Differential abundance analysis of Promise samples

Ece Kartal

01.07.2023

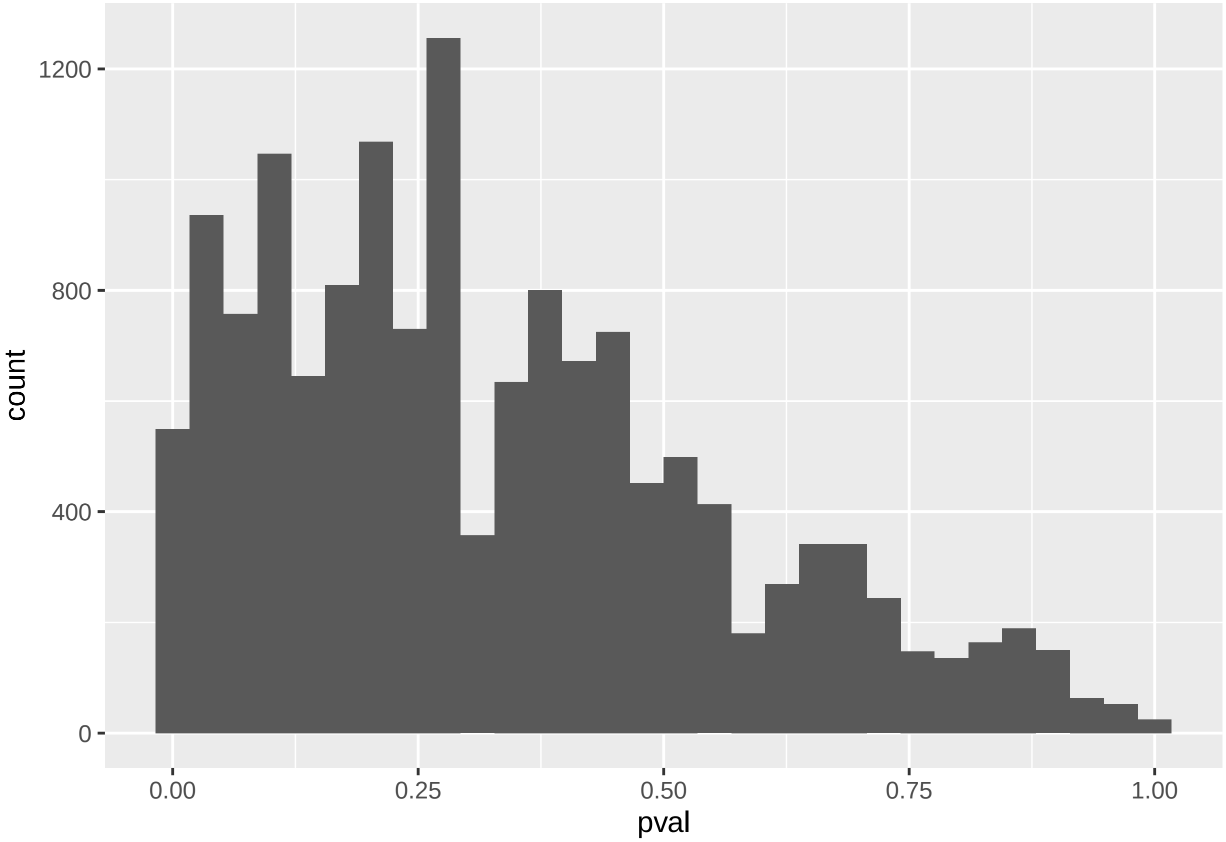
Differential Abundance Analysis (DAA) with Wilcoxin Test

Next, we can analyze the differences between the microbiome profiles of both groups. To do so, we use `wilcoxin test` which is a statistical test where all species in the count matrix are compared between the sample groups of interest. `wilcoxin test` is not the only option to perform Differential Abundance Analysis. Common alternatives include `edgeR` and `DESeq2`.

A **Wilcoxon test** estimates the difference in an outcome between two groups. It is a non-parametric alternative to a t-test, which means that the Wilcoxon test does not make any assumptions about the data.

Here I only perform the DAA between timepoints without eliminating the replicates since I dont have the full metadata. Once the metadata is available, the same analysis can be done for any combination by changing directly **timepoints** variable or adding a new one on top of that.

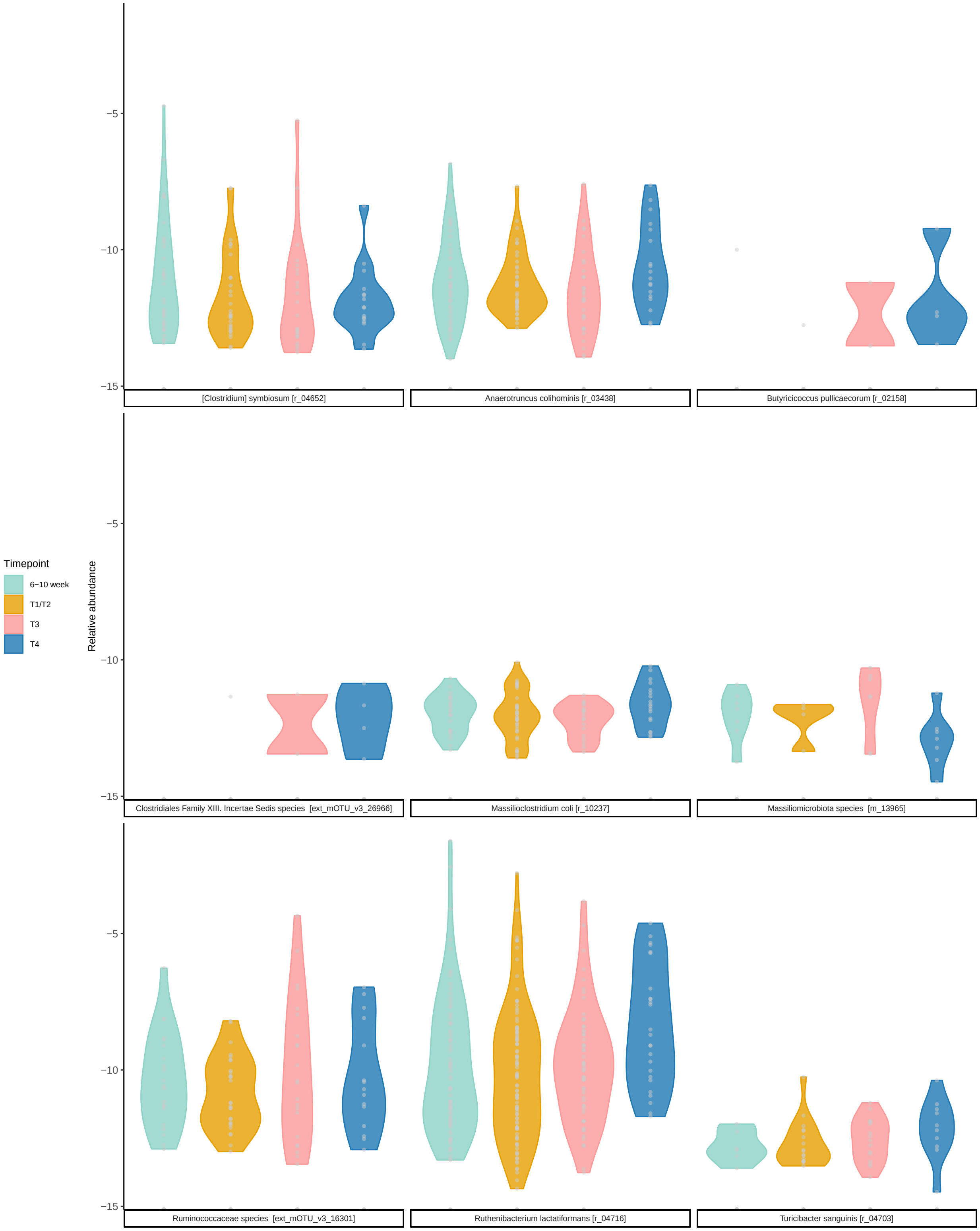
And look at the distribution of the adjusted P values:



We can visualize the abundance changes using one of the most common plots to explore Differential Abundance Analysis results, the **volcano plot**:

```
# Define the list of comparisons
comparisons <- specific_comparisons <- c("T1/T2 / 6-10 week", "T1/T2 / T3", "T1/T2 / T4",
                     "6-10 week / T3", "6-10 week / T4", "T3 / T4")


# Set up a list to store the plots
plot_list <- list()

# Loop over each comparison
for (comp in comparisons) {
  # Filter data for the current comparison
  filtered_data <- df.cal %>%
    filter(comparison == comp)

  # Create the volcano plot for this comparison
  plot <- ggplot(filtered_data, aes(x = fc, y = log10.p, color = sig)) +
    geom_point(alpha = 0.7) +
    scale_color_manual(values = c("black", "red")) +
    geom_hline(yintercept = -log10(p_cutoff)) +
    ggtitle(paste("Differentially abundant species for", comp)) +
    xlab("Generalized fold change") +
    ylab("Log10 adjusted p-value") +
    theme_classic() +
    theme(legend.position = "left",
          text = element_text(size = 10),
          axis.text.x = element_text(size = 15),
          axis.text.y = element_text(size = 15))

  # Add the plot to the plot list
 plot_list[[comp]] <- plot

  # Generate the filename based on the specific comparison
  comparison_name <- gsub("/", "_", comp)  # Replace "/" with underscores
  filename <- paste0(PARAM$folder.output, Sys.Date(), ".", comparison_name, ".wilcox.volcano.plot.pdf")

  # Save the plot as a PDF file
  ggsave(plot, filename = filename, height = 10, width = 13, unit = "cm")

  # Print the plot
  print(plot)
}
```

Now, we can focus on the species with differential abundace across timepoints (not FDR due to low number of samples, will be doen for final cohort)

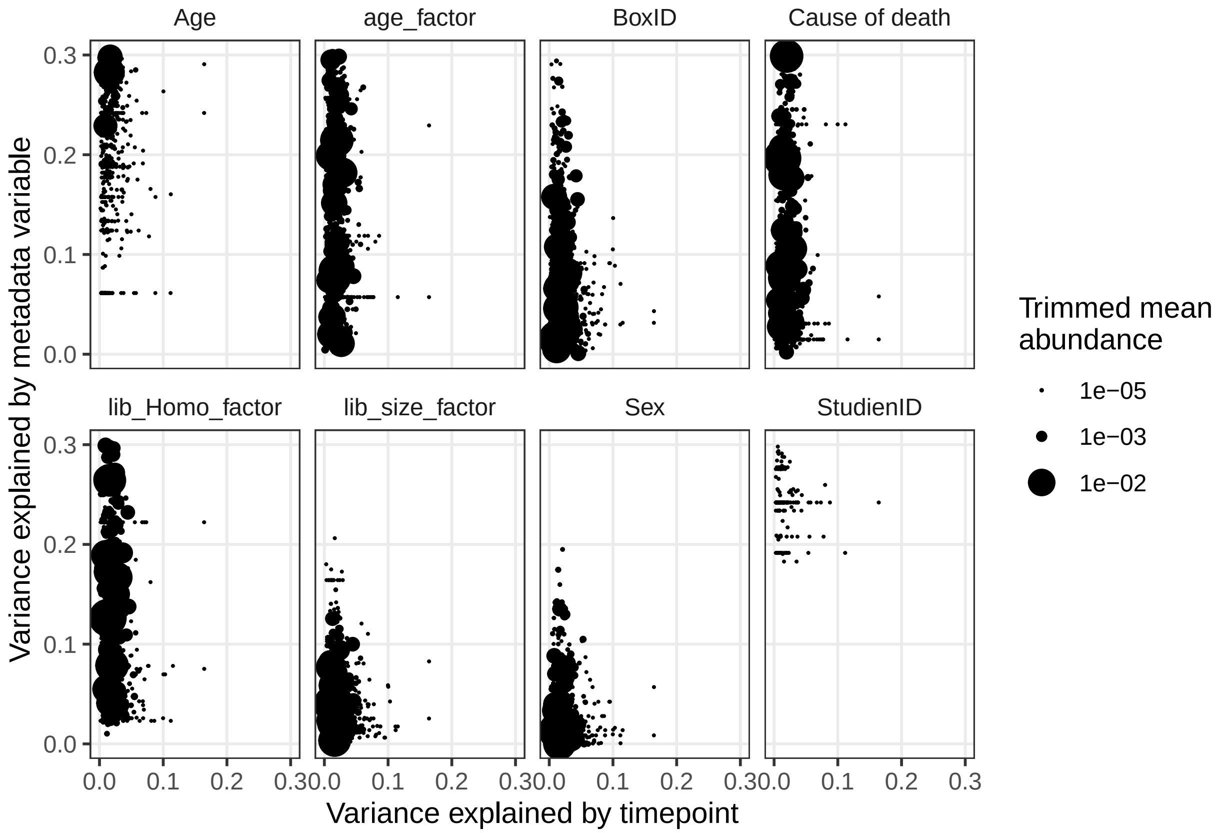


I couldnt focus here too much since the current metadata is a bit confusing, I assinged response per patient per timepoint but each patient have different response that is not consistent in the same timepoint (e.g. same patient has both progressive disease and stable disease response for T3). Once the metadata is fixed, this result will make more sense.

However there are soem trends like some bacteria shows only abundance in progressive diseases patients. But we have very limited patient number for stable and positive response group.

Confounder analysis

```
## ###############################
##  BoxID
## After filtering, the distribution of variables is:
##
##               1 2 3 4
##   T1/T2       9 9 8 8
##   6-10 week   7 7 7 7
##   T2 extended 1 0 3 2
##   T3          4 5 4 3
##   T3 extended 0 0 0 0
##   T4          3 2 3 3
##   Control     0 0 0 0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ###############################
##  lib_size_factor
## After filtering, the distribution of variables is:
##
##                1  2  3  4
##   T1/T2        9 11 11  3
##   6-10 week    4 11 10  3
##   T2 extended  2  2  2  0
##   T3           2  5  5  4
##   T3 extended  0  0  0  0
##   T4           3  7  0  1
##   Control      0  0  0  0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ###############################
##  StudienID
## After filtering, the distribution of variables is:
##
##               1-001 1-002 1-003 1-004 1-005 1-006 1-008 1-009 1-010 1-011 1-013
##   T1/T2           1     1     1     1     1     1     1     1     1     1     1
##   6-10 week       1     1     1     1     0     1     1     1     1     0     0
##   T2 extended     1     0     0     0     0     0     0     0     0     0     0
##   T3              1     0     0     1     0     1     0     1     1     0     1
##   T3 extended     0     0     0     0     0     0     0     0     0     0     0
##   T4              1     0     0     1     0     1     0     0     1     0     0
##   Control         0     0     0     0     0     0     0     0     0     0     0
##
##               1-014 1-015 1-016 1-017 1-018 1-019 1-020 1-021 1-022 1-024 1-025
##   T1/T2           1     1     1     1     1     1     1     1     1     1     1
##   6-10 week       1     1     1     1     0     1     1     0     1     1     1
##   T2 extended     0     0     0     0     0     0     0     0     0     0     0
##   T3              1     1     0     0     0     1     0     1     1     0     1
##   T3 extended     0     0     0     0     0     0     0     0     0     0     0
##   T4              1     0     0     0     0     0     0     1     0     0     1
##   Control         0     0     0     0     0     0     0     0     0     0     0
##
##               1-026 1-027 1-028 1-029 1-030 1-031 1-032 1-033 1-035 1-036 1-037
##   T1/T2           0     1     1     1     1     1     1     1     1     1     1
##   6-10 week       1     1     1     1     1     1     0     1     1     0     1
##   T2 extended     0     1     1     0     0     1     1     1     0     0     0
##   T3              0     1     0     0     1     0     0     0     1     0     1
##   T3 extended     0     0     0     0     0     0     0     0     0     0     0
##   T4              0     1     0     0     1     0     0     0     1     0     1
##   Control         0     0     0     0     0     0     0     0     0     0     0
##
##               1-038 1-083 1-084
##   T1/T2           1     0     1
##   6-10 week       1     1     0
##   T2 extended     0     0     0
##   T3              0     0     0
##   T3 extended     0     0     0
##   T4              0     0     0
##   Control         0     0     0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ################################
##  Age
## After filtering, the distribution of variables is:
##
##               46 48 50 51 52 54 55 57 59 61 62 63 64 65 67 69 70 71 72 75 76 77
##   T1/T2        1  1  1  1  3  1  2  1  1  1  1  1  2  2  2  4  1  1  2  1  1  1
##   6-10 week    1  1  1  1  1  1  2  0  1  1  1  1  1  1  2  3  1  2  1  1  1  1
##   T2 extended  1  0  0  0  0  0  2  0  0  0  0  0  1  0  0  1  0  0  0  1  0  0
##   T3           1  1  0  0  2  1  0  0  0  1  0  1  0  0  1  3  1  1  1  0  1  0
##   T3 extended  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##   T4           1  1  0  0  1  0  0  0  0  1  0  0  0  0  1  3  1  1  0  0  1  0
##   Control      0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0  0
##
##               79
##   T1/T2        1
##   6-10 week    1
##   T2 extended  0
##   T3           1
##   T3 extended  0
##   T4           0
##   Control      0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ################################
##  Sex
## After filtering, the distribution of variables is:
##
##               Female Male
##   T1/T2           16   17
##   6-10 week       10   17
##   T2 extended      4    2
##   T3               7    9
##   T3 extended      0    0
##   T4               4    7
##   Control          0    0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ###############################
##  Cause of death
## After filtering, the distribution of variables is:
##
##               Other cause (suicide, accident, etc.) Tumor conditional Unknown
##   T1/T2                                           1                 8       2
##   6-10 week                                       0                 5       2
##   T2 extended                                     0                 1       0
##   T3                                              0                 4       2
##   T3 extended                                     0                 0       0
##   T4                                              0                 2       1
##   Control                                         0                 0       0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ###############################
##  age_factor
## After filtering, the distribution of variables is:
##
##               1 2 3 4
##   T1/T2       4 5 1 1
##   6-10 week   3 3 1 0
##   T2 extended 0 1 0 0
##   T3          3 1 1 1
##   T3 extended 0 0 0 0
##   T4          1 1 1 0
##   Control     0 0 0 0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
##
## ###############################
##  lib_Homo_factor
## After filtering, the distribution of variables is:
##
##               1 2 3 4
##   T1/T2       6 4 1 0
##   6-10 week   2 0 3 2
##   T2 extended 1 0 0 0
##   T3          1 5 0 0
##   T3 extended 0 0 0 0
##   T4          3 0 0 0
##   Control     0 0 0 0
## Calculating variance explained by meta-variable...
## Calculating association with the meta-variable...
```



I had to use timepoint because we dont have response data organised yet. This plot should be updated to **response** vs all meta variables once the metadat is completed to have a proper view if there is any confounders in the data

Session info

```
sessionInfo()
```

```
## 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 v
##
## 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] pROC_1.18.5     ggrepel_0.9.4   reshape2_1.4.4  vegan_2.6-4
##  [5] lattice_0.22-5  permute_0.9-7   car_3.1-2       carData_3.0-5
##  [9] ggpubr_0.6.0    knitr_1.45      readxl_1.4.3    lubridate_1.9.3
## [13] forcats_1.0.0   stringr_1.5.1   dplyr_1.1.4     purrr_1.0.2
## [17] tidyr_1.3.0     tibble_3.2.1    ggplot2_3.5.0   tidyverse_2.0.0
## [21] 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     fastmap_1.1.1     grid_4.3.1        colorspace_2.1-0
## [33] cli_3.6.1         magrittr_2.0.3    utf8_1.2.4        broom_1.0.5
## [37] withr_2.5.2       scales_1.3.0      backports_1.4.1   timechange_0.2.0
## [41] rmarkdown_2.25    ggsignif_0.6.4    cellranger_1.1.0  ragg_1.2.6
## [45] hms_1.1.3         evaluate_0.23     mgcv_1.9-0        rlang_1.1.3
## [49] Rcpp_1.0.11       glue_1.6.2        rstudioapi_0.15.0 plyr_1.8.9
## [53] R6_2.5.1          systemfonts_1.0.5
```