Diversity analysis of fecal metagenomics dataset

Ece Kartal

5.10.2023

Results

In this dataset, I analyzed fecal shotgun samples from NSCLC cancer patients. The total read count in these samples ranged from min(meta$'ReadCount(Microbiome)')tomax(meta$'ReadCount (Microbiome)') million. The overall host DNA contamination and low-quality reads were less than 1%.

```
## Warning in geom_col(stat = "identity", alpha = 0.5, position =
## position_dodge2(width = 0.9, : Ignoring unknown parameters: ‘stat‘
```

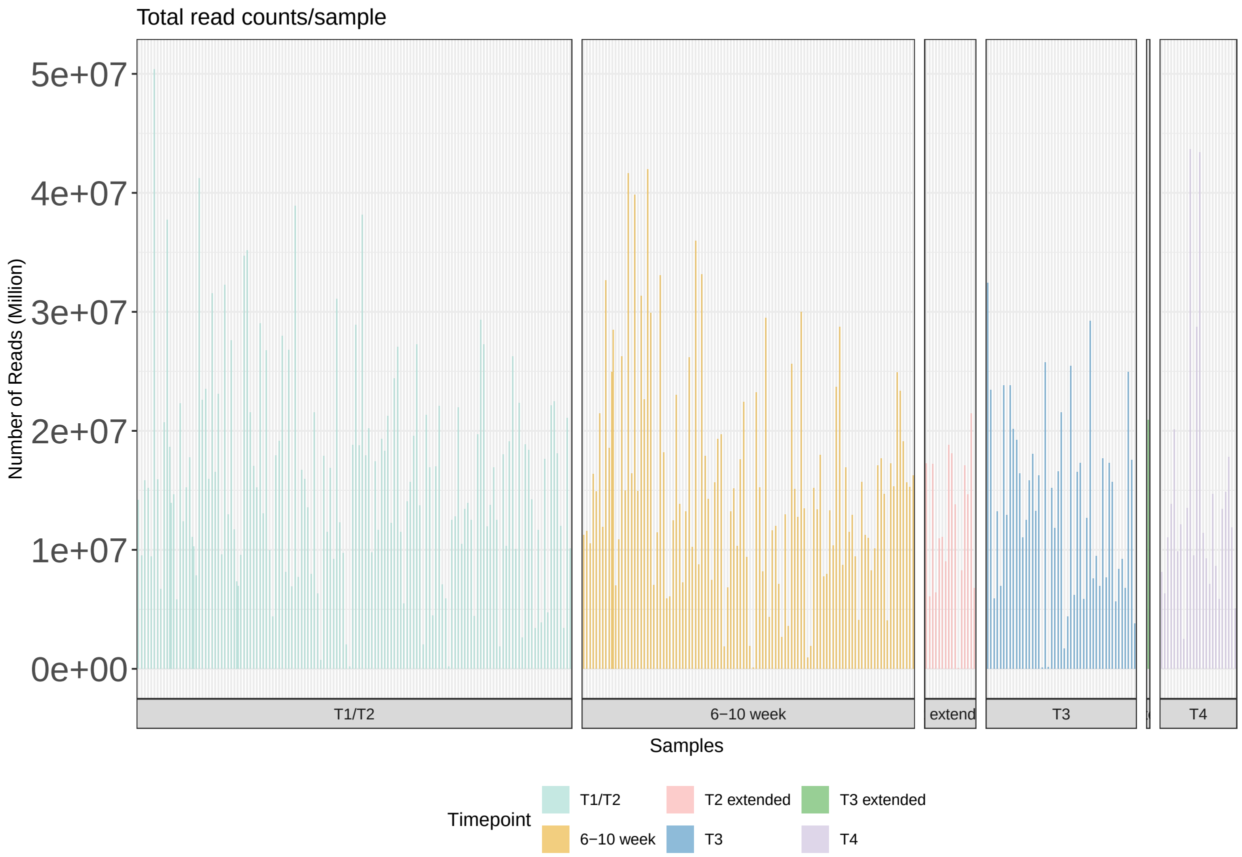


Figure 1. Total read count per sample

These results show that we can pool samples except "Promo106". A significant p-value will mean that, in terms of composition, variation within-duplicates will be smaller than variation between the duplicates and all other samples, so both duplicates should be equivalent

Diversity Analysis

Diversity in the ecological sense is intuitively understood as the complexity of a community of organisms.The two main categories of methods are known as **alpha diversity** and **beta diversity**

```
## [1] 2215  326
```

```
## [1] 2215  326
```

```
## [1] 2215  288
```

Alpha (within sample) Diversity Analysis

Alpha diversity measures the diversity within a single sample and is generally based on the number and relative abundance of taxa at some rank **Shannon**: How difficult it is to predict the identity of a randomly chosen individual. **Simpson**: The probability that two randomly chosen individuals are the same species. **Inverse Simpson**: This is a bit confusing to think about. Assuming a theoretically community where all species were equally abundant, this would be the number of species needed to have the same Simpson index value for the community being analyzed.

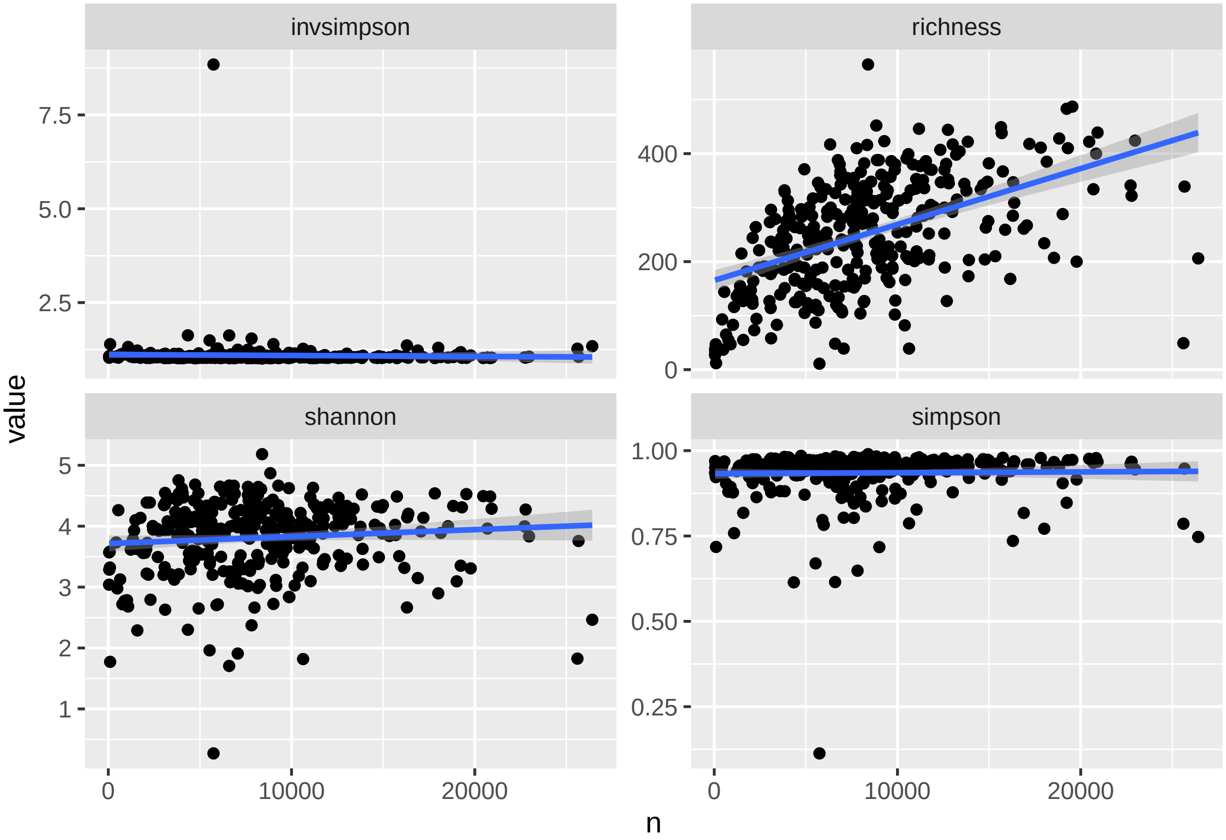
The `diversity` function from the vegan package can be used to calculate the alpha diversity of a set of samples.

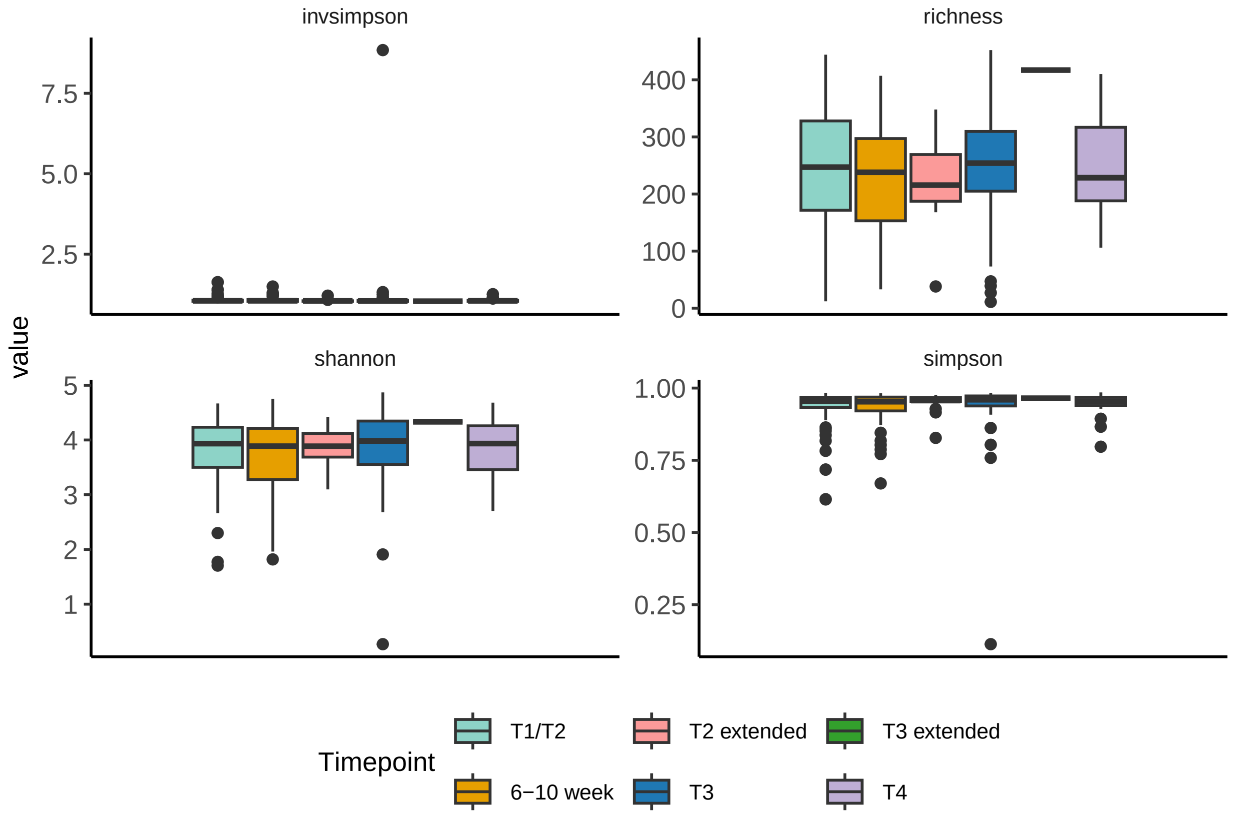
Figure 2. Comparison of alpha diversity measurements with read counts.

As we can see, a higher read count increases the likelihood of detecting a greater diversity of species in the sample (richness).

We will use `rrarefy` function from vegan package for rarefaction

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```


invsimpson
richness
shannon
simpson
value
n
7.5
5.0
2.5
400
200
0
5
4
3
2
1
1.00
0.75
0.50
0.25
0
10000
20000
0
10000
20000



```
## Warning in rrarefy(motu.abs, min_seq): some row sums < 'sample' and are not
## rarefied
```

In general, you will see roughly normal distribution for Shannon's diversity as well as most richness metrics. Simpson's diversity, on the other hand, is usually skewed. So most will use inverse Simpson (1/Simpson) instead. This not only increases normalicy but also makes the output more logical as a higher inverse Simpson value corresponds to higher diversity.

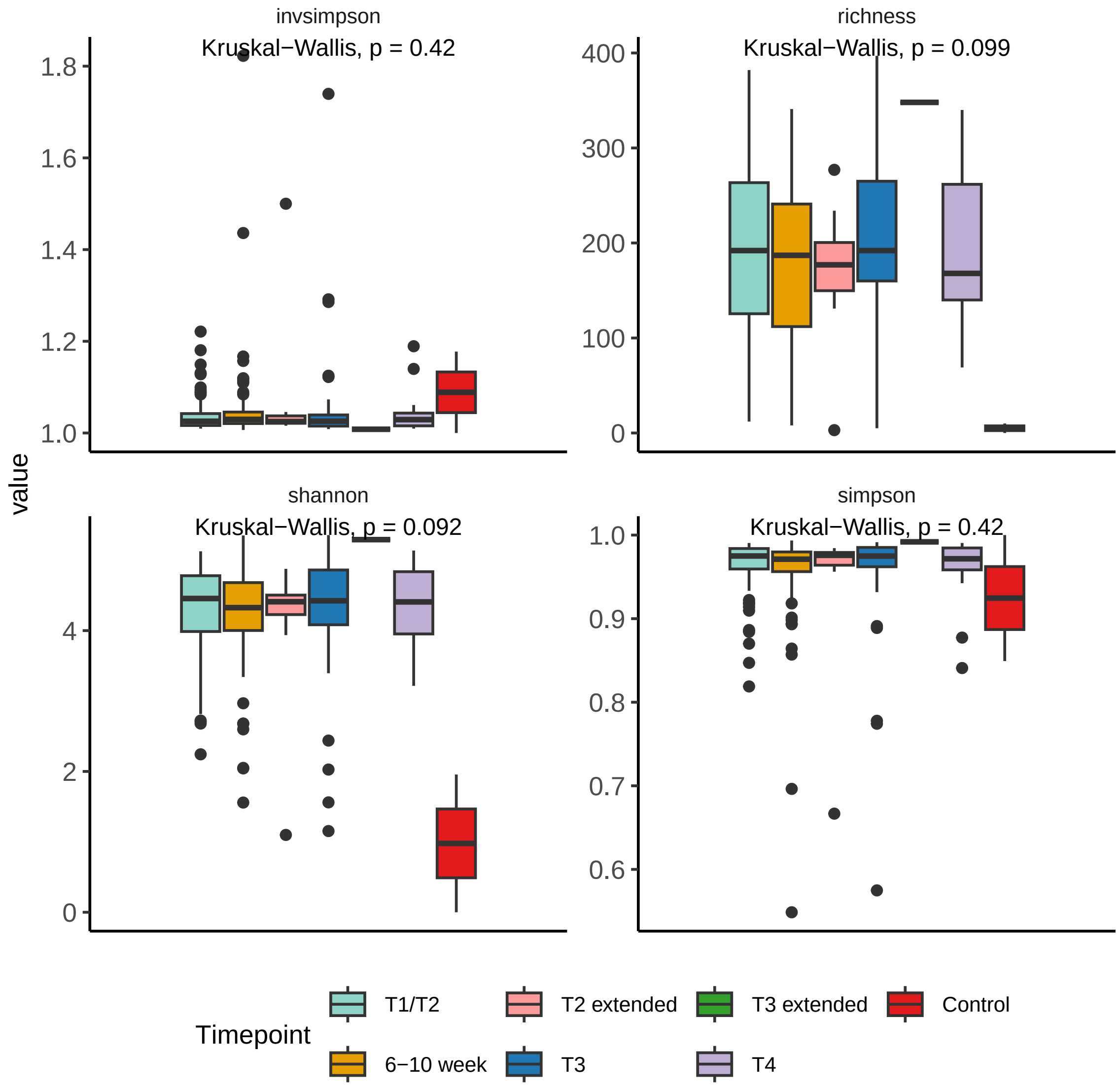


Figure 3. Comparison of alpha diversity measurements between non-rarefied and rarefied reads.

As seen in the figure, when we rarefy samples, we loose a lot of low abundant species, that would change the diversity measurements.Samples with high read counts by default have a higher species richness that we should be aware when comapring richness between 2 conditions

```
## Warning in printHypothesis(L, rhs, names(b)): one or more coefficients in the hypothesis include
##      arithmetic operators in their names;
##   the printed representation of the hypothesis will be omitted
```

```
## # A tibble: 9 x 7
##   term                 sumsq    df statistic      p.value metric         p.adj
##   <chr>                <dbl> <dbl>     <dbl>        <dbl> <chr>          <dbl>
## 1 CHEMO                2543.     1     0.213  0.647       richness 1
## 2 IMMUN                1548.     1     0.129  0.721       richness 1
## 3 TARGET              11030.     1     0.939  0.338       richness 1
## 4 Age                703525.    30     2.81   0.0000228   richness 0.000183
## 5 Sex                12664.     1    1.16   0.284        richness 1
## 6 Vital_status         610.     1    0.0554 0.814        richness 1
## 7 Timepoint          56781.     5    1.18   0.317        richness 1
## 8 days_to_death    1203642.    62    3.10   0.0000000990 richness 0.000000891
## 9 ReadCount (Homo) 2669668.   275    1.37   0.306        richness 1
```

Here we can see the significance of meta variables over the richness metric. As seen, `Age` and `days_to_death` show a high significant p value which mean that affects the richness of samples.

Beta (between sample) Diversity Analysis

Beta diversity is a way to quantify the difference between two communities. There are many metrics that are used for this, but we will only mention a few of the more popular ones.

- Indexes used with presence/absence data: *Jaccard*: the number of species common to both communities divided by the number of species in either community. *Unifrac*: The fraction of the phylogenetic tree branch lengths shared by the two communities.
- Indexes used with count data: *Bray–Curtis*: The sum of lesser counts for species present in both communities divided by the sum of all counts in both communities. This can be thought of as a quantitative version of the Sørensen index. *Weighted Unifrac*: The fraction of the phylogenetic tree branch lengths shared by the two communities, weighted by the counts of organisms, so more abundant organisms have a greater influence.

The vegan function `vegdist` is used to compute dissimilarity indexes. **Bray-Curtis** takes into account species presence/absence, as well as abundance, whereas other measures (like Jaccard) only take into account presence/absence and UniFrac incorporates phylogenetic information. Due to this reason, I used the Bray-Curtis metric in this analysis

```
## Run 0 stress 0.2448269
## Run 1 stress 0.2486586
## Run 2 stress 0.247666
## Run 3 stress 0.2484668
## Run 4 stress 0.2550776
## Run 5 stress 0.2490407
## Run 6 stress 0.2499789
## Run 7 stress 0.2616178
## Run 8 stress 0.2521451
## Run 9 stress 0.2458873
## Run 10 stress 0.25272
## Run 11 stress 0.2449896
## ... Procrustes: rmse 0.01636624  max resid 0.14451
## Run 12 stress 0.2479203
## Run 13 stress 0.2569692
## Run 14 stress 0.2482055
## Run 15 stress 0.2460981
## Run 16 stress 0.2490347
## Run 17 stress 0.2459354
## Run 18 stress 0.2546772
## Run 19 stress 0.2513102
## Run 20 stress 0.2459402
## *** Best solution was not repeated -- monoMDS stopping criteria:
##     12: no. of iterations >= maxit
##      8: stress ratio > sratmax
## Too few points to calculate an ellipse
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_path()‘).
```

Calculating Rarefied Beta Diversity `avgdist` function computes the dissimilarity matrix of a dataset multiple times using `vegdist` while randomly subsampling the dataset each time. All of the subsampled iterations are then averaged (mean) to provide a distance matrix that represents the average of multiple subsampling iterations.

```
## Run 0 stress 0.2531533
## Run 1 stress 0.263634
## Run 2 stress 0.2721521
## Run 3 stress 0.2531394
## ... New best solution
## ... Procrustes: rmse 0.03032977  max resid 0.2325071
## Run 4 stress 0.2554134
## Run 5 stress 0.2510859
## ... New best solution
## ... Procrustes: rmse 0.01692354  max resid 0.155948
## Run 6 stress 0.2592575
## Run 7 stress 0.2585391
## Run 8 stress 0.2635212
## Run 9 stress 0.2531122
## Run 10 stress 0.2564929
## Run 11 stress 0.2683577
## Run 12 stress 0.2538617
## Run 13 stress 0.2584344
## Run 14 stress 0.2524851
## Run 15 stress 0.2538294
## Run 16 stress 0.267128
## Run 17 stress 0.2559587
## Run 18 stress 0.2615322
## Run 19 stress 0.2580279
## Run 20 stress 0.2527381
## *** Best solution was not repeated -- monoMDS stopping criteria:
##     16: no. of iterations >= maxit
##      4: stress ratio > sratmax
```

Non-metric Multi-dimensional Scaling (NMDS) is a way to condense information from multidimensional data (multiple variables/species/OTUs), into a 2D representation or ordination. The closer the points/samples are together in the ordination space, the more similar their microbial communities.

- NMDS plots are non-metric, meaning that among other things, they use data that is not required to fit a normal distribution (there are only a few abundant species, and many, many species with low abundance).
- What makes an NMDS plot non-metric is that it is rank-based. This means that instead of using the actual values to calculate distances, it uses ranks.

```
## Too few points to calculate an ellipse
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (‘geom_path()‘).
## Too few points to calculate an ellipse

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_path()`).

## Too few points to calculate an ellipse

## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_path()`).
```

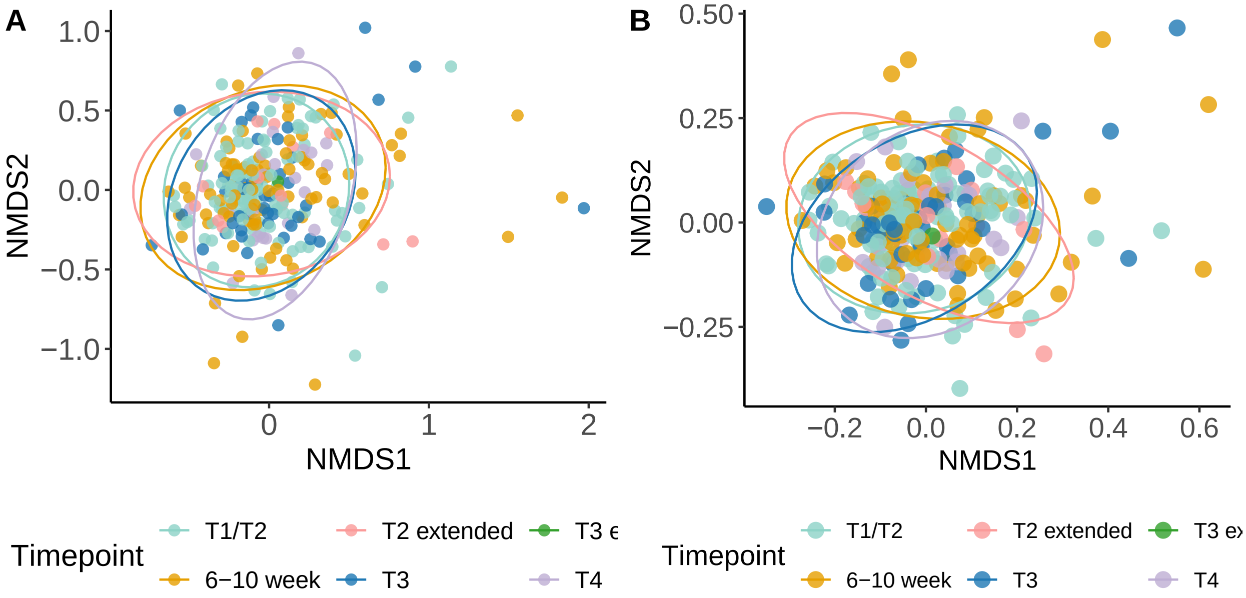


Figure 4. Comparison of beta diversity measurements between non-rarefied and rarefied reads.

Comparison of Bray-Curtis measurements by using non-rarefied and rarefied reads showed no significant difference between the two sets of measurements. This suggests that the rarefication process did not have a significant impact on the beta diversity calculated using the Bray-Curtis metric.

Biological biases in taxonomic composition

Are there systematic biases between detected gram-positive/ gram-negative bacterial ratio?

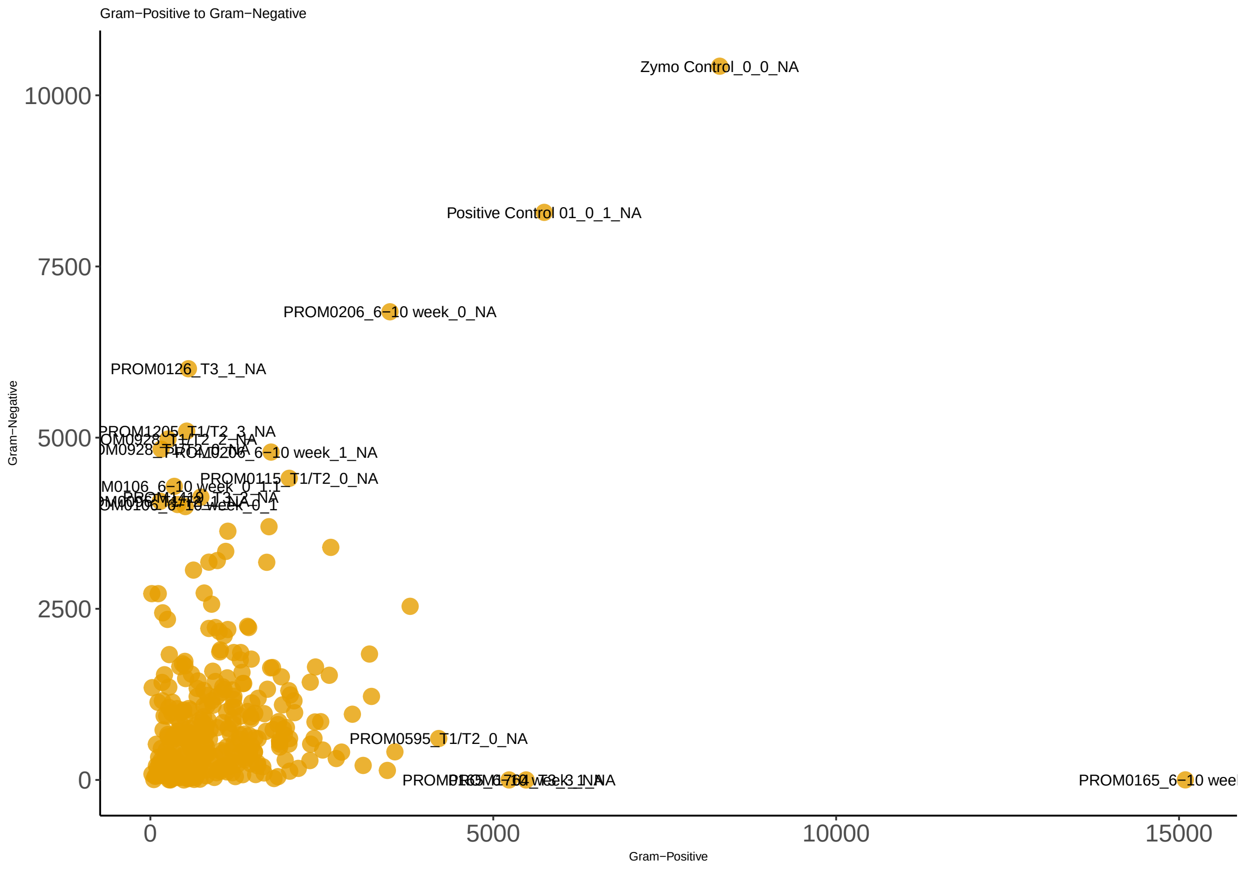


Figure 5. Comparison of most abundant gram positive and gram negative phylum

Majority of the samples in the pilot show a skew towards gram-negative bacteria. It is important to carefully examine any outliers, such as the absence of gram-negative bacteria in sample 165. This could be due to a variety of factors, including errors in the sampling or analysis process, or it could indicate the presence of unique characteristics or conditions in that sample that are not present in the others. We should carefully check the detailed metadata to understand the underlying reasons.

```
## R version 4.3.1 (2023-06-16)
## Platform: aarch64-apple-darwin20 (64-bit)
## Running under: macOS Sonoma 14.0
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.3-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## time zone: Europe/Berlin
## tzcode source: internal
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
```

```
##  [1] vegan_2.6-4     lattice_0.22-5  permute_0.9-7   usedist_0.4.0
##  [5] car_3.1-2       carData_3.0-5   ggpubr_0.6.0    knitr_1.45
##  [9] readxl_1.4.3    lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
## [13] dplyr_1.1.4     purrr_1.0.2     tidyr_1.3.0     tibble_3.2.1
## [17] ggplot2_3.5.0   tidyverse_2.0.0 readr_2.1.4     gtools_3.9.5
##
## loaded via a namespace (and not attached):
##  [1] gtable_0.3.4      xfun_0.41         rstatix_0.7.2     tzdb_0.4.0
##  [5] vctrs_0.6.4       tools_4.3.1       generics_0.1.3    parallel_4.3.1
##  [9] fansi_1.0.5       highr_0.10        cluster_2.1.4     pkgconfig_2.0.3
## [13] Matrix_1.6-3      lifecycle_1.0.4   farver_2.1.1      compiler_4.3.1
## [17] textshaping_0.3.7 munsell_0.5.0     htmltools_0.5.7   yaml_2.3.7
## [21] pillar_1.9.0      MASS_7.3-60       abind_1.4-5       nlme_3.1-163
## [25] tidyselect_1.2.0  digest_0.6.33     stringi_1.8.1     labeling_0.4.3
## [29] splines_4.3.1     cowplot_1.1.1     fastmap_1.1.1     grid_4.3.1
## [33] colorspace_2.1-0  cli_3.6.1         magrittr_2.0.3    utf8_1.2.4
## [37] broom_1.0.5       withr_2.5.2       scales_1.3.0      backports_1.4.1
## [41] timechange_0.2.0  rmarkdown_2.25    ggsignif_0.6.4    cellranger_1.1.0
## [45] ragg_1.2.6        hms_1.1.3         evaluate_0.23     mgcv_1.9-0
## [49] rlang_1.1.3       glue_1.6.2        rstudioapi_0.15.0 R6_2.5.1
## [53] systemfonts_1.0.5
```