16S and cytokines combination

Check: Evaluation of integrative clustering methods for the analysis of multi-omics data https://academic.oup.com/bib/advance-article-abstract/doi/10.1093/bib/bbz015/5316049?redirectedFrom=fulltext

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
```

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

The function momspiCytokines will make a SummarizedExperiment containing cytokine data

```
momspiCyto <- momspiCytokines()
momspiCyto
```

```
## class: SummarizedExperiment
## dim: 29 872
## metadata(0):
## assays(1): ''
## rownames(29): 1 2 ... 28 29
## rowData names(0):
## colnames(872): EP036702_K10_MVAX EP062329_K10_MP1P ...
##   EP936022_K10_MVAX EP949081_K10_MVAX
## colData names(9): sample_id subject_id ... project_name file
```

The cytokine data contains data for 29 cytokines over 872 samples.

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))
```

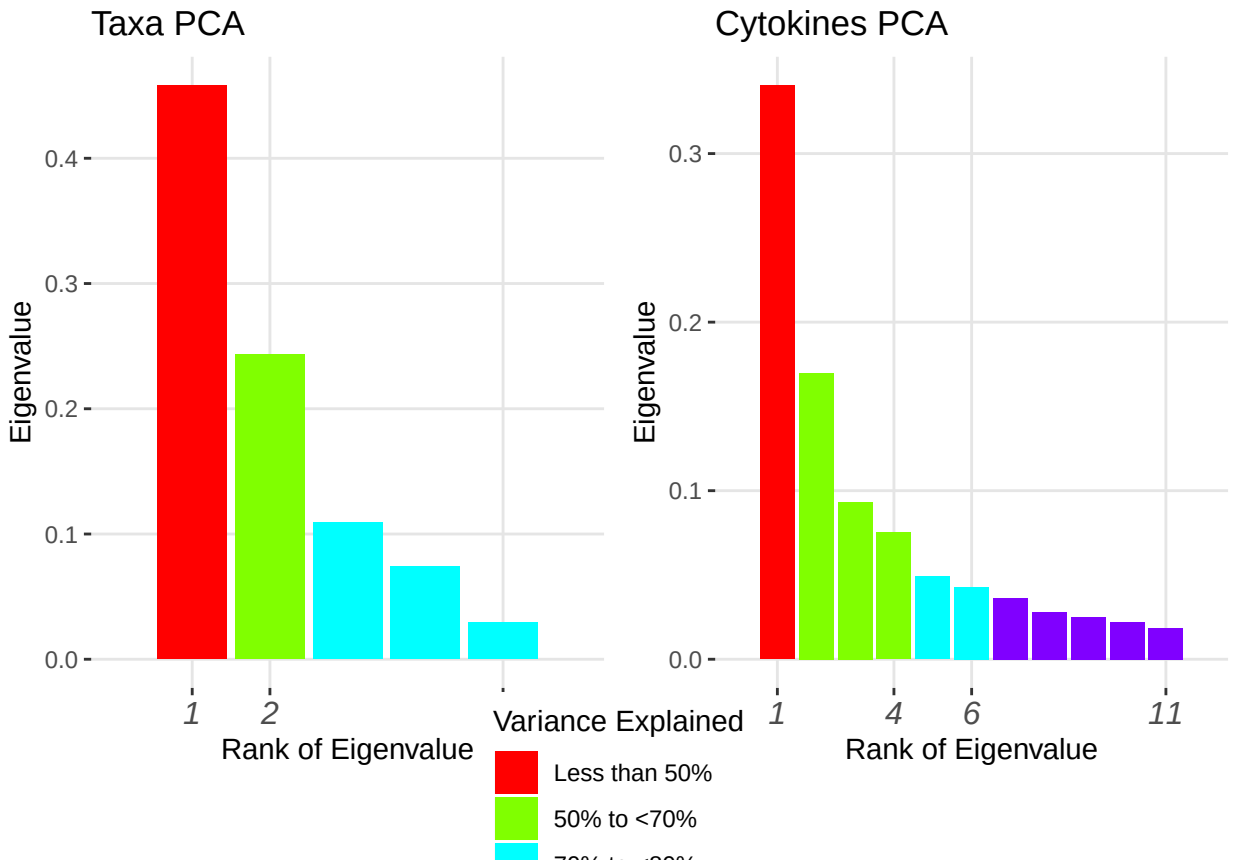
```
## [1] TRUE
```

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 λ_k are the eigevalues of Σ_X and r is the rank of X. This is the standartization approach used in multiple co-inertia analysis, which combines several tables.

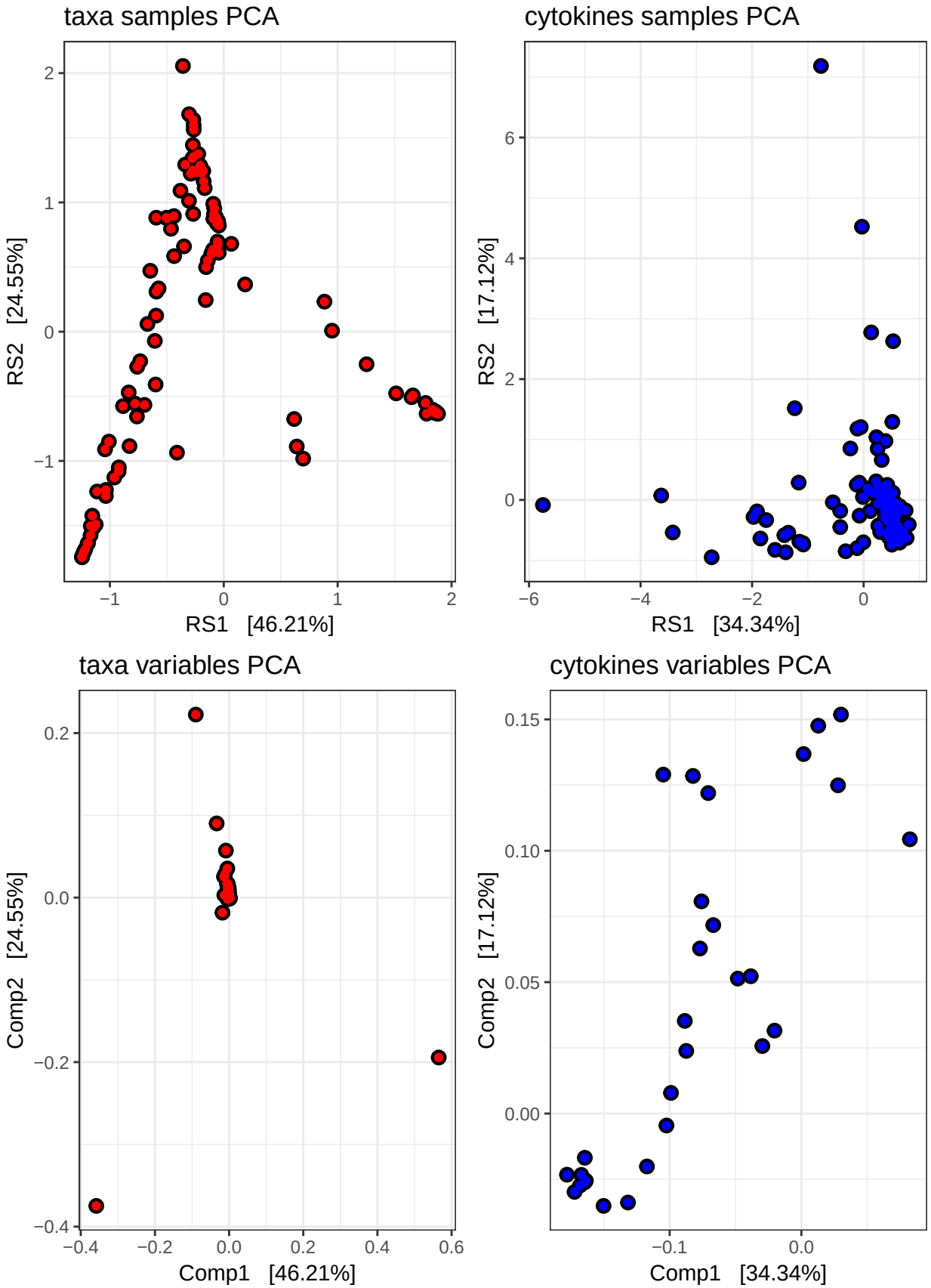
```
taxa_mtx <- scale(combined_16S_mtx, center = TRUE, scale = FALSE)
#use fast trace computation formula: tr(A^B) = sum(A*B), where '*' operator refers to elemetwise produc
taxa_tr <- sum(taxa_mtx*taxa_mtx)/(dim(taxa_mtx)[1]-1)
taxa_mtx <- taxa_mtx/sqrt(taxa_tr)
taxa.pca <- dudi.pca(taxa_mtx, scannf=FALSE, nf =61,
                     center = FALSE, scale = FALSE)
breaks <- c(50, 70, 80, 90)
scree.taxa.pca <- screePlot(taxa.pca, breaks) +
           theme( panel.background = element_rect(fill = "white"),
                  panel.grid.major = element_line(colour = "grey90"))
```

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)
breaks <- c(50, 70, 80, 90)
scree.cyto.pca <- screePlot(cyto.pca, breaks) +
           theme( panel.background = element_rect(fill = "white"),
                  panel.grid.major = element_line(colour = "grey90"))
```




```
## TableGrob (2 x 1) "arrange": 2 grobs
##   z     cells    name              grob
## 1 1 (1-1,1-1) arrange   gtable[arrange]
## 2 2 (2-2,1-1) arrange gtable[guide-box]
```


taxa samples PCA
RS2 [24.55%]
RS1 [46.21%]
cytokines samples PCA
RS2 [17.12%]
RS1 [34.34%]
taxa variables PCA
Comp2 [24.55%]
Comp1 [46.21%]
cytokines variables PCA
Comp2 [17.12%]
Comp1 [34.34%]

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

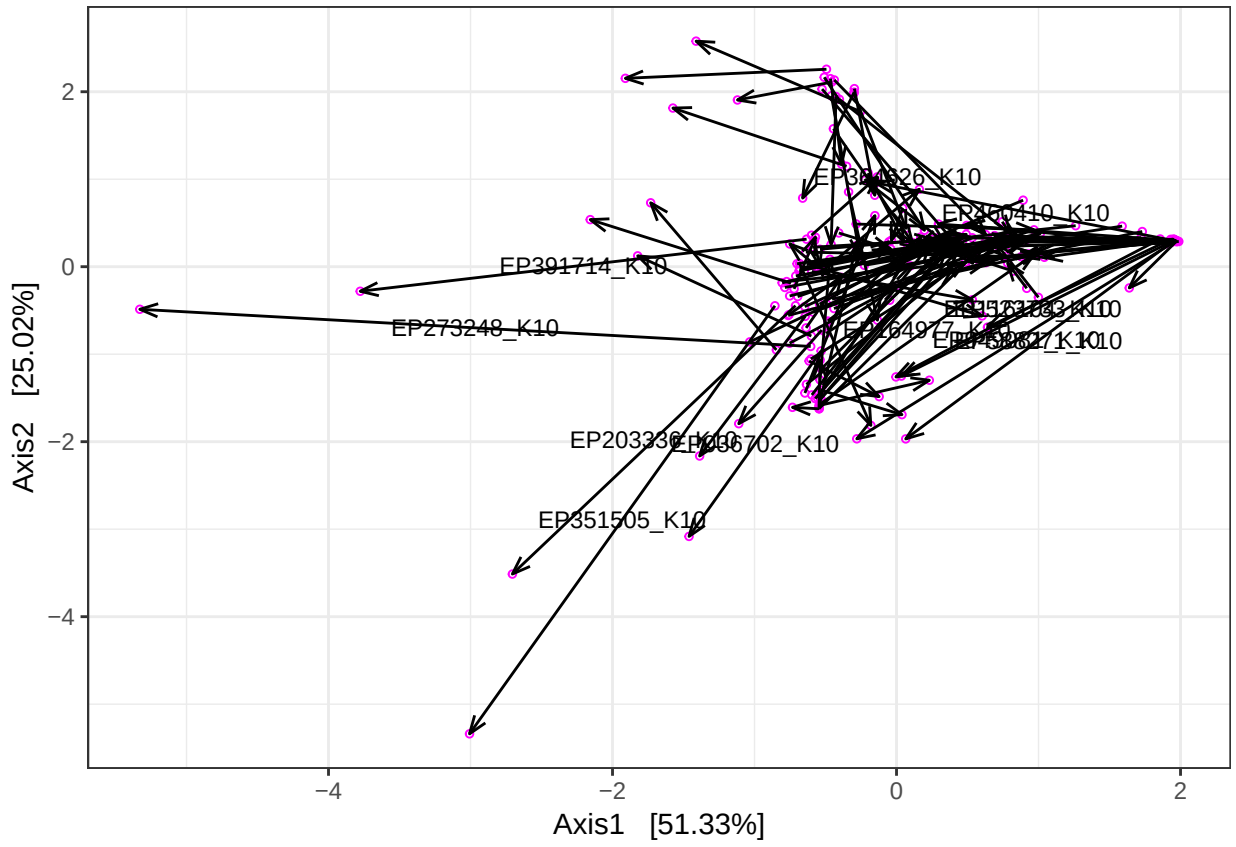
```
RV<- coin$RV
RV
```

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

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

```
Samp.coin <- CoinertiaPlot(coin = coin,
                   Quant = 0.9, Prop.Var = 0.9,
                   Env_Var=NULL,
                   Env_Var2 = NULL,
                   color=NULL, shape=NULL,
                   PtColor= "magenta",PtShape=1, PtSize=1,
                   linetype=1, LblSize=3,
                   LabelsOpt = NULL,
                   ArrLen=0.10, ArrAngle=20)

Samp.coin$p + theme_bw()
```



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

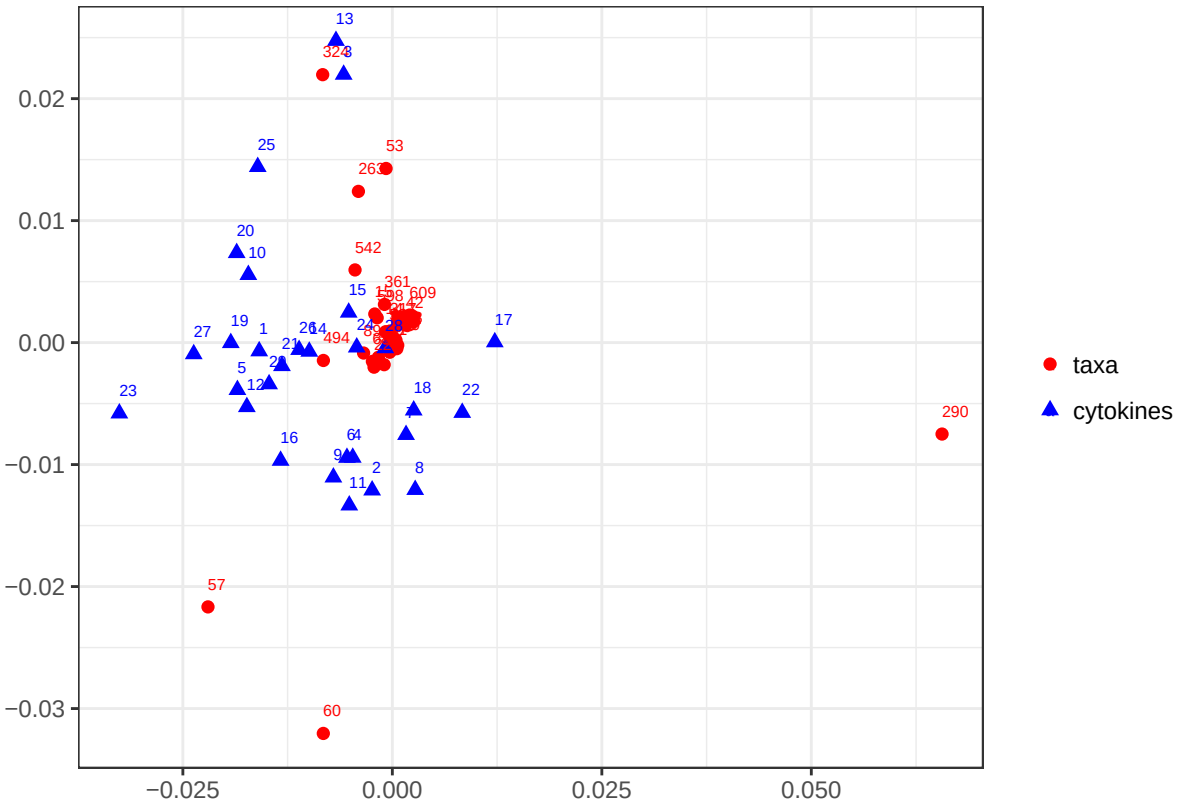
```
#Taxa with major differences across two sets
rownames(Samp.coin$Dissimilarity[Samp.coin$Dissimilarity$Quantile >= 0.9, ])
```

```
##  [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"
```

Plots of taxa weights

```
#Plot Canonical weights based on No BV taxa as reference point
CW.coin.ca <- PlotCW(coin = coin, name="Coin", path = pathtoplots, color = "magenta",
            Title1 = "Canonical weights for taxa",
            Title2 = "Canonical weights for cytokines",
            Labels1 = colnames(combined_16S_mtx), Labels2 = colnames(combined_Cyto_mtx),
            scale = TRUE,
            PtShape=2, PtSize=2,linesize=0.4,
            linetype=1, LblSize=2.5,
            ArrLen=0.15, ArrAngle=10,
            TitleSize = 10)
```

```
p.vars <- PlotCoinVars(coin, tab1 = "taxa", tab2 = "cytokines",
             Labels1 = NULL, #colnames(combined_16S_mtx)
             Labels2 = colnames(combined_Cyto_mtx),
             label = TRUE,PtSize=2,LblSize=2,
             hjust = 0, vjust = -1.5)
p.vars
```



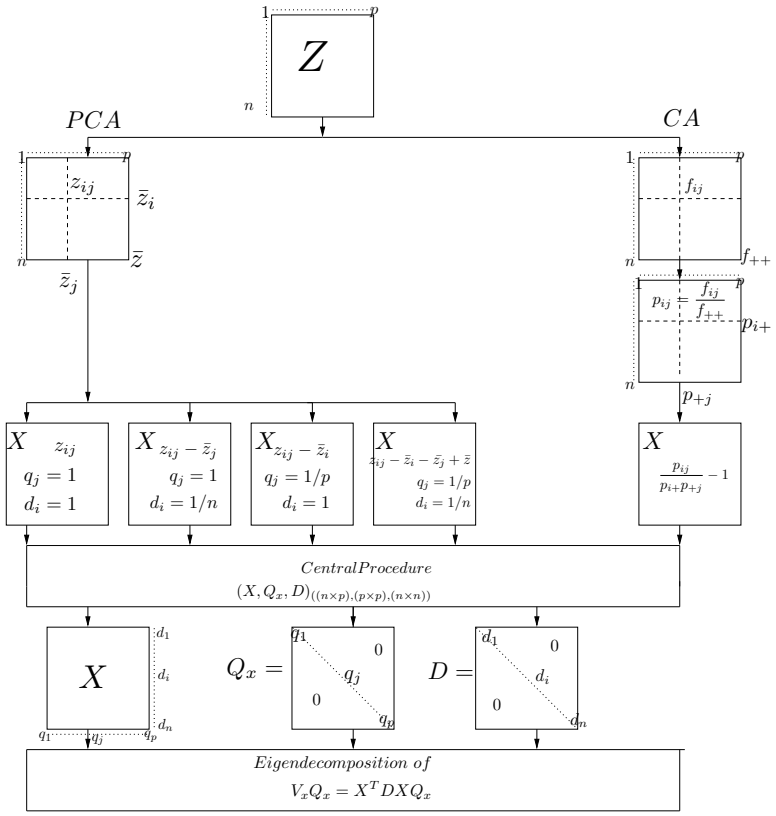
Taxa that correspond to larger lodings: 324, 53, 263, 542, 290, 494, 57, 60

```r
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
```

Additional details about the method.

- $Z_{n\times p}$ data matrix
  - n samples
  - p variables (taxa)
- Transformation $Z_{n\times p} \to X_{n\times p}$
  - Different transformation for each ordination method
  - Column weights: $Q_{p\times p}$
  - Row weights: $D_{n\times n}$
- Statistical triplet: (X, Q, D)
- Eigen decomposition of $(WD)_{n\times n} = XQX^TD$ or $(VQ)_{p\times p} = X^TDXQ$
- Write: $X = K\Lambda^{1/2}A^T$
  - $K^TDK = \mathbb{I}_r$, K - samples scores
  - $A^TQA = \mathbb{I}_r$, A - taxa loadings
  - r rank of X



Given statistical triplet (X, Q, D) the PCA is

- $X_{n\times p}$ - column centered and scaled matrix
  - n number of samples
  - p number of taxa
- $Q = \mathbb{I}_p$
- Correlation PCA: $Q = diag\{sd(\boldsymbol{z}_1), \ldots, sd(\boldsymbol{z}_p)\}$, where $sd(\boldsymbol{z}_j)$ standard deviation for j^{th} column of original data matrix Z
- $D = diag\{\frac{1}{n}, \ldots, \frac{1}{n}\} = \frac{1}{n}\mathbb{I}_n$
- Then eigendecomposition of
$$VQ = X^T DXQ = \frac{1}{n} X^T X = \Sigma$$
- Inertia: $\mathcal{I} = tr(\Sigma) = \sum_{i=1}^{p} \sigma_i^2 = \sum_{i=1}^{r} \lambda_i$

Co-inertia analysis idea:

- Analysis of $Z_1, Z_2, \ldots, Z_K$. Same rows, but different columns
- Z_1 – 16S data
- Z_2 – cytokines data

1. Co-Inertia Analysis
    - Seeks a common structure between two datasets Z_1 and Z_2
    - Two tables: Z_1 and Z_2 of dimensions $n \times p_1$ and $n \times p_2$
    - Use any one table method to obtain triplets: (X, Q_X, D) and (Y, Q_Y, D)
    - coinertia – coinertia analysis function in ade4
    - Co-inertia triplet: $(Y^T DX, Q_X, Q_Y)$
    - Decomposition: $Y^T DX = K\Lambda^{1/2} A^T$
    - Project X and Y onto A and K respectively
    - Let $A = XQ_X X^T D$ and $B = YQ_Y Y^T D$
    - RV coefficient: $RV(A, B) = \frac{tr(AB^T)}{\sqrt{tr(AA^T)tr(BB^T)}} = \frac{\mathcal{I}_{XY}}{\sqrt{\mathcal{I}_X}\sqrt{\mathcal{I}_Y}}$
    - $RV \in [0, 1]$ with values closer to 1 indicating stronger similarity among X and Y

Objects in ade4

SVD of a transformed table $X = K\Lambda^{1/2} A^T$

k – rank of X

matrix	dimensions	ade4	description
Z	$n \times p$		raw data; argument for dudi functions
X	$n \times p$	\$tab	transformed data
Q	$p \times p$	\$cw	diagonal matrix of column weights
D	$n \times n$	\$lw	diagonal matrix of row weights
A	$p \times k$	\$c1	the principal axes or column normed scores
K	$n \times k$	\$l1	the row normed scores
Λ	$k \times k$	\$eig	diagonal matrix of the k largest eigenvalues
$A\Lambda^{1/2}$	$p \times k$	\$co	column coordinates
$K\Lambda^{1/2}$	$n \times k$	\$li	the row coordinates