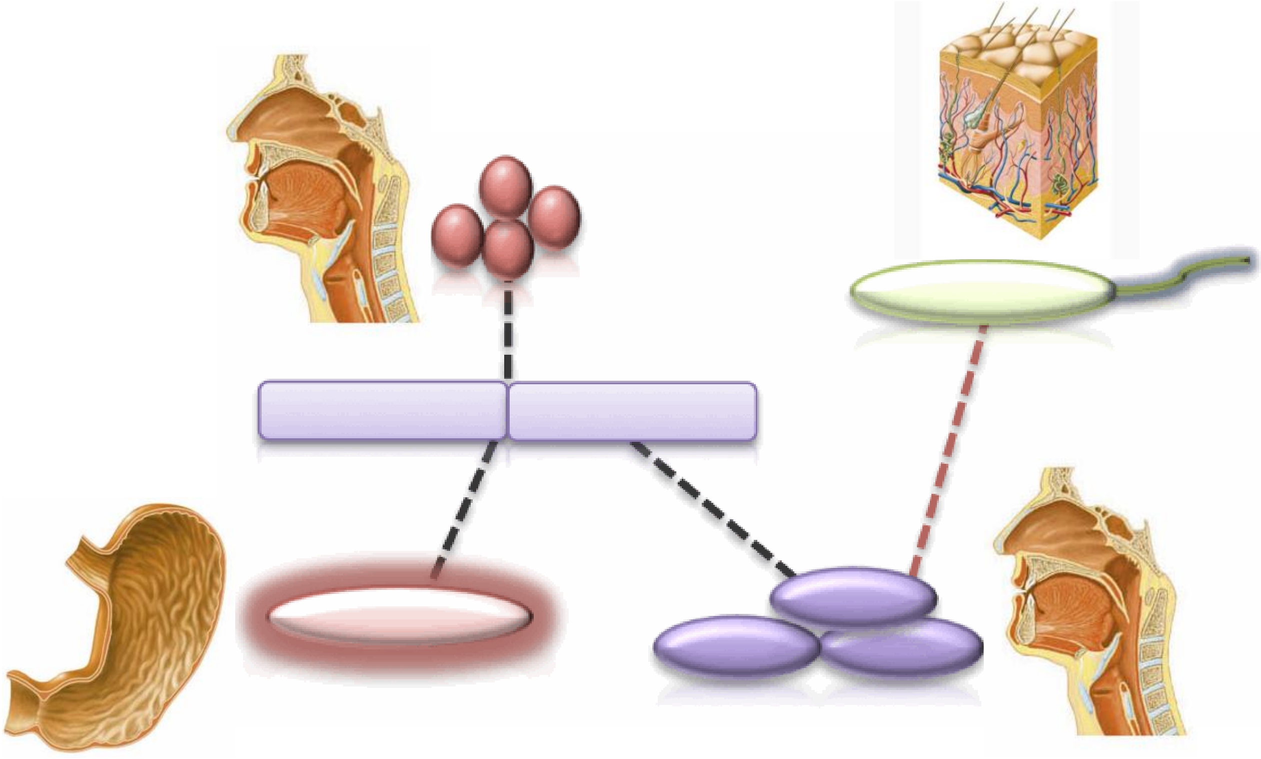
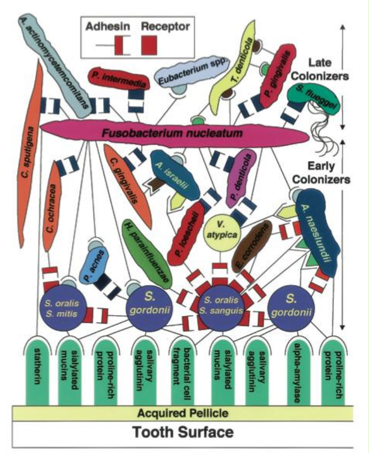
Karoline Faust
Raes Lab (Bioinformatics and (Eco-)Systems Biology)

10th October 2012
Bertinoro Computational Biology

Detecting bacterial associations in the human microbiome

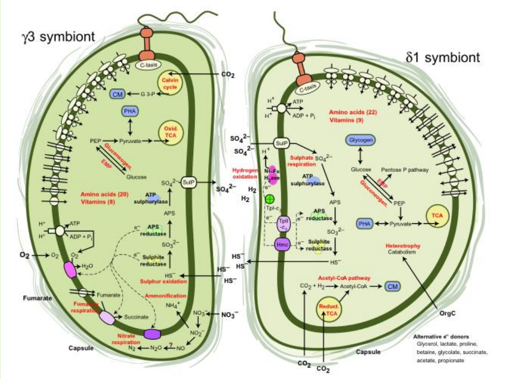


Examples for microbial relationships



dental plaque formation (Kolenbrander et al.)

sulfur oxidizer sulfate reducer

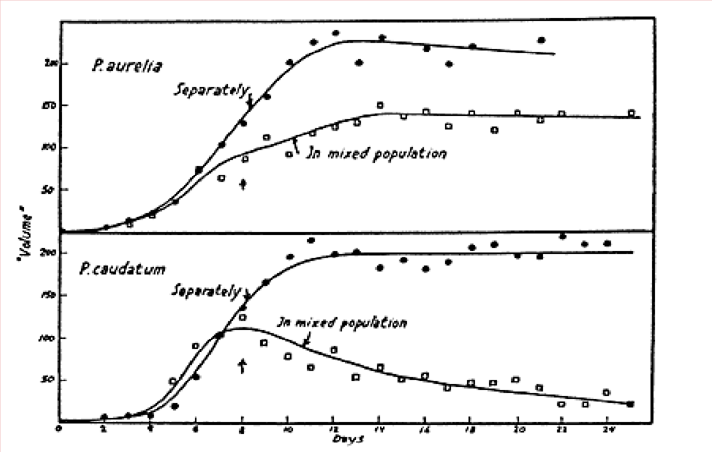


cross-feeding between bacterial symbionts of a marine worm (Woyke et al.)

artist’s rendering of human skin bacteria

Amoeba proteus feeding on algae

Bacteriophages infecting a bacterium



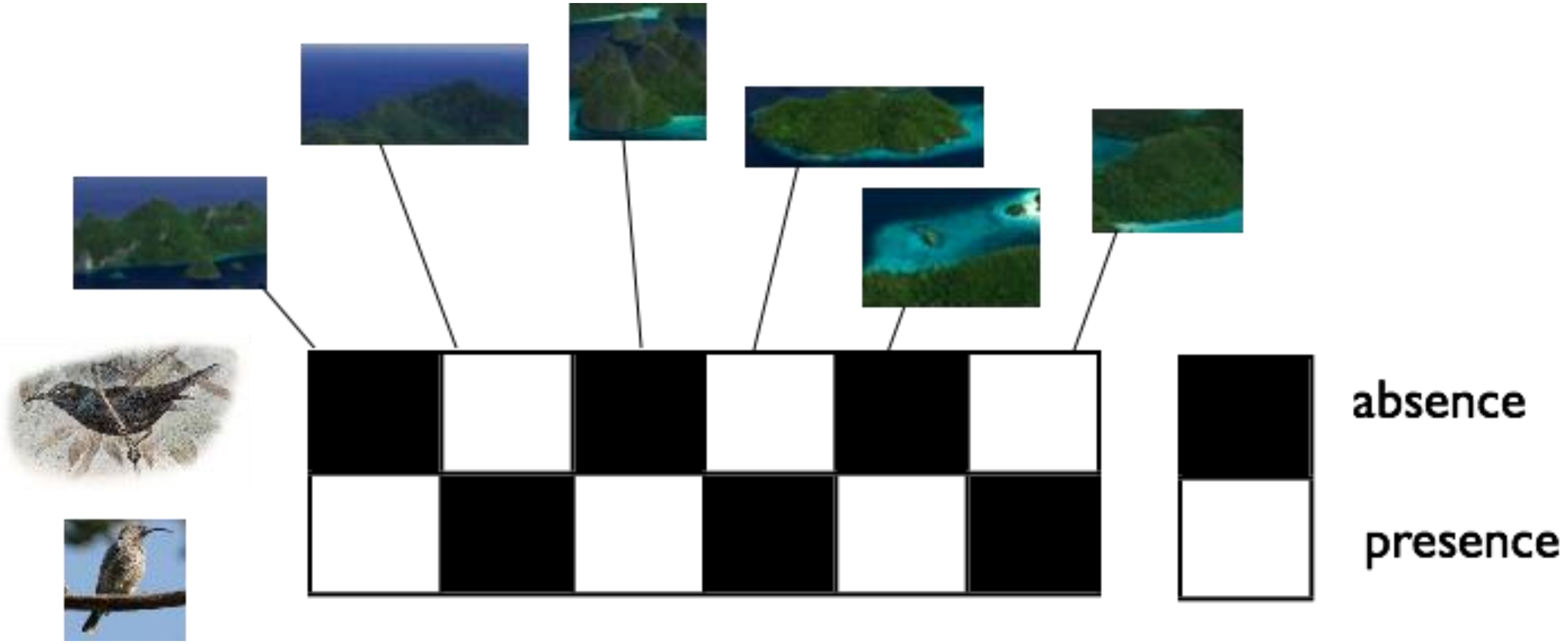
competition between two species of Paramecium (Gause)

algae bloom killing off other organisms

Gause (1934) “The Struggle for Existence”, Williams & Wilkins.
Kolenbrander et al. (2002) “Communication among Oral Bacteria”, Microbiol. and Mol. Biol. Reviews 66, pp. 486-505.
Woyke, T. et al. (2006) “Symbiotic insights through metagenomic analysis of a microbial consortium”, Nature 443, pp. 950-955.

Detecting ecological relationships from presence/absence data

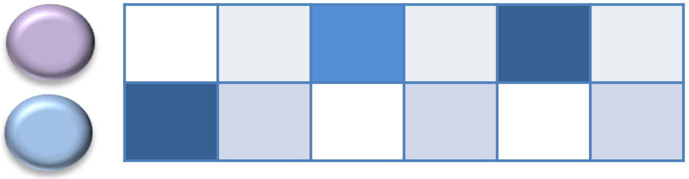
- Jared Diamond suggested that competition between species could be seen from their presences/absences across habitats (checkerboard pattern)
- checkerboard-like co-occurrence patterns have been found for micro-organisms as well (Horner-Devine et al.)



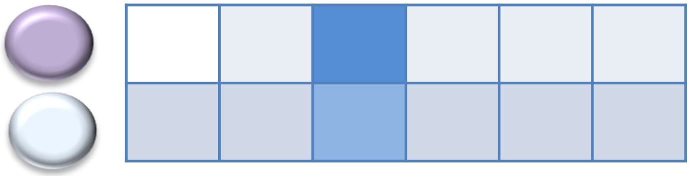
Diamond, J. (1975) “Assembly of species communities”, pp. 342-444 in “Ecology and evolution of communities” edited by Cody and Diamond, Harvard University Press.
Horner-Devine M.C. et al. (2007) “A Comparison Of Taxon Co-Occurrence Patterns For Macro- And Microorganisms” Ecology 88, pp. 1345-1353.

1. Introduction

Co-occurrence analysis in a nut shell



mutual exclusion (checker board)/anti-correlation



co-occurrence/correlation

1. Introduction

Reasons for association

Why would two taxa consistently occur together or avoid each other across samples?

ecological relationships

predator/parasite (win-loss)

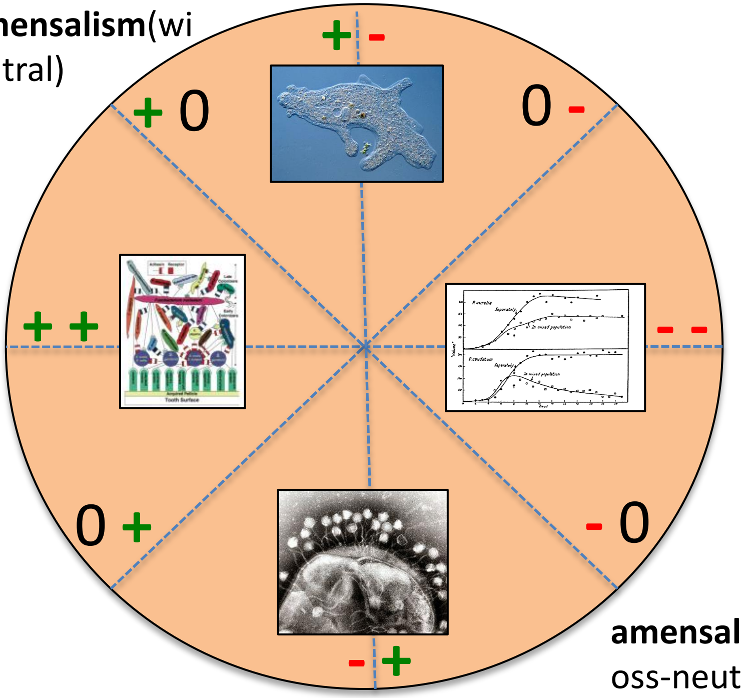
commensalism(win-neutral)

mutualism (win-win)

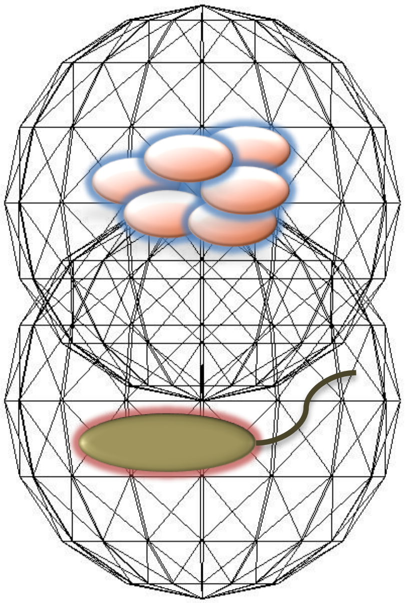
competition (loss-loss)

amensalism(loss-neutral)

prey/host (loss-win)



niche overlap

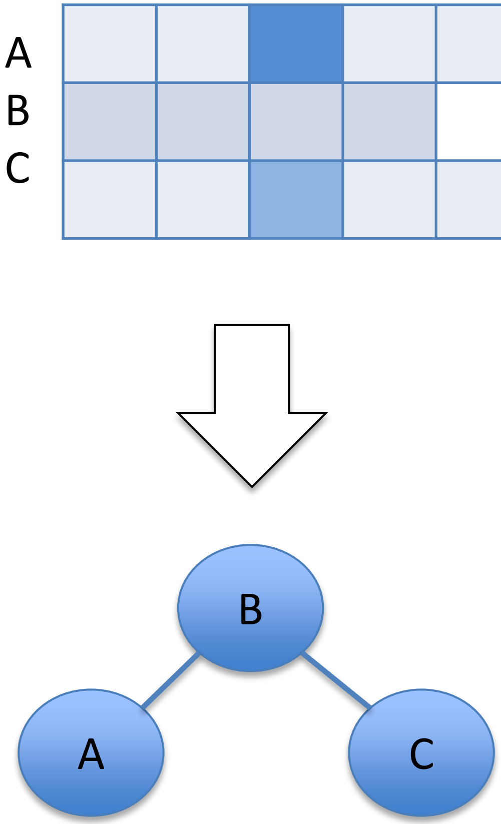


Adapted from Lidicker, W.Z. (1979) “A Clarification of Interactions in Ecological Systems”, BioScience 29, pp. 475-477.

Hutchinson, G.E. (1957) “Concluding remarks”, Cold Spring Harbour Symposium on Quantitative Biology 22, pp. 415-427.

1. Introduction

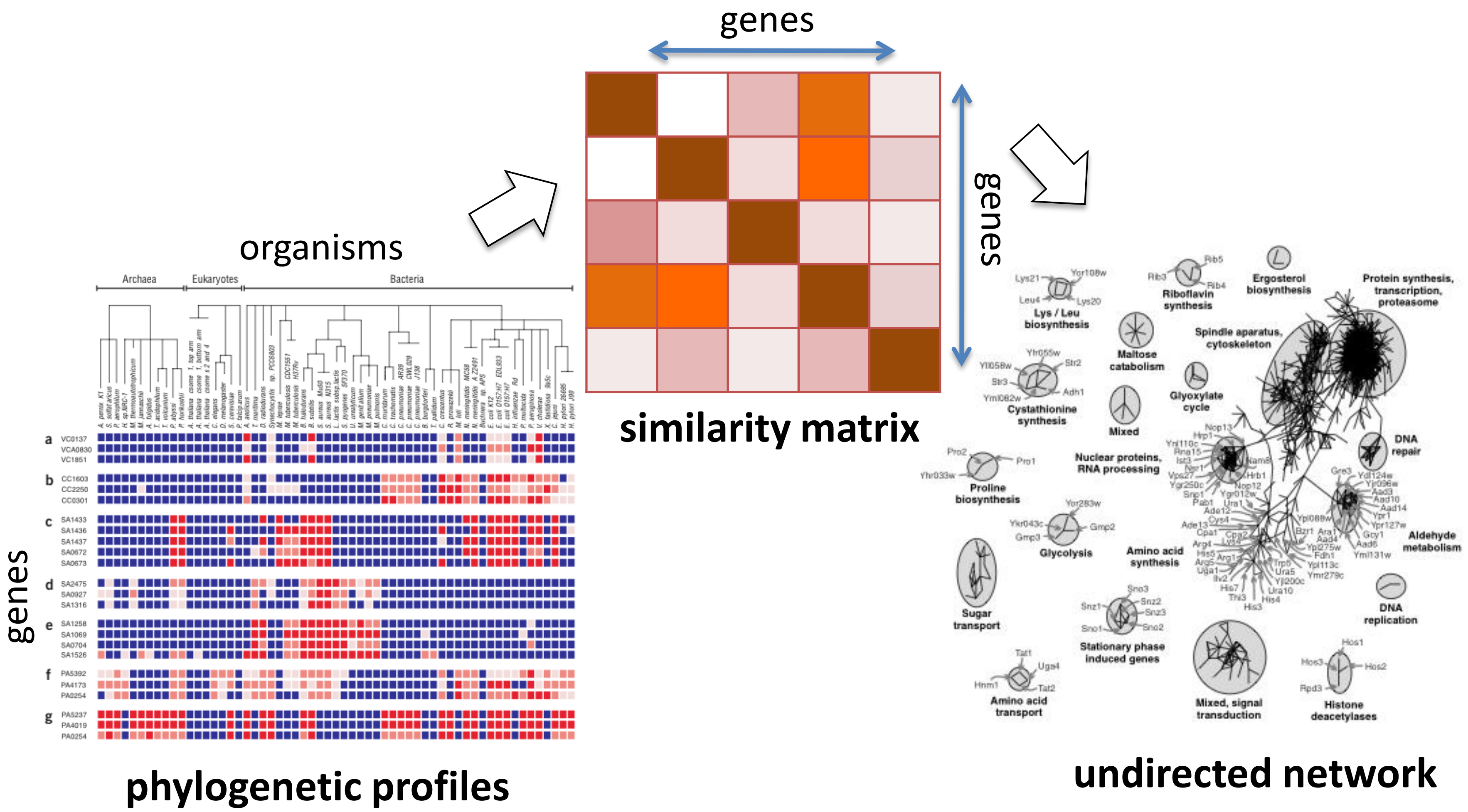
Inferring networks



- **network inference**: the problem of finding relationships between objects (genes, proteins, metabolites, species...) whose presence/absence or abundance was observed repeatedly

Example for similarity-based network inference

- task: obtain functional protein modules from co-occurrences of genes



genes

genes

similarity matrix

organisms

genes

phylogenetic profiles

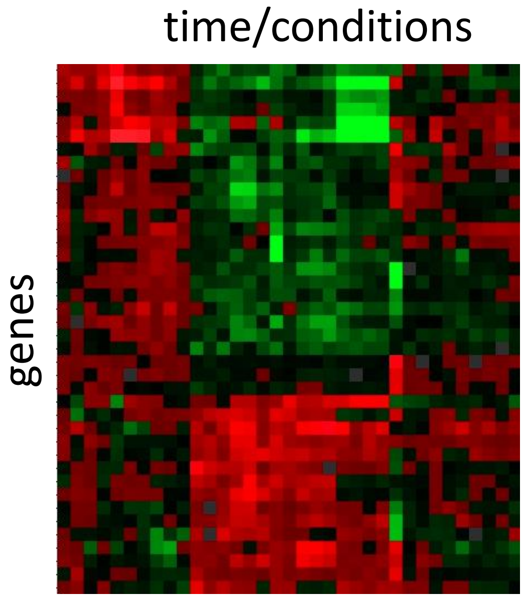
undirected network

Date, S.V. and Marcotte, E.M. (2003) “Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages”, Nature Biotechnology 21, pp. 1055-1062.

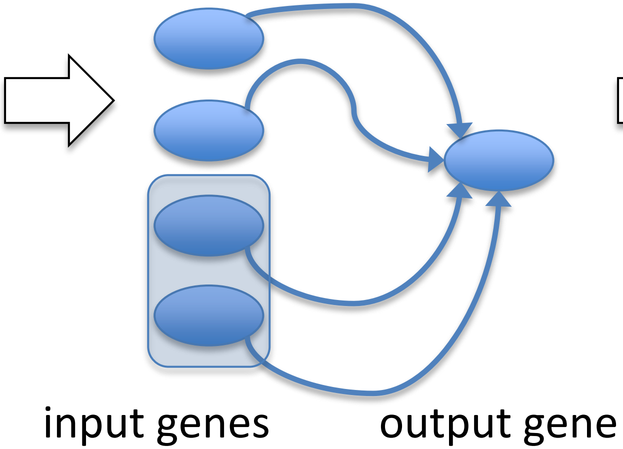
Example for sparse regression-based network inference

- task: identify gene regulatory network from microarray data

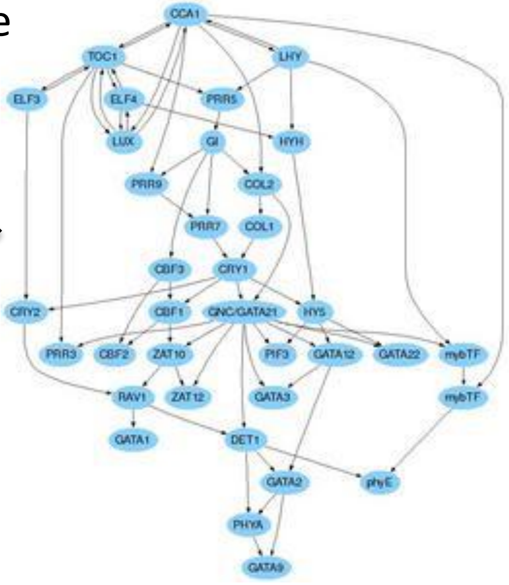
for each gene, find the regulators of that gene among all other genes: do sparse regression (using regression trees) to select the subset of input genes that predicts best the behavior of the output gene



microarray data



sparse regression

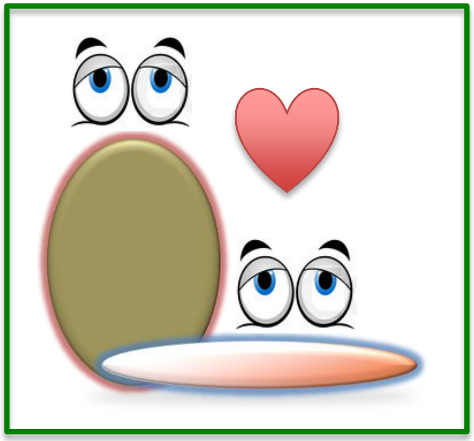


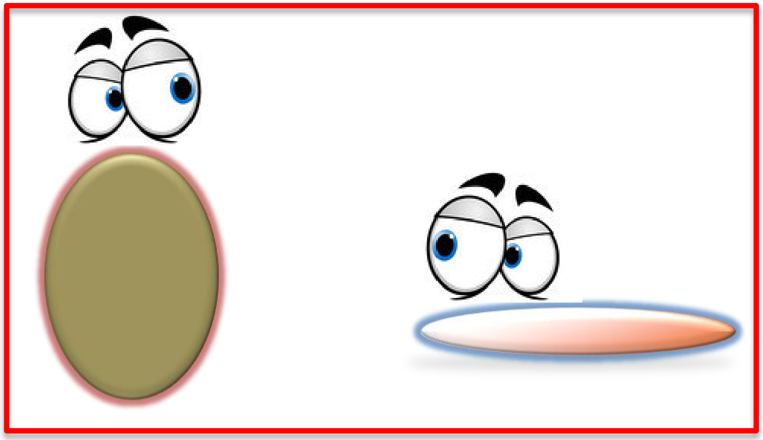
directed network

Huynh-Thu et al. (2010) “Inferring regulatory networks from expression data using tree-based methods”, PLoS one 5, e12776.

Goal: Infer network of microbial relationships

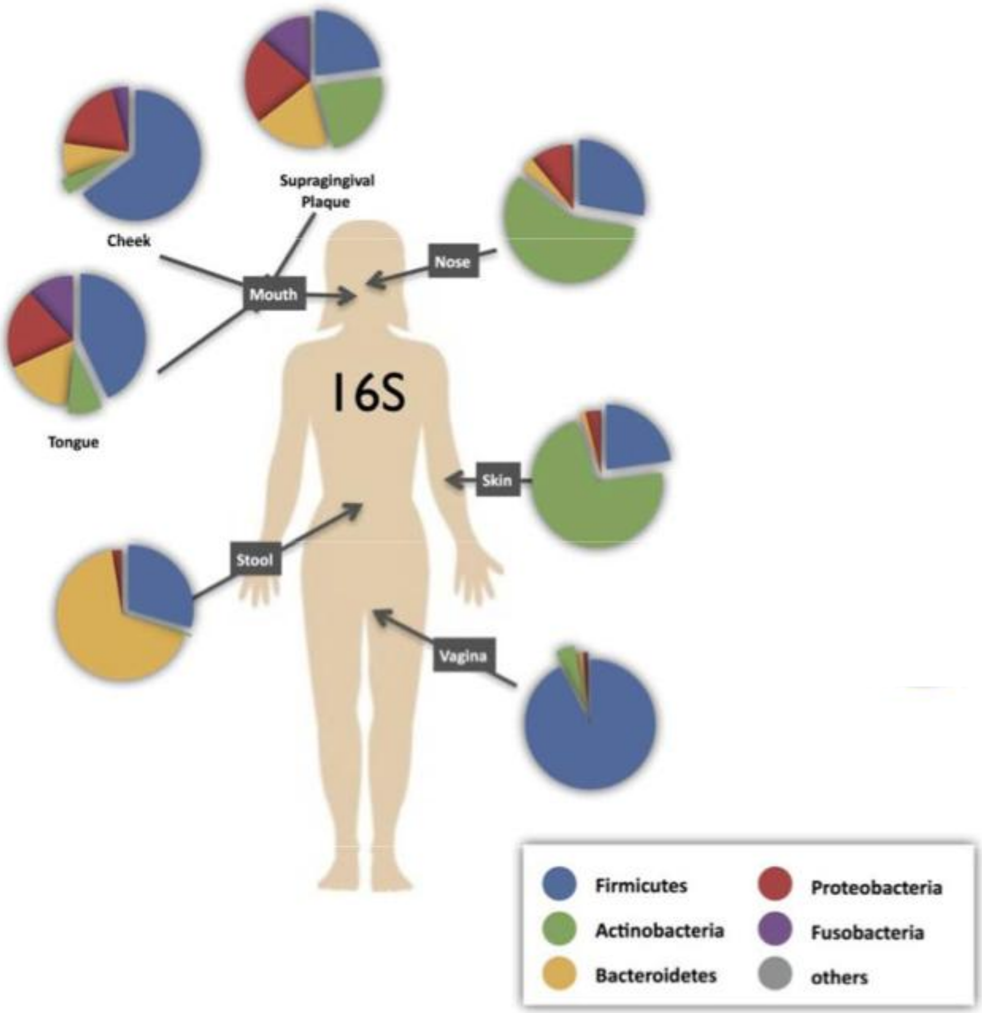
- several recent metagenomic data sets measure microbial abundance across a large number of samples
- network inference techniques can identify significant relationships between microorganisms from these data
- significant co-presence (co-occurrence of two microbes across samples) can be interpreted as niche overlap, mutualism, commensalism etc.
- significant mutual exclusion (avoidance of two microbes across samples) can be interpreted as alternative niche preference, competition, amensalism etc.





The Human Microbiome Project

- 18 body sites (15 sites in males)
- 242 healthy individuals sampled up to three times
- 5,177 samples 16S RNA-sequenced
- > 3.5 TB metagenomic sequences
- Metadata collected (sex, age, ethnicity, BMI, pulse, medication, smoking behavior, vaginal pH, etc.)



distribution of phyla across human body sites, according to 16S sequencing

The Human Microbiome Project Consortium (2012) “A framework for human microbiome research”, Nature 486, pp. 215-221.

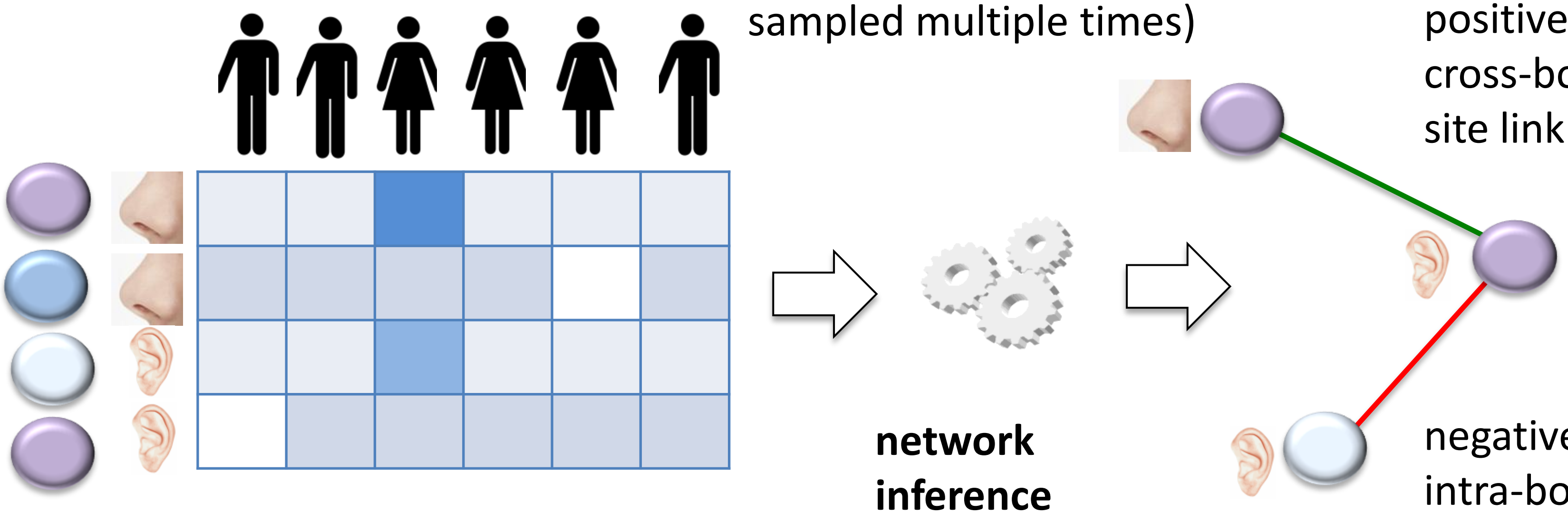
16S sequencing and processing

- 5,177 samples pyro-sequenced (454 GS FLX Titanium) in 4 different centers (for V1-V3, **V3-V5** and V6-V9 regions of 16S rRNA)
- 16S rRNA sequencing benchmarked on **mock communities** of known composition
- raw 16S rRNA reads were processed with mothur and Qiime pipelines
- mothur assigned reads to ~730 **phylotypes** and to ~9,450 OTUs (operational taxonomic units) using the RDP (Ribosomal Database Project) phylogenetic tree
- likely mislabeled samples removed using a machine learning approach (Knights, 2010)

Human Microbiome Project Data Generation Working Group (2012) “Evaluation of 16S rDNA-Based Community Profiling for Human Microbiome Research” PLoS ONE 7(6) e39315.
Schloss, P. et al. (2009) “Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.” Appl. Environ. Microbiol. 75, pp. 7537-7541.
Jumpstart Consortium Human Microbiome Project Data Generation Working Group “Evaluation of 16S rDNA-based Community Profiling for Human Microbiome Research”, PLoS one 7, e39315.
Cole, J.R. et al. (2009) “The Ribosomal Database Project: improved alignments and new tools for rRNA analysis”, Nucleic Acid Research 37, pp. D141-D145.
Knights, R. et al. (2010) “Supervised classification of microbiota mitigates mislabeling errors.” ISME 5, pp. 570-573.

Network inference from HMP data - Overview

- apply network inference strategies to predict relationships between bacterial taxa from the 16S HMP V35 phylotype data set (**genus** level)



... (392 columns, subjects sampled multiple times)

... (12,450 rows, taxa in body sites)

network inference

positive cross-body-site link

negative intra-body-site link

🟦 high abundance

⬜ low abundance

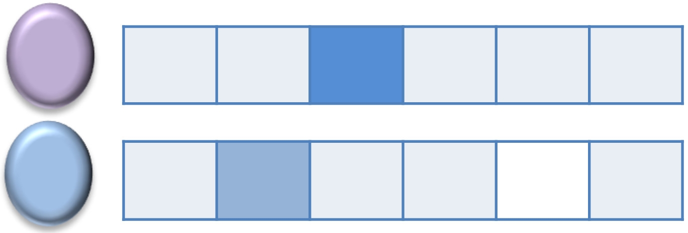
count matrix

network

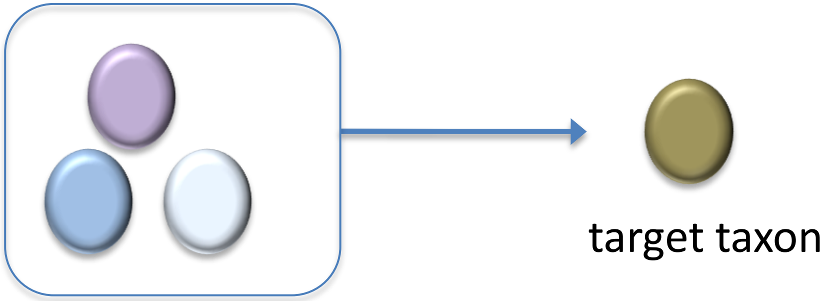
joint work with Huttenhower lab

4. Methods

Assessing strength of relationships between microorganisms



abundance profiles across samples



target taxon

source taxa

Pair-wise relationships (similarity)

- Pearson correlation
- Spearman correlation
- Kullback-Leibler dissimilarity (KLD)
- Bray Curtis dissimilarity (BC)

Complex relationships (sparse regression)

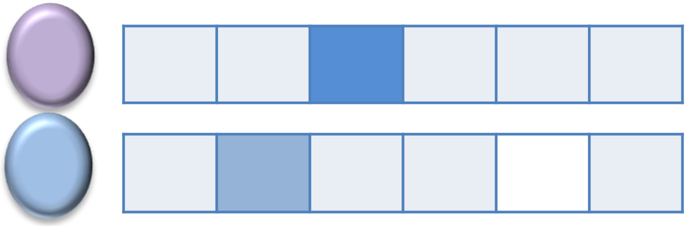
- GLBM (generalized, linear boosted models) to predict a target taxon from a set of source taxa by regression
- score: the goodness of fit (how well combined source taxa profiles predict target taxon profile)

Fah Sathirapongsasuti and Curtis Huttenhower

4. Methods

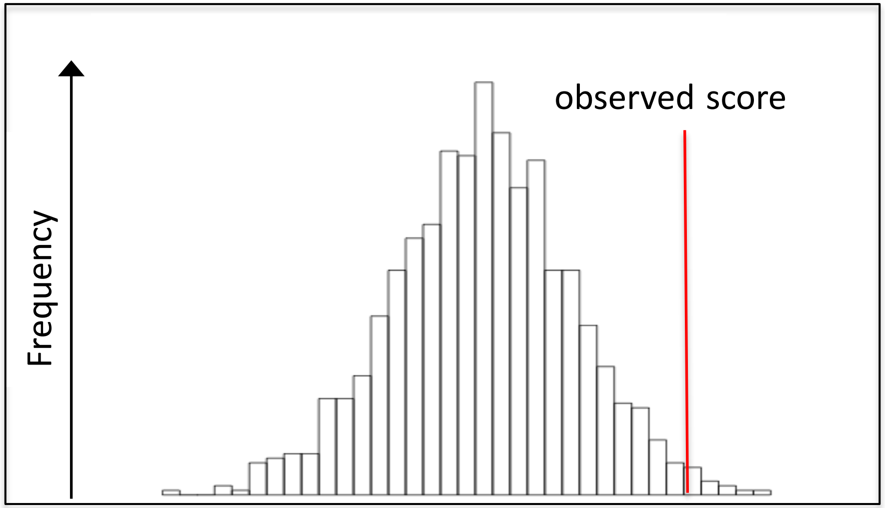
Computing significance of relationships I

- for each of the five methods (Pearson, Spearman, Kullback-Leibler, Bray Curtis, GLBM), compute permutation and bootstrap edge scores

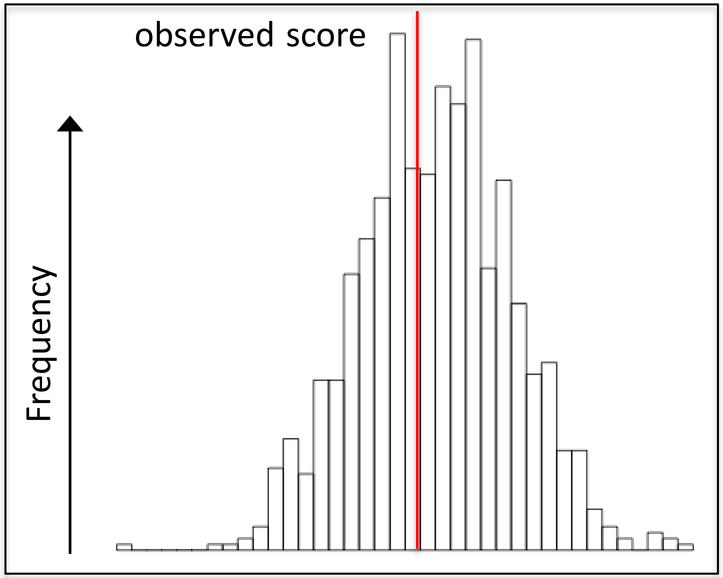


5 scores per edge, for each score:

permutation (null) distribution of method-specific edge score



bootstrap distribution of method-specific edge score (confidence interval)

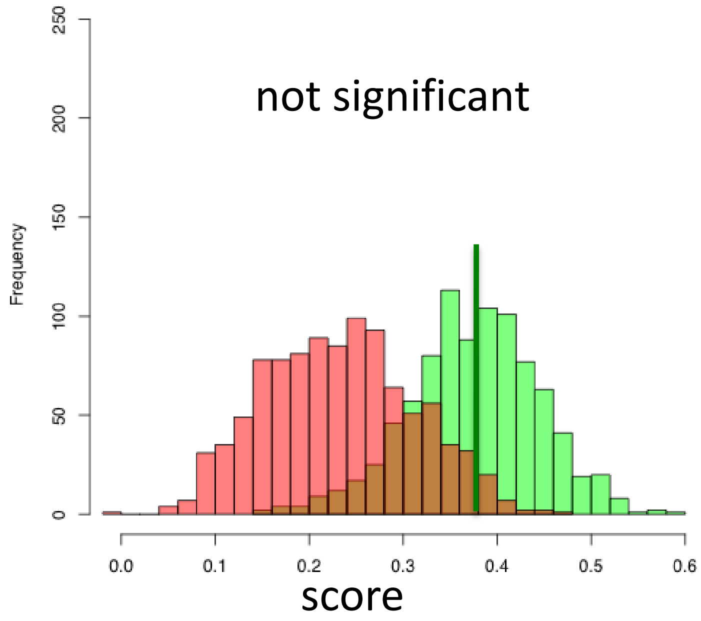


Computing significance of relationships II

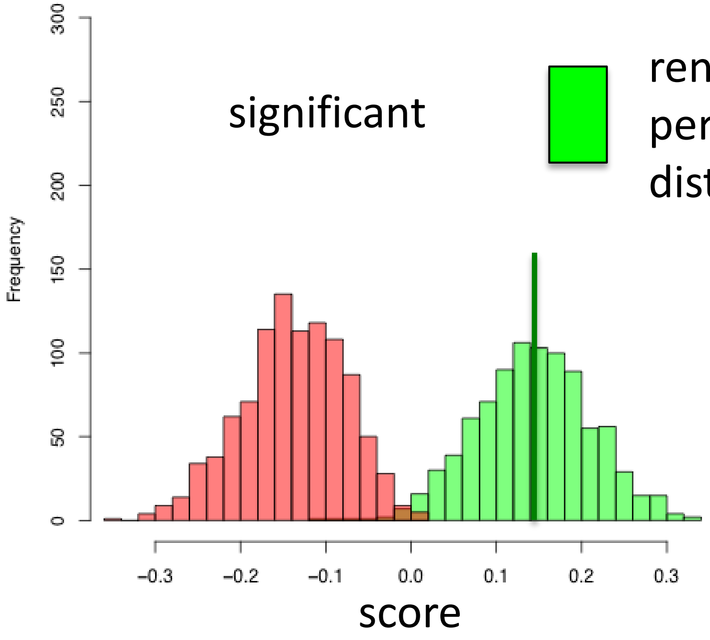
Edge- and method-specific p-value is computed with a **Z-test** (p-value of the null distribution mean given the bootstrap distribution, assuming normality for the bootstrap distribution)

🟥 bootstrap distribution

🟩 renormalized permutation distribution



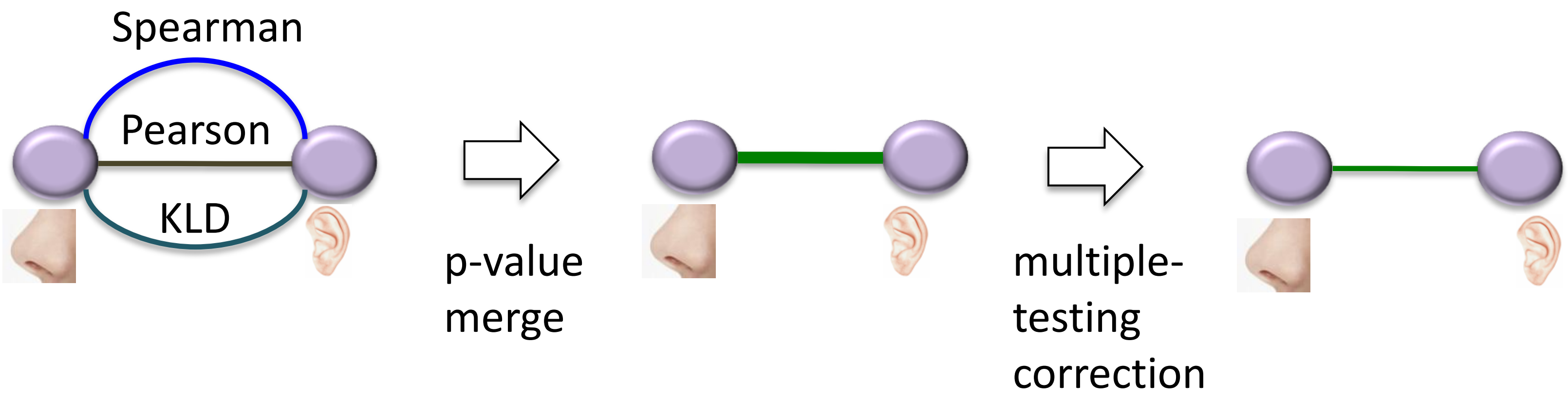
Fusobacteriales versus Streptococcaceae in buccal mucosa (Pearson)



Actinobacteria versus Bacteroidetes in subgingival plaque (Spearman)

4. Methods

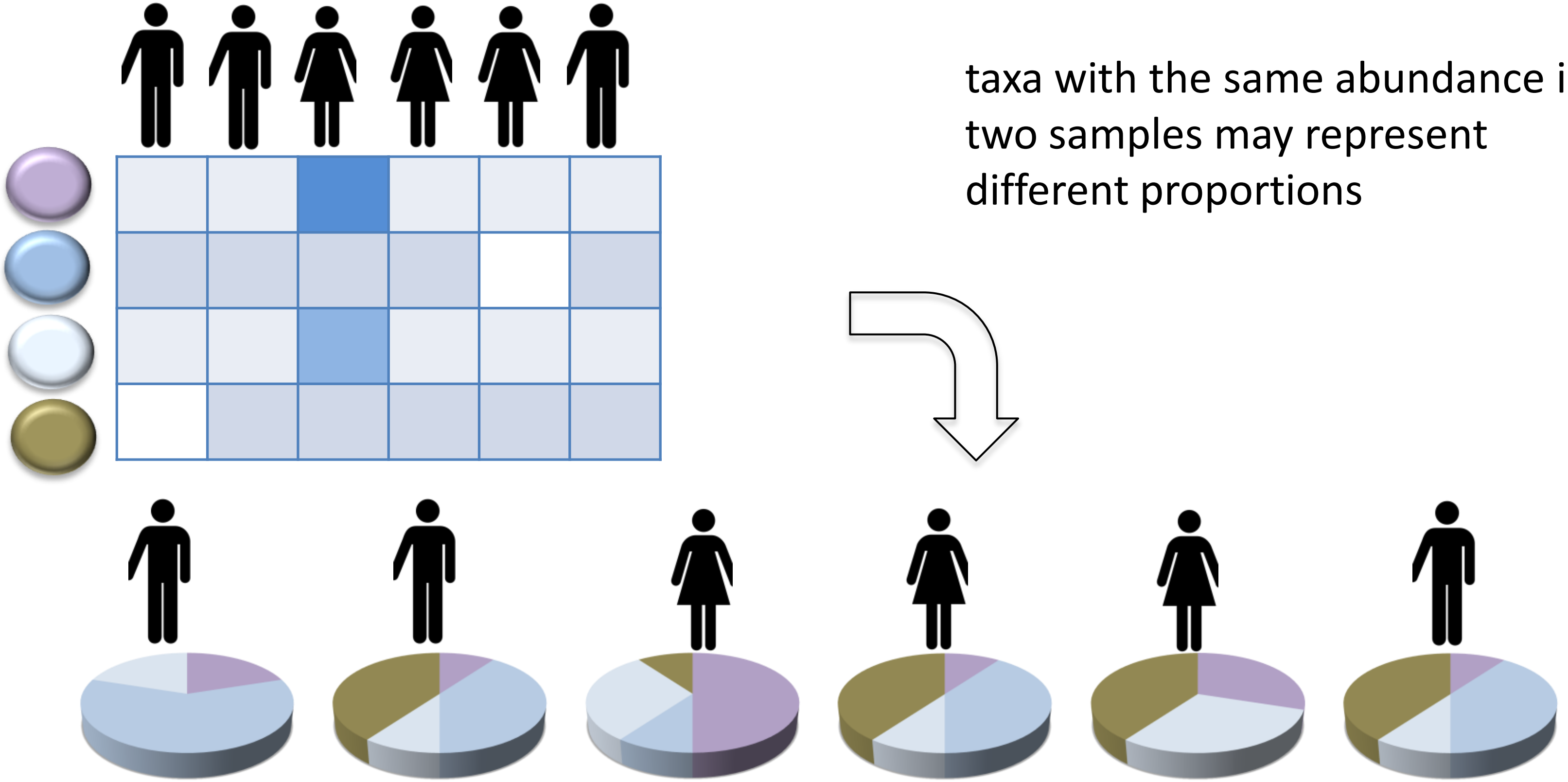
Network building



- merge method-specific p-values using Sime’s method
- apply Benjamini-Hochberg Yekutieli False Discovery Rate correction on merged p-values
- after correction, remove all p-values above the threshold (set to 0.05)
- represent remaining relationships as a network

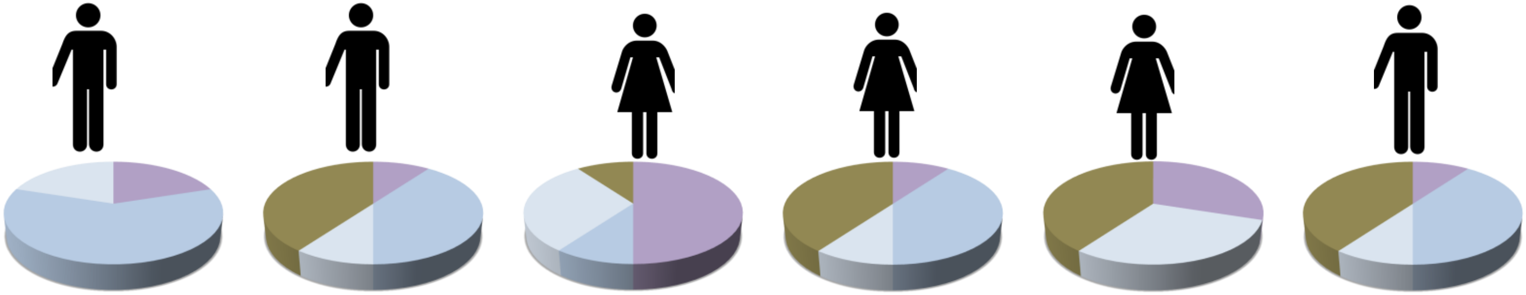
Problem: Data normalization and compositionality

- technical errors/differences in processing lead to different total abundances across samples
- sample-wise normalization necessary (i.e. division of abundances in a sample by this sample's total abundance sum)
- absolute abundances are converted into proportions



taxa with the same abundance in two samples may represent different proportions

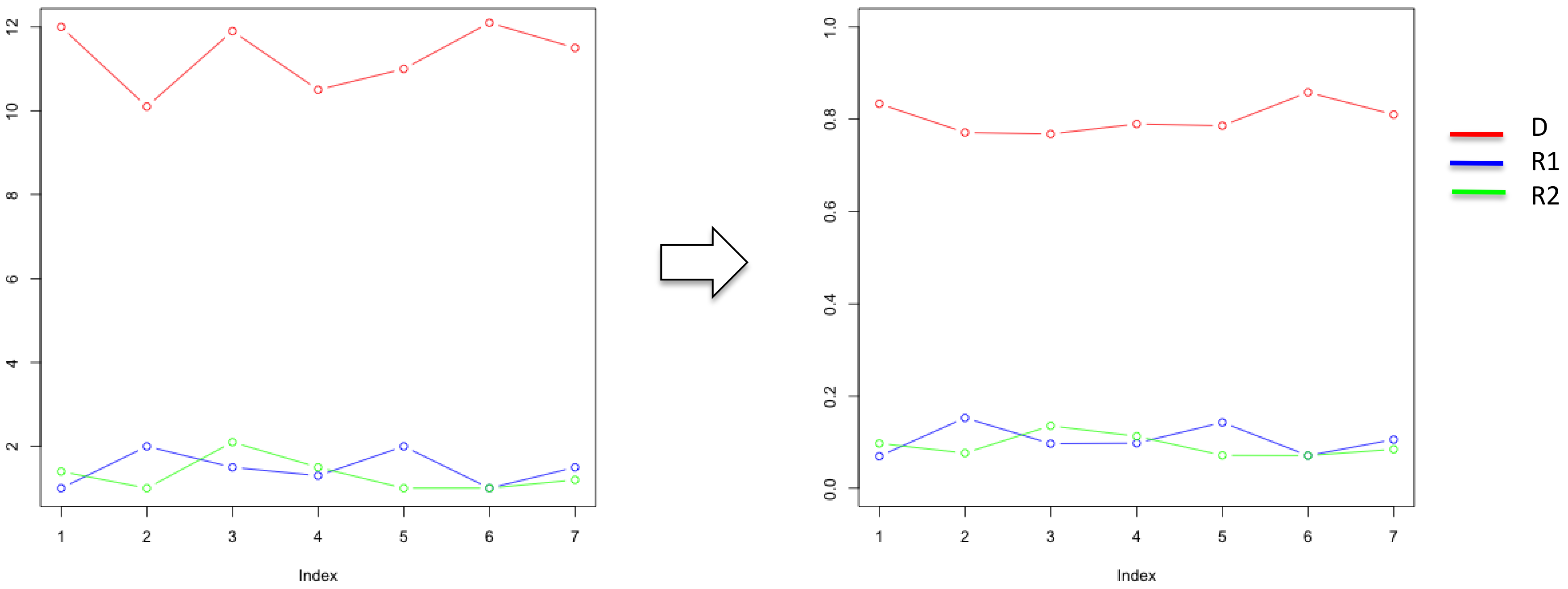
Problem: Data normalization and compositionality



- Pearson and Spearman can be severely distorted, because they consider “absolute” values
- measures based on ratios or log-ratios (KLD, BC) are not affected by data compositionality, since the ratio between two abundances in the same sample is not changed by the normalization

Aitchison J (1982) “The Statistical Analysis of Compositional Data.” Journal of the Royal Statistical Society Series B (Methodological) 44, pp. 139-177.

Data normalization and compositionality - Example


raw data

normalized data

	R1	R2	D
R1	1	-0.24	-0.69
R2		1	0.31
D			1

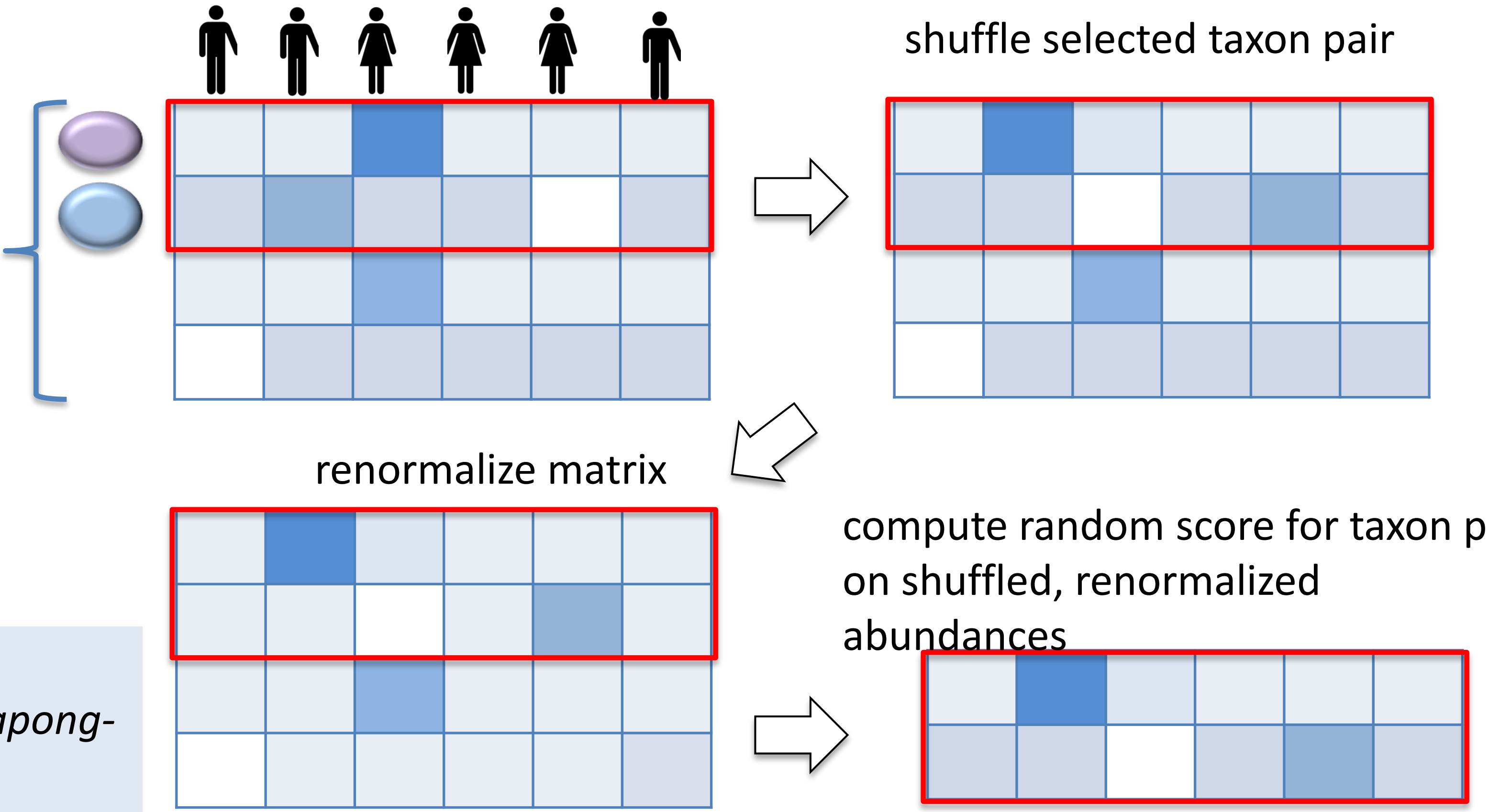
	R1	R2	D
R1	1	-0.32	-0.73
R2		1	-0.41
D			1

Pearson correlation

Adjust null distribution to mitigate the compositionality bias

- Permutation test: removes correlation, but also any bias due to compositionality
- Permutation with **renormalization**: for each pair of taxa, permute their abundances and then normalize the matrix (body-site-wise)

all taxa in one body site



shuffle selected taxon pair

renormalize matrix

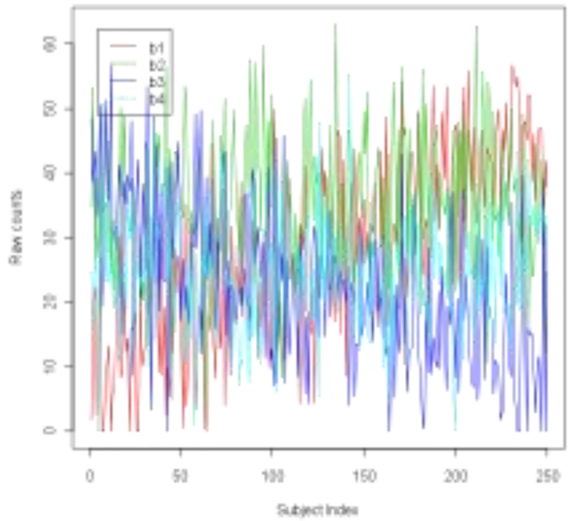
compute random score for taxon pair on shuffled, renormalized abundances

Fah Sathirapong-sasuti

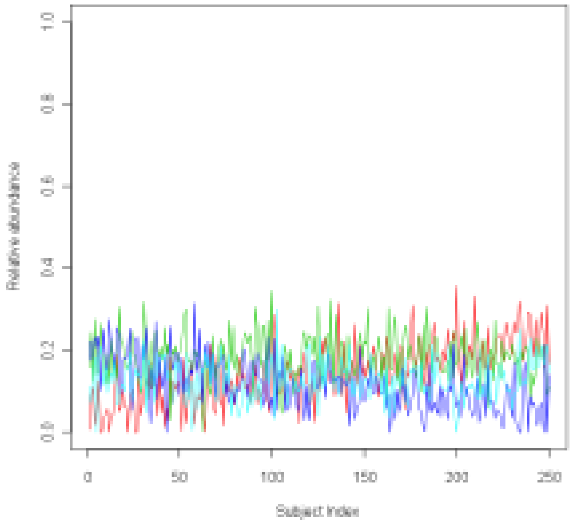
4. Methods

Renormalization mitigates compositionality bias

raw data



normalized data



true correlation between b1 and b3

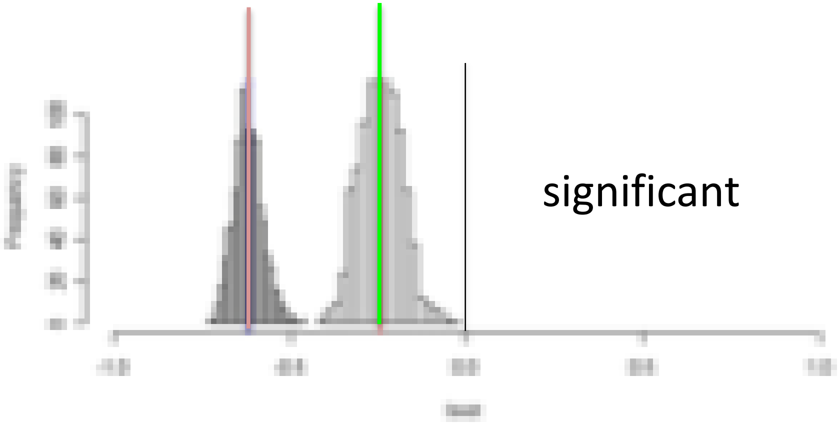
spurious correlation between b2 and b4 introduced by normalization

b1-b3

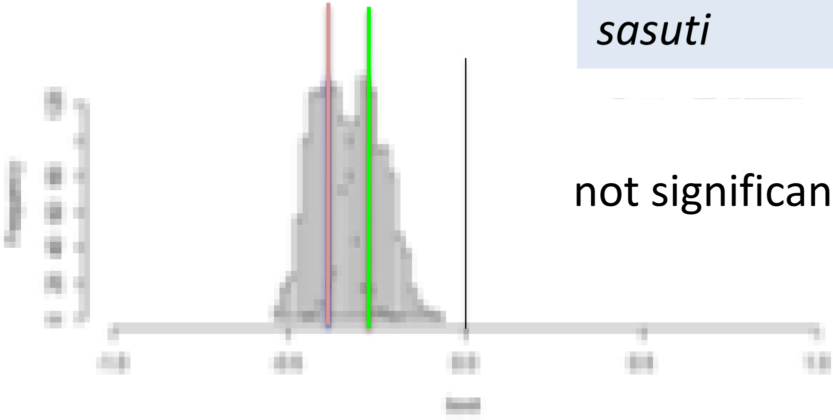
🟥 bootstrap distribution mean
🟩 renormalized permutation distribution mean

b2-b4

Fah Sathirapong-sasuti

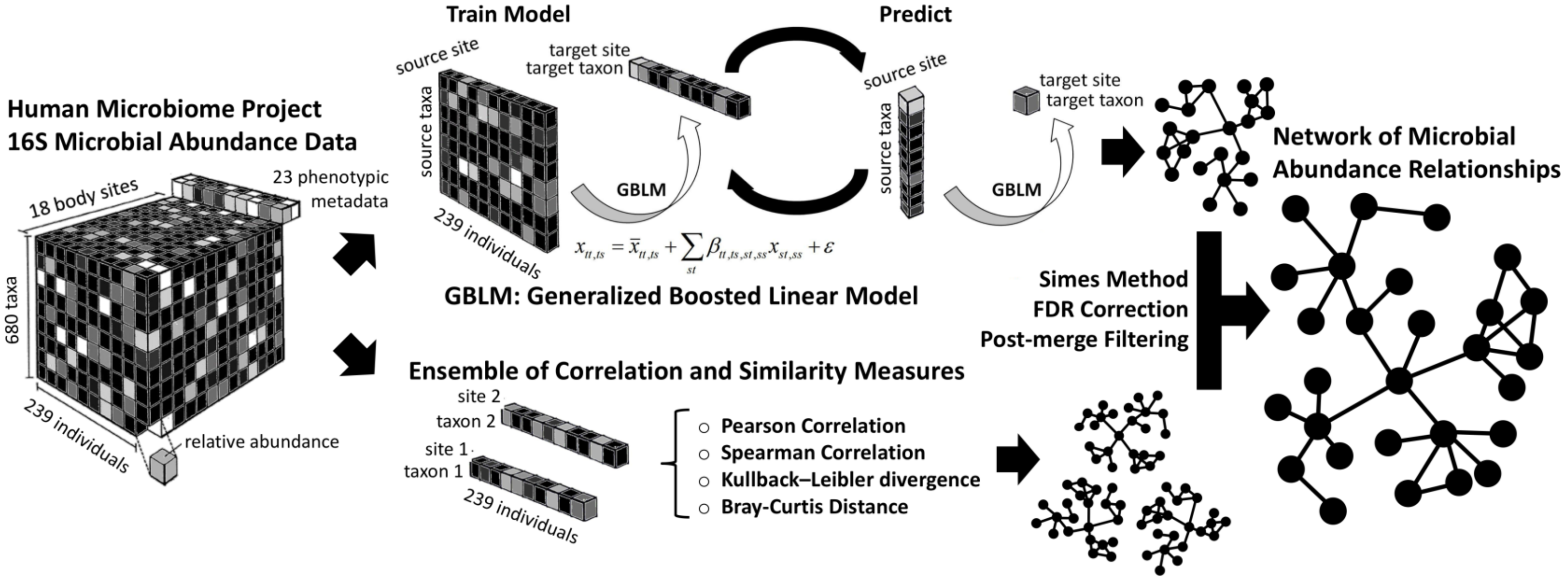


significant



not significant

Methodology overview

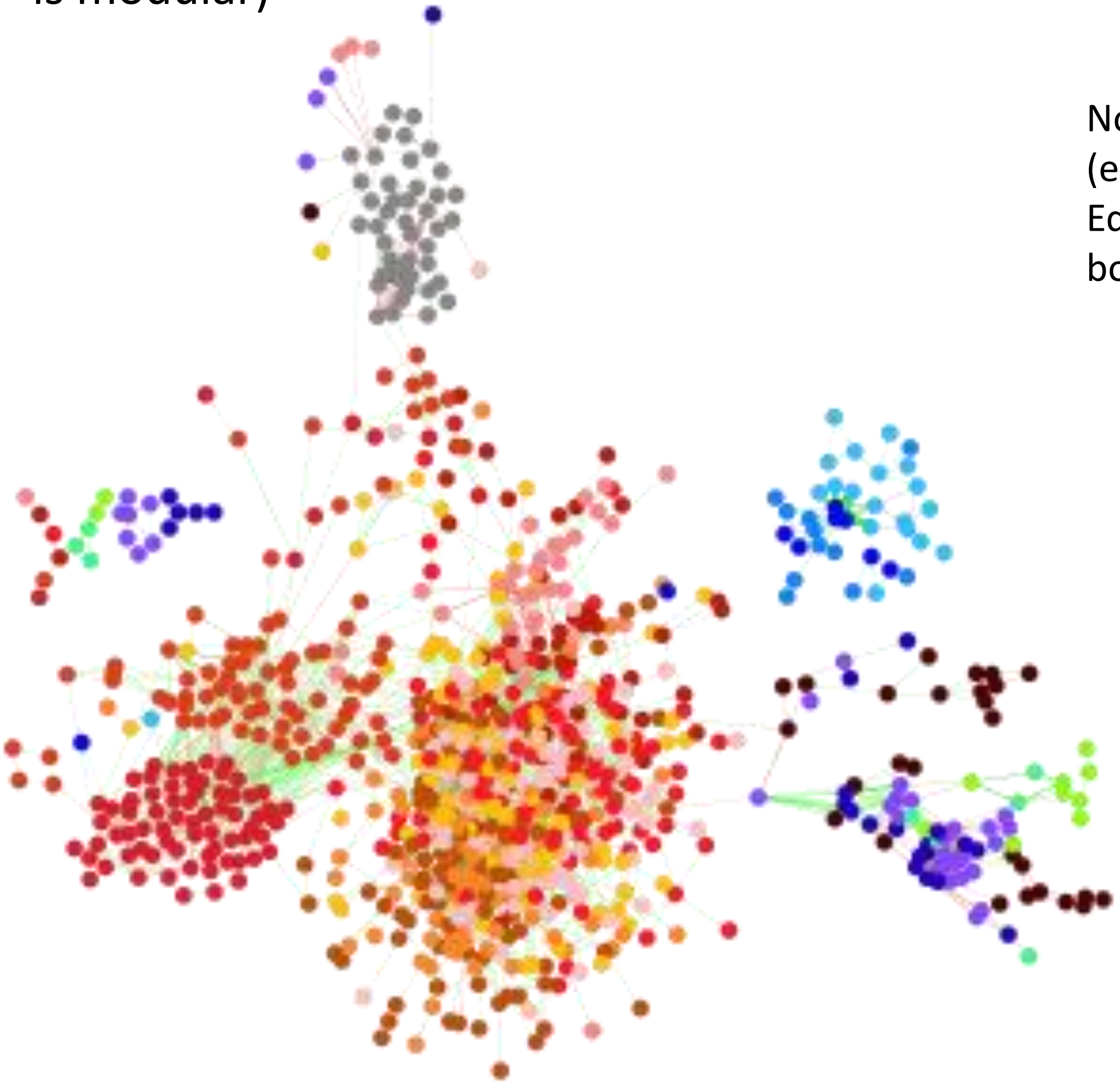


4. Methods

Network inferred for HMP 16S phylotypes

- most edges connect phylotypes within the same body area (e.g. vagina), but some edges link phylotypes across body areas (network is modular)

Nodes: body-site-specific phylotypes (e.g. *Ruminococcaceae* in Stool)
Edges: significant score between body-site-specific phylotypes



Node color code

Anterior nares

Buccal mucosa
Hard palate
Keratinized gingiva
Palatine tonsils
Saliva
Subgingival plaque
Supragingival plaque
Throat
Tongue dorsum

Left retroauricular crease
Right retroauricuar crease

Left antecubital fossa
Right antecubital fossa

Stool

Mid vagina
Posterior fornix
Vaginal introitus

Edge color code

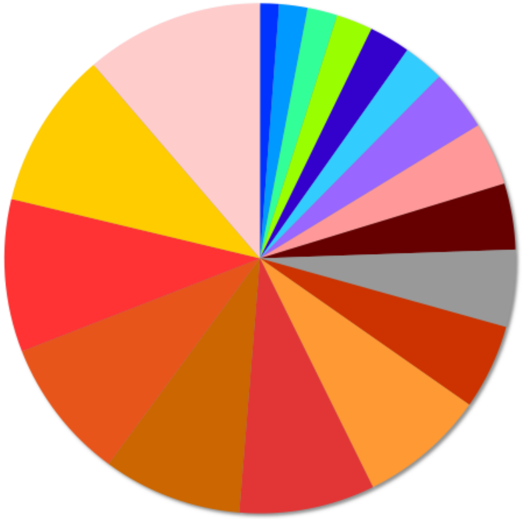
positive

negative

5. Results

HMP 16S Network - composition

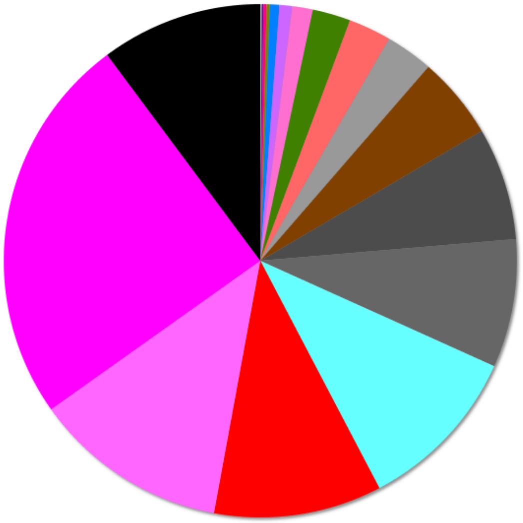
Body-site-specific node proportions



Posterior fornix
Mid vagina
Right antecubital fossa
Left antecubital fossa
Right retroauricuar crease
Vaginal introitus
Left retroauricular crease
Keratinized gingiva
Anterior nares
Stool

Buccal mucosa
Throat
Subgingival plaque
Palatine tonsils
Supragingival plaque
Hard palate
Saliva
Tongue dorsum

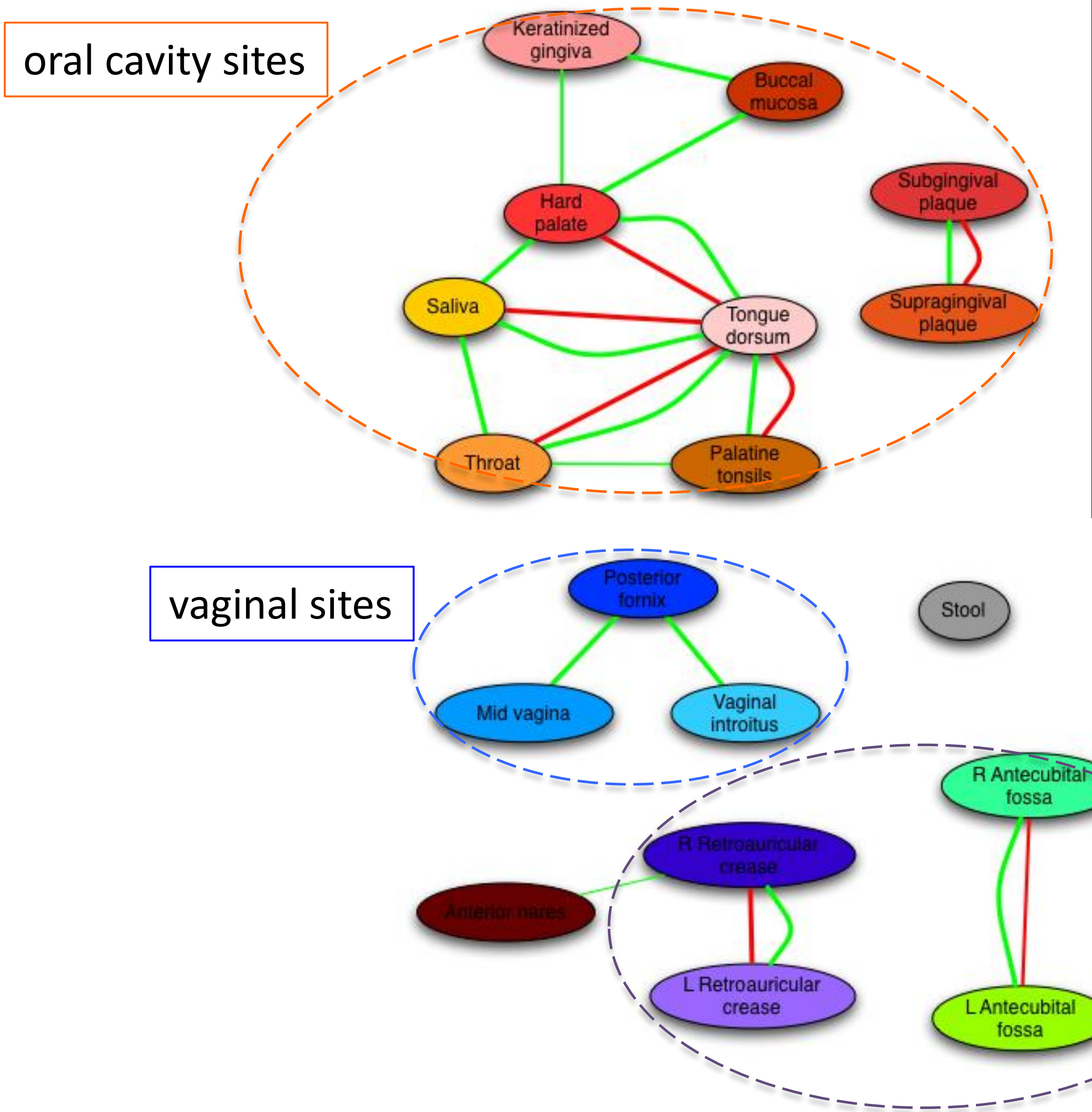
Class-specific node proportions



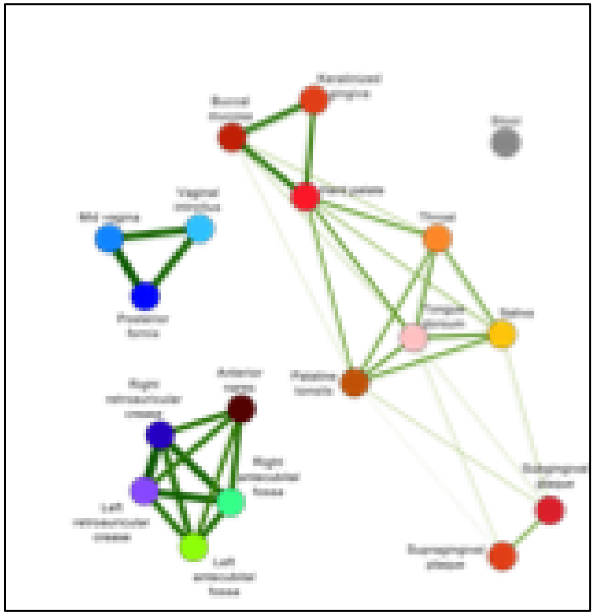
Alphaproteobacteria
Lentisphaeria
Verrucomicrobiae
Synergistia
Mollicutes
Negativicutes
Erysipelotrichi
Spriochaetes
Flavobacteria

Epsilonproteobacteria
Fusobacteria
Gammaproteobacteria
Betaproteobacteria
Actinobacteria
Bacteroidia
Bacilli
Clostridia
Above class-level

HMP 16S network – body-site relationships

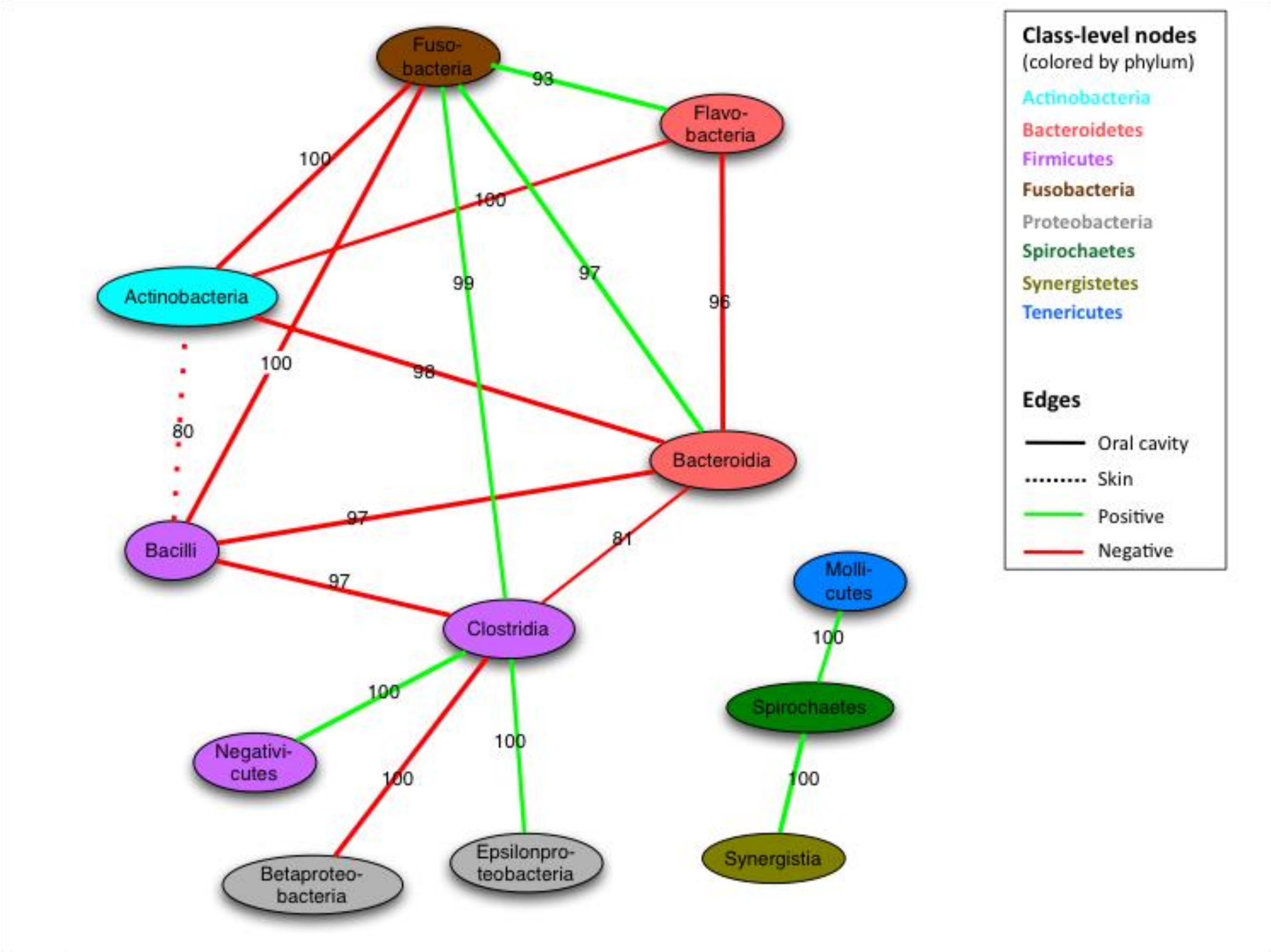


skin sites

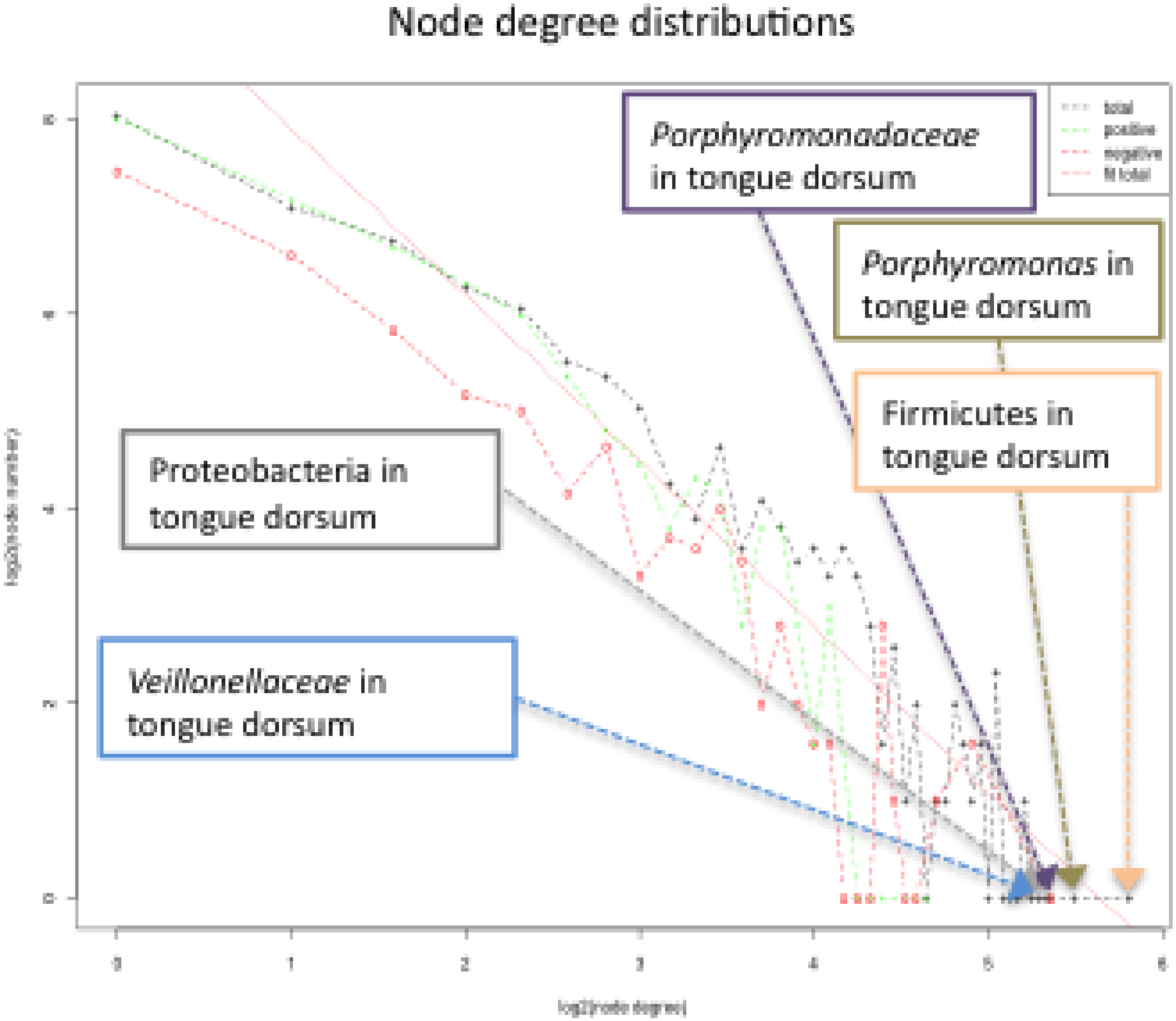


weighted 16S UniFrac beta diversity (Huttenhower et al., Nature 486, 207-214)

HMP 16S network – class relationships



HMP 16S network analysis



Porphyromonadaceae in tongue dorsum

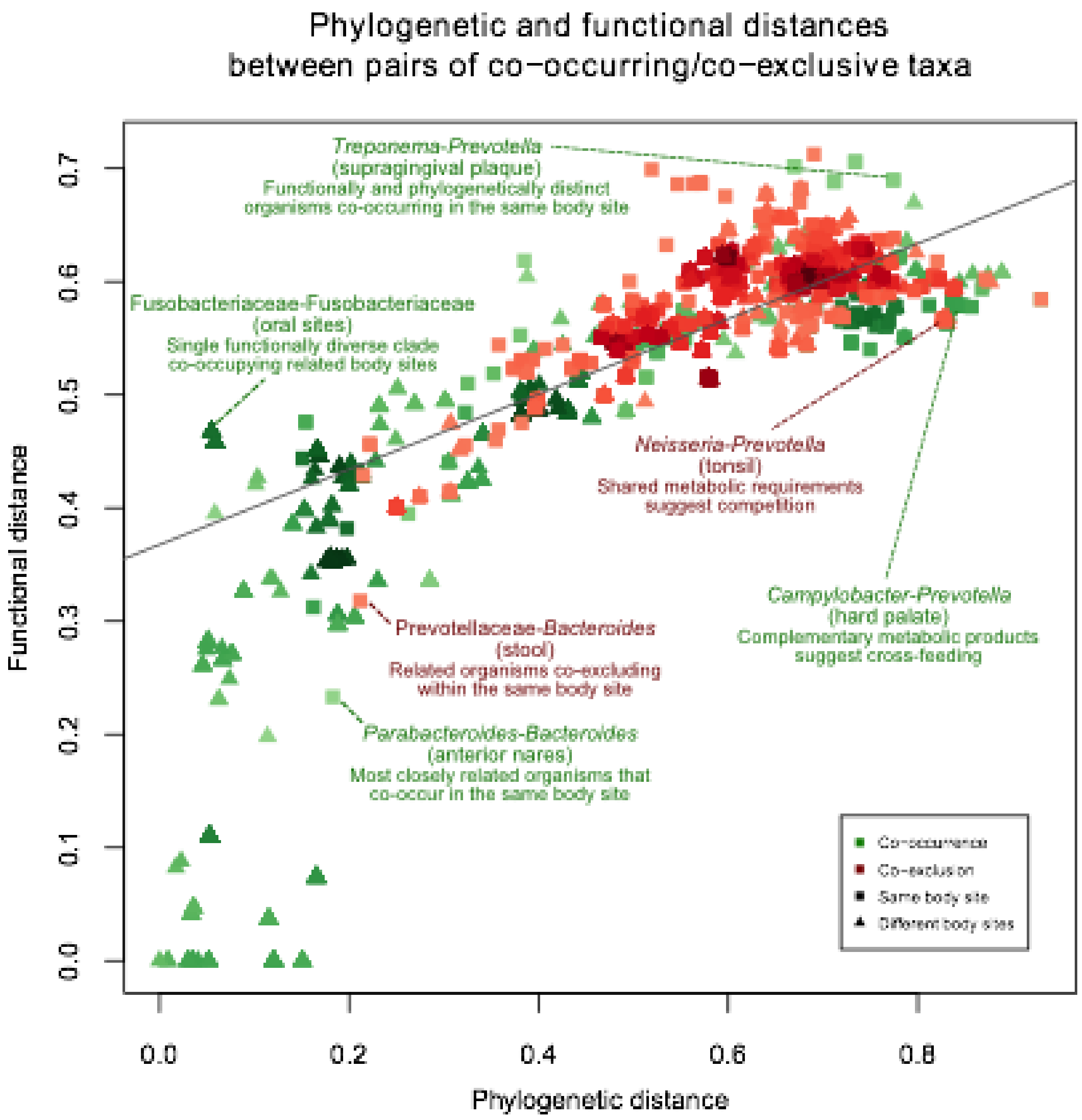
Porphyromonas in tongue dorsum

Firmicutes in tongue dorsum

Proteobacteria in tongue dorsum

Veillonellaceae in tongue dorsum

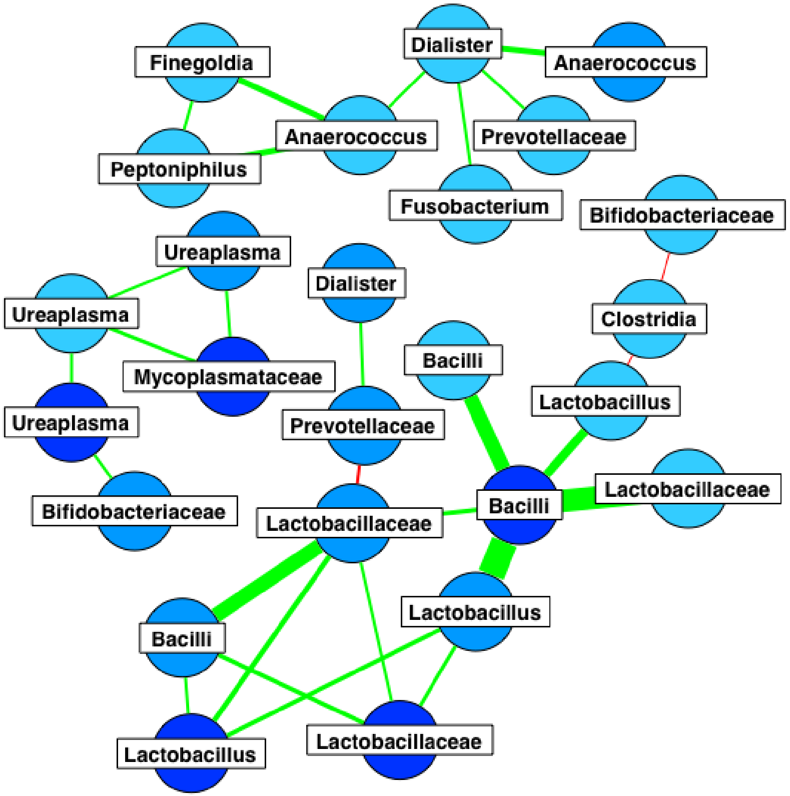
HMP 16S network functional analysis



Fah Sathirapongsasuti and Nicola Segata

Vaginal sub-network of HMP 16S network

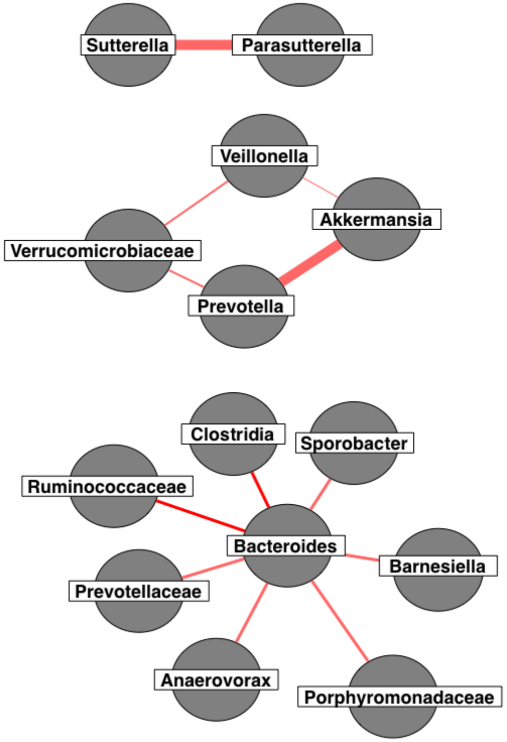
- Ravel et al. (2011): 5 vaginal community types identified
- 4 (I, II, III and V) of these dominated by *Lactobacillus* species
- 1 (IV) is diverse and contains members of Actinobacteria, Bacteroidetes and other phyla
- exclusion between *Prevotellaceae* (Bacteroidetes) and *Lactobacillaceae* as well as co-occurrence of anaerobic taxa (*Finegoldia*, *Dialister*, *Peptoniphilus,* *Prevotellaceae*), which are members of community IV



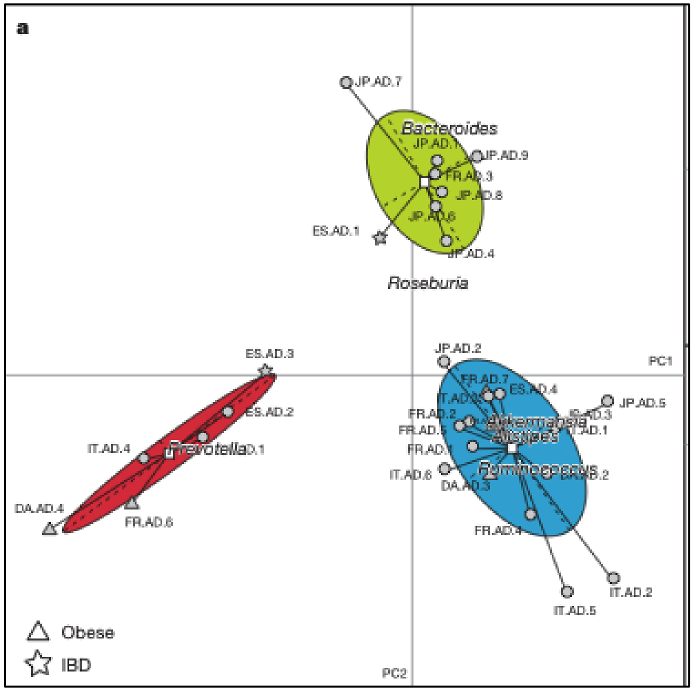
taxonomic levels shown: genus, family and class

Ravel, J. et al. (2011) “Vaginal microbiome of reproductive-age women”, PNAS, vol. 108, pp. 4680-4687.

Stool sub-network of HMP 16S network



taxonomic levels shown: genus, family and class

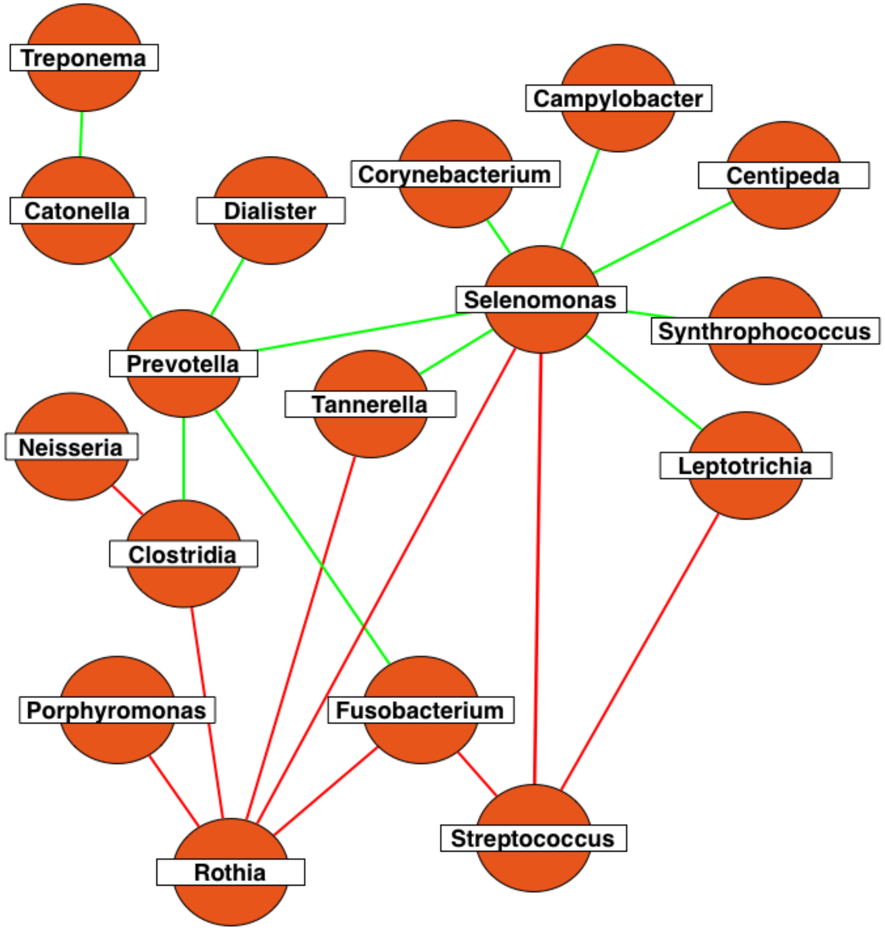


- Arumugam et al. (2011): three different gut communities identified
- driven by: *Prevotella*, *Bacteroides* (both *Bacteroidetes*) and *Ruminococcus* (*Firmicutes*)
- *Ruminococcaceae* and *Bacteroides* as well as *Prevotellaceae and Bacteroides* exclude each other in the stool sub-network

Arumugam, M., Raes, J. et al. (2011) “Enterotypes of the human gut microbiome”, Nature 473, pp. 174-180.

Supragingival plaque sub-network of HMP 16S network

- negative relationship between early colonizers of the tooth surface (*Streptococcaceae*) and intermediate colonizers (*Fusobacterium*)
- positive relationships between late colonizers (Selenomonas, Tannerella)



taxonomic levels shown: genus

Kolenbrander, P.E. et al. (2010) “Oral multispecies biofilm development and the key role of cell-cell distance”, Nature Reviews Microbiology 8, pp. 471-480.

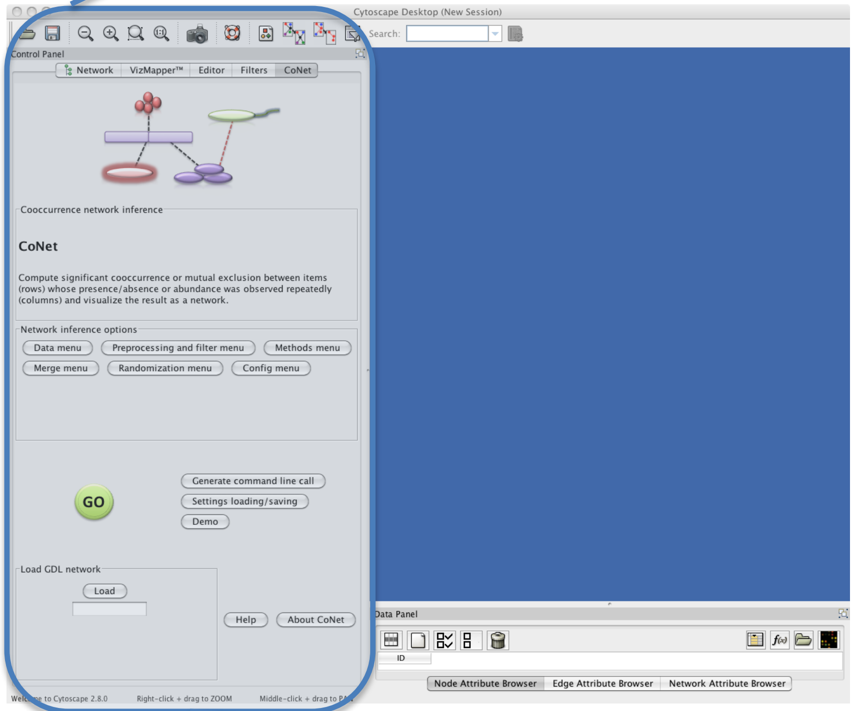
Conclusions

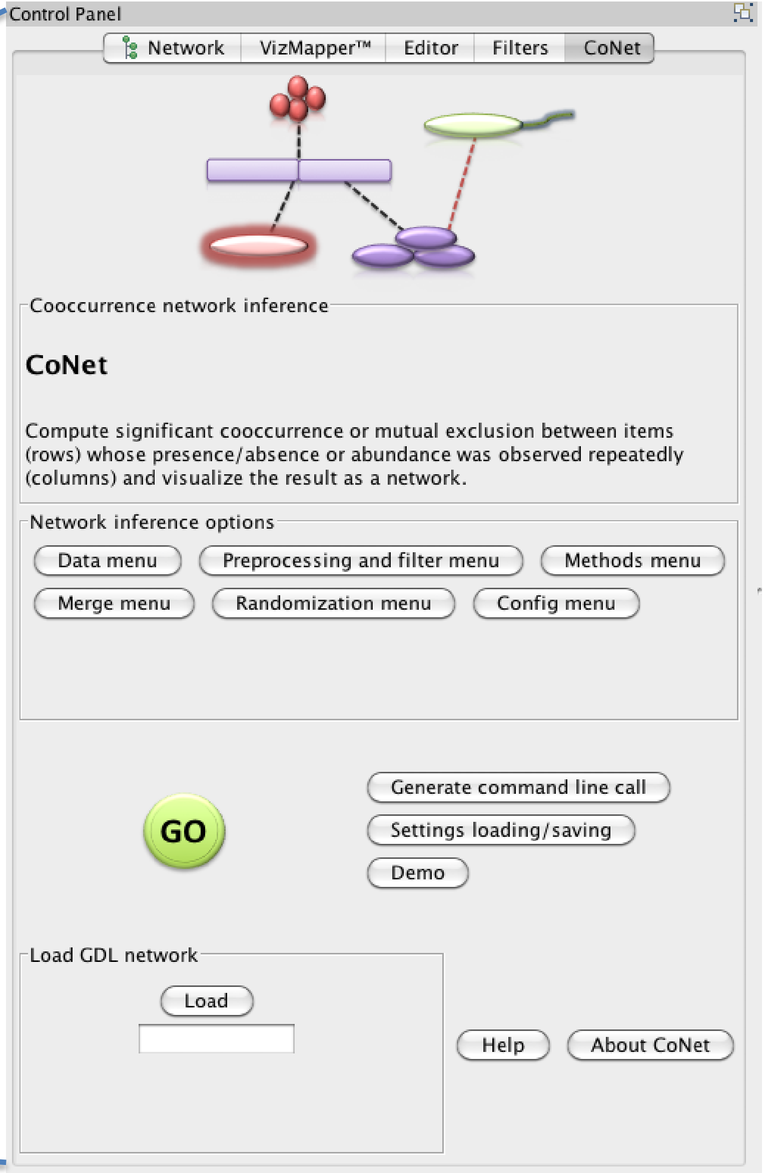
- few cross-body-area relationships: different body areas harbor distinct microbiota
- body sites can be grouped based on cross-links between their microbiota): oral, skin and vaginal sites form separate clusters, airways and stool separated from the oral cavity: clusters can be interpreted as different microbial niches
- alternative microbial communities observed in the vagina and the gut detected
- stages of dental plaque formation captured
- closely related microbes tend to co-occur in body sites with similar conditions
- negative relationships occur between more distantly related microbes

Sathirapongsasuti*, Faust* et al. (2012) “Microbial Co-occurrence Relationships in the Human Microbiome”, PLoS Computational Biology 8 (7) e1002606.

CoNet – Similarity-based network inference with multiple measures

Cytoscape main window


Cytoscape Desktop (New Session)
Search:
Control Panel
Network
VizMapper™
Editor
Filters
CoNet
Data Panel
ID
Node Attribute Browser
Edge Attribute Browser
Network Attribute Browser
Welcome to Cytoscape 2.8.0
Right-click + drag to ZOOM
Middle-click + drag to PAN



Cooccurrence network inference

CoNet

Compute significant cooccurrence or mutual exclusion between items (rows) whose presence/absence or abundance was observed repeatedly (columns) and visualize the result as a network.

Network inference options

Load GDL network

7. Tool

CoNet – Features

http://systemsbiology.vub.ac.be/conet

- runs as Cytoscape plugin or on command line
- allows combining several measures, either in a multigraph or by merging their scores or p-values
- supports abundance as well as for presence/absence matrices
- implements various randomization and multiple test correction routines
- integrates external network inference packages, e.g. minet (mutual information based network inference) and apriori (association rule mining algorithm)
- plots score distributions
- offers preprocessing, missing value treatment, grouping rows
- settings loading/saving
- well documented (manual, tutorials, FAQ)

7. Tool

Outlook

- Dynamic network inference to decipher relationships among microorganisms in recent metagenomic time series data

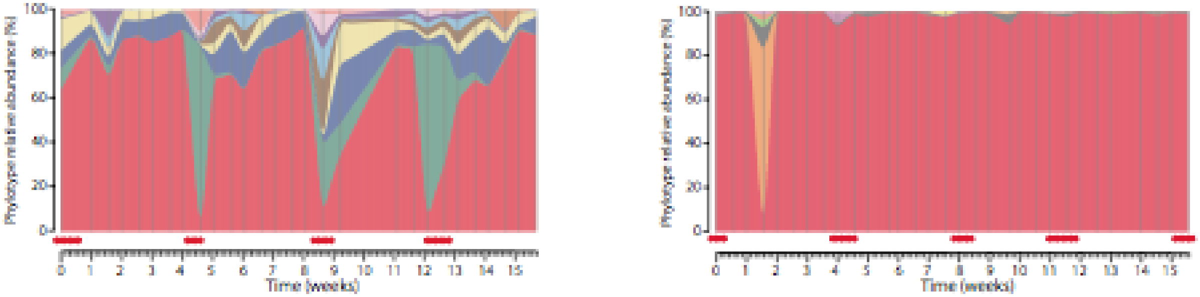
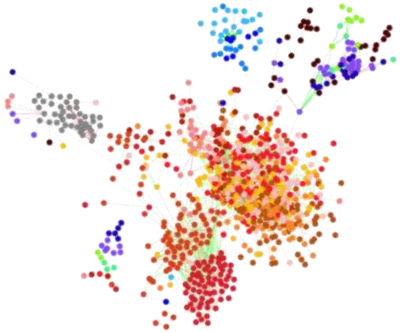


image taken from Gajer et al. (2012) Sci. Transl. Med 4, 132ra52



Acknowledgement

Bioinformatics and (Eco-)Systems Biology (BSB) lab

The Huttenhower Lab
Department of Biostatistics, Harvard School of Public Health

Curtis Huttenhower

Fah Sathirapongsasuti

Nicola Segata

HMP

HMP Consortium for data access

...and Dirk Gevers, Jacques Izard, Alvin Lo and Dominique Maes

Funding

FWO
VLAANDEREN

ODYSSEUS
Top science in Flanders
Nieuwe oproep Odysseus

Bacterial abundances from 16S reads

- raw 16S rRNA reads were processed by Pat Schloss with his **mothur** pipeline

- processing steps included sequence trimming (primers and barcodes removal), filtering (of ambiguous bases, homo-polymers and redundant sequences) and chimera removal (with ChimeraSlayer)

- mothur assigned reads to ~730 phylotypes using the Ribosomal Database Project (RDP) reference 16S rRNA sequences and the RDP phylogenetic tree

- mothur also assigned reads to ~9,450 OTUs (operational taxonomic units), by first clustering reads based on alignments and then assigning a consensus taxonomy to the groups using the RDP phylogenetic tree and reference sequences

- likely mislabeled samples were detected by Dirk Gevers using a machine learning approach (Knights, 2010)

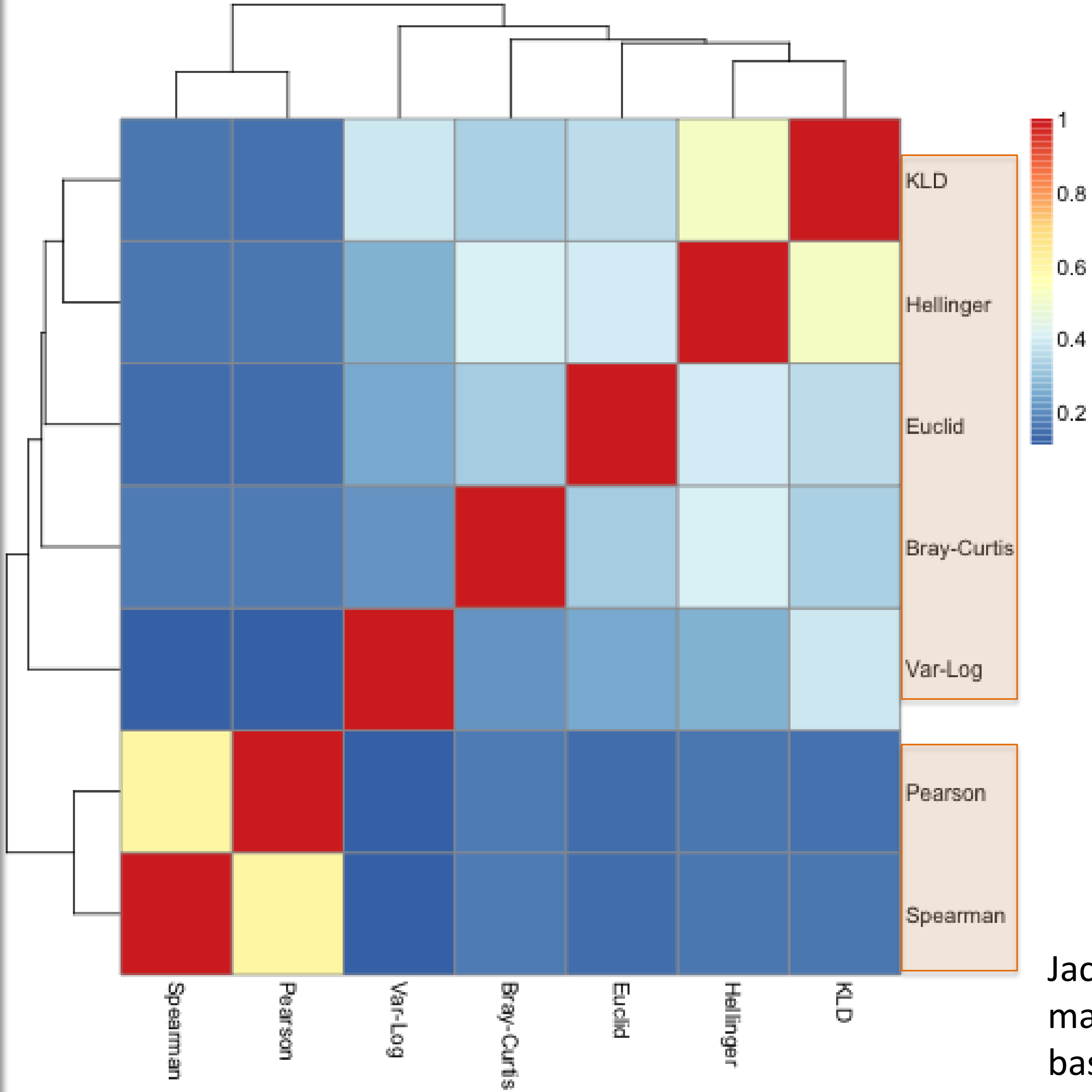
Appendix

Schloss, P. et al. (2009) “Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities.” Appl. Environ. Microbiol., vol. 75, pp. 7537-7541

Cole, J.R. et al. (2009) “The Ribosomal Database Project: improved alignments and new tools for rRNA analysis”, Nucleic Acid Research, vol. 37, pp. D141-D145

Knights, R. et al. (2010) “Supervised classification of microbiota mitigates mislabeling errors.” ISME, vol. 5, pp. 570-573

Selection of measures



Experiment: Select 1,000 top-ranked and 1,000 bottom-ranked measure-specific edges in Houston data subset

Jaccard similarity heat map (Ward clustering) based on edge overlap

Definition of measures

Hellinger
(*x* and *y each* sum up to 1)

$$d(x,y)=\sqrt{\sum\left(\sqrt{x_i}-\sqrt{y_i}\right)^2}$$

Kullback-Leibler
(*x* and *y* each sum up to 1)

$$d(x,y)=\sum\left(x_i\log\left(\frac{x_i}{y_i}\right)+y_i\log\left(\frac{y_i}{x_i}\right)\right)$$

Logged Euclidean

$$d(x,y)=\sqrt{\sum\left(\log(x_i)-\log(y_i)\right)^2}$$

Recommended for compositional data (absolute values are not of interest)

Require pseudo-counts or smoothing because log(0) = -Inf

Hellinger distance and Kullback-Leibler divergence are mathematically related measures.

Euclidean distance

$$d(x,y)=\sqrt{\sum\left(x_i-y_i\right)^2}$$

Bray Curtis
(Steinhaus is the corresponding similarity)

$$d(x,y)=1-\frac{2\sum\min(x_i,y_i)}{\sum x_i+\sum y_i}$$

Recommended for taxon abundance data

Bray-Curtis dissimilarity is computed on row-wise normalized data (i.e. x and y each sum up to 1)

Supplement

Definition of measures continued

Variance of log-ratios

$$d(x,y) = \mathrm{var}(\log(\frac{x_i}{y_i}))$$

Aitchison proposed a scaling between 0 and 1, where 1 corresponds to maximal similarity:

$$d(x,y) = 1 - e^{-\sqrt{d(x,y)}}$$

Variance of log-ratios, conceived for compositional data

Require pseudo-counts or smoothing because log(0) = -Inf

Pearson

$$d(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

Spearman

$$d(x,y) = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}, d_i = x_i - y_i(ranks)$$

For Pearson, vectors *x* and *y* are standardized (subtraction of mean, division by standard deviation) and for Spearman, ranks are considered, so vector-wise standardization is not necessary for either of these measures.

Appendix

Generalized Boosted linear models (GBLM)

$$x_{tt,\,ts} = \bar{x}_{tt,\,ts} + \sum_{st} \beta_{tt,\,ts,\,st,\,ss} x_{st,\,ss}$$

$x_{tt,ts}$ = target taxon at target site
$x_{st,ss}$= source taxon at source site
β = coefficients (interaction strengths)

Multiple regression: more than one source taxon may predict the target taxon’s abundance
Boosting: a form of **sparse regression** (coefficients with small contributions are set to zero)

In practice, all source taxa of a body site are considered to predict the abundance of a target taxon in the same or another body site. Then, the optimal sub-set of source taxa is selected by boosting (sparsity enforcement).

Appendix

Generalized Boosted linear models (GBLM)

Prefiltering

- only source taxa correlating with target taxon with Spearman p-value < 0.05 considered (to enforce sparsity and avoid over-fitting)

Scoring

Regression scoring: adjusted R^2 (AR)
R^2 = root mean square error between prediction and observation

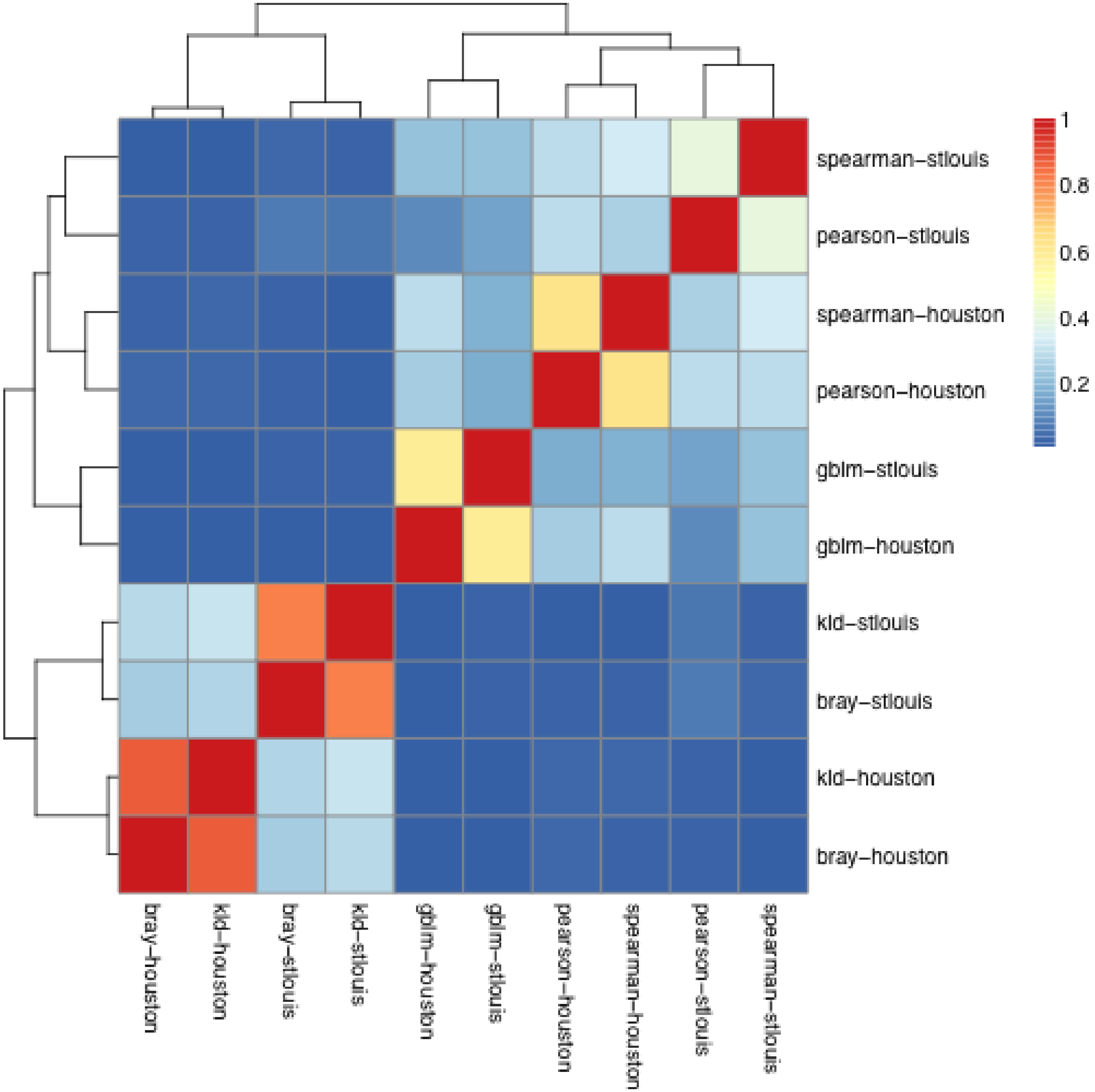
$$AR^2 = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

n = sample number
p = number of source taxa with non-zero coefficient

Cross-validation

- boosting was carried out with three different iteration numbers (50, 100, 150)
- the most accurate (according to AR^2) selected among the three
- 10-fold cross-validated and minimum AR^2 retained as regression score

Agreement between data and methods



Appendix