**Methods**

*Microbiome sequencing and analysis*

16S rRNA gene library preparation and sequencing was performed by GenomeQuebéc (Canada) with the Illumina MiSeq platform (PE 250bp). We targeted the V4 region with the primers 515f (5’-GTGCCAGCMGCCGCGGTAA) and 806r(5’-GGACTACHVGGGTWTCTAAT) and included PNA clamps to block plant DNA amplification. Low quality ends of the sequences were removed with Trimmomatic v 0.4. Primer sequence removal with Cutadapt, sequence merging and denoising with dada2, and taxonomy assignment with classify-sklearn were all performed in Qiime2 2021.2. We utilized the SILVA 138 SSU database to train the classifier with rescript according standard configurations. ASVs with less than 8 occurrences were discarded. Library sizes were normalized with Comulative Sum Scaling of the “MetagenomeSeq” package for beta diversity analysis, and rarefied to 47.000 reads per sample with the “phyloseq” package for alpha diversity analysis. A PCoA ordination of bray-curtis distances between microbial communities was calculated with the betadisper function of the vegan package. Significantce of community difference was evaluatated with the functions adonis2 from vegan and Adonis.pair from “EcolUtils” package. Fisher diversity index was calculated with the diversity function of the “microbiome” package, significance was evaluated with the functions “aov” from stats package and LSD.test function from the “agricolae” package. Differential abundance analysis was performed with the DESeq2 package. Random forest feature selection was performed on rarefied data with the “Boruta” package, and evaluated in a 5-fold cross validation repeated 200 times with the “caret” package. The networks were constructed with the “bdgraph” package. Once stable networks were achieved, we only kept the 5% most probable edges of a network based on the ASVs higheted as important by the “Boruta” package. Correlations of plant metadata with network modules were calculated with the “WGCNA” package. The full analysis code with raw data is available at <https://github.com/PedroBeschoren/Katherine_thesis>.

**Results**

We obtained 3.76M pair-ended 250bp sequences from 46 samples, with median phred quality score of 37 in the last base of the sequence after primer removal (SUP\_1\_phed\_quality). 3.09M sequences were filtered, denoised, and merged into 2.48M non-chimeric sequences. After removing plastid and chloroplast ASVs, ASVs occurring less than 8 times, and 6 samples with less than 6000 sequences, the library size per sample was 61,070±9,661 reads (SUP\_2\_rarefaction\_curve)

Microbial communities composition differed according to treatment (**1\_Beta\_Dipersion\_PCoA**). Beta dispersion tests indicated homogeneous multivariate variance (F = 1.907, p=0.1187). There were clear effects for treatment (F = 3.528, R² = 0.341, p = 0.001) and all pairwise treatment comparisons (SUP\_pairwise\_adonis\_table1). The most similar communities were in the MW and HC treatments, as can be observed by the proximity of the centroids. The BSF treatment was the most dissimilar from either controls, accentuating the community differences between CHT and C and between OF and MW/HC.

Differential abundance analysis also indicates clear bi-clustering of samples over 62 different ASVs in 25 Orders (**2\_deseq\_order**). ASV cluster 1 separates C and CHT from other treatments, and cluster 2 shows ASVs missing in OF. Cluster 3 shows Paenibacilalles missing in C and CHT and Catenulisporales missing in C and OF. Cluster 4 has ASVs missing in HC, and cluster 5 has a large group of ASVs, spread out in different samples. Cluster 6 is dominated by high-abundance ASVs which are less frequent in BSF and MW. As a general pattern, Pseudomonadales (10 ASVs), are more present in OF and less present in BSF treatments; Paenibacilalles (7 ASVs) are less present in C but more abundant in HC, MW and BSF; Bacillales (5 ASVs) are less frequent in OF and C; and Burkholderiales (5 ASVs) were spread out but slightly more frequent in HC. There is a large group of 29 ASVs from 14 Orders whose presence characterize HC, BSF and MW and is well populated by Paenibacillales and Catenulisporales. This groups shares Kineosporiales, with C and CHT and 9 ASVs from 7 Orders with OF. The higher presence of Catenulisporales in CHT and ASVs from cluster 4 in C best differentiate these treatments. The NSF and MW treatments are slightly differentiate by abundances of ASVS from Cluster 3.

Alpha diversity levels, measured by the Fisher Diversity index, also clearly differed between treatments (F = 8.81, p <0.001). The BSF treatment presented the highest diversity, similar only to MW. CHT treatment presented an intermediate diversity, as HC and C presented less diversity than MW. OF presented the lowest diversity, comparable only to C (**3\_fihser\_diversity**).Fisher Diversity index was used in place of Shannon diversity index due to better homogeneity of variances by the level test (p = 0.986 for Fisher and 0.058 for Shannon).

Random Forest found 94 ASV features as important predictors of treatment with accuracy of 0.951 and Kappa of 0.940. The most represented taxonomic groups are Family Paenibacillaceae (12 ASVs in 3 genera, 9 of them in Genus Paenibacillus) and Class Actinobacteria (27 ASVs in 7 orders, 7 of them in Order Streptomycetales).From the 15 ASVs with highest importance to the model, 9 are in class Bacilli and 5 in Class Actinobacteria, with an ASV from genus Kitasatospora showing the highest importance. Genus Pseudomonas and Family Blastocatellaceae are also well represented, with 5 ASVs each. As can be seen **4\_RF\_features\_heat\_tree**, BSF has the most Paenibacillus and Controls the least. HC and MW are well represented by Actinobacteria while OF has the most Pseudomonas. CHT is the most similar to control, but does not have large groups of clearly different ASVs when compared to other treatments.

We also detected 23 ASVs as RF predictors of number of flowers in a regression model (RMSE = 8,97, r² = 0.59). While Firmicutes were the most represente phyla (with 8 ASVs), Class Bacilli and Phylum Proteobacteria presented more importance in the model. Notably, Burkholderiales had a much more relevant role as predictors of number of flowers than of treatment (**5\_RF\_features\_flowers**). 13 of these ASVs were also predictors of the different treatments (**RF\_bi-importance\_table1**), with 2 ASVs from Order Bacialialles presenting the highest summed importance. Boruta could not find any ASV that was important to predict plant biomass..

Network analysis of the 94 RF-selected features was employed with the dbgraph package to further identify ASVs of interest. The network was stable after 4.5 million iterations (**SUP\_3\_network\_stability**). There were 101 edges across 85 nodes selected in the final network(**6\_network\_by\_module**), based on link probability of 0.621 (median+ 2SD of link probalibity distribution). ASV\_13 (class Bacilli, family Planococcaceae) was detected as a keystone taxa and module hub, while ASV\_513 (class Bacilli, genus Paenibacillus), was detected as a module connector (**SUP\_4\_network\_support A-B**). 9 modules were detected in the network, with module 1 being the most well-connected with higher number of edges and containing both ASV\_13 and ASV\_513. The ASVs from module 6 were positively correlated to plant biomass and number of flowers (**SUP\_4\_network\_support C**). This module contains ASV\_6, ASV\_26, and ASV\_30, that were also identified as important predictors of the number of flowers in the RF regression. The distribution of these 5 highlighted ASVS (**SUP\_4\_network\_support D**) indicates that their abudancne in BSF is higher than in C and CHT treatments, except for ASV\_3 (*Pseudomonas*), which is also much higher in the OF treatment. Except for ASV\_513, the HC treamtment presented higher median abudancens for the ASVs when compared to the BSF median. Taken together, this results indicate that the most impactful components in community structure are from class bacilli, as their abundances can strongly affect several modules and ASVS within the network, and have the potential to drive signficant community changes. The ASVs that were more correlated to plant phenotype however where from Actinobacteria and Gammaproteobacteria, are very peripheral to the network, and are positively correlated to each other.

**MAIN POINTS FOR RESULTS AND DISCUSSION**

* We have clear alpha and beta diversity profiles (has to be shown!)
* We see the taxonomies that better separate them apart (in RF:bacillus, actino, pseudomonas)
* We see that plant phenotype may be more related to bacterial groups other than bacilli (actino, gamma)
* We see some ASVs overlapping between multiple methods; we can also highlight the taxonomies of these key ASVs
* Not all class Bacilli are or act the same. HC treatments more strongly stimulates Baciliae, while BSF more strongly stimulates Paenibacillus,

**Tables**

**RF\_bi-importance\_table1: Taxanomies and RF model importance of the 13 ASVs detected as important for predicting both treatments and number of flowers**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| ASV\_id | Domain | Phylum | Class | Order | Family | Genus | Mean Importance for Treatment | Mean Importance for number of flowers | Summed importance |
| ASV\_3 | d\_\_Bacteria | p\_\_Proteobacteria | c\_\_Gammaproteobacteria | o\_\_Pseudomonadales | f\_\_Pseudomonadaceae | g\_\_Pseudomonas | 13.999 | 9.005 | 23.004 |
| ASV\_10 | d\_\_Bacteria | p\_\_Actinobacteriota | c\_\_Actinobacteria | o\_\_Micrococcales | f\_\_Intrasporangiaceae | g\_\_Terrabacter | 20.543 | 13.906 | 34.450 |
| ASV\_26 | d\_\_Bacteria | p\_\_Actinobacteriota | c\_\_Actinobacteria | o\_\_Micrococcales | f\_\_Intrasporangiaceae | g\_\_Phycicoccus | 9.798 | 8.259 | 18.057 |
| ASV\_30 | d\_\_Bacteria | p\_\_Actinobacteriota | c\_\_Actinobacteria | o\_\_Propionibacteriales | f\_\_Nocardioidaceae | g\_\_Nocardioides | 22.446 | 9.648 | 32.095 |
| ASV\_36 | d\_\_Bacteria | p\_\_Firmicutes | c\_\_Bacilli | o\_\_Bacillales | f\_\_Bacillaceae | g\_\_Bacillus | 22.620 | 20.943 | 43.563 |
| ASV\_72 | d\_\_Bacteria | p\_\_Actinobacteriota | c\_\_Actinobacteria | o\_\_Streptomycetales | f\_\_Streptomycetaceae | NA | 13.633 | 13.856 | 27.490 |
| ASV\_204 | d\_\_Bacteria | p\_\_Firmicutes | c\_\_Bacilli | o\_\_Bacillales | f\_\_Planococcaceae | NA | 19.869 | 21.862 | 41.731 |
| ASV\_254 | d\_\_Bacteria | p\_\_Proteobacteria | c\_\_Gammaproteobacteria | o\_\_Burkholderiales | f\_\_Oxalobacteraceae | g\_\_Massilia | 10.075 | 19.546 | 29.622 |
| ASV\_266 | d\_\_Bacteria | p\_\_Firmicutes | c\_\_Bacilli | o\_\_Paenibacillales | f\_\_Paenibacillaceae | g\_\_Paenibacillus | 21.457 | 13.942 | 35.399 |
| ASV\_283 | d\_\_Bacteria | p\_\_Firmicutes | c\_\_Bacilli | o\_\_Alicyclobacillales | f\_\_Alicyclobacillaceae | g\_\_Tumebacillus | 8.565 | 15.596 | 24.161 |
| ASV\_346 | d\_\_Bacteria | p\_\_Firmicutes | c\_\_Bacilli | o\_\_Paenibacillales | f\_\_Paenibacillaceae | g\_\_Paenibacillus | 15.606 | 20.875 | 36.480 |
| ASV\_449 | d\_\_Bacteria | p\_\_Actinobacteriota | c\_\_Actinobacteria | o\_\_Corynebacteriales | f\_\_Nocardiaceae | g\_\_Nocardia | 17.099 | 14.758 | 31.857 |
| ASV\_511 | d\_\_Bacteria | p\_\_Actinobacteriota | c\_\_Actinobacteria | o\_\_Micromonosporales | f\_\_Micromonosporaceae | NA | 7.710 | 11.242 | 18.952 |

**SUP\_pairwise\_adonis\_table1:** pairwise comparsion of microbial community composition between treatments. OF = Organic fertilizer…..

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Combination | | | SumsOfSqs | MeanSqs | F.Model | R2 | P.value | P.value.adjs |
| BSF | <-> | C | 0.5012 | 0.5012 | 4.9698 | 0.2766 | 0.0020 | 0.0033 |
| BSF | <-> | CHT | 0.2944 | 0.2944 | 3.2914 | 0.2152 | 0.0020 | 0.0033 |
| BSF | <-> | HC | 0.3344 | 0.3344 | 4.7594 | 0.2840 | 0.0020 | 0.0033 |
| BSF | <-> | MW | 0.1960 | 0.1960 | 2.6510 | 0.1942 | 0.0020 | 0.0033 |
| BSF | <-> | OF | 0.6569 | 0.6569 | 6.9115 | 0.4087 | 0.0030 | 0.0041 |
| C | <-> | CHT | 0.2252 | 0.2252 | 2.0067 | 0.1337 | 0.0120 | 0.0120 |
| C | <-> | HC | 0.3434 | 0.3434 | 3.6325 | 0.2184 | 0.0010 | 0.0033 |
| C | <-> | MW | 0.2849 | 0.2849 | 2.8514 | 0.1920 | 0.0040 | 0.0043 |
| C | <-> | OF | 0.3852 | 0.3852 | 3.1702 | 0.2237 | 0.0040 | 0.0043 |
| CHT | <-> | HC | 0.2558 | 0.2558 | 3.0968 | 0.2051 | 0.0010 | 0.0033 |
| CHT | <-> | MW | 0.1876 | 0.1876 | 2.1469 | 0.1633 | 0.0040 | 0.0043 |
| CHT | <-> | OF | 0.4119 | 0.4119 | 3.7496 | 0.2727 | 0.0020 | 0.0033 |
| HC | <-> | MW | 0.1588 | 0.1588 | 2.3895 | 0.1785 | 0.0010 | 0.0033 |
| HC | <-> | OF | 0.3116 | 0.3116 | 3.5876 | 0.2640 | 0.0020 | 0.0033 |
| MW | <-> | OF | 0.3761 | 0.3761 | 4.0374 | 0.3097 | 0.0030 | 0.0041 |

**Legends**

**1\_Beta\_Dipersion\_PCoA:** PCoA ordination of Bray-Curtis distances between microbial communities. Different colors and shapes indicate different treatments, and the treatment label indicates the centroid. OF = Organic fertilizer…..

**2\_deseq\_order:** Bi-cluster heatmap indicating log-transformed the frequency of differentially abundant ASVs (rows) across the different samples (columns). ASVs clusters are numbered and alternating in grey scale to facilitate identification. Asterisks indicate ASVs that were also defined as important to separate treatments by the random forest model. Orders with a single ASV representative are also in grey scale.

**3\_Fisher\_diversity**: Fisher diversity index indicating treatment effect. Letters denote significant pairwise differences under the LSD test. Colors denote different treatment. OF = Organic Fertilizer…

**4\_RF\_features\_heat\_tree:** Matrix of heat trees comparing ASVs that were important according the Boruta RF model across the treatments. The large grey tree to the left shows a key to the legend. Each node represet a taxonomic group, from domain to ASV. Colors indicate log2-transformed relative abundances in pairwise comparsions across the rows and columsn that represent treatments.

**5\_RF\_features\_flowers:** Heat tree showing ASVs that were important according the Boruta RF model across number of flowers. Each node represet a taxonomic group, from domain to ASV. Colors indicate log2-transformed relative abundances.

**6\_network\_by\_module:** ASVnetwork and sub-communities. Only ASVs detected as important in the Ranodm Forest algorithm were included in the calculations for the network. Each circle represents an OUT, and groups of circles represent a network module. Red lines indicate positive weights, and blue edges represent negative weights. ASV\_521 was classified as a module connector and is slightly outside module 1. ASV\_13 was classified as a module hub as as a keystone taxa, and is in the center of module 1. Edges from these two ASVs are represented with thicker lines. Node color indicates different classes, with low-diversity classes greyed out

**SUP\_1\_phed\_quality:** phred quality scores of forward and reverse reads before (a) and after (b) trimming of low quality ends with Trimmomatic and of primer sequences with Cutadapt.

**SUP\_2\_rarefaction\_curve:** Rarefaction curve showing library size and ASV diversity per sample (black lines). Six samples with less than 6.000 reads were removed from the analysis. Libraries were rarefied to at least 47.124 reads per sample (red dashed line)

**SUP\_3\_network\_stability:** Stabiltiy of networks created with 4.5M iterations on the dbgraph package. A= posterior link probability of 100 edges in the last iterations, where the flat lines indicates stability for edge probability has been reached. B = summary of the bdgraph network, indicating mean network size. The normal distribution of network size at link probability of 50% indicates stability has been reached. Probability link cut-off was based on the probability median +2 standard deviations.

**SUP\_4\_network\_support:** highlighted ASVs and modules in the network. A = z-transformed node degree, betweenness centrality and closeness centrality. The dashed lines indicates mean + 2 SD; A single ASV above both dashed lines in each panel was classified as keystone taxa. B = Pi (inter-module connectivity) and Zi (intra-module connectivity), with dashed lines indicating module connectors (Pi>0.62) and module hubs (Zi >2.5). colors indicate different classes. C = correlation of plant biomass and number of flowers with each module, D = Log-transformed rarefied abundance of noteworthy ASVs in the network analysis