16S and cytokines combination

```
#token 35f0903098e0c9ef30faf7514e382e6bbd5c7179
#BiocManager::install("HMP2Data")

#devtools::install_github("jstansfield0/HMP2Data", auth_token = "35f0903098e0c9ef30faf7514e382e6bbd5c71
```

```
library(HMP2Data)
library(phyloseq)
library(SummarizedExperiment)
library(MultiAssayExperiment)
library(dplyr)
library(magrittr)
library(Hmisc)
library(colorspace)
library(ade4)
source("ScreePlot.R")
source("CIAPlots.R")

pathtoplots <- file.path(getwd(), "Plots/")#change me
```

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 using distances

Principal coordinates analysis (PCoA)

PCoA, usually called MDS in the statistics literature, is an eigen-decomposition of a distance matrix. Pairwise distances between samples (rows) of $\mathbf{X}$ are computed with respect to some distance metric; popular metrics are Bray-Curtis, Jaccard, Jensen-Shannon, UniFrac, and weighted UniFrac distances. The distance matrix, denoted here by the $n \times n$ matrix $\mathbf{B}$, is then squared element-wise, $\mathbf{B}^{(2)}$, and centered according to a procedure proposed by *Gower*, to give $\mathbf{H}$ as

$$\mathbf{H} = -\frac{1}{2} D^{1/2} \mathbf{F} \mathbf{B}^{(2)} \mathbf{F} \mathbf{D}^{1/2} = \mathbf{V} \mathbf{\Lambda} \mathbf{V}^T, \text{ where } \mathbf{F} = (\mathbf{I}_n - \mathbf{D}\mathbf{J}), \quad (1)$$

where $\mathbf{D}$ is the row weights matrix from above and $\mathbf{J}$ is an $n \times n$ matrix of ones. The $\mathbf{H}$ in eqn. (1) can be rewritten as $\mathbf{H} = \tilde{\mathbf{X}}\tilde{\mathbf{X}}^T$ with $\tilde{\mathbf{X}} = \mathbf{V}\mathbf{\Lambda}^{1/2}$ giving *approximate Euclidian distances for each sample from the origin*. Since PCoA eigen-decomposes a standardized matrix of distances between samples, only sample scores as a function of the eigenvectors are available for further examination and information about individual taxa is unavailable,

$$\begin{cases} \mathbf{A} = \mathbf{I} \\ \mathbf{K} = \mathbf{D}^{-1/2}\mathbf{V} \quad \text{sample scores,} \end{cases}$$

with the taxa loadings matrix $\mathbf{A}$ as the identity matrix.

General Framework

As shown in Figure below, the ordination method-specific matrices of sample scores $\mathbf{K}$, taxa loadings $\mathbf{A}$ and eigenvalues $\mathbf{\Lambda}$ derived from the eigen-decomposition of $\mathbf{H}$ correspond to the singular value decomposition of $\tilde{\mathbf{X}}$. The results of this one–table decomposition in *ade4* contain a table of sample scores, denoted \$l1, and a table of taxa loadings, denoted \$c1. The scores and loadings are used to construct lower dimensional visualizations of samples and taxa, especially via biplots. The entries in $\mathbf{\Lambda^{1/2}}$ (these are standard deviations in PCA) reflect the importance of the dimensions in the singular value decomposition; mathematically they are the lengths of the axes. For visualization purposes they are used to scale \$l1 and \$c1 to obtain the scaled sample scores (\$li) and scaled taxa loadings (\$co) which represent the configuration of samples or taxa in various plots.

$$\tilde{\mathbf{X}} = \underbrace{\mathbf{K}\ \mathbf{\Lambda}^{\frac{1}{2}}}_{\$li}\ \underbrace{\mathbf{A}^T}_{(\$c1)^T} = \underbrace{\mathbf{K}}_{\$l1}\ \underbrace{\mathbf{\Lambda}^{\frac{1}{2}}\ \mathbf{A}^T}_{(\$co)^T}$$

Co-inertia analysis (CoIA)

Co-inertia analysis allows for simultaneous exploratory analysis of two multidimensional tables by linking the common rows (samples) so that data across the columns (eg. taxa) of the two tables can be visualized. CoIA can thus be used for visualization of microbiome data collected longitudinally on the same samples at two different times. Implementations of CoIA in the literature *Dray3* rely on PCA, CA, or NSCA. Covariance PCA and NSCA are highly biased towards abundant taxa, while correlation PCA and CA give a lot more weight to rare taxa. Distance based methods, especially those relying on UniFrac or Jensen-Shannon divergence, are less prone to either of these extremes, often have biological interpretation, and are widely used in microbiome data analyses. Here we extend CoIA using distance matrices derived from OTU tables on the same subjects at two different time periods.

First, from the OTU tables at two different times (eg. baseline and disease state, or different stages of pregnancy), we construct distances matrices from separate PCoAs to give $\tilde{\mathbf{X}}_1$ and $\tilde{\mathbf{X}}_2$. Next, as suggested in *Dray3*, the two matrices are coupled to give $\mathbf{H}$ which is then eigen-decomposed,

$$\mathbf{H} = (\tilde{\mathbf{X}}_2^T\mathbf{D}\tilde{\mathbf{X}}_1)^T(\tilde{\mathbf{X}}_2^T\mathbf{D}\tilde{\mathbf{X}}_1) = \mathbf{U}\Lambda\mathbf{U}^T = (\mathbf{V}\Lambda^{1/2}\mathbf{U}^T)^T(\mathbf{V}\Lambda^{1/2}\mathbf{U}^T), \tag{2}$$

with $\mathbf{D}$ as the row weights matrix. The eigen-decomposition in eq. (2) is used to obtain co-inertia axes that are *common* to both OTU tables,

$$\begin{cases} \mathbf{V} = (\tilde{\mathbf{X}}_2^T \mathbf{D} \tilde{\mathbf{X}}_1)\mathbf{U}\mathbf{\Lambda}^{-1/2} & \text{OTU 1 co-inertia axes} \\ \mathbf{U} & \text{OTU 2 co-inertia axes.} \end{cases}$$

Finally, the common representation of samples from time 1 to 2 are given by the co-inertia sample scores

$$\begin{cases} \mathbf{G}_1 = \tilde{\mathbf{X}}_1 \mathbf{V} & \text{OTU 1 co-inertia sample scores} \\ \mathbf{G}_2 = \tilde{\mathbf{X}}_2 \mathbf{U} & \text{OTU 2 co-inertia sample scores.} \end{cases} \tag{3}$$

The distance between the i^{th} sample's co-inertia sample scores is obtained by comparison, on the r axes, of the co-inertia sample scores for two tables 1 and 2,

$$d_i = \sqrt{\Big(\sum_{j=1}^{k} (g_{ij}^1 - g_{ij}^2)^2 \Big)}, \tag{4}$$

where $\boldsymbol{g}_i^1 = (g_{i1}^1, \ldots, g_{ir}^1)$ and $\boldsymbol{g}_i^2 = (g_{i1}^2, \ldots, g_{ir}^2)$ are the i^{th} rows of sample score matrices $\mathbf{G}_1$ and $\mathbf{G}_2$ respectively. Smaller values for the distance d_i indicates more similarity in i^{th} sample's distance information across the two tables.