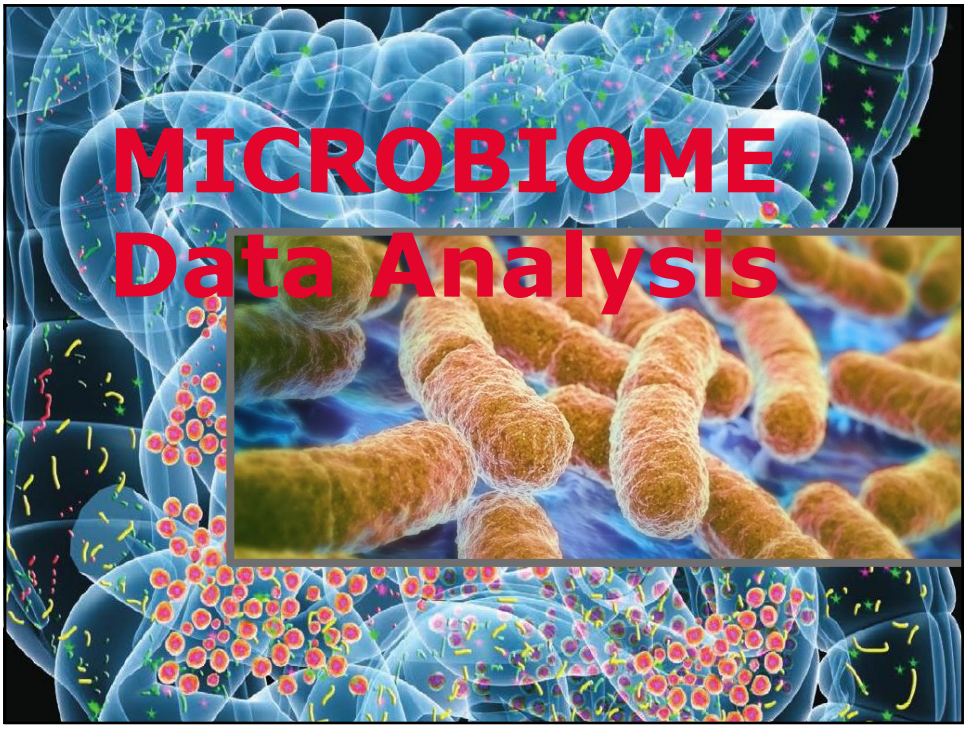


1

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

2

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 Morning Preprocessing Alpha diversity Beta diversity	**Day2 Morning** Testing design versus Beta diversity
Day1 Afternoon Differential Abundance testing	**Day2 Afternoon** Multiomics with heatmaps and CCA

Preprocessing

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

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

Diversity metrics

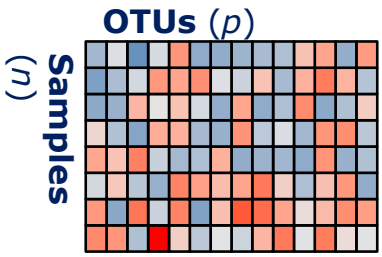
Alpha diversity

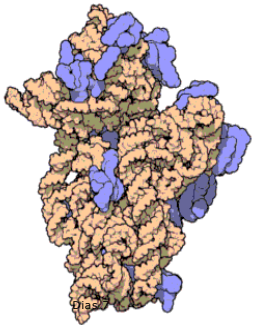
Within sample characteristics

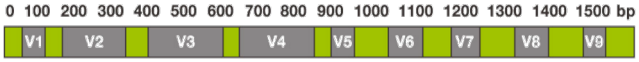
Beta diversity

Between sample characteristics

Amplicon







CONSERVED REGIONS: unspecific applications
VARIABLE REGIONS: group or species-specific applications

Alpha diversity

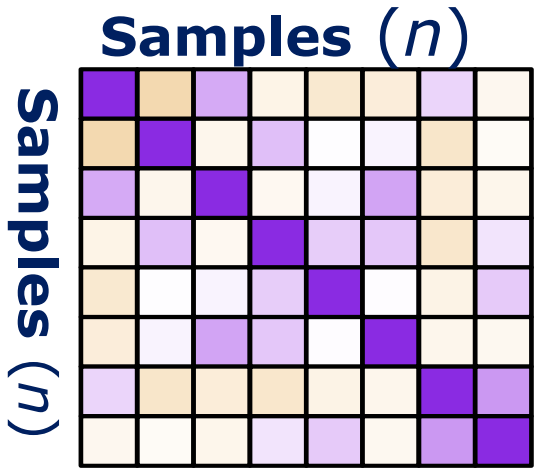
Number of different taxa

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

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

11

Bray Curtis

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

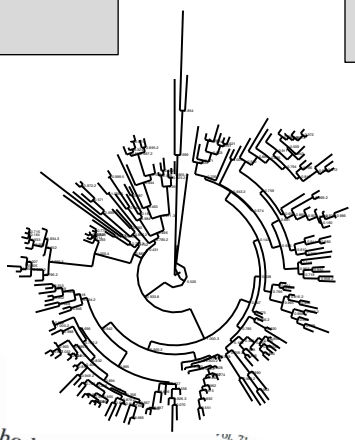
X_{ij}, X_{ik} Number of individuals in species i in each sample (j, k)
n Total number of species in samples.

12

UNIFRAC

OTU table

Dist



AND ENVIRONMENTAL MICROBIOLOGY, Dec. 2005, p. 8228–8235
0/05/$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, [1] and Department of Ch... Colorado 80309, and Department of ... Colorado, ...

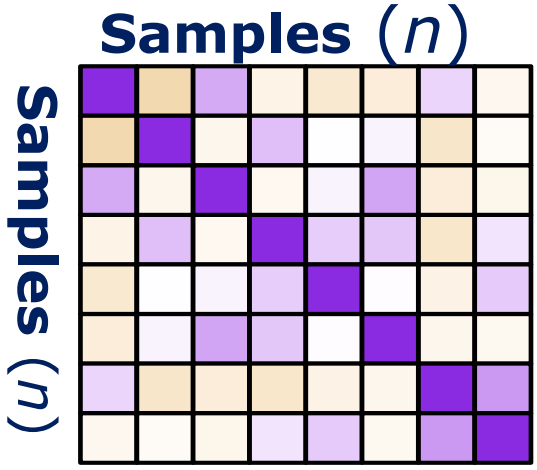
...ol. 71, No. 12

13

Ordination

14

β diversity *to* PCoA

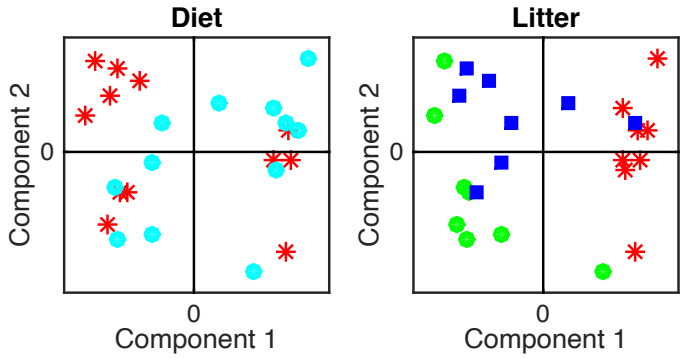


$= U\Lambda U'$

15

Multidimensional scaling

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



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

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

16

Differential Abundance Testing *or* OTUWAS

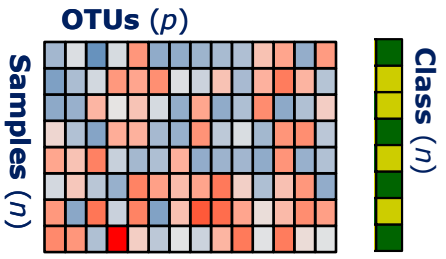
17

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.



18

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)

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

21

Controlling Type I error

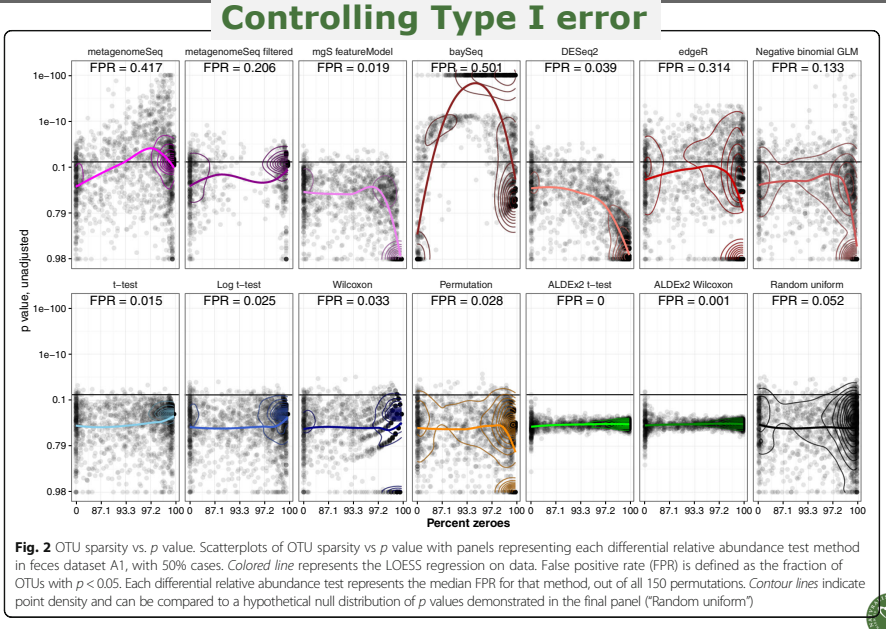


Fig. 2 OTU sparsity vs. *p* value. Scatterplots of OTU sparsity vs *p* value with panels representing each differential relative abundance test method in feces dataset A1, with 50% cases. *Colored line* represents the LOESS regression on data. False positive rate (FPR) is defined as the fraction of OTUs 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")

22

Correct ordering

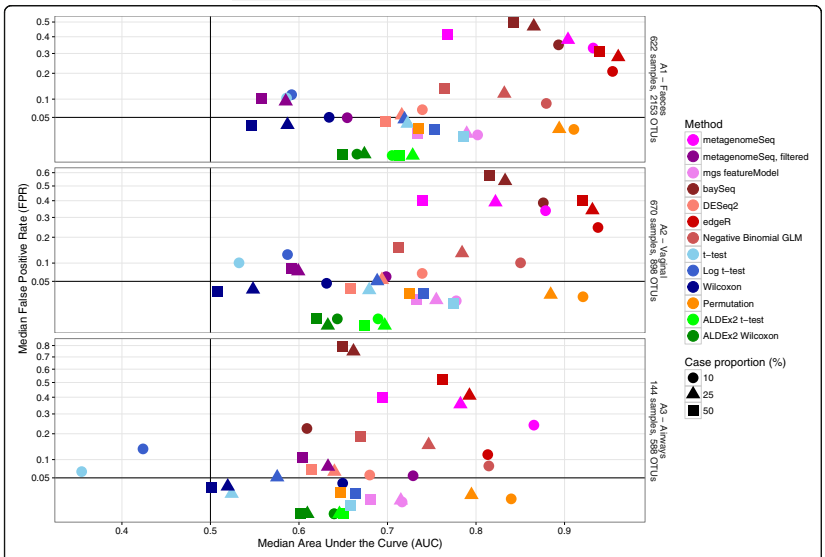


Fig. 3 Median test AUC vs median test FPR. Scatterplots of median test area under the curve (AUC) vs median test false positive rate (FPR), for the datasets A1–A3; three different compartments in the $COPSAC_{2010}$ cohort. FPR is defined as the fraction of OTUs with $p < 0.05$. *Dot color* represents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

23

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*

24

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

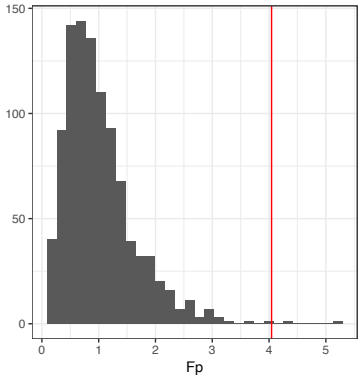
For univariate ANOVA models, we rely on independence and homoschedastic and gaussian *like* residuals.

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.

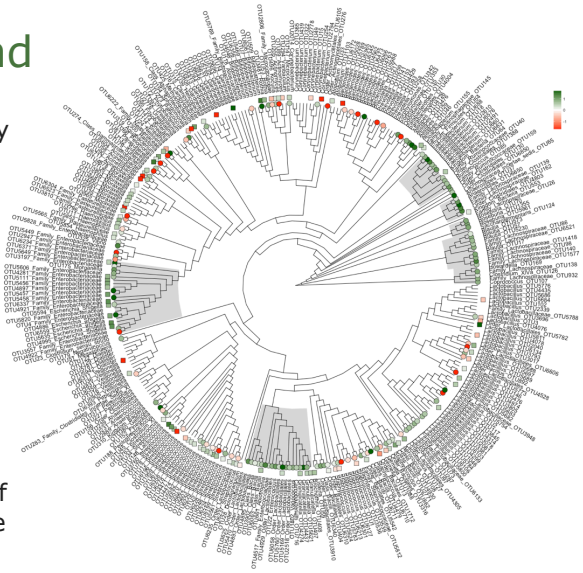


27

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

28

MultiOmics

29

Data integration

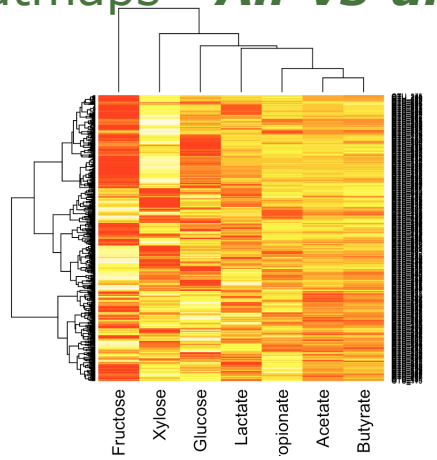
-Omics	OTU table
$_n\mathbf{X}^{p_X}$	$_n\mathbf{Y}^{p_Y}$

30

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

31

Data integration

-Omics	OTU table
$_{n}\mathbf{X}^{p_X}$	$_{n}\mathbf{Y}^{p_Y}$

32

32

Functionally Supervised

Normal CCA

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

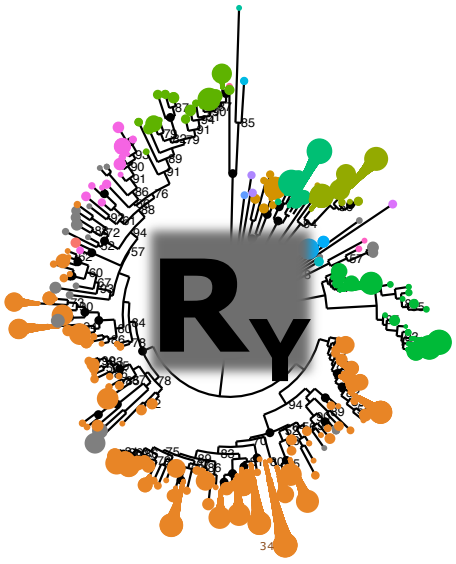
Supervised CCA

- Let $\mathbf{R_Y}$ be a symmetric matrix reflecting similarity between the features in **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**



UNIVERSITY OF COPENHAGEN

Functionally Supervised

Supervised CCA

Re-formalize the main objective

Kernel Smoothing

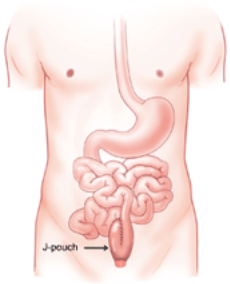
$$\max_{\mathbf{w_X},\mathbf{w_Y}} \mathbf{w_X}^T \mathbf{X}^T \mathbf{Y} \mathbf{R_Y} \mathbf{w_Y}$$

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

35

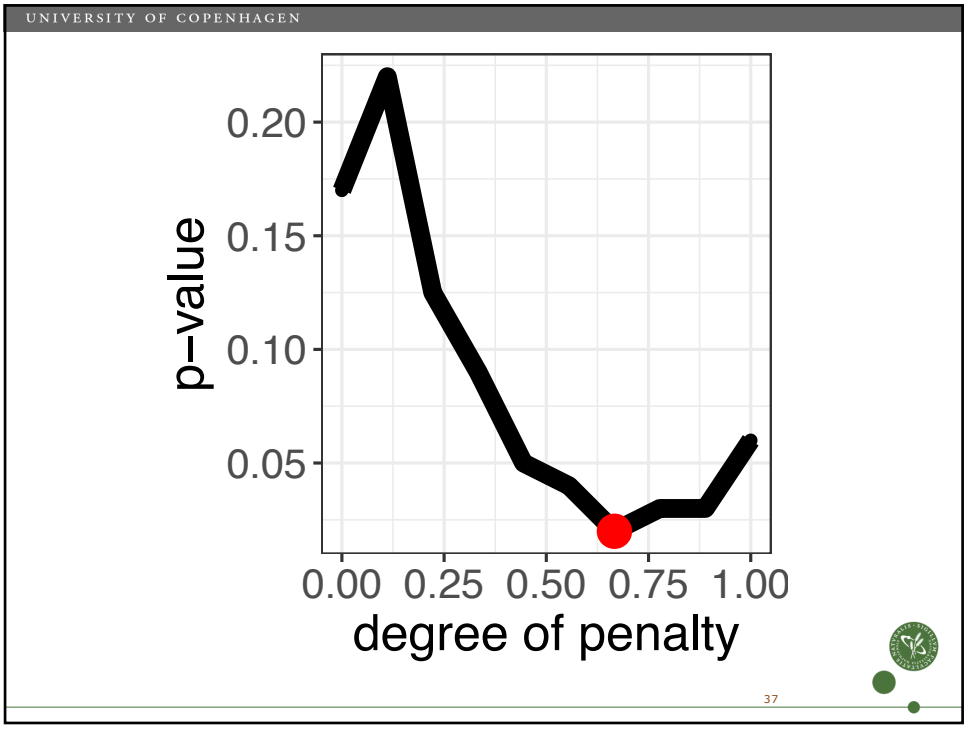
UNIVERSITY OF COPENHAGEN

Example Pouchitic cohort

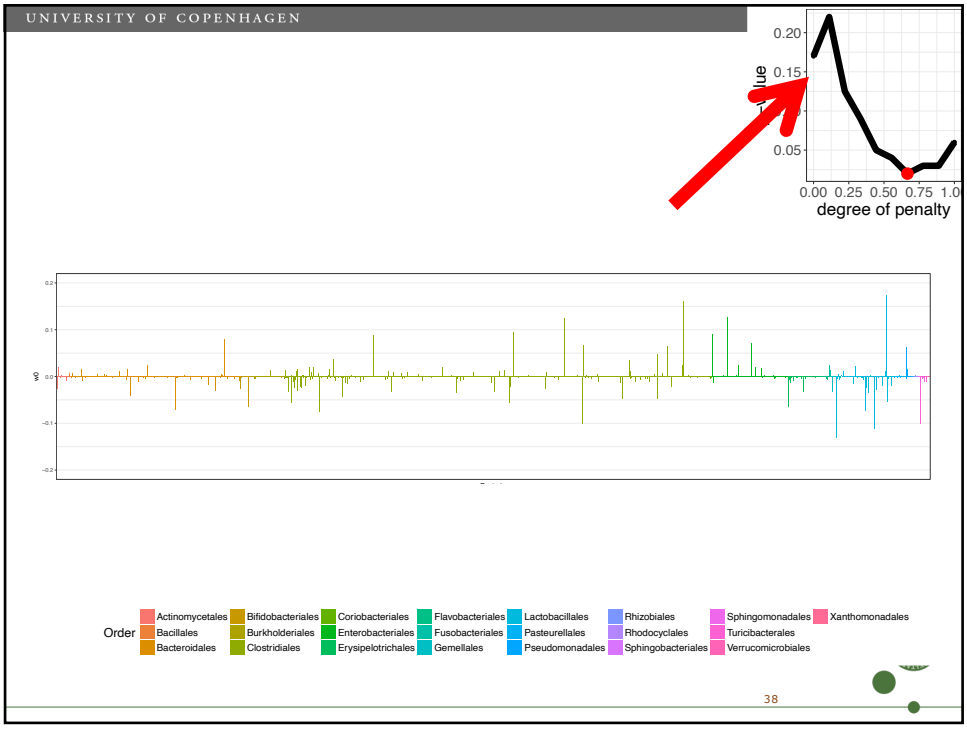


Gene Expression	OTU table
$248 \mathbf{X}^{1991}$	$248 \mathbf{Y}^{1380}$

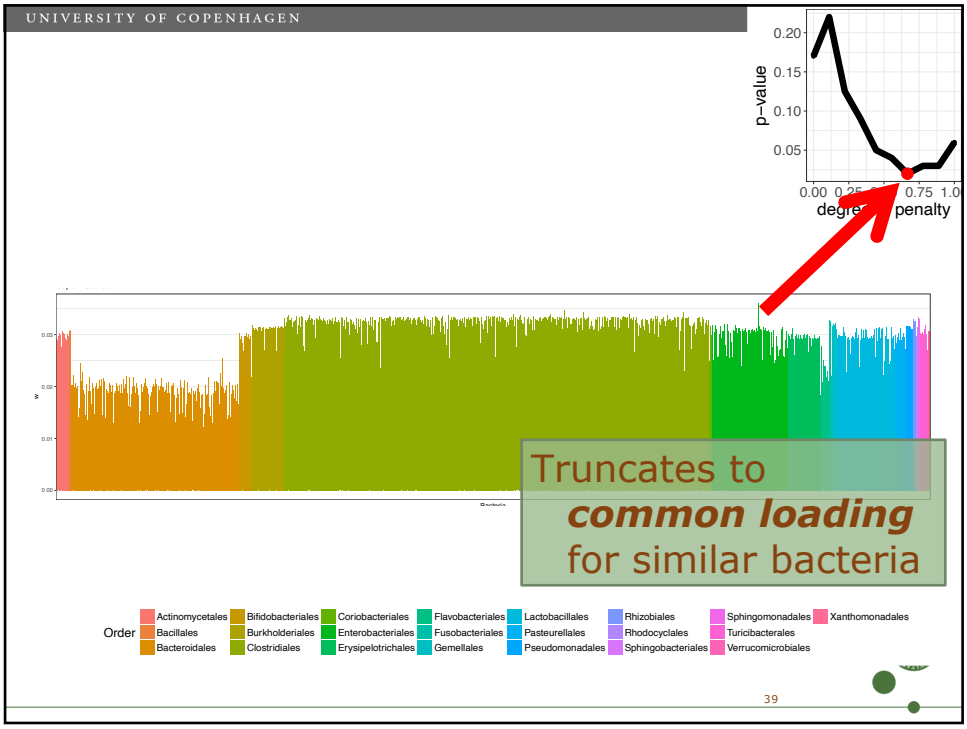
36


UNIVERSITY OF COPENHAGEN
0.20
0.15
0.10
0.05
p–value
0.00 0.25 0.50 0.75 1.00
degree of penalty
37

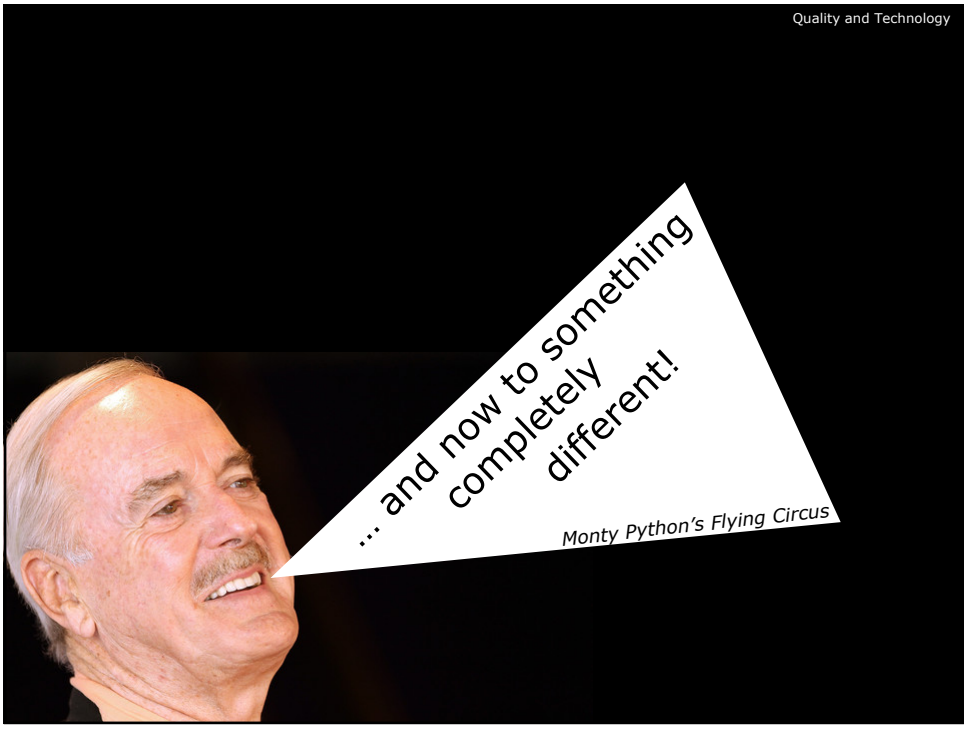
37


UNIVERSITY OF COPENHAGEN
0.20
0.15
0.05
0.00 0.25 0.50 0.75 1.0
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
38

38


UNIVERSITY OF COPENHAGEN
p−value
0.20
0.15
0.10
0.05
0.00
0.75
1.0
penalty
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
39

39

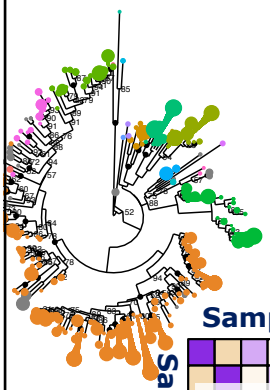

Quality and Technology
... and now to something completely different!
Monty Python's Flying Circus

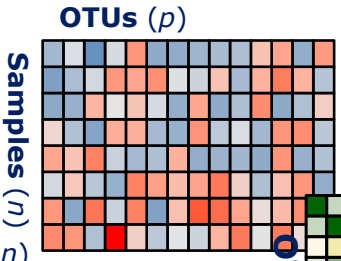
40

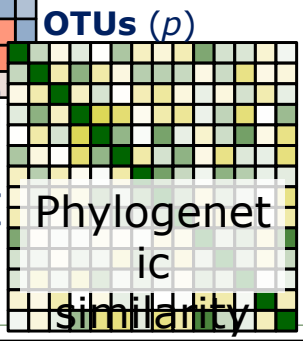
UNIVERSITY OF COPENHAGEN

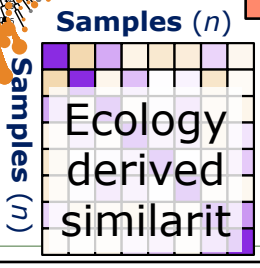
Data analysis

We know *more* about the features than *just* their number/name!





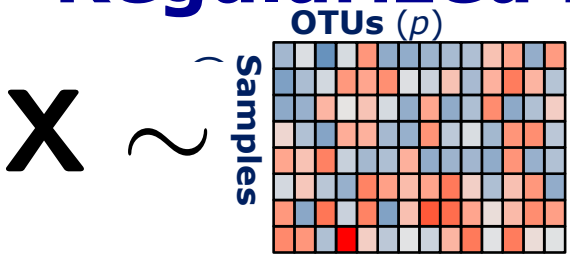


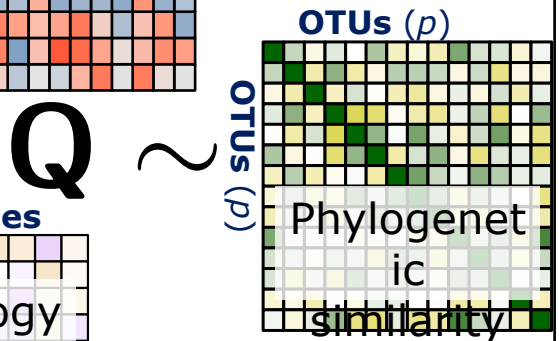


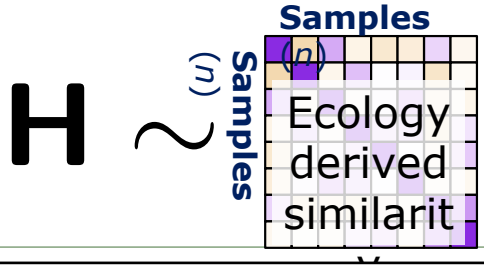
41

UNIVERSITY OF COPENHAGEN

Regularized PLS







42

Regularized PLS

Objective

Find a PLS solution such that

1. The covariance between **X** and **Y** are high
2. The weights resemble known feature2feature structure
3. The scores resemble ecological structure
4. (…including variable selection)

43

Regularized PLS

Objective

1. Covariance btw **X** and **Y**
2. Resemble feature structure
3. Resemble ecological structure
4. Variable selection

$$\max_{\mathbf{w}}(\mathbf{w}^T\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{w})$$

$$\max_{\mathbf{w}}(\mathbf{w}^T\mathbf{Q}\mathbf{w})$$

$$\max_{\mathbf{w}}(\mathbf{w}^T\mathbf{X}^T\mathbf{H}\mathbf{X}\mathbf{w})$$

$$\|\mathbf{w}\|_1 \leq c \qquad \|\mathbf{w}\|_2 = 1$$

44

Regularized PLS

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} (\mathbf{w}^T(\alpha_1 \mathbf{Q}\mathbf{X}^T\mathbf{Y}\mathbf{Y}^T\mathbf{X}\mathbf{Q} + (1-\alpha_1)\mathbf{X}^T\mathbf{H}\mathbf{X})\mathbf{w})$$

$$\hat{\mathbf{w}} \leftarrow \mathit{softthr}(\hat{\mathbf{w}}, \lambda)$$

$$\mathbf{Q} = \alpha_2 \mathbf{Q}^* + (1-\alpha_2)\mathbf{I}$$

$$\|\mathbf{w}\|_2 = 1$$

45

46

UNIVERSITY OF COPENHAGEN

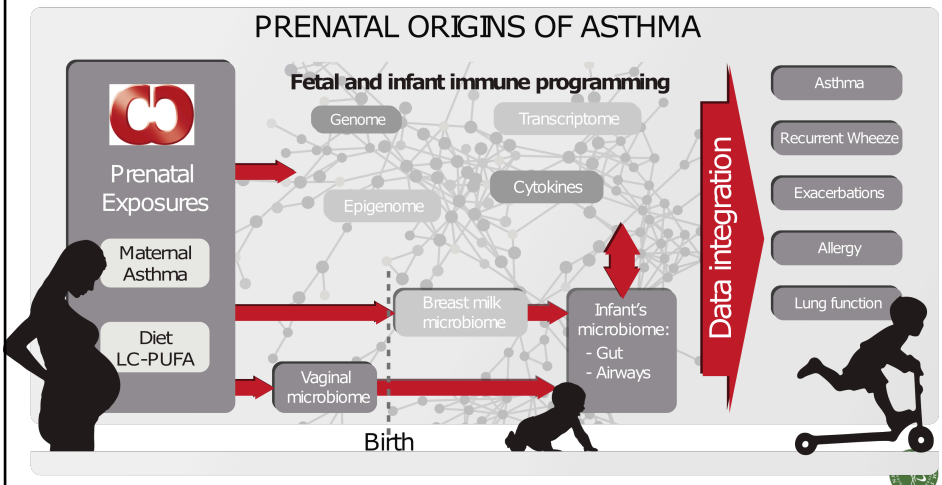
FAST

Birth mode
C Section vs Vaginal
fecal @1week

47

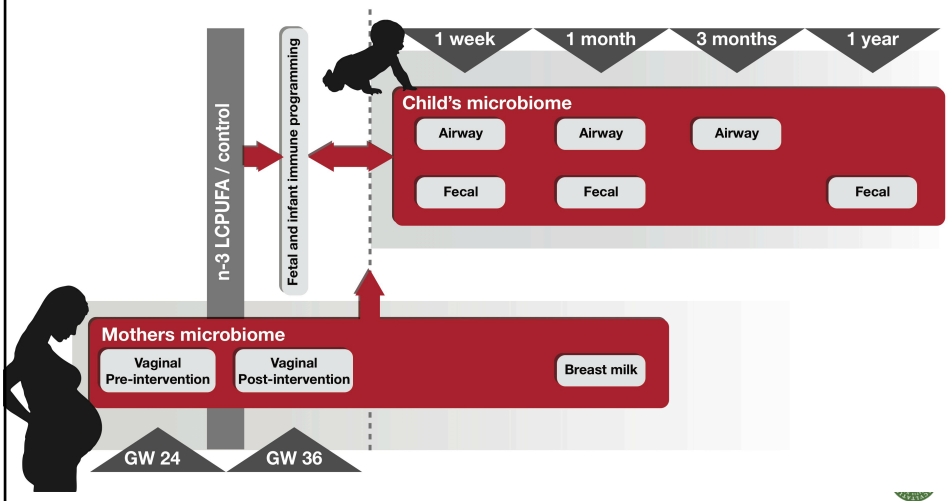
UNIVERSITY OF COPENHAGEN

COPSAC$_{2010}$ Cohort

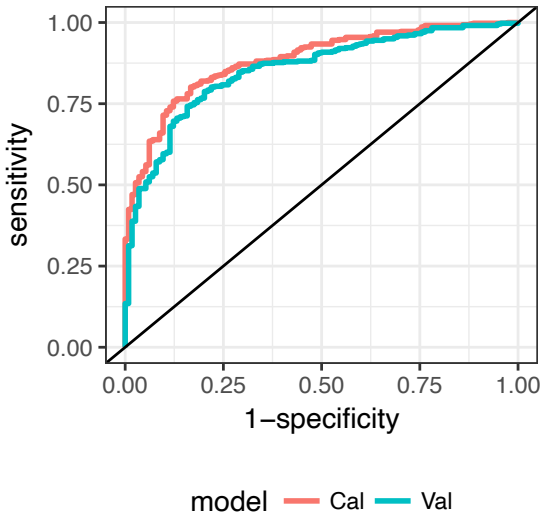


48

Microbiome data



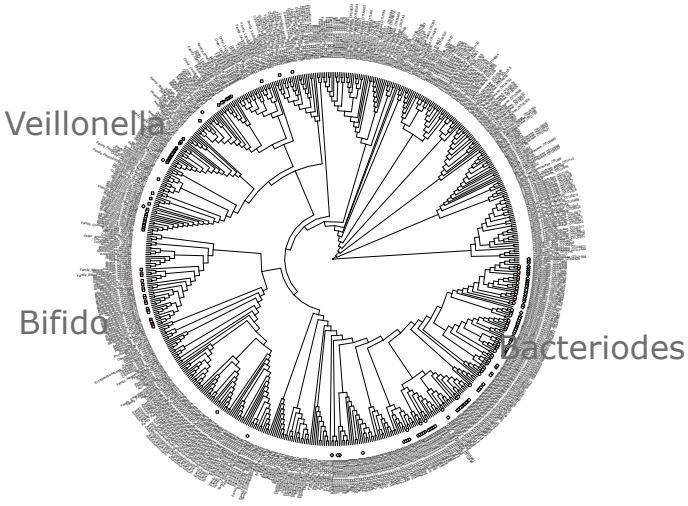
49



50

50

Loadingplot



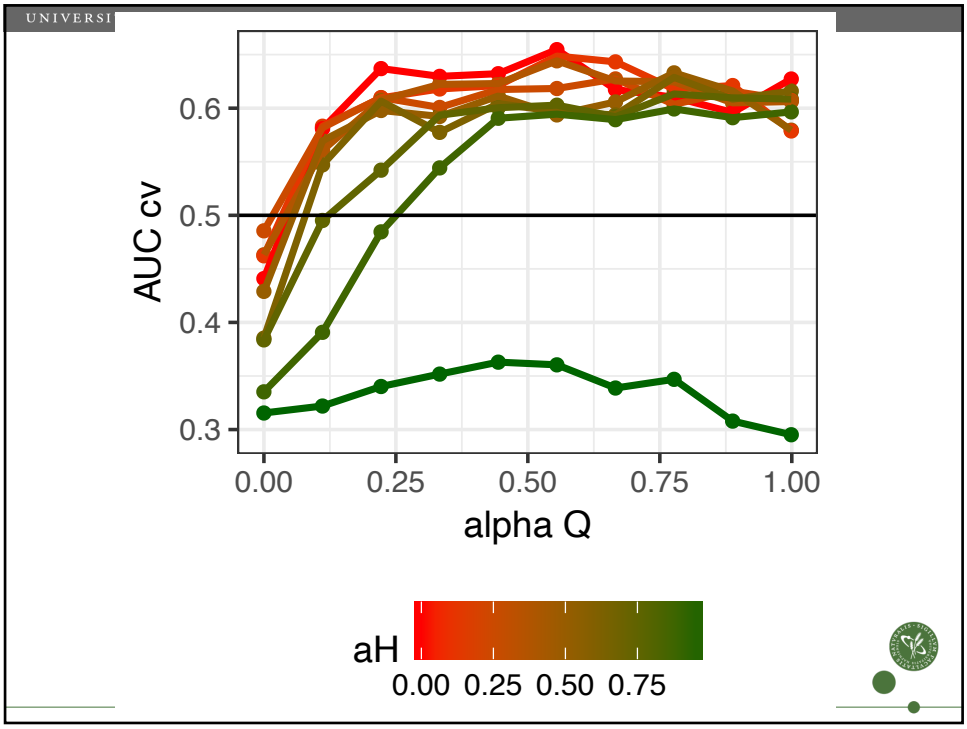
51

DIFFICULT

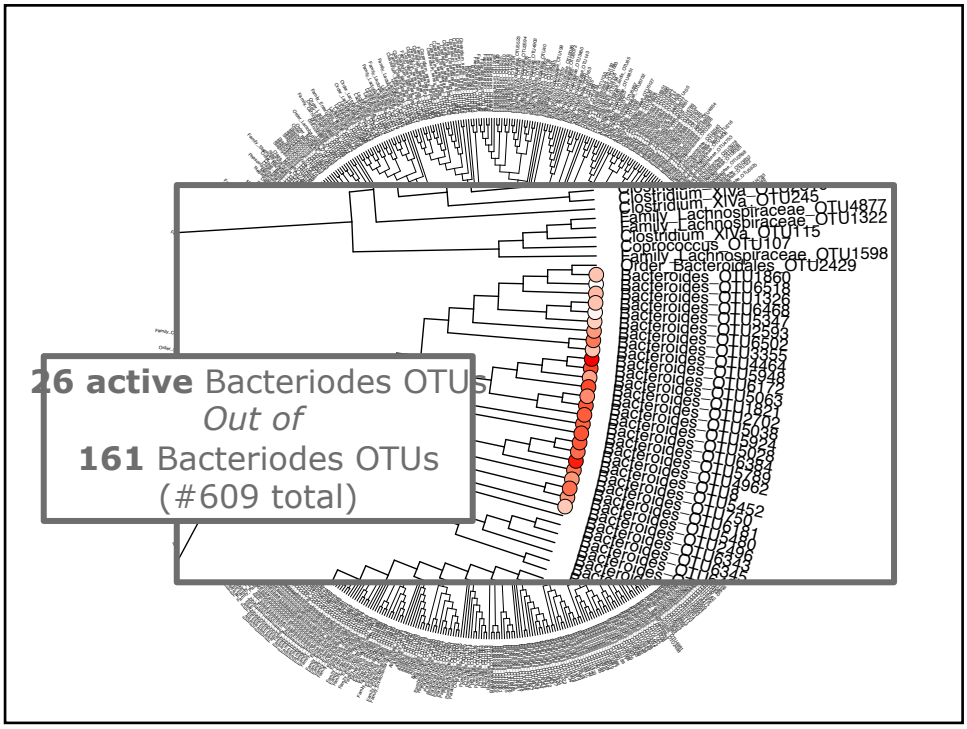
Asthma 6y fecal @1year

- in asthmatic mothers-

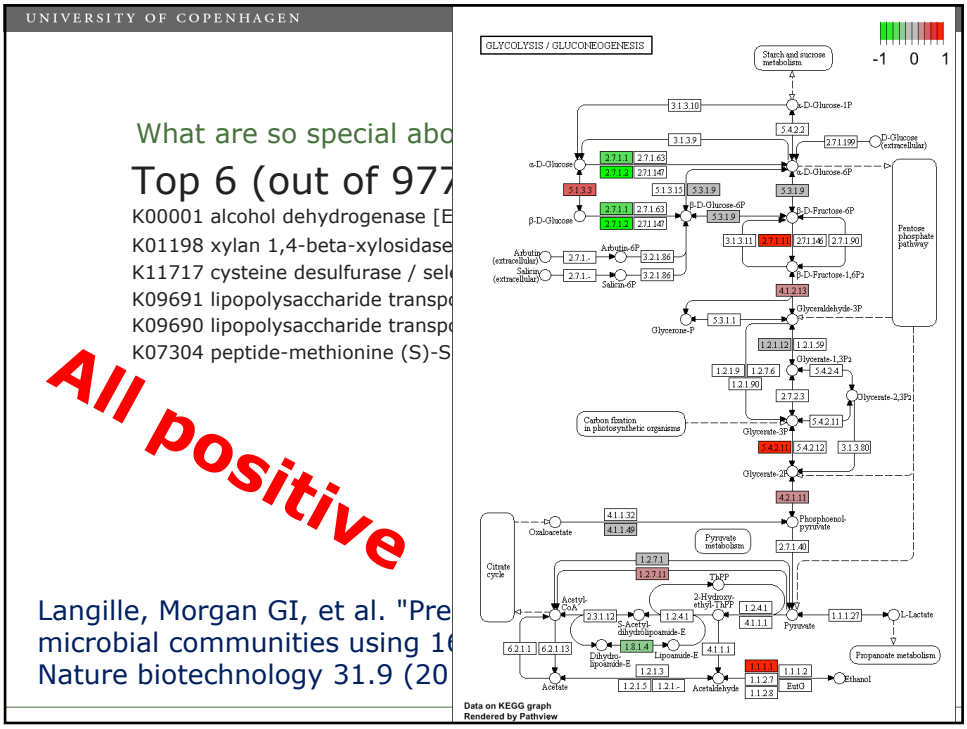
52


UNIVERSI
AUC cv
0.6
0.5
0.4
0.3
0.00
0.25
0.50
0.75
1.00
alpha Q
aH
0.00 0.25 0.50 0.75

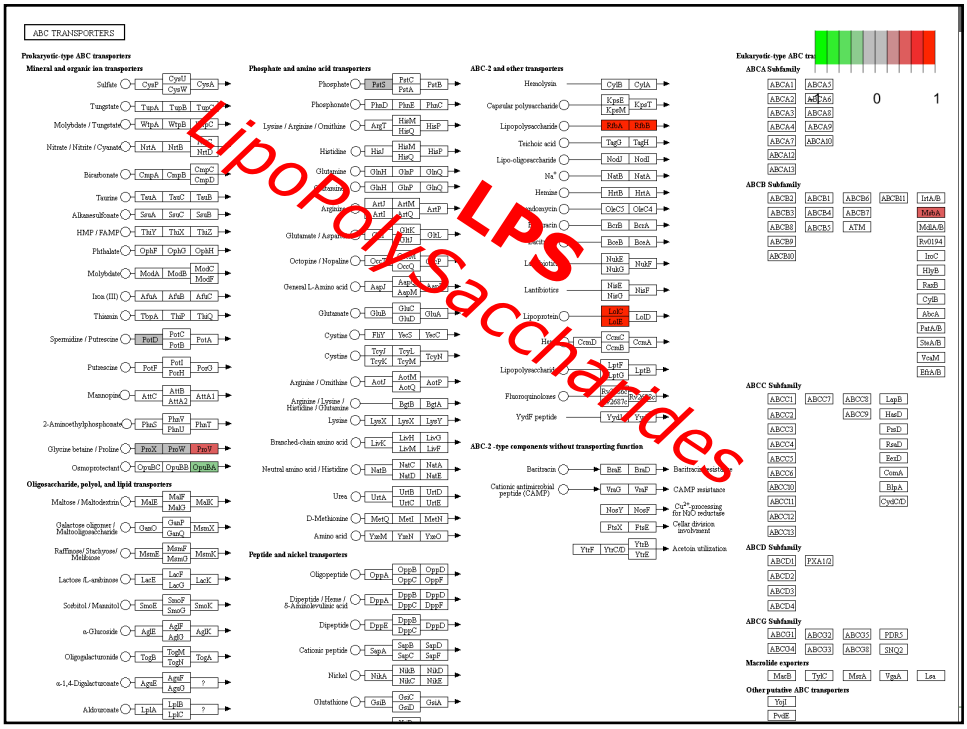
53


26 active Bacteriodes OTUs
Out of
161 Bacteriodes OTUs
(#609 total)

54


UNIVERSITY OF COPENHAGEN
What are so special abo
Top 6 (out of 977
K00001 alcohol dehydrogenase [E
K01198 xylan 1,4-beta-xylosidase
K11717 cysteine desulfurase / sel
K09691 lipopolysaccharide transp
K09690 lipopolysaccharide transp
K07304 peptide-methionine (S)-S
All positive
Langille, Morgan GI, et al. "Pre
microbial communities using 16
Nature biotechnology 31.9 (20
GLYCOLYSIS / GLUCONEOGENESIS
-1 0 1
Data on KEGG graph
Rendered by Pathview

55


ABC TRANSPORTERS
Lipopolysaccharides
LPS
0 1

56