to adjacent bulk soil; (ii) NRPSs and PKSs composition and expression in the root as a function of plant type; (iii) the sequence and inferred SM structures of whole bacterial BGCs that are highly abundant or expressed in root environments; and (iv) the occurrence of root-enriched bacterial BGCs in other ecosystems.

Results

1. Composition and diversity of NRPSs and PKSs genes in roots vs. bulk soil

To determine the composition and diversity of NRPSs and PKSs genes in tomato and lettuce root samples relative to bulk soil (previous studies targeting this controlled lysimeter system showed that bulk soil from tomato and lettuce microbiomes were almost identical and therefore only tomato soil was analyzed here), we applied a previously described amplicon sequencing approach to amplify the conserved adenylation (AD) and ketosynthase (KS) domains of NRPSs and PKSs, respectively (23). Overall, sequencing yielded a total of 1,850,442 and 2,174,020 raw KS and AD reads with average read lengths of 280 bp and 235 bp, respectively (**Table S1**). Further filtering steps using QIIME2 and DADA2 denoising methods, resulted in 2,980 and 3,269 non-redundant KS and AD domain sequences, respectively.

We observed significantly higher diversity of both AD and KS domains in the bulk tomato soil compared to the adjacent roots (**Fig. S1- A and C**). In contrast, no difference in diversity was observed between tomato and lettuce roots for either of the SM encoding domains (**Fig. S1- B and D**). To assess differences in AD and KS domain diversity between samples, a principal coordinate analysis (PCoA) and analysis of similarity (ANOSIM) using the Bray-Curtis

similarity index were performed (Fig. **1A** and **1B**). The KS and AD domain profiles of the roots (from both tomato and lettuce) formed distinct clusters, which were significantly different from the adjacent bulk soil ($R=0.332$, $p< 0.05$ for PKS, $R=0.308$, $p<0.01$ for NRPS).

To explore the potential novelty of root-associated SM-encoding genes, the amplified AD and KS domain sequences from the root and soil samples were first aligned against the MIBiG (Minimum Information about a Biosynthetic Gene cluster) repository (25) using blastp, with >50% amino acid sequence identity cut-off, and then grouped according to their identity to the MIBiG reference genes (**Table 1**). On average, more than 25% of the AD and almost 13% of the KS domain sequences in the root environment showed less than 50% amino acid identity with genes found in the MIBiG database (characterized as 'unassigned'), whereas less than 1% and 6% of the AD and KS sequences (respectively) shared over 85% similarity to the reference MIBiG genes. These results demonstrate the profusion of potentially novel SMs in both root and soil environments.

2. Pinpointing predicted root-enriched NRPs and PKs

As SMs are known to play critical roles in bacteria-bacteria and bacteria-plant interactions, we were interested in the associated metabolites synthesized by BGCs whose AD or KS domains were highly abundant and enriched in tomato or lettuce roots relative to bulk tomato soil. To do so, the MIBiG-aligned amplicons were annotated to corresponding BGC-associated metabolites, using a cut-off e-Value of $< 10^{-40}$. Sequences that did not meet these criteria were defined as 'unknown'. Previous analyses have shown that when compared to reference KS or AD domain

sequences, amplicons with e-Values of < 10^{-40} are likely derived from the same BGC family as the reference sequence, and thus may be inferred to encode a similar function (26-29).

A differential abundance analysis using DESeq2 of the top20 highly abundant AD and KS amplicons revealed that 55% (11/20) and 70% (14/20) of the amplicons in both of the plant root samples (tomato and lettuce respectively, adjusted p-value < 0.1, log2FoldChange>5), were not associated with known BGCs, thus their associated metabolites cannot be inferred (**Fig. 2**). Nonetheless, several root-enriched AD and KS domains sequences (9 in tomato and 6 in lettuce) were above these threshold values, thus can be considered as congener to known metabolites. These included the non-ribosomal peptides stenothricin (30) and griselimycin (31) whose BGC analogues were highly abundant in the tomato and lettuce roots, respectively, and were less profuse in the bulk soil. While we cannot determine the actual role of these enriched metabolites in the root environment, both of them are known for their antimicrobial activity.

Next, we calculated the relative abundance of amplicons that were associated with known metabolites (based on the criteria above) in the different root and soil samples (**Fig. S2**).To pinpoint associated metabolites that may play a role in adaptation to the root environment, we focused our analysis on metabolites that were present in at least four of the root samples (tomato and lettuce) and in no more than one soil sample (**Fig. 3**). In addition, to identify metabolites specifically relevant to soil, we also selected metabolites that were present in all three soil samples and in no more than one root-associated sample. For NRPs, we found four associated metabolites that were highly abundant in both of the root samples (*e.g.* the Streptomyces-derived antibiotic macrolide family streptovaricin). For PKs, we again found several highly abundant Streptomyces-derived predicted metabolites, among others. These included lasalocid, sanglifehrin A and azalomycin A. Interestingly, amplicons associated with the two former were

also found to be highly enriched in roots relative to soil in lettuce (Fig. 2B). Overall, we found 26 associated metabolites were present in at least one of the root-associated samples, and completely absent in the soil samples, *e.g*. diaphorin (in lettuce) and basiliskamides (present in both tomato and lettuce, **Fig. S2**).

Due to the potential biases associated with the above PCR-based approach, we analyzed previously published (32) shotgun metagenomes of the same tomato and lettuce root samples (n=3 each). Assembled open reading frame (ORF) sequences from the metagenomes identified using Prodigal were aligned against the MIBiG database, generating a list of the fifty most abundant NRPSs and PKSs in each of the root datasets (representing the normalized abundance within-samples by plotting the coefficient of variance (CV) for each gene, **Fig. S3A** and **B**). In addition, as gene clusters are often silent or expressed under very specific conditions, we evaluated gene expression in parallel to gene occurrence to uncover active BGCs with ecological importance in the highly dynamic root ecosystem. Thus, in parallel to the shotgun metagenome analysis described above we applied a similar analytical approach using the previously collected shotgun meta-transcriptomics (32) to identify NRPSs and PKSs with enhanced expression in lettuce and/or tomato root microbiomes. Interestingly, 60% (30/50) and 46% (23/50) of the AD and KS domains that were highly abundant in the tomato and lettuce root samples (respectively) were highly expressed as well (**Fig. S**3).

Next, we filtered the highly abundant hits based on their CV value (<50) in order to analyze sequences with lower dispersion levels within tomato and lettuce root-associated samples, followed by taxonomical annotation using MEGAN. The resulting 42 sequences were clustered into two main phyla- Actinobacteria (13/42) and Proteobacteria (25/42). Several sequences could not be assigned taxonomic affiliation and one was assigned to the Bacteriodetes phylum (**Fig. 4**).

While we could not infer the associated metabolites of most of these highly abundant sequences (including all of those assigned to the phylu Proteobacteria), suggesting their potential novelty, we did manage to annotate several of the Actinobacteria-associated metabolites. These were distantly associated with Ossamycin (5 hits) and Polycyclic Tetramate Macrolactams (PTMs; 2 hits, including the most highly abundant NRPSs/PKSs-related sequence and the 5th most expressed). Ossamycin is a known antifungal and cytotoxic macrocyclic polyketide originally isolated from soil-associated Streptomyces hygroscopicus var. ossamyceticus (33, 34), while PTMs are a family of biologically important metabolites including HSAF, ikarugymcin and clifednamides, generally associated with different isolates of Actinobacteria and Gammaproteobacteria (35).

Finally, to evaluate the extent to which the amplicon sequencing method was able to detect NRPSs and PKSs in the targeted samples relative to the PCR-independent shotgun-metagenomic analyses, we analyzed the distribution of total MIBiG-associated genes in all four datasets (lettuce and tomato NRPSs and PKSs amplicons, tomato and lettuce metagenomes, **Fig. S4**). In general, approximately 34% and 25% of the MIBiG-characterized genes were found in both amplicon sequences and metagenomes of the lettuce and tomato roots, respectively. In contrast, approximately 33% and 20% of the tomato and lettuce genes, respectively, were only detected in the shotgun metagenomic datasets. Less than 4% and 5% of the NRPSs and PKSs were found only within the tomato and lettuce amplicon sequencing datasets (28 and 35 genes in tomato and lettuce, respectively) and less than a fifth (139 genes or 17.2%) were common to all four culture-independent datasets.

3. <u>Extraction and environmental distribution of whole SM-encoding gene clusters</u>

The identification of NRPSs and PKSs that were either enriched in roots relative to adjacent bulk soil or were abundant and/or highly expressed in lettuce and/or tomato roots, encouraged us to capture whole BGCs associated with these sequences, in order to shed light on their phylogenetic affiliation and potentially infer their function and chemical structure. This was achieved by screening the root-associated NRPSs and PKSs candidate sequences identified in this study against a large set of previously collated soil and rhizosphere cosmid libraries using the bioinformatic platform eSNaPD (29)(see Materials and Methods for full pipeline). Five cosmid library targets showed low e-Values and high nucleotide sequence identity (>75%) to candidate AD or KS sequences. Sequencing and annotation of the metagenomic insert captured in each cosmid revealed three NRPS and two hybrid NRPS/PKS gene clusters (**Fig. 5**). Based on gene content and sequence identity, the identified gene clusters were not identical to BGCs associated with known metabolites. The NRPS and PKS ORFs of two recovered clones (B326 and B385) were not affiliated with any known bacterial taxa (<50% nucleic acid identity to the NCBI database), while the other three clones were related to genes from Actinobacteria **(Table 2**).Of the cosmids recovered from the metagenomic libraries, clone B481 was nearly identical to an uncharacterized NRPS BGC found in the genome of *Streptomyces cyaneogriseus* (**Fig. 6A**). The only predicted chemical structure we could infer from the recovered BGCs was for clone B893, which was related to an uncharacterized PKS gene cluster found in the genome of *Saccharothrix saharansis*, a filamentous Actinobacteria isolated from desert soil. A detailed bioinformatic analysis of its PKS domains revealed that the gene cluster likely encodes for an extended polyene substructure (**Fig. 6B**). The seven PKS modules captured on the clone all contain dehydratase (DH) and ketoreduction (KR) domains indicating that each module introduces a double bond into the polyketide backbone (**Fig. 6B**). While polyene substructures like this are

seen in a number of natural products (36, 37), they are most commonly seen in polyene antifungal agents, including many that are derived from *Streptomyces* species (e.g., cyphomycin, nystatin, filipin and pimaricin). This may suggest that the BGC encode for an antifungal compound.

The five recovered BGCs were initially targeted due to their abundance in tomato and lettuce root datasets, suggesting a link to root ecosystems. To test this hypothesis, we assessed the abundance of the five BGCs in a large collection of publicly available shotgun metagenomes (20 metagenomes from each environment) from four distinct environments, targeting gut (animal and human), aquatic (freshwater and marine), soil (different soil types) and root-associated (various plant species) datasets. Our analysis demonstrated that the recovered BGCs are ubiquitous in most of the queried root samples and in some of the soil samples (**Fig. 7A**). Only B893 showed a significantly higher abundance in the root samples when compared with the soil samples ($P<0.05$, Wilcoxon test, **Fig. S5**). B893 was found in 16/20 of the root-associated datasets we examined (compared with 8/20 soil datasets, **Fig. 7B**). In contrast, none of the five BGCs were found in any of the gut microbiome communities analyzed, and very few were detected in the aquatic environments (**Fig. 7B**).

Discussion

NRPs and PKs produced by root-associated microbial communities play an important role in plant root ecosystems (38, 39). Several studies have identified and characterized BGCs and/or associated metabolites in prominent plant growth promoting and biocontrol agents originally isolated from plant roots (40, 41). However, the large fraction of uncultivated bacteria in root

ecosystems and the limitations associated with culturing bacteria encouraged us to examine the composition of genes encoding NRPS and PKS in root environments using culture-independent approaches. These methods have been applied before to understand SMs diversity and distribution in bulk soil (17, 26, 42, 43), but NRPSs and PKSs have not been comprehensively explored in root ecosystems. Furthermore, to the best of our knowledge, this is the first study to explore expression of secondary metabolites associated genes in root ecosystems.

Our results demonstrate a distinct composition of NRPSs and PKSs in plant (tomato and lettuce) roots relative to adjacent bulk soil; that these genes in root microbial communities are less diverse than those found in soil microbiomes; and that a fraction of these NRPSs and PKSs are highly expressed in root ecosystems. This is consistent with several previous studies showing that the phylogeny and functionality of root-associated microbial communities are significantly different from those in adjacent soil (44-46). It is well established that plants interact with soil bacterial communities through secretion of root exudates (47, 48), resulting in selection of specific microbial populations from the soil microbiome. This appears to be the case in the recruitment of plant-growth promoting rhizobacteria (PGPR), which are known to harbor specific SM encoding genes (49-51). Thus, while at this point we cannot infer the actual function of the highly abundant and expressed NRPs and PKs encoding genes in the root environment, we can infer that these NRPSs and PKSs likely play a role in various processes, *e.g.* competition and root colonization (52, 53).

While most of the detected SM-associated genes were too novel to link to any known metabolites, a few of the highly abundant, expressed and/or root-enriched NRPSs and PKSs were associated with known metabolites. Azalomycin F, for instance, found in both of the plant microbiome samples, is a polyketide with reported antifungal activity against a variety of

phytopathogens, which is produced by different soil- and root-associated Streptomyces isolates (54, 55). Diaphorin, found only in the tomato root microbiome, is a pederin-analog known to be produced by the psyllid *Diaphorina citri* endosymbiont *Candidatus* Profftella armatura (Betaprotoebacteria), with potential antifungal and cytotoxic activity (56). In this regard, the described culture-independent approaches are a promising platform for identification of novel BGCs and for elucidating their roles in soil ecosystems and within the framework of drug discovery efforts, despite their current limitations (16, 17).

A large fraction of the highly abundant and highly expressed NRPSs and PKSs identified in this study were not assigned to, or had low identity (50-70%), when compared to previously characterized genes in the MIBiG repository. Recently, a previously unidentified hybrid NRPS/PKS gene cluster was found to be essential for Rhizoctonia suppression by an endophyte *Flavobacterium* sp. (57), highlighting the vast amount of root-associated SMs with unidentified functional roles, that undoubtedly play a pivotal role in bacterial-plant interactions.

We screened a large set of soil cosmid libraries with candidate sequences from our amplicon sequencing and metagenomic analyses that were enriched (relative to bulk soil), and/or highly abundant and expressed in tomato and lettuce roots, taking advantage of a unique culture-independent platform capable of extracting and analyzing long NRPSs and PKSs (27, 58). Five clones containing uncharacterized gene clusters with no known function, including two that were not associated with any known taxonomic group were identified. The fact that all of these BGCs were rather common in various root-associated environmental metagenomes, but rare or completely absent in other environments, suggests their potential importance in these habitats. While we can only speculate as to their synthesized metabolite's actual activity, they were associated with bacterial groups well-known for their antimicrobial capacity. *Saccharohrix*

*saharansis (*Actinobacteria*)*, for instance, which contains a BGC closely related to clone B893, is a soil-dwelling bacteria known to produce an array of antimicrobials (59). This BGC likely encodes for a polyene substructure, often seen in antifungal agents, as it capable of directly disrupting the fungal membrane (60). Of particular interest is clone B481, associated with SM-associated genes from *Streptomyces cyaneogrisues*, known for its ability to produce the biopesticide Nemadectin (61). We speculate that this BGC may be associated with bacterial-fungal competition in the root ecosystems.

At the broader level, our results emphasize the need to look beyond basic descriptive diversity and composition information regarding SM capacity of microbial communities. The pipeline adopted in this study, where potentially important AD and KS sequences are first identified, followed by extraction of BGCs from cosmid libraries in order to identify potentially novel BGCs, has the advantages of being resource-efficient, while spawning deeper knowledge regarding potential important gene clusters. Future studies will focus on expressing these cosmid library BGCs in suitable hosts, enabling us to characterize their encoded metabolites and test their *in vitro* and *in planta* activity against various phytopathogens. Our results coincide with studies conducted in other plants, showcasing the yet unexplored diversity of NRPSs and PKSs (43, 62).

Overall, this study indicates that the root microbiome harbors a unique, diverse and potentially novel array of SM synthesizing genes, which are significantly different from those in the bulk soil microbiome. To enhance our current understanding, future research should focus on identifying additional factors shaping the occurrence and expression of SMs in the root microbiome (*e.g.*, plant health, presence of phytopathogens and plant growth). This will

undoubtedly help expose the ecological role of SMs in root ecosystems and provide a platform for drug discovery and novel and environmentally sustainable compounds for plant protection.

Materials and Methods

Fig. S6 present a conceptual description of the pipeline applied in this study.

Amplicon sequencing, shotgun metagenomic and meta-transcriptomics analyses

Tomato soil and root samples and lettuce root samples were collected as previously described (63). Briefly, tomato (*Solanum lycopersicum*-Heinz 4107) and lettuce (*Lactuca sativa*-Romaine-Assaph) seedling were planted and grown for 42 days in a random block lysimeters experiment, at Lachish agricultural research station in Kiryat-Gat, Israel. Each sample type (soil, tomato roots and lettuce roots) was analyzed in triplicate samples from three different lysimeters (thus n=3 for each sample type). Each triplicate consisted of a composite sample collected from 2-4 plants or 3 soil samples, taken from the distant edges of the lysimeters and away from plant roots. As the same soil was used throughout the experiment, and given soil sample collection was taken for sufficient distance from growing plants, bulk soil from the tomato lysimeters served as a reference point for both tomato and lettuce soils. Previous study showed that they harbored almost identical microbial communities (Zolti et al. 2019). Soil samples were collected, frozen in liquid nitrogen on site, and stored at −80 °C for until further analysis. Root samples were collected intact, and soil particles were removed by shaking and briefly rinsing. The roots were then lightly dried and immediately frozen in liquid nitrogen on site and kept at −80 °C until processed. In this study, extracted DNA was used as template for NRPSs and PKSs PCR amplification using degenerated primers (A3F/A7R for AD-NRPS and degKS2F/degKS2R for

KS-PKS domains, as previously reported (23). Resulted barcoded libraries were pooled and sequenced on an Iluumina MiSeq, employing V2 chemistry at the University of Illinois Chicago Sequencing Core (UICSQC). A total of 18 samples were sequenced, which included sampling location (tomato soil vs. tomato and lettuce roots) and SM family (AD and KS domains) with three replicates for each treatment. Resulting sequences were processed and demultiplexed using QIIME2 pipeline and the integrated DADA2 method (64). Exact Sequence Variants (ESVs) represented by fewer than 3 sequences were removed from downstream analyses. Raw amplicon sequences are available via the MG-RAST data depository, under mgm4862150.3. In addition, shotgun metagenome and metatranscriptome datasets of lettuce and tomato roots (n=3 for each dataset type, hence 6 for tomato and 6 for lettuce), previously generated and analyzed from same samples were also used for NRPSs and PKSs identification as described below (32). Shotgun sequences data is available via the NCBI Sequence Read Archive (SRA) data depository, with the project number PRJNA602301.

Identification and annotation of NRPSs and PKSs

For chemical diversity analysis of NRPS and PKS gene clusters, the different datasets (ESVs generated via amplicon sequencing and metagenome/meta-transcriptome assembled genes) were aligned against the Minimum Information about a Biosynthetic Gene cluster repository (MIBiG, version August 6th, 2018). Only core NRPSs and PKSs genes (AD and KS domains) were included in the analysis. Alignment was performed using diamond blastx command line, with >50% amino acid sequence identity. To associate each NRPS or PKS hit with its potentially synthesized metabolite, e-Value $<10^{-40}$ was used as a cut-off. Hits that did not pass this threshold were regarded as 'unknown'.

For taxonomical annotation, sequences were aligned against the non-redundant (nr) Blast NCBI database, followed by lowest-common-ancestor (LCA) classification using MEGAN 6.15 Ultimate edition by taking the top 10% hits and filtering for a minimum score of 50 and maximum eValue of 0.01 (65). Conversion of gene identifiers to taxonomical path was done using the mapping files provided by MEGAN as of October 2016.

Soil libraries amplicons generation screening

Metagenomic libraries were constructed as previously reported (66). Briefly, in each library, crude environmental DNA (eDNA) was extracted directly from field collected soil, gel purified, blunt ended, ligated into cosmid pWEB::TNC (Epicenter), packaged into lambda phage and transferred into Eco EC100 (Lucigen). Each library was expanded to contain 20 x106 unique cosmid clones with ~30-45 kb eDNA (environmental DNA) inserts and then arrayed into 768 sub-pool (2×384-well plates) containing ~25,000 unique cosmid clones per well. Each sub-pool was then stored as a glycerol stock for the clone recovery of interesting hits and as cosmid DNA to facilitate PCR-based screening. To generate an amplicon sequence database of NRPSs and PKSs, two set of degenerate primers (AD and KS) were applied to amplify the conserved region in Adenylation and Keto-synthase domains in the biosynthetic gene cluster: AD (forward) 5'-SATBTAYACSTCVGGHWCSAC-3' and (reverse) 5'-CCANRTCNCCBGTSYKGTACA-3'; KS (forward) 5'-TGYTCSDSSTCGCTSGTSGCS-3' (reverse) 5'-GTNCCSGTSCCRTGBGCYTCS-3'. The 5' ends of the primers were augmented with Miseq sequencing adapters followed by unique 8 bp barcode sequences identifying the soil metagenome from which they were amplified. Amplicons were pooled as collections of 96 samples and cleaned using magnetic beads. Cleaned, pooled amplicons were used as template in a second PCR. Prior to sequencing, all PCR amplicons were quantified by gel electrophoresis and mixed

in an equal molar ratio. The resulting pool was fluorometrically quantified with a HS D1000 ScreenTape and sequenced on an Illumina MiSeq instrument. Reads were debarcoded and trimmed to 240 bp. The reads from each sample were clustered using UCLUST (67) to generate the 95% tags.

Recovery of BGC Clones from metagenomic library pools

The library well locations for target AD or KS domains were identified using well-specific barcodes incorporated in the degenerate primers (27). Specific primers with Tm value ~60°C (18 to 20 bp) were designed to amplify each unique conserved sequence of interest. To recover the single cosmid clone from each library subpool, a serial dilution of whole-cell PCR strategy was used (17). Briefly, library glycerol stocks that contain target hits from eSNaPD analysis were inoculated into 3 ml LB broth (Kanamycin & Chloramphenicol) and grow overnight at 37 °C to confluence. The overnight cultured cells were diluted to 2000 CFU ml-1 calculated by OD600nm. 384 well plates were inoculated with 50 ul of the resulting dilution (600 CFU/well) with an Eppendorf epMotion 5075 liquid hander, grown overnight and screened using real-time PCR with touchdown PCR program, to identify wells containing target clones. Target positive wells were diluted to a concentration of ~5 CFU ml-1 and the process repeated to identify new wells containing target clones. Five clone pools were then plated on solid agar plate and target single clones were identified by clone PCR.

Analysis of Recovered Gene Clusters

Recovered single cosmid clones were miniprepped by QIAprep kit and sequenced using Miseq technology. M13-40FOR and T7 universal primers were utilized to sequence both ends of the insert sequences. Reads, amplicons and End sequences were assembled together to generate

constant contigs using Newbler 2.6 (68). Fully assembled contigs were then analyzed using an in-house annotation script, consisting of open reading frame (ORF) prediction with MetaGeneMark, HMM Scan and BLAST search. The annotation script was developed using Python and is available in the open source repository: https://github.com/brady-lab-rockefeller/gene_annotation. Putative function and source organism of genes in the BGC were assigned based on the closest characterized gene in NCBI. KnownClusterBlast in antiSMASH 5.0 (69) was utilized to analyze the relationship between known characterized gene clusters and recovered BGC. The structure prediction of adenylation domain and keto-synthase domain in BGC was given by the antiSMASH prediction which employ three prediction algorithms, NPRSPredictor2, Stachelhaus code, and SVM prediction. These predicted building blocks were then utilized to predict final structure combined with known characterized BGC in cultured bacteria.

Recovered clones search in environmental shotgun metagenomes

AD and KS domains from all five recovered gene clusters were searched against shotgun metagenomes from four different environments- animal and humans feces (gut), aquatic, soil and root-associated. We selected 20 Illumina-sequenced shotgun metagenomes from each of these ecosystems using the JGI IMG/MER advanced search option, followed by a blastn search using the IMG website online tool. Additional filtering was performed based on eValue threshold (10^{-40}) and identity (>85%). For each hit, counts were normalized using gyrB and rpoB housekeeping genes counts (obtained via the IMG/MER platform). A relative abundance heatmap created using pheatmap R package (70). For gene clusters with more than one AD/KS domain, results are shown only for datasets that contained all cluster-belonging domains. Gene clusters

presence/absence plot created using ggplot2 R package. Further information regarding selected metagenome datasets is shown in table S2.

Statistical Analyses

Alpha (Simpson index) and beta (Bray-Curtis) diversity indices across environments (bulk soil vs. roots and soil vs. tomato vs. lettuce) were calculated using R package vegan (71). Variation of NRPS/PKS-associated genes was visualized by Principal Coordinate Analysis (PCoA) using same R package. To obtain this figure, we performed ordination on ESV count table (constructed by qiime2) using Bray-Curtis distances, followed by plotting using R ggplot2 package. Difference significance between groups was determined using vegan ANOSIM (Analysis of similarities) analysis.

Enrichment of ESVs between soil and tomato and lettuce roots and of NRPS/PKS-related sequences between roots shotgun metagenomes and metatranscriptomes was determined using DESeq2 (72). Only sequences with a corrected-adjusted P-value <0.1 (Wald test p-values corrected for multiple testing by Benjamini-Hochberg method) were chosen.

Acknowledgments

We thank Avihai Zolti, Asaf Levi, Jonathan Friedman and Noa Sela for providing excellent technical and methodological help.

Funding

This manuscript was supported by the US-Israel Binational Agricultural Research and Development Fund (BARD) Grant #: IS-5177-19F.

Literature Cited

466 1. Lynch J, Whipps J. 1990. Substrate flow in the rhizosphere. Plant and soil 129:1-10.
467 2. Pinton R, Varanini Z, Nannipieri P. 2007. The rhizosphere: biochemistry and organic substances
468 at the soil-plant interface. CRC press.
469 3. Huang X-F, Chaparro JM, Reardon KF, Zhang R, Shen Q, Vivanco JM. 2014. Rhizosphere
470 interactions: root exudates, microbes, and microbial communities. Botany 92:267-275.
471 4. Mendes R, Kruijt M, De Bruijn I, Dekkers E, van der Voort M, Schneider JH, Piceno YM, DeSantis
472 TZ, Andersen GL, Bakker PA. 2011. Deciphering the rhizosphere microbiome for disease-
473 suppressive bacteria. Science:1203980.
474 5. Vacheron J, Desbrosses G, Bouffaud M-L, Touraine B, Moënne-Loccoz Y, Muller D, Legendre L,
475 Wisniewski-Dyé F, Prigent-Combaret C. 2013. Plant growth-promoting rhizobacteria and root
476 system functioning. Frontiers in plant science 4:356.
477 6. Tyc O, Song C, Dickschat JS, Vos M, Garbeva P. 2017. The ecological role of volatile and soluble
478 secondary metabolites produced by soil bacteria. Trends in microbiology 25:280-292.
479 7. Philippot L, Raaijmakers JM, Lemanceau P, Van Der Putten WH. 2013. Going back to the roots:
480 the microbial ecology of the rhizosphere. Nature Reviews Microbiology 11:789.
481 8. Reddy BVB, Kallifidas D, Kim JH, Charlop-Powers Z, Feng Z, Brady SF. 2012. Natural product
482 biosynthetic gene diversity in geographically distinct soil microbiomes. Applied and
483 environmental microbiology:AEM. 00102-12.
484 9. Payne JA, Schoppet M, Hansen MH, Cryle MJ. 2017. Diversity of nature's assembly lines–recent
485 discoveries in non-ribosomal peptide synthesis. Molecular BioSystems 13:9-22.
486 10. Hertweck C. 2009. The biosynthetic logic of polyketide diversity. Angewandte Chemie
487 International Edition 48:4688-4716.
488 11. Newman DJ, Cragg GM. 2016. Natural products as sources of new drugs from 1981 to 2014.
489 Journal of natural products 79:629-661.
490 12. Chowdhury SP, Hartmann A, Gao X, Borriss R. 2015. Biocontrol mechanism by root-associated
491 Bacillus amyloliquefaciens FZB42–a review. Frontiers in microbiology 6:780.
492 13. Pham VH, Kim J. 2012. Cultivation of unculturable soil bacteria. Trends in biotechnology 30:475-
493 484.
494 14. Alain K, Querellou J. 2009. Cultivating the uncultured: limits, advances and future challenges.
495 Extremophiles 13:583-594.
496 15. Abrudan MI, Smakman F, Grimbergen AJ, Westhoff S, Miller EL, van Wezel GP, Rozen DE. 2015.
497 Socially mediated induction and suppression of antibiosis during bacterial coexistence.
498 Proceedings of the National Academy of Sciences 112:11054-11059.
499 16. Crits-Christoph A, Diamond S, Butterfield CN, Thomas BC, Banfield JF. 2018. Novel soil bacteria
500 possess diverse genes for secondary metabolite biosynthesis. Nature 558:440.
501 17. Hover BM, Kim S-H, Katz M, Charlop-Powers Z, Owen JG, Ternei MA, Maniko J, Estrela AB,
502 Molina H, Park S. 2018. Culture-independent discovery of the malacidins as calcium-dependent
503 antibiotics with activity against multidrug-resistant Gram-positive pathogens. Nature
504 microbiology 3:415.
505 18. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, Suarez Duran HG, de Los Santos EL,
506 Kim HU, Nave M. 2017. antiSMASH 4.0—improvements in chemistry prediction and gene cluster
507 boundary identification. Nucleic acids research 45:W36-W41.
508 19. Skinnider MA, Merwin NJ, Johnston CW, Magarvey NA. 2017. PRISM 3: expanded prediction of
509 natural product chemical structures from microbial genomes. Nucleic acids research 45:W49-
510 W54.
511 20. Lemetre C, Maniko J, Charlop-Powers Z, Sparrow B, Lowe AJ, Brady SF. 2017. Bacterial natural
512 product biosynthetic domain composition in soil correlates with changes in latitude on a
513 continent-wide scale. Proceedings of the National Academy of Sciences:201710262.

514 21. Chevrette MG, Gutiérrez-García K, Selem-Mojica N, Aguilar-Martínez C, Yañez-Olvera A, Ramos-
515 Aboites HE, Hoskisson PA, Barona-Gómez F. 2020. Evolutionary dynamics of natural product
516 biosynthesis in bacteria. Natural Product Reports 37:566-599.
517 22. Amos GC, Borsetto C, Laskaris P, Krsek M, Berry AE, Newsham KK, Calvo-Bado L, Pearce DA,
518 Vallin C, Wellington EM. 2015. Designing and implementing an assay for the detection of rare
519 and divergent NRPS and PKS clones in European, Antarctic and Cuban soils. PloS one
520 10:e0138327.
521 23. Charlop-Powers Z, Pregitzer CC, Lemetre C, Ternei MA, Maniko J, Hover BM, Calle PY, McGuire
522 KL, Garbarino J, Forgione HM. 2016. Urban park soil microbiomes are a rich reservoir of natural
523 product biosynthetic diversity. Proceedings of the national academy of sciences 113:14811-
524 14816.
525 24. Dror B, Jurkevitch E, Cytryn E. 2020. State-of-the-art methodologies to identify antimicrobial
526 secondary metabolites in soil bacterial communities-A review. Soil Biology and
527 Biochemistry:107838.
528 25. Kautsar SA, Blin K, Shaw S, Navarro-Muñoz JC, Terlouw BR, van der Hooft JJ, Van Santen JA,
529 Tracanna V, Suarez Duran HG, Pascal Andreu V. 2020. MIBiG 2.0: a repository for biosynthetic
530 gene clusters of known function. Nucleic acids research 48:D454-D458.
531 26. Wu C, Shang Z, Lemetre C, Ternei MA, Brady SF. 2019. Cadasides, calcium-dependent acidic
532 lipopeptides from the soil metagenome that are active against multidrug-resistant bacteria.
533 Journal of the American Chemical Society 141:3910-3919.
534 27. Owen JG, Charlop-Powers Z, Smith AG, Ternei MA, Calle PY, Reddy BVB, Montiel D, Brady SF.
535 2015. Multiplexed metagenome mining using short DNA sequence tags facilitates targeted
536 discovery of epoxyketone proteasome inhibitors. Proceedings of the National Academy of
537 Sciences 112:4221-4226.
538 28. Peek J, Lilic M, Montiel D, Milshteyn A, Woodworth I, Biggins JB, Ternei MA, Calle PY, Danziger
539 M, Warrier T. 2018. Rifamycin congeners kanglemycins are active against rifampicin-resistant
540 bacteria via a distinct mechanism. Nature communications 9:1-15.
541 29. Reddy BVB, Milshteyn A, Charlop-Powers Z, Brady SF. 2014. eSNaPD: a versatile, web-based
542 bioinformatics platform for surveying and mining natural product biosynthetic diversity from
543 metagenomes. Chemistry & biology 21:1023-1033.
544 30. Liu W-T, Lamsa A, Wong WR, Boudreau PD, Kersten R, Peng Y, Moree WJ, Duggan BM, Moore
545 BS, Gerwick WH. 2014. MS/MS-based networking and peptidogenomics guided genome mining
546 revealed the stenothricin gene cluster in Streptomyces roseosporus. The Journal of antibiotics
547 67:99-104.
548 31. Terlain B, Thomas J. 1971. Structure of griselimycin, polypeptide antibiotic extracted
549 Streptomyces cultures. I. Identification of the products liberated by hydrolysis. Bulletin de la
550 Societe chimique de France 6:2349-2356.
551 32. Zolti A, Green SJ, Sela N, Hadar Y, Minz D. 2019. The microbiome as a biosensor: functional
552 profiles elucidate hidden stress in hosts. bioRxiv:752261.
553 33. Bilyk O, Samborskyy M, Leadlay PF. 2019. The biosynthetic pathway to ossamycin, a macrocyclic
554 polyketide bearing a spiroacetal moiety. PloS one 14:e0215958.
555 34. Schmitz H, KE C, KE P, Lein J. 1965. Ossamycin, a new cytotoxic agent. The Journal of Antibiotics,
556 Series A 18:82-88.
557 35. Blodgett JA, Oh D-C, Cao S, Currie CR, Kolter R, Clardy J. 2010. Common biosynthetic origins for
558 polycyclic tetramate macrolactams from phylogenetically diverse bacteria. Proceedings of the
559 National Academy of Sciences 107:11692-11697.
560 36. Hamilton-Miller J. 1973. Chemistry and biology of the polyene macrolide antibiotics.
561 Bacteriological reviews 37:166.

562 37. Chevrette MG, Carlson CM, Ortega HE, Thomas C, Ananiev GE, Barns KJ, Book AJ, Cagnazzo J,
563 Carlos C, Flanigan W. 2019. The antimicrobial potential of Streptomyces from insect
564 microbiomes. Nature communications 10:1-11.
565 38. Wang H, Fewer DP, Holm L, Rouhiainen L, Sivonen K. 2014. Atlas of nonribosomal peptide and
566 polyketide biosynthetic pathways reveals common occurrence of nonmodular enzymes.
567 Proceedings of the National Academy of Sciences 111:9259-9264.
568 39. Vannier N, Agler M, Hacquard S. 2019. Microbiota-mediated disease resistance in plants. PLoS
569 pathogens 15.
570 40. Loper JE, Gross H. 2007. Genomic analysis of antifungal metabolite production by Pseudomonas
571 fluorescens Pf-5, p 265-278, New Perspectives and Approaches in Plant Growth-Promoting
572 Rhizobacteria Research. Springer.
573 41. Tran H, Ficke A, Asiimwe T, Höfte M, Raaijmakers JM. 2007. Role of the cyclic lipopeptide
574 massetolide A in biological control of Phytophthora infestans and in colonization of tomato
575 plants by Pseudomonas fluorescens. New Phytologist 175:731-742.
576 42. Charlop-Powers Z, Owen JG, Reddy BVB, Ternei MA, Guimarães DO, de Frias UA, Pupo MT,
577 Seepe P, Feng Z, Brady SF. 2015. Global biogeographic sampling of bacterial secondary
578 metabolism. Elife 4:e05048.
579 43. Aleti G, Nikolić B, Brader G, Pandey RV, Antonielli L, Pfeiffer S, Oswald A, Sessitsch A. 2017.
580 Secondary metabolite genes encoded by potato rhizosphere microbiomes in the Andean
581 highlands are diverse and vary with sampling site and vegetation stage. Scientific reports 7:2330.
582 44. Uroz S, Buée M, Murat C, Frey‐Klett P, Martin F. 2010. Pyrosequencing reveals a contrasted
583 bacterial diversity between oak rhizosphere and surrounding soil. Environmental Microbiology
584 Reports 2:281-288.
585 45. García‐Salamanca A, Molina‐Henares MA, van Dillewijn P, Solano J, Pizarro‐Tobías P, Roca
586 A, Duque E, Ramos JL. 2013. Bacterial diversity in the rhizosphere of maize and the surrounding
587 carbonate‐rich bulk soil. Microbial biotechnology 6:36-44.
588 46. Yan Y, Kuramae EE, de Hollander M, Klinkhamer PG, van Veen JA. 2017. Functional traits
589 dominate the diversity-related selection of bacterial communities in the rhizosphere. The ISME
590 journal 11:56.
591 47. Zhalnina K, Louie KB, Hao Z, Mansoori N, da Rocha UN, Shi S, Cho H, Karaoz U, Loqué D, Bowen
592 BP. 2018. Dynamic root exudate chemistry and microbial substrate preferences drive patterns in
593 rhizosphere microbial community assembly. Nature microbiology 3:470-480.
594 48. Yang Y, Wang N, Guo X, Zhang Y, Ye B. 2017. Comparative analysis of bacterial community
595 structure in the rhizosphere of maize by high-throughput pyrosequencing. PLoS One
596 12:e0178425.
597 49. Meena VS, Meena SK, Verma JP, Kumar A, Aeron A, Mishra PK, Bisht JK, Pattanayak A, Naveed
598 M, Dotaniya M. 2017. Plant beneficial rhizospheric microorganism (PBRM) strategies to improve
599 nutrients use efficiency: a review. Ecological Engineering 107:8-32.
600 50. Wilkinson SW, Magerøy MH, López Sánchez A, Smith LM, Furci L, Cotton TA, Krokene P, Ton J.
601 2019. Surviving in a hostile world: plant strategies to resist pests and diseases. Annual review of
602 phytopathology 57:505-529.
603 51. Mendes R, Kruijt M, De Bruijn I, Dekkers E, van der Voort M, Schneider JH, Piceno YM, DeSantis
604 TZ, Andersen GL, Bakker PA. 2011. Deciphering the rhizosphere microbiome for disease-
605 suppressive bacteria. Science 332:1097-1100.
606 52. Schlatter D, Kinkel L, Thomashow L, Weller D, Paulitz T. 2017. Disease suppressive soils: new
607 insights from the soil microbiome. Phytopathology 107:1284-1297.

608 53. Bach E, dos Santos Seger GD, de Carvalho Fernandes G, Lisboa BB, Passaglia LMP. 2016.
609 Evaluation of biological control and rhizosphere competence of plant growth promoting
610 bacteria. Applied soil ecology 99:141-149.
611 54. Cheng J, Yang SH, Palaniyandi SA, Han JS, Yoon T-M, Kim T-J, Suh J-W. 2010. Azalomycin F
612 complex is an antifungal substance produced by Streptomyces malaysiensis MJM1968 isolated
613 from agricultural soil. Journal of the Korean Society for Applied Biological Chemistry 53:545-552.
614 55. Liu C, Zhuang X, Yu Z, Wang Z, Wang Y, Guo X, Xiang W, Huang S. 2019. Community structures
615 and antifungal activity of root-associated endophytic actinobacteria of healthy and diseased
616 soybean. Microorganisms 7:243.
617 56. Yamada T, Hamada M, Floreancig P, Nakabachi A. 2019. Diaphorin, a polyketide synthesized by
618 an intracellular symbiont of the Asian citrus psyllid, is potentially harmful for biological control
619 agents. PLoS One 14:e0216319.
620 57. Carrión VJ, Perez-Jaramillo J, Cordovez V, Tracanna V, De Hollander M, Ruiz-Buck D, Mendes LW,
621 van Ijcken WF, Gomez-Exposito R, Elsayed SS. 2019. Pathogen-induced activation of disease-
622 suppressive functions in the endophytic root microbiome. Science 366:606-612.
623 58. Owen JG, Reddy BVB, Ternei MA, Charlop-Powers Z, Calle PY, Kim JH, Brady SF. 2013. Mapping
624 gene clusters within arrayed metagenomic libraries to expand the structural diversity of
625 biomedically relevant natural products. Proceedings of the National Academy of Sciences
626 110:11797-11802.
627 59. Lamari L, Zitouni A, Boudjella H, Badji B, Sabaou N, Lebrihi A, Lefebvre G, Seguin E, Tillequin F.
628 2002. New Dithiolopyrrolone Antibiotics from Saccharothrix sp. SA 233. The Journal of
629 antibiotics 55:696-701.
630 60. te Welscher YM, van Leeuwen MR, de Kruijff B, Dijksterhuis J, Breukink E. 2012. Polyene
631 antibiotic that inhibits membrane transport proteins. Proceedings of the National Academy of
632 Sciences 109:11156-11159.
633 61. Wang H, Li C, Zhang B, He H, Jin P, Wang J, Zhang J, Wang X, Xiang W. 2015. Complete genome
634 sequence of Streptomyces cyaneogriseus ssp. noncyanogenus, the thermotolerant producer of
635 commercial antibiotics nemadectin. Journal of biotechnology 204:1-2.
636 62. Zhao B, Gao Z, Shao Y, Yan J, Hu Y, Yu J, Liu Q, Chen F. 2012. Diversity analysis of type I
637 ketosynthase in rhizosphere soil of cucumber. Journal of basic microbiology 52:224-231.
638 63. Zolti A, Green SJ, Mordechay EB, Hadar Y, Minz D. 2019. Root microbiome response to treated
639 wastewater irrigation. Science of The Total Environment 655:899-907.
640 64. Bolyen E, Rideout JR, Dillon MR, Bokulich NA, Abnet C, Al-Ghalith GA, Alexander H, Alm EJ,
641 Arumugam M, Asnicar F. 2018. QIIME 2: Reproducible, interactive, scalable, and extensible
642 microbiome data science. PeerJ Preprints,
643 65. Huson DH, Beier S, Flade I, Górska A, El-Hadidi M, Mitra S, Ruscheweyh H-J, Tappu R. 2016.
644 MEGAN community edition-interactive exploration and analysis of large-scale microbiome
645 sequencing data. PLoS computational biology 12:e1004957.
646 66. Brady SF. 2007. Construction of soil environmental DNA cosmid libraries and screening for
647 clones that produce biologically active small molecules. Nature protocols 2:1297.
648 67. Edgar RC. 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics
649 26:2460-2461.
650 68. Zhang T, Luo Y, Chen Y, Li X, Yu J. 2012. BIGrat: a repeat resolver for pyrosequencing-based re-
651 sequencing with Newbler. BMC research notes 5:567.
652 69. Blin K, Shaw S, Steinke K, Villebro R, Ziemert N, Lee SY, Medema MH, Weber T. 2019. antiSMASH
653 5.0: updates to the secondary metabolite genome mining pipeline. Nucleic acids research
654 47:W81-W87.
655 70. Kolde R, Kolde MR. 2015. Package 'pheatmap'. R Package 1.

71. Oksanen J, Kindt R, Legendre P, O'Hara B, Stevens MHH, Oksanen MJ, Suggests M. 2007. The vegan package. Community ecology package 10:631-637.
72. Love M, Anders S, Huber M. 2014. Differential gene expression analysis based on the negative binomial distribution. Genome Biol 15:550.

Tables legend

Table 1. Abundance of NRPS (AD) and PKS (KS) amplicons based on levels of similarity to MIBiG NRPSs and PKSs reference genes, in the tomato and cucumber root and tomato soil samples. Identity groups were determined as unassigned (< 50% amino acid sequence identity), 50-70%, 70-85%, 85-95% and 95-100% sequence identity. Numbers outside parentheses indicate % of sequences associated with each group, numbers inside parentheses indicate total number of hits for each group.

Table 2. Predicted taxonomy and source dataset of recovered cosmids. UA- unassigned, only affiliated organisms with >80% identity are shown.

Figures legend

Figure 1. Principal coordinate analysis (PCoA) of lettuce roots (green circles), tomato roots (red squares) and soil (from tomato lysimeters, brown triangles) samples using the Bray-Curtis similarity index. (A) Analysis of KS (from PKS) domain amplicons (roots vs. soil ANOSIM statistic R=0.332, p-Value<0.05, tomato vs. lettuce vs. soil ANOSIM R=0.432, NS). (B)

Analysis of AD (NRPS) domain amplicons (Roots vs. soil ANOSIM statistic R=0.308, p-Value<0.01, tomato vs. lettuce vs. soil ANOSIM R=0.6872, p-Value<0.01).

Figure 2. Differential abundance of the top20 highly abundant NRPS and PKS amplicons in tomato (A) and lettuce (B) root microbial communities relative to bulk tomato soil. DESeq 2 was used to calculate log2 fold changes and test for significance (adjusted p-value (FDR) of <0.1). Associated metabolites (based on the MIBiG repository) are presented only for sequences with e-Value<10-40 and >50% AA identity. Identity bars represent amino acid identity to MIBiG reference sequence (0-not identified).

Figure 3. Relative abundance of NRPS (A) and PKS (B) associated metabolites based on MIBiG annotations. Only AD and KS amplicon hits to reference MIBiG sequences over >50% amino acid identity and e-Value<10-40 were included. Counts were normalized and log10-transformed. Only associated metabolites that were present in >2 different samples are shown.

Figure 4. Relative abundance and taxonomy of highly abundant NRPSs/PKSs in tomato and lettuce roots metagenome (MG) and metatranscriptome (MT). Shapes represent BGC-associated chemical class. Taxonomy was assigned using MEGAN 5.10 and the lowest common ancestor (LCA) classification algorithm. Only low CV sequences were chosen.

Figure 5. Overview of retrieved (from archived soil cosmid libraries) biosynthetic gene clusters containing AD and KS domain contigs that were abundant and/or highly expressed in tomato and cucumber root microbiomes.

Figure 6. A) Comparison of B481 cluster and the related BGC in *S. cyaneogriseus.* Gray area indicate the aligned region. B) Structure prediction for PKS modules present on clone B893. Examples of known polyene antifungal agents are shown below.

Figure 7. A) Relative abundance of clones-recovered AD and KS domains in shotgun metagenomes (n=20) from 4 environmental ecosystems- gut (animal and human), aquatic (freshwater and marine), soil and root-associated. Analysis was conducted using IMG BlastN feature with eValue<10-5 and identity >85%. Counts were normalized with rpoB and gyrB for each dataset. B) Presence/absence of each recovered gene cluster in the different environment. Only gene clusters where all AD/KS domains were found are considered present.

Supplemental material legend

Figure S1. Alpha diversity (1-Simpson similarity index) of soil and root (tomato and lettuce) samples. **A and B**- AD sequences (NRPS). **C and D**- KS sequences (PKS). Asterisks represent statistical significance between samples (by Tukey HSD, *- pValue<0.001, **- pValue<0.0001, NS- not significant).

Figure S2. Relative abundance of NRPS (A) and PKS (B) associated metabolites based on MIBiG annotations. Only AD and KS amplicon hits to reference MIBiG sequences over >50% amino acid identity and e-Value<10^{-40} were included. Counts were normalized and log10-transformed. Only associated metabolites that were present in >2 different samples are shown.

Figure S3. The fifty most abundant NRPSs and PKSs in tomato (A) and lettuce (B) root microbiomes extrapolated from shotgun metagenomic data. Relative abundance represent the

mean of three repeats per sample. CV (coefficient of variance) was determined as STD/Mean per sequence. Green-circled points represent sequences that are highly abundant in the corresponding meta-transcriptome datasets per sample type indicating that these are also highly expressed.

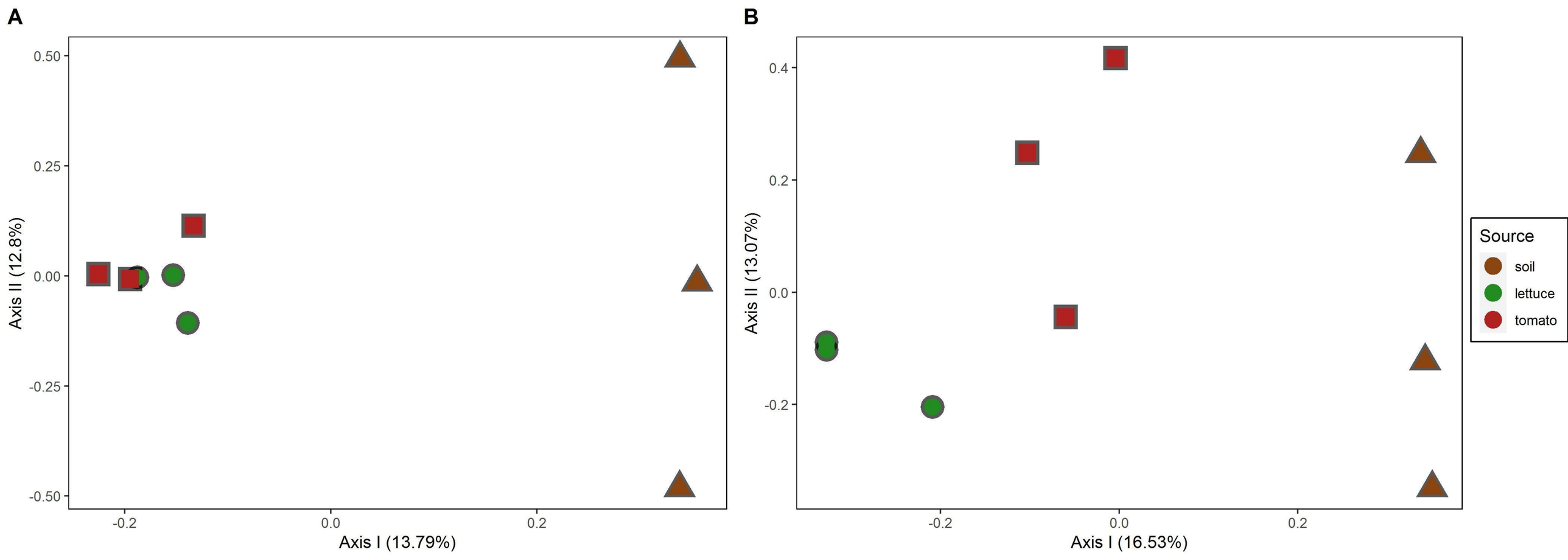
igure S4. Mined NRPSs and PKSs from amplified AD and KS domains in tomato and lettuce roots, tomato metagenome and lettuce metagenome. AS-amplicon sequencing, MG-metagenome. Numbers in circles represent number of unique MIBiG-derived genes found within each dataset, while numbers in parentheses represent percent of unique genes from total.

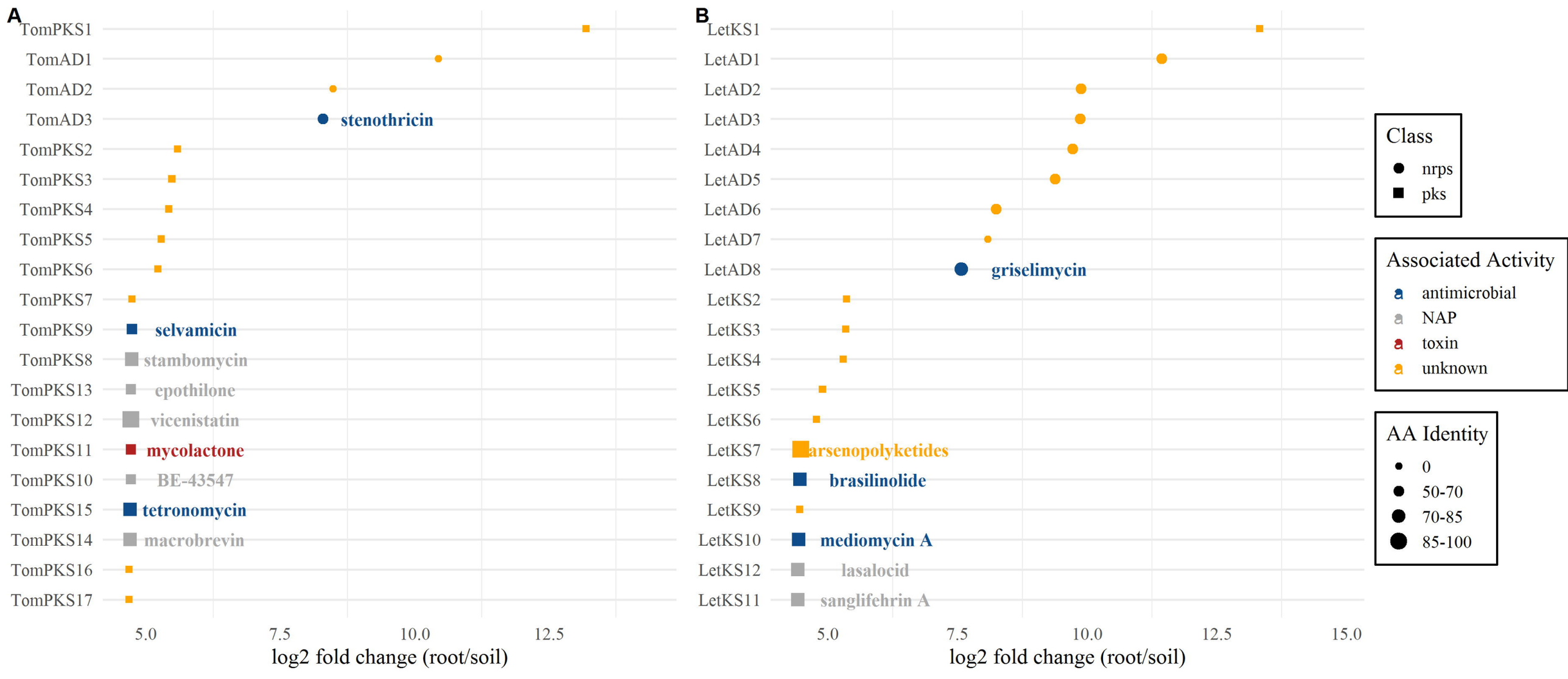
Figure S5. Relative abundance of recovered gene clusters in root-associated and soil metagenomic samples. Abundance was normalized to gyrB and rpoB housekeeping genes within each dataset (n=20 per environment). Statistical difference was calculated using Wilcoxon non-parametric test. *= $p<0.05$, NS- not-significant

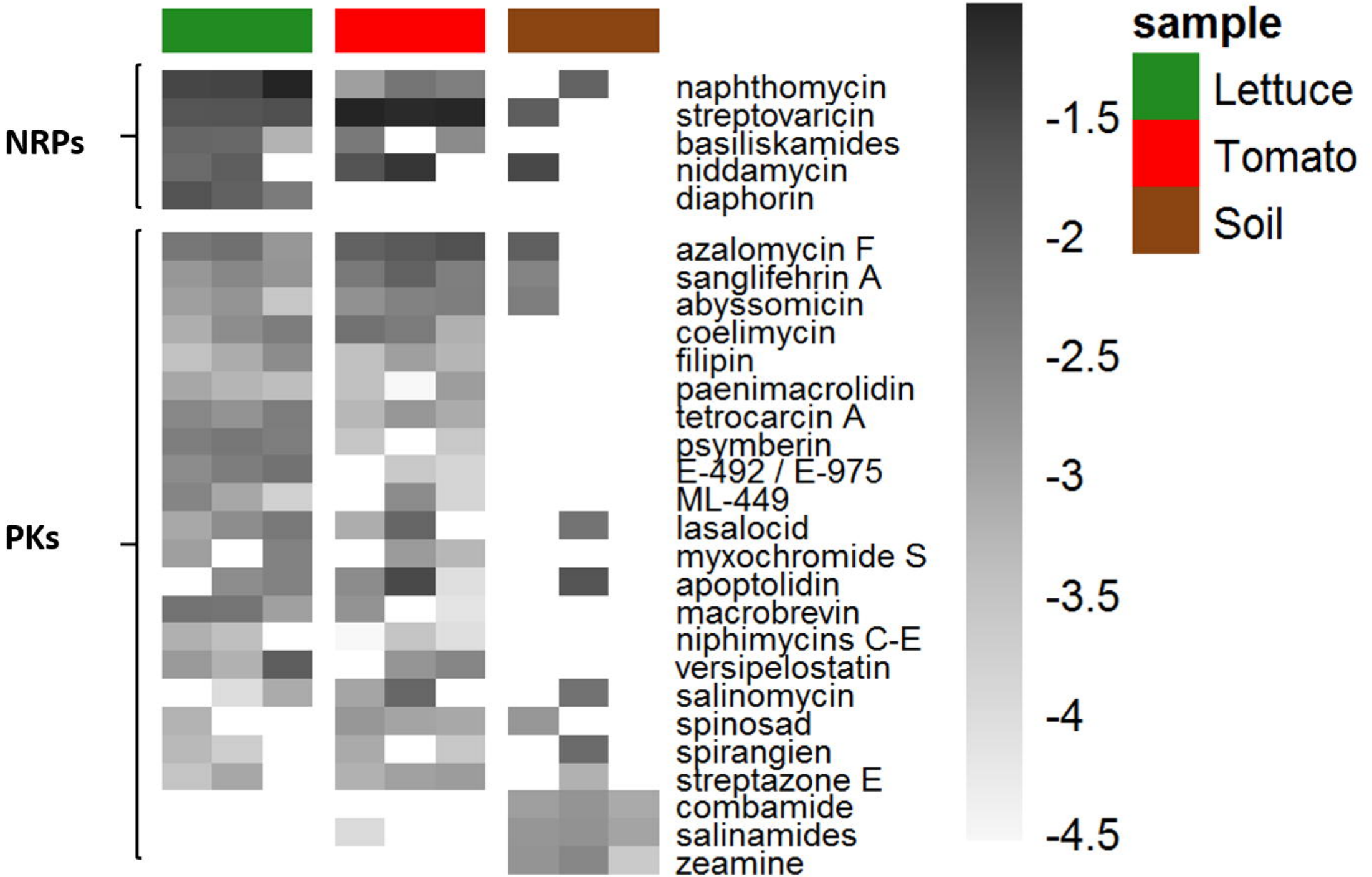
Figure S6. Study conceptual pipeline scheme

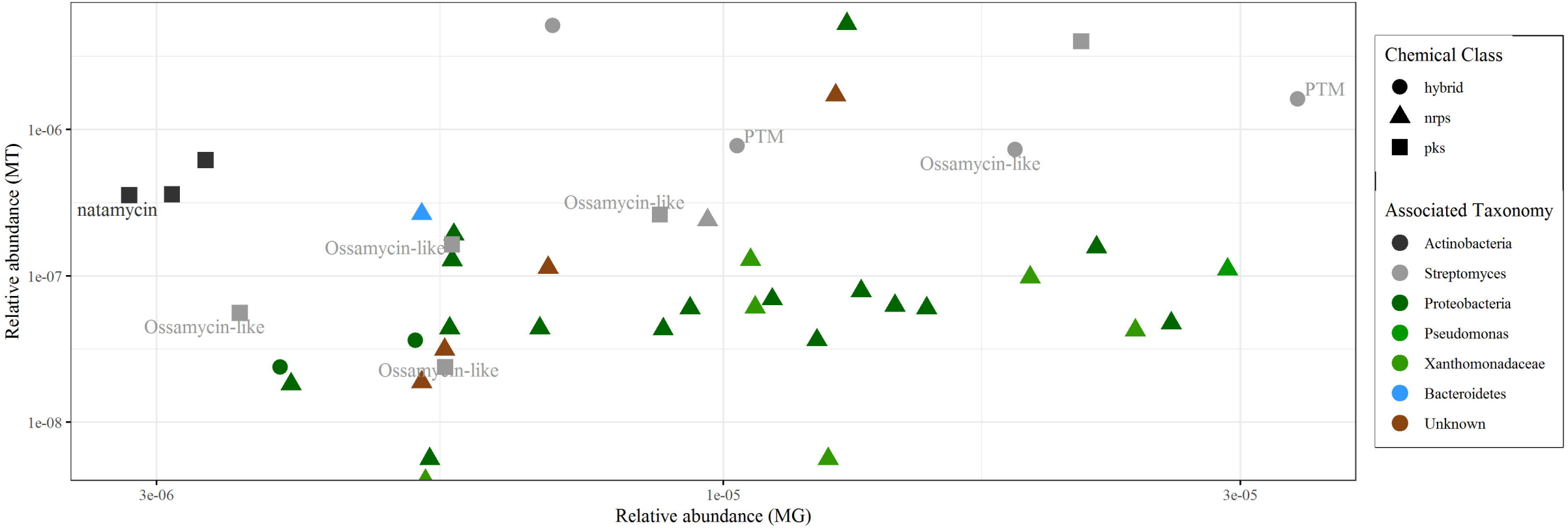
Table S1. Amplicon sequencing processing information.

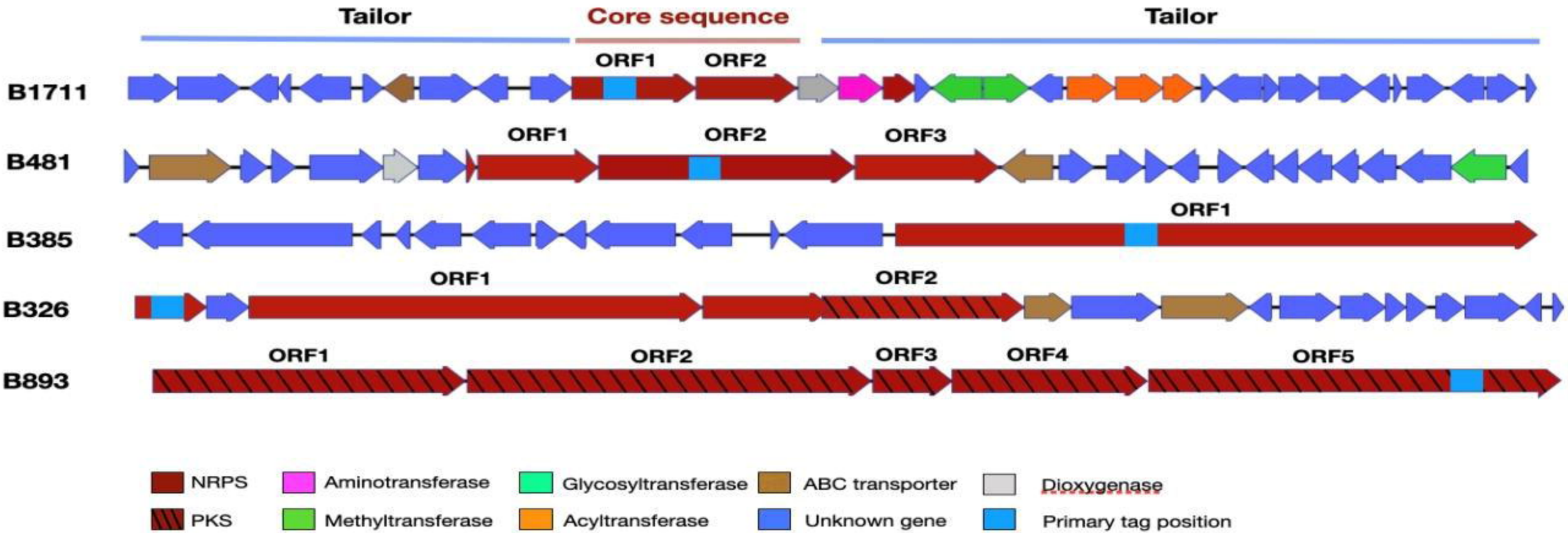
Table S2. Public shotgun metagenome of microbial communities from four different ecosystems (animal and human gut, aquatic, soil and root-associated).

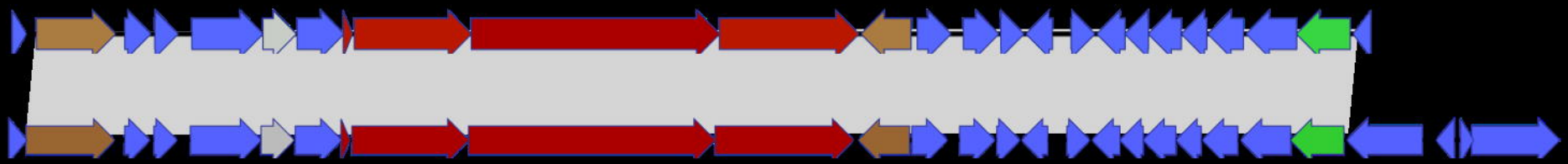

A
0.50
0.25
0.00
-0.25
-0.50
Axis II (12.8%)
-0.2
0.0
0.2
Axis I (13.79%)
B
0.4
0.2
0.0
-0.2
Axis II (13.07%)
-0.2
0.0
0.2
Axis I (16.53%)
Source
soil
lettuce
tomato

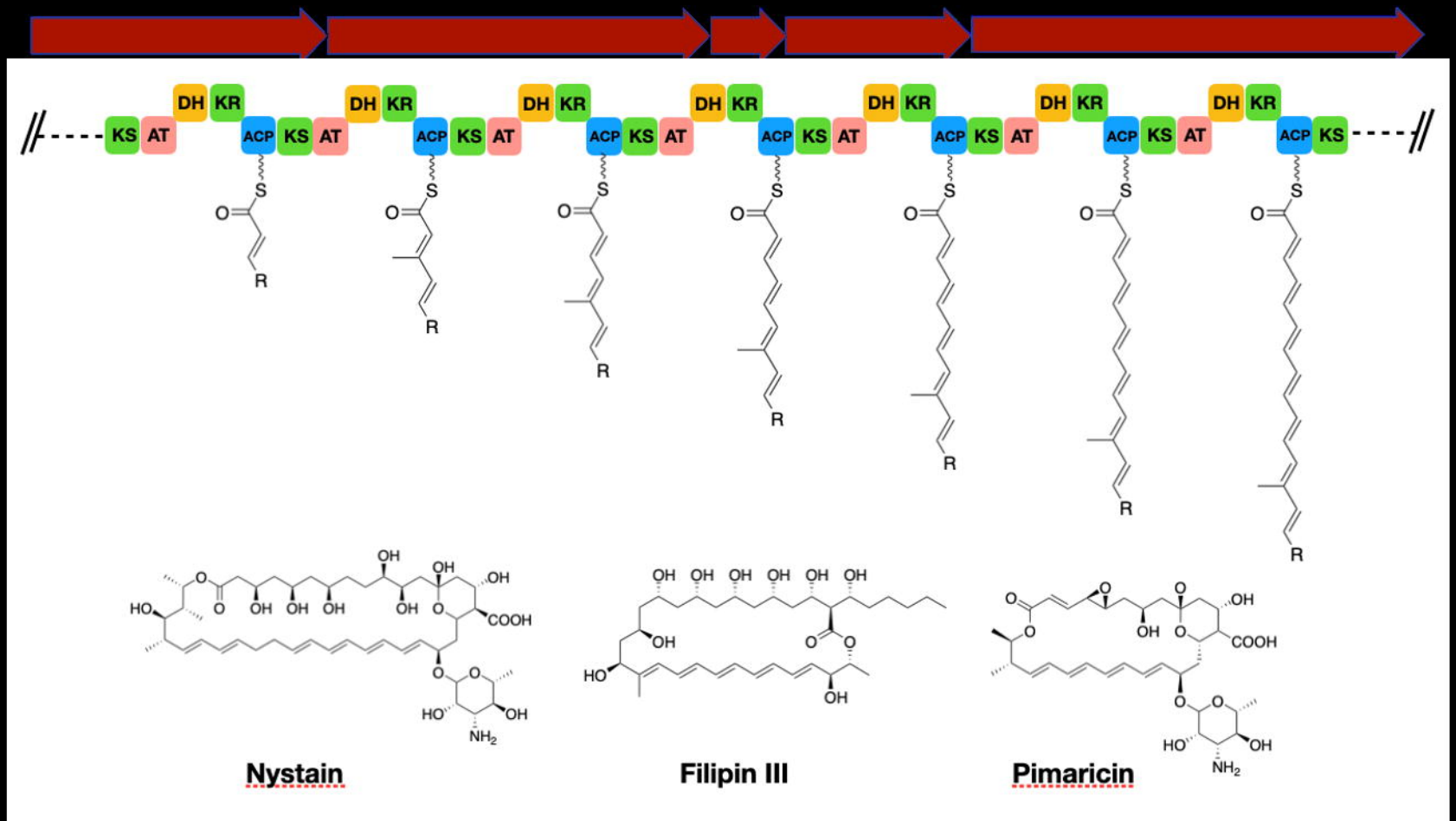

A
TomPKS1
TomAD1
TomAD2
TomAD3
TomPKS2
TomPKS3
TomPKS4
TomPKS5
TomPKS6
TomPKS7
TomPKS9
TomPKS8
TomPKS13
TomPKS12
TomPKS11
TomPKS10
TomPKS15
TomPKS14
TomPKS16
TomPKS17
stenothricin
selvamicin
stambomycin
epothilone
vicenistatin
mycolactone
BE-43547
tetronomycin
macrobrevin
5.0
7.5
10.0
12.5
log2 fold change (root/soil)
B
LetKS1
LetAD1
LetAD2
LetAD3
LetAD4
LetAD5
LetAD6
LetAD7
LetAD8
LetKS2
LetKS3
LetKS4
LetKS5
LetKS6
LetKS7
LetKS8
LetKS9
LetKS10
LetKS12
LetKS11
griselimycin
arsenopolyketides
brasilinolide
mediomycin A
lasalocid
sanglifehrin A
5.0
7.5
10.0
12.5
15.0
log2 fold change (root/soil)
Class
nrps
pks
Associated Activity
antimicrobial
NAP
toxin
unknown
AA Identity
0
50-70
70-85
85-100

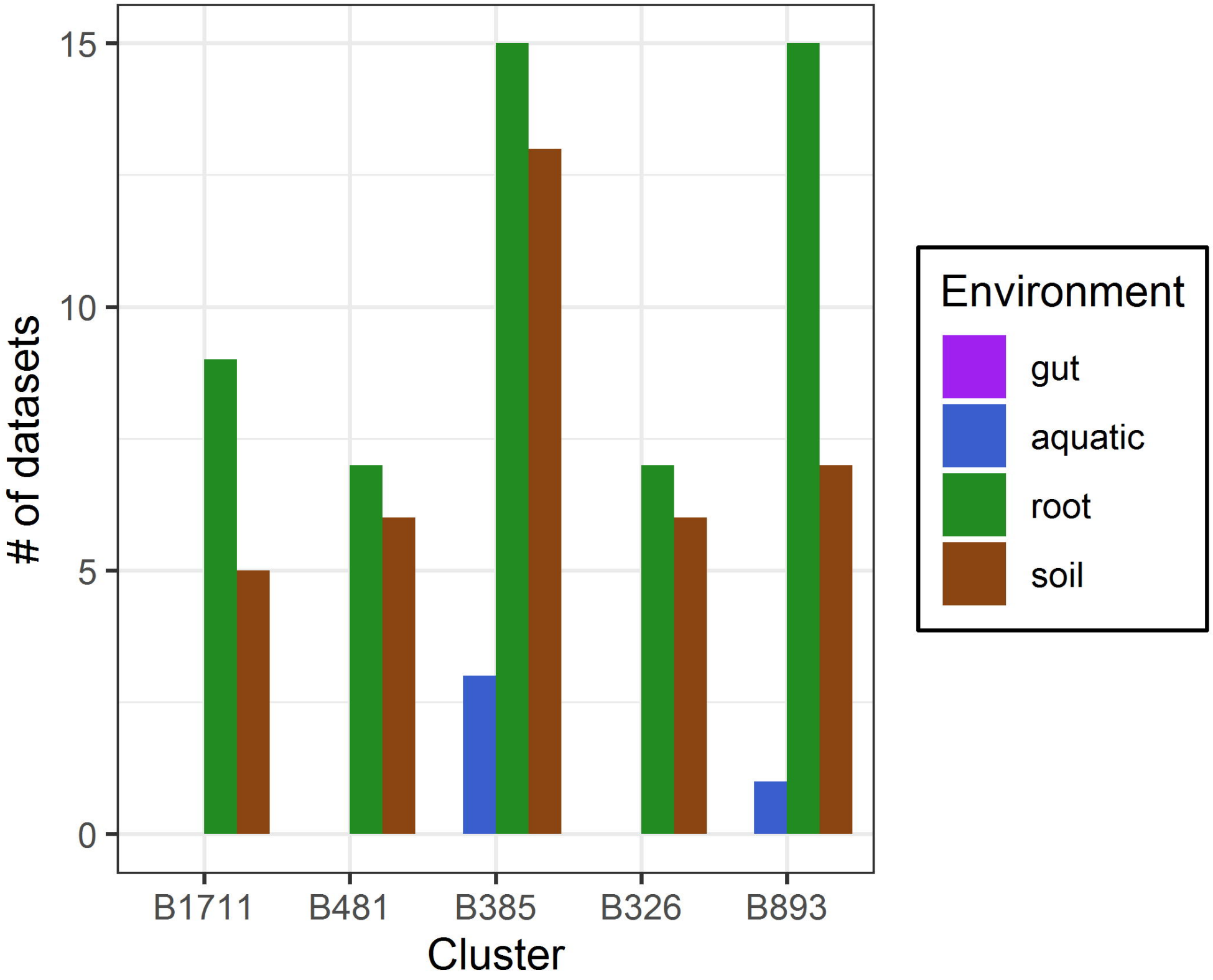

sample
Lettuce
Tomato
Soil
NRPs
PKs
naphthomycin
streptovaricin
basiliskamides
niddamycin
diaphorin
azalomycin F
sanglifehrin A
abyssomicin
coelimycin
filipin
paenimacrolidin
tetrocarcin A
psymberin
E-492 / E-975
ML-449
lasalocid
myxochromide S
apoptolidin
macrobrevin
niphimycins C-E
versipelostatin
salinomycin
spinosad
spirangien
streptazone E
combamide
salinamides
zeamine
-1.5
-2
-2.5
-3
-3.5
-4
-4.5

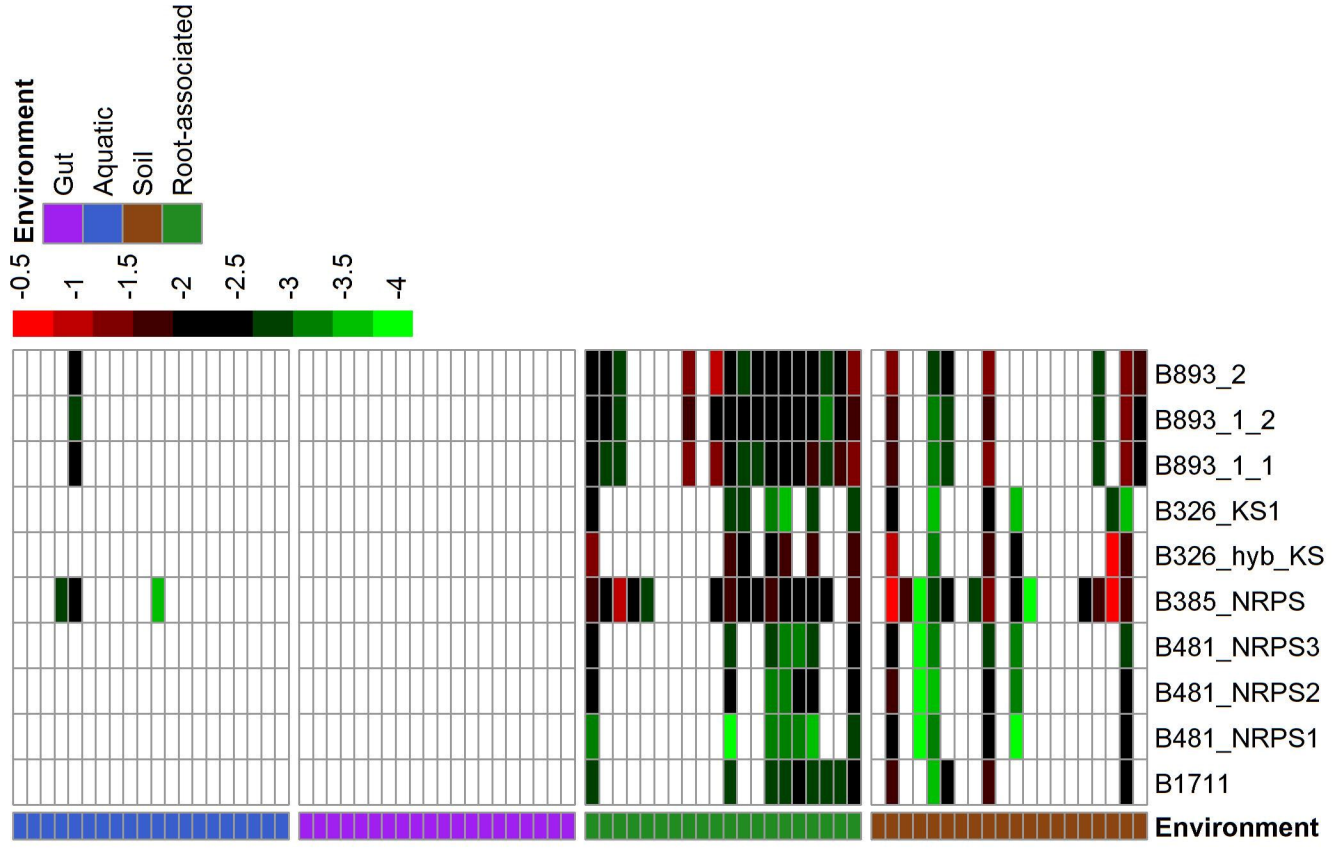

Relative abundance (MT)
1e-06
1e-07
1e-08
Relative abundance (MG)
3e-06
1e-05
3e-05
natamycin
PTM
PTM
Ossamycin-like
Ossamycin-like
Ossamycin-like
Ossamycin-like
Ossamycin-like
Chemical Class
hybrid
nrps
pks
Associated Taxonomy
Actinobacteria
Streptomyces
Proteobacteria
Pseudomonas
Xanthomonadaceae
Bacteroidetes
Unknown


Tailor
Core sequence
Tailor
B1711
ORF1
ORF2
B481
ORF1
ORF2
ORF3
B385
ORF1
B326
ORF1
ORF2
B893
ORF1
ORF2
ORF3
ORF4
ORF5
NRPS
PKS
Aminotransferase
Methyltransferase
Glycosyltransferase
Acyltransferase
ABC transporter
Unknown gene
Dioxygenase
Primary tag position




KS
AT
DH
KR
ACP
S
O
R
HO
OH
COOH
NH2
Nystain
Filipin III
Pimaricin


15
10
5
0
# of datasets
B1711
B481
B385
B326
B893
Cluster
Environment
gut
aquatic
root
soil


Environment
Gut
Aquatic
Soil
Root-associated
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
B893_2
B893_1_2
B893_1_1
B326_KS1
B326_hyb_KS
B385_NRPS
B481_NRPS3
B481_NRPS2
B481_NRPS1
B1711
Environment