CINEMA TUTORIAL

Microbiome analysis for Clinical Research

Sara Vieira-Silva

2023-07-06

UNIVERSITÄTSmedizin.
MAINZ

Setup workspace

packages

```
#libraries to load
library("corrplot")
library("phyloseq")
library("ggplot2")
library("vegan")
library("Hmisc")
library("tidyverse")

#installing phyloseq (Handling high-throughput microbiome census data)
# if (!require("BiocManager", quietly = TRUE))
#     install.packages("BiocManager")
# BiocManager::install("phyloseq")
```

directories, data, functions

Tutorial outline

Sections

- data-driven approaches
  - dimensionality reduction (ordination)
  - constrained ordination
- hypothesis testing for biomarker discovery
  - model design
  - nested models for confounder analysis
- data stratification
  - enterotyping using Dirichlet multinomial mixtures
  - stratification in clinical associations

Dataset

Reanalysing a part of the dataset of the **BMIS dataset**

Reference

Vieira-Silva et al. Nature. 2020
Statin therapy is associated with lower prevalence of gut microbiota dysbiosis.
doi: 10.1038/s41586-020-2269-x.

Hypothesis

Is gut dysbiosis associated to cardiovascular disease?

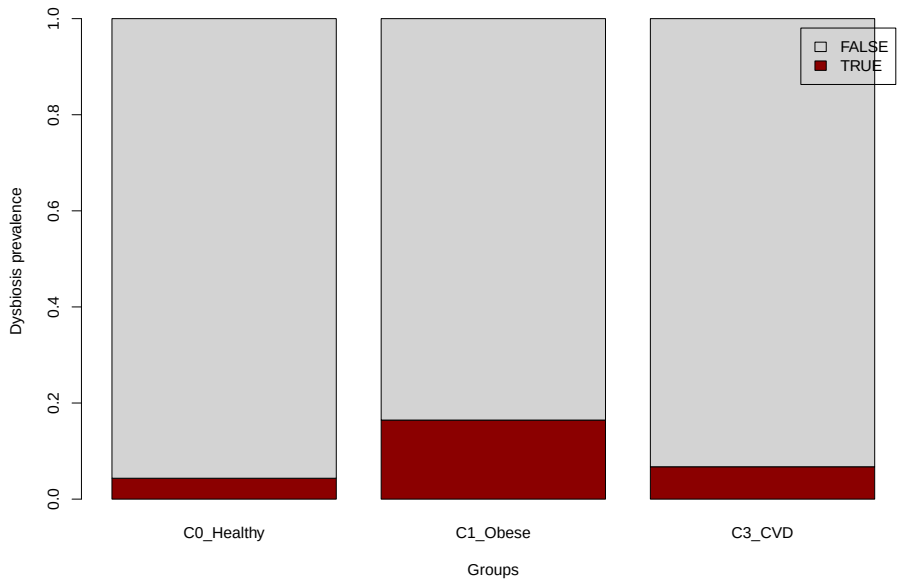
Let's dive right in!

First instinct: test the main hypothesis.

```
mydata=EXTENDED_DATA

# tabulate dysbiosis (defined as Bact2-enterotype) per patient group
contingency_table <- table(mydata$Health_status_refined,mydata$Bact2)
# reorder
contingency_table <- contingency_table[,c(2,1)]
# compute prevalence table
contingency_table_prop <-  prop.table(contingency_table,
                                      margin = 1)
```

Test the main hypothesis: plot and test

```
# plot contingency table
barplot(t(contingency_table_prop), beside = F, legend = T, xlab = "Groups"
        , ylab = "Dysbiosis prevalence", col=c('darkred','lightgrey'))
```




```
# Test differences: pairwise chi-squared test
paired_test <- pairwise.prop.test(contingency_table,
                                  p.adjust.method = "holm")
```

What is your conclusion?

Is gut dysbiosis associated to cardiovascular disease?

```
# Test differences: pairwise chi-squared test
paired_test <- pairwise.prop.test(contingency_table,
                                  p.adjust.method = "holm")
print(paired_test$p.value)
```

```
##             C0_Healthy     C1_Obese
## C1_Obese 3.852117e-08           NA
## C3_CVD   2.272452e-01 0.0003515885
```

Taking a step back

Decomposing your research question into tractable steps

1. Find a quantitative biomarker of the pathomechanism
2. Get to know your data and potential confounders in your design
3. Quantify the contribution of dysbiosis to disease risk
4. Identify modulators of this contribution

Taking a step back

Decomposing your research question into tractable steps

1. Find a quantitative biomarker of the pathomechanism
   - BMI
2. Get to know your data and potential confounders in your design
   - Confounders in clinical panel and medication history
3. Quantify the contribution of dysbiosis to disease risk
4. Identify modulators of this contribution

Get to know your data

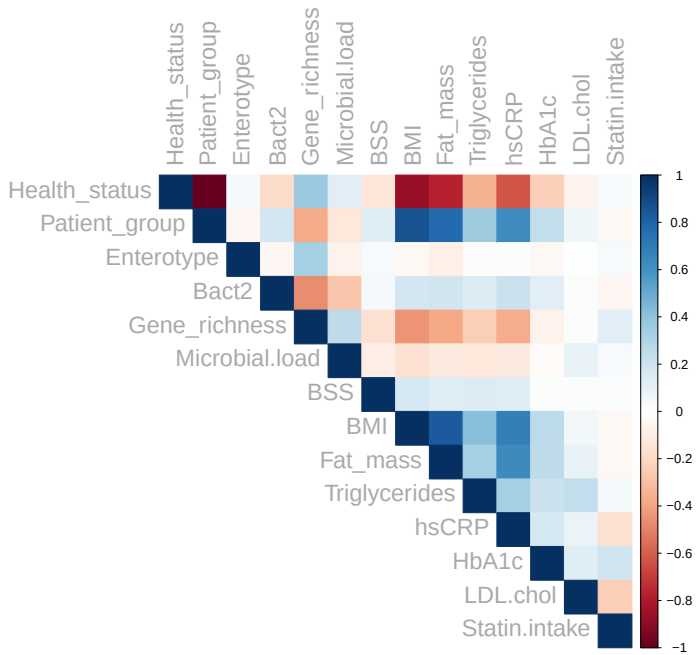
Are there confounding associations between outcome and predictor variables?

```
mydata=METADATA_1
# Use summary() to generate descriptive statistics
mysum <- summary(mydata)

# Convert factor/logical variables to numeric
numeric_data <- mydata
factor_columns <- sapply(numeric_data, is.factor)
logical_columns <- sapply(numeric_data, is.logical)
numeric_data[factor_columns] <- lapply(numeric_data[factor_columns]
                                       , as.numeric)
numeric_data[logical_columns] <- lapply(numeric_data[logical_columns]
                                        , as.numeric)

# Calculate the correlation matrix
correlation_matrix <- rcorr(as.matrix(numeric_data), type = "spearman")
# create human-readable table
correlations <- flattenCorrMatrix(correlation_matrix$r,correlation_matrix$P
```

Get to know your data

Are there confounding associations between predictor variables?

```
# Read and plot the correlation matrix
corrplot(correlation_matrix$r, method = "color", type = "upper",
         tl.col = "darkgrey", tl.cex = 1.5)
```



Data-driven approach: unconstrained ordination

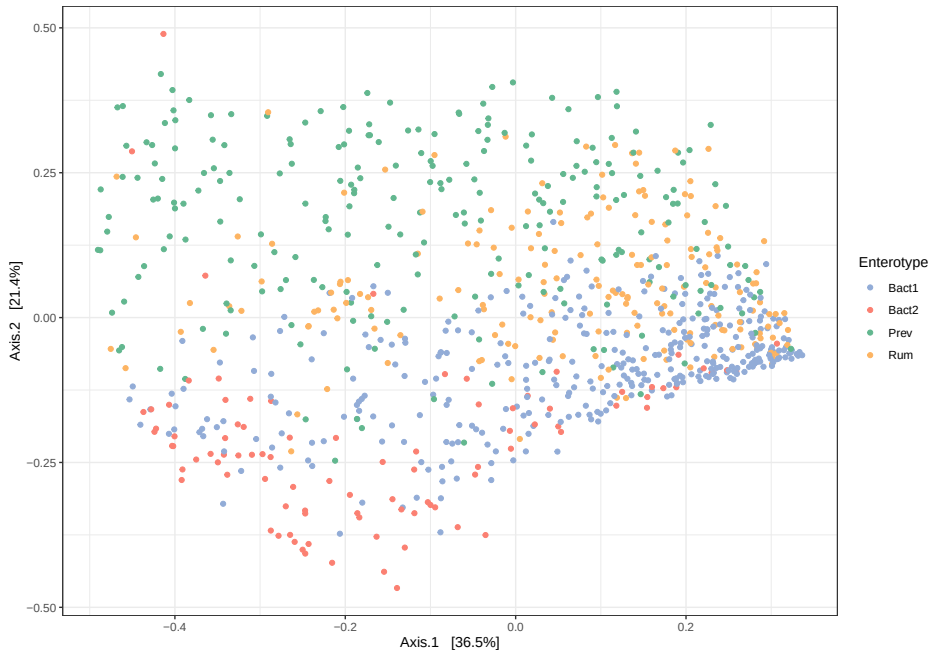
Visualizing your microbiome data

```
# Define a color palette
currentPAL=c("#92ACD7","#FB8072","#62B78F","#FDB462","light grey")

# Use phyloseq to manipulate and plot microbiome data
PHY=phyloseq(otu_table(QMP_genera,taxa_are_rows=FALSE),
             sample_data(mydata))

# Dimensionality reduction (aka ordination): PCoA
bc_ord <- ordinate(PHY,method = "PCoA",distance = "bray")
pPCoA=plot_ordination(PHY,bc_ord,color="Enterotype",
                      shape="Health.status") +
  theme_bw()  + scale_color_manual(values = currentPAL )
```

Data-driven approach: unconstrained ordination

Visualizing your microbiome data

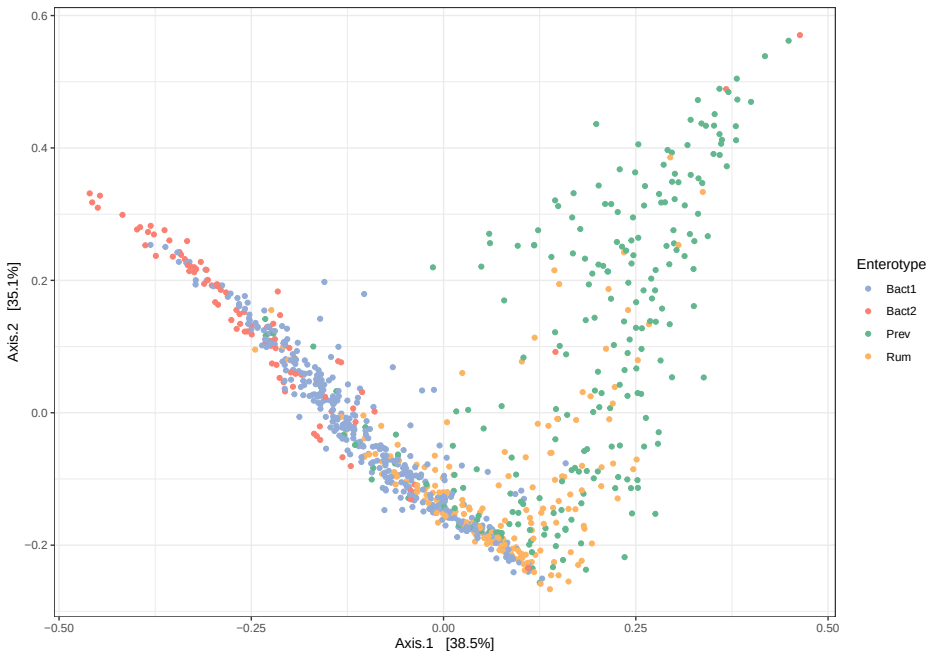
```
# Dimensionality reduction (aka ordination): PCoA
print(pPCoA)
```



Data-driven approach: unconstrained ordination

More used to seeing relative microbiome profiles in PCoA?

```
# change to relative microbiome data
RMP <- t(apply(QMP_genera,1, function(x) x/sum(x)))
PHY=phyloseq(otu_table(RMP,taxa_are_rows=FALSE),sample_data(mydata))
bc_ord <- ordinate(PHY,method = "PCoA",distance = "bray")
plot_ordination(PHY,bc_ord,color="Enterotype",shape="Health.status") + them
```



Data-driven approach: post-hoc fitting

Fitting clinical variables to unconstrained ordination

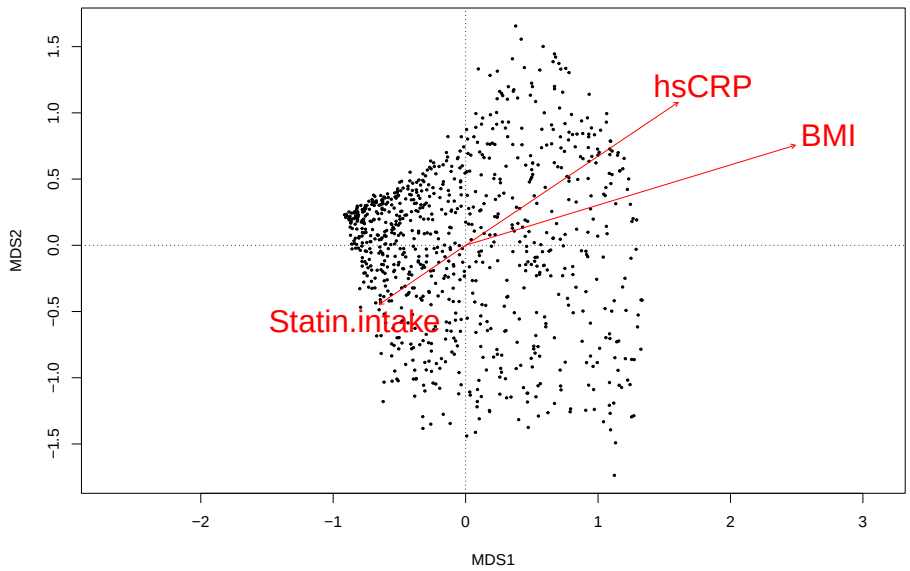
```
# unconstrained ordination
genera_bc <- vegdist(QMP_genera, "bray")
genera_cap  <- capscale(genera_bc ~ 1)
# post-hoc fit on ordination
envfit(genera_cap ~ .,data=numeric_data[,c(8,11,14,9)], na.rm=TRUE)
```

```
##
## ***VECTORS
##
##                    MDS1     MDS2     r2 Pr(>r)
## BMI             0.96273  0.27047 0.0715  0.001 ***
## hsCRP           0.83793  0.54578 0.0377  0.001 ***
## Statin.intake -0.82520 -0.56485 0.0061  0.067 .
## Fat_mass        0.74442  0.66771 0.0477  0.001 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## Permutation: free
## Number of permutations: 999
##
## 23 observations deleted due to missingness
```

Data-driven approach: post-hoc fitting

Fitting clinical variables to unconstrained ordination

```
# plotting post-hoc fit on ordination
plot(genera_cap,type="none") #plot PCoA
points(genera_cap, pch=20, cex=0.5)
plot(envfit(genera_cap ~ .,data=numeric_data[,c(8,11,14)], add=T,na.rm=T),c
```



Data-driven approach: constrained ordination

Constraining an ordination to clinical data (dbRDA)

```
# Constraining an ordination by dbRDA
var <- "BMI"
capsc <- capscale(genera_bc ~ METADATA_1[,var], na.action=na.omit)
#constrain effect size and significance (R2 and p-value)
signif_dbRDA <- anova.cca(capsc)
ES_dbRDA <- RsquareAdj(capsc)
paste0(var,", adjR2=",round(ES_dbRDA$r.squared,5),
       ",P=",signif_dbRDA$`Pr(>F)`[[1]])
```

```
## [1] "BMI, adjR2=0.03112,P=0.001"
```

Data-driven approach: multivariable constrained ordination

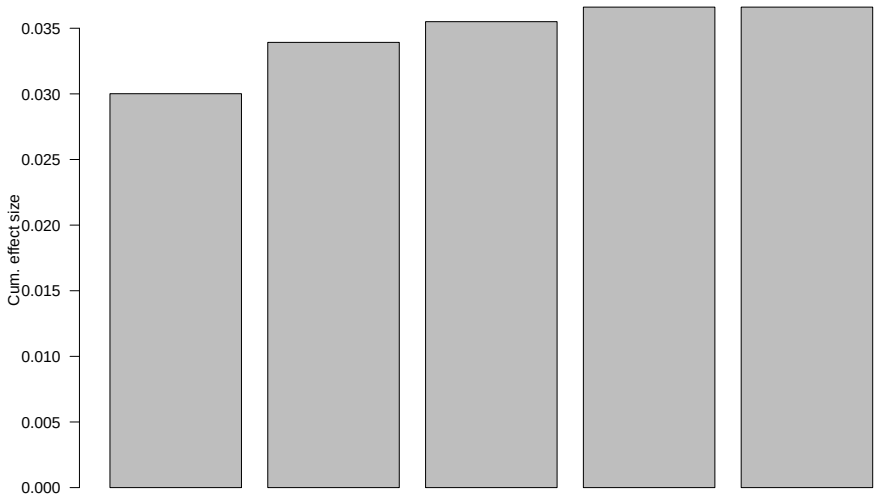
Constraining an ordination to more clinical variables (stepwise dbRDA)

```
# using nested dbRDAs for greedy selection of best forward model
vars <- c("BMI","Statin.intake","BSS","hsCRP")
mydata <- METADATA_1[,vars]
#impute missing data
mydata <- mydata %>% mutate_if(is.numeric,
            function(x) ifelse(is.na(x), median(x, na.rm = T), x))
# using nested dbRDAs for greedy selection of best forward model
attach(mydata)
mod0 <- capscale(genera_bc ~ 1) #H0: unconstrained ordination
mod1 <- capscale(genera_bc ~ ., mydata) #H1: constrained ordination
#running stepwise forward dbRDA
step.res <- ordiR2step(mod0, scope=formula(mod1), data=mydata ,
                       direction="forward", Pin = 1, R2scope = TRUE,
                       trace = F)
res <- step.res$anova
```

Data-driven approach: multivariable constrained ordination

Constraining an ordination to more clinical variables (stepwise dbRDA)

```
# Plotting stepwise dbRDA results
vars <- gsub("<All variables>", "all",
             gsub("+ ", "",labels(res)[[1]],fixed = T))
barplot(res$R2.adj, names.arg = vars,cex.names = 1,
        ylab= "Cum. effect size",las=2)
```



Hypothesis testing for biomarker discovery: model design

Hypothesis: Are some taxa abundances associated with BMI?

Model: BMI ~ Taxa

```
#select some taxa (usually minimum prevalence 25%)
mytaxa = c("Akkermansia","Bacteroides","Bilophila","Eggerthella","Escherich
myQMP=QMP_genera[,mytaxa]
summary(lm(mydata$BMI ~ rank(myQMP[,"Akkermansia"])))
```

```
##
## Call:
## lm(formula = mydata$BMI ~ rank(myQMP[, "Akkermansia"]))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.256  -8.537  -1.583   6.912  44.304
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     27.422684   0.717964  52.123  < 2e-16 ***
```

Hypothesis testing (biomarker discovery): model design

Hypothesis: Are some taxa abundances associated with BMI?

Model: BMI ~ Taxa

```
lm.res[order(lm.res$AdjP),]
```

```
##                              R2             P          AdjP
## Akkermansia        5.162195e-02 6.102735e-12 3.661641e-11
## Faecalibacterium   8.306056e-03 3.998489e-03 1.199547e-02
## Bacteroides       -2.044615e-04 3.650989e-01 3.650989e-01
## Bilophila         -1.931461e-05 3.217091e-01 3.650989e-01
## Eggerthella        5.019965e-04 2.305507e-01 3.650989e-01
## Escherichia       -8.780180e-06 3.194523e-01 3.650989e-01
```

Hypothesis testing (biomarker discovery): model refinement

Refine your hypothesis and model:

- Instead of "Are some taxa abundances associated with BMI?"
- Best: "Do certain taxa contribute to inflammatory load within obesity?" Model: hsCRP $\sim$ BMI + Taxa

Nested models

```
#Create nested models
neutral_model <- lm(mydata$hsCRP ~ mydata$BMI)
hyp_model <- lm(mydata$hsCRP ~ mydata$BMI + rank(myQMP[,"Akkermansia"]))
#test significance of super model
#anova(neutral_model,hyp_model)

myQMP=QMP_genera[,mytaxa]
nested.lm.res <- sapply(colnames(myQMP),function(x) {
  hyp_model <- lm(mydata$hsCRP ~ mydata$BMI + rank(myQMP[,x]));
  anova(neutral_model,hyp_model, test = "LRT")[2,5]})
```

Hypothesis testing (biomarker discovery): Nested models

Refine your hypothesis and model:

- Instead of "Are some taxa abundances associated with BMI?"
- Best: "Do certain taxa contribute to inflammatory load within obesity?" Model: hsCRP $\sim$ BMI + Taxa

Nested models

```
nested.lm.res
```

```
##       Akkermansia       Bacteroides         Bilophila       Eggerthella
##         0.4545174         0.1554496         0.6751050         0.9975358
##       Escherichia Faecalibacterium
##         0.2643297         0.3039230
```

Hypothesis testing: nested models

Exercise 1

Show that dysbiosis increases with obesity, and show that statins modulate this association

1. design your H0 and H1 models
2. adapt the code using glm(family = binomial())
3. compare extended model with models stratified by statin intake

- use: EXTENDED_DATA

Stratification for clinical discovery

why stratify?

- Dysbiosis and disease are not collinear
- Not all patients harbour dysbiotic microbiomes

stratification as model refinement

- Stratify by microbiome separately from diagnosis
- Quantify the eventual contribution of microbiome classes to disease (risk or pathomechanistic biomarker)

Stratification for clinical discovery

Exercise 2

Quantify the contribution of dysbiosis to inflammatory load in obesity

1. design your model of inflammatory load (hsCRP) and obesity (BMI)
2. Show and test deviations from this model in dysbiosis vs eubiosis

- use: METADATA_1