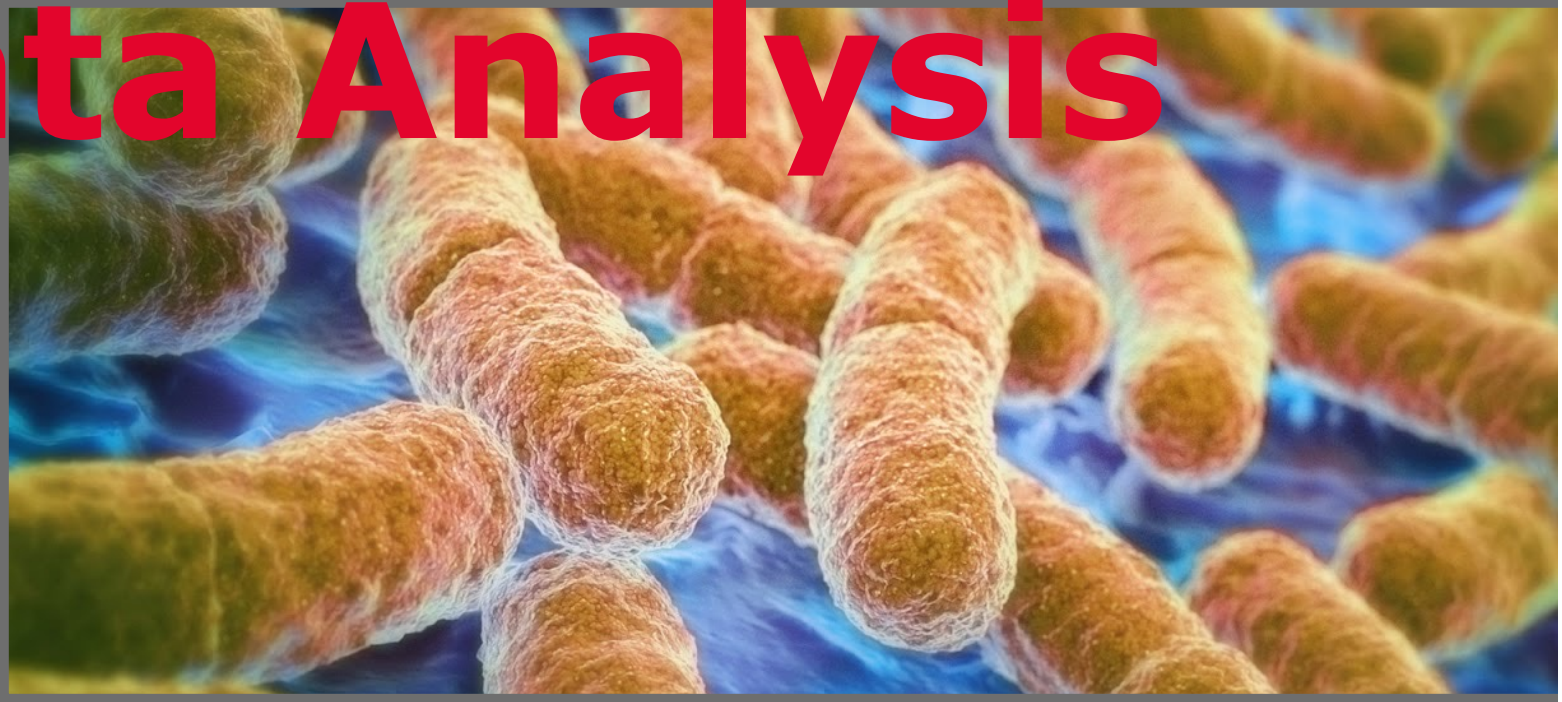
MICROBIOME Data Analysis



* A site which points you towards data analysis resources for microbiome data

https://ucph-foodmicro.github.io/UCPH-FOODMICRO/

* The site used for this course including tutorial material and exercises

https://mortenarendt.github.io/MicrobiomeDataAnalysis/index.html

Purpose

- To descriptive describe the individual communities.
- To compare with external data
- To integrate with other layers of *-omics* type data.

Problem

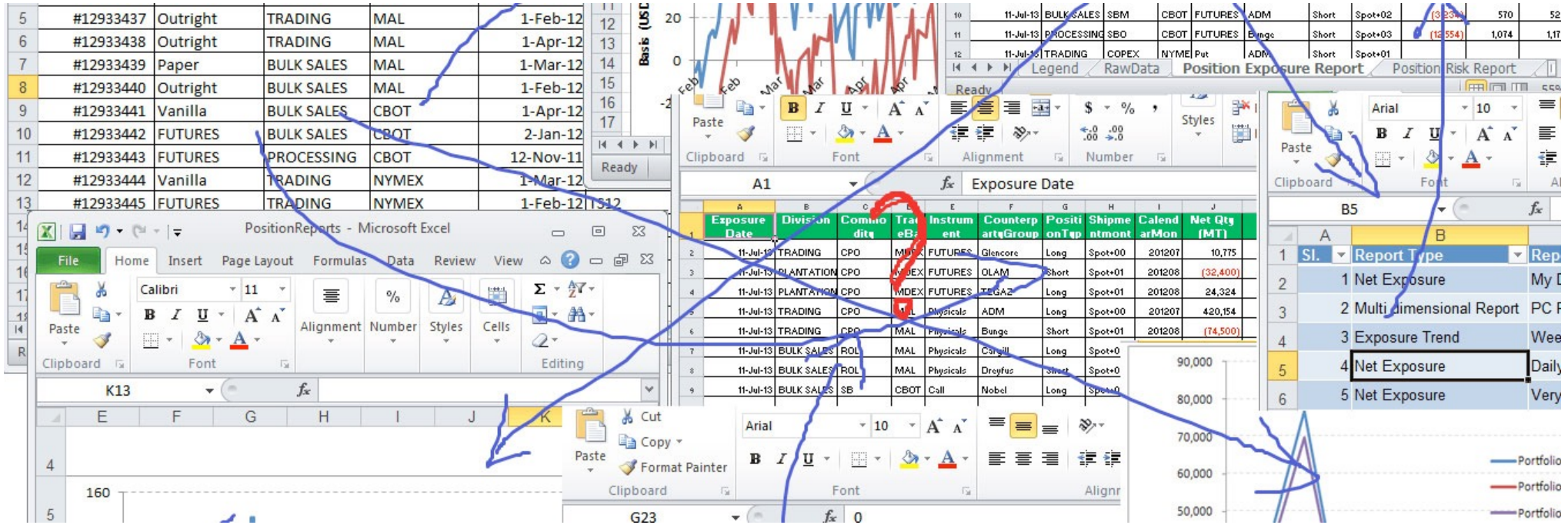
- Sparse / Zero inflated
- Compositional
- Over dispersed

Outline

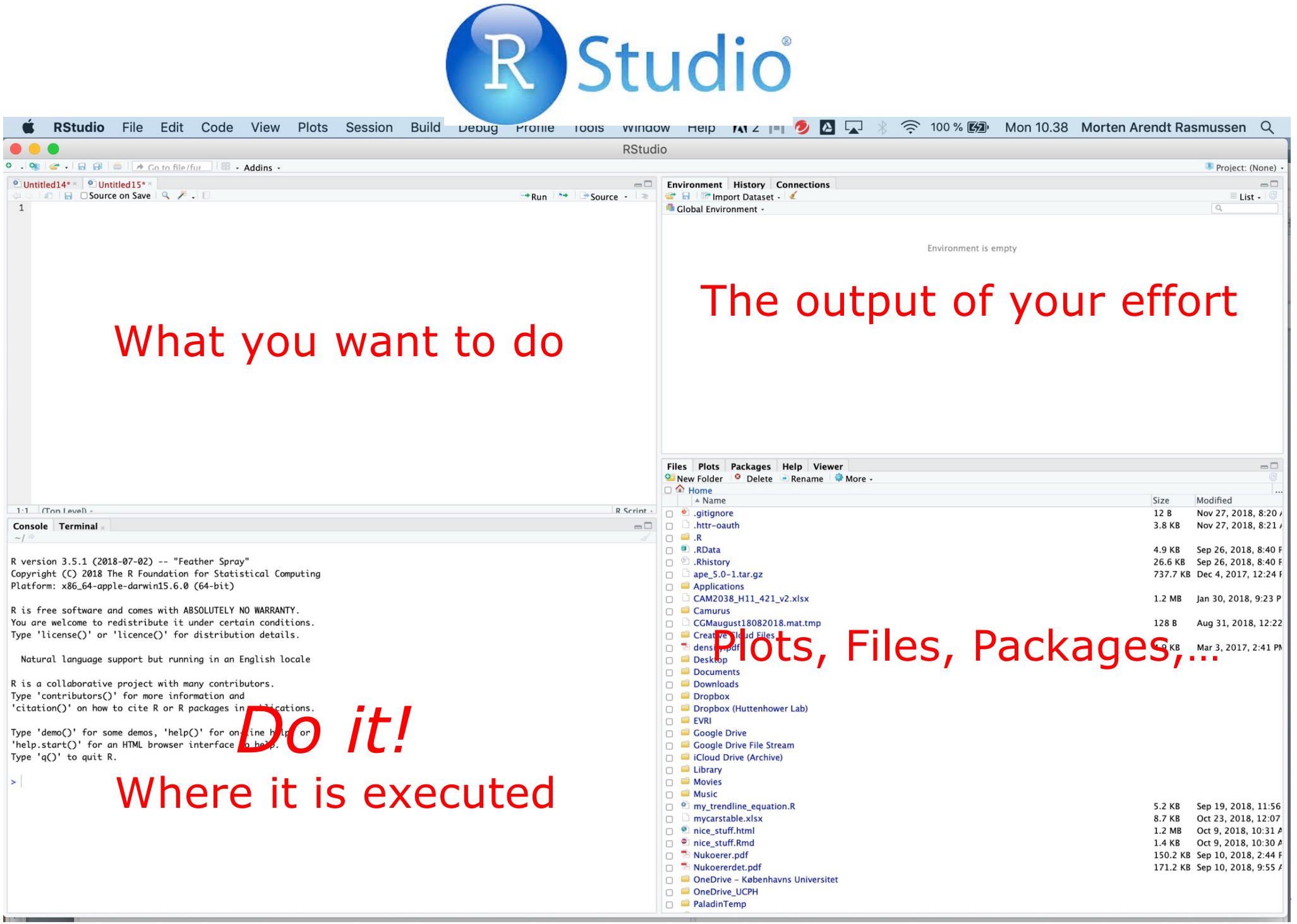
Day1 STAT101 recap Preprocessing	**Day2 Morning** Alpha diversity Beta diversity Testing design versus Beta diversity Permanova	**Day3 Morning** More DA Multiomics with heatmaps
	Day2 Afternoon Differential Abundance testing	**Day3 Afternoon** Multiomics - CCA Mediation and Propensity scoring

Why R??

- Digitalization *is everywhere*
- Reproducibility
- Get *quicker insight* and *More knowledge* out of your data


#12933437 Outright TRADING MAL 1-Feb-12
#12933438 Outright TRADING MAL 1-Apr-12
#12933439 Paper BULK SALES MAL 1-Mar-12
#12933440 Outright BULK SALES MAL 1-Feb-12
#12933441 Vanilla BULK SALES CBOT 1-Apr-12
#12933442 FUTURES BULK SALES CBOT 2-Jan-12
#12933443 FUTURES PROCESSING CBOT 12-Nov-11
#12933444 Vanilla TRADING NYMEX 1-Mar-12
#12933445 FUTURES TRADING NYMEX 1-Feb-12
PositionReports - Microsoft Excel
File Home Insert Page Layout Formulas Data Review View
Paste Cut Copy Format Painter
Clipboard Font Alignment Number Styles Cells Editing
Exposure Date
Position Exposure Report
Position Risk Report
Net Exposure
Multi dimensional Report
Exposure Trend
Portfolio

Excel Hell


RStudio
What you want to do
The output of your effort
Do it!
Where it is executed
Plots, Files, Packages,...

Packages

R comes with some functionality, but not everything is covered.

Therefor we need additional functions.

There comes in the form of ***packages*** which is installed directly from within R.

> install.packages('ggplot2')

From
CRAN (http://cran.r-project.org/)
(13713 available packages)

> BiocManager::install('metagenomeSeq')

From
Bioconductor (https://www.bioconductor.org/)

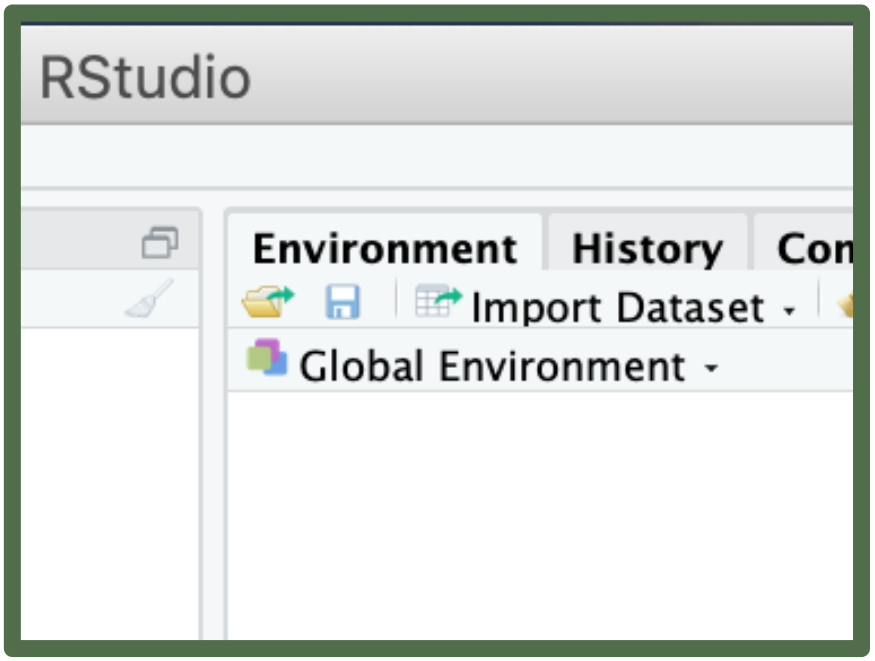
> devtools::install_github('vqv/ggbiplot')

From
github (https://github.com/)

Data import

Use the *point-n-click* method in Rstudio

…. Eventually copy paste the commands produced into the script



Or

Use the **rio** package

```
> library(rio)
> X <-
import('myExcelFile.xlsx')
```

data.frame()

You should store your data in a data.frame

A data frame is as an excel sheet with first row being variable names.

To be aware of:

Avoid using space, leading numerics and repetitions in names.

Some useful functions to look at the data.frame

> head() / tail()
> View()
> rownames()
> str()
> colnames()

Descriptive stats

Calculate descriptive stats – *a single number describing a distribution of numbers* – splitted according to grouping information.

Functions

> mean(), sd(), min(), max(), range(), length()

> aggregate()

Overlook of a design

How many observations are there for a combination of (design) variables

> table()

STAT 101 recap

A Linear model

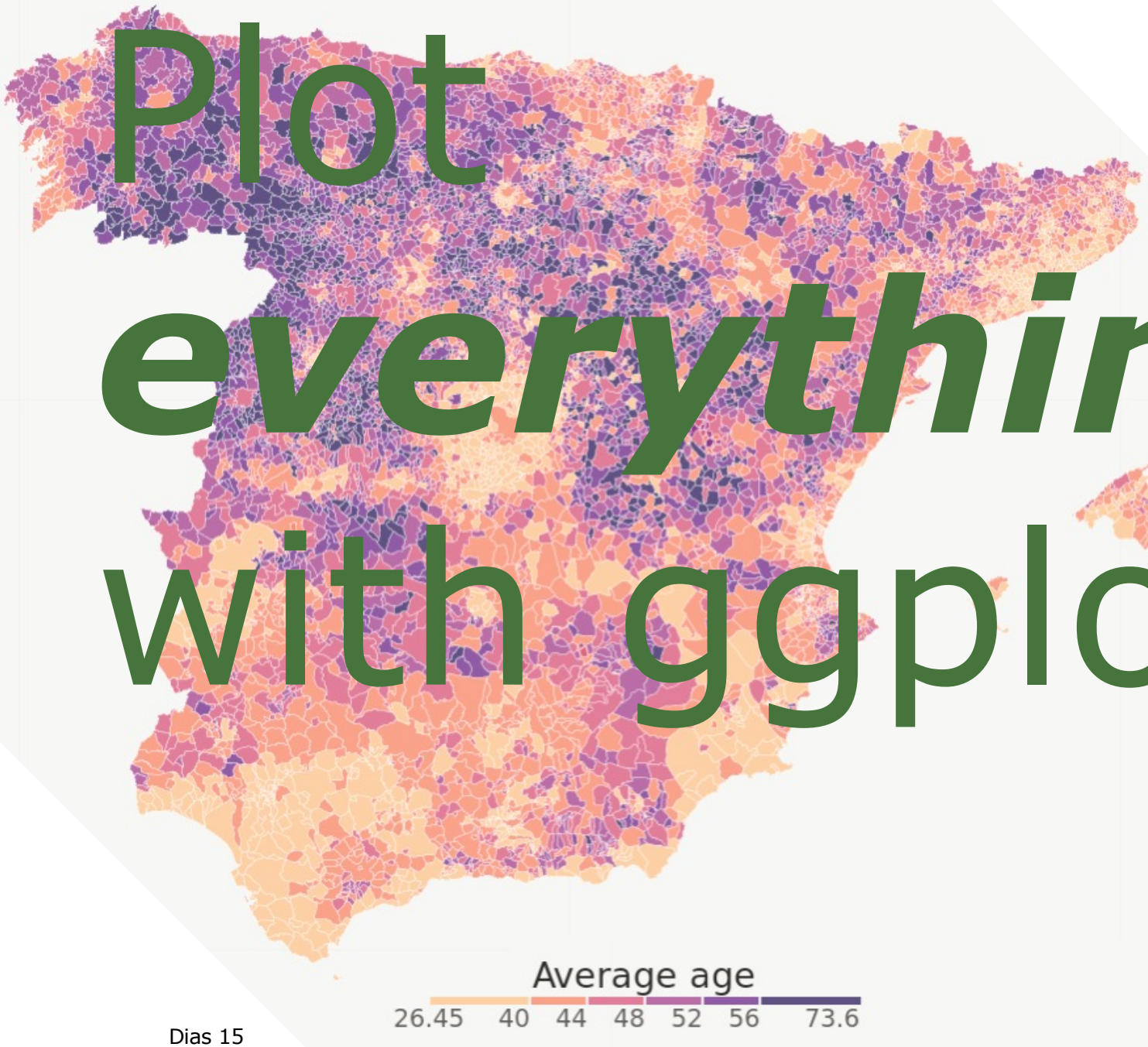
y = a(treatment) + b(time) + k(batch) + e

The workhorse of understanding effects of interventions, handling covariates and confounders, and getting uncertainties on all this

Test statistics

A metric/scalar measure of observed differences, which we can construct a test for

Plot *everything* with ggplot2



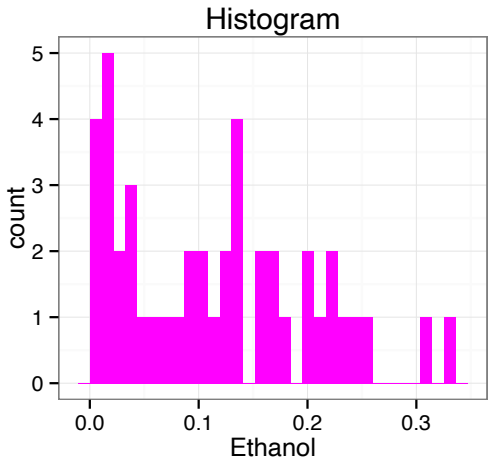
PLOTS

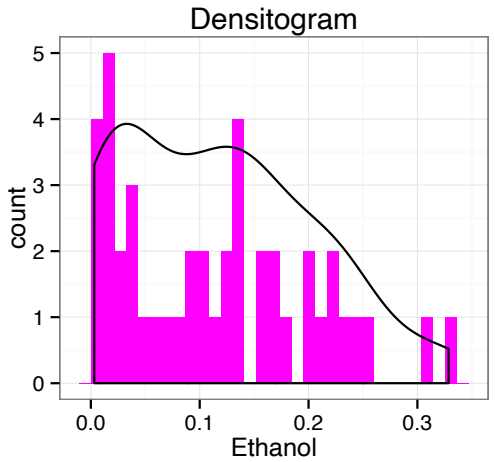
- Histogram, Densitogram, Boxplot, Jitterplot, Violinplot
    - Continuous data
    - Describes distribution
- Lineplot, Stem plot
    - Continuous data
    - Ordinal in e.g. time
- Bar chart, pie chart
    - Categorical data or compositional data
- Scatterplot
    - Bi variate continuous data
- Spline plot
    - Smooth relation between two (or more) variables

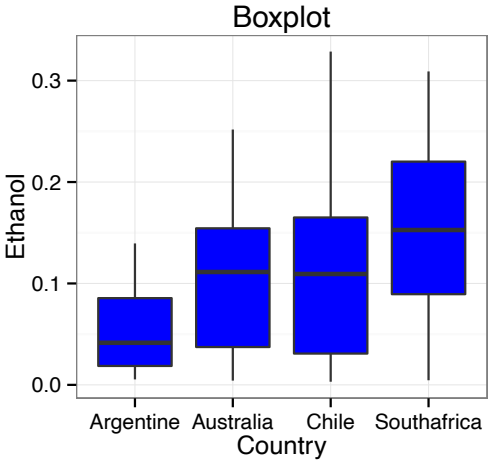
Histogram, Densitogram, Boxplot, Jitterplot, Violin plot,

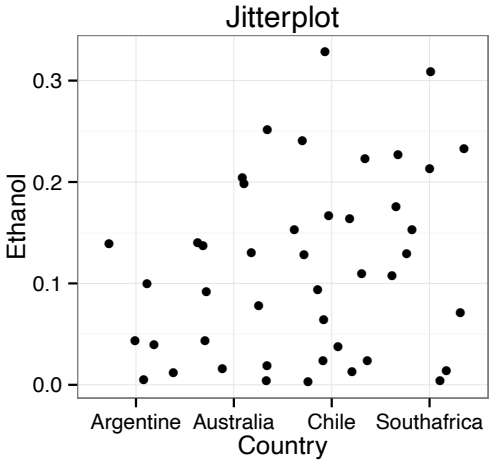
- Jitter shows raw data
- Boxplot and violin can handle many observations Densitogram and Violin may cheat in the representation of raw data

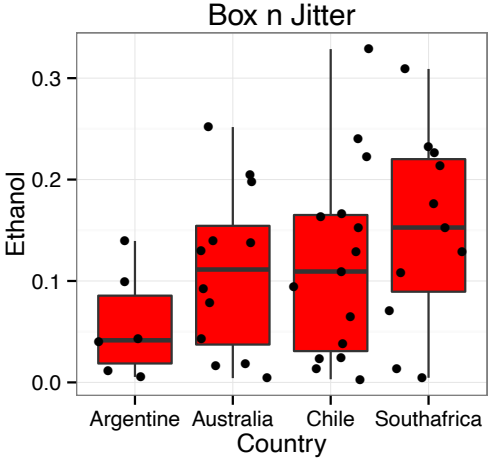
Continuous data
Describes distribution

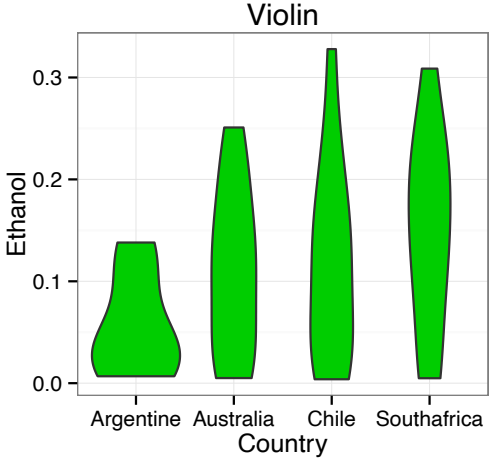






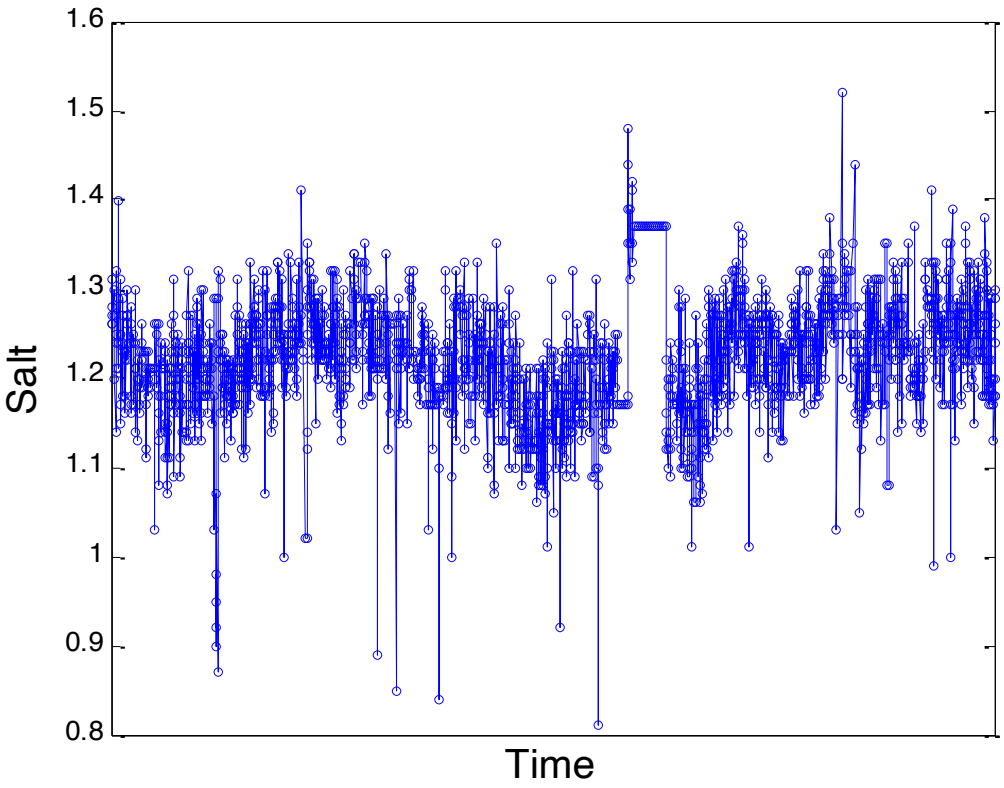






Lineplot

- Is used in e.g. Statistical Process Control (SPC)
- Excellent for capturing drift (systematical change over time) in the laboratory of in the production.

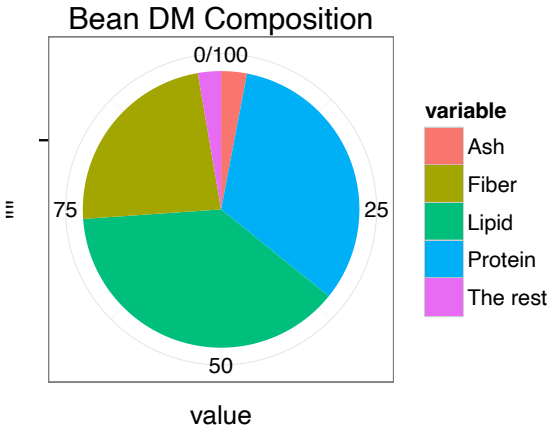


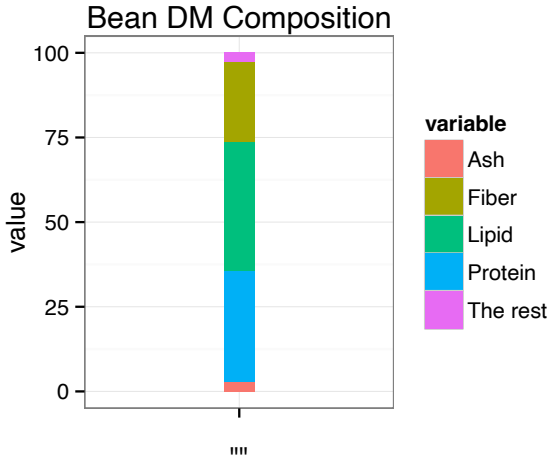
Continuous data
Ordinal in e.g. time

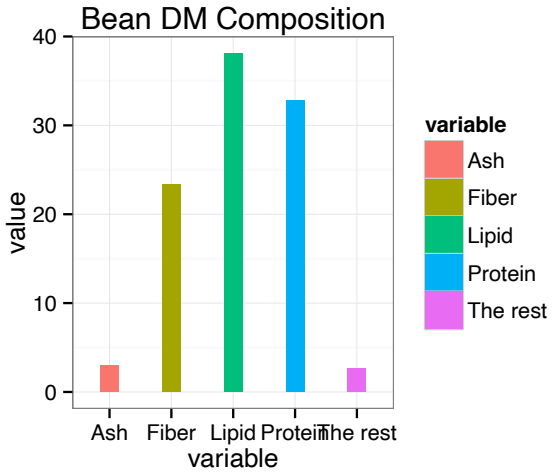
Bar charts
Pie charts

- OBS: Do not replace a boxplot with a bar chart!

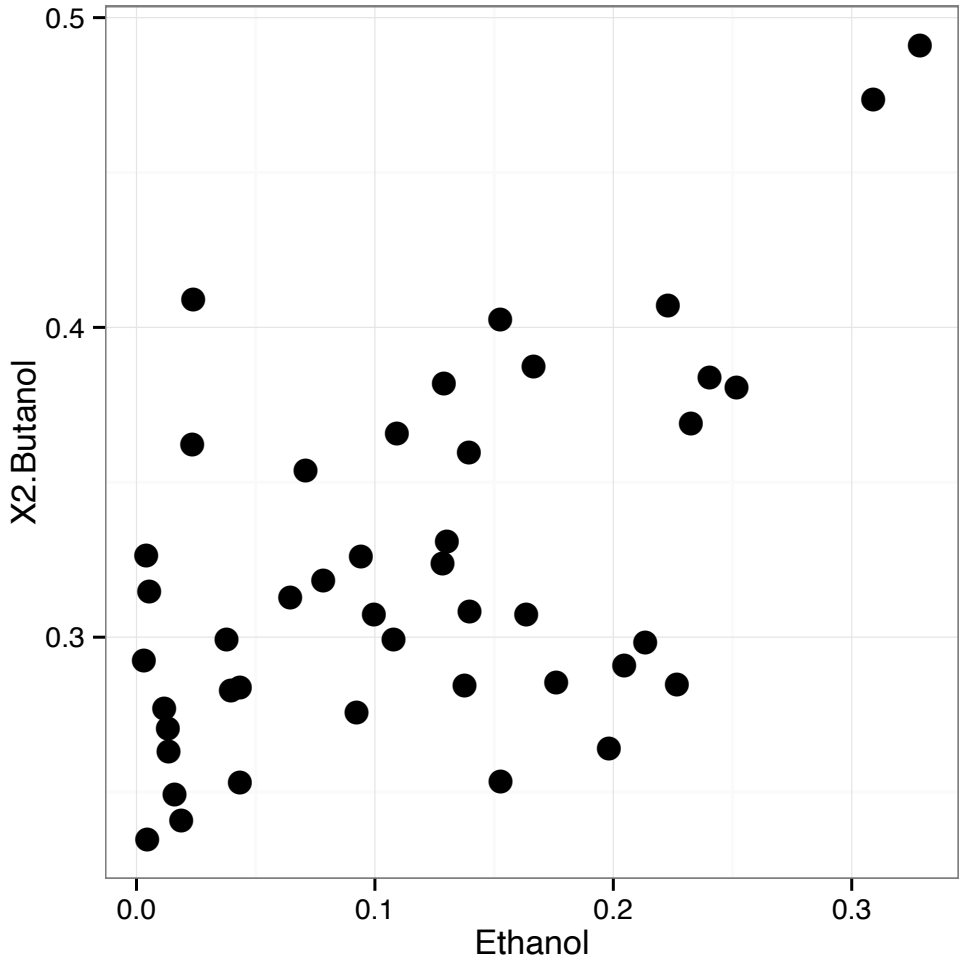
Compositional data
Categorical data





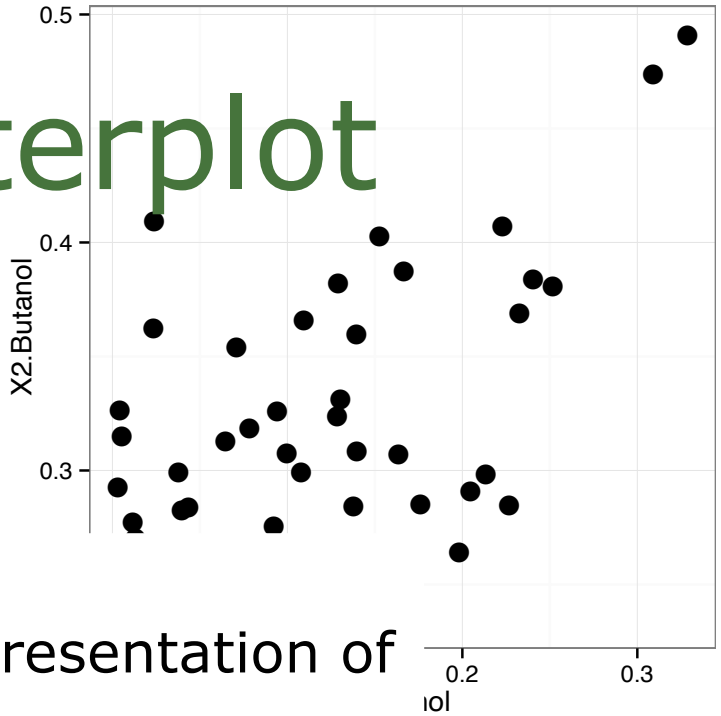


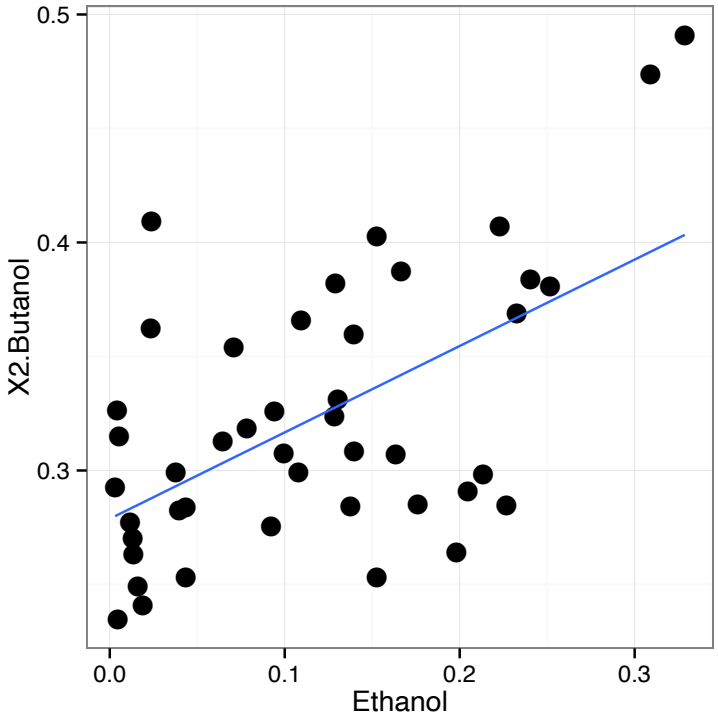
Scatterplot



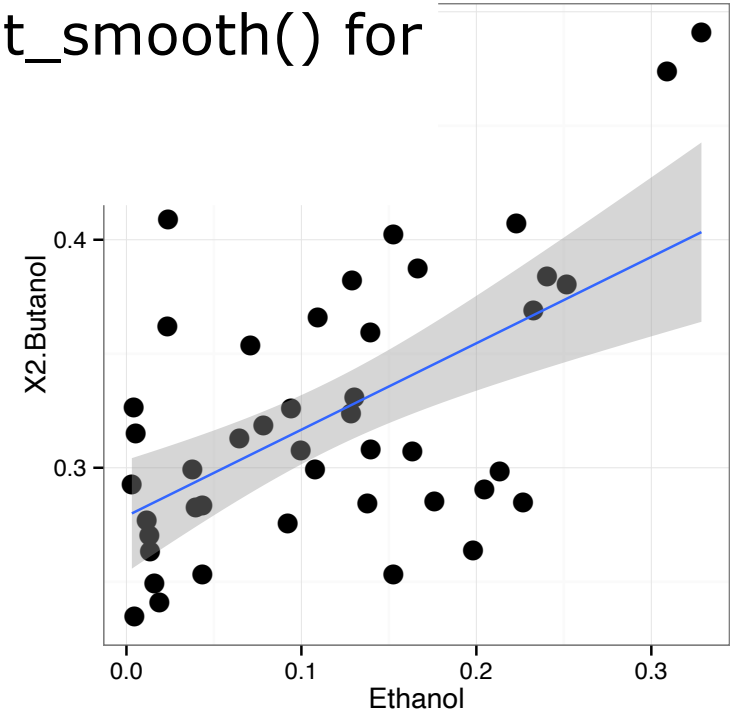
- Two response variables versus each other
- Very usefull for Multivariate data

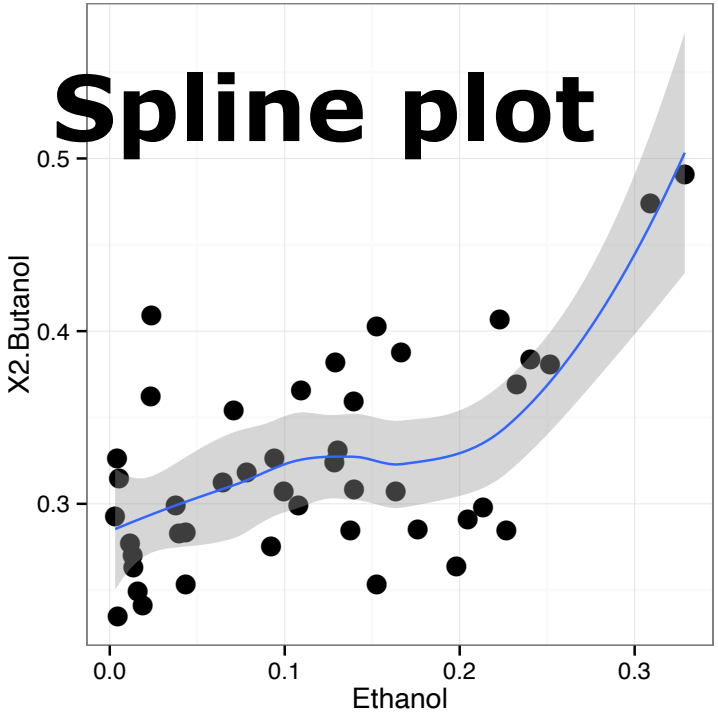
Scatterplot





- Same data!
- Different representation of the relation
- Check: +stat_smooth() for details





Preprocessing

Problem

- Sparse / Zero inflated
- Compositional
- Over dispersed
- High dimensional

- Normalization/rarefaction
- Filter off rare taxa
- Agglomeration
- Transformation (e.g. log())

Diversity metrics

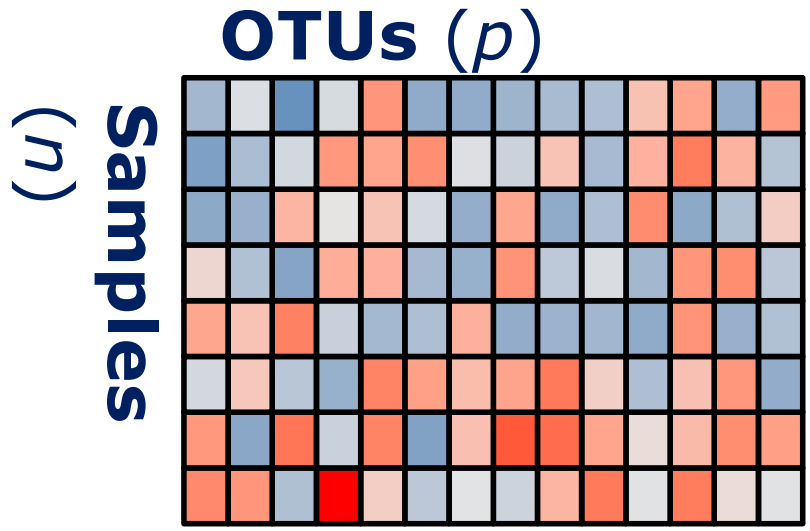
Alpha diversity

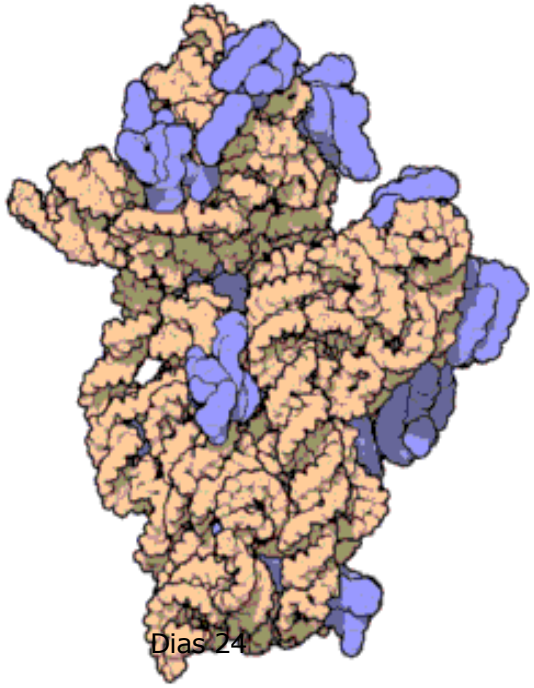
Within sample characteristics

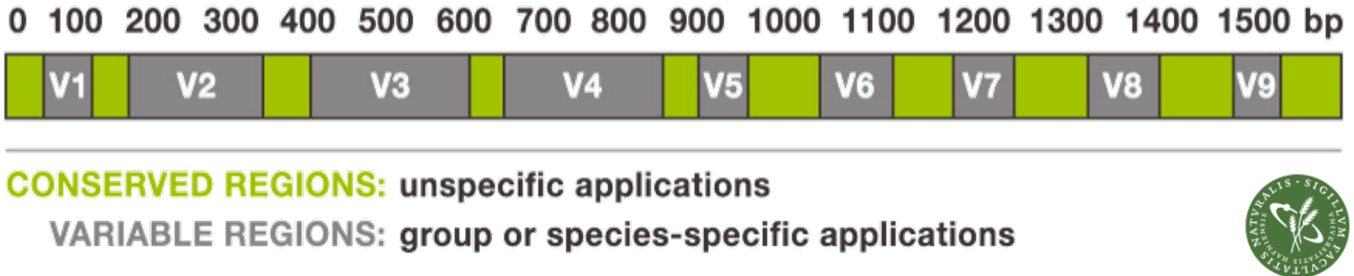
Beta diversity

Between sample characteristics

Amplicon







CONSERVED REGIONS: unspecific applications

VARIABLE REGIONS: group or species-specific applications

Alpha diversity

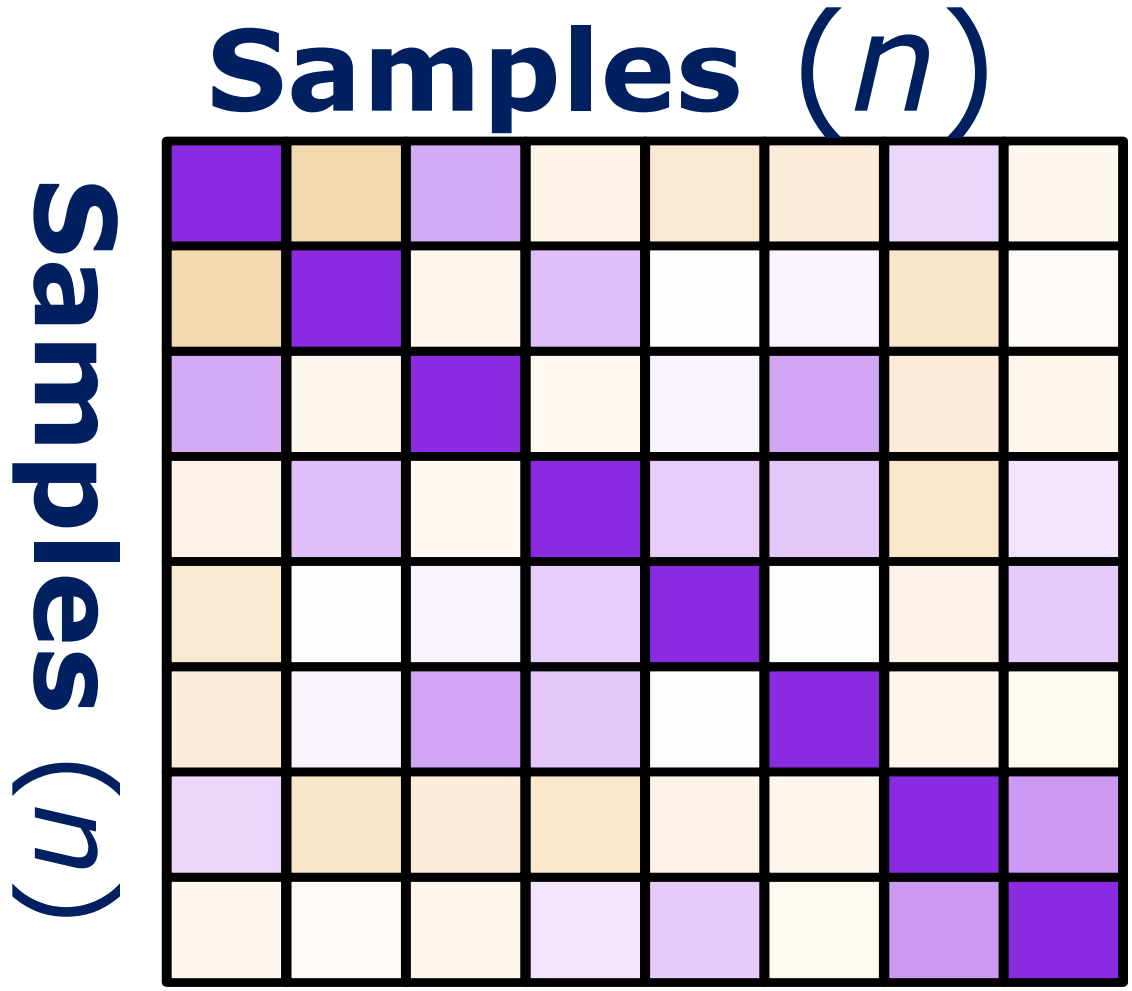
Number of different taxa

Shannon diversity $$H = -\sum_{i=1}^{p} ra_i \cdot ln(ra_i)$$

Simpson $$D = \frac{1}{\sum_{i=1}^{p} ra_i^2}$$

kinda'~ $\mathbf{XX^T}$

β diversity



β diversity

	Presence/absense	Abundance
+Phylo	UNIFRAC PINA	wUNIFRAC wPINA
No-phylo	Jaccard Sørensen …	Bray-Curtis Euclidian Manhatten …

Jaccard

		Sample A	
		No. of species present	No. of species absent
Sample B	No. of species present	a	b
	No. of species absent	c	d

$$S_j = \frac{a}{a+b+c}$$

Bray Curtis

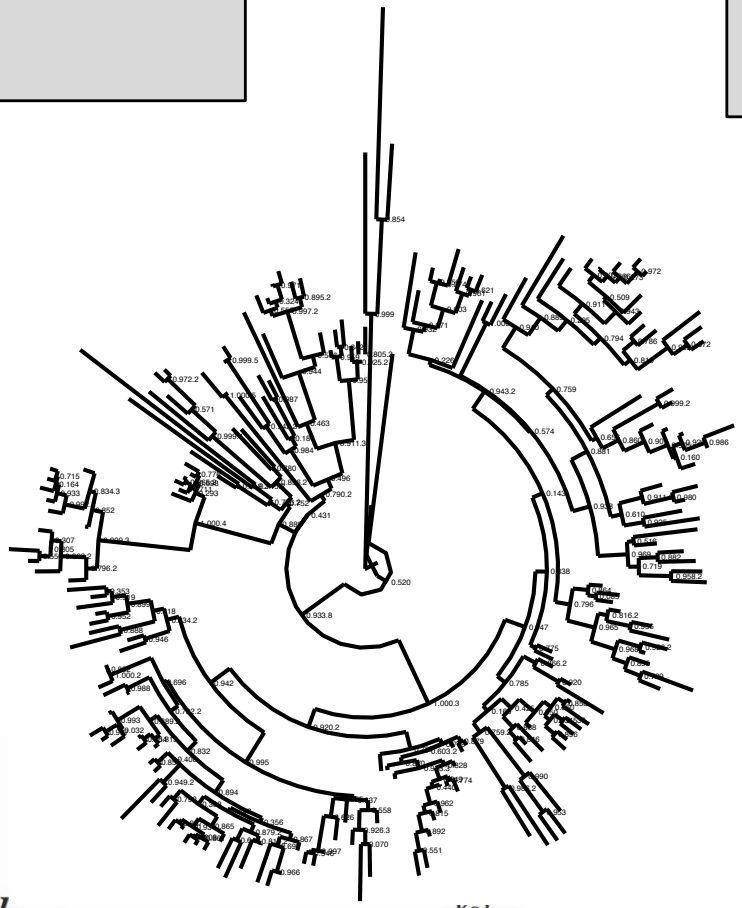
$$BC = \frac{\sum_{i=1}^{n} \left| X_{ij} - X_{ik} \right|}{\sum_{i=1}^{n} \left(X_{ij} + X_{ik} \right)}$$

X_{ij}, X_{ik}	Number of individuals in species i in each sample (j, k)
n	Total number of species in samples.

UNIFRAC

OTU table

Dist



AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 8228–8235
/05/$08.00+0 doi:10.1128/AEM.71.12.8228–8235.2005

UniFrac: a New Phylogenetic Method for Comparing Microbial Communities

Catherine Lozupone[1] and Rob Knight[2]
Department of Molecular, Cellular, and Developmental Biology, Colorado 80309,[1] and Department of Chemistry and Biochemistry, Colorado

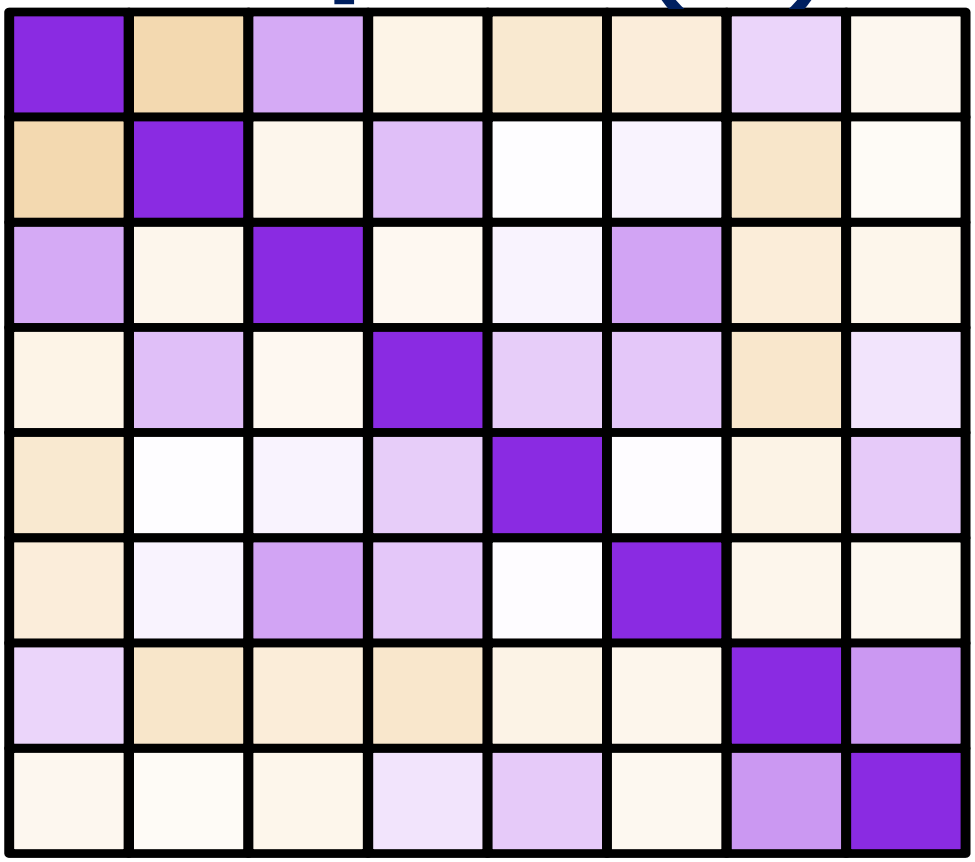
Vol. 71, No. 12

Ordination

β diversity *to* PCoA

Samples (*n*)

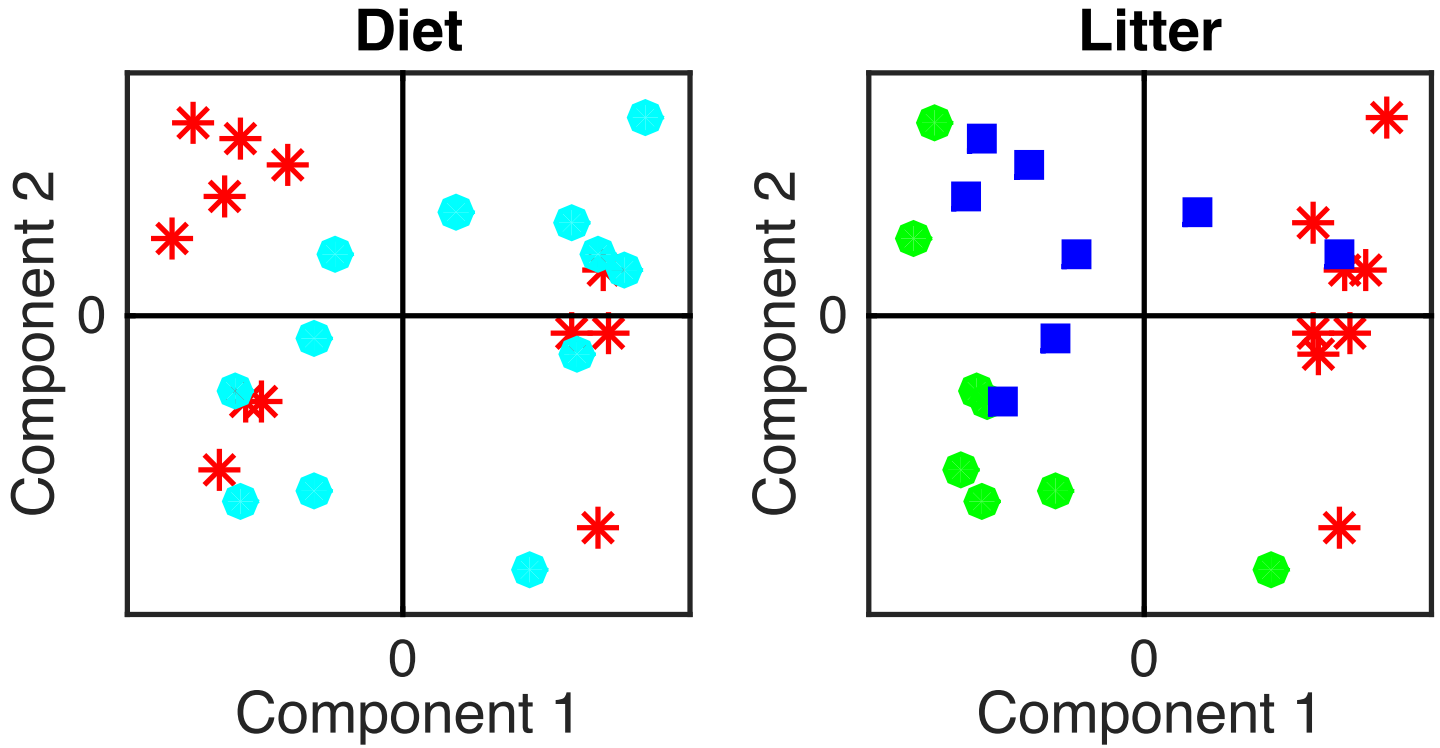
Samples (*n*)



$= U \Lambda U'$

Multidimensional scaling

$$U\Lambda U^T = M\exp(-Dist)M^T$$



Diet	Litter 1	2	3
A	4	4	4
B	4	4	4

$$M = (I - \frac{1}{n}11^T)$$

Testing Beta diversity *PermANOVA*

Partitioning when we do not see the original data

The underlying model

$$Y = XB + E = \hat{Y} + E$$

Variance partitioning

$$tr(Y^T Y) = tr(\hat{Y}^T \hat{Y}) + tr(E^T E)$$

$$= tr(YY^T) = tr(\hat{Y}\hat{Y}^T) + tr(EE^T)$$

$$\hat{Y}\hat{Y}^T = H(YY^T)H$$

$$H = X(X^T X)^{-1} X^T$$

$$YY^T \propto \exp(-Dist)$$

McArdle, B. H., & Anderson, M. J. (2001). Fitting multivariate models to community data: a comment on distance-based redundancy analysis. Ecology, 82(1), 290-297.

Establishing the *null* distribution

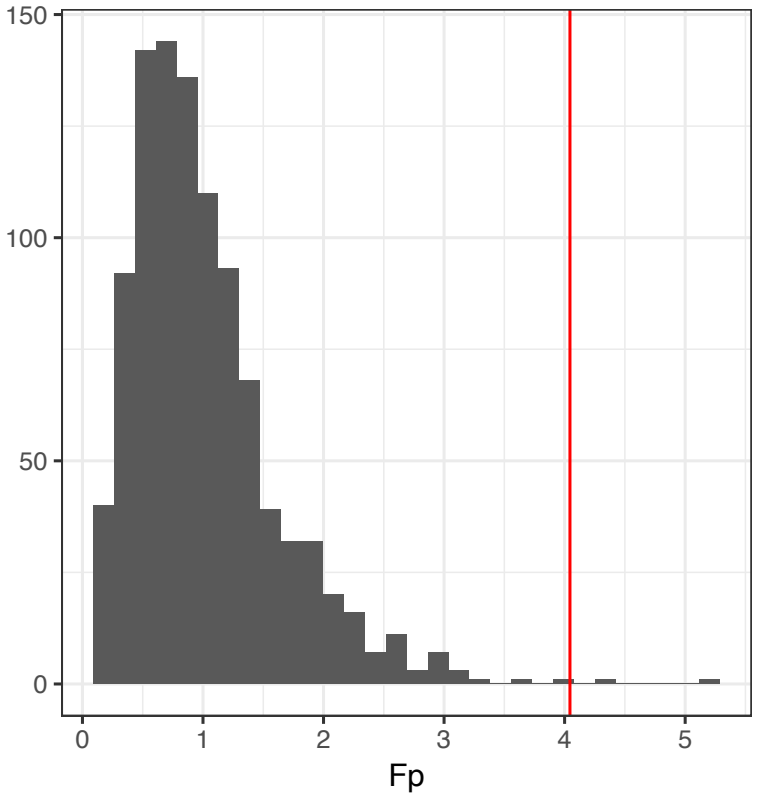
For univariate ANOVA models, we rely on independence and homoschedastic and gaussian *like* residuals.

This make testing easy, as the F-statistic follows the F-distribution under the null hypothesis

For adonis / permanova, we do not know exactly which distribution to use.

What to do?

This one is established by **permuting** the design matrix many times, each time calculating the F-statistics.

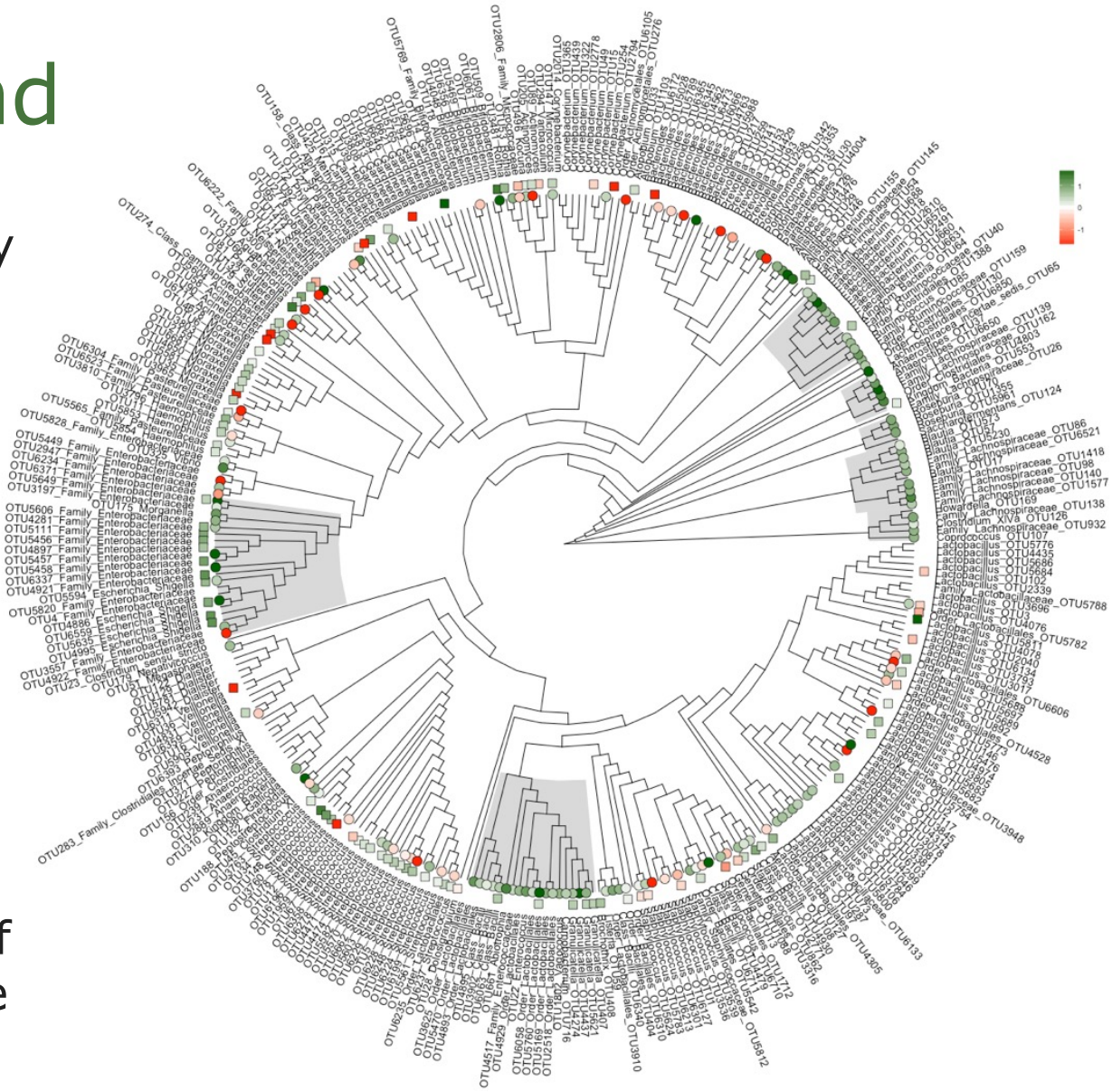


Flip it around

Just as the beta diversity reflects sample similarity.

The phylogenetic tree reflects evolutionary similarity.

We can use this in a similar fashion with adonis to answer questions related to the dependency on the phylogenetic structure of some relevant univariate results



ggtree

Differential Abundance Testing

or

OTUWAS

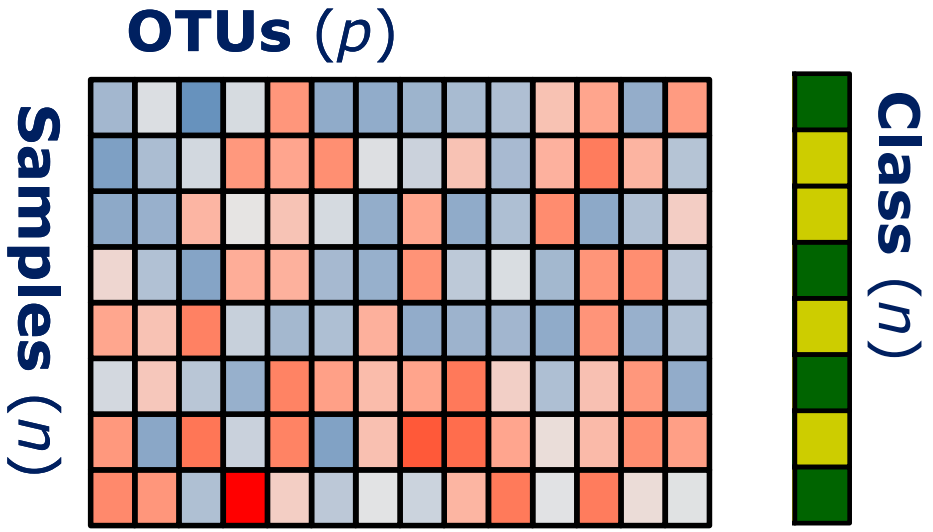
Idea

1. Perform *p* univariate tests recording an inferential statistics (e.g. the p-value)

2. Arrange the *p* (OTUs) from the most different wrt classes to the least different

$$pv_1 < pv_2 < \ldots < pv_p$$

3. Figure out a thresshold to separate the p OTUs into discoveries and non-discoveries.



What to consider?

Choose a powerful statistical method

That is: avoid methods which are wrong in distributional assumptions.

Go parametric if you can!

Utilize actively the multiple estimation to *robustify* the individual estimates.

That is: in stead of using maximum likelihood for each of the *p* variables, shrink these towards a common value.

Problem

- Sparse / Zero inflated
- Compositional
- Over dispersed
- High dimensional

DESeq2

Developed for RNAseq

Based on log2 fold changes between (two) groups

Zeros are handled by regularized logarithm (shrinkage of low abundance towards common value)

Uses **empirical bayes** on

- Dispersion parameter to shrink towards theoretical distribution
- Fold Change (central parameter) to shrink towards zero

MetagenomeSeq

Handles the zero inflation explicitly by a mixture model of

1)The zeros and (fitted across OTUs)

2)The biological model (fitted for each OTU)

Uses **empirical bayes** on central- and dispersion parameters to shrink towards common value

(uses cumulative sum scaling for prepro)

$$f_{zig}(y_{ij}; s_j, \beta, \mu_i, \sigma_i^2) = \pi_j(s_j) \cdot I_{\{0\}}(y_{ij}) + (1 - \pi_j(s_j)) \cdot f_{count}(y_{ij}; \mu_i, \sigma_i^2)$$

Controlling Type I error

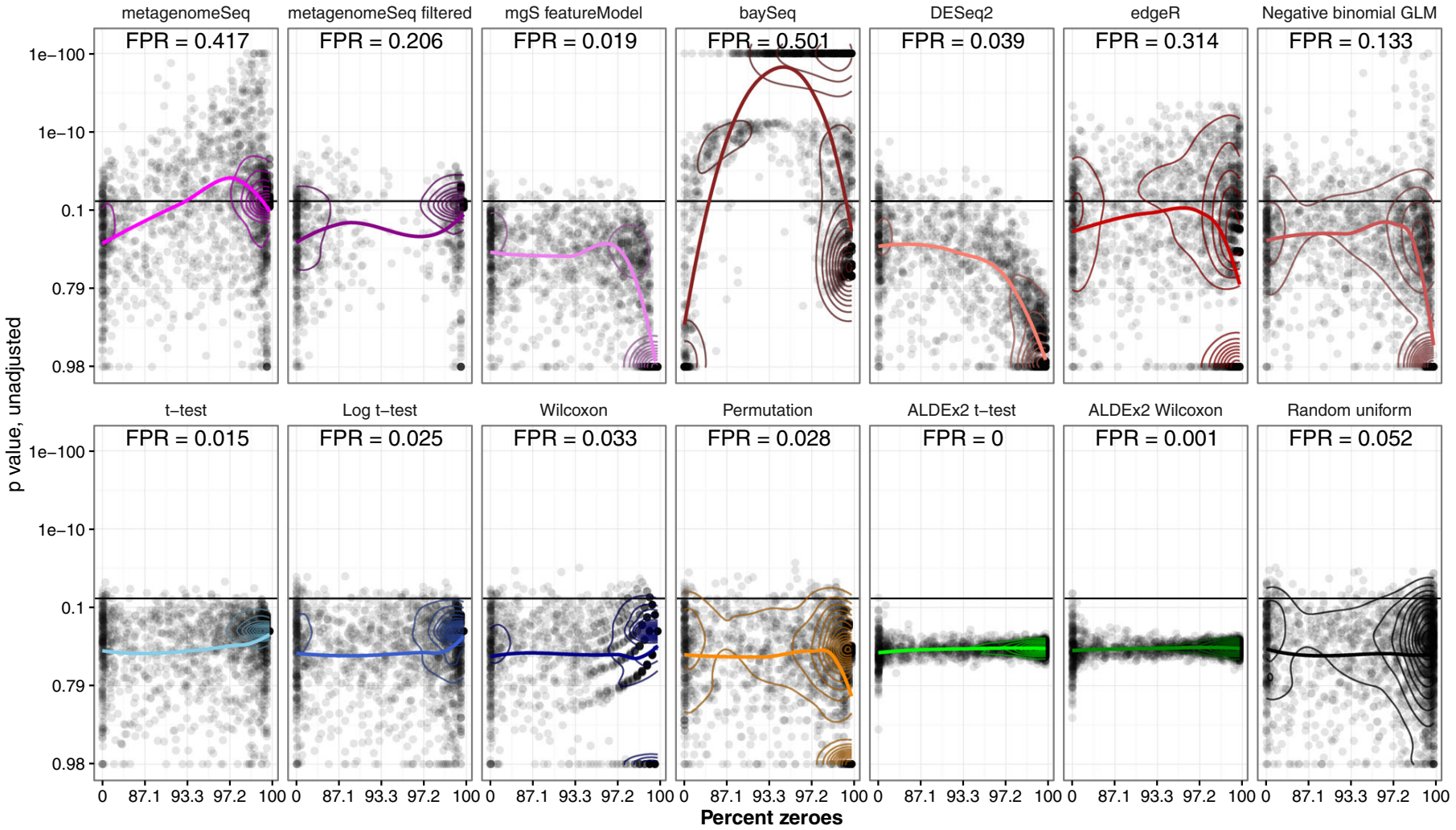


Fig. 2 OTU sparsity vs. *p* value. Scatterplots of OTU sparsity vs *p* value with panels representing each differential relative abundance test method in feces dataset A1, with 50% cases. *Colored line* represents the LOESS regression on data. False positive rate (FPR) is defined as the fraction of OTUs with $p < 0.05$. Each differential relative abundance test represents the median FPR for that method, out of all 150 permutations. *Contour lines* indicate point density and can be compared to a hypothetical null distribution of *p* values demonstrated in the final panel ("Random uniform")

Correct ordering

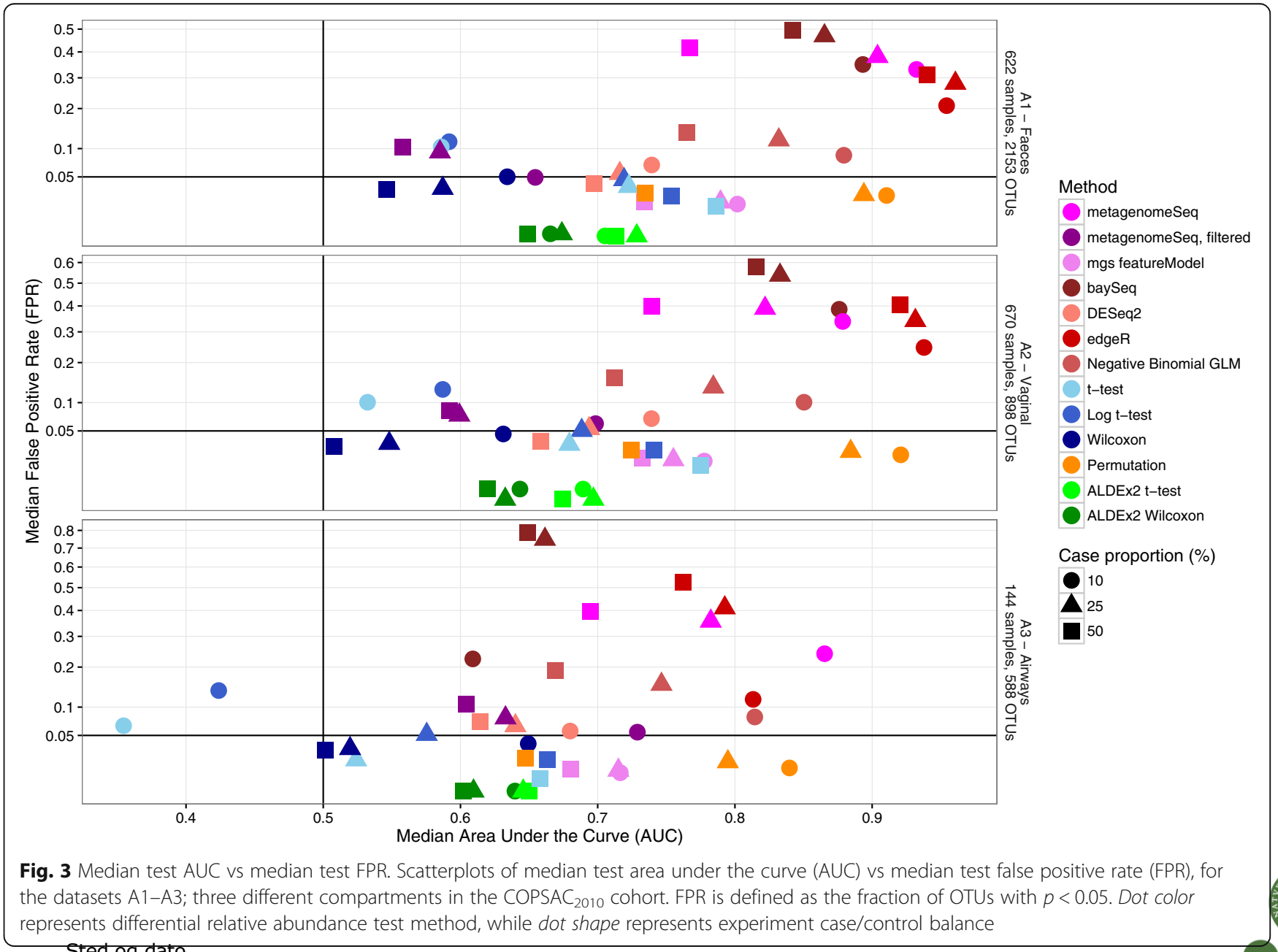


Fig. 3 Median test AUC vs median test FPR. Scatterplots of median test area under the curve (AUC) vs median test false positive rate (FPR), for the datasets A1–A3; three different compartments in the $COPSAC_{2010}$ cohort. FPR is defined as the fraction of OTUs with $p < 0.05$. *Dot color* represents differential relative abundance test method, while *dot shape* represents experiment case/control balance

Where to cut?

Bonferoni's Family Wise Error Rate (FWER)

Any **pv < alpha / p** is a discovery

False Discovery Rate (FDR) control

Any $\mathbf{pv_k < k * alpha / p}$ is a discovery
Where k is the order.

These are under independence assumptions… Which is almost never fulfilled

Alternatives: *permutation testing*

MultiOmics

Data integration

-Omics

OTU table

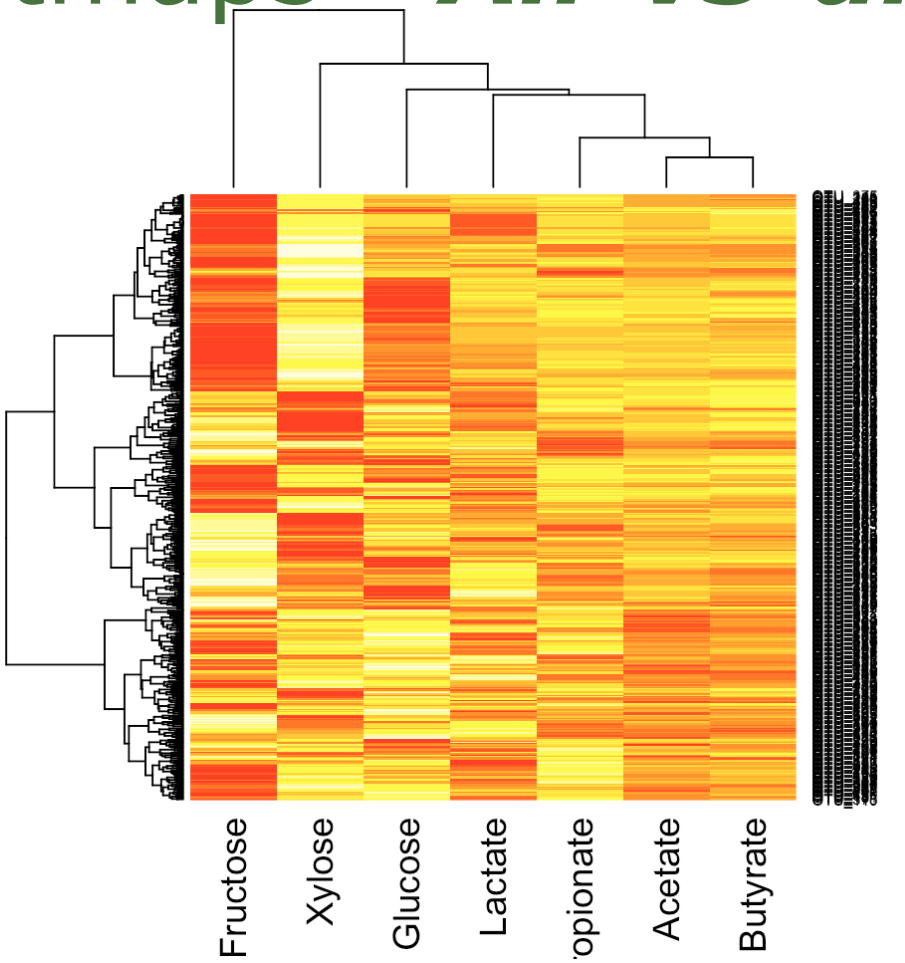
$$_{n}\mathbf{X}^{p_X}$$

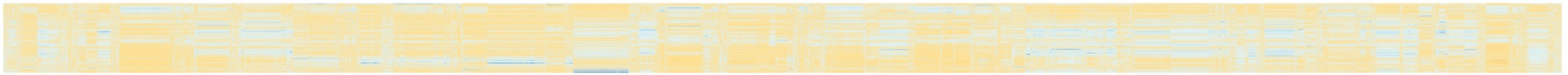
$$_{n}\mathbf{Y}^{p_Y}$$

Descriptive Heatmaps - ***All-vs-all***

Easy and intuitive representation for multi-omics data

munivariate as it is basically *univariate* correlations interpreted multivariate visually





Corr matrix of 1500 by 79000

Data integration

-Omics

OTU table

$$_{n}\mathbf{X}^{p_X}$$

$$_{n}\mathbf{Y}^{p_Y}$$

Functionally Supervised

Normal CCA

$$\max_{\mathbf{w_X},\mathbf{w_Y}} \mathbf{w_X}^T \mathbf{X}^T \mathbf{Y} \mathbf{w_Y} \quad \textbf{s.t.} \quad \begin{matrix} \mathbf{w_X}^T \mathbf{X}^T \mathbf{X} \mathbf{w_X} = 1 \\ \mathbf{w_Y}^T \mathbf{Y}^T \mathbf{Y} \mathbf{w_Y} = 1 \end{matrix}$$

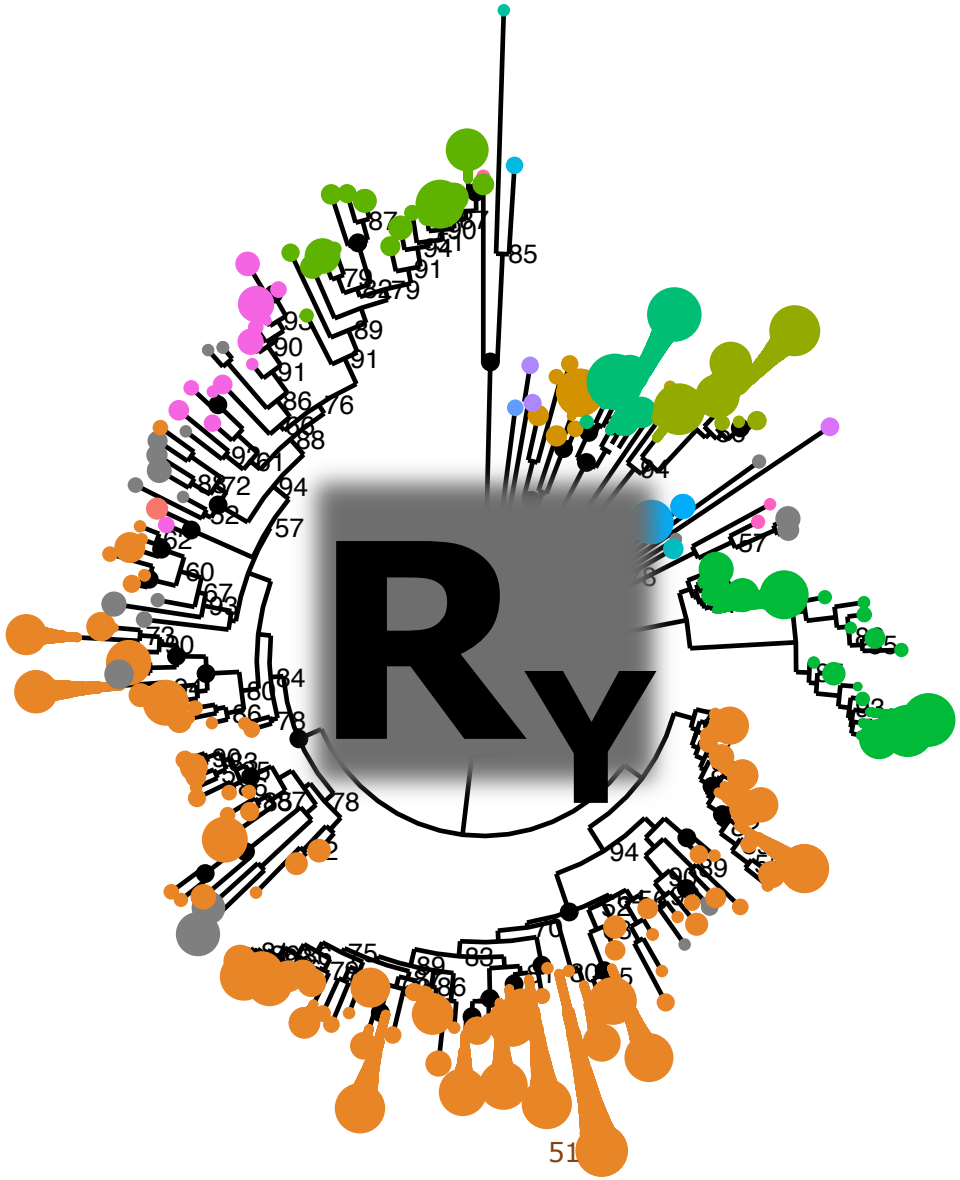
Supervised CCA

- Let $\mathbf{R_Y}$ be a symmetric matrix reflecting similarity between the features in $\mathbf{Y}$ based on **external knowledge**

Functionally Supervised

Supervised CCA

- Let $\mathbf{R_Y}$ be a symmetric matrix reflecting similarity between the features in **Y** based on **external knowledge**



Functionally Supervised

Supervised CCA

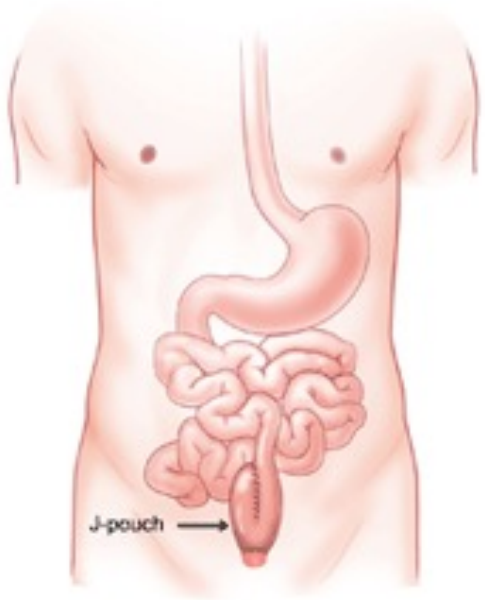
Re-formalize the main objective

Kernel Smoothing

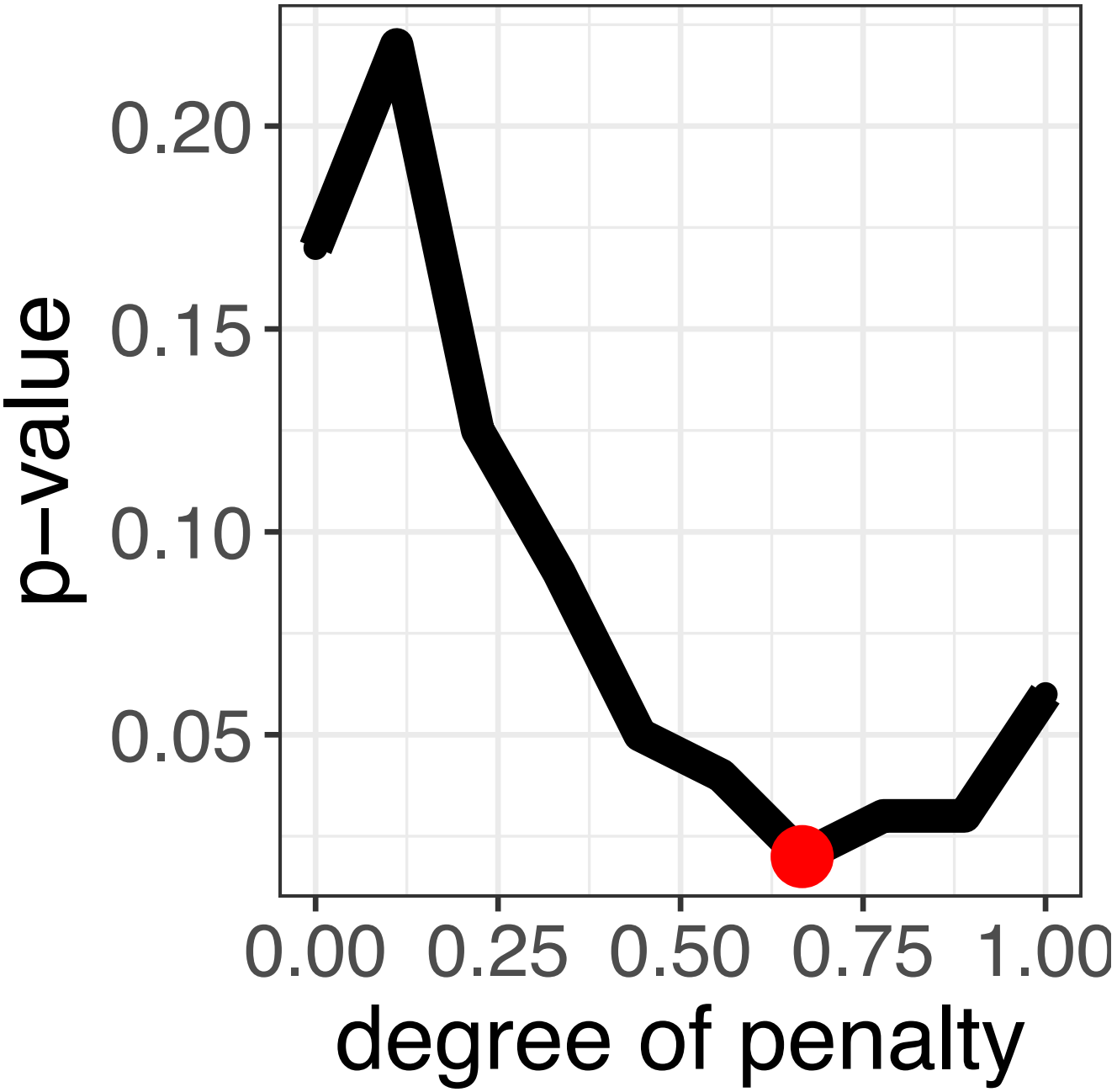
$$\max_{\mathbf{w_X},\mathbf{w_Y}} \mathbf{w_X}^T \mathbf{X}^T \mathbf{Y} \mathbf{R_Y} \mathbf{w_Y}$$

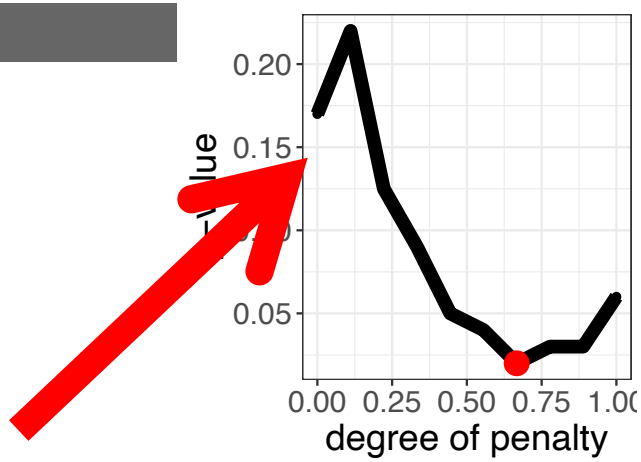
$$\mathbf{R_Y} = \alpha \mathbf{S} + (1 - \alpha)\mathbf{I}$$

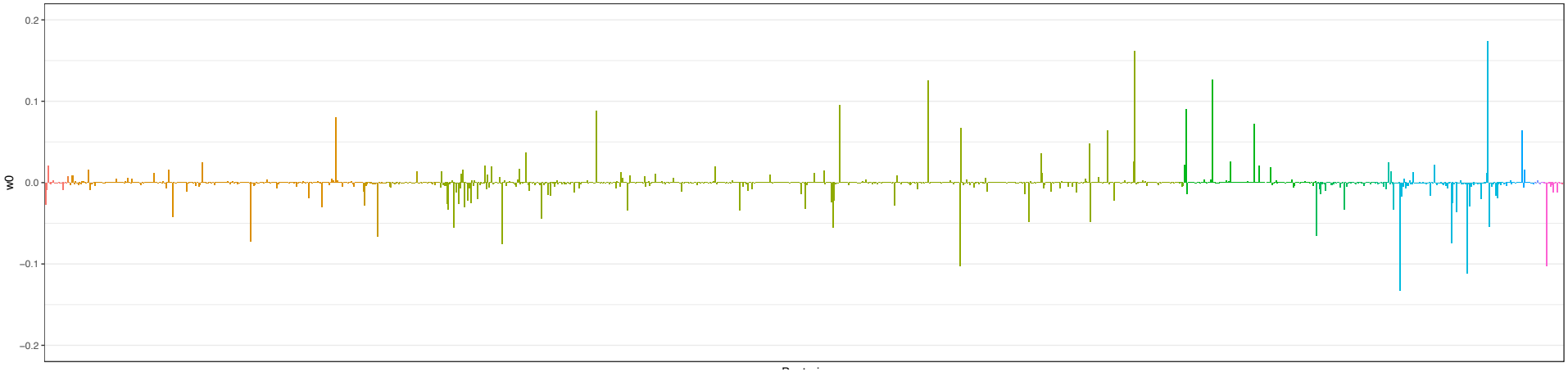
Example Pouchitic cohort



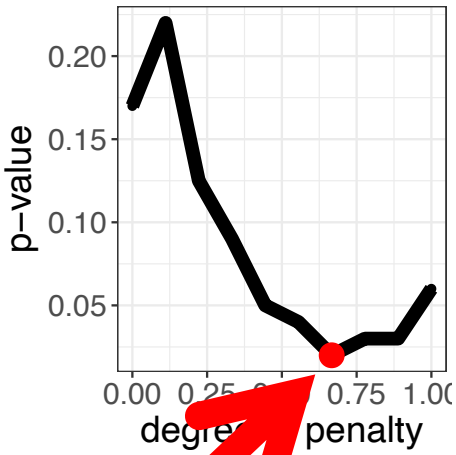
Gene Expression	OTU table
248 X 1991	248 Y 1380

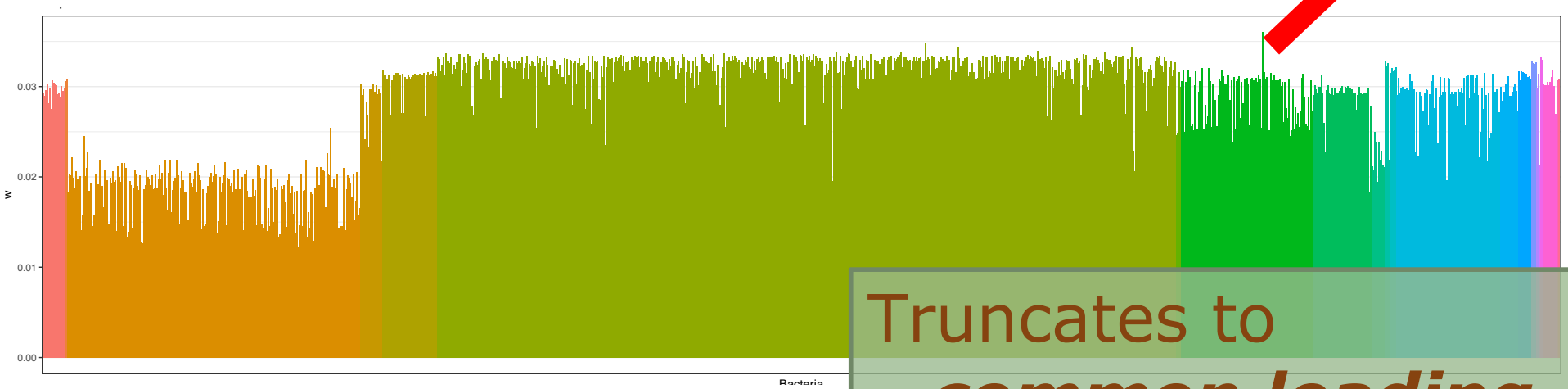

p–value
0.20
0.15
0.10
0.05
0.00
0.25
0.50
0.75
1.00
degree of penalty





Order
Actinomycetales
Bacillales
Bacteroidales
Bifidobacteriales
Burkholderiales
Clostridiales
Coriobacteriales
Enterobacteriales
Erysipelotrichales
Flavobacteriales
Fusobacteriales
Gemellales
Lactobacillales
Pasteurellales
Pseudomonadales
Rhizobiales
Rhodocyclales
Sphingobacteriales
Sphingomonadales
Turicibacterales
Verrucomicrobiales
Xanthomonadales

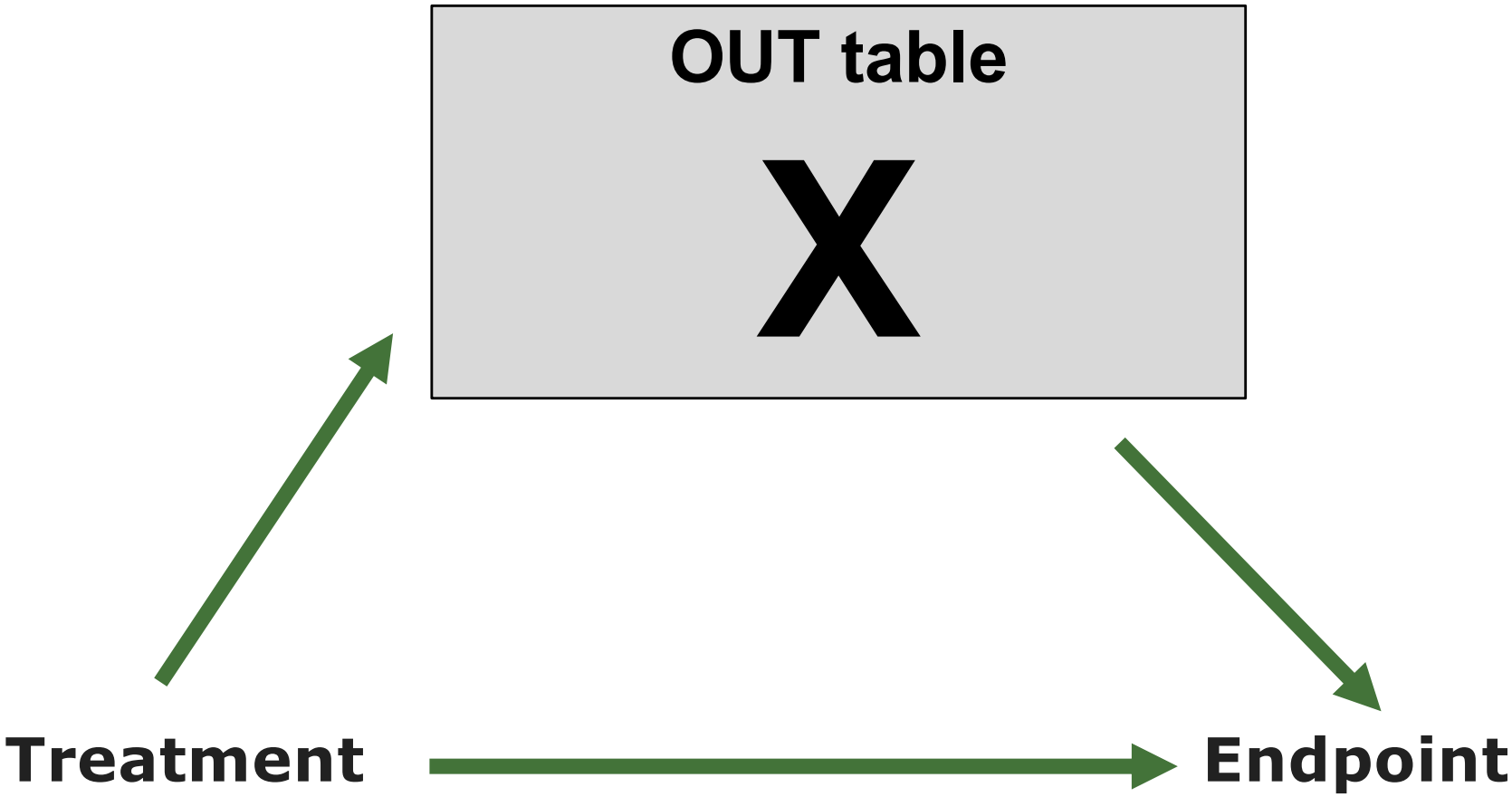




Truncates to
common loading
for similar bacteria

Order
Actinomycetales
Bacillales
Bacteroidales
Bifidobacteriales
Burkholderiales
Clostridiales
Coriobacteriales
Enterobacteriales
Erysipelotrichales
Flavobacteriales
Fusobacteriales
Gemellales
Lactobacillales
Pasteurellales
Pseudomonadales
Rhizobiales
Rhodocyclales
Sphingobacteriales
Sphingomonadales
Turicibacterales
Verrucomicrobiales
Xanthomonadales

Mediation analysis



Exposure Propensities

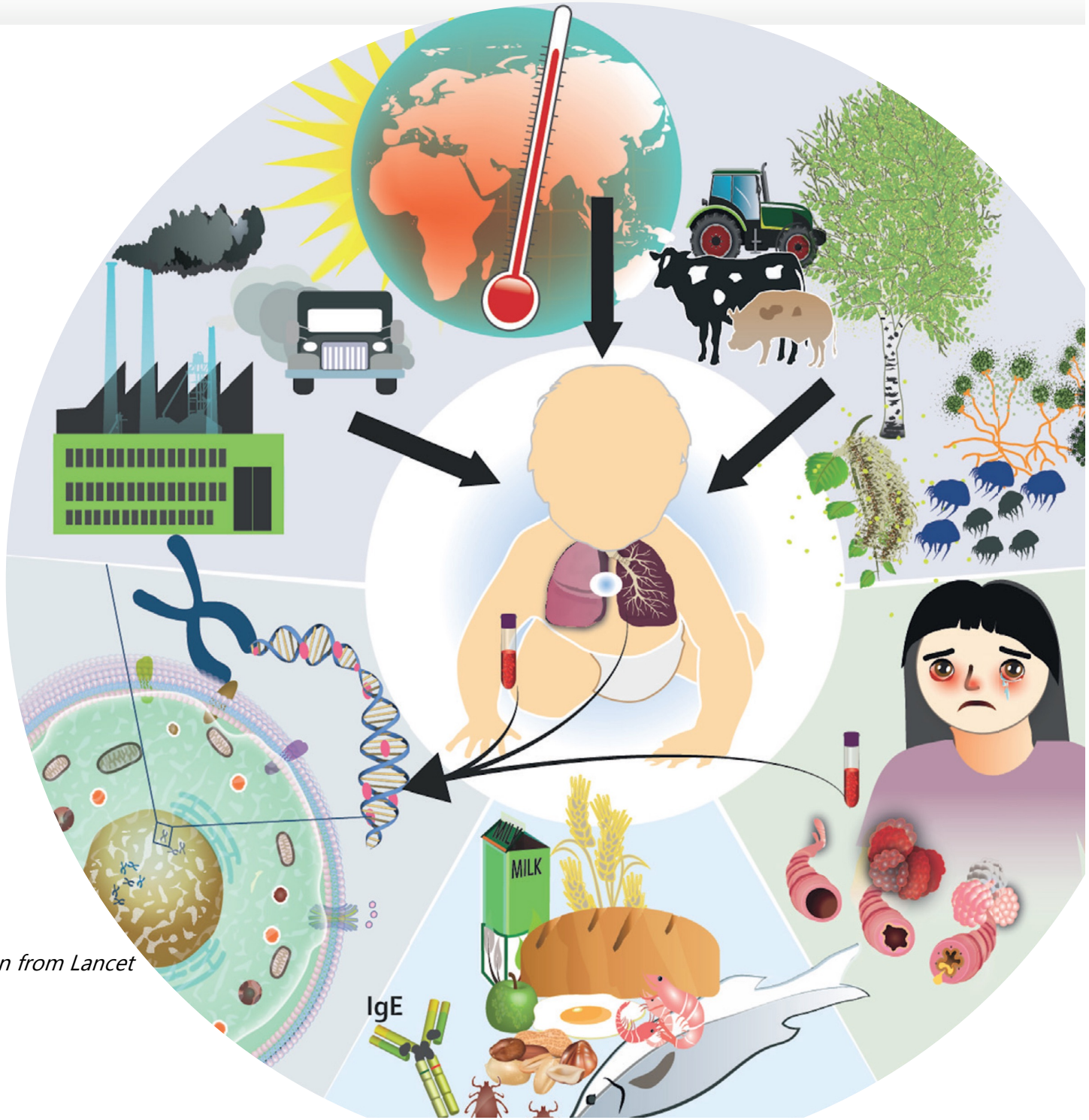
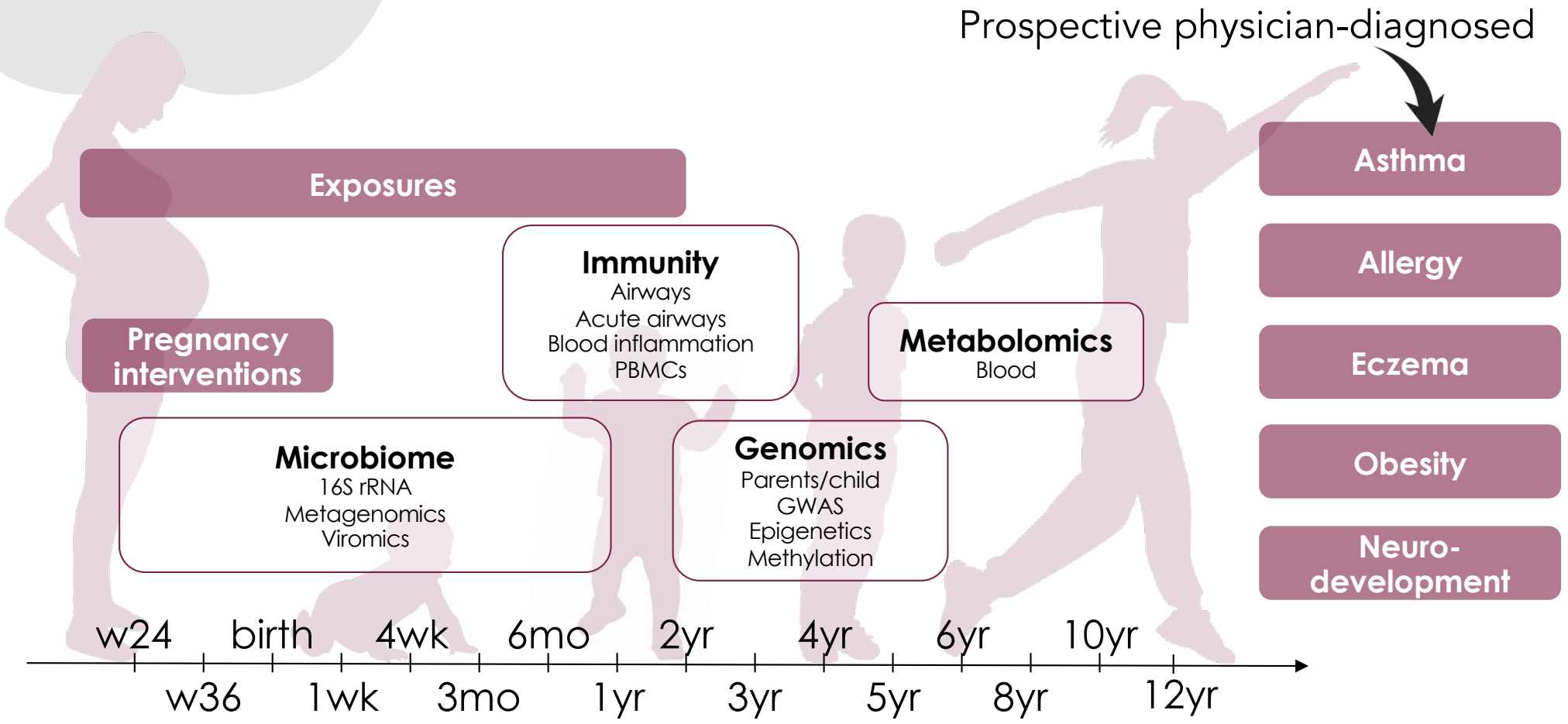
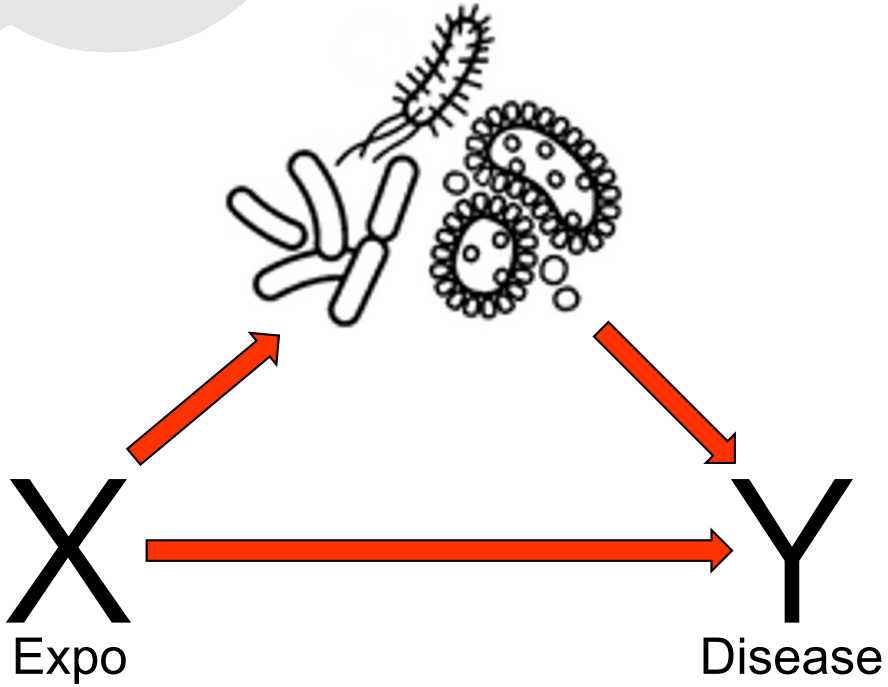


Image stolen from Lancet

COPSAC$_{2010}$ cohort design



Triangulation



- Uncovers mediation
- Comes with the benefit of statistically controlling the *noisy* omics layer

Stokholm et al, Sci Trans Med 2020

A common question structure

- Exposure -> Omics -> Disease

- Siblings -> Microbiome -> Allergy
- Rural living -> Blood Metabolome -> Infections
- C-section -> Microbiome -> Asthma
- Diet -> Blood Metabolome -> Neuro

Establishing an:
Omics-derived Exposure Propensity Score
Fingerprint

- Build
  betw
  expo
- Use C
  mode
- Use th
  predi

Example
Cesarean section score

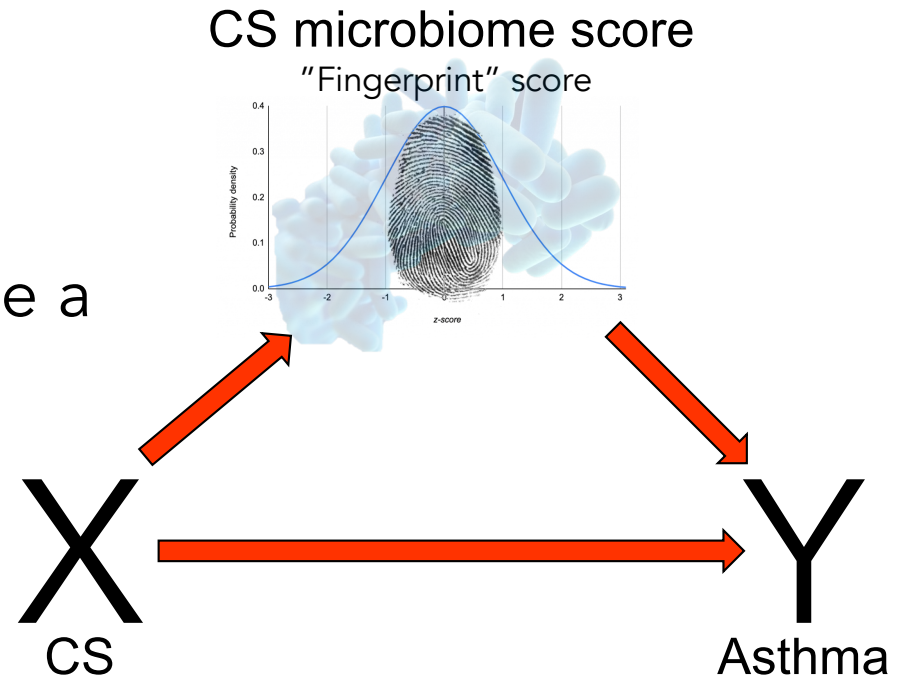
High values = your microbiome looks like you were born by cesarean section

Low values = your microbiome looks like you were born vaginally

Fingerprint mediation analysis

- 1) Establish clinical association
- 2) Derive fingerprint
- 3) Examine mediation

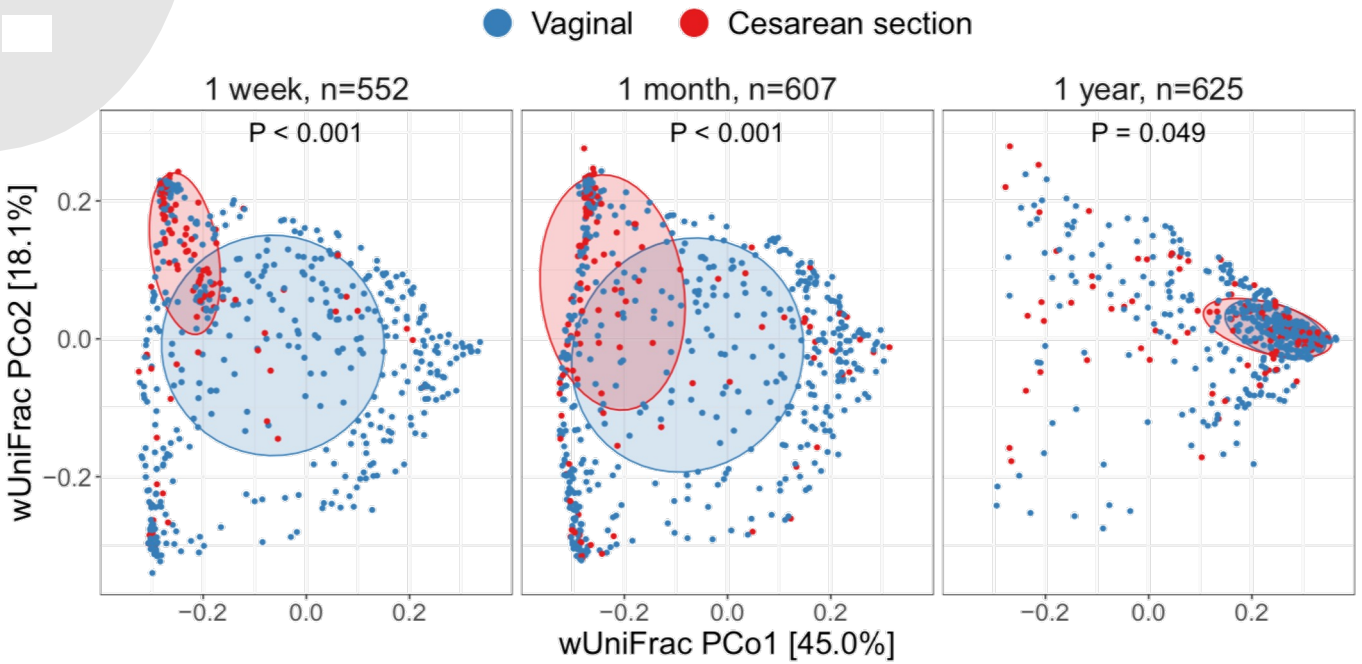
- Conclusion: Is the microbiome a mediator of the association between X and Y?



Stokholm et al, Sci Trans Med 2020

Cesarean section and gut microbiome

COPSAC$_{2010}$ cohort

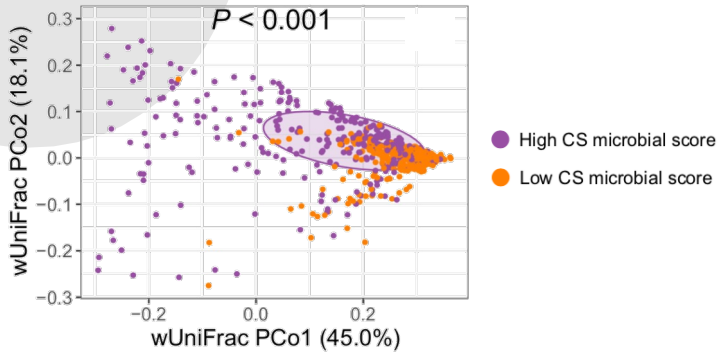


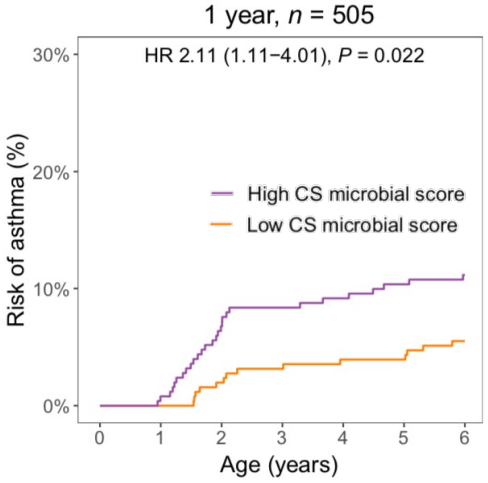
Is the CS-asthma association mediated by differences in the gut microbiome?

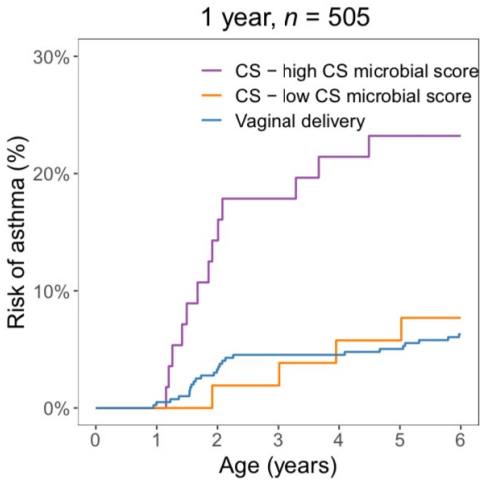
Stokholm et al, *Science Trans. Med.* 2020

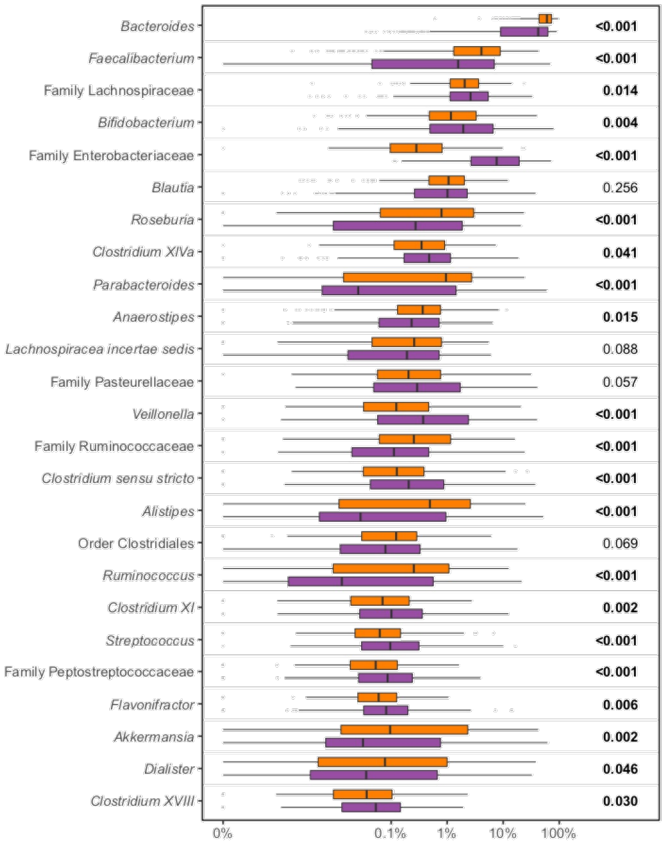
Cesarean section and gut microbiome

COPSAC$_{2010}$ cohort







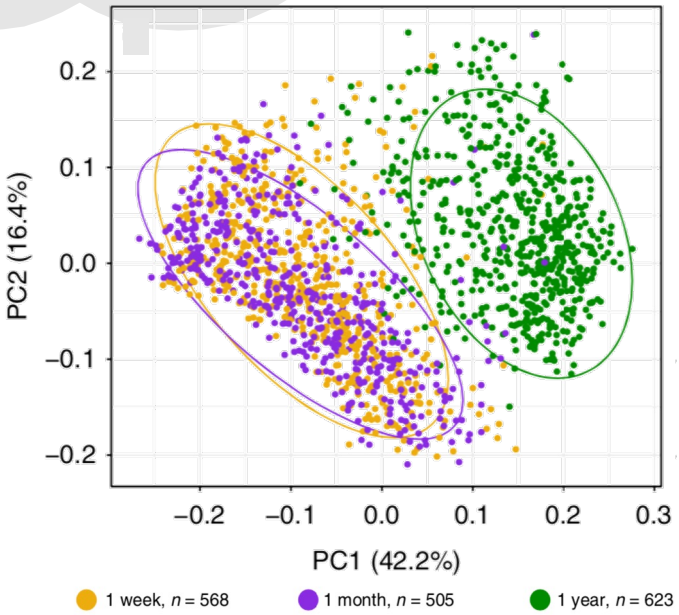


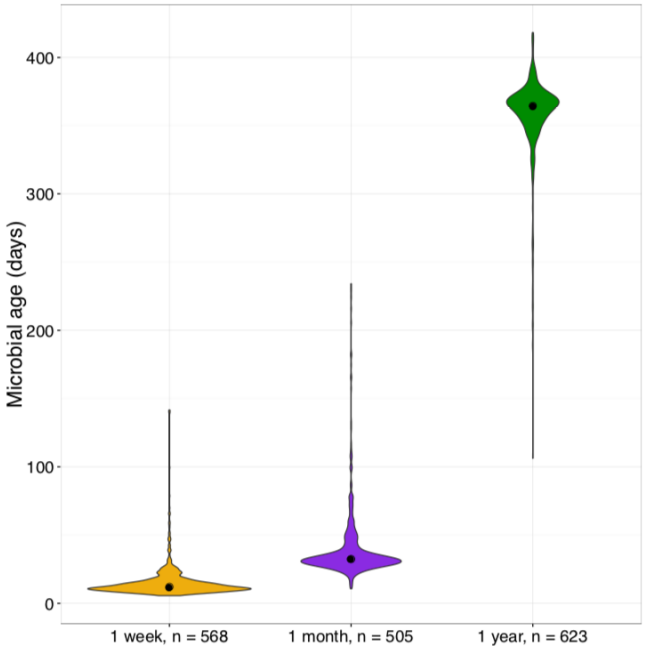
Stokholm et al, *Science Trans. Med.* 2020

Gut microbiome maturation

- $COPSAC_{2010}$ cohort, N=700

$COPSAC_{2010}$ cohort





Microbiome maturation
MAZ – Microbiota-by-Age Z-scores
Random Forest model

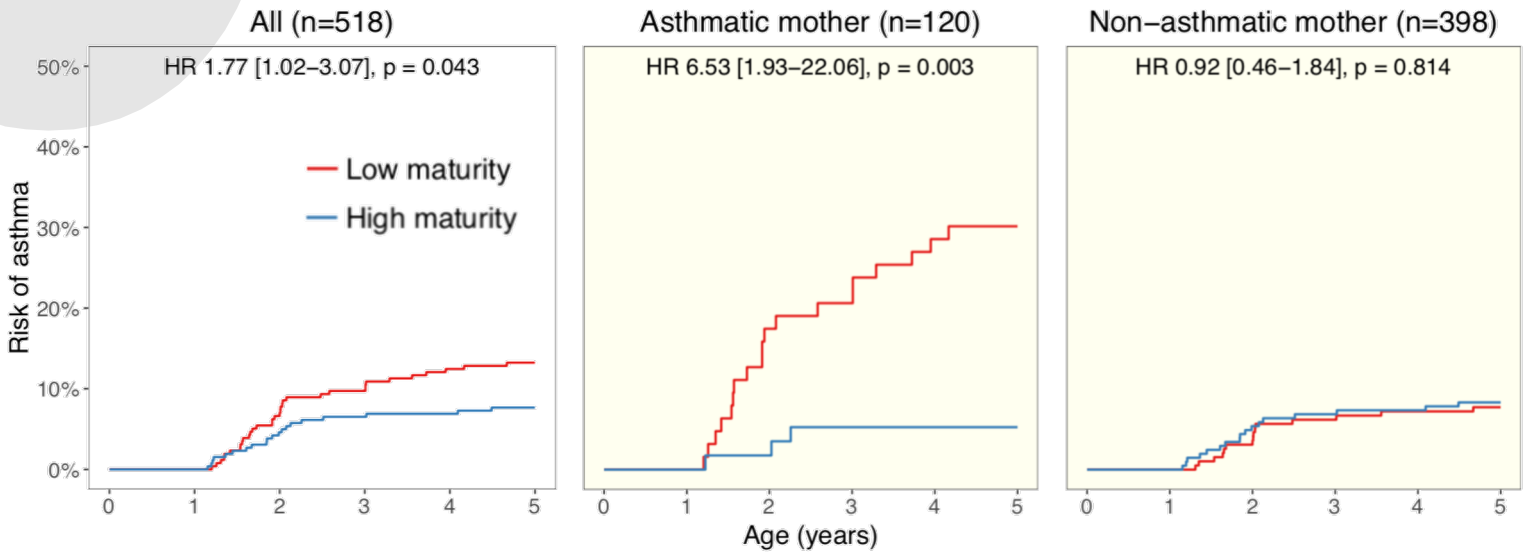
Modified by

- Delivery mode
- Breastfeeding
- Antibiotics
- Siblings
- Pets

Stokholm et al, *Nat Comm* 2018; Subramanian et al, *Nature* 2014;
Stewart et al, *Nature* 2018; Bokulich et al, *Sci Trans Med* 2016

Maturation and asthma

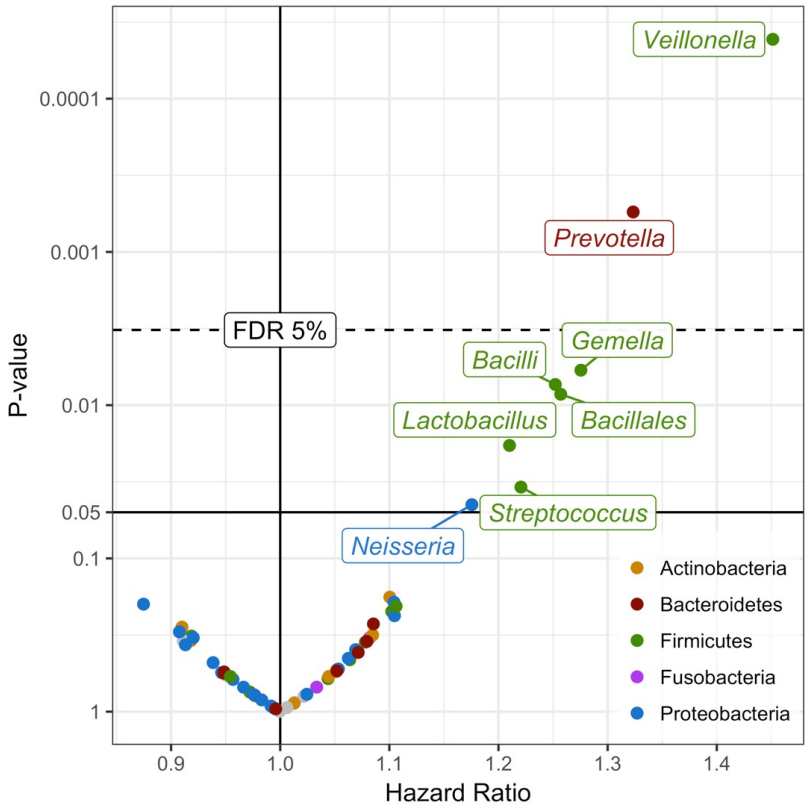
COPSAC$_{2010}$ cohort

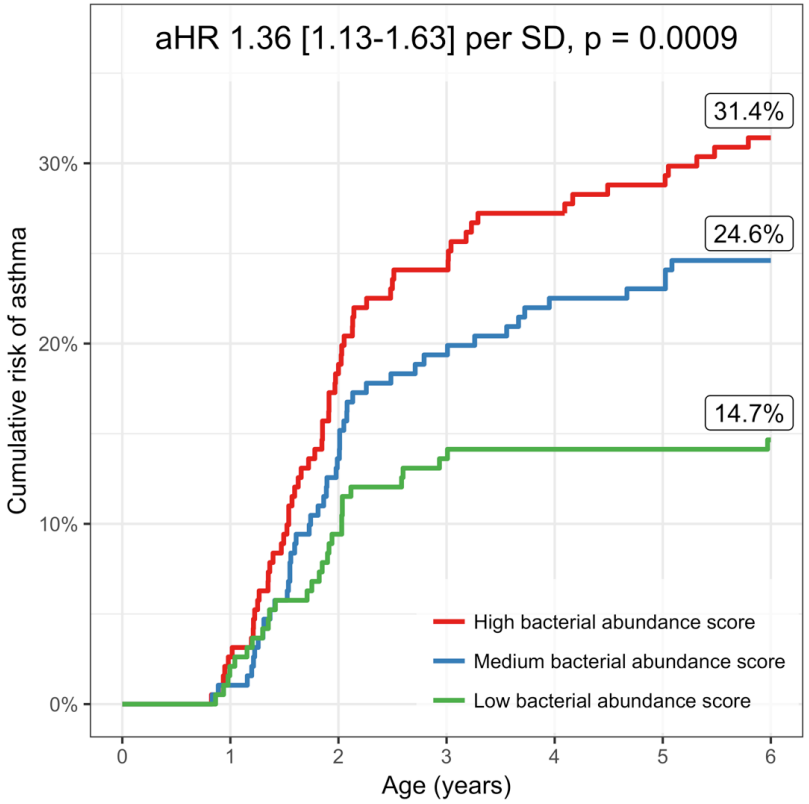


1-year microbiome maturity and asthma by age 5

Stokholm et al, *Nat Comm* 2018

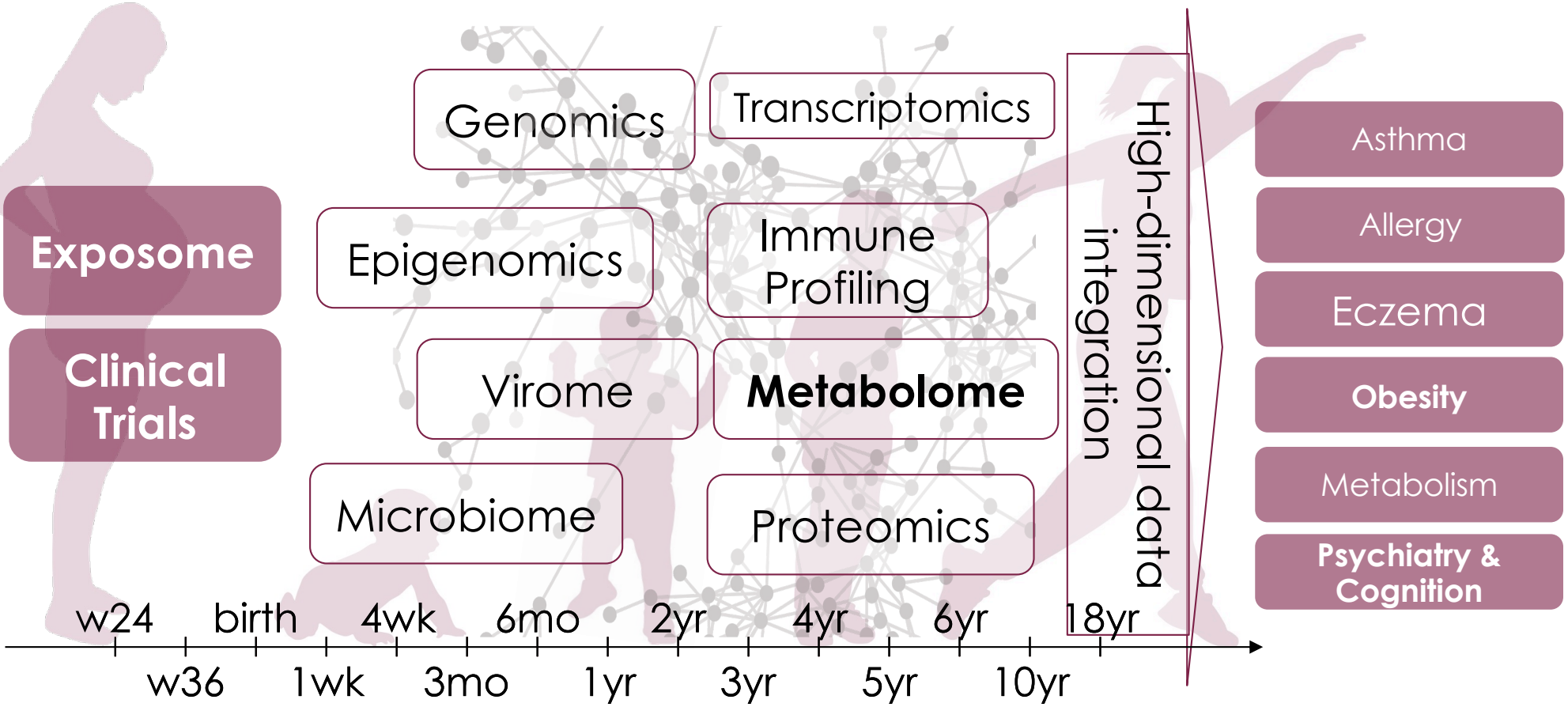
Infant airway microbiome and asthma



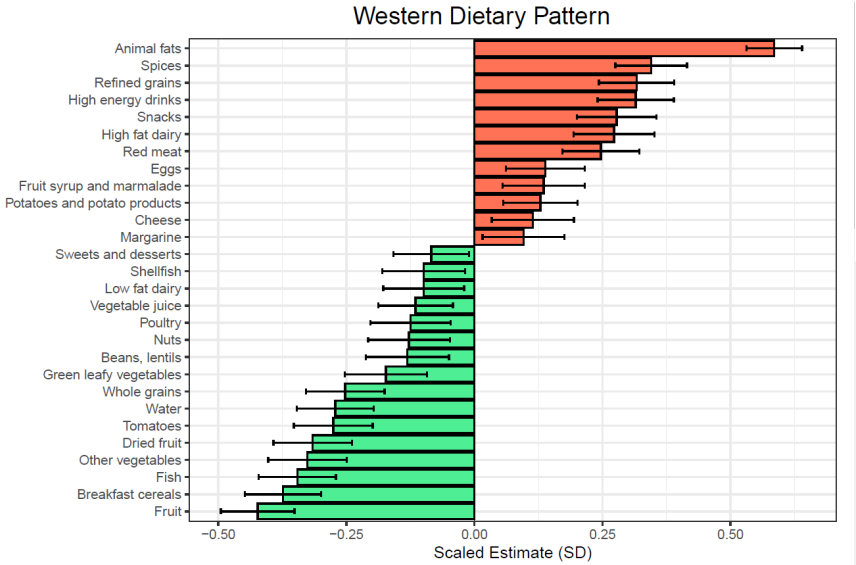


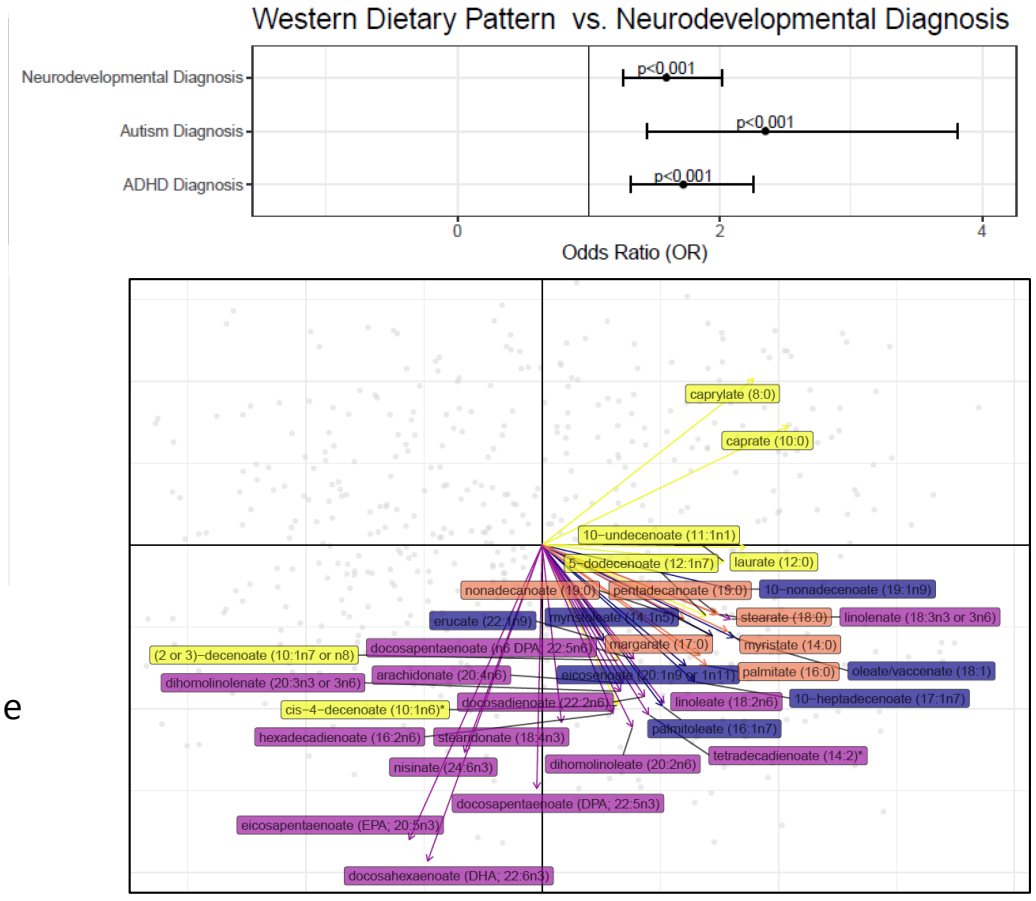
Thorsen et al, *Nature Comm.* 2019

Another example from COPSAC2010



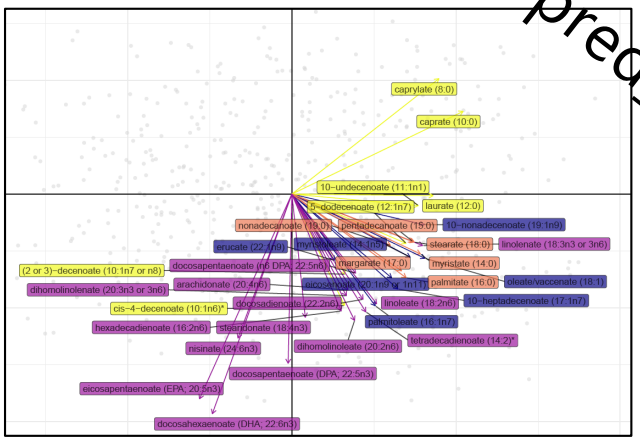
Diet in Pregnancy and offspring health





By the use of **blood metabolomics** in both pregnancy and childhood we can point towards the **critical window** in time is **during pregnangy** and not just a marker of a constant poor family diet. Further, elucidate mechanisms of inflammatory metabolic products

Metabolomics as mediator



SPLS

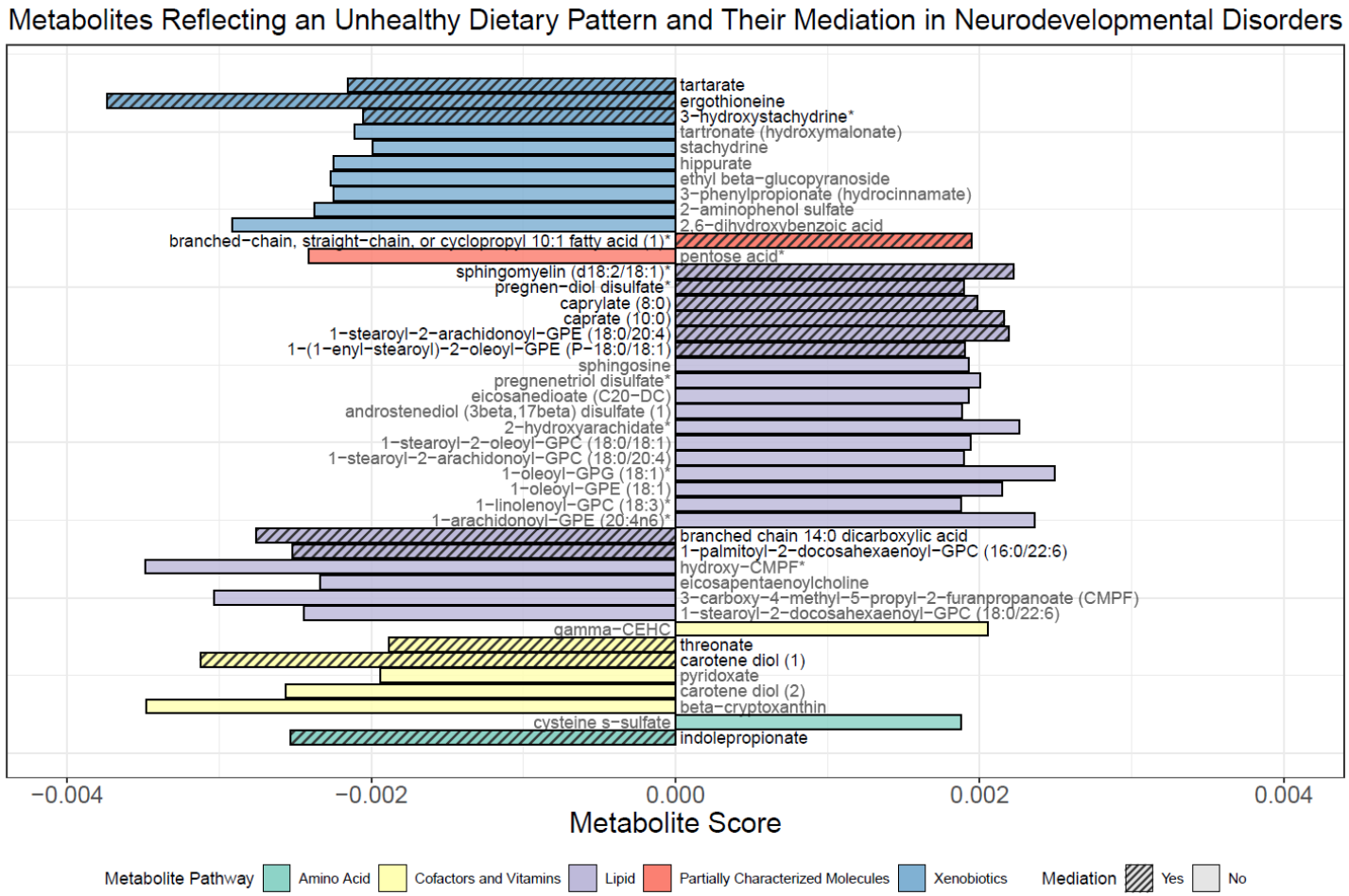
Metabolism

Diet

Neuro development

Metabolomics as mediator

Boolean backward elimination *Keep/kill without re-estimation*



Cook-Book

- Set up data in phyloseq
- Take appropriate preprocessing choices and maybe remove samples with low seq-depth
- Perform alpha diversity analysis versus design
- Perform beta diversity analysis versus design
- Do DA and report overall results as e.g. volcano-plot and maybe reference the results against phylogenetics
    - Maybe tax agglomerate to a higher taxonomic level, and repeat analysis to see at which taxonomic level the associations are pronounced
- Do omics-omics analysis as correlation heatmaps
- Perform a supervised omics-omics using CCA or (s)PLS2 including cross-validation.
- In case of a triangulation setup, use propensity scoring and mediation to add some “causality”