Functional analyses of the rhizosphere microbiome in strigolactone-deficient tomatoes

Susana López Lemarroy

VU: 2730794 UvA: 14108550

Supervisor: Dr. Anna Heintz-Buschart
Academic Advisor: Dr. Marten Postma
Internship Host: Swammerdam Institute for Life Sciences (SILS) at University of Amsterdam (UvA)
Duration: September 2022 to January 2023

A thesis presented for the degree of
Masters of Science in Bioinformatics and Systems Biology

Faculty of Science
Vrije Universiteit van Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands

Faculty of Natural Sciences Mathematics and Computer Science
Universiteit van Amsterdam
Science Park 904, 1098 XH Amsterdam
The Netherlands

Abstract

Strigolactones are plant hormones involved in the interaction of plant roots with the rhizosphere microbiome, influencing plant growth and development. Their biosynthesis and secretion are upregulated under phosphate starvation conditions since they play a role in the recruitment of beneficial microorganisms to ease the stress of the condition (Zhang et al. (2015)). Shotgun metagenome sequencing data was obtained from tomato rhizosphere under phosphate starvation, where the plants' expression of hormone-synthetizing enzymes was altered through VIGS, targeting genes in the strigolactone biosynthesis gene cluster. The targeted genes include P450 CYP712, P450 CYP722C, and CCD8, which result in concentration fluxes of the various strigolactone products. This is relevant to investigate the effect of tomato strigolactones on the microbial communities of the rhizosphere and in return their functional potential to impact the plant. The latter accomplished by secondary metabolite BGC analysis through antiSMASH6 (Blin et al. (2021)). Data processing was performed through the IMP3 pipeline (Narayanasamy et al. (2016)) to obtain contigs from *de novo* assembly, to annotate and count mapped read counts, followed by statistical analyses between the treatment lines of the functional profiles built from the counts. The analysis methods revealed that the CYP712 silenced line results in the most differential abundant KOs from the KEGG annotation. The main differential abundance methods used were DESeq2, MaAslin2, and ANCOM2, which revealed an overlap of some annotated KOs and CAZymes which were mostly involved in energy metabolism, biosynthesis of amino acids, nucleotide sugars, and cofactors for the former, and depolymerization and recycling cell wall polysaccharides for the latter (Amengual et al. (2022)). Finally for the potential BGCs in the microbiome, antiSMASH6 results indicated the differential presence of siderophore and betalactone types of clusters, which are involved in iron transport into the cell for growth and survival (Saha et al. (2016)), and regulation of root development and phosphate intake, respectively. Further research to understand the effect of the microbial diversity change in the plant could include treatment of bulk-soil with the strigolactones involved in the observed differences, in this case Orobanchol.

Introduction

The microbial diversity in habitats, like the soil in close proximity to the roots (rhizosphere) or the tissue inside the roots (endosphere), is highly influential to the health and growth of a plant (Bulgarelli et al. (2013)). The recruitment of microorganisms to the rhizosphere of the plant is accomplished through root exudates that serve as signaling ligands, as well as nutrients. Recognizing these signals is an evolved mechanism by the microorganisms to localize close to the plant. The microorganisms themselves are specifically recruited because of their functional benefits to the plant; their metabolisms can aid plants to obtain nutrients from the soil that would otherwise be difficult to attain (Reinhold-Hurek et al. (2015)). Previous experimental studies using rRNA analysis have demonstrated that the microbiome diversity is differentially enriched at the genus level between rhizosphere soil compared to bulk-soil (where the soil is not associated with plant roots). While the type and conditions of the soil have a greater impact on soil microbial diversity, plant species have been found to have significant influence in the 'active' microbiome composition in the rhizosphere. Increasing influence and specialization towards the inside of the root, resulting in a more host-specific community composition for the endosphere (Reinhold-Hurek et al. (2015)).

Plants produce primary metabolites such as amino acids or carbohydrates which have key roles in the physiology of the plant and also its interaction with the environment. Plants, as well as microorganisms, produce other natural products classified as *secondary metabolites*. These are low molecular weight products with various functions and external impacts. These include human health benefits, from anti-tumor to cholesterol-lowering and antibiotic effects (Medema et al. (2011)). For which they have been studied due to their potential to develop medicines. They also participate in the interaction between plants and microorganisms, as they can provide protection from pathogens and pests or can be substrates to ease stress from environmental conditions. In microbes their study has been historically experimental, and therefore limited to microorganisms that can be cultured, including isolation from cultures, purification and characterization (Blin et al. (2021)). However, advances in sequencing efforts and the lowering costs of sequencing microbial genomes has allowed for the application of other methods such as genome mining and analysis of metagenome data to identify the biosynthesis pathways involved in their production. Furthermore, secondary metabolites biosynthesis pathways genes are found in clusters of two or more genes that encode the production of the metabolite. These are called Biosynthesis Gene Clusters (BGCs) (Medema et al. (2015)). They follow a characteristic organization of genes with regulatory domains, as well as core motifs that allow their identification and classification into the known classes of secondary metabolites. Annotation and classification of new gene cluster types and their metabolite products can shed light onto the chemicals at play in important interactions between plants and microorganisms. Finally uncovering the functional capabilities of the soil microbiome according to its diverse composition.

Secondary metabolites secreted by plants can have multiple functions in the interaction between the plant and the (biotic and abiotic) environment. Plant hormones (phytohormones) are an important class of metabolites involved in the interaction between plants and microbes, influencing plant growth and development. Strigolactones, a recently identified class of phytohormones, are specifically important during plant phosphorous

and nitrogen starvation (Zhang et al. (2015)). Their biosynthesis and secretion are upregulated under these conditions as they play a role in recruiting beneficial microorganisms to ease these stresses. They have various other roles involving plant functions and the rhizosphere, with some very beneficial to their growth and other not so beneficial such as stimulation of root parasitic microbe germination. Strigolactones have a wide variety of structures and stereoisomers, each with a different potential for combinations of their varied functions (Wang and Bouwmeester (2018)). The intermediates of the biosynthesis pathway also have their own function as a type of strigolactone. Identifying the isomers that have a positive functional impact on the organisms could contribute to define the plant's potential to shape the microbial habitat in the rhizosphere.

Apocarotenoids are carotenoid-derived compounds resulting from a cleavage event at any double carbon bond in the carotenoid backbone. This event is an oxidation that can be catalyzed by an enzyme or can happen nonenzymatically. The products depend on the carbon bonds oxidized as it can result in multiple products of different lengths (Al-Babili and Bouwmeester (2015)). Strigolactones are downstream from the 9-cis-β-apo-10'-carotenal product of carotenoid cleavage oxygenase (CCO) CCD7 (Alder et al. (2012)). **Figure 1** shows the biosynthesis pathway including characterized enzymes involved in catalyzing different structures and strigolactone stereoisomers throughout the pathway. The pathway for strigolactones starts at the CCD8 enzyme which converts the apocarotenoid into Carlactone (CL). This step is crucial for synthesis of all strigolactones and disruption would result in low levels of strigolactones. Carlactone is then converted into Carlactonioc acid (CLA) by MAX1, belonging to the P450 CYP711A plant superfamily. Homologs of MAX1 have been found to perform enzymatic steps downstream, such as the 2 step process to get to Orobanchol via 4DO (Al-Babili and Bouwmeester (2015)). A less traditional mechanism is catalysis by P450 CYP722C directly converting CLA to Orobanchol, which was characterized by Wakabayashi et al. (2019). Disruption of the gene encoding this enzyme could result in accumulation of CLA and low concentrations of downstream strigolactones, but not as low as if MAX1 was disrupted since there is an alternate path to obtain Orobanchol. Another similar enzyme of the P450 superfamily is CYP712 which has been hypothesized for a couple of steps in the pathway by Wang et al. (2022), as downstream of Orobanchol and its disruption would potentially result in accumulation of Orobanchol and therefore a different signaling response. The focus of this study is on the effect that disruption of the strigolactone biosynthesis pathway has on the microbial diversity of the rhizosphere. Gene silencing was performed for CCD8, P450 CY=722C, and P450 CYP712, which results in a change of biosynthesis product fluxes.

Metagenomics is a large array of widely used techniques for microbiome analysis. It has become increasingly accessible with the emergence of Next Generation Sequencing (NGS), providing a description of the combined genomes present in the totality of the sample. This allows further exploration into the functional potential of the microbial community in the samples, including the non-cultured microorganisms (Zancarini et al. (2021)). Furthermore, using an automated bioinformatics pipeline to incorporate all processing of the metagenomic data is crucial for handling such large datasets and for making the results

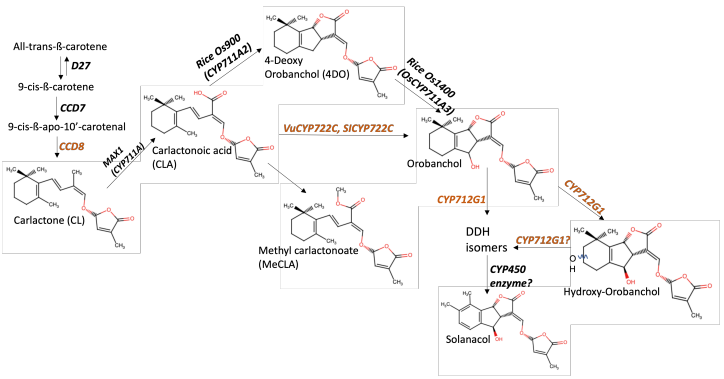

Figure 1 Biosynthesis pathway for Strigolactones from apocarotenoid 9-cis-β-apo-10'-carotenal. Colored steps indicate the enzymes that were targeted in order to generate the different experimental treatment lines.

reproducible. The Integrated Meta-omic Pipeline (IMP3) is a pipeline for the *de novo* assembly, annotation and analysis of metagenomic data. Reference-independent or *de novo* assembly methods, such as MEGAHIT (Li et al. (2015)), allow for novel microbial genomic features to be found since the assembly process does not rely on aligning to a reference. Reference-based methods frequently result in discarding reads that are not close enough to the chosen reference. *De-novo* assemblies enable the reconstruction of actual genomes from the samples as well as increasing mapping rates for further analysis (Narayanasamy et al. (2016)). Additionally, for specific analyses, tools like antibiotics & Secondary Metabolite Analysis Shell (antiSMASH6) can be applied to the *de novo* assembly. antiSMASH6 is a pipeline in itself, focused on identification and annotation of genome sequences for secondary metabolite biosynthesis gene clusters. The antiSMASH6 methodology relies on homology to identify new BGCs, by using a database built based on aligning the potential regions to their nearest relatives of known gene clusters. In the handled regions whole clusters of genes with a core motif are identified and classified as a type of secondary metabolite BGC. Furthermore, it uses the homology database to visualize in a phylogenetic tree the location of the gene cluster among its closest relatives, enabling rapid prediction of function. Finally, with the same database of gene clusters it constructs secondary metabolism Clusters of Orthologous Groups (smCOGs). Using profile Hidden Markov Models (pHMMs) each gene is classified into a secondary-metabolism-specific gene family specific for the conserved sequence region (Blin et al. (2021)). This is all for the goal of giving a prediction of function and phylogenetic analysis of genes. It can be expected that the long contiguous sequences (contigs) from IMP3 enable a more in depth search by the antiSMASH6 pipeline, recovering novel BGCs that would otherwise not be annotated or identified as a whole. For example, without assembly the read-level analysis would be limited to the reference-based alignment and what was already annotated on that.

antiSMASH6 is the golden standard of its category (Blin et al. (2021)) and its combination with IMP3 can elucidate the types of microbial functions involved in the interactions between plants and microorganisms. The question that set the direction

for the research, and guided the analysis and interpretation of the results throughout the study is: *How do the antiSMASH6 implemented IMP3 pipeline results improve or broaden the functional analysis of the rhizosphere microbiome* and *What is the effect of tomato strigolactones on the microbial communities of the rhizosphere and in particular on their functional potential to impact the plant?* The expected results include: to find differentiating functional profiles between the types of samples since the different strigolactones found in the samples should have different levels of concentration, resulting in varying aggregation of microorganisms into the rhizosphere.

Materials & Methods

Processing and analyzing the samples was focused on the presented enzymes involved in the biosynthesis pathway of strigolactones. The bioinformatics tools implemented allowed for a metagenomics complete analysis from sequenced reads, through assembly and annotation, to secondary metabolite biosynthesis gene cluster search.

Plant Experimental & Metagenomics Sequencing

An experiment was designed to observe the differences in microbial diversity through functional profiles as a result of strigolactones differing fluxes and differing structures, from phosphate deficient plant growing conditions. The plant sample data was provided by the Plant Hormone Biology laboratory of Dr. Harro Bouwmeester at UvA's SILS. The samples were from 6 different tomato plant lines (**Table** 1). Tomato seeds were sowen, and seedlings were transferred after 4 days to a growth medium, and in turn transferred after 2 days to a 3:1:1 clay, sand and compost soil mix, watered with demineralized water. The plants were treated with virus-induced gene silencing (VIGS) for the target genes. Phosphate starvation was introduced after VIGS by watering with half-strength Hoagland solution media and substitution of 0.4mM $K_2HPO_4 * 2H_2O$ to 0.8mM KCl. After 5 weeks of this treatment the plants were harvested and separated into: roots, root exudate, and rhizosphere. First the root exudate was obtained during the process of separating the roots from the rhizosphere, in order to perform strigolactones analysis. DNA was extracted from the rhizosphere for shotgun metagenomics sequencing. The sequencing was performed using DNBSEQ-T7 technology, with an average of 600,000,000 reads per sample. Gene silencing was controlled using RNA from the roots. The experimental design and sample manipulation was performed by PhD candidate Bora Kim.

The rhizosphere metagenome data consisted of 3 experimental treatment lines and 3 controls. The GUS line is a negative control for the transformation, since an empty vector was used. The MMA line is another negative control for the transformation since the plants were exposed to transformation solution but no vector was included. Comparison of GUS and MMA can indicate any effect that the vector might have that is not related to the gene silencing objective. Finally, the EP is only bulk soil, without the plant, to have a baseline for microbial diversity.

In addition, exudate data and information on plant were available. The variables measured include: Shoot FreshWeight (grams), Root FreshWeight (grams), Orobanchol concentration (pmol/ml·g), Solanacol concentration (pmol/ml·g), Hydroxy-Orobanchol response (/ml·g), and Didehydro-Orobanchol response (/ml·g). From this data, 712 differed strongly by having higher Orobanchol levels, while the other treatments showed similar levels between each other.

Table 1 Tomato lines and number of replicates. The objective of each line in terms of strigolactones (SLs) levels secreted into the rhizosphere is shown.

Line	# Replicates	Effect
712	11	More Orobanchol
722	10	More CLA
CCD8	11	Less SLs
GUS	10	Normal SLs (vector but no transformation)
MMA	8	Normal SLs (mock transformation/no vector)
EP	4	Bulk-soil (no tomato)

Metagenomics Data Processing in IMP3

IMP3 (Narayanasamy et al. (2016)) is implemented in the workflow manager Snakemake (Mölder et al. (2021), https://imp3.readthedocs.io/en/latest/index.html). Custom configuration files per sample were provided for the following steps of the pipeline (**Figure 2**):

Preprocessing & Taxonomy. These steps were performed by PhD candidate Bora Kim. The raw reads were entered as paired-end reads in fastq format and filtered using adapter sequences for the Illumina Paired-End (PE) sequencing, as well as the host genome from the tomato4mp reference (https://solgenomics.net/ftp/tomato_genome/assembly/build_4.00/). For the taxonomy, the Kraken2 (Wood et al. (2019)) tool was used, utilizing the 'kraken_pfp8' parameter, which uses Kmers indexing to search for and classify query sequences into a taxonomy class. Other specific read lengths and quality filtering parameters can be found in the configuration files available in the documentation repository (**REPOSITORY LINK**).

Assembly & Mapping MEGAHIT was applied for *de novo* assembly for metagenomic data (Li et al. (2015)). The parameters used include minimum kmer size of 25, maximum kmer size of 99, and step of 4 for increment of kmer size per iteration. This step generated contigs from the metagenomic data through a Bruijn graph-based approach. The output included FASTA files containing the contigs assembled which were in turn ran through mapping and annotation. Mapping of processed reads to the assembled contigs was performed using BWA (Li and Durbin (2010)), yielding an alignment in .bam format.

Annotation. The software used in the annotation step of the pipeline is prokka (Seemann (2014)) which is a software tool for the rapid annotation of prokaryotic genomes. It is able to predict genes, tRNAs, rRNAs, and other functional elements from the assembled contigs. For coding sequences (CDS) it uses Prodigal_v2.6.3 (Hyatt et al. (2010)). The output of this step includes fasta files with the predicted gene sequences and a GFF (General Feature Format) file (`annotation.filt.gff`), including all identified elements

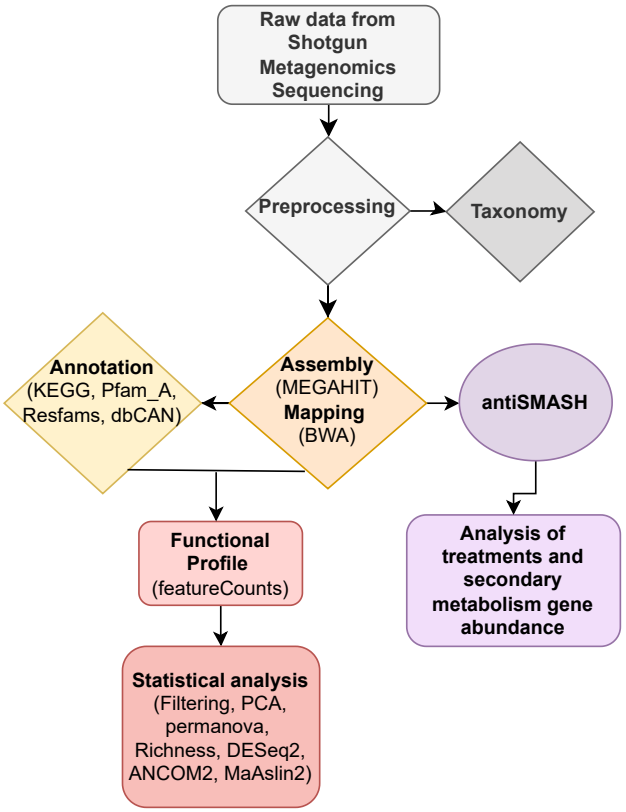

Figure 2 Workflow of the Data Processing, Annotation, and Analysis. Main pipeline processes in rhombus, new steps in the pipeline are shown as ovals, and statistical and various other analyses in rectangular shape. Grey coloring indicates previously completed steps by PhD candidate Bora Kim.

(https://www.ensembl.org/info/website/upload/gff.html). The gene sequences are then annotated using HMMER (Finn et al. (2011)). The annotation tool HMMER (Hidden Markov Models (HMMs) for Biological Sequence Analysis) uses HMMs to search for and align sequences to databases of profile HMMs. The annotation databases used are: KEGG, Pfam_A, Resfams, and dbCAN. The KEGG (Kyoto Encyclopedia of Genes and Genomes) (Kanehisa et al. (2016)) and the Pfam_A (Protein Domain Families Part A) (Sonnhammer et al. (1997)) are large and widely used databases to annotate orthologous groups of proteins. The first is a collection of databases concerning genomes, biological pathways, diseases, drugs, and chemical substances, and it is annotated on the basis of KEGG Orthology (KO) with an identifier for cross-species annotation matches. Pfam contains a comprehensive collection of protein domains and their associated functional and structural information. The Part A is curated through a process of seed alignments that are manually checked for quality and then expanded using HMMs to find and align all members of the family (Sonnhammer et al. (1997)). Of the remaining databases, Resfams contains information about antibiotic resistance genes and mobile genetic elements (such as transposons, integrons, and plasmids) that can mediate the spread of resistance genes (Gibson et al. (2015)). dbCAN contains information about carbohydrate-active enzymes (CAZymes) (Yin et al. (2012)). These are enzymes that are involved in the degradation, modification, and synthesis of carbohydrates, providing information on their domains and structural features.

The output from this step included a GFF file containing information on the metagenomic location and functional potential of the genes. Using another, database-specific GFF, the number of reads mapping to each functional group was calculated by the featureCounts tool (Liao et al. (2014)), based on the alignment tool. The output includes tab-separated-value (tsv) files for each database with reads counts that mapped back from the sequencing data.

Secondary Metabolite Gene Cluster Annotation in antiSMASH6

To integrate antiSMASH6 (Blin et al. (2021)) into the IMP3 pipeline, three scripts were developed as intermediary steps to connect the tools, which can be found in the attached repository. First, the output of assembly gets filtered by a function from the BBMap function: reformat.sh. BBMap is an open source short read aligner and other bioinformatics tools (https://sourceforge.net/projects/bbmap/). The reformat.sh tool filters FASTA files based on multiple parameters and was applied for the FASTA file to filter the contigs based on basepair length with a threshold of 10,000 base pairs. Filtering contigs based on length can help to reduce the number of false positive annotations that may occur when analyzing short contigs. This in turn can improve the accuracy of the analysis and increase the chances of identifying important biosynthesis gene clusters. Then the filtered assembly file was used to filter the annotation.filt.gff annotation file, to obtain an annotation file containing only the same contigs as the filtered assembly file. This is performed with the `gff.filtering.py` script. These are then input for the antiSMASH6 tool. The GFF file was entered with the `--genefinding-gff3` parameter to indicate the existing annotation and file type. The default features include a general KnownClusterBlast analysis on the MiBig (Medema et al. (2015)) repository, SubsclusterBlast analysis from an operon database and Active-Site Finder for highly conserved enzymes detection on variation of active sites. The additional parameters are: `--clusterhmmer` which runs a cluster-limited hmmer analysis, `--cb-subcluster` to perform a ClusterBlast search and compare the identified potential clusters with known subclusters responsible for biosynthesis, `--cb-knownclusters` with a ClusterBlast search on the MIBig repository against known gene clusters of the database, and `--allow-long-headers` to conserve the original contig names during processing. The output includes an .html report that summarizes the regions found and sequence specific details, and GenBank (GBK) files with the annotations from the databases. From the GBK files the information is parsed with another python script (`gbk_parsiong.py`) that translates it into a GFF format for further analysis. The BGCs annotation (GFF) files obtained from antiSMASH6 include databases and tools: Pfam, gene_functions, MiBig, and aSDomain, for which featureCounts can also be applied to obtain mapping read counts. (https://docs.antismash.secondarymetabolites.org/)

Statistical Analysis of Functional Profiles

Analyzing large-scale metagenomic datasets can be challenging due to the unique characteristics of the data such as high dimensionality, count and compositional structure, sparsity, overdispersion and dependencies among samples. To effectively analyze these datasets, specialized methods that can address these challenges are needed. These methods should be implemented in user-friendly and reproducible software to accurately understand microbial communities within population studies while still preserving the ability to detect subtle differences and avoiding false positives. The following processes were applied to the resulting data objects to analyze the microbial communities in the samples.

Joining of Functional Profiles Building matrices of the count data from each annotation database is accomplished by a python script (`mapp_funct_prof.py`), which takes as input the tsv files from the featureCounts. It joins all samples into a tsv file in a matrix format with all samples in the horizontal axis and KOs, functional groups, and Pfam IDs, depending on the database, on the vertical axis. This matrix is filled with the total mapped count reads per functional group for all samples. The following steps include functional analysis on the count data database of annotation mapped reads to the functional groups, per treatment line to examine significant difference in the mapped read counts between the types of samples. From the IMP3 annotation databases mentioned before, the ones used to compare functional profiles were KEGG and dbCAN. This and following steps were performed on R studio and documentation can be found in the attached repository.

Filtering A filtering function was used to remove rare features, which had little information value and whose zeros could disrupt downstream analyses. It filters functional groups from the functional profiles by first calculating the percentage of zero counts per functional group within each treatment line (`per_treat_cutoff`)and then by calculating the percentage of treatment lines that surpass this threshold, removing functional groups with less treatment lines than `funct_cutoff` of acceptable zero percentage. The filtering cutoffs were: `per_treat_cutoff` $>$ 25% and `funct_cutoff` $<$ 85%.

Richness & TukeyHSD test Richness is a measure of the number of different functional groups present in a sample, and it is commonly used to assess the diversity of microbial communities (Stirling and Wilsey (2001)). To calculate Richness, the number of different functional groups present in each sample was summed. To determine statistical significance among the groups, a post-hoc test was performed using the Tukey Honest Significant Differences (TukeyHSD) test. The TukeyHSD test is a multiple comparison procedure that allows for the comparison of all possible pairs of treatment lines while controlling for type I error. The test is based on the Studentized range distribution, also known as the Tukey distribution, and it calculates the critical difference (q*). It was used to compare the mean differences between pairs of groups. The R function used is `TukeyHSD(aov(Richness Treatment))` where the "Richness" variable contains the sample richness values and the "Treatment" variable indicates the different treatment lines. A boxplot was generated using the Richness values of each sample along each treatment line.

Normalization & Transformation To analyze the count data of functional groups in the dataframe, a normalization step was first performed. Two methods were applied separately: the Centered Log Ratio (CLR) transformation and the Sum normalization. The CLR transformation aims to stabilize the variance of the data by removing the effects of the total count and proportion of each sample. The formula for CLR normalization is:

$$clr(x_j) = [ln\frac{x_{1j}}{g(x_j)}, ..., ln\frac{x_{Dj}}{g(x_j)}]$$

The function belongs to the 'compositions' R package, where ratios between the variables are calculated to the geometric mean. This is done by multiplying all elements and taking the nth root of the number of elements used. This is followed by transforming the ratios to logarithmic scale. The Sum normalization method, also known as Total Sum Scaling (TSS), is used to adjust the total count of each sample to a constant value. The formula for Sum normalization is

$$x_i' = x_i / sum(x)$$

where x is the count data and x_i' is the normalized count data. The R function `decostand(method='total')` is used to normalize count data by dividing each count value by the total count for that sample, yielding relative abundance. Both methods allow to compare the abundance of functional groups across different samples and treatments by removing the effects of sequencing depth and total count.

PCA & PERMANOVA Principal Component Analysis (PCA) was applied to the transformed data to identify patterns of variation among the different treatment lines. PCA is a technique for reducing the dimensionality of the data by transforming the original variables into a new set of uncorrelated variables called principal components (PCs). These PCs are linear combinations of the original variables and are chosen in such a way that the first PC explains the greatest amount of variation in the data, the second PC explains the second-greatest amount of variation, and so on.To visualize the results of the PCA, a scatter plot was generated using the first two principal components, which represented the majority of the variation in the data. In the scatter plot, each point represents a sample, and the position of the point represents the sample's scores on the first two principal components.

To test whether the treatment has an effect on the distribution of the data by testing whether the within-group distances are smaller than the between-group distances, a Permutational Multivariate Analysis of Variance (PERMANOVA, Anderson (2014)) test was performed using the `adonis()` function in R. The test is based on permutations of the data and it provides an estimation of the p-value, which indicates the probability of observing difference in composition among the groups by chance. The test is a way to complement the visual representation of the PCA and it can help to understand whether the differences in the position of the points in the scatter plot are significant or not. In addition to the p-value, the result of a PERMANOVA test includes the test statistics, such as the F-value and the R-squared value. The F-value is a measure of the ratio of between-group variance to within-group variance, and it can be used to assess the strength

of the effect. The R-squared value is a measure of the proportion of the total variation in the data that can be explained by the differences among the groups or treatments.

DESeq2 Differential abundance analysis was performed using the DESeq2 package in R (Love et al. (2014)). The data used were non-transformed pair-wise zero-filtered dataframes of the experimental plant treatment lines 712, 722, and CCD8, and control group MMA, each against the empty control, GUS. This package is based on a negative binomial model and it is useful for identifying which functional groups are differentially abundant among the different treatment lines. The package was used to fit a negative binomial model to the count data using the function `DESeqDataSetFromMatrix()` and the design formula was defined using the variable indicating the treatment lines. Then, the function `estimateSizeFactors()` was used to estimate the size factors for each sample, which are necessary to normalize the counts by the sequencing depth. After that, the function `DESeq()` was used to estimate the dispersion estimates for each functional group, which are required to perform the statistical tests. Finally, the function `results()` was used to calculate the p-value, log2 fold change and adjusted p-value for each functional group, indicating whether the group is differentially abundant in the treatment compared to GUS. The adjusted p-value is calculated using the Benjamini-Hochberg correction, which controls the false discovery rate.

ANCOM2 ANCOM2 (Analysis of Composition of Microbiomes) is a statistical method for identifying differentially abundant features (e.g. OTUs, genes, etc.) across multiple samples or groups (Kaul et al. (2017)). It uses a negative binomial model of log-transformed count ratios to calculate the W statistic, which is a measure of the difference in relative abundance between groups. It was applied on a pair-wise analysis of non-transformed zero-filtered data of treatment line 712 against control group GUS. The W statistic is calculated as follows:

$$W = (y_i - y_{base}) * log(y_i / y_{base})$$

Where y_i is the count of feature i in group i, and y_{base} is the count of feature i in a baseline group. The W statistic can be interpreted as the log-fold change in relative abundance between the group i and the baseline group, which should deal with compositionality.

The CLR mean difference, which is used in the volcano plot, is calculated by taking the logarithm of the W statistic and then scaling it by the square root of the mean of the W statistic. The CLR mean difference allows for a more symmetric representation of the log-fold change in relative abundance, which makes it easier to compare across different features. The CLR mean difference is calculated as follows:

$$CLRmeandifference = sqrt(mean(W)) * log(W)$$

The volcano plot generated by ANCOM2 plots the CLR mean difference on the y-axis, and the p-value on the x-axis. Features with a high CLR mean difference and a low p-value are considered to be differentially abundant between groups. Features with a high CLR mean difference and a low p-value are considered to be differentially abundant between groups.

MaAslin2 Differential abundance analysis was also performed using the MaAslin2 (Multivariate Association with Linear Models) package in R (Mallick et al. (2021)). This package is a flexible pipeline that can handle large-scale metagenomic datasets with complex characteristics such as high-dimensionality, sparsity, and compositional data structure. The input was a dataframe with the count data, which in this case was also pair-wise zero-filtered between treatment line 712 and control GUS. Two runs were executed. For the first the minimal prevalence was set to 0 to keep all functional groups. The transform option was set to "NONE" to not perform any data transformation. Normalization was performed using the Centered Log Ratio (CLR) method, and the analysis method was set to Linear Model (LM). Finally, the reference group was set to GUS. The configuration of the first run could deal with the compositionality of the data. For the second run, the minimal prevalence, transform and reference group options were also the same. However, the normalization was performed using the Cumulative Sum Scaling (CSS) method, and the analysis method was set to Negative Binomial (NEGBIN), which in this case could deal with data counts structure.

This method is a statistical approach for analyzing compositional data, such as microbiome data. The input data is transformed and normalized, in this case, the data is normalized by CLR (Centered Log Ratio) method which helps to stabilize the variance and make the data additive. The data is then analyzed using a linear model (LM) or a negative binomial model (NEGBIN) and adjusted for fixed effects, such as treatment line.

Results

The following results aim to explain and understand the differences in functional groups among the tomato treatment lines using high-throughput metagenomic data. These were achieved through several bioinformatics methods and statistical analyses applied on the dataset. All of this to answer the research questions: *What is the effect of tomato strigolactones on the microbial communities of the rhizosphere and in particular on their functional potential to impact the plant?* and *How do the antiSMASH6 implemented IMP3 pipeline results improve or broaden the functional analysis of the rhizosphere microbiome?*

De novo Assembly Improves Mapped Read Counts

The previously obtained results of mapped read counts from reference-based assembly by PhD candidate Bora Kim indicated a mapping rate of 11-20% of reads mapping back to the assembled contigs. On the other hand, the performed *de novo* assembly yielded a mapping rate of approximately 50%, indicating an improved mapping efficiency. This increase in mapped read counts can be attributed to the improved contiguity and completeness of the *de novo* assembly. This improved mapping rate has significant implications for downstream functional annotation and gene prediction, as it increases the accuracy and confidence of the results. Overall, the use of *de novo* assembly in this study demonstrated a clear advantage over reference-based assembly in terms of mapping efficiency and the potential for more accurate functional annotation.

Further values to assess the sequencing quality of the data

Table 2 Assembly output information. CDS are the average number of coding sequences found with potential to be annotated by functional databases, KEGG are the average number of reads that mapped to those CDS KEGG functionally annotated genes, and dbCAN are the average number of reads that mapped to those CDS dbCAN functionally annotated genes. Averaged per treatment line and for all samples together, and with the standard deviation for all samples.

	CDS	KEGG	dbCAN
712	2,952,791	1,753,763	97,158
722	3,361,408	1,987,808	110,909
CCD8	3,393,515	1,995,493	112,402
GUS	3,580,395	2,122,702	120,767
MMA	3,378,798	2,003,746	111,064
EP	3,302,686	1,853,687	112,384
Mean ALL	3,323,491	1,959,105	110,370
SD ALL	286,940	169,215	10,281

and contiguity of the assembly include the N50, the input number of reads after filtering, the number of coding sequences (CDS) identified, and the number of those coding sequences that were annotated with a specific database. The N50 is the length of the shortest contig in the assembly such that half of the total assembly length is contained in contigs of that length or longer. It reflects the number of base pairs (bp) assembled per contig and the number of contigs in the assembly of a certain length or longer. The average N50 of all the samples was 672 bp with a standard deviation of 12. The sequencing depth was observably lower for treatment line 712 (**Figure 3A**), where the average number of reads for 712 was about 42 billion reads, while the rest averaged above 50.5 billion reads. Furthermore, in terms of CDS or potential genes annotated, there was an average 3,323,491 between all samples (**Table 2**). The 712 line had the lowest average number and GUS had the highest average of CDSs, where all other treatment lines deviated in about 0.07 billion from the average. Additionally, the identified CDS regions were then annotated with the mentioned databases and mapped back to the reads to calculate the number of reads mapping to functionally annotated genes. The comparison between the treatment lines for KEGG and dbCAN databases (**Table 2**) follow what was observed for the previous measure, with line 712 having the lowest average and GUS the highest in both cases.

Functional Richness Comparison with KEGG Profile, EP vs. Rhizosphere Lines

The KEGG collection of databases is very extensive and its application on analysis of annotated samples provides insight into their functional composition. Comparing the treatment lines based on functional group (KO) diversity (**Figure 3B**), the distribution was similar among the plant treatments, while for Empty-Pot it contained higher diversity of KOs. The differences in means between treatment lines GUS and EP was significant (TukeyHSD test p-adj<0.05, **Table 3**). Furthermore, despite the lower mean sequencing depth, the functional richness of 712 did not differ from the other plant treatment lines (**Figure 3B, Table 3**). Additionally, treatments 712, 722, GUS, and MMA had similar or equal lower bound Richness values, also with their upper bound distribution reaching about the same values, indicating a similar distribution of range as well as values. Finally, CCD8 had a much wider spread which overlaps with all other treatment lines.

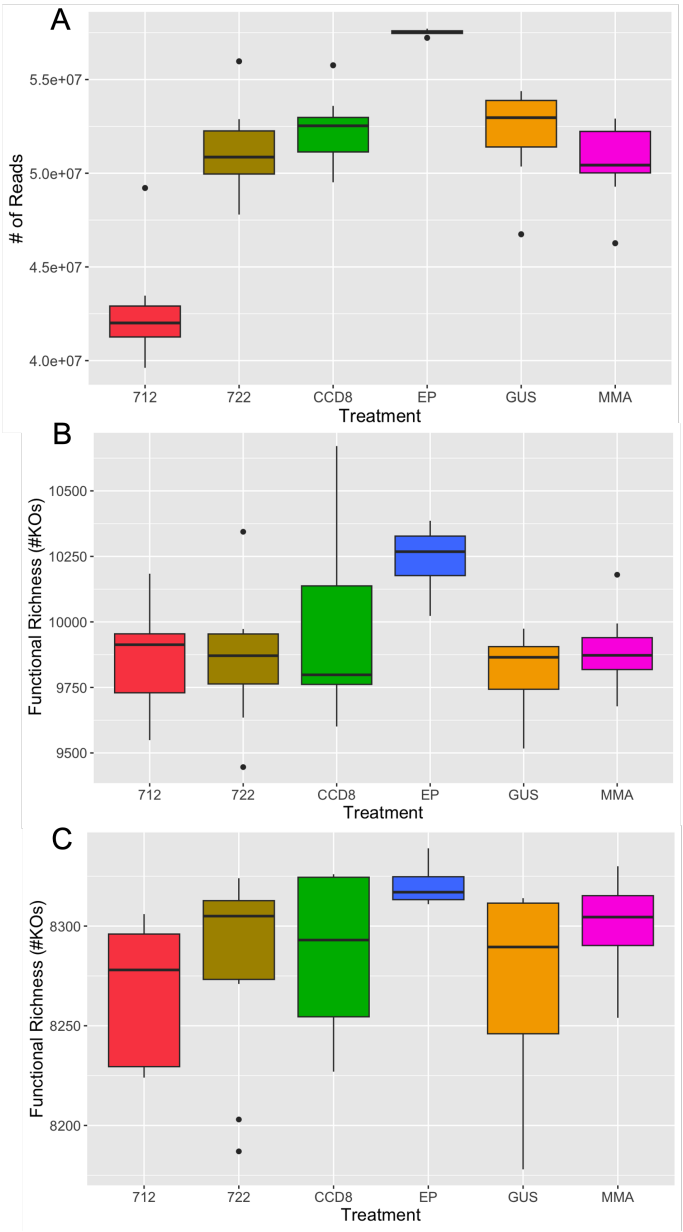

Figure 3 Box Plots of treatment lines sequencing depth and the number of detected KEGG orthologous genes before and after filtration. A) Number of output filtered reads after sequencing B) Raw data Richness. C) Filtered data Richness (cutoff parameters of 25% < `per_treat_cutoff` and `funct_cutoff` < 85%) of different functional groups per sample grouped by treatment.

For the following steps of the analysis the functional profile data was filtered to remove zeros with a rather strict threshold.

The filtering effect can be observed with an initial percentage of zero values over functional groups of 28.89% and with a post-filtering of 1.09%. This filtration yielded an absence of difference in Richness between the groups (**Figure 3C**). This makes the treatment lines more comparable.

Overall Comparison of Functional Profiles: PCA & PERMANOVA for KEGG

Further explorative analysis on the KEGG profile include PCA plots and PERMANOVA test on CLR transformed data and Sum normalized data, including all samples lines. Sum-normalization and CLR-transformation are two commonly used methods in pre-treating functional profiles, which correct for differences in sequencing depth and counter compositionality effects, respectively, before applying any other methods. A higher percentage of variation was explained in the first two principal components in the case of Sum-normalization, reflecting the different scales of the processed datasets. This clear separation between the plant groups and EP group, suggests there is a difference in the functional profile between plant associated soil and bulk soil (**Figure 4B**). Upon looking carefully at the Sum normalized PCA it is possible to distinguish the GUS samples locating further away from the other treatments. The PERMANOVA also indicated significant differences between the plant treatments (**Table 4**).

Table 3 TukeyHSD results indicating difference and P-adjusted values for Richness analysis of the KEGG functional groups. Treatment section contains the treatments that were paired in order to obtain the data. Red asterisk marks P-values that indicate a significant difference.

Treatment Pairs	Diff	P-adj
722-712	-5.427273	0.99999992
CCD8-712	117.727273	0.81036024
EP-712	372.522727	0.06091314
GUS-712	-58.227273	0.99032958
MMA-712	29.522727	0.99971808
CCD8-722	123.154545	0.79681241
EP-722	377.950000	0.06036523
GUS-722	-52.800000	0.99447128
MMA-722	34.950000	0.99941886
EP-CCD8	254.795455	0.37175791
GUS-CCD8	-175.95455	0.46184380
MMA-CCD8	-88.20455	0.95434853
GUS-EP	-430.75000	0.02160189*
MMA-EP	-343.00000	0.13459760
MMA-GUS	87.75000	0.95901408

In the field of microbial ecology, dissimilarities between microbial profiles are commonly represented by measures, such as Bray-Curtis-Index. A PERMANOVA on this distance mostly returned similar results (**Table 5**). For the plant only samples the

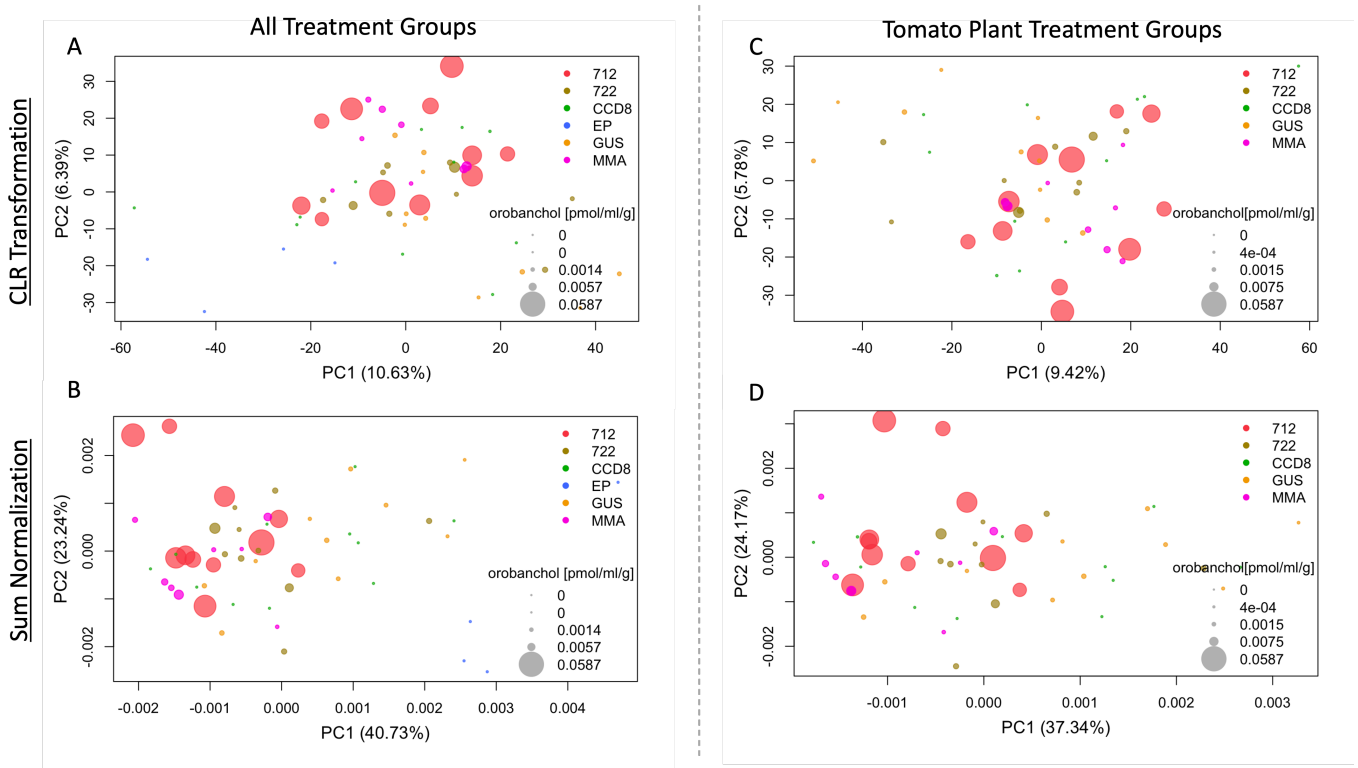

Figure 4 PCA plots for KEGG functional profiles. Left plots include all treatment lines. Right plots include only plant treatment lines (no EP). CLR transformed data was used for the top row of graphs, bottom row were performed with data transformed with Sum Normalization. Data points are replicates colored based on treatment line and size is relative Orobanchol concentration from root exudate analysis.

Table 4 Permanova test results of KEGG functional groups using CLR transformed data and euclidean distance method. Upper rows present all treatment lines, wereas bottom 3 rows showcase only plant treatments (EP is excluded).

	Df	SumSqs	MeanSqs	F.Model	R2	Pr(F)
Treatment	5	28088.36	5617.672	1.538781	0.13815	0.01
Residuals	48	175234.97	3650.729	-	0.86185	-
Total	53	203323.33	-	-	1.00000	-
Treatment	4	19926.82	4981.706	1.213911	0.09739	0.01
Residuals	45	184673.22	4103.849	-	0.90261	-
Total	49	204600.04	-	-	1.00000	-

variation decreased while the high percentage of explained variation was maintained, indicating significant difference between the plant only groups. In plots C and D of **Figure 4**, the separation of GUS is more noticeable in both cases of data processing. In these plots the clustering of the 712 is also visible to a defined region of the plot. The main overlapping treatment samples with this group are 722 and MMA. This shows that even with different data processing and treatment lines, the principal components cluster these into the same range of coordinates in the space. In these cases the clustering of the samples per treatment were also significant with the F statistic lowering only 0.3 values for the CLR data and almost half for the sum normalized data (**Tables 4 & 5**).

Table 5 Permanova test results of KEGG functional groups using Sum Normalized data and Bray-Curtis dissimilarity method. Upper rows present all treatment lines, wereas bottom 3 rows showcase only plant treatments (EP is excluded).

	Df	SumSqs	MeanSqs	F.Model	R2	Pr(F)
Treatment	5	0.022124	0.00442470	4.9686	0.34105	0.01
Residuals	48	0.042746	0.00089053	-	0.65895	-
Total	53	0.064869	-	-	1.00000	-
Treatment	4	0.009216	0.00230390	2.717472	0.19456	0.01
Residuals	45	0.038151	0.00084781	-	0.80544	-
Total	49	0.047367	-	-	1.00000	-

It is important to mention that although the percentages of explained variation are quite different between the two forms of data processing, the observable clustering in the plots follow similar trends. The Sum Normalization allows for correlation between the features when calculating and plotting a PCA. This can be observed in the drop of the F statistic from all treatments to plant only treatments in **Table 5**. On the other hand, the CLR transformation allows for relative abundance in the data to be more evenly spread out, granting proportionality which avoids the pitfall of spurious correlation. In brief, this leads to lower explained variation but mainly because features with higher abundance in the counts data will lead to more separation on that basis, hiding away the effect of features with lower counts that could be substantial to explain the variation between the treatments.

Differential Analysis of KEGG Features

To determine which KOs' abundances differed between the plant treatment lines, DESeq2 differential abundance method were applied. The first differential analysis applied was on the KEGG functional profile of pairs of treatments. The pair-wise comparison of MMA and GUS control lines was performed to evaluate the difference between the controls as a result of the VIGS treatment. The analysis resulted in 1097 differentially abundant KOs between the controls based on an adjusted p-value < 0.05 (**Figure 5**). Indicating that a better control to continue comparing the 3 VIGS plant lines would be GUS, since it was also subject to the vector transformation, having the same underlying effect, and therefore fulfilling comparison assumptions of same genetic conditions. Consequently, the plant treatment lines, 712, 722 and CCD8, were compared to the GUS control to find significant differences in KOs based on the silenced gene for each treatment. The VIGS treatment line with the most differential KOs from GUS was 712. A functional group was considered significant if it has a high log2 fold change and a low adjusted p-value.

Of the number of significant KOs, as determined by a padj-value<0.05 (in the upper right corner of each volcano plot in **Figure 5**), 712 and MMA had an overlap of 452 KOs from each pair-wise comparison against GUS, from the 1180 KOs between 712 and GUS, and the 1097 KOs between MMA and GUS. Furthermore, there were no KOs with significant abundance different between CCD8 and GUS, and only 16 between 722 and GUS. Finally, with a more rigorous threshold of padj-value<0.0005, the DESeq2 count data output of 128 KOs were plotted against the Orobanchol concentration to observe any correlation in the distribution. The obtained plots (**Figure S4**) showed no clear trend for any of the KOs.

To test whether the differentially abundant functional groups would be found by alternative methods, the MaAslin2 function was applied to the raw data on a pair-wise analysis between 712 and GUS. Two runs were performed with different transformation and model parameters, as explained in Methods. The output of the MaAslin2 method includes p-values, effect sizes, and adjusted p-values for each feature (e.g. functional group). The features with a significant p-value and effect size are considered differentially abundant between the treatments.

The output variable that represents the effect size (coefficient) of each feature represents the estimated change in the log-transformed mean of the feature for a one-unit change in the predictor variable, after adjusting for other fixed effects and adjusting for multiple testing. In addition, the p-value associated with the coefficient value can be used to determine the statistical significance of the association.

Significant differentially abundant functions were defined by a p-value<0.05 and a coefficient threshold of ± 1, this only for purposes of visualization (**Figure 6**).

The first run yielded a total of 2038 significant KOs based on a cutoff of p-value<0.05. For the second run with CSS transformation and negative binomial model the total number of KOs with significant differences was 36. For further analysis and comparison the extracted KOs from the DESeq2 method and MaAslin2 were compared to observe any overlapping results. The First MaAslin2 run yielded an overlap of 957 KOs with the

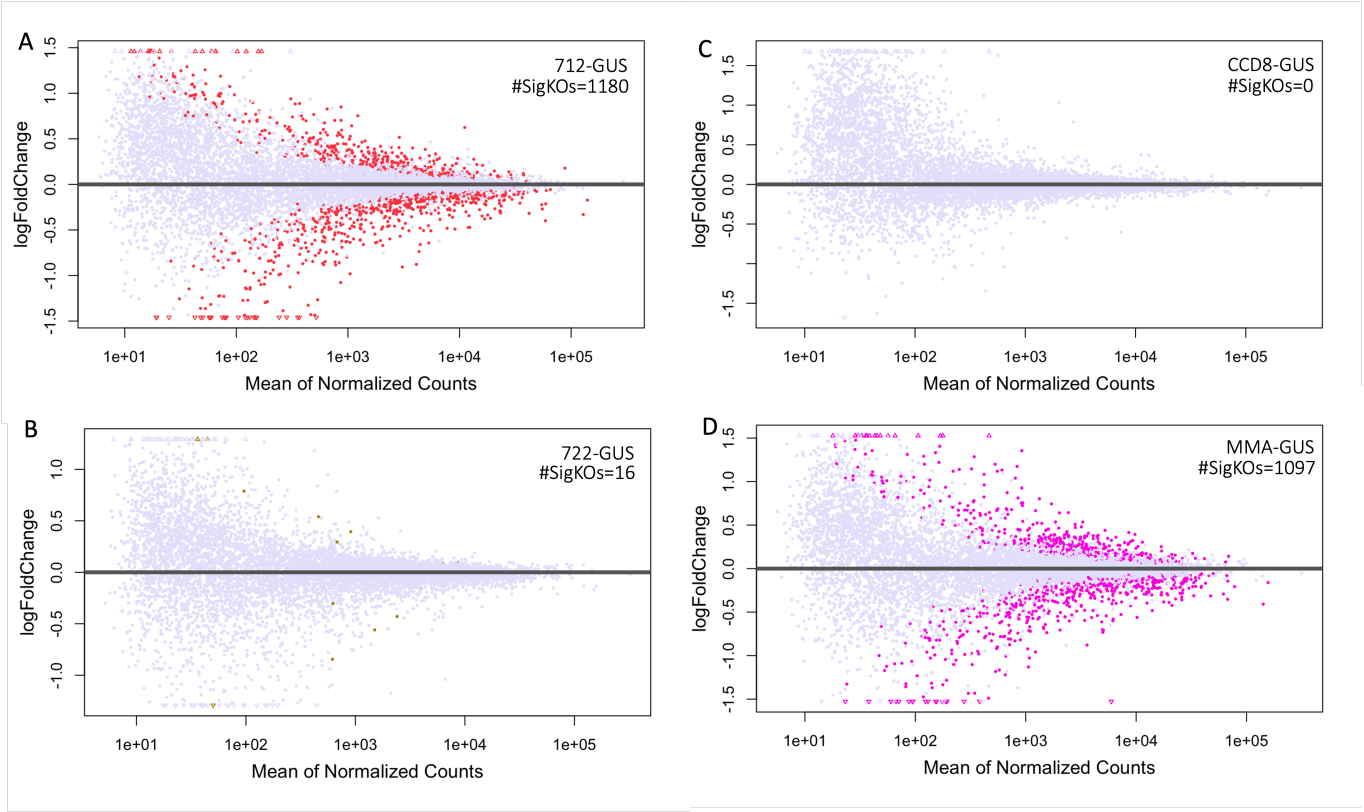

Figure 5 Volcano plots of DESeq2 pair-wise analysis on KEGG functional profiles. Data points are KOs, with lavender color being non-significant and corresponding treatment color based on a significant threshold of adj-p-value < 0.05. A) 712-GUS shows 1180 KOs with significant difference. B) 722-GUS shows 16 KOs with significant difference. C) CCD8-GUS shows no KOs with significant difference. D) MMA-GUS with 1097 KOs with significant difference.

DESeq2 results, while the second MaAslin2 run had 34 overlapping KOs. These were merged into a list with a total of 981 KOs, since there were 10 KOs duplicated in the DESeq2 comparisons of the two MaAslin2 runs. These results were further analyzed by submitting them into the KEGG website Mapper Reconstruction tool to observe the metabolic pathways where these significant KOs can be found (**Figure S1**). The hits indicated 348 KOs in the main metabolic pathway, 134 KOs in the secondary metabolism pathway (**Figure S2**), and 84 KOs in the microbial metabolism in diverse environment pathway (**Figure S3**). The matches between the pathways are observable in the KOs description table in the reference repository. Between the general metabolism and secondary metabolism there were 86 KOs, between the general metabolism and microbial metabolism in diverse environment there were 78 and between all 3 there were 40.

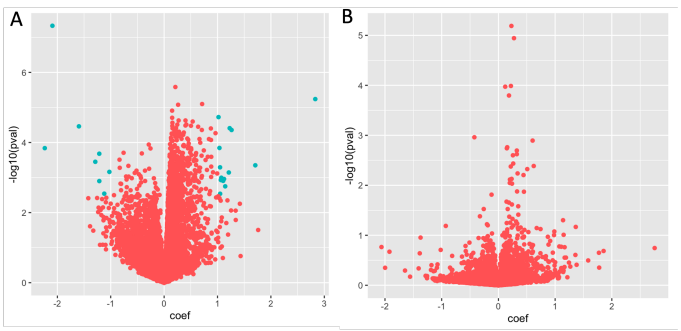

Figure 6 Volcano plots for MaAslin2. Aqua dots correspond to P-value < 0.05 and Coeff $> \pm 1$. The x-axis represents the effect size of each feature, and the y-axis represents the -log10(p-value). The magnitude and direction of the coefficient can be used to infer the strength and direction of the association between the feature and the predictor variable. A) CLR transformed linear model run. B) CSS normalized negative binomial model.

Functional Richness Comparison with dbCAN Profile, EP vs. Rhizosphere Lines

In the analysis of functional groups, dbCAN is a database that was used to annotate the presence of carbohydrate-active enzymes (CAZymes) within the samples. These enzymes play a crucial role in the breakdown and utilization of carbohydrates within the microbial community, such as root exudate rhizodeposits. By utilizing dbCAN, insight was gained into the functional capabilities of the microbes present in the samples

and further analysis was performed to identify potential differences among treatment lines. A similar sequence of analyses as for KEGG was performed on dbCAN in order to maintain a reproducible pipeline on the exploratory and statistical analysis of the data.

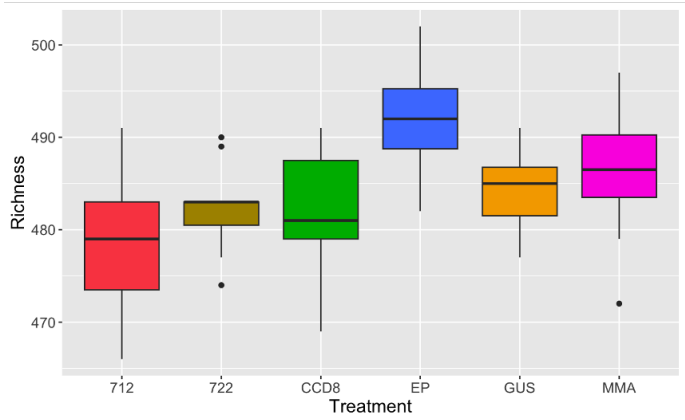

Figure 7 Richness Box Plots of the obtained dbCAN functional profiles. Raw count data measure of different functional groups per sample grouped by treatment (cutoff parameters of 25% for `per_treat_cutoff` and `funct_cutoff` equal to 85%).

First, the Richness assessment of the enzymatic annotations can be observed in **Figure 7**, where the difference between the treatments resembles the previous Richness observations for the KEGG profile. A key difference between the databases is the size, with the KEGG profile containing about 13,000 KO annotations and dbCAN containting about 500 CAZymes. This makes the data more prone to effect by its compositionality and sequencing depth. The main difference within the raw count reads for the profile is found between EP and 712 where a significant difference was measured by the TukeyHSD test, marked by a red asterisk in the P-adj (**Table 10**, in supplementary). Similarly to the KEGG functional profile, the diversity in 712 is not affected by its lower sequencing depth.

As mentioned above, continuing with the statistical analysis of the treatment lines, filtration was applied onto the datasets to remove features with a high percentage of zeros. The strictness of the filtering has a larger impact with this functional profile mainly because there are fewer groups to filter through. This is also observable in the percentage of zero counts per functional group with an initial 13.9% for the raw data but results with a 0.72% after filtration.

Overall Comparison of Functional Profiles: PCA & PERMANOVA for dbCAN

For the exploratory analysis step of this functional profile, the PCA plot in **Figure 8** demonstrates the clustering of the treatment lines along the principal components calculated by the method. In this case both Sum Normalization and CLR transformation have a slightly larger percentage of explained variation than for KEGG, becasue there are less features. Furthermore, the clustering of the treatment lines is more evident in these plots. In plots A and B it is again clear that the EP group separates very well from all the other groups, which is expected due to the fact that the samples are not associated with plant roots and therefore are not subjected to the diversifying or restricting forces upon microbial community diversity.

Statistical support of the PCA plots includes the PERMANOVA test shown in **Tables 6 and 7**. The results of these tests show that in all cases the difference between treatment lines is significant. In both data processing cases, the removal of EP to compare only plant treatment lines results in proportional decrease of the F statistic.

Table 6 Permanova test results of dbCAN functional groups using CLR transformed data using euclidean distance method. Upper rows present all treatment lines, wereas bottom 3 rows showcase only plant treatments (EP is excluded).

	Df	**SumSqs**	**MeanSqs**	**F.Model**	**R2**	**Pr(F)**
Treatment	5	1561.972	312.3943	2.207312	0.18694	0.01
Residuals	48	6793.296	141.5270	-	0.81306	-
Total	53	8355.268	-	-	1.00000	-
Treatment	4	904.732	226.1830	1.5250	0.11923	0.01
Residuals	45	6683.123	148.5138	-	0.88077	-
Total	49	7587.8552	-	-	1.00000	-

Table 7 Results of Permanova test of dbCAN functional groups using transformed data with Sum Normalization and using Bray-Curtis dissimilarity method. Upper rows present all treatment lines, wereas bottom 3 rows showcase only plant treatments (EP is excluded).

	Df	**SumSqs**	**MeanSqs**	**F.Model**	**R2**	**Pr(F)**
Treatment	5	0.029426	0.0058853	4.3036	0.30953	0.01
Residuals	48	0.065641	0.0013675	-	0.69047	-
Total	53	0.095067	-	-	1.00000	-
Treatment	4	0.012889	0.0032222	2.5295	0.18357	0.01
Residuals	45	0.057323	0.0012738	-	0.81643	-
Total	49	0.070212	-	-	1.00000	-

Differential Analysis of dbCAN Features

Finally for the differential analysis of the dbCAN functional profile, both DESeq2 and MaAslin2 were applied as with KEGG, solely on pair-wise analyses between 712 and GUS (**Figure 9**).

Additionally, a fourth differential abundance method was applied for the dbCAN functional profile on the raw zero-filtered counts for treatment lines 712 and GUS. The ANCOM test is a non-parametric method that uses a Wilcoxon rank sum test to compare the relative abundance of features between groups while accounting for compositional data.

For these tests mentioned, DESeq2 resulted in 105 CAZymes with significant difference between the treatment lines tested. For MaAslin2 the results indicate 148 CAZymes with significant differences for the first run, out of which 102 overlapped with the DESeq2 results, and for the second run 117 were found significant

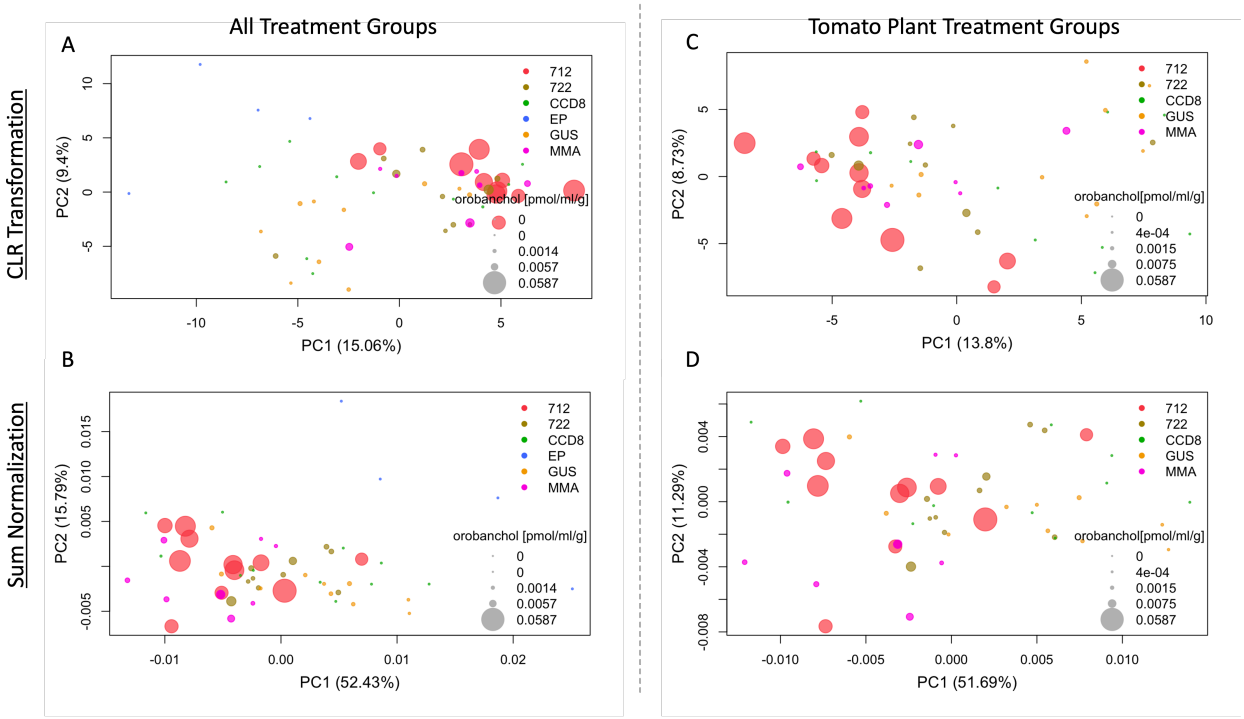

Figure 8 PCA plots for dbCAN transformed functional profiles. Left plots include all treatment lines. Right plots include only plant treatment lines (no EP). CLR transformed data was used for the top row of graphs, bottom row were performed with data transformed with Sum Normalization. Data points are replicates/treatments colored based on treatment line and size is relative Orobanchol concentration from root exudate analysis.

with 89 overlapping with the DESeq2 CAZymes. Furthermore, the ANCOM test resulted in 7 CAZymes with significant difference, which are annotated on the volcano plot (**Figure 9B**. Of these, only 3 overlapped with the MaAslin2 results (GH45.hmm, PL1_4.hmm, CBM59.hmm), and only 2 overlapped among all methods (PL1_4.hmm, CBM59.hmm).

Datashape of antiSMASH6 Results Includes Too Many zeros

In the analysis of functional profiles from the antiSMASH6 annotations, 3 different functional profiles were obtained: Pfam, MiBig, and aSDomain. The goal was to apply similar statistical analyses as with the previous KEGG and dbCAN datasets. However, upon explorative observation it was discovered that these datasets did not contain enough data to be able to draw any differential abundance conclusions through comparative analysis between the treatment lines. The features with counts per sample were very sparse, with counts only for one or two samples per feature. These datasets rather than have a comparative potential, had information about the identity of the genes contained in the biosynthesis gene clusters annotated by the tool. To be able to compare the treatment lines it is more important to compare the types of clusters identified by the tool in each of the treatment lines.

Number of antiSMASH6 Clusters & Sort of Found Clusters

The output of the antiSMASH6 tool resulted in multiple annotations as previously discussed. One important annotation was the identity of the cluster core, which determines the type of secondary metabolite to be synthesized by the gene cluster. The tool identifies 4 main types of BGCs: Terpenes, PKS(polyketide synthases), NRPS(nonribosomal peptide synthetases), and RiPP(ribosomally synthesized and post-translationally modified peptides). The rest of the types of BGCs are classified as 'other' on general description but the core is identified on the whole cluster description (), and some overlapping clusters in the same contigs are classified as 'hybrid'. The results of the antiSMASH6 annotation included varying numbers of regions for each sample with again a clear lower count for the 712 due to the sequencing depth difference. For the type of cluster count (**Table 8**) comparison, the sample per treatment line with the highest number of regions identified was selected. The counts of the main cluster types showed that there is a baseline for the main types of clusters for all treatment lines, showing that they were all present in all the samples. However, there was some difference observed with NRPS having a low count of 5 regions of this type for 712 compared to all other treatment lines, ranging between 10 and 14. Additionally, the hybrid NRPS/PKS was only found in one region of 712 line, while in at least 3 different regions of the other lines, including EP. Interestingly, a hybrid region that was only found in 712 and MMA was RiPP/terpene.

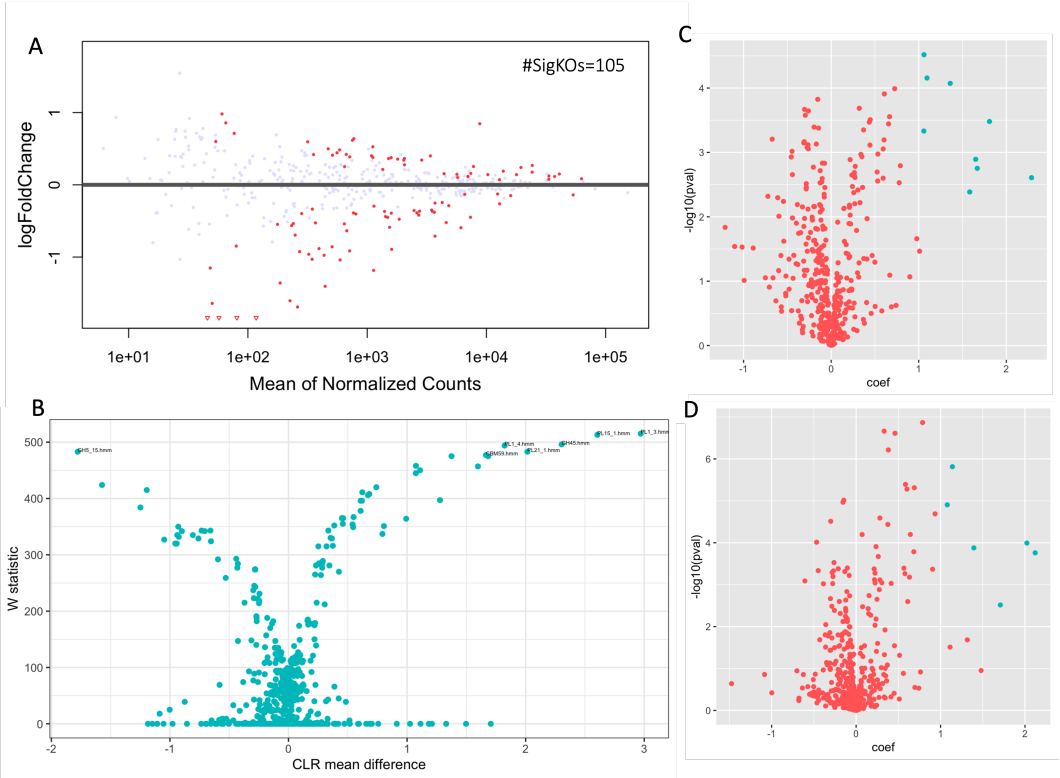

Figure 9 Volcano Plots of differential abundance analysis methods. A) DESeq2 C) & D) 2 MaAslin2 runs, and B) ANCOM, the y-axis represents the -log10 of the W statistic and the x-axis value represents the CLR transformed mean difference in abundance of a given function between the 712 and GUS treatment lines. Points that fall above a certain threshold on the y-axis are considered significant differentially abundant. x-axis position indicates the direction of the difference, with points to the left represent a decrease in relative abundance and points to the right represent an increase in relative abundance. Only functions with reject null-hypothesis >95% are labelled.

Table 8 antiSMASH6 main biosynthesis gene cluster types counts between treatment lines. Sample with the most regions found per treatment used. Total counts included: 712 with 85 regions, 722 with 89 regions, CCD8 with 99 regions, GUS with 91 regions, MMA with also 91 regions, and EP with 84 regions.

Lines	Terpene	PKS	NRPS	RiPP	hybrids
712	9	15	5	27	NRPS/PKS, RiPP/terpene, NRPS/ladderane, NRPS/acyl_amino_acids, siderophore/hserlactone
722	9	8	10	35	NRPS/PKS(3), NRPS/ladderane
CCD8	12	12	13	30	NRPS/PKS(4), NRPS/ladderane
GUS	8	9	14	30	NRPS/PKS(5), NRPS/ladderane
MMA	7	15	11	28	NRPS/PKS(3), resorcinol/arylpolyene, RiPP/terpene, NRPS/ladderane
EP	5	13	11	31	NRPS/PKS(8)

On the additional types of clusters annotated by antiSMASH6, the counts showed siderophore BGC type in more regions of 712 than the other plant lines. Counter-wise, a beta-lactone region was not found in 712, while in 2 regions for 722, CCD8, and GUS, and 5 regions for MMA. Additionally, ladderane was also not found individually in 712, but once in each plant line, and in 3 regions of EP (**Table 9**). However, ladderane was found in a hybrid cluster with RiPP of 712 line. The results also showed a clear difference between the tomato lines and EP, with many of the additional BGC types not being annotated in the EP, while at least once in the other lines. The counts of the additional BGCs were more evenly distributed between the treatment lines, not indicating any trends or distinctions, only presence.

Discussion

The present study had the aim to investigate the effect of tomato strigolactones on the microbial communities of the rhizosphere and in return their functional potential to impact the plant. To achieve this goal, different analysis tools were applied on the shotgun metagenomic sequencing data. The following discussion will highlight the key findings from these analyses and their implications for understanding the relationship between strigolactones,

Table 9 Other antiSMASH6 biosynthesis gene cluster types.

Lines	hserlactone	indole	arylpolyene	ectoine	siderophore	phosphonate	acyl_amino_acids	betalactone	ladderane
712	5	2	7	1	4	2	1	0	0
722	6	1	5	2	2	2	2	2	1
CCD8	3	2	9	1	2	1	2	2	1
GUS	6	2	5	1	1	1	3	2	1
MMA	2	1	6	2	2	1	1	5	1
EP	3	3	4	0	0	1	0	0	3

the rhizosphere microbiome, and the functional potential of the microbiome in its interaction with the tomato plant.

Evaluating Control Lines

The results showed that there was a significant difference in functional group diversity between the treatment lines GUS and EP (**Figure 3B**), as well as a clear separation between the plant groups and EP group in the PCA plots (**Figure 4**). Evaluating the viability of the control lines to yield significant results, it was clear that EP was a control to confirm the overall diversity difference compared to plant presence condition. The findings suggest that there was a diversity difference between plant-associated soil and bulk soil. Furthermore, the MMA control line was also evaluated as a control but also as a possible indication of an effect from the transformation vector. The results stated that there was a significant difference between GUS and MMA lines, from the DESeq2 analysis (**Figure 5D**). This confirmed that GUS was the best control to compare the VIGS lines and that the VIGS vector, even without a gene silencing target, had an effect on the expression of plant genes, which in turn influenced the diversity of the soil microbiome. A literature survey was done on VIGS effect on microbial diversity and function, but it didn't yield any results. Nonetheless the data is showing that applying transformation approach results in significant changes in terms of functional groups.

712 Line Sequencing Depth Limits

It is important to note that the study was limited by the sequencing depth of the treatment samples in 712, which may have affected the results (**Table 2**). The number of output sequenced reads was lower than the rest of the lines, but as observed by the Richness distribution (**Figure 3B**), this did not affect the number of different functional groups found in 712 compared to the other treatment lines (**Figure 7**). To make sure that the differential abundance analysis was robust, to lessen the effect of the sequencing depth difference, multiple differential methods were applied. A comparison between the methods demonstrated an overlap between the significant differentially abundant functional groups, as well as some difference, showing how distinct methods and data handling result in similar results.

Differential Abundance Biological Results from KEGG and dbCAN

The results of the KEGG analysis indicated that there was a significant difference in the functional group diversity between the 712 treatment line and GUS, as well as the other treatment lines. This difference could be attributed to the gene that was silenced in the 712 treatment line, which caused the accumulation of Orobanchol in the plant. The accumulation of Orobanchol has a direct impact on the microbial community in the rhizosphere, as it attracts specific microorganisms that are beneficial for the plant (Kim et al. (2022)). A literature survey was done on VIGS effect on microbial diversity and function, but it didn't yield any results. Nonetheless the data is showing that applying a transformation approach results in significant changes in terms of functional groups.

When comparing the results from MaAslin2 and DESeq2, it was observed that they were similar in the sense that they both identified specific functional groups that were differentially abundant between the treatment lines (**Figure 9**. However, the methods used to obtain these results are different, with MaAslin2 using a linear mixed-effects model and DESeq2 using a negative binomial model. Despite these differences, the results from both methods were comparable as they both provided a measure of the effect size of each feature and a P-value for significance. Additionally, both methods used a threshold for the coefficient to determine significance, with a commonly used threshold for the P-value being 0.05. It is also worth mentioning that -log10 of the q-value can also be used as a threshold for significance.

The results from the KEGG analysis showed that the main pathways where the KOs were involved in include: metabolic pathways, biosynthesis of secondary metabolites, and microbial metabolism in diverse environments (**Figure S1**). Specifically, most differential KOs were mainly in carbohydrate metabolism, energy metabolism, and the biosynthesis of amino acids, nucleotide sugars, and cofactors.

The results from the dbCAN analysis revealed that the most differentially abundant families of CAZymes were GH45, PL1_4, and CBM59. The enzymes encoded by these genes are known to be involved in depolymerization, with CBM59 (carbohydrate-binding module 59) and GH45 (Glycoside Hydrolase Family 45) being necessary for fungi to restructure and recycle their cell wall polysaccharides (Amengual et al. (2022)), and PL14 (Polysaccharide Lyase Family 14) having functions in bacterial metabolism (Alfaro et al. (2016)). By breaking down xylan, CBM59 CAZymes play a role in the decomposition of plant material, making important nutrients available to microorganisms in the environment (Belda et al. (2011)). A previous study on these CAZymes includes fungi secretome functional annotation, where GH45 and PL1_4 were found to have very high expression in the *in vitro* conditions of the studied fungi. The paper concluded that the secretome of the fungi is indicative of the lifestyle or conditions the fungi lives in, rather than the phylogeny. In terms

of the current study, this could indicate that the microbes in the rhizosphere had these functionally important genes involved in the conditions under which they were existing.

Relevant and current research efforts have found that the presence of strigolactones produced by non-mycorrhizal plants affect the diversity of the fungi, but not of the bacteria, in the rhizosphere (Carvalhais et al. (2019)). This relates to the current efforts by indicating that even though the aim was not to prove a differing diversity, a differing functional potential was observed. This could also be a result of the microbial interactions of bacteria and fungi, if in their case their diversity was affected. Another research stated that the strigolactone Orobanchol had a higher effect on the microbial community composition in rice-associated rhizosphere (Kim et al. (2022)). This goes in hand with the results presented in this study, where Orobanchol concentration had higher levels for samples from the 712 treatment line, in which Orobanchol was accumulated due to the gene silencing of downstream synthesis genes that would otherwise convert it into another strigolactone or metabolite.

Biosynthesis Gene Cluster Differences and Similarities Indicating Microbial Functional Potential

The results from the antiSMASH6 tool indicated that the genes individually could not be statistically analyzed for differential abundance between the treatment lines, or even samples, due to lack of sufficient count data. This is expected from this analysis due to the fact that the annotation targets are whole clusters rather than individual genes. The gene clusters contain multiple genes involved in the biosynthesis pathway of the metabolite or metabolites it produces, and the identification of the genes in proximity to one another allows for the identification of a cluster. In many cases, not all of the genes involved in the biosynthesis must be annotated in order to classify a cluster as a type of secondary metabolite BGC. Having the core motif that indicates which type secondary metabolite is synthesized along with a couple other genes is sufficient.

Furthermore, the results indicated that there are mostly similarities between the types of biosynthesis gene clusters of the plant treatment lines. The most important differences to note were count differences of the siderophore and betalactone cluster types, the first present in the 712 line and the latter not present. Siderophores are microbial metabolites that chelate iron ions and help transport them into the cells, which is essential for their growth and survival (Saha et al. (2016)). They are often produced by bacteria and fungi in response to iron limitation, especially in low-nutrient soils, including those that are phosphate deficient. The low availability of iron in these soils can induce the production of siderophores as a way for microbes to acquire the necessary iron for growth and metabolism. On the other hand, Betalactones are a group of compounds found in plants, which may play a role in the regulation of root development and phosphate uptake (Lee et al. (2019)). In phosphate-deficient soil, plants could produce higher amounts of betalactones in order to increase their ability to scavenge for limited phosphate resources. These compounds may act as signals to nearby microbes, inducing them to produce enzymes that solubilize mineral-bound phosphates and make them available to the plant.

Conclusion & Future Implications

In conclusion, the study provides valuable insights into the functional potential of microbial communities in the rhizosphere in response to strigolactones. However, further research is needed to fully understand the role of CAZymes in the rhizosphere and their potential impact on plant growth and development. Additionally, more research is needed to explore the diversity of functional groups in the rhizosphere and their potential impact on plant health.

In terms of future research, it would be interesting to further investigate the impact of Orobanchol accumulation on the microbial community and its functional potential to impact the plant. This could be done through bulk-soil treatment with Orobanchol and other types of strigolactones to observe the impact without the other secreted metabolites of the plant. Additionally, further research could be conducted to explore the effect of other plant hormones and metabolites on the microbial communities of the rhizosphere and endosphere. This could include a more in-depth analysis of the biosynthesis pathways involved in the production of secondary metabolites, and the identification of new biosynthetic types and their metabolites. Furthermore, it would be useful to investigate the functional potential of the soil microbiome according to its diverse composition, and the chemicals at play in the interactions between plants and microorganisms.

Acknowledgements

I would like to thank Dr. Harro Bouwmeester and PhD Candidate Bora Kim for kindly providing the shotgun metagenomic sequencing data of the tomato rhizosphere. Also the Crunchomics Cluster admins for providing access and space to the cluster, in order to perform analytical processes. Finally, my supervisor Dr. Anna Heintz-Buschart for her guidance and comprehension during the duration of the internship research project.

Availability

The repository contains scripts, functional profiles, and additional information about the samples:

https://github.com/susylpzly/tomato-rhizosphere

References

Yanxia Zhang, Carolien Ruyter-Spira, and Harro J Bouwmeester. Engineering the plant rhizosphere. *Current opinion in biotechnology*, 32:136–142, 2015.

Kai Blin, Simon Shaw, Alexander M Kloosterman, Zach Charlop-Powers, Gilles P Van Wezel, Marnix H Medema, and Tilmann Weber. antismash 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Research*, 49(W1):W29–W35, 2021.

Shaman Narayanasamy, Yohan Jarosz, Emilie EL Muller, Anna Heintz-Buschart, Malte Herold, Anne Kaysen, Cédric C Laczny, Nicolás Pinel, Patrick May, and Paul Wilmes. Imp: a pipeline for reproducible reference-independent integrated metagenomic and metatranscriptomic analyses. *Genome biology*, 17(1):1–21, 2016.

Neus Gacias Amengual, Florian Csarman, Lena Wohlschlager, and

Roland Ludwig. Expression and characterization of a family 45 glycosyl hydrolase from fomitopsis pinicola and comparison to phanerochaete chrysosporium cel45a. *Enzyme and Microbial Technology*, 156:110000, 2022.

Maumita Saha, Subhasis Sarkar, Biplab Sarkar, Bipin Kumar Sharma, Surajit Bhattacharjee, and Prosun Tribedi. Microbial siderophores and their potential applications: a review. *Environmental Science and Pollution Research*, 23:3984–3999, 2016.

Davide Bulgarelli, Klaus Schlaeppi, Stijn Spaepen, Emiel Ver Loren Van Themaat, and Paul Schulze-Lefert. Structure and functions of the bacterial microbiota of plants. *Annual review of plant biology*, 64:807–838, 2013.

Barbara Reinhold-Hurek, Wiebke Bünger, Claudia Sofía Burbano, Mugdha Sabale, and Thomas Hurek. Roots shaping their microbiome: global hotspots for microbial activity. *Annual review of phytopathology*, 53:403–424, 2015.

Marnix H Medema, Kai Blin, Peter Cimermancic, Victor De Jager, Piotr Zakrzewski, Michael A Fischbach, Tilmann Weber, Eriko Takano, and Rainer Breitling. antismash: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic acids research*, 39(suppl_2):W339–W346, 2011.

Marnix H Medema, Renzo Kottmann, Pelin Yilmaz, Matthew Cummings, John B Biggins, Kai Blin, Irene De Bruijn, Yit Heng Chooi, Jan Claesen, R Cameron Coates, et al. Minimum information about a biosynthetic gene cluster. *Nature chemical biology*, 11(9):625–631, 2015.

Yanting Wang and Harro J Bouwmeester. Structural diversity in the strigolactones. *Journal of experimental botany*, 69(9):2219–2230, 2018.

Salim Al-Babili and Harro J Bouwmeester. Strigolactones, a novel carotenoid-derived plant hormone. *Annual review of plant biology*, 66:161–186, 2015.

Adrian Alder, Muhammad Jamil, Mattia Marzorati, Mark Bruno, Martina Vermathen, Peter Bigler, Sandro Ghisla, Harro Bouwmeester, Peter Beyer, and Salim Al-Babili. The path from β-carotene to carlactone, a strigolactone-like plant hormone. *Science*, 335(6074): 1348–1351, 2012.

Takatoshi Wakabayashi, Misaki Hamana, Ayami Mori, Ryota Akiyama, Kotomi Ueno, Keishi Osakabe, Yuriko Osakabe, Hideyuki Suzuki, Hirosato Takikawa, Masaharu Mizutani, et al. Direct conversion of carlactonoic acid to orobanchol by cytochrome p450 cyp722c in strigolactone biosynthesis. *Science advances*, 5(12): eaax9067, 2019.

Yanting Wang, Janani Durairaj, Hernando G Suárez Duran, Robin van Velzen, Kristyna Flokova, Che-Yang Liao, Aleksandra Chojnacka, Stuart MacFarlane, M Eric Schranz, Marnix H Medema, et al. The tomato cytochrome p450 cyp712g1 catalyzes the double oxidation of orobanchol en route to the rhizosphere signaling strigolactone, solanacol. *New Phytologist*, 2022.

Anouk Zancarini, Johan A Westerhuis, Age K Smilde, and Harro J Bouwmeester. Integration of omics data to unravel root microbiome recruitment. *Current Opinion in Biotechnology*, 70:255–261, 2021.

Dinghua Li, Chi-Man Liu, Ruibang Luo, Kunihiko Sadakane, and Tak-Wah Lam. Megahit: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de bruijn graph. *Bioinformatics*, 31(10):1674–1676, 2015.

Felix Mölder, Kim Philipp Jablonski, Brice Letcher, Michael B Hall, Christopher H Tomkins-Tinch, Vanessa Sochat, Jan Forster, Soohyun Lee, Sven O Twardziok, Alexander Kanitz, et al. Sustainable data analysis with snakemake. *F1000Research*, 10, 2021.

Derrick E Wood, Jennifer Lu, and Ben Langmead. Improved metagenomic analysis with kraken 2. *Genome biology*, 20(1):1–13, 2019.

Heng Li and Richard Durbin. Fast and accurate long-read alignment with burrows–wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

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

Doug Hyatt, Gwo-Liang Chen, Philip F LoCascio, Miriam L Land, Frank W Larimer, and Loren J Hauser. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC bioinformatics*, 11(1):1–11, 2010.

Robert D Finn, Jody Clements, and Sean R Eddy. Hmmer web server: interactive sequence similarity searching. *Nucleic acids research*, 39 (suppl_2):W29–W37, 2011.

Minoru Kanehisa, Yoko Sato, Masayuki Kawashima, Miho Furumichi, and Mao Tanabe. Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462, 2016.

Erik LL Sonnhammer, Sean R Eddy, and Richard Durbin. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins: Structure, Function, and Bioinformatics*, 28 (3):405–420, 1997.

Molly K Gibson, Kevin J Forsberg, and Gautam Dantas. Improved annotation of antibiotic resistance determinants reveals microbial resistomes cluster by ecology. *The ISME journal*, 9(1):207–216, 2015.

Yanbin Yin, Xizeng Mao, Jincai Yang, Xin Chen, Fenglou Mao, and Ying Xu. dbcan: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic acids research*, 40(W1):W445–W451, 2012.

Yang Liao, Gordon K Smyth, and Wei Shi. featurecounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*, 30(7):923–930, 2014.

Gray Stirling and Brian Wilsey. Empirical relationships between species richness, evenness, and proportional diversity. *The American Naturalist*, 158(3):286–299, 2001.

Marti J Anderson. Permutational multivariate analysis of variance (permanova). *Wiley statsref: statistics reference online*, pages 1–15, 2014.

Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome biology*, 15(12):1–21, 2014.

Abhishek Kaul, Siddhartha Mandal, Ori Davidov, and Shyamal D Peddada. Analysis of microbiome data in the presence of excess zeros. *Frontiers in microbiology*, 8:2114, 2017.

Himel Mallick, Ali Rahnavard, Lauren J McIver, Siyuan Ma, Yancong Zhang, Long H Nguyen, Timothy L Tickle, George Weingart, Boyu Ren, Emma H Schwager, et al. Multivariable association discovery in population-scale meta-omics studies. *PLoS computational biology*, 17(11):e1009442, 2021.

Bora Kim, Johan A Westerhuis, Age K Smilde, Kristỳna Floková, Afnan KA Suleiman, Eiko E Kuramae, Harro J Bouwmeester, and Anouk Zancarini. Effect of strigolactones on recruitment of the rice root-associated microbiome. *FEMS Microbiology Ecology*, 98 (2):fiac010, 2022.

Manuel Alfaro, Raúl Castanera, José L Lavín, Igor V Grigoriev, José A Oguiza, Lucía Ramírez, and Antonio G Pisabarro. Comparative and transcriptional analysis of the predicted secretome in the

lignocellulose-degrading basidiomycete fungus pleurotus ostreatus. *Environmental microbiology*, 18(12):4710–4726, 2016.

Eugeni Belda, Laia Pedrola, Juli Pereto, Juan F Martinez-Blanch, Arnau Montagud, Emilio Navarro, Javier Urchueguia, Daniel Ramon, Andres Moya, and Manuel Porcar. Microbial diversity in the midguts of field and lab-reared populations of the european corn borer ostrinia nubilalis. *PLoS One*, 6(6):e21751, 2011.

Lilia C Carvalhais, Vivian A Rincon-Florez, Philip B Brewer, Christine A Beveridge, Paul G Dennis, and Peer M Schenk. The ability of plants to produce strigolactones affects rhizosphere community composition of fungi but not bacteria. *Rhizosphere*, 9:18–26, 2019.

Shin Ae Lee, Bashistha Kumar Kanth, Hyeon Su Kim, Tae-Wan Kim, Mee Kyung Sang, Jaekyeong Song, and Hang-Yeon Weon. Complete genome sequence of the plant growth-promoting endophytic bacterium rhodanobacter glycinis t01e-68 isolated from tomato (solanum lycopersicum l.) plant roots. *The Microbiological Society of Korea*, 55(4):422–424, 2019.

Supplementary section

Table 10 TukeyHSD test results. Difference and P-adjusted values for Richness analysis of the dbCAN functional groups. Treatment section contains the treatments that were paired in order to obtain the data. Red asterisk marks P-values that indicate a significant difference.

Treatment Pairs	Diff	P-adj
722-712	4.30909091	0.662349829
CCD8-712	4.36363636	0.626556341
EP-712	13.90909091	0.008317082*
GUS-712	6.30909091	0.254727115
MMA-712	7.78409091	0.128095284
CCD8-722	0.05454545	1.000000000
EP-722	9.60000000	0.151345783
GUS-722	2.00000000	0.983044040
MMA-722	3.47500000	0.871332341
EP-CCD8	9.54545455	0.145694855
GUS-CCD8	1.945455	0.9833727
MMA-CCD8	3.420455	0.8690624
GUS-EP	-7.600000	0.3791238
MMA-EP	-6.125000	0.6489391
MMA-GUS	1.475000	0.9968219

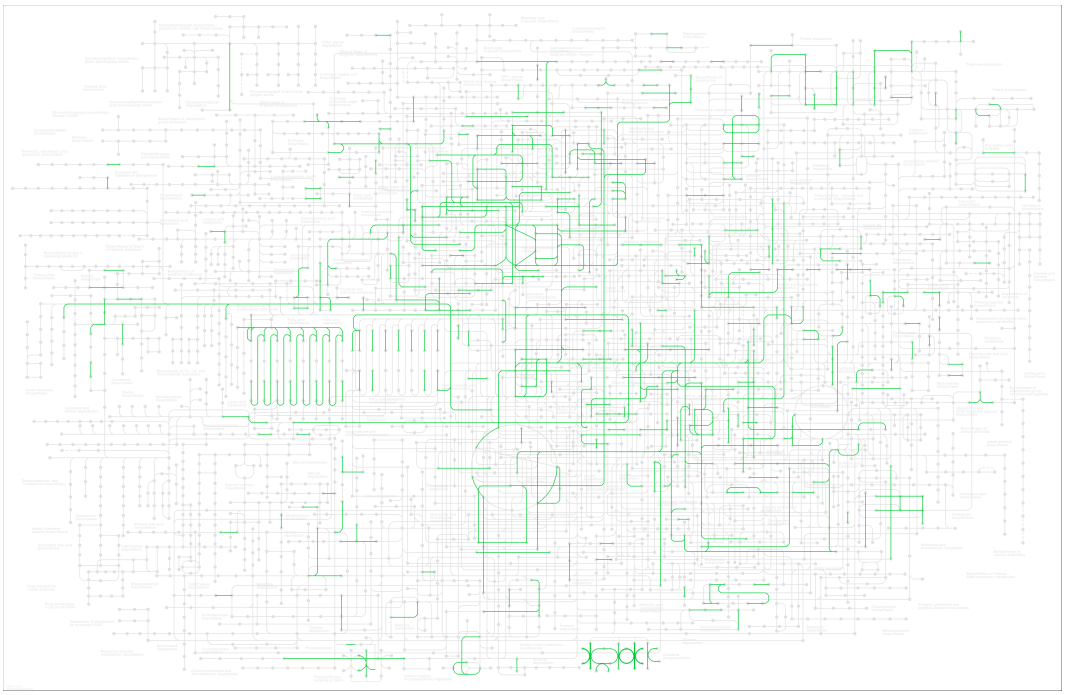

Figure S1 Metabolic pathway map obtained from KEGG database. The data used for generating this pathway match came from the overlap between DESeq and MaAslin2 differentially abundant KOs, which corresponds to 981 KOs. Green highlight indicates matched KOs from the input data to the methabolic pathway, which corresponds to 348.

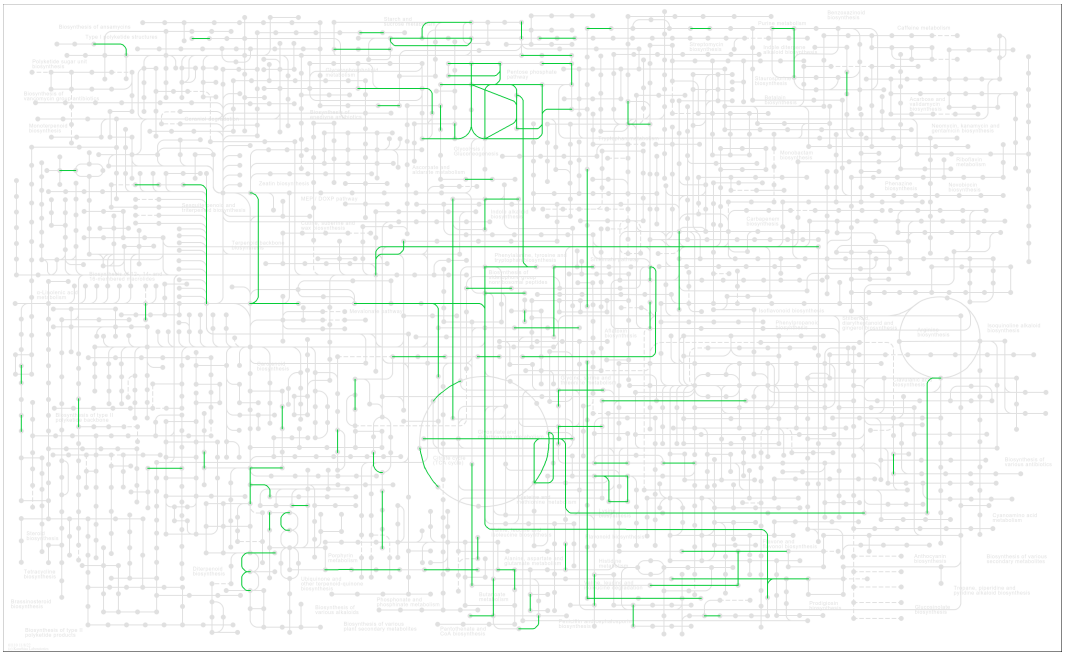

Figure S2 Secondary metabolism pathway map obtained from KEGG database. The data used for generating this pathway match came from the overlap between DESeq and MaAslin2 differentially abundant KOs, which corresponds to 981 KOs. Green highlight indicates matched KOs from the input data to the methabolic pathway, which corresponds to 134.

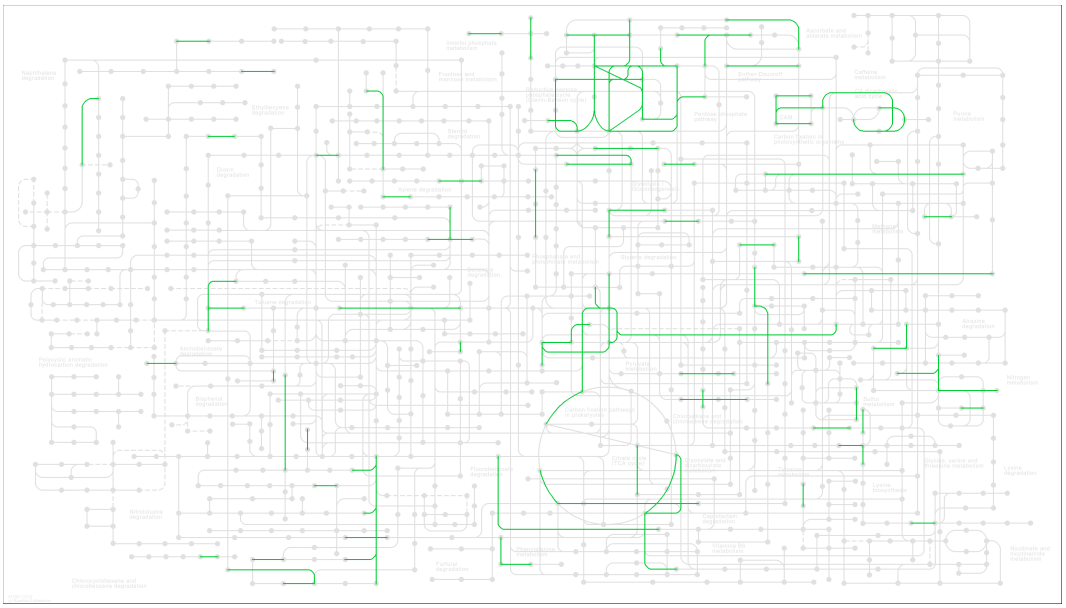

Figure S3 Microbe Metabolism in diverse environemtnpathway map obtained from KEGG database. The data used for generating this pathway match came from the overlap between DESeq and MaAslin2 differentially abundant KOs, which corresponds to 981 KOs. Green highlight indicates matched KOs from the input data to the methabolic pathway, which corresponds to 84.

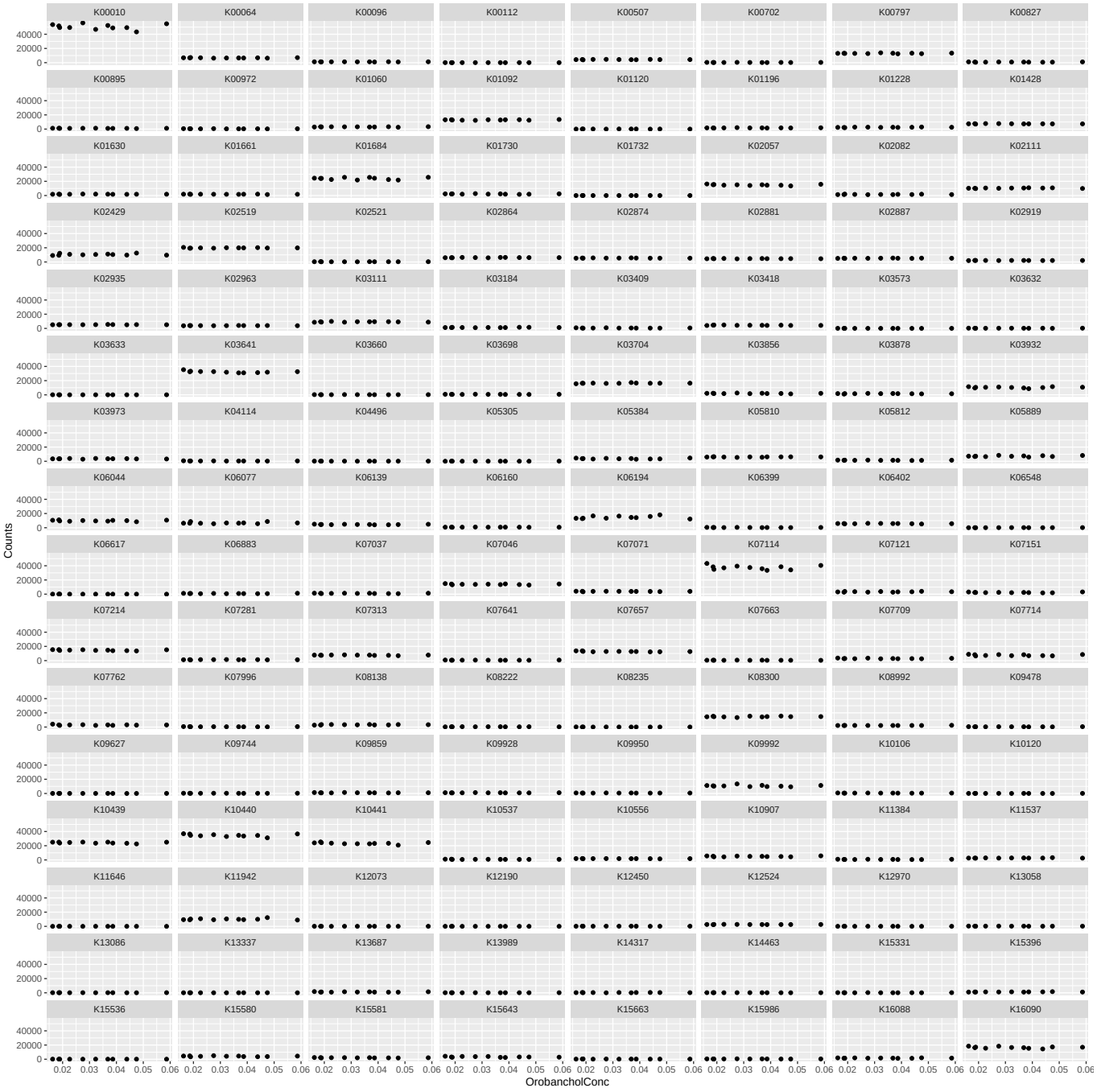

Figure S4 Plots between 712 and GUS significantly different KOs from DESeq2 analysis, with padj<0.0005, against Oronachol concentration in root exudate .