Ravel_2013_BV

```
library(SummarizedExperiment)
library(dplyr)
library(tidyr)
library(tibble)
library(purrr)
library(ggplot2)
```

The data comes from the article of Ravel et. al. 2013.

Reference:

Ravel, J., Brotman, R.M., Gajer, P. et al. Daily temporal dynamics of vaginal microbiota before, during and after episodes of bacterial vaginosis. Microbiome 1, 29 (2013). https://doi.org/10.1186/2049-2618-1-29

```
se <- readRDS("Ravel_2013_BV_se.rds")
se
```

```
## class: SummarizedExperiment
## dim: 151 1657
## metadata(0):
## assays(2): counts relative_abundance
## rownames(151): Lactobacillus iners Gardnerella vaginalis ... Aerococcus
##    viridans Streptococcus mutans
## rowData names(6): superkingdom phylum ... family genus
## colnames(1657): s3.w1d1 s3.w1d2 ... s135.w10d7 s135.w11d1
## colData names(20): subjectID dayInStudy ... MENSTRU3 Total.Reads
```

Exploring sample metadata

```
sample_metadata <- se %>%
    colData() %>%
    as.data.frame() %>%
    rownames_to_column(var = "Sample") %>%
    as_tibble() %>%
    mutate(
        pH = as.double(pH),
        Nugent_Score = as.integer(Nugent_Score),
        amsel_pos = case_when(SBV == "1" | ABV == "1" ~ 1, TRUE ~ 0),
        amsel_res = case_when(
            SBV == "1" ~ "SBV", ABV == "1" ~ "ABV", TRUE ~ "Neg"
        ),
        nugent_res = case_when(
            Nugent_Score <= 3 ~ "Negative",
            Nugent_Score >=4 & Nugent_Score <= 6 ~ "Intermediate",
            Nugent_Score >= 7 ~ "Positive"
        ),
        nugent_res = factor(
```

```
        nugent_res, levels = c("Negative", "Intermediate", "Positive")
      )
    )

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion

## Warning in mask$eval_all_mutate(quo): NAs introduced by coercion
glimpse(sample_metadata)

## Rows: 1,657
## Columns: 24
## $ Sample        <chr> "s3.w1d1", "s3.w1d2", "s3.w1d3", "s3.w1d5", "s3.w1d6", "~
## $ subjectID     <dbl> 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3, 3,~
## $ dayInStudy    <dbl> 1, 2, 3, 5, 6, 8, 9, 10, 11, 12, 14, 15, 16, 17, 18, 19,~
## $ DIA_WEEK      <dbl> 1, 1, 1, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 4,~
## $ DIA_DAY       <dbl> 1, 2, 3, 5, 6, 1, 2, 3, 4, 5, 7, 1, 2, 3, 4, 5, 6, 7, 1,~
## $ AGE           <dbl> 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, ~
## $ ETHa          <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ SBV           <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ ABV           <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ BV_medication <dbl> 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0,~
## $ pH            <dbl> NA, 4.0, 4.0, 4.0, 5.3, 4.7, 4.0, 4.0, 4.0, 4.0, 4.0, 4.~
## $ Nugent_Score  <int> 1, 4, 1, 6, 6, 7, 7, 7, 4, 4, 1, 3, 3, 3, 4, 3, 1, 1, 1,~
## $ VAG_ODOR      <chr> "NA", "0", "0", "0", "0", "1", "1", "0", "0", "0", "0", ~
## $ VAG_IRR       <chr> "NA", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", ~
## $ VAG_ITCH      <chr> "NA", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", ~
## $ VAG_BURN      <chr> "NA", "0", "0", "0", "0", "0", "0", "0", "NA", "0", "0",~
## $ VAG_DIS       <chr> "NA", "0", "0", "0", "0", "1", "0", "0", "NA", "1", "1",~
## $ MENSTRU1      <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "N~
## $ MENSTRU2      <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "N~
## $ MENSTRU3      <chr> "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "NA", "N~
## $ Total.Reads   <dbl> 206, 328, 253, 156, 131, 131, 142, 253, 293, 280, 320, 4~
## $ amsel_pos     <dbl> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,~
## $ amsel_res     <chr> "Neg", "Neg", "Neg", "Neg", "Neg", "SBV", "Neg", "Neg", ~
## $ nugent_res    <fct> Negative, Intermediate, Negative, Intermediate, Intermed~
```
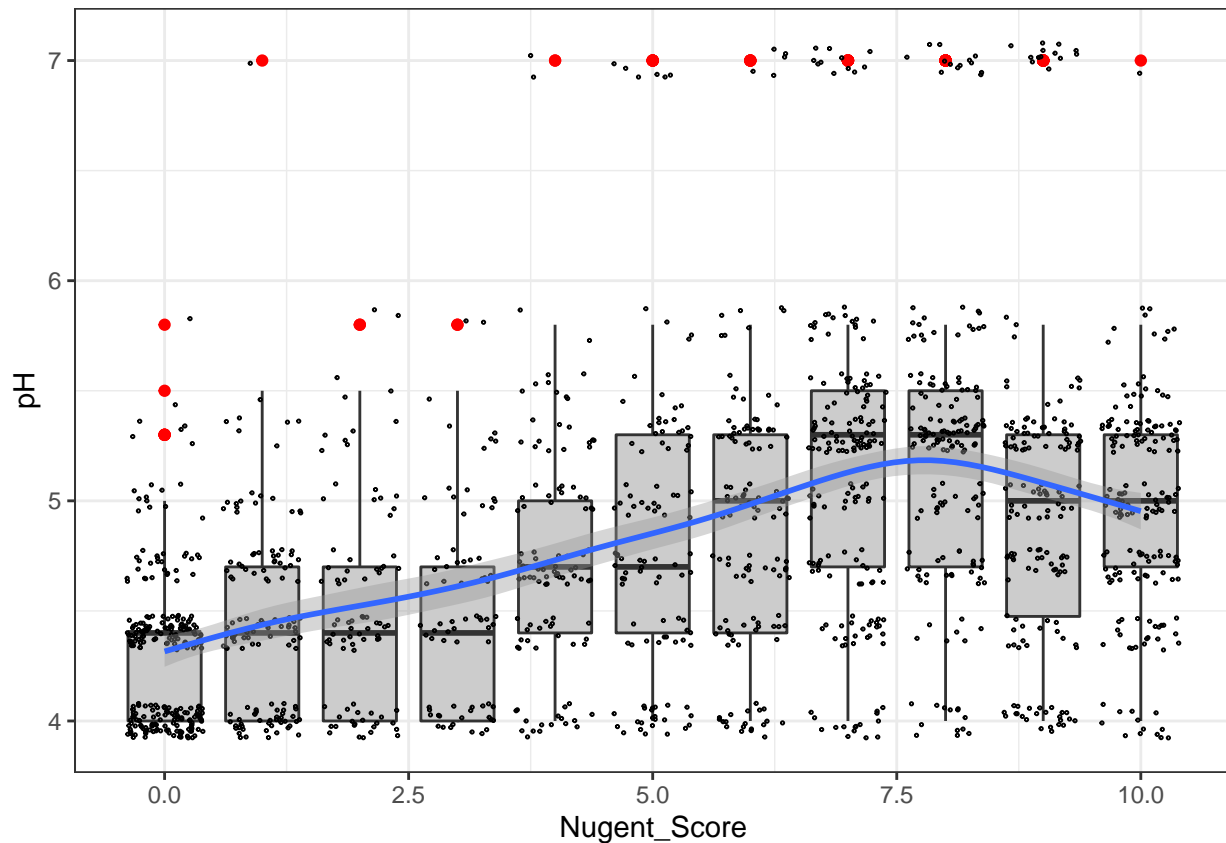
Exploring the relationship between Nugent Scores and pH

```
sample_metadata %>%
    mutate(pH = as.double(pH), Nugent_Score = as.integer(Nugent_Score)) %>%
    suppressWarnings() %>%
    filter(!is.na(pH), !is.na(Nugent_Score)) %>%
    ggplot(aes(Nugent_Score, pH )) +
    geom_boxplot(
        aes(group = Nugent_Score), fill = "gray80",
        outlier.color = "red"
        ) +
    geom_point(size = .25, position = "jitter", shape = 1) +
    geom_smooth() +
    theme_bw()
```

Select subjects

- Keep only subjects from the same ethnical group (white):

```
subjects_data <- sample_metadata %>%
    select(subjectID, AGE, amsel_pos, ETHa) %>%
    group_by(subjectID) %>%
    slice_max(order_by = amsel_pos, with_ties = FALSE) %>%
    ungroup() %>%
    filter(ETHa == 0)

subjects <- unique(subjects_data$subjectID)
```

```
subjects_data %>%
    modify_at(.at = c("subjectID", "amsel_pos", "ETHa"), .f = as.factor) %>%
    summary()
```

```
##     subjectID        AGE         amsel_pos ETHa
## 3        : 1   Min.   :19.0   0: 5      0:20
## 5        : 1   1st Qu.:21.0   1:15
## 7        : 1   Median :26.0
## 15       : 1   Mean   :26.9
## 23       : 1   3rd Qu.:29.0
## 27       : 1   Max.   :45.0
## (Other):14
```

According to the article, only four subjects never experienced some form of BV. However, in the summary

above there are five subjects.

```
negative_subjects <- subjects_data %>%
    filter(amsel_pos == 0) %>%
    pull(subjectID)
negative_subjects
```

```
## [1]  5  7 40 52 53
```

```
sample_metadata %>%
    filter(subjectID %in% negative_subjects) %>%
    select(subjectID, SBV, ABV) %>%
    distinct()
```

```
## # A tibble: 5 x 3
##    subjectID   SBV   ABV
##        <dbl> <dbl> <dbl>
## 1          5     0     0
## 2          7     0     0
## 3         40     0     0
## 4         52     0     0
## 5         53     0     0
```

Assays

```
count_matrix <- assays(se)$counts
count_matrix[1:5, 1:5]
```

```
##                        s3.w1d1 s3.w1d2 s3.w1d3 s3.w1d5 s3.w1d6
## Lactobacillus iners        140     222     173      79      31
## Gardnerella vaginalis       11       4       5      36      75
## Lactobacillus crispatus     10      26      33       1       0
## Atopobium vaginae            0       0       0       0       2
## Lactobacillus jensenii      45      73      41       2       2
```

```
relative_abundance_matrix <- assays(se)$relative_abundance
relative_abundance_matrix[1:5, 1:5]
```

```
##                           s3.w1d1   s3.w1d2   s3.w1d3    s3.w1d5    s3.w1d6
## Lactobacillus iners     67.961165 67.682927 68.379447 50.6410256 23.664122
## Gardnerella vaginalis    5.339806  1.219512  1.976285 23.0769231 57.251908
## Lactobacillus crispatus  4.854369  7.926829 13.043478  0.6410256  0.000000
## Atopobium vaginae        0.000000  0.000000  0.000000  0.0000000  1.526718
## Lactobacillus jensenii  21.844660 22.256098 16.205534  1.2820513  1.526718
```

```
max(colSums(relative_abundance_matrix))
```

```
## [1] 100
```

Summarize by genus

```
se_genus <- mia::agglomerateByRank(se, rank = "genus")
se_genus
```

```
## class: SummarizedExperiment
```

4

```
## dim: 63 1657
## metadata(0):
## assays(2): counts relative_abundance
## rownames(63):
##   Firmicutes_Bacilli_Lactobacillales_Lactobacillaceae_Lactobacillus
##   Actinobacteria_Actinomycetia_Bifidobacteriales_Bifidobacteriaceae_Gardnerella
##   ...
##   Proteobacteria_Gammaproteobacteria_Moraxellales_Moraxellaceae_Acinetobacter
##   Proteobacteria_Gammaproteobacteria_Pseudomonadales_NA_NA
## rowData names(6): superkingdom phylum ... family genus
## colnames(1657): s3.w1d1 s3.w1d2 ... s135.w10d7 s135.w11d1
## colData names(20): subjectID dayInStudy ... MENSTRU3 Total.Reads
```

Get taxonomy table:

```r
taxa_genus <- se_genus %>%
    rowData() %>%
    as.data.frame %>%
    rownames_to_column(var = "tax_name") %>%
    as_tibble()
```

Combine sample metadata, relative abundance, and taxonomy into a single tibble:

```r
tidy_table_genus <- assays(se_genus)$relative_abundance %>%
    as.data.frame() %>%
    rownames_to_column(var = "Taxa") %>%
    as_tibble() %>%
    pivot_longer(names_to = "Sample", values_to = "Abundance", 2:last_col()) %>%
    left_join(sample_metadata, by = "Sample") %>%
    right_join(taxa_genus, by = c("Taxa" = "tax_name")) %>%
    relocate(Sample, Taxa, Abundance) %>%
    filter(subjectID %in% subjects)
```

Nugent score vs relative abundance of Lactubacillus

```r
taxa_genera <- c(
    "Lactobacillus$",
    "Gardnerella$"
)
taxa_regex <- paste0("(", paste0(taxa_genera, collapse = "|"), ")")

tidy_table_genus %>%
    filter(grepl(taxa_regex, Taxa)) %>%
    select(Sample, Taxa, Abundance, pH, Nugent_Score, nugent_res) %>%
    mutate(
        Taxa = sub("^.+_", "", Taxa),
        Taxa = factor(Taxa, levels = sub("\\$$", "", taxa_genera)),
    ) %>%
    drop_na() %>%
    suppressWarnings() %>%
    ggplot(aes(Nugent_Score, Abundance)) +
    geom_boxplot(
        aes(group = Nugent_Score, color = nugent_res), outlier.shape = 4
    ) +
    # geom_point(size = .25, position = "jitter") +
    geom_smooth(color = "red") +
```

```
facet_wrap(~Taxa, ncol = 2) +
scale_x_continuous(breaks = seq(0,10,1)) +
scale_color_discrete(name = "BV diagnostic (Nugent score)") +
labs(
    y = "Relative abundance", x = "Nugent score",
    title = "Full dataset (Ravel et. al., 2013)"
) +
theme_bw() +
theme(
    legend.position = "bottom"
)
```


Full dataset (Ravel et. al., 2013)