Sys BioMed HW5 - Microbiome

Group 2

What is the difference between a microbial community and a monoculture?

- monoculture: cells of the same microbial species
- microbial community: cells of at least two microbial species (could be a lot more)

Why is the 16S rRNA gene ideal for profiling microbial communities with amplicon sequencing?

It has highly conserved parts, which are good for amplification, but also highly varied parts, which can be used to differentiate between species. It is also present in all bacteria.

Name five factors which impact our microbiome.

1. Birth mode
2. Diet
3. Exercise
4. Age
5. Disease

Name five omics datatypes.

1. Proteomics
2. Metabolomics
3. Metatranscriptomics
4. Culturomics
5. Metagenomics

What is an important advantage of assembly-based methods?

They can identify novel genes and genomes, since they do not depend on reference genomes etc.

Why is sequencing data compositional?

The number of reads is several orders of magnitude smaller than the number of cells. With random sampling we get accurate proportions, but not exact counts. Therefore we only know the relative abundance of reads.

What characterizes microbial dysbiosis?

Microbial dysbiosis is a disruption of the microbiome, linked to many human diseases. Pathobiont expansion, reduced microbe diversity and loss of beneficial microbes are all forms of dysbiosis.

The table “hmp2_IBD_abd.csv” provides relative abundances of a metagenomic dataset comparing the gut microbiome of IBD patients to healthy individuals and the “hmp2_IBD_metadata.csv” file provides some metadata information for this cohort.

```
# libraries & data load

library(data.table)
library(fossil)
library(vegan)
library(dplyr)
library(ggplot2)

gen_data <- fread("microbiome_ex/hmp2_IBD_abd.csv", sep=";")
metadata <- fread("microbiome_ex/hmp2_IBD_metadata.csv", sep=";")
```

a) Write a script (e.g. R, python) that calculates the median alpha diversity based on the Shannon and chao1 index for the IBD and healthy group, respectively.

```
long_gen_data <- melt(gen_data, id.vars = "V1", variable.name = "sample_id", value.name = "abundance")
setnames(long_gen_data, old="V1", new="species")

samples <- metadata$sample_id
shannon_div <- c()
chao1_div <- c()
for (sample in samples){
  shannon_div <- append(shannon_div, diversity(long_gen_data[sample_id==sample]$abundance))
  chao1_div <- append(chao1_div, chao1(long_gen_data[sample_id==sample]$abundance))
}
metadata$shannon_div <- shannon_div
metadata$chao1_div <- chao1_div

ibd_shan <- median(metadata[disease=="IBD"]$shannon_div)
healthy_shan <- median(metadata[disease=="healthy"]$shannon_div)
ibd_chao <- median(metadata[disease=="IBD"]$chao1_div)
healthy_chao <- median(metadata[disease=="healthy"]$chao1_div)

print(paste("Median shannon index for IBD:", ibd_shan))
```

```
## [1] "Median shannon index for IBD: 2.23130704744467"
```

```
print(paste("Median shannon index for healthy", healthy_shan))
```

```
## [1] "Median shannon index for healthy 2.42088127579023"
```

```
print(paste("Median chao1 index for IBD:", ibd_chao))
```

```
## [1] "Median chao1 index for IBD: 48"
```

```
print(paste("Median chao1 index for healthy", healthy_chao))
```

```
## [1] "Median chao1 index for healthy 58"
```

b) Write another script and calculate the beta diversity using the Bray Curtis dissimilarity between patients from the IBD and healthy group, respectively. Show your results in a boxplot and describe them.

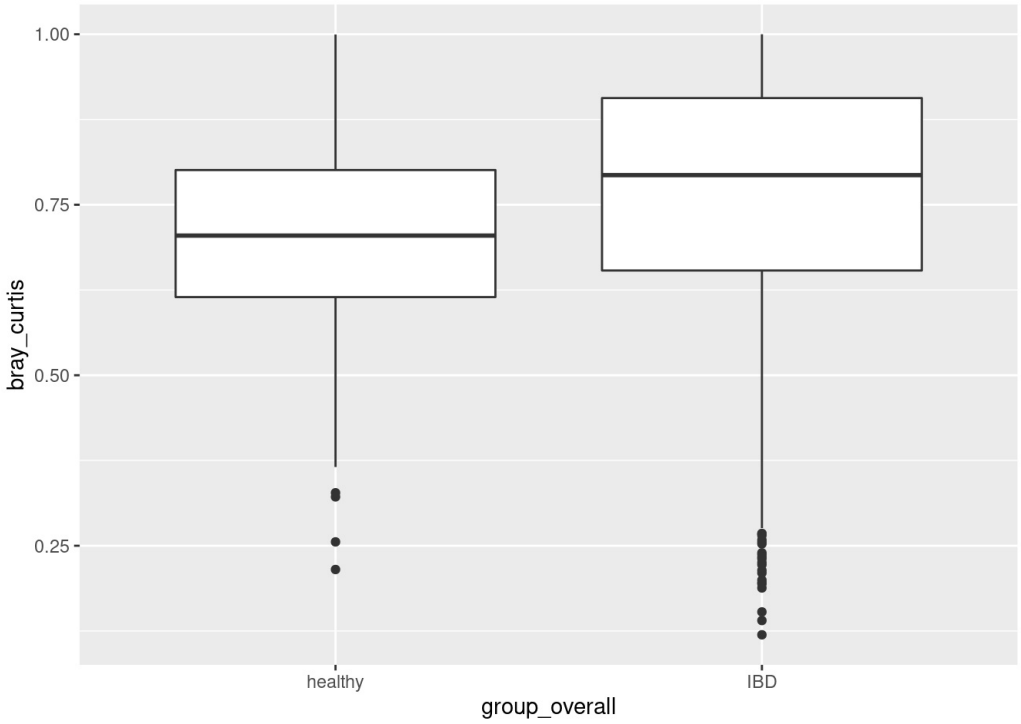
```
bray <- vegdist(t(gen_data[, -"V1"])) %>% as.matrix()
bray[lower.tri(bray)] <- NA
bray <- as.data.table(bray, keep.rownames=T)

long_bray_values <- melt(bray, id.vars = "rn", variable.name = "sample_2", value.name = "bray_curtis")
setnames(long_bray_values, old = "rn", new="sample_1")
long_bray_values <- na.omit(long_bray_values)
long_bray_values <- long_bray_values[sample_1!=sample_2]

long_bray_group_1 <- c()
long_bray_group_2 <- c()
long_bray_group_overall <- c()
for (i in 1:nrow(long_bray_values)){
  g1 <- metadata[sample_id==long_bray_values[i,]$sample_1]$disease
  long_bray_group_1 <- append(long_bray_group_1, g1)
  g2 <- metadata[sample_id==long_bray_values[i,]$sample_2]$disease
  long_bray_group_2 <- append(long_bray_group_2, g2)
  if (g1 == g2){long_bray_group_overall <- append(long_bray_group_overall, g1)}
  else {long_bray_group_overall <- append(long_bray_group_overall, "mixed")}
}
long_bray_values$group_sample_1 <- long_bray_group_1
long_bray_values$group_sample_2 <- long_bray_group_2
long_bray_values$group_overall <- long_bray_group_overall

internal_comp_dt <- long_bray_values[group_overall!="mixed"]

ggplot(internal_comp_dt, aes(group_overall, bray_curtis))+geom_boxplot()
```



The IBD group has a slightly higher median variance for the bray-curtis dissimilarity, hinting at greater diversity of the microbiome between patients with IDB compared to healthy individuals.There are however many outliers, suggesting, that while overall the diversity is greater, there are still many patients within the group with a similar microbiome composition. Additionally, even the healthy group shows a fairly high median dissimilarity, which suggests, that most of the diversity seen in the IBD group is likely not caused by the disease but naturally occuring.

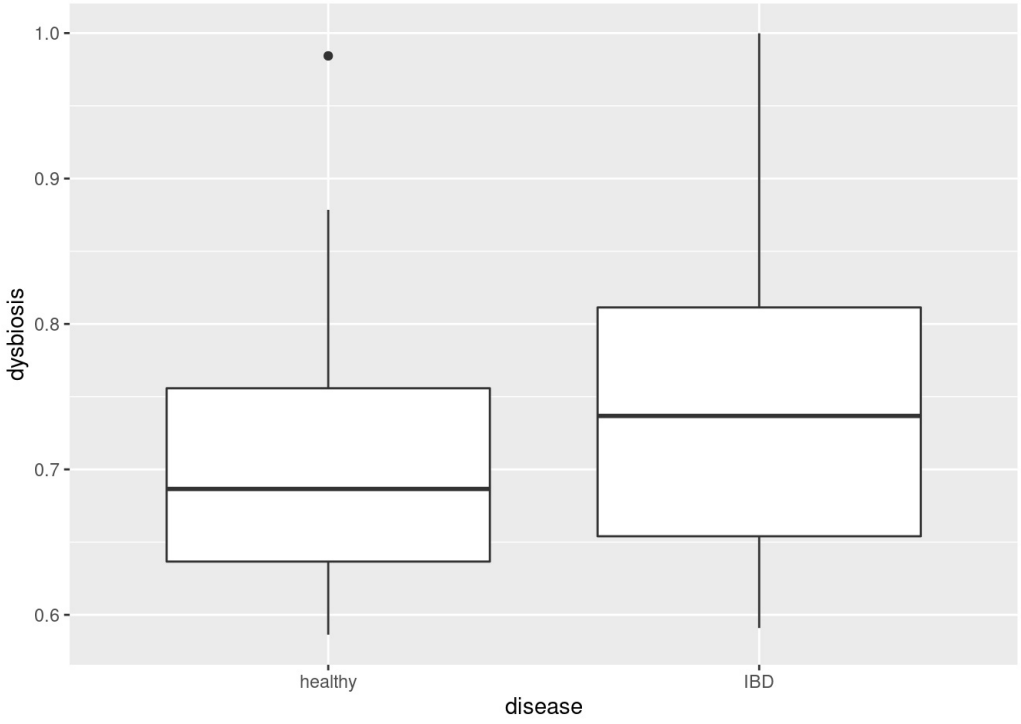
c) Write a script and calculated the dysbiosis score for each patient, draw a boxplot for the healthy group and IBD group and describe the results.

```
# in case this cell is run more than once
# metadata$dysbiosis <- NULL

metadata$dysbiosis <- double()
for (sample in metadata$sample_id){
  values <- long_bray_values[((sample_1==sample)&(group_sample_2=="healthy")) | ((sample_2==sample)&(group_sample
_1=="healthy"))]$bray_curtis

  metadata[sample_id==sample]$dysbiosis <- median(values)
}

ggplot(metadata, aes(disease, dysbiosis))+geom_boxplot()
```



Individuals with IBD show a greater median bray-curtis dissimilarity to healthy individuals, than healthy individuals do to each other. However, the difference is once again fairly small, suggesting most of the microbiome species differences to be natural.