# 16S and cytokines combination

## MOMS-PI

The MOMS-PI data can be loaded as follows.

### 16S data

Load 16S data as a matrix, rows are Greengene IDs, columns are sample names:

```
data("momspi16S_mtx")
```

Load the Greengenes taxonomy table as a matrix, rows are Greengene IDs, columns are taxonomic ranks:

```
data("momspi16S_tax")
# Check if Greengene IDs match between the 16S and taxonomy data
# all.equal(rownames(momspi16S_mtx), rownames(momspi16S_tax)) # Should be TRUE
```

Load the 16S sample annotation data as a matrix, rows are samples, columns are annotations:

```
data("momspi16S_samp")

# Check if sample names match between the 16S and sample data
# all.equal(colnames(momspi16S_mtx), rownames(momspi16S_samp)) # Should be TRUE
```

The `momspi16S` function assembles those matrices into a `phyloseq` object.

```
momspi16S_phyloseq <- momspi16S()
momspi16S_phyloseq
```

```
## phyloseq-class experiment-level object
## otu_table()   OTU Table:         [ 7665 taxa and 9107 samples ]
## sample_data() Sample Data:       [ 9107 samples by 9 sample variables ]
## tax_table()   Taxonomy Table:    [ 7665 taxa by 7 taxonomic ranks ]
```

### Cytokine data

The MOMS-PI cytokine data can be loaded as a matrix, rownames are cytokine names, colnames are sample names:

```
data("momspiCyto_mtx")
dim(momspiCyto_mtx)
```

```
## [1]  29 872
```

Load the cytokine sample annotation data as a matrix, rows are samples, columns are annotations:

```
data("momspiCyto_samp")
dim(momspiCyto_samp)
```

```
## [1] 872   9
```

## Multi-table analysis

Combine 16S and cytokines data

```
#select data collected at the same visit
combined_samp <- merge(momspi16S_samp, momspiCyto_samp,
                       by = c("subject_id", "sample_body_site",
                        "project_name", "study_full_name",
                        "subject_gender", "subject_race",
                        "visit_number"))

table(combined_samp$visit_number)
```

```
##
##   1   2   3   4   5   6   7   8   9  10  11  12  13  14  15
## 115  86  80  99  78  60  51  48  59  39  19  10   3   2   1
```

Select first visit data, this assures that samples we look at were taken at the same time and at the first or second trimester. We don't have trimesters information in the data, but know it from the study design.

```
#select data from first visit only
combined_samp <- combined_samp[combined_samp$visit_number ==  1,]

table(combined_samp$sample_body_site)#all vaginal samples
```

```
##
## vagina
##    115
```

The two objects we use for combined 16S and cytokines analysis are: 'combined_16S_mtx' and 'combined_Cyto_mtx'. Phylogenetic information for those OTUs is available in 'tax_table(combined_16S_phyloseq)' object.

```
#select 16S data for those samples
combined_16S_phyloseq <- subset_samples(momspi16S_phyloseq, file %in% combined_samp$file.x)

#get rif of otus that are not observed in any sample for this subset
combined_16S_phyloseq %<>%
    taxa_sums() %>%
    is_greater_than(0) %>%
    prune_taxa(combined_16S_phyloseq)

combined_16S_mtx <- otu_table(combined_16S_phyloseq)

combined_Cyto_mtx <- momspiCyto_mtx[, colnames(momspiCyto_mtx) %in% combined_samp$file.y ]
dim(combined_Cyto_mtx)
```

```
## [1]  29 115
```

We match the samples (contained in columns of both tables) by the file names contained in colnames of each table.

In 'combined_samp' object the names of matched files names for 16S data are recorded in column 'file.x' and for cytokines data in column 'file.y'.

```
#make sure all samples across 3 tables are in the same order
combined_samp <- combined_samp[order(combined_samp$subject_id),]
#reorder cytokines samples
```

```
combined_Cyto_mtx <- combined_Cyto_mtx[,combined_samp$file.y]
#reorder taxa samples
combined_16S_mtx <- combined_16S_mtx[,combined_samp$file.x]
```

# Co-inertia analysis

## Basics:

- Let $Z_1$ and $Z_2$ be 16S and cytokines tables respectively
- rows: same $n$ women at first visit
- Columns: $p_1$ taxa, $p_2$ cytokines
- To visualize differences/similarities among taxa and cytokines data sets we view samples as rows while taxa and cytokines as columns
- PCA analysis for each table: $(X, Q_X, D)$ and $(Y, Q_Y, D)$
- Co-inertia axes: $Y^T DX = K\Lambda^{1/2}A^T$ of decomposition $(Y^T DX, Q_X, Q_Y)$
- Plot $F_X = XA$ and $F_Y = YK$

```
combined_16S_mtx <- t(combined_16S_mtx)
combined_16S_mtx <- combined_16S_mtx/apply(combined_16S_mtx, 1, sum)
combined_Cyto_mtx <- t(combined_Cyto_mtx)

#cut the last 5 characters that correspond to the -omics type identifier
rownames(combined_Cyto_mtx) <- substr(
  rownames(combined_Cyto_mtx), 1,nchar(rownames(combined_Cyto_mtx))-5)

rownames(combined_16S_mtx) <- substr(
  rownames(combined_16S_mtx), 1,nchar(rownames(combined_16S_mtx))-5)

#make sure all rownames match
all(rownames(combined_16S_mtx) == rownames(combined_16S_mtx))
```

```
## [1] TRUE
```

We first center 16S data to work with PCA on the covariance matrix $\Sigma_X = Cov(X)$ and then, to normalize the data, we divide each value of $X$ by the total variance: $\sqrt{\text{tr}(\Sigma_X)}$, which is equivalent to dividing the matrix by $\sqrt{\sum_{k=1}^{r} \lambda_k}$, where $\lambda_k$ are the eigevalues of $\Sigma_X$ and $r$ is the rank of $X$. This is the standartization approach used in multiple co-inertia analysis, which combines several tables.

```
taxa_mtx <- scale(combined_16S_mtx, center = TRUE, scale = FALSE)
#use fast trace computation formula: tr(A^B) = sum(A*B), where '*' operator refers to elemetwise produc
taxa_tr <- sum(taxa_mtx*taxa_mtx)/(dim(taxa_mtx)[1]-1)
taxa_mtx <- taxa_mtx/sqrt(taxa_tr)
taxa.pca <- dudi.pca(taxa_mtx, scannf=FALSE, nf =61,
                     center = FALSE, scale = FALSE)
```

Cytokines PCA on centered and scaled data, also normalized by the square root of total variances.

```
cyto_mtx <- scale(combined_Cyto_mtx, center = TRUE, scale = TRUE)
cyto_tr <- sum(cyto_mtx*cyto_mtx)/(dim(cyto_mtx)[1]-1)
cyto_mtx <- cyto_mtx/sqrt(cyto_tr)
```

```
cyto.pca <- dudi.pca(cyto_mtx, scannf=FALSE, nf =61,
                     center = FALSE, scale = FALSE)
```

## Combine the tables using co-inertia

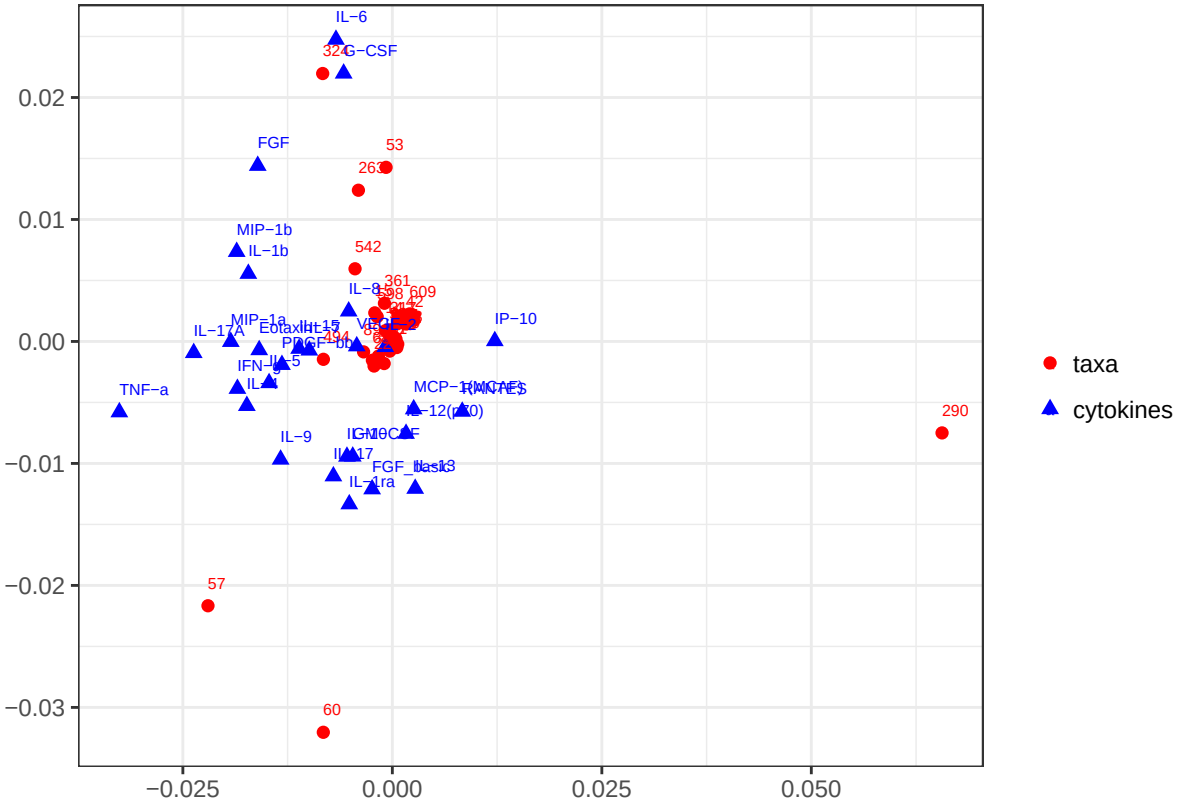
```
coin <- coinertia(taxa.pca, cyto.pca, scannf = FALSE, nf = 2)
```

RV coefficient – measure of similarity between 16S and cytokines tables

```
RV<- coin$RV
RV
```

```
## [1] 0.04494507
```

Plots of variables weights: interpretation is similar to interpretation of PCA variables plots. Cytokines (blue) projected in the same direction as taxa (red) have more similarity.

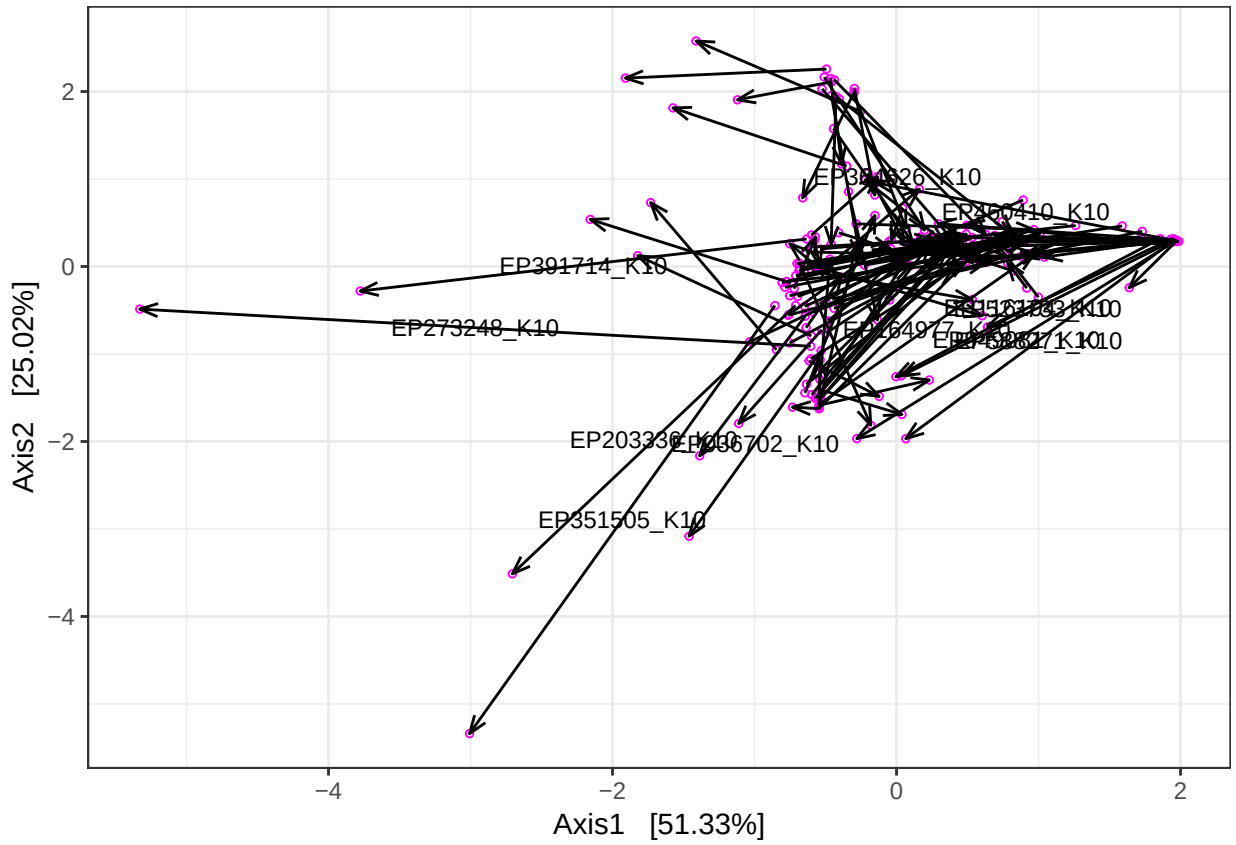



Taxa that correspond to larger lodings: 324, 53, 263, 542, 290, 494, 57, 60

```
taxa.inx <- c(324, 53, 263, 542, 290, 494, 57, 60)
taxa.ids <- colnames(combined_16S_mtx)[taxa.inx]
#look up these taxa ids in phyloqes
tax_table(momspi16S_phyloseq)[rownames(tax_table(momspi16S_phyloseq)) %in% taxa.ids,c("Genus", "Species
```

```
## Taxonomy Table:     [8 taxa by 2 taxonomic ranks]:
##        Genus           Species
## 134467 "Lactobacillus" NA
## 137183 "Gardnerella"   NA
## 137580 "Lactobacillus" NA
## 318320 "Lactobacillus" NA
## 332718 "Lactobacillus" NA
## 354905 "Lactobacillus" NA
## 469663 "Atopobium"     "vaginae"
## 529233 "Streptococcus" NA
```

Sample scores plots. Length of the arrows indicates the samples that have larger differences across two data sets.




Samples with largest difference across two data sets. Samples with arrow lengths in 0.9 quatile are chosen.

```
#Taxa with major differences across two sets
rownames(Samp.coin$Dissimilarity[Samp.coin$Dissimilarity$Quantile >= 0.9, ])
```

```
##  [1] "EP460410_K10" "EP364626_K10" "EP164977_K10" "EP036702_K10"
##  [5] "EP116104_K10" "EP523733_K10" "EP588271_K10" "EP949081_K10"
##  [9] "EP391714_K10" "EP203336_K10" "EP273248_K10" "EP351505_K10"
```