Microbiome Profiling Report

Biomcare ApS

27/12/2024

Customer	Tove Pedersen
Customer ID	DA00206-23
Project	Processing and analysis of 16S region
Sample Type	Soil
Number of samples	14 samples
Type of data	Sequencing of 16S region

Overview of microbiome profiles

Processing of the bacterial 16S rRNA gene sequencing data across the 14 samples resulted in the detection of 78,425 unique bacterial sequences and 1408 unique archaeal sequences. When assigning the sequences at higher taxonomic levels, a total of 898 genera, 371 families, 263 orders, 126 classes and 49 phyla were detected. At the lower taxonomic levels, an increasing number of sequences did not match a known microorganism in the reference database, potentially due to the high complexity of soil samples with many organisms only now getting identified and characterized (e.g. 55,996 unassigned sequences at the genus level).

The mean read depth across samples was 64,755 with a minimum assigned read count of 45,345. In total, 14 out of 14 samples samples passed a requirement for a minimum read count to be above 5,000 reads for further processing and analysis.

Summary statistics of the microbiome profiles across all samples, and the changes imposed at each step of filtering are listed here:

```
## Overview of the microbiome profile before filtering. At this step, only the quality-contro
l samples have been removed.
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 79905 taxa and 14 samples ]
## sample_data() Sample Data:       [ 14 samples by 20 sample variables ]
## tax_table()   Taxonomy Table:    [ 79905 taxa by 6 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 79905 reference sequences ]
```

```
## Overview of the microbiome profile after
##     - filtering of samples to remove any sample with a total read count less than 5000.
##     - filtering to remove reads where primers have mis-aligned (chloroplast or mitochondri
a).
##     - if spike-in was used to obtain total abundance, the two spiked bacteria were removed
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 79833 taxa and 14 samples ]
## sample_data() Sample Data:       [ 14 samples by 20 sample variables ]
## tax_table()   Taxonomy Table:    [ 79833 taxa by 6 taxonomic ranks ]
## refseq()      DNAStringSet:      [ 79833 reference sequences ]
```

Visualization of the bacterial 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/Illustrations, names: StackedBarPlot_xxx.png"). Below you find two stacked barplots, for phyla and class level profiles. The illustrations show the microbiome profiles of individual samples (up to 40 samples). If your project includes more samples than included in the illustrations, please find the illustrations saved in your project folder.

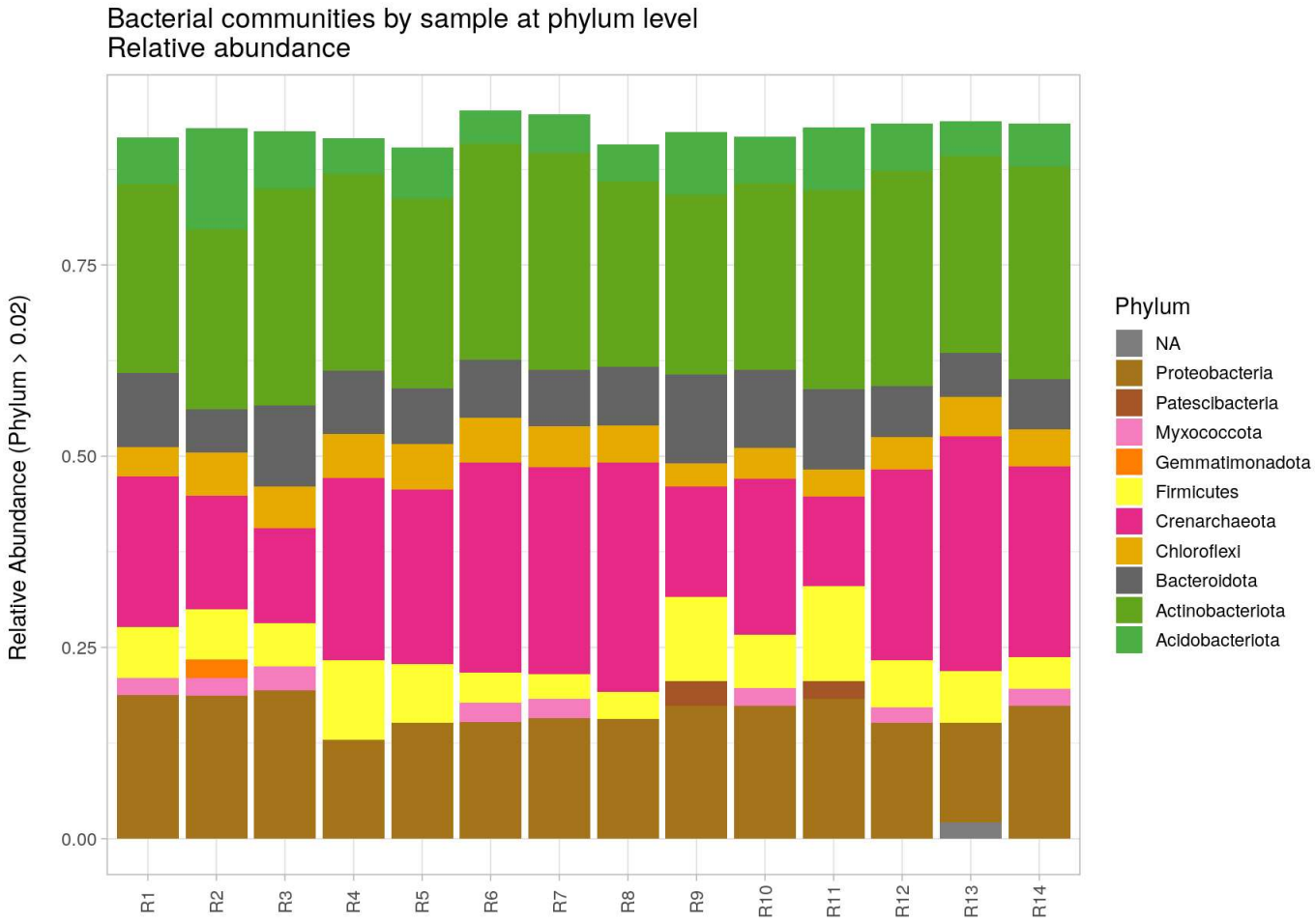

Figure 1: Stacked barplot of bacterial communities at the taxonomic level of phylum. The top abundant phyla (comprising >2% of the community) were selected and shown with their relative abundances. Occasionally, some sequences cannot be assigned to a known microorganism in the reference database and in these cases, "NA" indicates that the sequences could not be assigned at the phylum level.

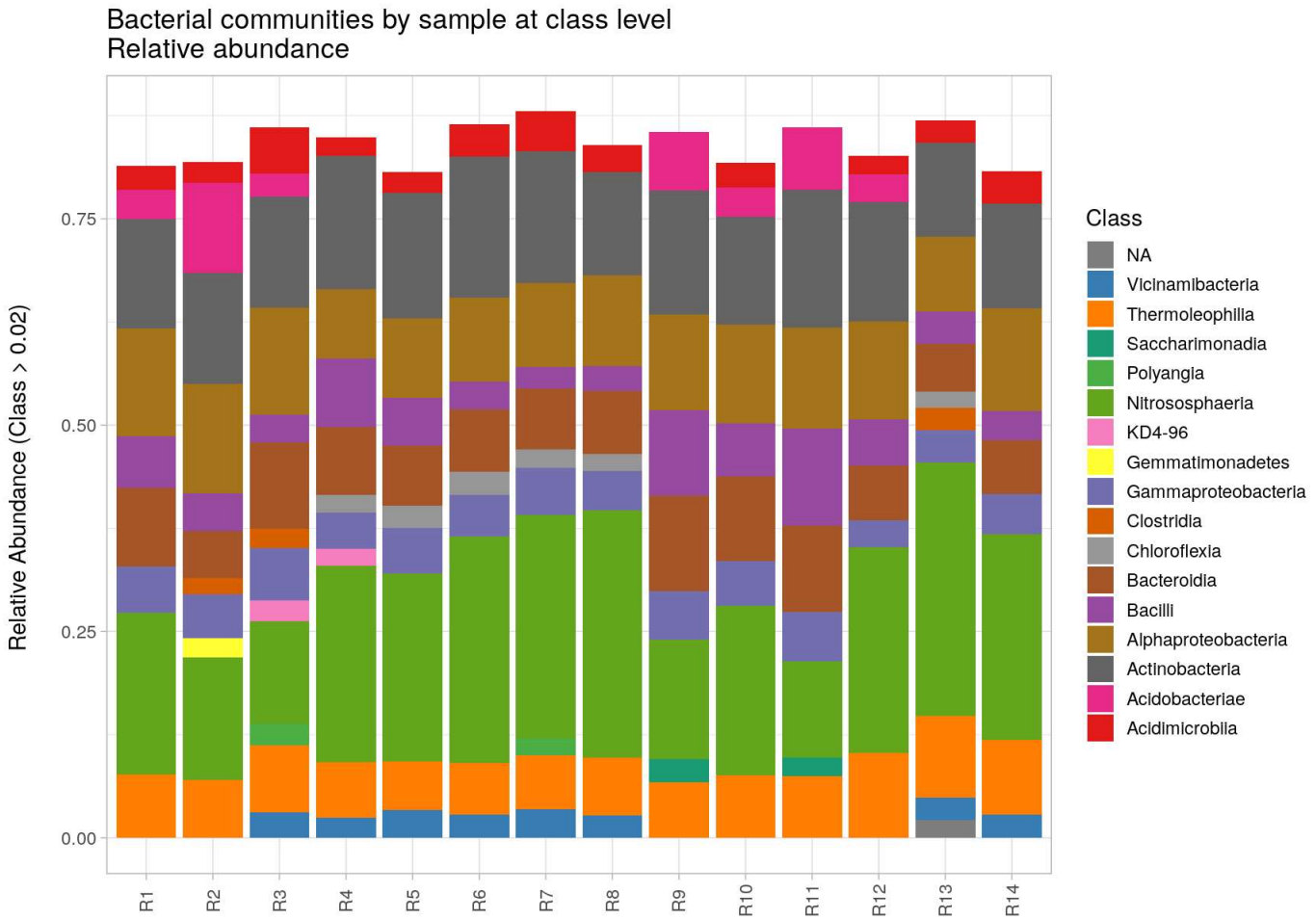

Figure 2: Stacked barplot of bacterial communities at the taxonomic level of class. The top abundant classes (comprising >2% of the community) were selected and shown with their relative abundances. Occasionally, some sequences cannot be assigned to a known microorganism in the reference database and in these cases, “NA” indicates that the sequences could not be assigned at the class level.

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 score, 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.

Different measures of alpha-diversity exist and each show slightly different aspects of the sample diversity. Among the most commonly used measures are the Observed (also called richness) and Shannon diversity measures. Each of these provide information on different aspects of the microbiome as explained below.

For these two measures, a plot is provided 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 microbes within the sample and is often called richness.

Shannon diversity reflects both richness and evenness of a microbiome community. It reflects the probability of predicting a random species from the dataset, and, thus, increases the more one species dominates the sample.

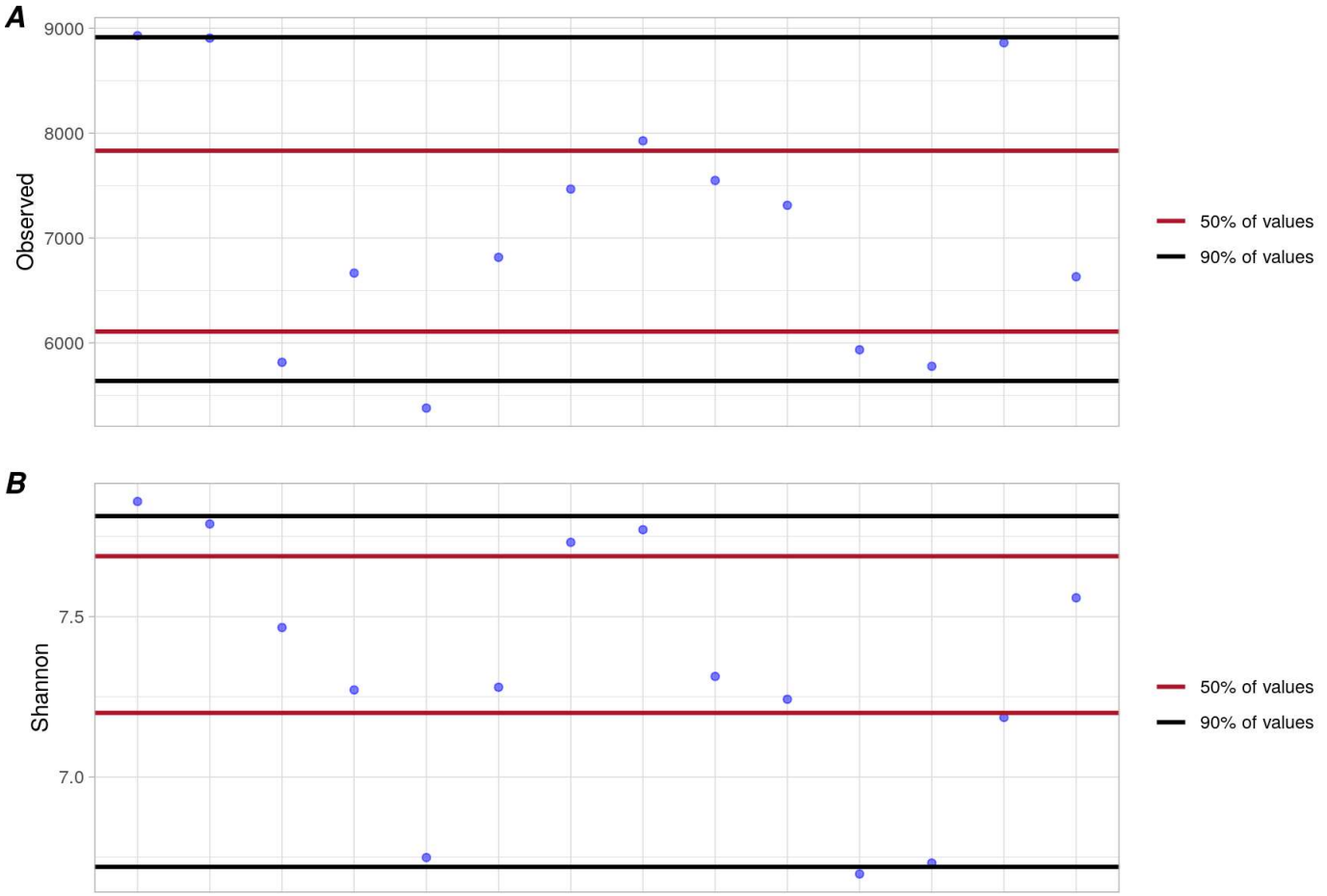

Figure 3: Observed diversity (A) and Shannon diversity (B) across samples. Rarefied abundance data was used for the calculation of observed diversity while count data was used for Shannon. 50% of sample points lie between the two red lines, while 90% of sample points lie between the two black lines.

A number of additional alpha diversity measures are calculated and saved to the alpha diversity file so that these are easily available if of interest in further analyses.

Correlation of alpha diversity metrics

It is not sensible to analyse all possible alpha diversity measures and thus one or a few must be selected for further analyses. Many alpha diversity measures will correlate and it can be ideal to select one that represent the general diversity or a few that capture different aspects of the diversity landscape.

```
##             Observed      Chao1   Shannon InvSimpson  Evenness
## Observed   1.0000000 0.97802198 0.6571429 0.17802198 0.2615385
## Chao1      0.9780220 1.00000000 0.5384615 0.05934066 0.1296703
## Shannon    0.6571429 0.53846154 1.0000000 0.81978022 0.8549451
## InvSimpson 0.1780220 0.05934066 0.8197802 1.00000000 0.9780220
## Evenness   0.2615385 0.12967033 0.8549451 0.97802198 1.0000000
```

Table 1: Correlation matrix for the calculated measures using the Spearman correlation method.

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. 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.

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. The two calculated beta-diversity measures are: Bray-Curtis dissimilarity and binary Jaccard distance.

Bray-Curtis includes the taxa abundance in its algorithm, while the binary Jaccard transforms the data to presence/absence (1 or 0). Using both measures allow 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

Relative abundance

```
## Sequences
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.9133  0.9422  0.9514  0.9530  0.9640  0.9813
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3870  0.6364  0.7192  0.7208  0.8068  0.9258
```

```
## Genera
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.4453  0.5431  0.5728  0.5726  0.5981  0.6676
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07486 0.23570 0.30524 0.31812 0.40981 0.52742
```

```
## Family
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3985  0.4783  0.5055  0.5072  0.5398  0.6058
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05275 0.19369 0.25735 0.26848 0.35407 0.46010
```

```
## Order
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3390  0.4389  0.4671  0.4705  0.4969  0.5825
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.04449 0.17045 0.23131 0.24032 0.32864 0.41094
```

```
## Class
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.3390  0.4245  0.4559  0.4567  0.4884  0.5732
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.03057 0.13627 0.16213 0.17695 0.22196 0.35228
```

```
## Phyla
```

```
## $jaccard
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.1364  0.3261  0.3810  0.3809  0.4331  0.5667
##
## $bray
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.02002 0.09529 0.11988 0.12864 0.16364 0.23451
```

Nonmetric Multidimensional Scaling (NMDS)

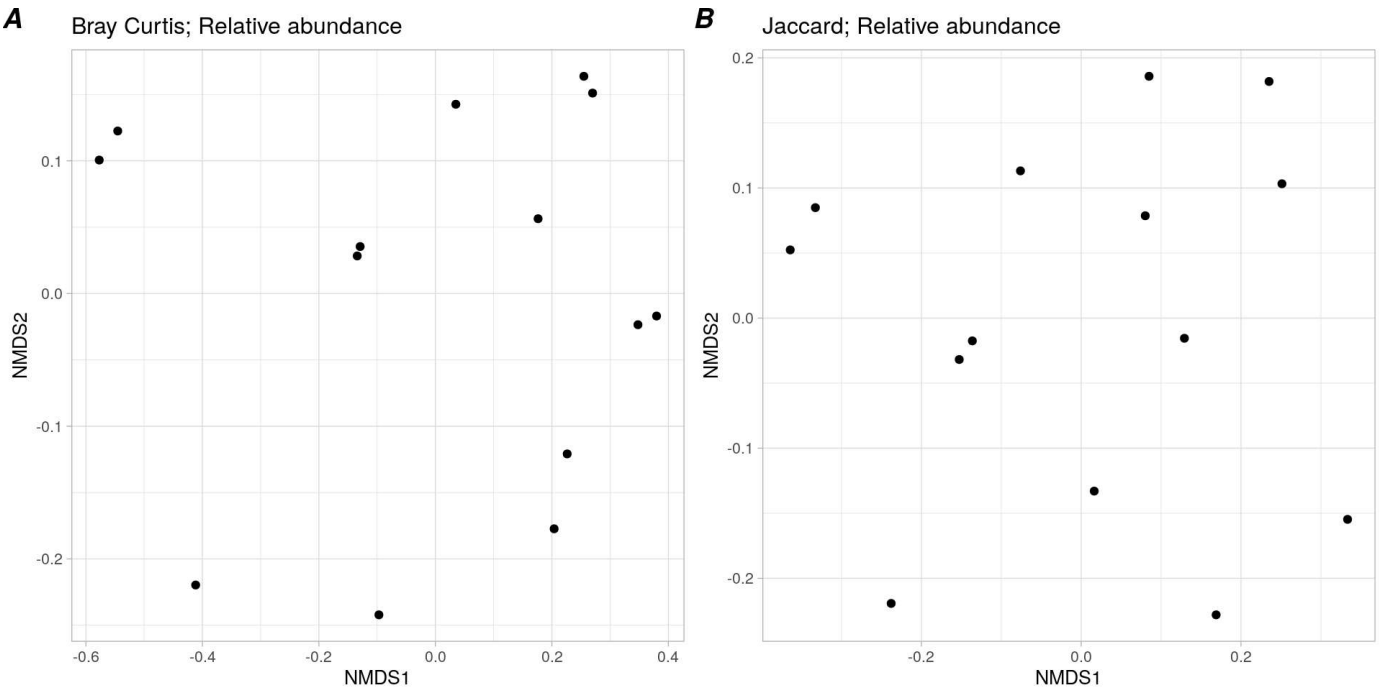

Figure 4: Community composition of project samples for further biostatistical analysis. Nonmetric multidimensional scaling over Bray-Curtis (A) and Jaccard distance (B) calculated for the genera level of the microbial communities. Here, there is no indication of specific sample names (see Report 3 for a plot including the sample names). The aim of the illustrations here is to inspect the general structure of the microbiome data and it allows us to identify if oddities are found in the data, such as unexpected shapes and outlines."

Version information

Table 1: List of used software including the used R-programming environment packages.

Package	Version	Package	Version
OS	Ubuntu 20.04.4 LTS	**timechange**	0.3.0
R	4.3.3	**tidyselect**	1.2.1
bitops	1.0-7	**rstudioapi**	0.16.0
rlang	1.1.4	**abind**	1.4-5
magrittr	2.0.3	**yaml**	2.3.9
ade4	1.7-22	**codetools**	0.2-20
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
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

Package	Version	Package	Version
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