Microbiome Profiling Report

Taxonomic profiles

Biomcare ApS

28/12/2024

Customer	Tove Mariegaard Pedersen
Customer ID	DA00206-23
Project	Regenerativt landbrug til videreudvikling af den økologiske planteproduktion
Sample Type	Soil
Number of samples	14 samples
Type of data	Shotgun metagenomic sequencing

Overview of microbiome profiles

Processing of the shotgun sequencing data through quality control and microbiome profiling across the 14 samples resulted in the detection of 2345 unique species. When assigning the sequences at higher taxonomic levels a total of 722 genera, 268 families, 140 orders, 69 classes and 30 phyla were detected. At the kingdom level, 26 Archaea, 2278 bacteria, 28 Eukaryota and 13 viruses were detected.

The taxonomic profiles were generated using the software Kraken 2, which uses exact k-mer matching to map each sequence in the data to a known reference (see report 1 for more details). The result is a table containing the number of reads assigned to each taxa in each of the samples, and information on the number of reads that did not map to a reference genome. The data on unclassified reads helps us to evaluate how big a fraction of the sample is composed of microbes not found in the reference database. For the current dataset, on average 90.51% of reads are unclassified (min: 89.12; max: 91.85). For soil samples, a high percentage (>80%) of unclassified reads is expected as the soil microbiome is very complex and not well-described in the reference databases.

For the identified organisms we calculate both the relative abundance for each taxonomic level (excluding the unmapped reads) and their read counts. Reads classified by Kraken 2 at one taxonomic level, but not at the lower level, will be pushed down to species level by the software Bracken and therefore, there are no ‘unclassified’ clades appearing in the profiles unless the reference database has the reference listed as unclassified. The k-mer matching approach is a very sensitive species-identification approach, and with the large reference database used, many different species will be identified in most microbiome samples. The diversity of the samples will therefore be high with many of the species being rare. In the rarefied count data, a total of 2333 species have a mean relative abundance across samples of less than or equal to 1%, while 12 species have a higher mean relative abundance. It can therefore be of particular relevance to filter the data to remove very rare species that will also be the less confident calls across the many identified species.

Filtering of the dataset

The dataset is filtered to remove the samples with a minimum number of mapping reads of 1e+05 as identified in the count data. Microbiome profiles presented as read counts are rarefied to the number of reads found in the sample with the lowest read count after sample filtering (min read count of 1,991,921). In further processing and analysis, we have included 14 samples. The microbiome profile data has been transformed into different phyloseq objects:

- One phyloseq object for relative abundances
- One phyloseq object with raw counts
- One phyloseq object with rarefied counts

These are available in the project folder for easy loading into the R programming environment.

Summary of the phyloseq objects

```
## Overview of the microbiome profile before any filtering of samples and/or taxa:
```

```
##  - number of samples:
```

```
## [1] 14
```

```
##  - number of microbes:
```

```
## [1] 2346
```

```
## Overview of the microbiome profile after filtering of samples and/or taxa:
```

```
##  - number of samples:
```

```
## [1] 14
```

```
##  - number of microbes:
```

```
## [1] 2345
```

Visualization of the microbial communities

To support an effective evaluation of the microbiome profile in each sample, stacked barplots for each taxonomic level are provided in separate files (location: 3_Microbiome_profiles/3_Illustrations, names: StackedBarPlot_xxx.png"). Below you find two stacked barplots, for phyla and class level. The illustrations show the microbiome profiles of individual samples (up to 48 samples). If your project include more samples than is shown in the illustrations, please find the illustrations saved in your project folder. In your project folder, you will also find the stacked barplots of microbiome profiles at lower taxonomic levels.

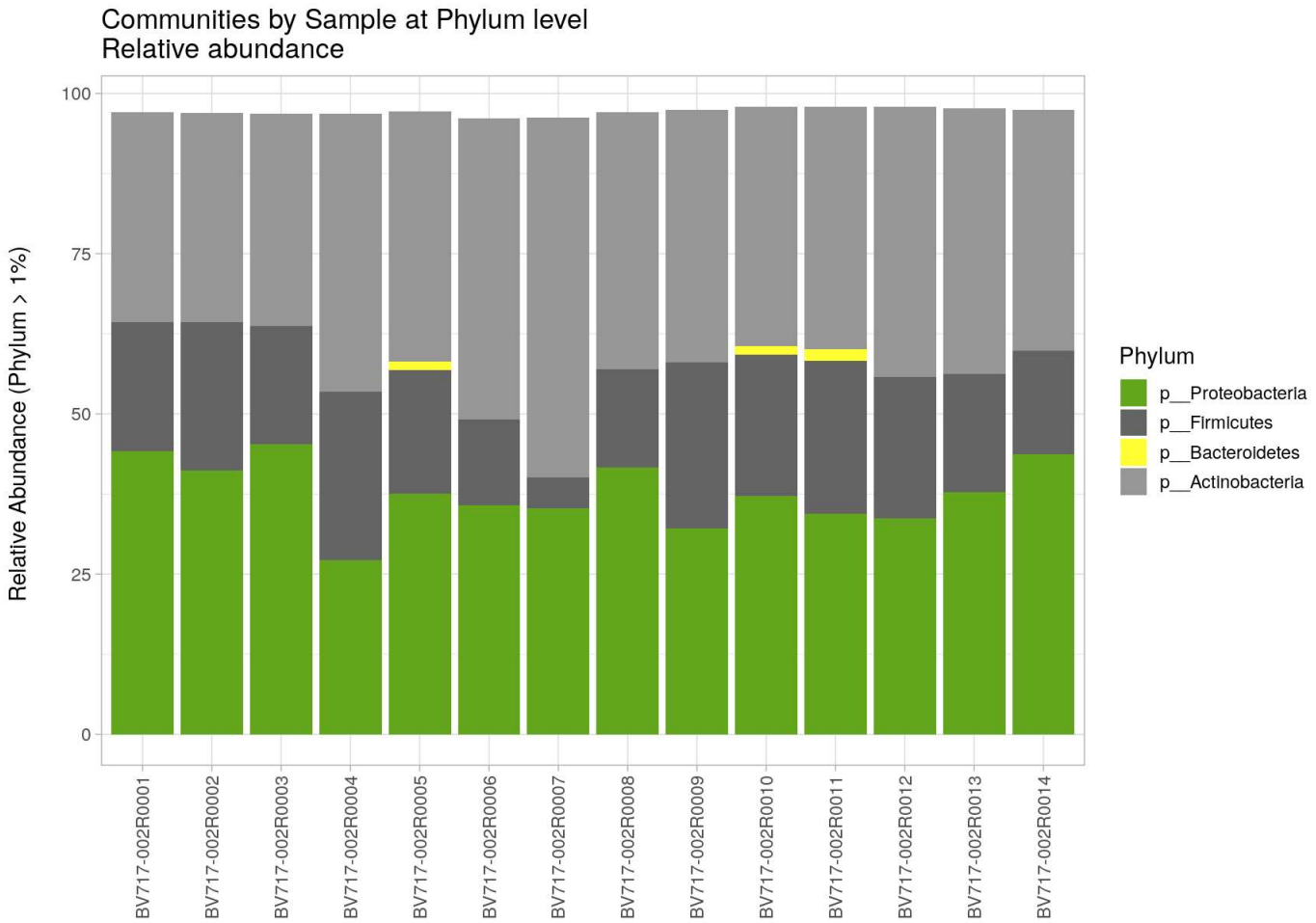

Figure 1: Stacked barplot of microbial communities at the taxonomic level of phylum. The top abundant phyla (comprising >1% of the community) were selected and shown with their relative abundances. The microbiome profiles generated using shotgun metagenomic sequencing include organisms in the sample that was found in the reference database. As such, there are no unknown organisms, and the data is scaled using detected organisms (leaving out unassigned reads). Note that some organisms are sparsely annotated and while having been annotated at species level might not have a known classification at higher levels. These species are causing the appearance of unnamed taxa in the figure such as ’c__’ or ’p__’.

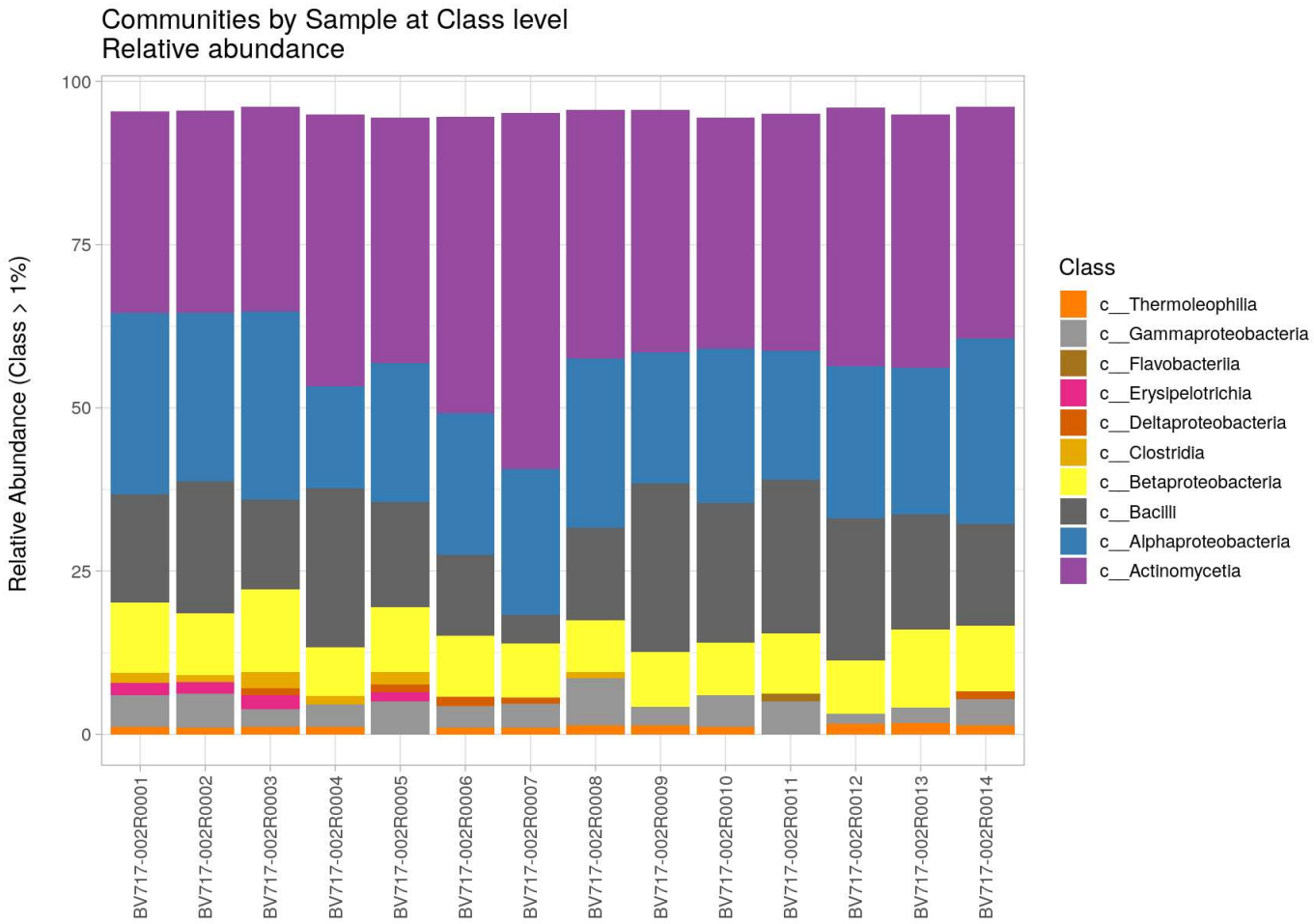

Figure 2: Stacked barplot of microbial communities at the taxonomic level of class. The top abundant classes (comprising >1% of the community) were selected and shown with their relative abundances. The microbiome profiles generated using shotgun metagenomic sequencing include organisms in the sample that was found in the reference database. As such, there are no unknown organisms, and the data is scaled using detected organisms (leaving out unassigned reads). Note that some organisms are sparsely annotated and while having been annotated at species level might not have a known classification at higher levels. These species are causing the appearance of unnamed taxa in the figure such as 'c__' or 'p__'.

Alpha-diversity

Alpha-diversity is a measure of the diversity (or complexity) within one microbiome community. Different samples' alpha-diversity scores are thus independent of any other sample's scores, unlike beta-diversity, described below. The alpha-diversity measures are calculated and saved in a matrix which can then be used for later statistical analysis and visualization (data is found in your project folder).

Different measures of alpha-diversity exist and each show slightly different aspects of the sample diversity. Among the most commonly used measures are Observed and Shannon, each providing information on different aspects of the microbiome as explained below.

One plot is provided for the two alpha diversity measures (Observed and Shannon), illustrating the diversity measures across samples. These plots are useful for distinguishing any outliers or trends in alpha-diversity across samples in the project.

Observed diversity is simply the number of observed species within the sample.

Shannon diversity reflects both richness and evenness of a microbiome community. The values increase as both richness and evenness increase. The uncertainty in determining the species of an individual explains diversity by the fact that if one species dominates the sample, it is easy to predict the species of an individual.

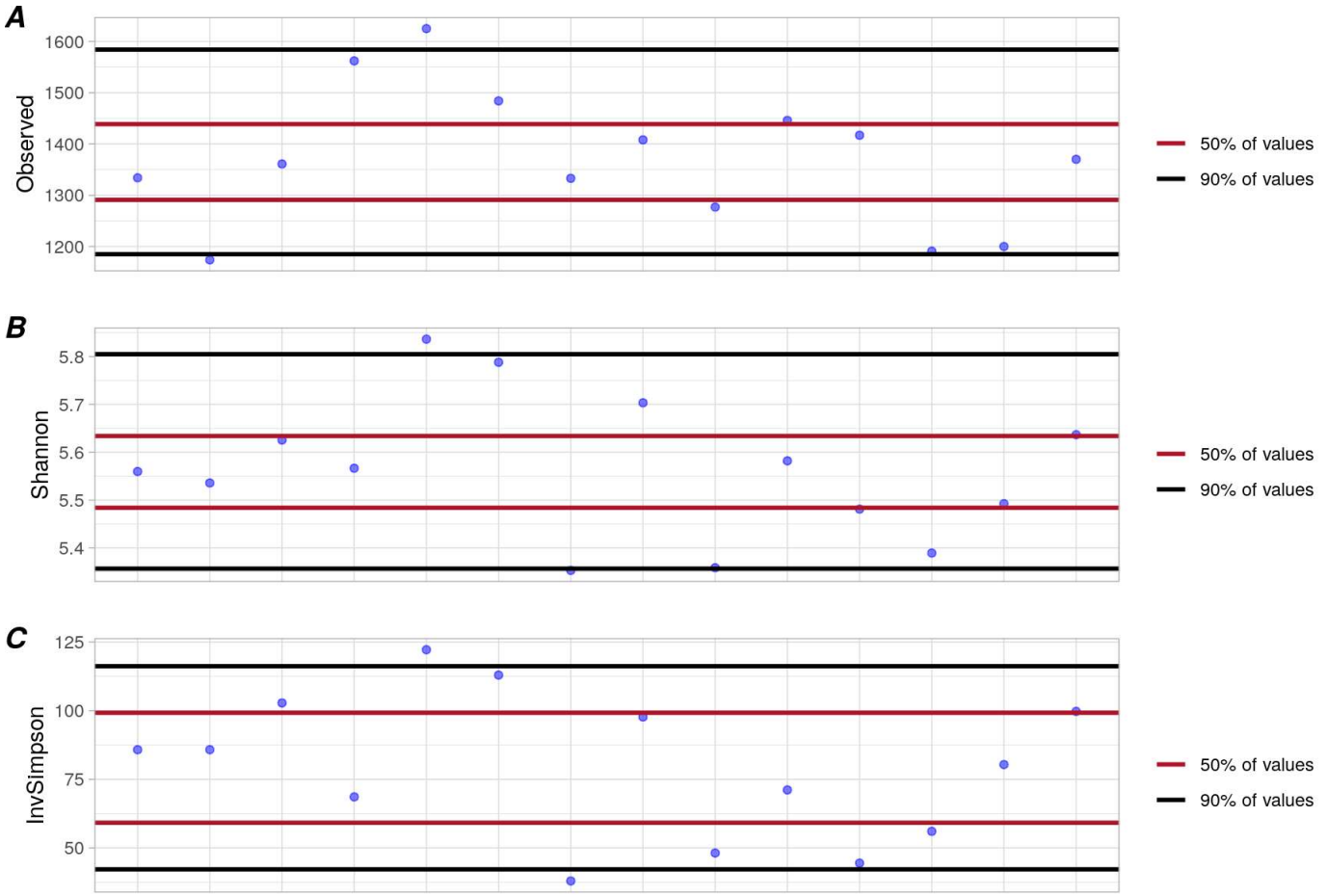

Figure 3: Observed diversity (A) and Shannon diversity (B) across samples. Rarefied count data was used for the calculation of observed diversity while count data was used for Shannon. 50% of sample points lie between the two red lines, while 90% of sample points lie between the two black lines.

Beta-diversity

Beta-diversity is a measure of the diversity (or complexity) of the microbiome community between samples and therefore differs from alpha-diversity which is a measure of the diversity within samples. Beta-diversity is a measure of how similar or dissimilar each pair of samples are. Two types of beta-diversity measures are calculated and stored in a distance matrix which can then be used for later statistical analysis and visualization.

Beta-diversity measures are often inspected in ordination plots, where each sample is a point and the distance between the points increases with increasing dissimilarity in the microbiome community. Such plots are useful in order to explore the data and see which variables explain the similarity or dissimilarity of groups of samples.

The two calculated beta-diversity measures are: *Bray-Curtis dissimilarity* and *Jaccard distance*. Bray-Curtis includes the taxa abundance in its algorithm, while Jaccard transforms the data to presence/absence (1 or 0). Using both measures allows us to evaluate if abundance is a more important driver than presence/absence of taxa in explaining the patterns in the data. We calculate both beta-diversity measures for the relative abundance data (normalized by total read count per sample) at multiple taxonomic levels. The summary statistics for the calculated matrices are listed below and can be used to evaluate the overall dissimilarity pattern across the dataset. Both Bray-Curtis and Jaccard range from 0 to 1, with 1 indicating maximum possible difference between samples (no shared sequences). A mean value for a beta-diversity matrix that is close to 1 therefore indicates high dissimilarity between all samples.

Summary statistics of the beta-diversity matrices

```
## Species
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3628  0.4629  0.4975  0.4962  0.5224  0.5998
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2041  0.3222  0.3894  0.3879  0.4491  0.6012
```

```
## Genera
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3220  0.4171  0.4383  0.4431  0.4680  0.5339
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1423  0.2277  0.2791  0.2872  0.3366  0.4698
```

```
## Family
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2368  0.3511  0.3784  0.3797  0.4101  0.4919
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1101  0.1847  0.2320  0.2410  0.2893  0.4327
```

```
## Order
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2034  0.3233  0.3582  0.3594  0.3972  0.5000
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.08268 0.14294 0.16947 0.18164 0.21584 0.33082
```

```
## Class
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1875  0.3189  0.3529  0.3571  0.4051  0.4878
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04685 0.09197 0.11261 0.12663 0.15930 0.25384
```

```
## Phyla
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09524 0.29630 0.33333 0.33351 0.40000 0.48485
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.01835 0.06145 0.08803 0.10191 0.12730 0.24417
```

Nonmetric Multidimensional Scaling (NMDS)

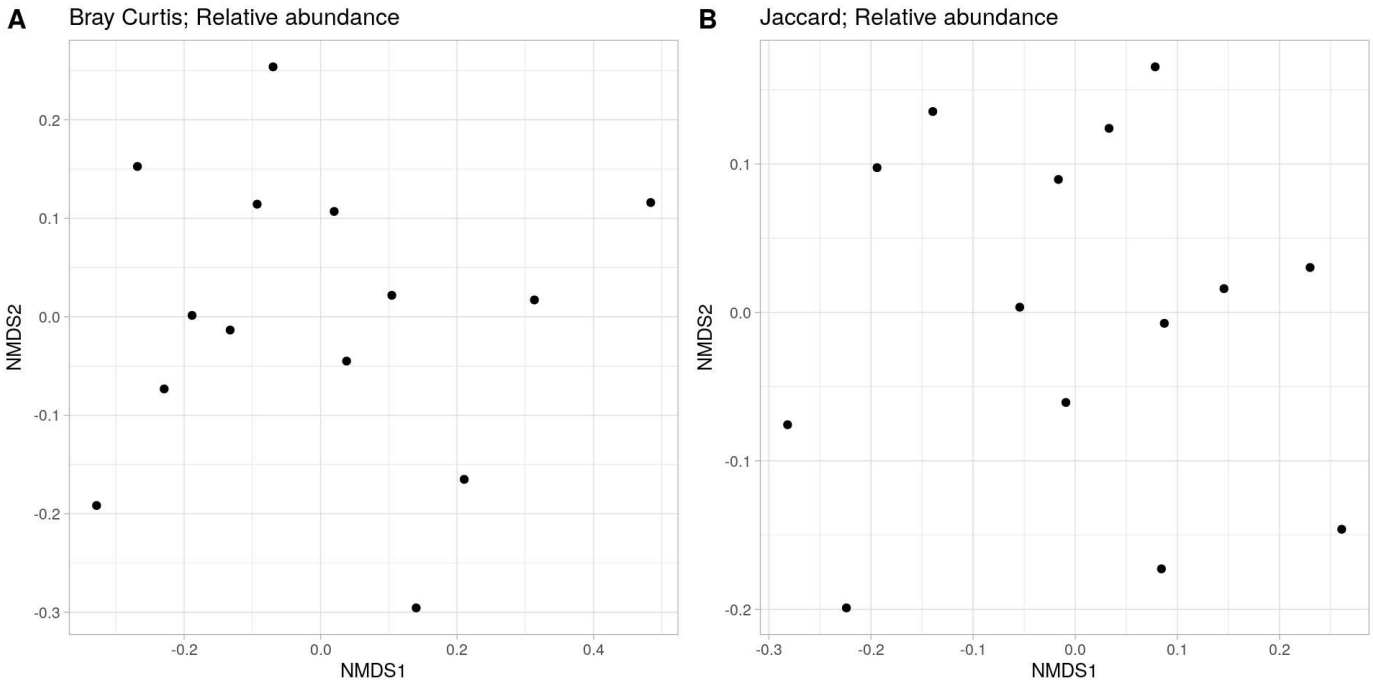

Figure 4: Nonmetric multidimensional scaling illustrating the inter-sample relationship in therms of their microbial community composition. Bray-Curtis (A) and Jaccard distance (B) calculated for the genera level of the microbiome communities.

Bacterial-to-fungal abundance ratio

The bacterial-to-fungal abundance ratio of each sample was computed as described in “Report 1: Sequencing and Data Processing Report”. In short, the ratios were calculated based on the number of Small Subunit (SSU) reads assigned to either bacteria or fungi. Summary statistics for the results can be seen in the following table.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SSU reads assigned as bacteria	46706	54105	55269	55365	57153	62483

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
SSU reads assigned as fungi	1282	1722	1979	2227	2712	3754
FB_ratio	0.025	0.030	0.037	0.040	0.049	0.061
BF_ratio	16.331	20.412	27.226	27.150	33.647	40.379
Percentage fungal reads	2.417	2.886	3.544	3.862	4.672	5.770

Table 1: Summary statistics for bacterial-to-fungal abundance ratios across samples. It was possible to detect fungal SSU reads in 14 samples. For these samples, the ratios were estimated and the summary statistics for the obtained results is presented in the table. The “Percentage fungal reads” is the percent of bacterial+fungal reads that are fungal reads

Version information

Table 2: List of used software including the used R-programming environment packages.

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	**tidyselect**	1.2.1
R	4.3.3	**rstudioapi**	0.16.0
bitops	1.0-7	**abind**	1.4-5
rlang	1.1.4	**yaml**	2.3.9
magrittr	2.0.3	**codetools**	0.2-20
ade4	1.7-22	**plyr**	1.8.9
compiler	4.3.3	**Biobase**	2.62.0
mgcv	1.9-1	**withr**	3.0.0
systemfonts	1.1.0	**evaluate**	0.24.0
vctrs	0.6.5	**survival**	3.7-0
reshape2	1.4.4	**zip**	2.3.1
pkgconfig	2.0.3	**xml2**	1.3.6
crayon	1.5.3	**Biostrings**	2.70.3
fastmap	1.2.0	**pillar**	1.9.0
backports	1.5.0	**carData**	3.0-5
XVector	0.42.0	**foreach**	1.5.2
labeling	0.4.3	**stats4**	4.3.3
utf8	1.2.4	**generics**	0.1.3
rmarkdown	2.27	**RCurl**	1.98-1.16
tzdb	0.4.0	**S4Vectors**	0.40.2

Package	Version	Package	Version
xfun	0.46	**hms**	1.1.3
zlibbioc	1.48.2	**munsell**	0.5.1
cachem	1.1.0	**glue**	1.7.0
GenomeInfoDb	1.38.8	**tools**	4.3.3
jsonlite	1.8.8	**ggsignif**	0.6.4
biomformat	1.30.0	**cowplot**	1.1.3
highr	0.11	**rhdf5**	2.46.1
rhdf5filters	1.14.1	**ape**	5.8
Rhdf5lib	1.24.2	**colorspace**	2.1-0
broom	1.0.6	**nlme**	3.1-165
parallel	4.3.3	**GenomeInfoDbData**	1.2.11
cluster	2.1.6	**cli**	3.6.3
R6	2.5.1	**fansi**	1.0.6
bslib	0.7.0	**viridisLite**	0.4.2
stringi	1.8.4	**svglite**	2.1.3
car	3.1-2	**gtable**	0.3.5
jquerylib	0.1.4	**rstatix**	0.7.2
Rcpp	1.0.13	**sass**	0.4.9
iterators	1.0.14	**digest**	0.6.36
knitr	1.48	**BiocGenerics**	0.48.1
IRanges	2.36.0	**farver**	2.1.2
Matrix	1.6-5	**htmltools**	0.5.8.1
splines	4.3.3	**multtest**	2.58.0
igraph	2.0.3	**lifecycle**	1.0.4
timechange	0.3.0	**MASS**	7.3-60.0.1