Statistical Analysis Report

Biomcare ApS

24/05/2023

Customer	Anton Rasmussen and Sidsel Schmidt
Customer ID	DA01102-22
Project	The effect of compost-based fertilizer on the microbial community
Sample Type	Soil
Number of samples	12 samples
Type of data	Shotgun metagenomic sequencing

Introduction to the biostatistical analysis

Objective

The project objective is to evaluate the effect of a de-gassed compost-based fertilizer on the soil microbiome. One field was included in the project, named Field D. The field was split into two blocks, and three replicates were taken from each. One block was included as a control where no fertilizer was applied and in the other block, one of the two fertilizers was added. Sampling was performed at two time points (TP), before fertilizer was applied (baseline) and 2 months after.

We aim to evaluate the effect of fertilizer treatment on the soil microbiome.

To support the analyses, we have named the two blocks per field as G1 and G2, indicating the block that remains untreated and the block that is treated between TP1 and TP2, respectively. These abbreviations, as well as the notation TP for time point, are used in illustrations.

Analysis

In this report, biostatistical analyses are performed and the results presented, building on the data generated and evaluated in the 2 prior reports (**Report 1: Sequencing and data processing report, Report 2: Microbiome profiling report**).
Through biostatistical analysis we compared the microbiome profiles between the different groups of samples. The focus here is to evaluate the effect of treatment and to use the untreated blocks as controls for the changes happening due to other factors.

A number of analyses were performed, as shortly introduced here:

1. The report initiates with a visual evaluation of the overall microbial community composition using ordination plots and a statistical analysis to evaluate how much of the compositional variation in the microbiome is explained by the variable(s) of interest and if the compositional differences are statistically significant (using Permutational Multivariate Analysis of Variance (ADONIS)).
2. If changes are seen in the overall community composition, it can be due to subtle changes in many microbes or more pronounced changes to few microbes. If the latter is causing the observed shifts in the overall microbiome, we can identify the specific microbes that have increased or decreased in proportion due to the variable of interest. The analysis thereby allows us to identify indicator organisms that may be of interest for further testing. We compared the abundance of organisms classified at one or more taxonomic levels, depending on the power and aim of the study.
3. Lastly, we investigate the alpha diversity, which is a measure of the within-sample diversity, often referred to as biodiversity, and we evaluate whether the alpha-diversity differs between groups using two different measures of alpha-diversity.
4. Lastly, we investigate the percentage of fungi detected, and we evaluate whether the measure differs between the defined groups in a similar manner to the analysis of alpha-diversity.

Differences in overall community composition

Visualization by ordination (beta-diversity)

As described in **Report 2**, beta-diversity is a measure of how similar or dissimilar the microbial community is between each pair of samples. The measures are useful for statistical analysis and visualization of the overall microbiome community. In ordination plots, each sample is a point and the distance between the points increases with increasing dissimilarity in the microbiome communities.

Here we evaluate the microbiome communities using the Bray-Curtis, Aitchison and Jaccard beta-diversity measures. If not all plots are shown in this report, you can find them in the project folder. We use a combination of measures as each measure highlights different properties of the microbiome. See more details in Report 2. We use the different measures in combination with different microbiome profiles (taxonomic levels and normalization) as follows:

- we use Bray Curtis and Jaccard on the relative abundance data, at the species level
- we use Aitchison distance on the total abundance data transformed with central-log-ratio (CLR), at the species level

The Aitchison distance is a simple euclidean distance calculated using CLR transformed microbiome profiles. An analysis of CLR transformed data will reveal how the organisms behave relative to the per-sample average microbiome. Values for a microbe can therefore be negative after CLR transformation - meaning that it makes up a smaller amount of the microbiome than the average abundant microbe. This is a very different way to view the microbiome than Bray-curtis and Jaccard that uses the data as relative proportions (i.e. how big a proportion of the sample’s microbiome does the individual microbe comprise). This might appear unnecessarily mathematical and unrelated to agrobiology but the CLR transformation has proved to be able to pinpoint patterns in microbiomes that are driven by environmental factors such as nutrient content or treatment applied to the samples. We therefore evaluate structures in the dataset using all three measures.

In the plot(s), samples are colored by treatment status and time point. When samples are split into one panel per (field)plot, coloring is only used to identify TP.

Bray-Curtis Jaccard Aitchison

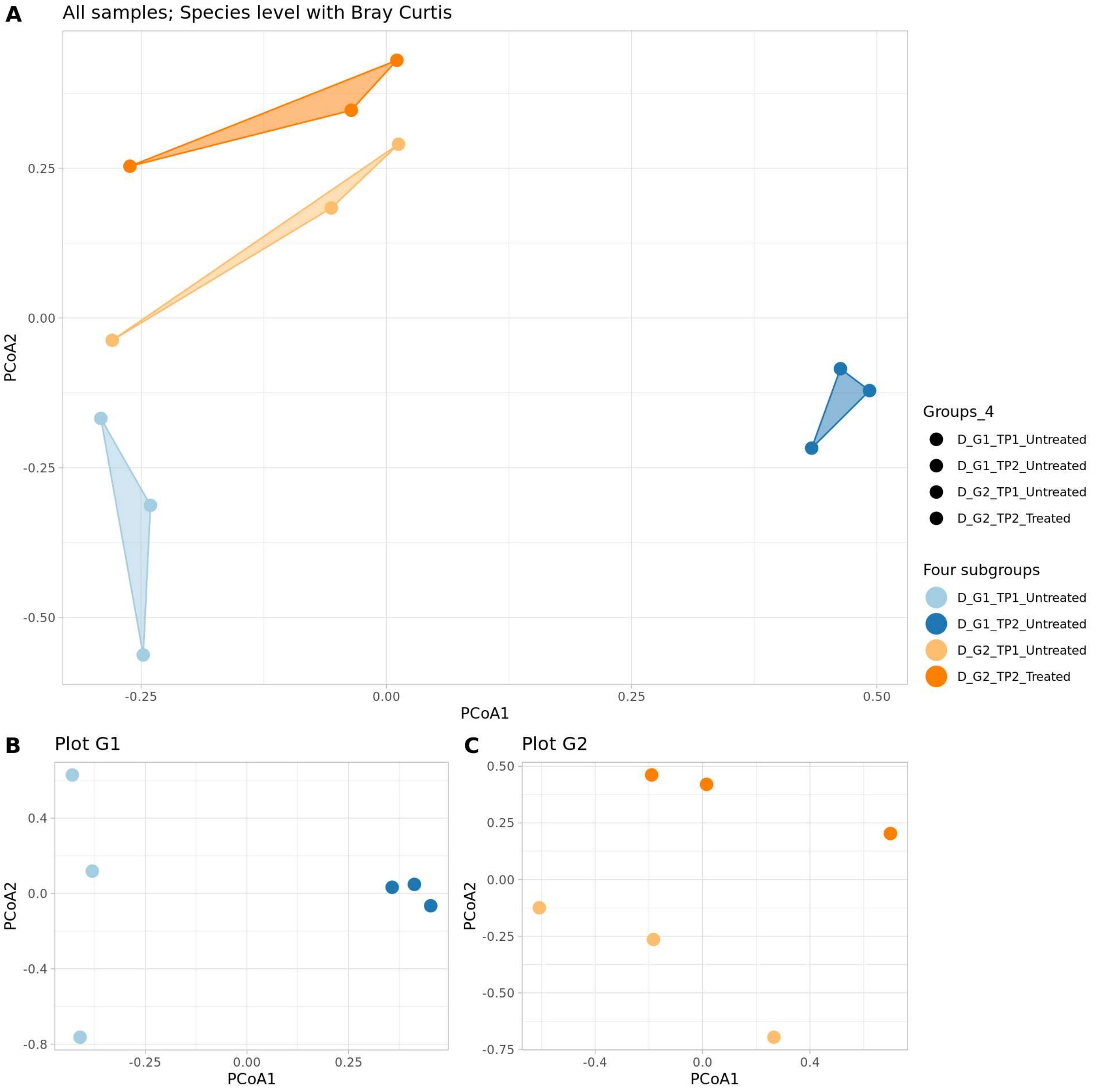

Figure 1: Visualization of structure of the microbial community between the samples. Ordination plots using Bray-Curtis for relative abundance transformed taxa and a NMDS ordination method. **A)** shows community composition for all samples across plots and TP, while samples have been subset to individual plots in **B)** and **C)**. Dots are colored by treatment status, TP and block as indicated. In **A)** we added “convex hull shapes” to guide the eye to where the samples of each group are located.

Permutational Multivariate Analysis of Variance

To evaluate if the compositional differences evaluated above using ordination plots explain much of the microbial variation and if they are statistically significant, we perform an analysis named Permutational Multivariate Analysis of Variance (ADONIS). ADONIS partitions sums of squares of a multivariate dataset and is analogous to MANOVA (multivariate analysis of variance) using beta-diversity measures. It partitions distance matrices among sources of variation and fits linear models to the distance matrices using a permutation test with pseudo-F ratios and can therefore be considered as a “permutational manova”. For the analysis we use Bray-Curtis and Jaccard beta-diversity measures and perform the analysis at the species up to the phylum level.

ADONIS results

After evaluating the ADONIS analysis in detail I have decided to not include the results, as the analysis is unstable tue to low number of samples (a simple repeated run and comparison gives very fluctuating statistics). Therefore, we must judge by visual inspection the effect of treatment on the microbiome overall composition. I leave the explanation of the analysis model in the report, to show what we can do to add statistical insight to the ordination plots, when we have the power to do so.

Differential abundance of single taxa

We now move from the evaluation of overall microbiome diversity and composition, to evaluate if the abundance of specific organisms are affected by the treatment. This analysis provides the first insight into potential indicator organisms and a first peek at the single organisms that drive compositional differences between the groups of samples. Due to the small number of samples per group (n=3) we have chosen to analyse organisms at the taxonomic level of family and to analyse the top 20 most abundant family clades. A number of statistical models are available for such analysis, and a model must be selected considering the study design and power available. We selected a linear regression with a so-called "change-score model" as the samples are collected from two separate blocks and therefore the groups are not composed of samples taken from the same population at baseline. The model therefore evaluates if the difference in abundance between the two time-points is significantly different between G1 and G2; meaning is the difference that occur over time different when the treatment is applied.

Below you find a table for the analysis of field D. The table provide summary statistics for the analysed clades and results from the statistical analyses. A boxplots is made for the most associated clades (P<0.1). These plots allow for visual inspection of the clades that show significant (p<0.05) or a trending association. Due to the low number of samples per group, we have a low power for obtaining a significant p-value and therefore, it is relevant to also inspect clades that show interesting trends.

Species	Phyla	Order	Family	Field	G1_TP1_Untreated	G1_TP2_Untreated	G2_TP
				Median	Median	Median	Mediar
ID_1404	p__Firmicutes	o__Bacillales	**f__Bacillaceae**	11.88	6.58	13.04	13.74
ID_29448	p__Proteobacteria	o__Hyphomicrobiales	**f__Bradyrhizobiaceae**	9.02	8.92	6.61	10.28
ID_1916	p__Actinobacteria	o__Streptomycetales	**f__Streptomycetaceae**	7.37	10.67	5.92	7.11
ID_71433	p__Proteobacteria	o__Hyphomicrobiales	**f__Phyllobacteriaceae**	5.88	4.92	3.38	6.13
ID_196162	p__Actinobacteria	o__Propionibacteriales	**f__Nocardioidaceae**	5.05	6.21	4.93	4.03
ID_1849032	p__Actinobacteria	o__Micrococcales	**f__Micrococcaceae**	4.95	5.49	3.93	5.04
ID_2792224	p__Proteobacteria	o__Burkholderiales	**f__Comamonadaceae**	3.99	3.28	5.46	2.76
ID_117567	p__Actinobacteria	o__Corynebacteriales	**f__Mycobacteriaceae**	3.43	4.17	3.09	3.59
ID_386874	p__Proteobacteria	o__Sphingomonadales	**f__Sphingomonadaceae**	3.02	2.78	3.03	3.02
ID_190721	p__Proteobacteria	o__Burkholderiales	**f__Burkholderiaceae**	2.36	1.33	2.36	2.36
ID_47850	p__Actinobacteria	o__Micromonosporales	**f__Micromonosporaceae**	2.2	2.36	10.66	1.74
ID_82380	p__Actinobacteria	o__Micrococcales	**f__Microbacteriaceae**	2.06	2.91	1.39	1.9
ID_2762611	p__Proteobacteria	o__Xanthomonadales	**f__Xanthomonadaceae**	1.83	1.97	1.01	1.91
ID_2006	p__Actinobacteria	o__Streptosporangiales	**f__Streptosporangiaceae**	1.5	1.22	1.32	1.93
ID_1833	p__Actinobacteria	o__Corynebacteriales	**f__Nocardiaceae**	1.46	1.66	0.84	1.36
ID_1852	p__Actinobacteria	o__Pseudonocardiales	**f__Pseudonocardiaceae**	1.46	1.59	1.41	1.51
ID_666685	p__Proteobacteria	o__Xanthomonadales	**f__Rhodanobacteraceae**	1.43	1.55	0.42	2.12
ID_2026	p__Firmicutes	o__Bacillales	**f__Thermoactinomycetaceae**	1.33	0.95	1.41	1.58
ID_1464	p__Firmicutes	o__Bacillales	**f__Paenibacillaceae**	0.66	5.61	0.64	0.87
ID_2742128	p__Actinobacteria	o__Streptosporangiales	**f__Thermomonosporaceae**	0.61	0.63	7.69	0.59

Table 2: Summary table of median abundance and statistical results for the top 20 most abundant family clades. The table shows all top 20 abundant clades with their NCBI taxonomic ID number and annotations at phyla, order and family level. Median values are given for the whole field and for each sample group. For the statistical analysis, estimates or beta-values are given for the difference between the groups and the corresponding p-values.

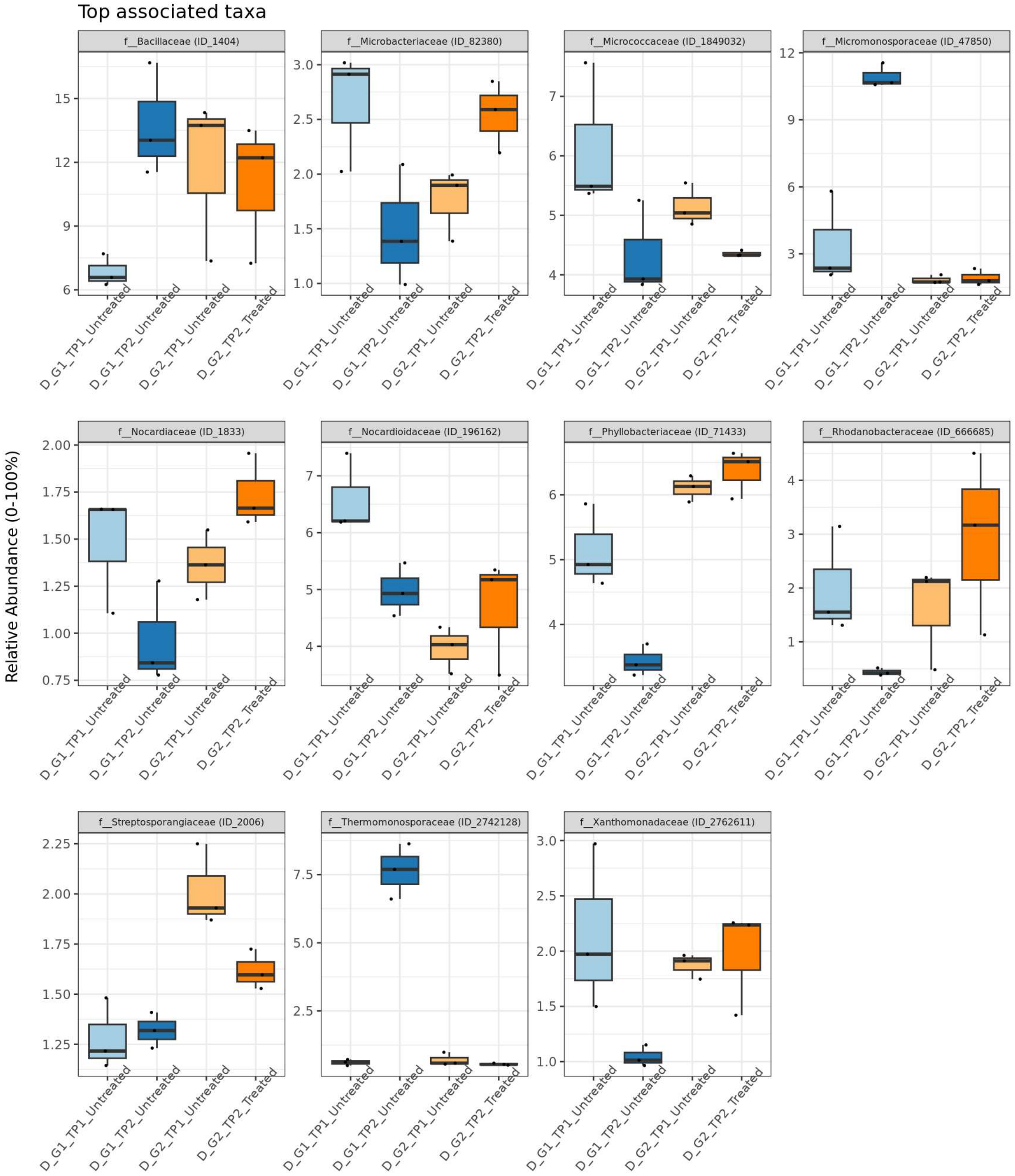

Figure 4: Differentially abundant taxa at family level. Top differentially abundant taxa are shown selected at a p-value<0.2.

Part conclusion of single taxa analyses:

Despite the limited power with three samples per group, the analyses detected quite some trends for interesting taxa. The pattern across the detected taxa, when looking at the boxplots, reflect the results form the ordination plots, of an effect of treatment that changes the microbes in a way that is very different form the effect of time, and in some cases appear to stabilize the microbiome. Again, we also see that there is a consistent difference between the two plots at TP1.

Differences in alpha-diversity

As described in **Report 2**, alpha diversity is a measure of the diversity within (or complexity within) one microbiome community. We here evaluate the two measures; Shannon and observed computed at the species level (a measure of richness). The measures are introduced in Report 2. We start with inspecting the summary statistics of the measures in the different subgroups of samples.

Group	n	median	mean	sd	min	max
D_G1_TP1_Untreated	3	5.682	5.704	0.057	5.660	5.769
D_G1_TP2_Untreated	3	5.524	5.492	0.113	5.367	5.586
D_G2_TP1_Untreated	3	5.572	5.608	0.093	5.539	5.714
D_G2_TP2_Treated	3	5.683	5.659	0.076	5.574	5.721

Table 3: Summary statistics of alpha diversity measure Shannon The table shows the different subgroups of samples, the number of samples in the group (n), the median and mean level in the group, the group standard deviation (sd) and minimum (min) and maximum (max) values.

Group	n	median	mean	sd	min	max
D_G1_TP1_Untreated	3	1548	1609.333	115.001	1538	1742
D_G1_TP2_Untreated	3	1506	1501.333	32.254	1467	1531
D_G2_TP1_Untreated	3	1365	1355.667	31.070	1321	1381
D_G2_TP2_Treated	3	1473	1469.667	44.095	1424	1512

Table 4: Summary statistics of alpha diversity measure Observed. The table shows the different subgroups of samples, the number of samples in the group (n), the median and mean level in the group, the group standard deviation (sd) and minimum (min) and maximum (max) values.

As for the single organisms, we use a linear regression with a so-called “change-score model” for the analysis of alpha diversity. The table below shows the effect of treatment within field D.

	Shannon				Observed			
	Estimate	Std. Error	t value	Pr(>|t|)	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.211	0.097	-2.179	9.48e-02	-108	65.755	-1.642	1.76e-01
TreatmentTreated	0.262	0.137	1.912	1.28e-01	222	92.991	2.387	7.54e-02

Table 5: Results from analysis of alpha diversity. The effect of treatment on alpha diversity within field D was analysed using a linear regression designed as a “change-score model”. Columns 1-4 gives results for Shannon diversity and columns 5-8 gives results for Observed (richness).

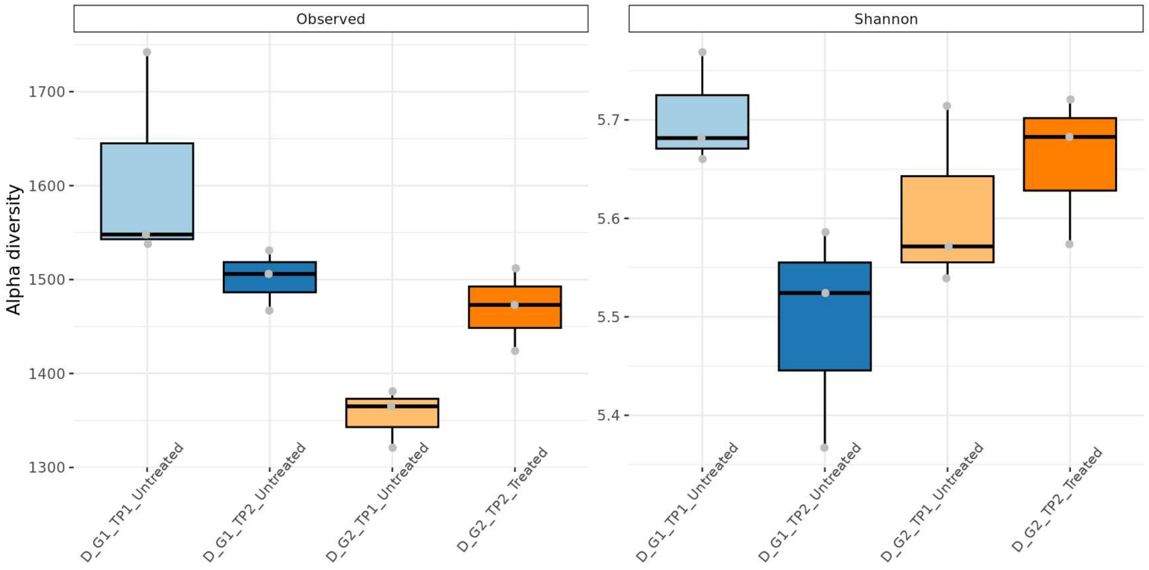

Figure 5: Illustration of the alpha diversity levels in each sample groups. The Observed and Shannon diversity measure is shown and the boxes are colored by sample group.

Part conclusion on alpha diversity:

The statistical analyses detected a trending difference in Shannon and Observed diversity as an effect of treatment. Time causes a decrease in diversity, and the treatment appear to cause a increase in diversity that overrules the time effect. Again, we also see that there is a consistent difference between the two plots at TP1.

Differences in percentage fungi

As described in **Report 2**, calculations are made of the relationship between fungi and bacteria in the samples. We here evaluate the percentage fungi detected in the samples as a measure of the percentage of the combined bacteria and fungi community that is fungi. We start with inspecting the summary statistics of the measures in the different subgroups of samples.

Group	n	median	mean	sd	min	max
D_G1_TP1_Untreated	3	4.443	4.385	0.601	3.758	4.955
D_G1_TP2_Untreated	3	3.850	4.177	0.620	3.789	4.892

Group	n	median	mean	sd	min	max
D_G2_TP1_Untreated	3	5.204	5.126	0.147	4.956	5.218
D_G2_TP2_Treated	3	3.981	4.076	1.486	2.639	5.606

Table 6: Summary statistics of percentage of fungi by treatment group and field The table shows the different subgroups of samples, the number of samples in the group (n), the median and mean level in the group, the group standard deviation (sd) and minimum (min) and maximum (max) values.

As for the single organisms, we use a linear regression with a “change-score model” for the analysis of percentage fungi. The table below shows results from comparing the fields at TP1 (Field diff), effects of treatment across fields (Across fields), and the effect of treatment within fields (Field Conv., Field Org1 and Field Org2).

Fungi_perc - Effect of treatment

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2	0.74	-0.3	0.794
TreatmentTreated	-0.8	1.05	-0.8	0.469

Table 7: Results from analysis of percentage of fungi. The effect of treatment on the percentage of fungi (row 2) was analysed using a linear regression designed as a “change-score model”.

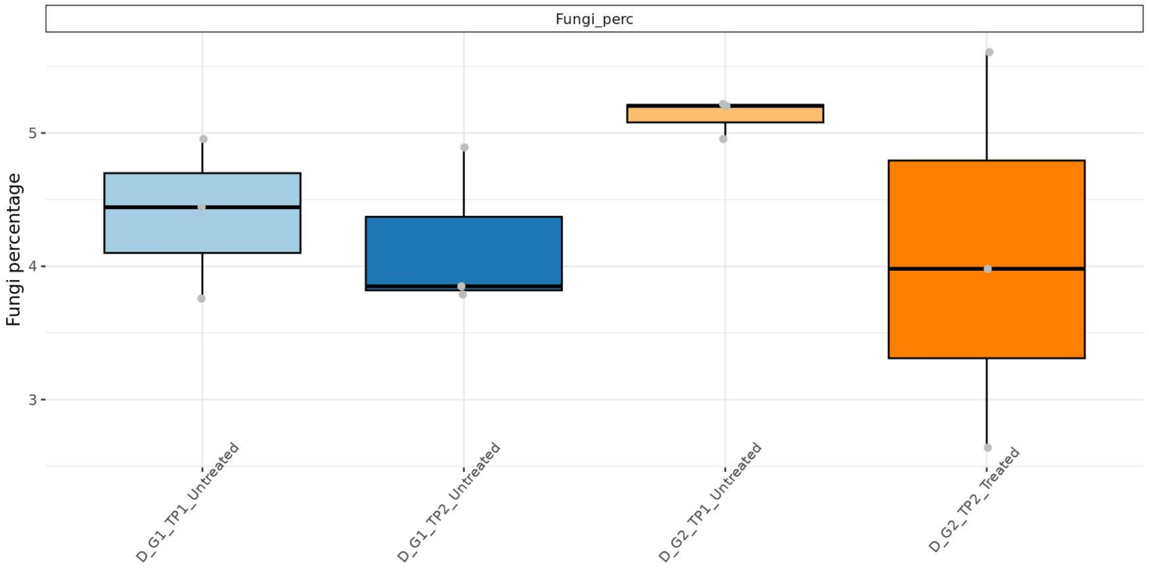

Figure 6: Illustration of the percentage of fungi levels in each sample. For each group of samples, the percentage of fungi measure is shown and the boxes are colored by sample group.

Part conclusion for percentage fungi:
We do not see an effect of time or treatment on the percentage of fungi in the samples.

Version information

Table 8: List of used software including the used R-programming environment packages.

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	**munsell**	0.5.0
R	4.2.0	**cellranger**	1.1.0
readxl	1.4.2	**tools**	4.2.0
backports	1.4.1	**cli**	3.6.0
systemfonts	1.0.4	**generics**	0.1.2
igraph	1.3.1	**ade4**	1.7-19
splines	4.2.0	**broom**	0.8.0
TH.data	1.1-1	**evaluate**	0.15
digest	0.6.29	**biomformat**	1.24.0
foreach	1.5.2	**fastmap**	1.1.0

Package	Version	Package	Version
htmltools	0.5.2	**yaml**	2.3.5
fansi	1.0.3	**fs**	1.5.2
magrittr	2.0.3	**zip**	2.2.0
cluster	2.1.3	**nlme**	3.1-157
tzdb	0.3.0	**xml2**	1.3.3
modelr	0.1.8	**compiler**	4.2.0
RcppParallel	5.1.5	**rstudioapi**	0.14
sandwich	3.0-1	**png**	0.1-7
svglite	2.1.0	**ggsignif**	0.6.3
timechange	0.2.0	**reprex**	2.0.2
jpeg	0.1-9	**bslib**	0.3.1
colorspace	2.0-3	**stringi**	1.7.6
rvest	1.0.2	**highr**	0.9
haven	2.5.0	**nloptr**	2.0.2
xfun	0.31	**microbiome**	1.18.0
crayon	1.5.1	**multtest**	2.52.0
RCurl	1.98-1.6	**vctrs**	0.5.2
survival	3.3-1	**pillar**	1.8.1
zoo	1.8-10	**lifecycle**	1.0.3
iterators	1.0.14	**rhdf5filters**	1.8.0
ape	5.6-2	**jquerylib**	0.1.4
glue	1.6.2	**estimability**	1.3
gtable	0.3.0	**cowplot**	1.1.1
zlibbioc	1.42.0	**bitops**	1.0-7
webshot	0.5.3	**R6**	2.5.1
DelayedArray	0.22.0	**latticeExtra**	0.6-29
car	3.0-13	**hwriter**	1.3.2.1
Rhdf5lib	1.18.2	**codetools**	0.2-18
abind	1.4-5	**boot**	1.3-28
mvtnorm	1.1-3	**MASS**	7.3-57
DBI	1.1.2	**assertthat**	0.2.1
rstatix	0.7.0	**rhdf5**	2.40.0
viridisLite	0.4.0	**withr**	2.5.0
xtable	1.8-4	**mnormt**	2.0.2
tmvnsim	1.0-2	**multcomp**	1.4-19
httr	1.4.5	**GenomeInfoDbData**	1.2.8
ellipsis	0.3.2	**mgcv**	1.8-40
farver	2.1.0	**parallel**	4.2.0
pkgconfig	2.0.3	**hms**	1.1.2
sass	0.4.1	**coda**	0.19-4
dbplyr	2.1.1	**minqa**	1.2.4
utf8	1.2.2	**rmarkdown**	2.14
labeling	0.4.2	**carData**	3.0-5
tidyselect	1.2.0	**Rtsne**	0.16
rlang	1.0.6	**ggpubr**	0.4.0
reshape2	1.4.4	**lubridate**	1.9.2