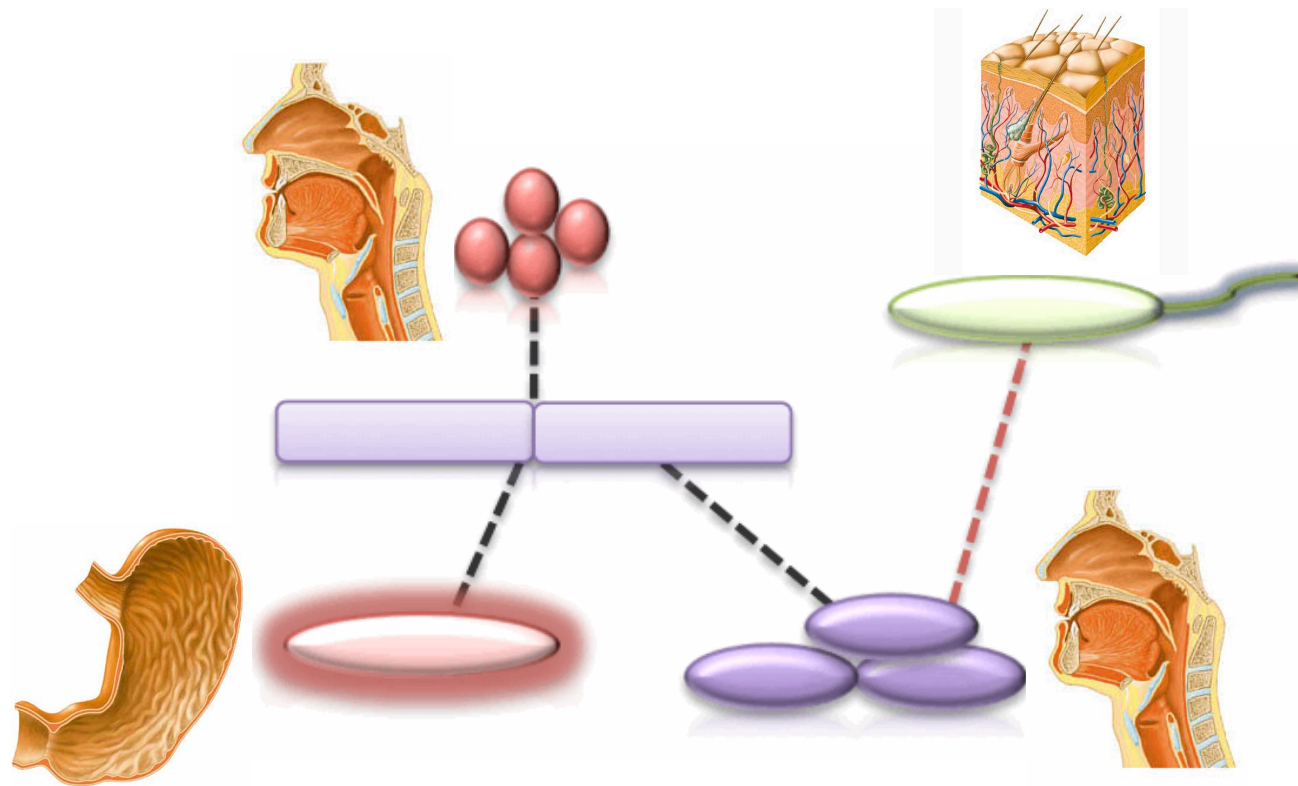
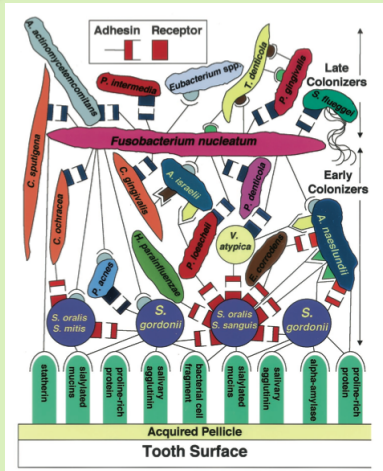


Detecting bacterial associations in the human microbiome

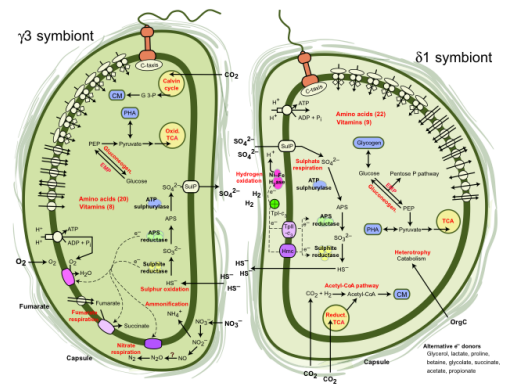


Examples for microbial relationships

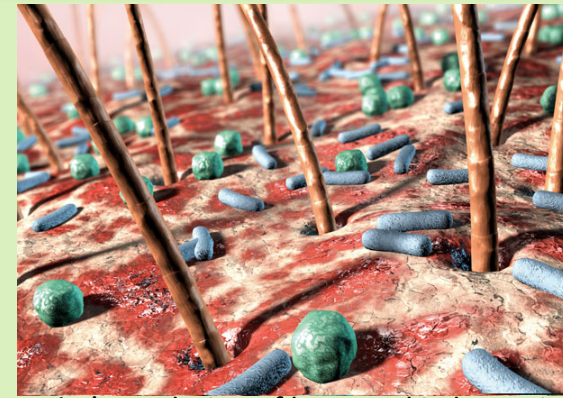


dental plaque formation (Kolenbrander et al.)

sulfur oxidizer sulfate reducer



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



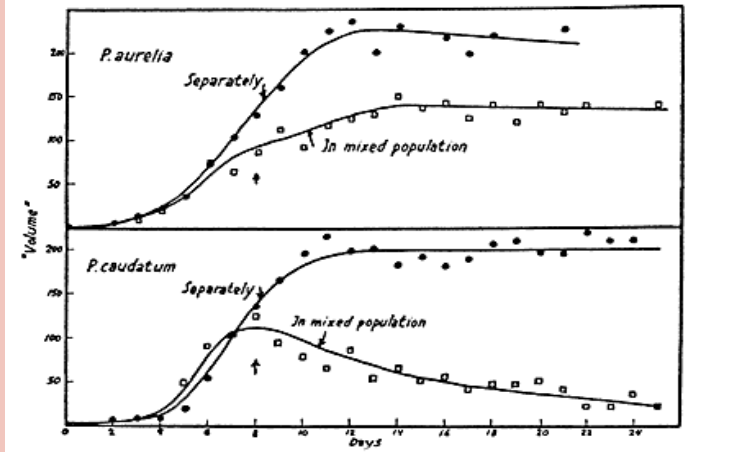
artist's rendering of human skin bacteria



Amoeba proteus feeding on algae



Bacteriophages infecting a bacterium



competition between two species of Paramecium (Gause)



algal bloom killing off other organisms

1. Introduction

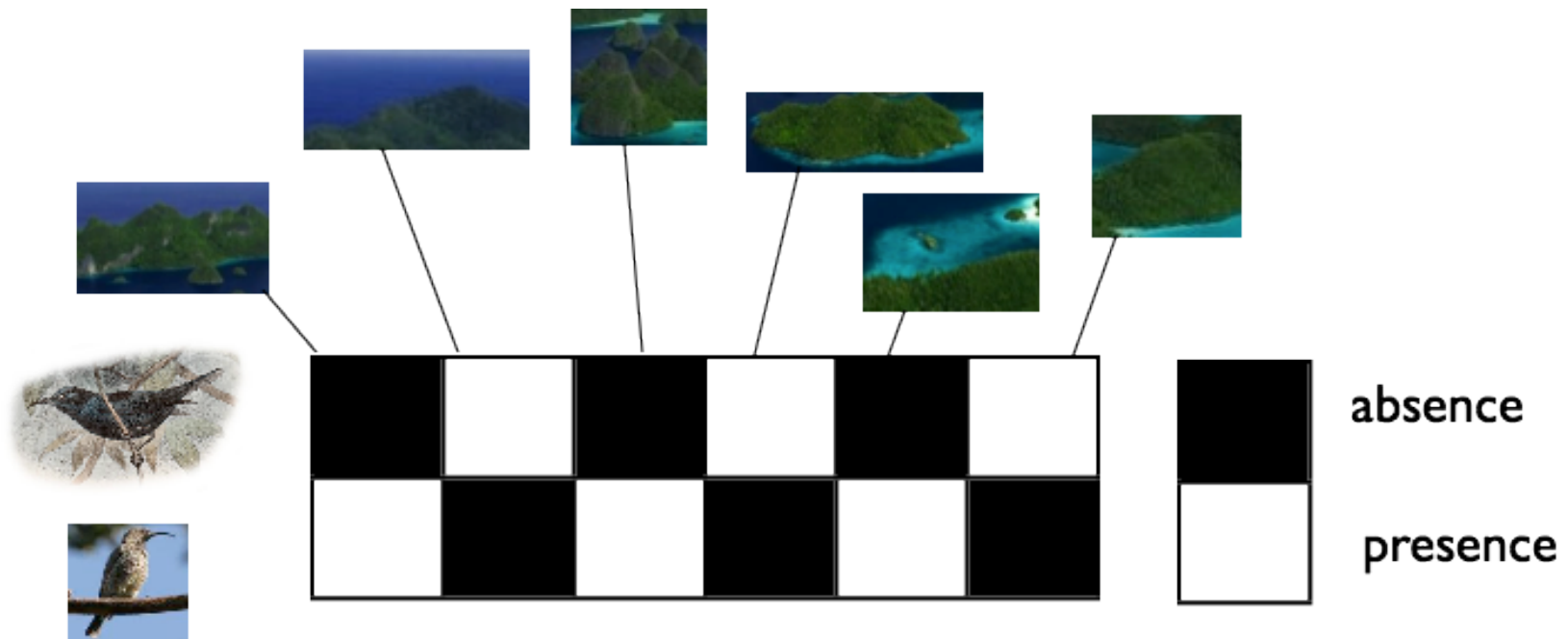
Gause (1934) "The Struggle for Existence", Williams & Wilkins.

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.

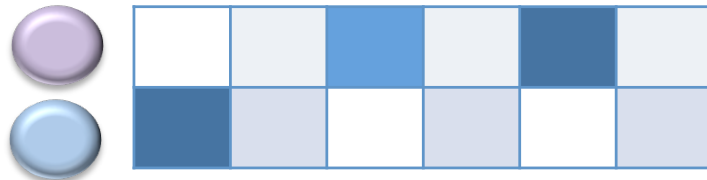
Detecting ecological relationships from presence/absence data

- 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.)

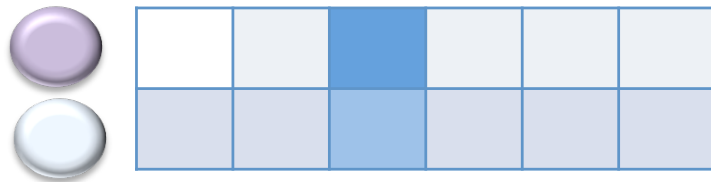


Diamond, J. (1975) "Assembly of species communities", pp. 342-444 in "Ecology and evolution of communities" edited by Cody and Diamond, Harvard University Press.
Horner-Devine M.C. et al. (2007) "A Comparison Of Taxon Co-Occurrence Patterns For Macro- And Microorganisms" Ecology 88, pp. 1345-1353.

Co-occurrence analysis in a nut shell



mutual exclusion (checker board)/anti-correlation

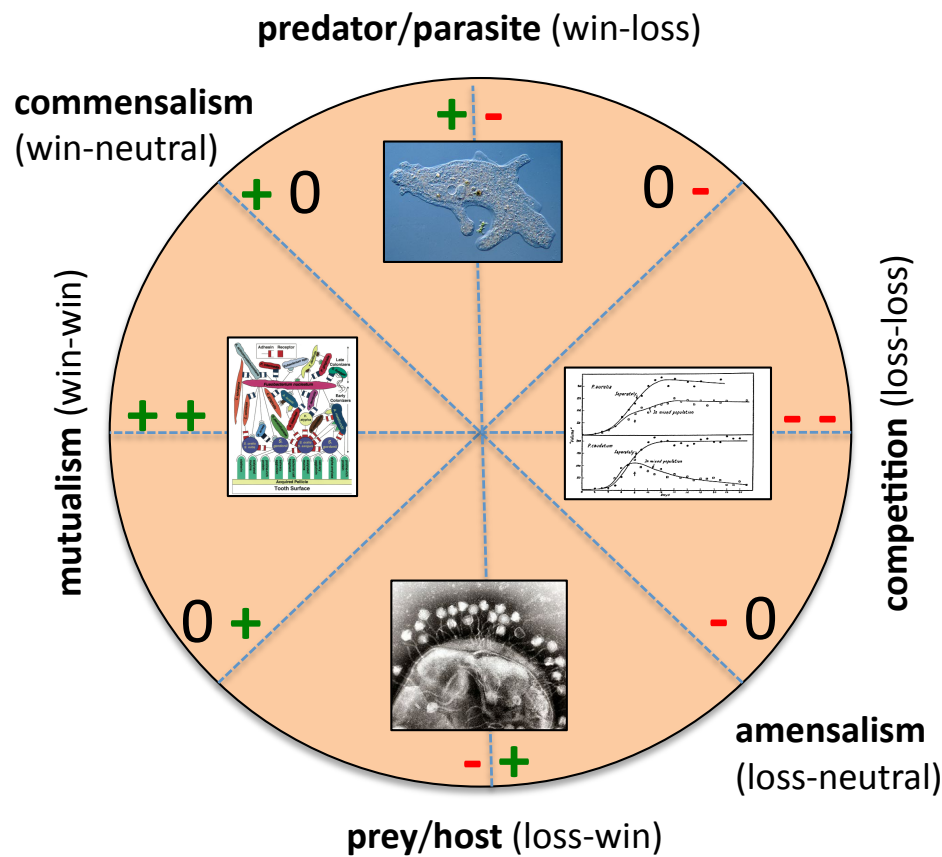


co-occurrence/correlation

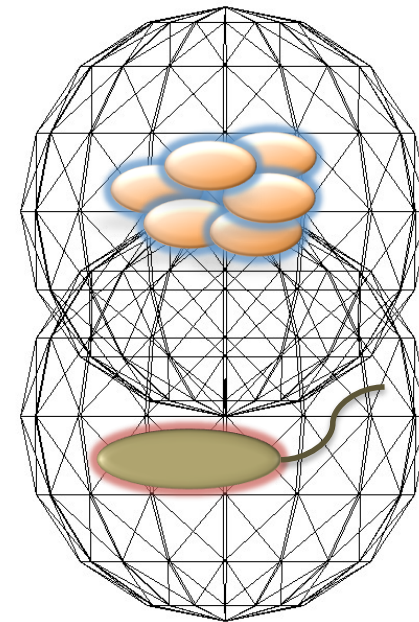
Reasons for association

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

ecological relationships

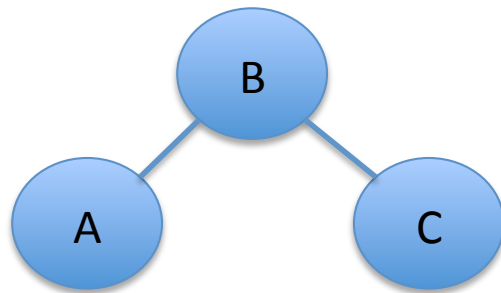
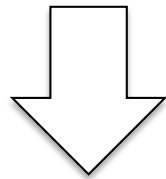


niche overlap



Inferring networks

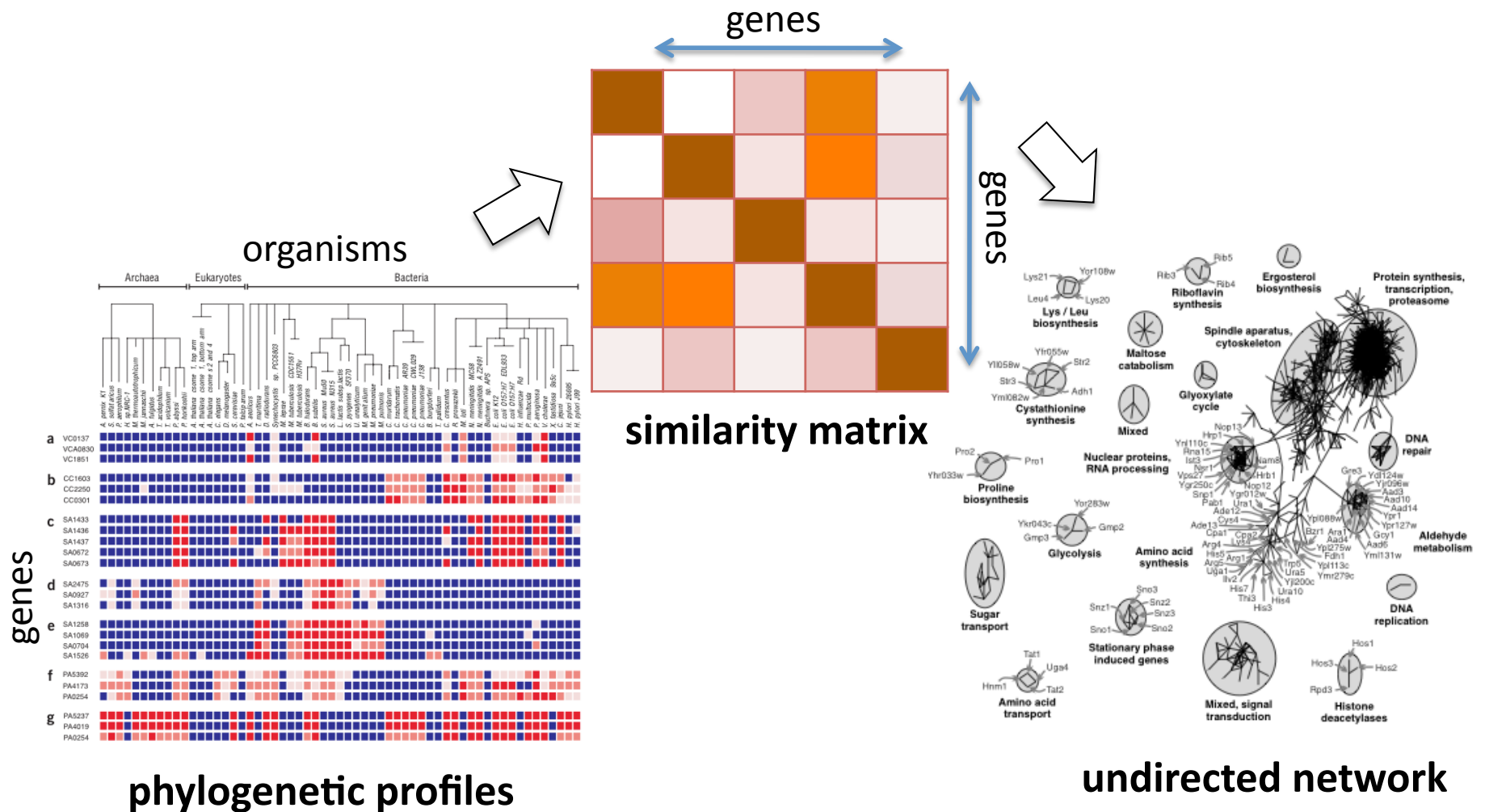
A					
B					
C					



- **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

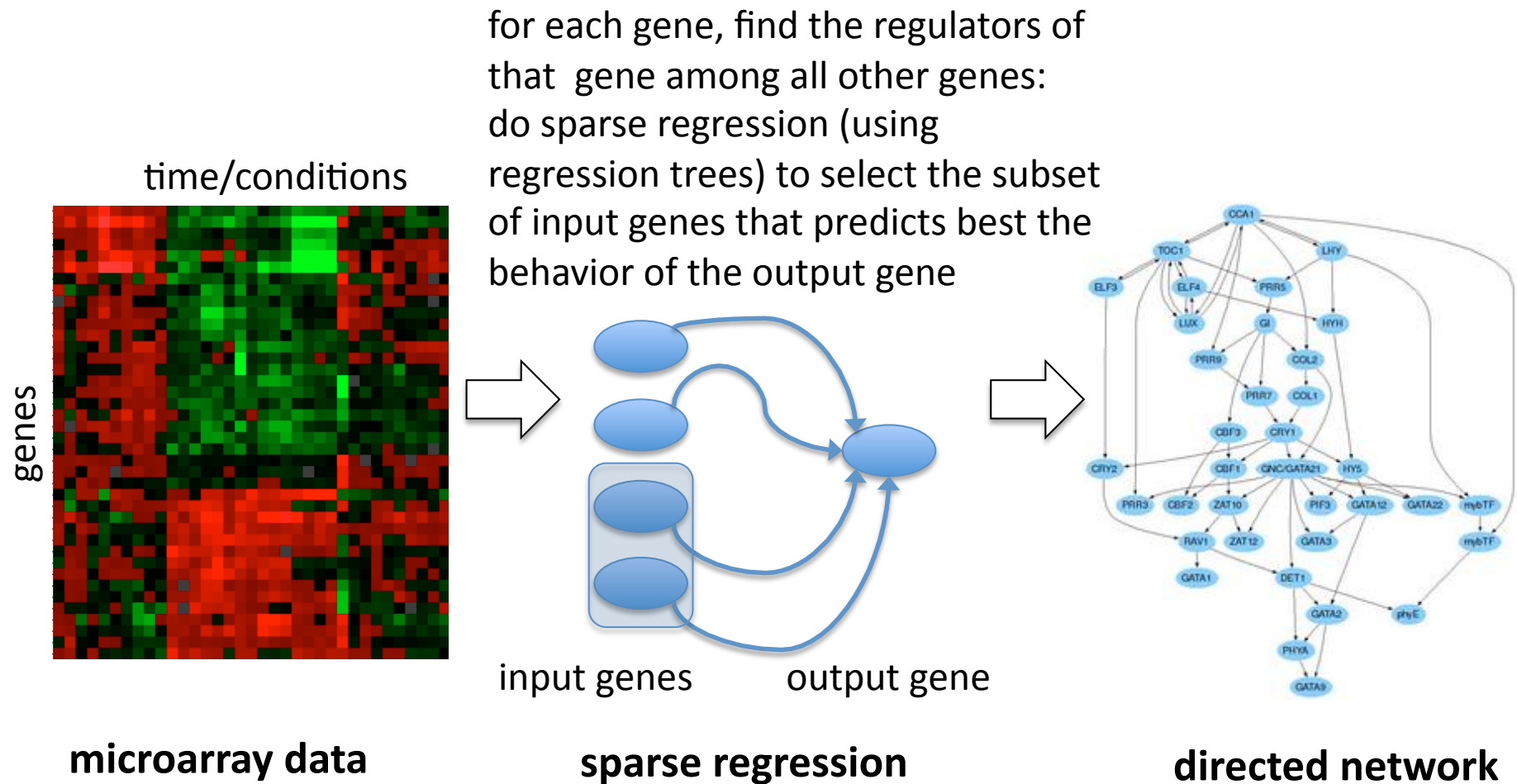
- task: obtain functional protein modules from co-occurrences of genes



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.

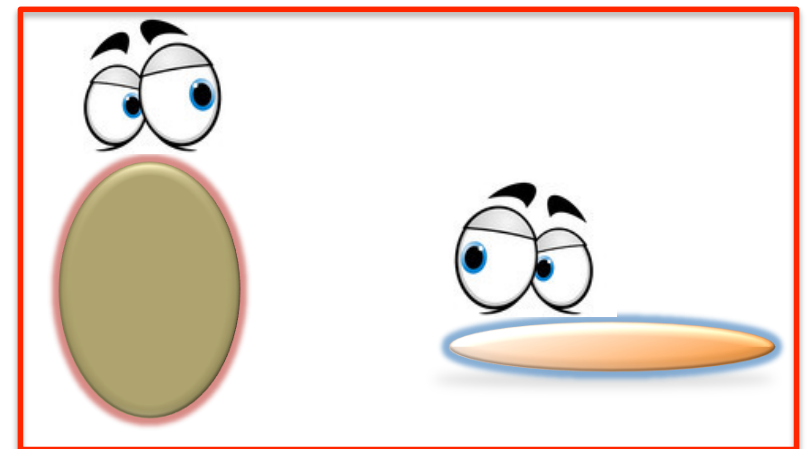
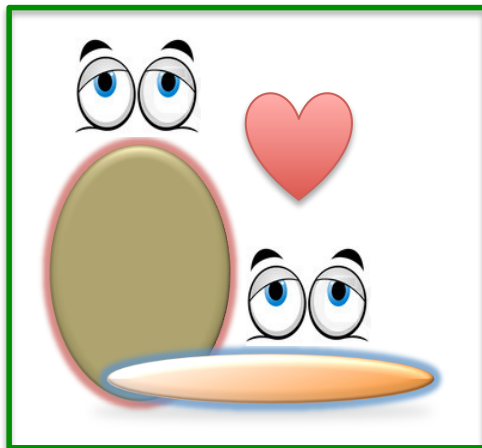
Example for sparse regression-based network inference

- task: identify gene regulatory network from microarray data



Goal: Infer network of microbial relationships

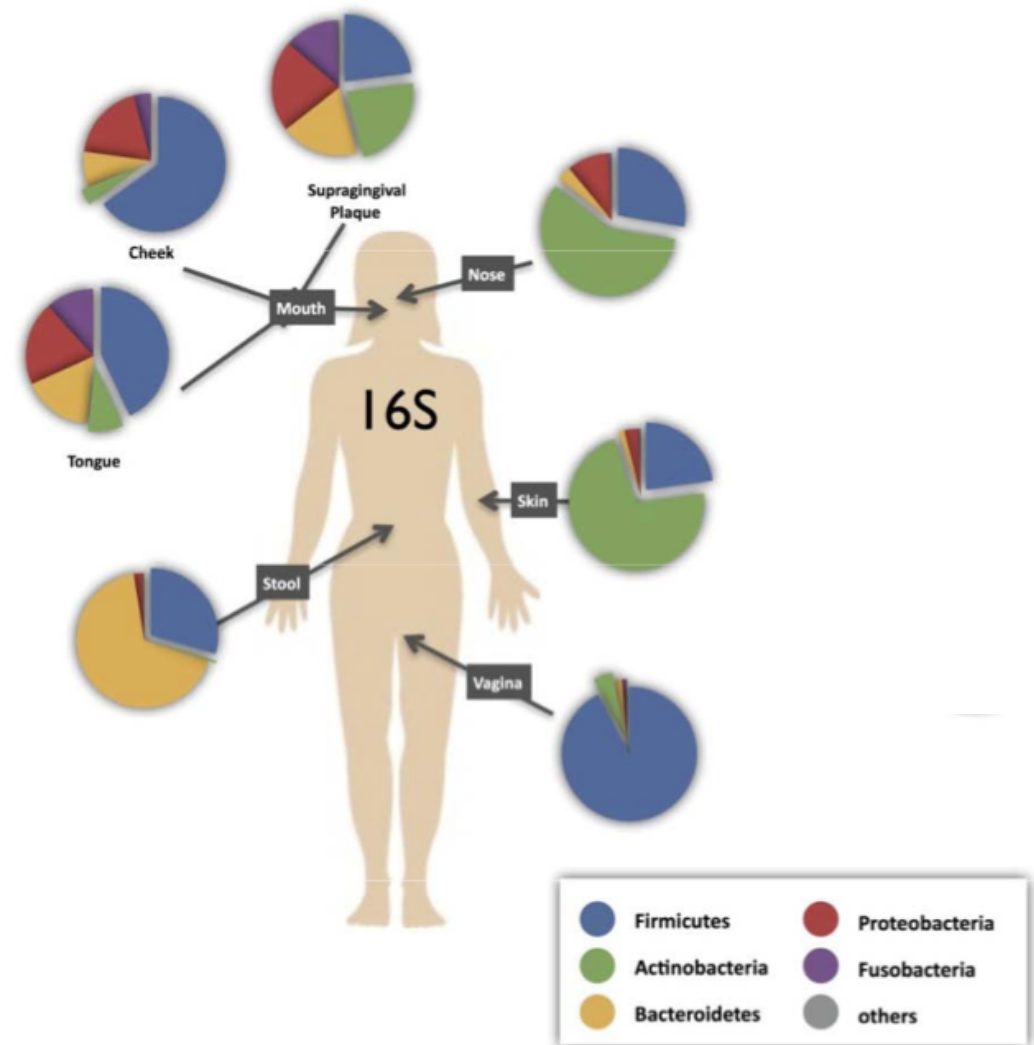
- several recent metagenomic data sets measure microbial abundance across a large number of samples
- network inference techniques can identify significant relationships between microorganisms from these data
- significant **co-presence** (co-occurrence of two microbes across samples) can be interpreted 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

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 **phlotypes** 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.

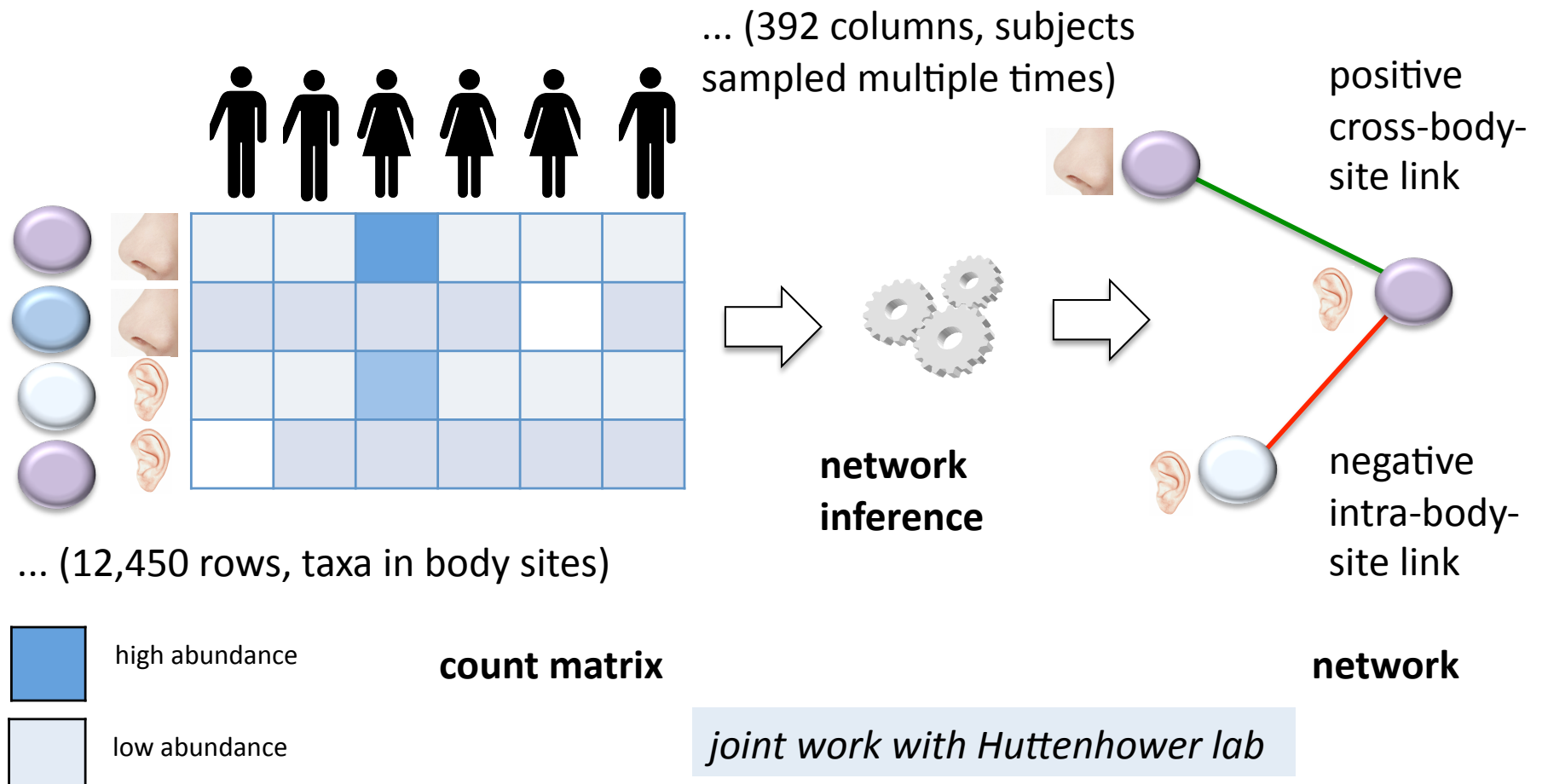
Jumpstart Consortium Human Microbiome Project Data Generation Working Group "Evaluation of 16S rDNA-based Community Profiling for Human Microbiome Research", PLoS one 7, e39315.

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.

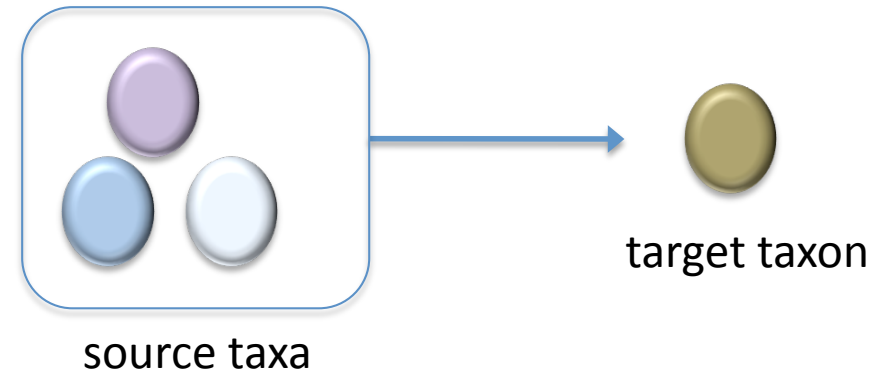
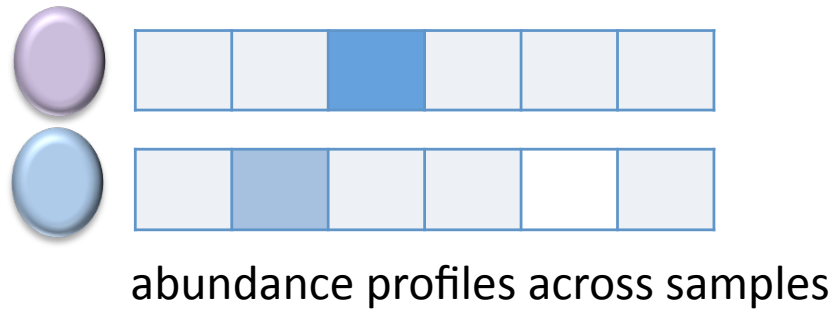
Knights, R. et al. (2010) "Supervised classification of microbiota mitigates mislabeling errors." ISME 5, pp. 570-573.

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)



Assessing strength of relationships between microorganisms



Pair-wise relationships (similarity)

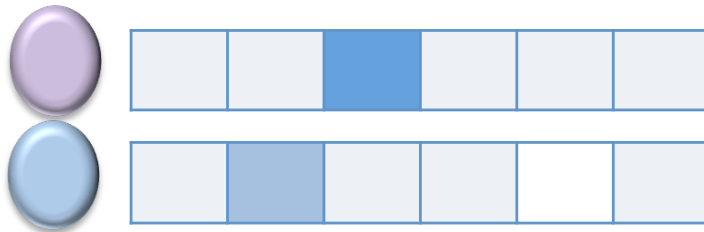
- Pearson correlation
- Spearman correlation
- Kullback-Leibler dissimilarity (KLD)
- Bray Curtis dissimilarity (BC)

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)

Computing significance of relationships I

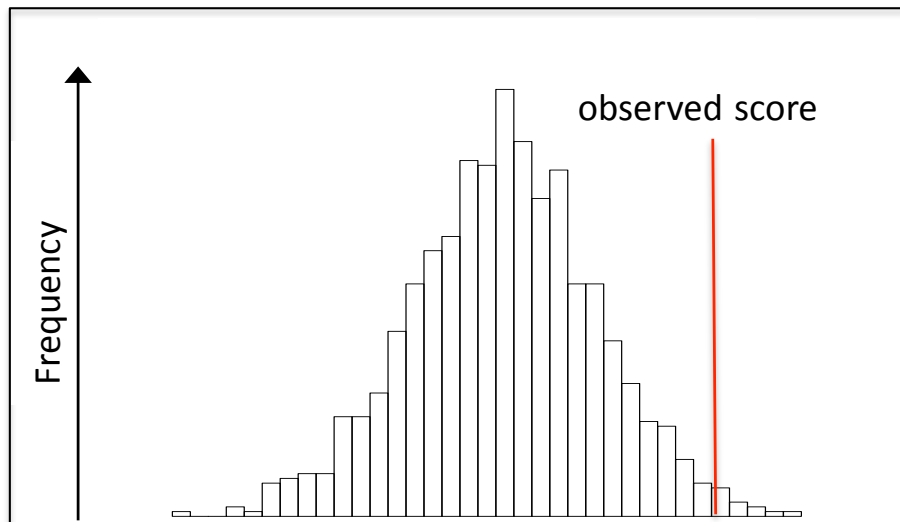
- for each of the five methods (Pearson, Spearman, Kullback-Leibler, Bray Curtis, GLBM), compute permutation and bootstrap edge scores



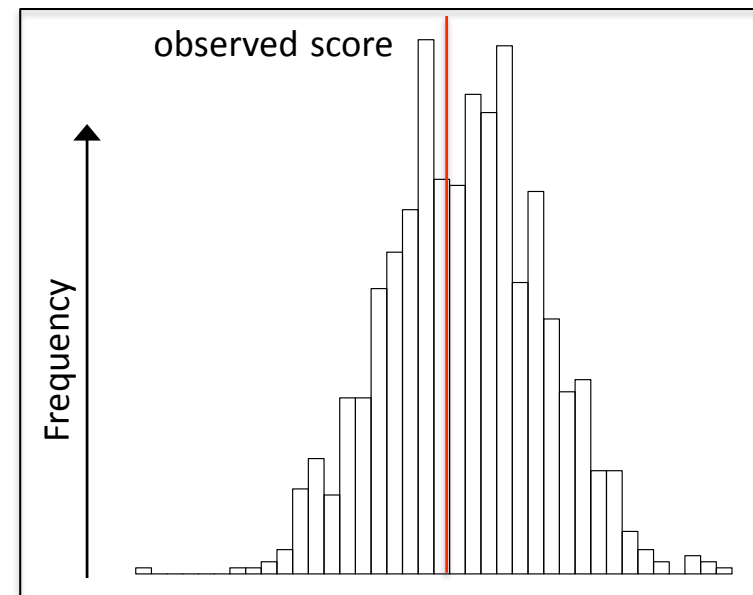
5 scores per edge, for each score:



permutation (null) distribution of method-specific edge score

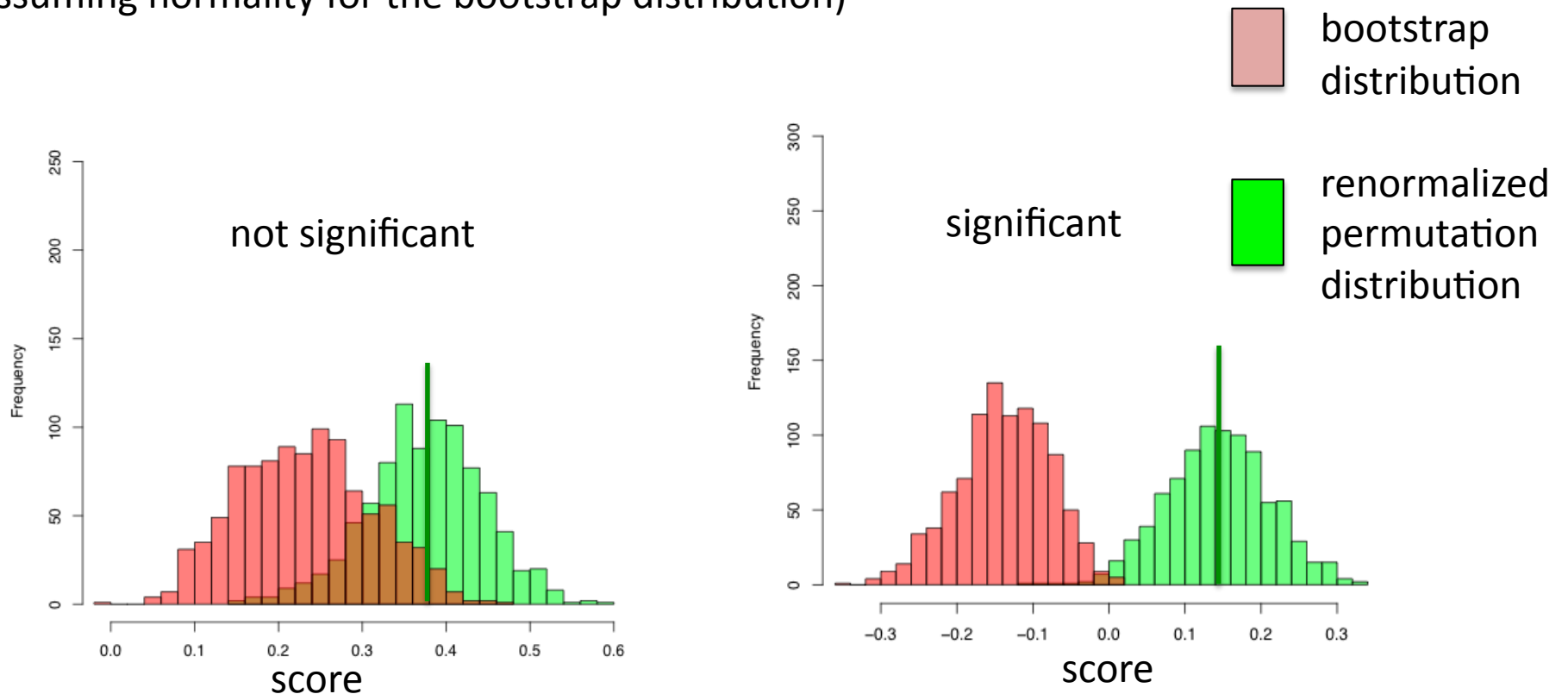


bootstrap distribution of method-specific edge score (confidence interval)



Computing significance of relationships II

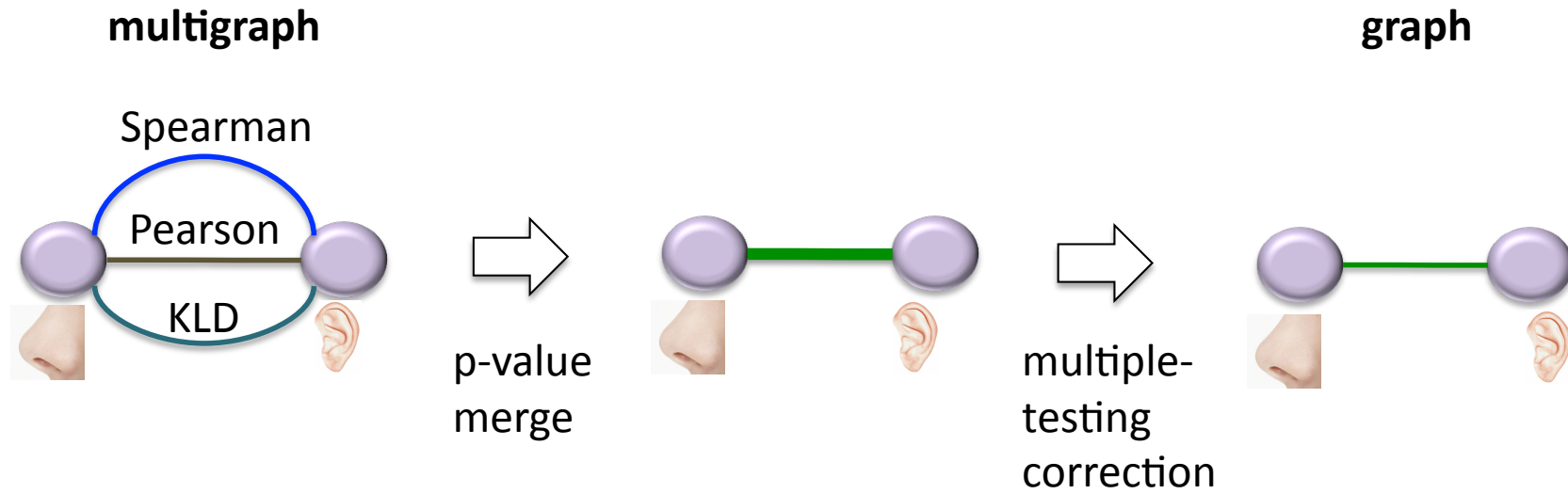
Edge- and method-specific p-value is computed with a **Z-test** (p-value of the **null distribution mean** given the **bootstrap distribution**, assuming normality for the bootstrap distribution)



Fusobacteriales versus
Streptococcaceae in buccal
mucosa (Pearson)

Actinobacteria versus
Bacteroidetes in subgingival
plaque (Spearman)

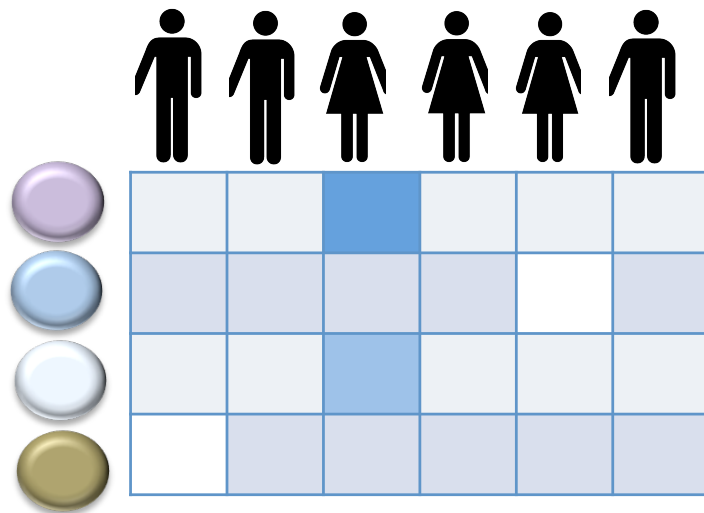
Network building



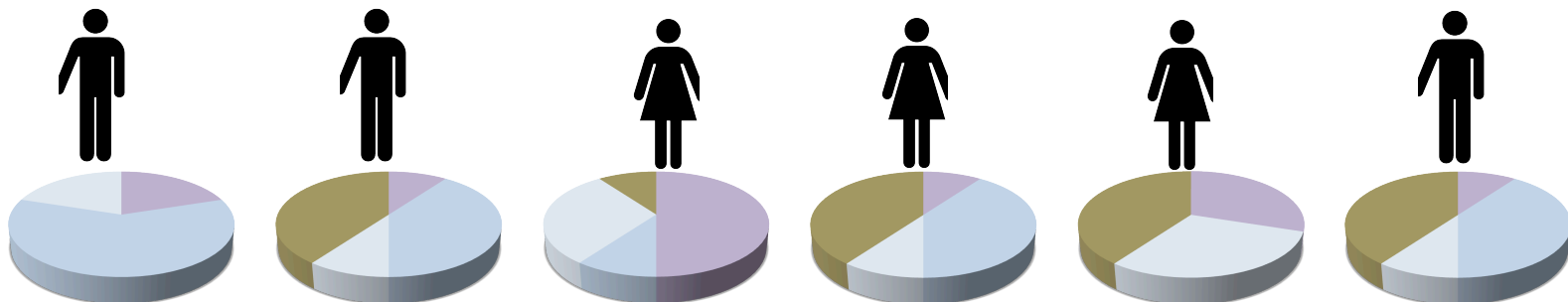
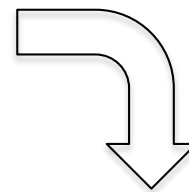
- merge method-specific p-values using Sime's method
- apply Benjamini-Hochberg Yekutieli False Discovery Rate correction on merged p-values
- after correction, remove all p-values above the threshold (set to 0.05)
- represent remaining relationships as a network

Problem: Data normalization and compositionality

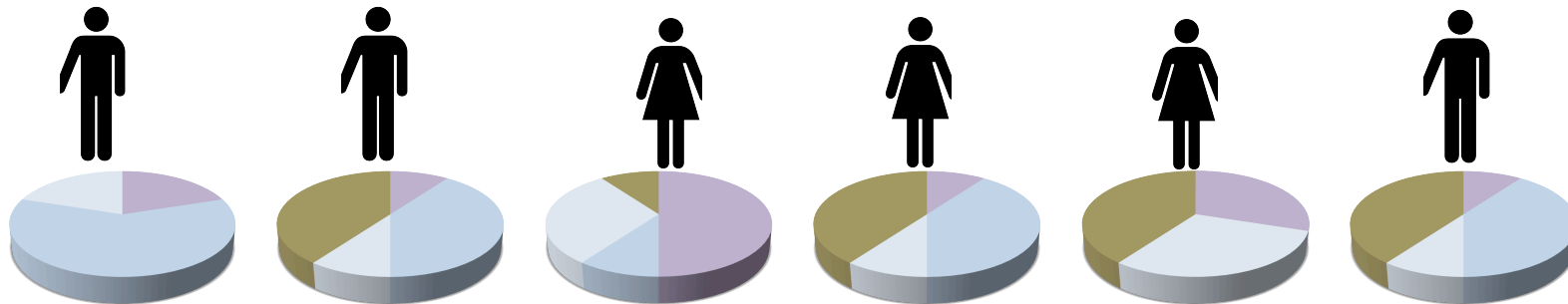
- technical errors/differences in processing lead to different total abundances across samples
- sample-wise normalization necessary (i.e. division of abundances in a sample by this sample's total abundance sum)
- absolute abundances are converted into proportions



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



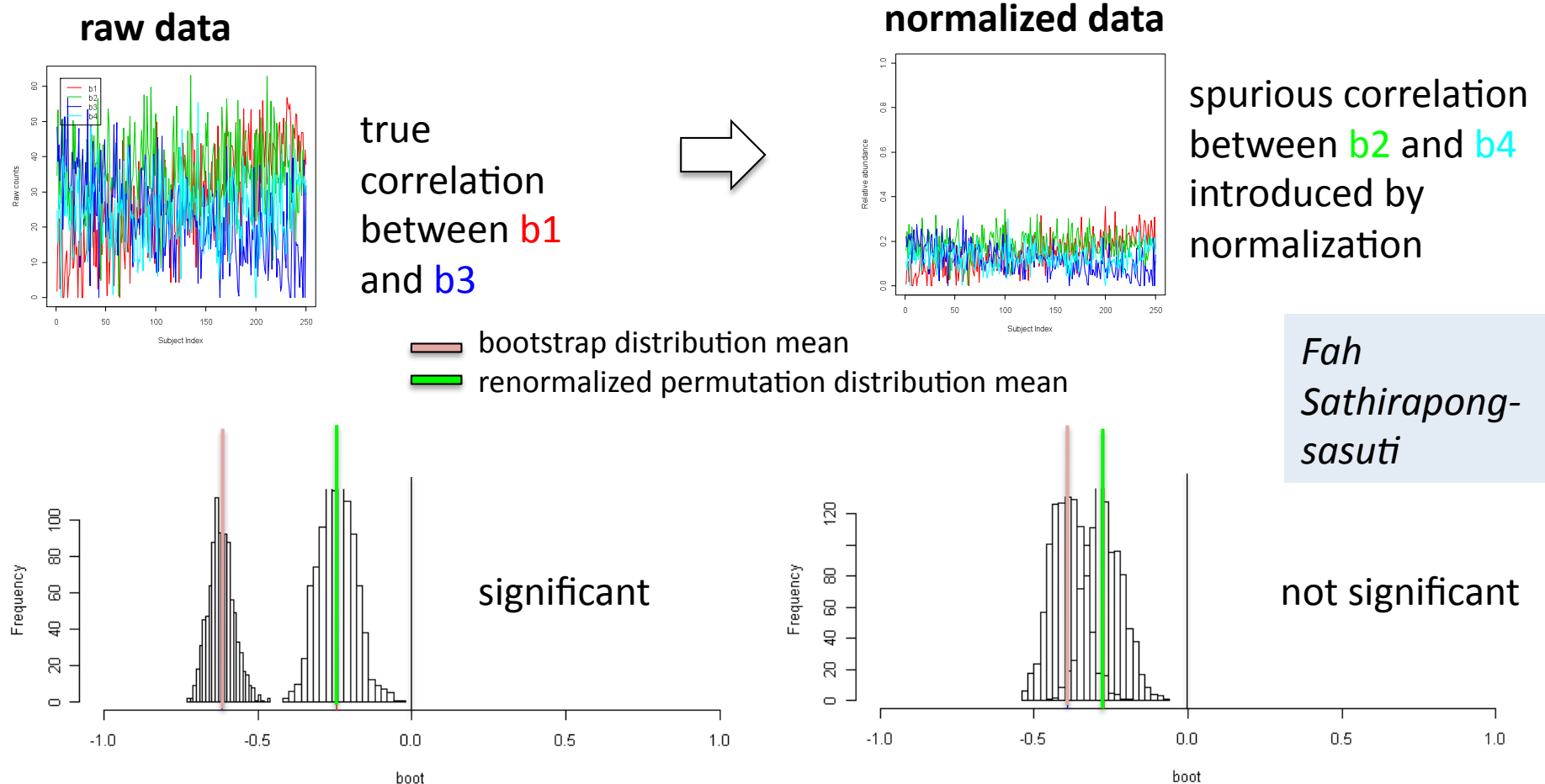
Problem: Data normalization and compositionality



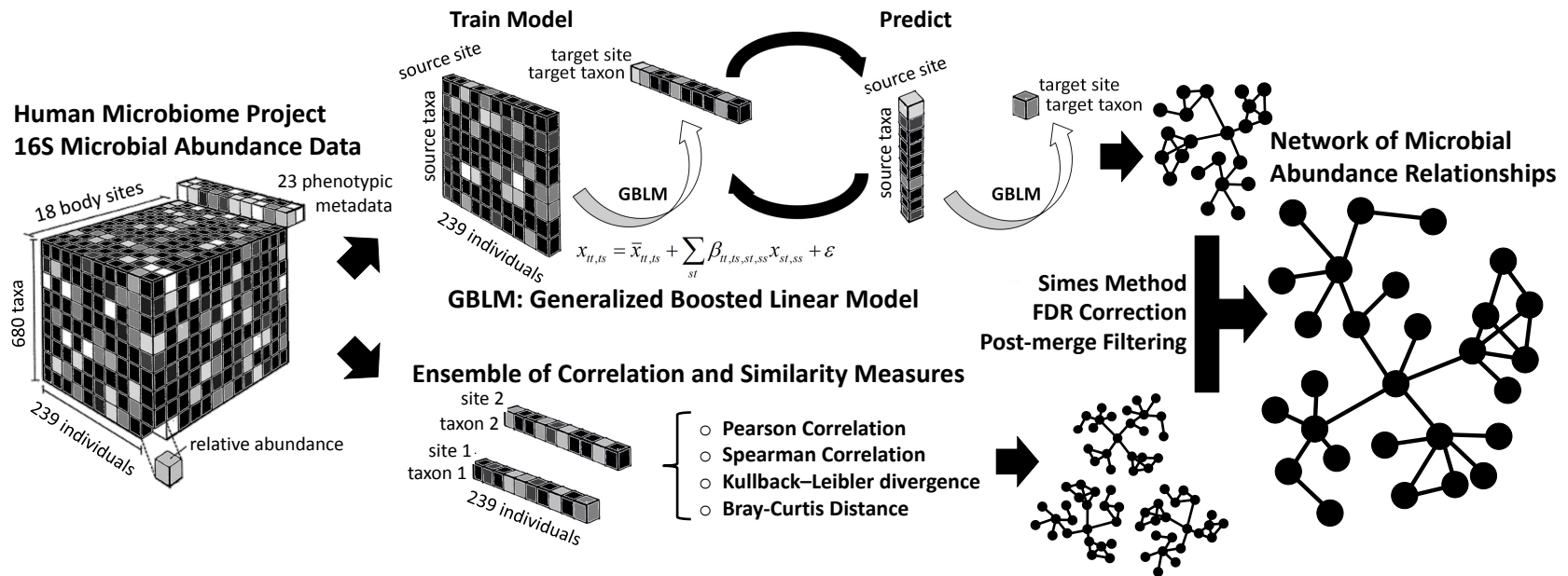
- Pearson and Spearman can be severely distorted, because they consider “absolute” values
- measures based on ratios or log-ratios (KLD, BC) are not affected by data compositionality, since the ratio between two abundances in the same sample is not changed by the normalization

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 renormalize the matrix (body-site-wise)

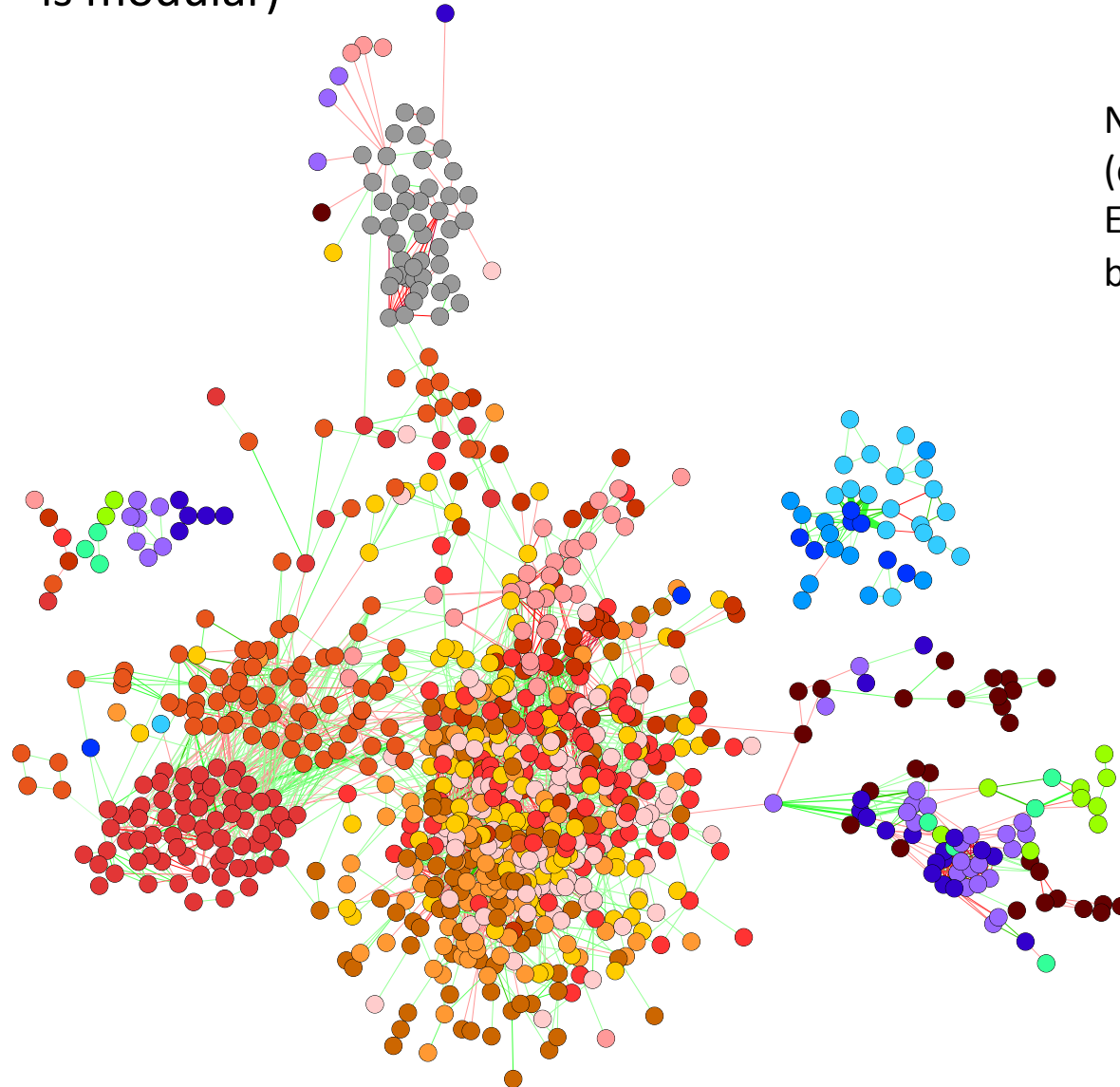


Methodology overview

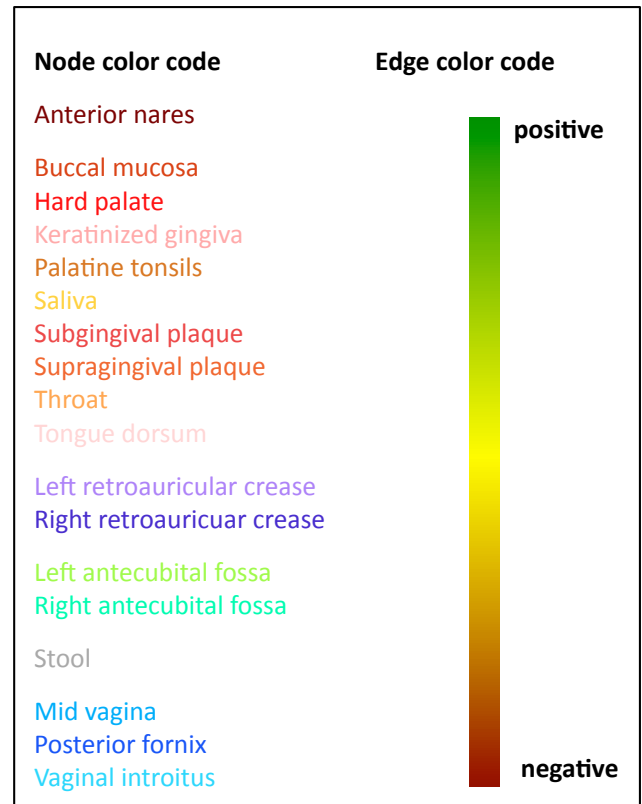


Network inferred for HMP 16S phylotypes

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

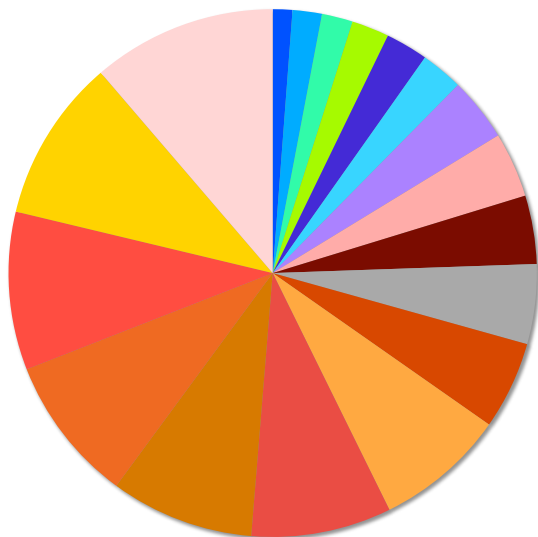


Nodes: body-site-specific phylotypes
(e.g. *Ruminococcaceae* in Stool)
Edges: significant score between
body-site-specific phylotypes



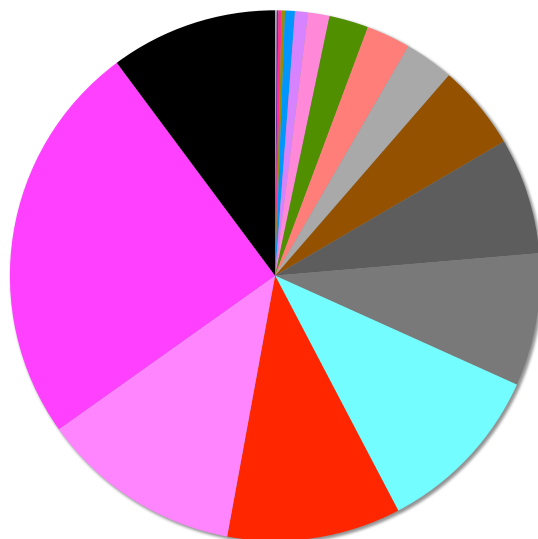
HMP 16S Network - composition

Body-site-specific node proportions



- | | |
|-----------------------------|----------------------|
| Posterior fornix | Buccal mucosa |
| Mid vagina | Throat |
| Right antecubital fossa | Subgingival plaque |
| Left antecubital fossa | Palatine tonsils |
| Right retroauricular crease | Supragingival plaque |
| Vaginal introitus | Hard palate |
| Left retroauricular crease | Saliva |
| Keratinized gingiva | Tongue dorsum |
| Anterior nares | |
| Stool | |

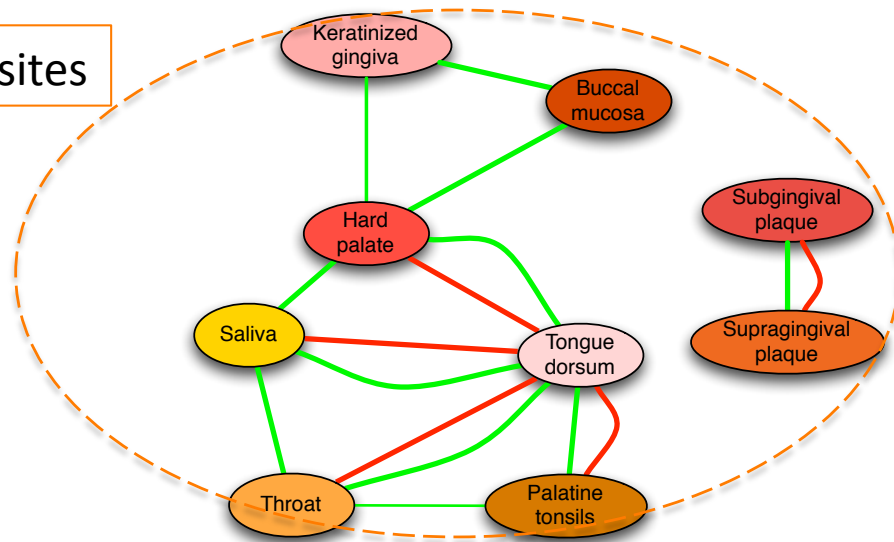
Class-specific node proportions



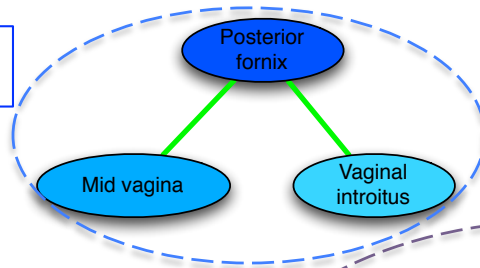
- | | |
|---------------------|-----------------------|
| Alphaproteobacteria | Epsilonproteobacteria |
| Lentisphaeria | Fusobacteria |
| Verrucomicrobiae | Gammaproteobacteria |
| Synergistia | Betaproteobacteria |
| Mollicutes | Actinobacteria |
| Negativicutes | Bacteroidia |
| Erysipelotrichi | Bacilli |
| Spriochaetes | Clostridia |
| Flavobacteria | Above class-level |

HMP 16S network – body-site relationships

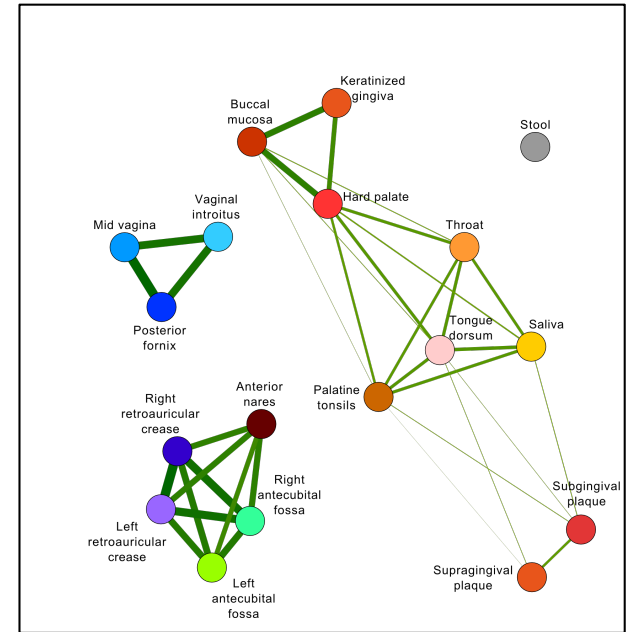
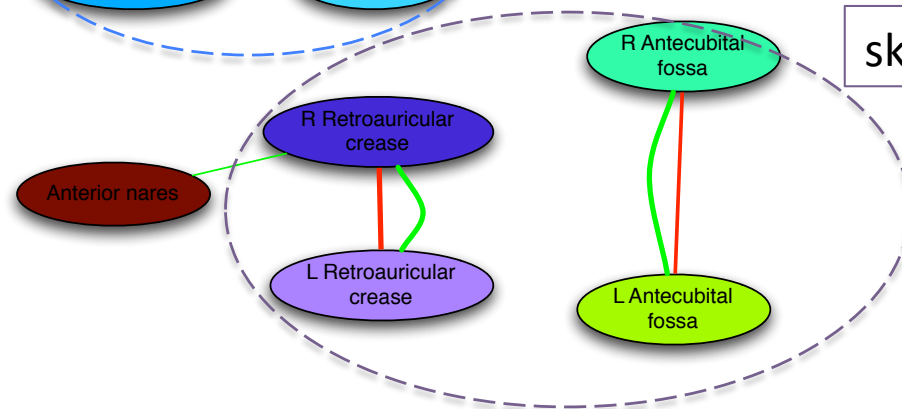
oral cavity sites



vaginal sites

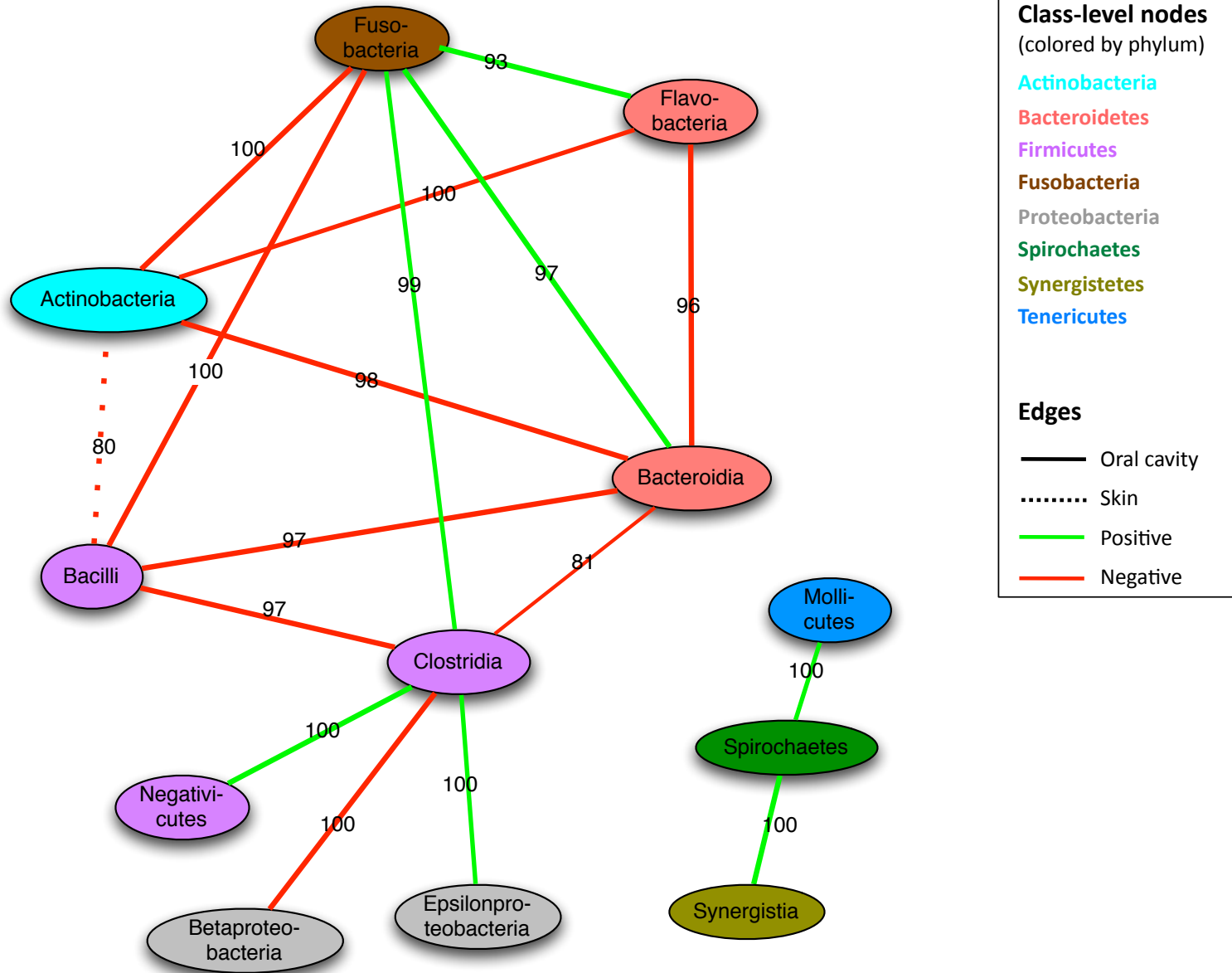


skin sites



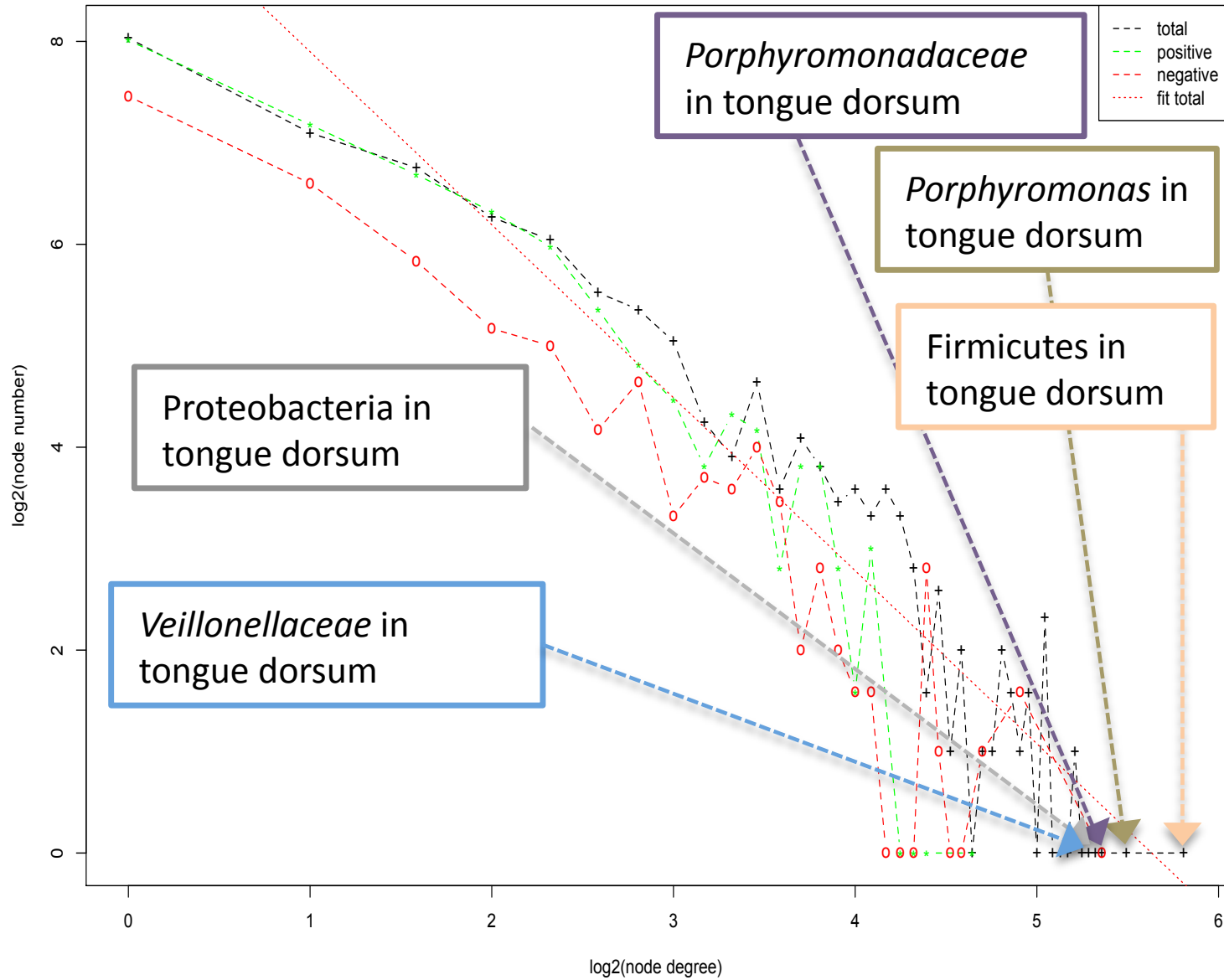
weighted 16S UniFrac beta diversity (Huttenhower et al., Nature 486, 207-214)

HMP 16S network – class relationships



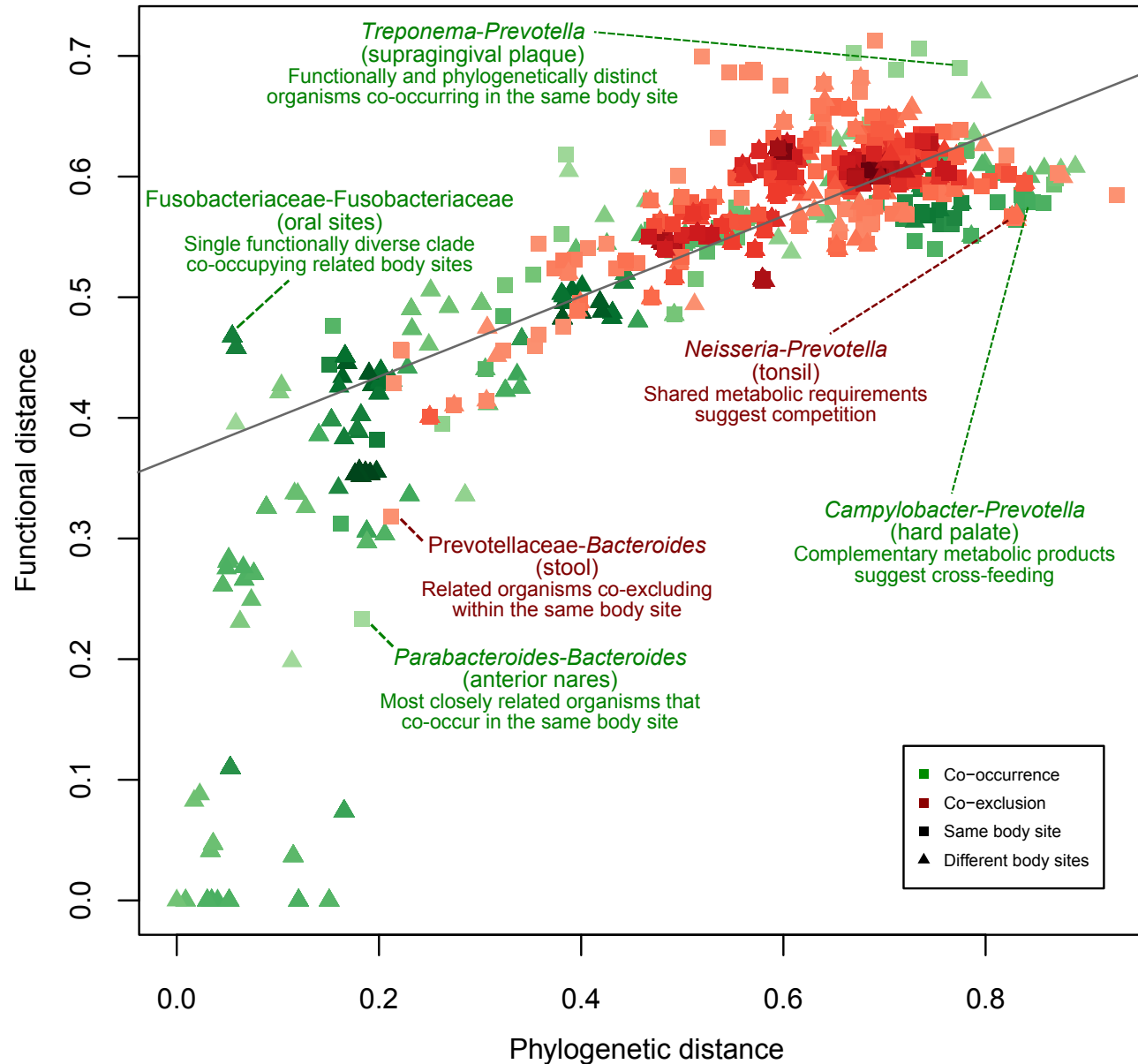
HMP 16S network analysis

Node degree distributions



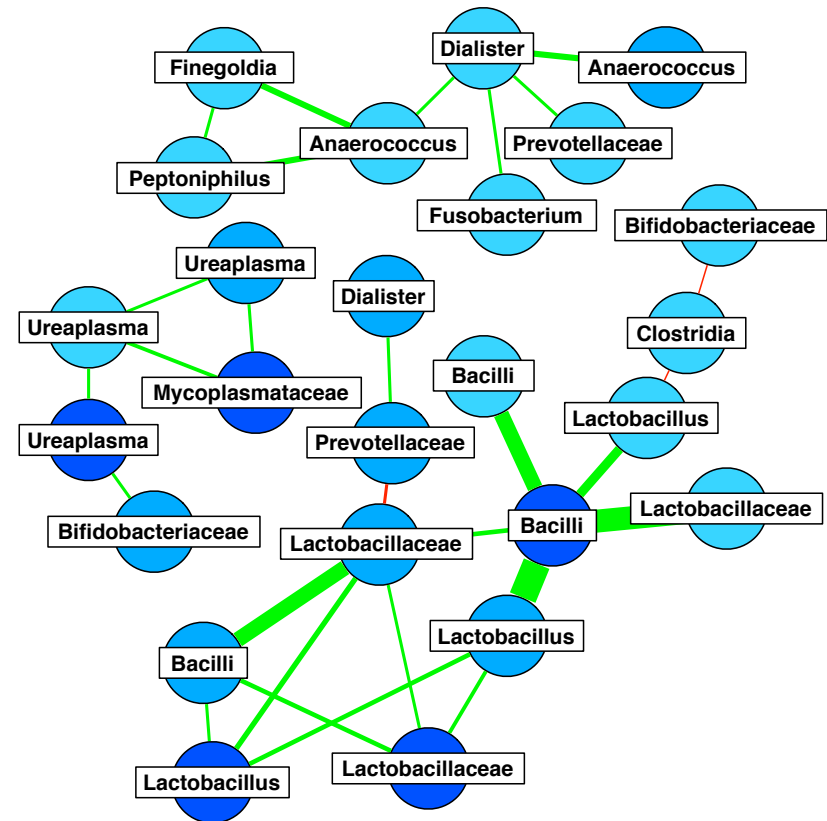
HMP 16S network functional analysis

Phylogenetic and functional distances
between pairs of co-occurring/co-exclusive taxa



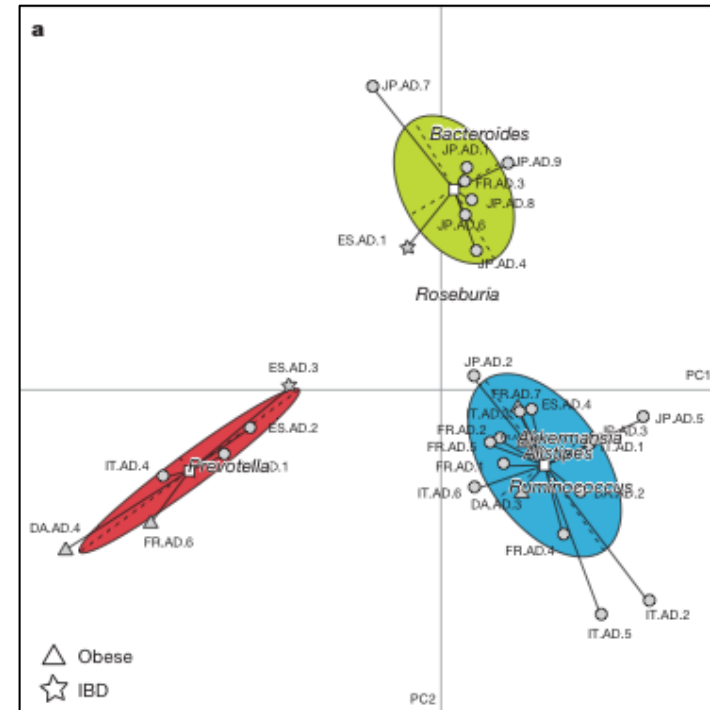
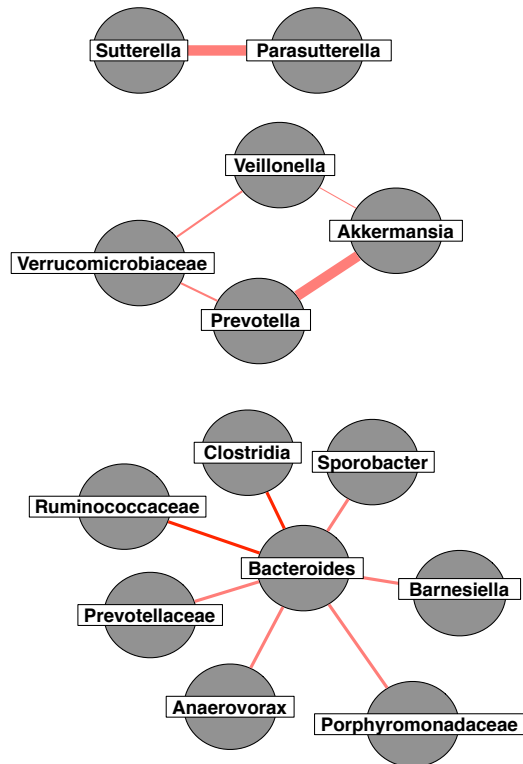
Vaginal sub-network of HMP 16S network

- 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

Stool sub-network of HMP 16S network



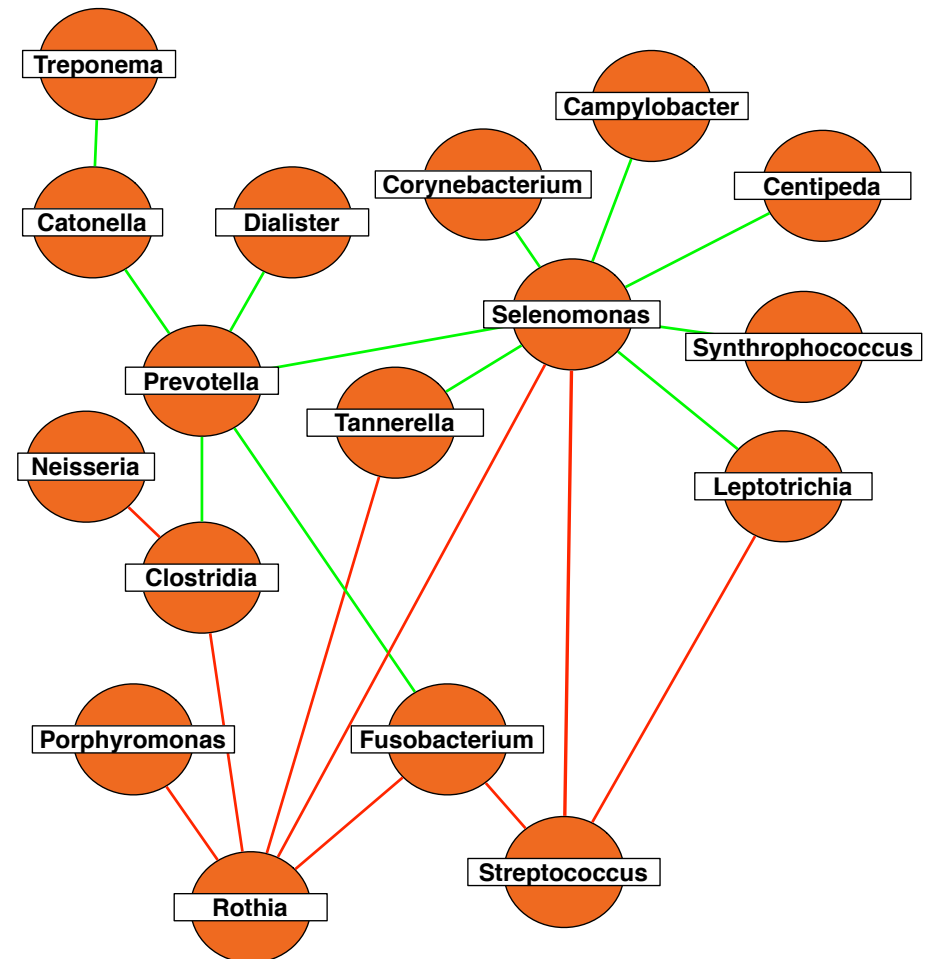
taxonomic levels shown: genus, family and class

- 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

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*)



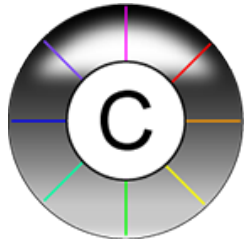
taxonomic levels shown: genus

Conclusions

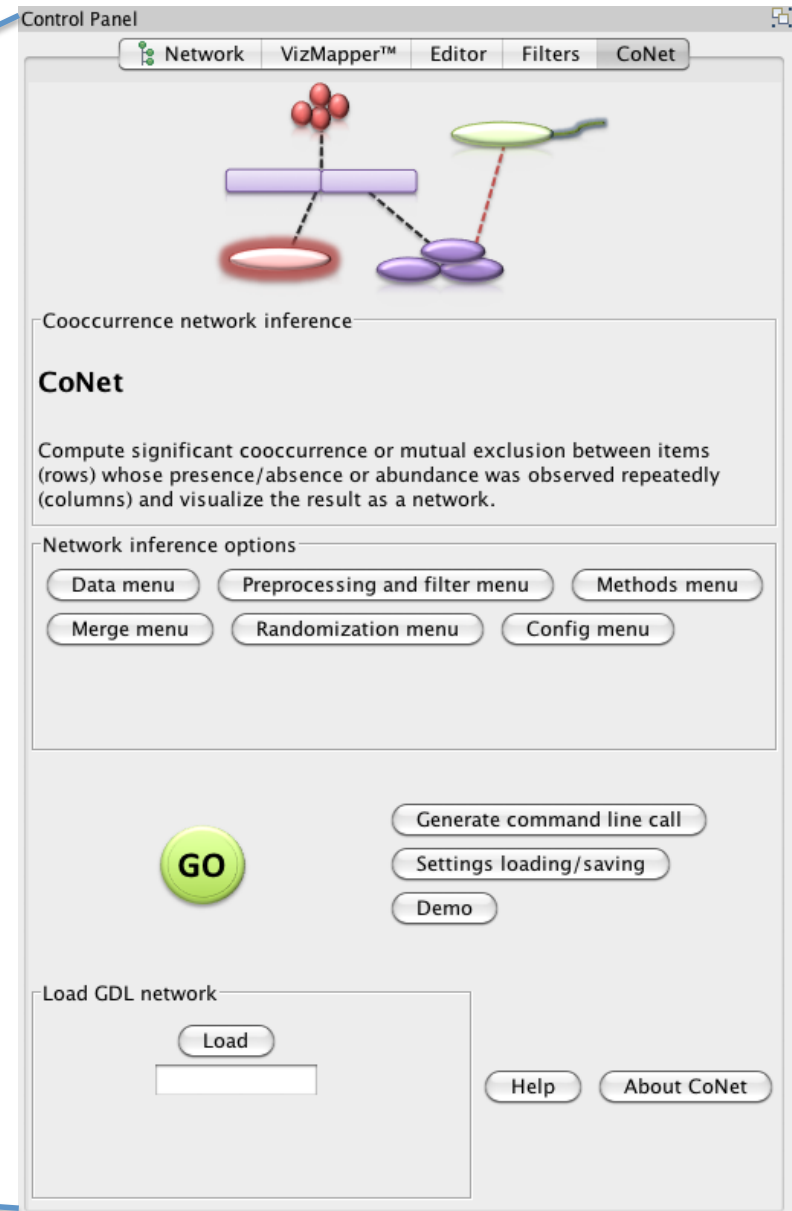
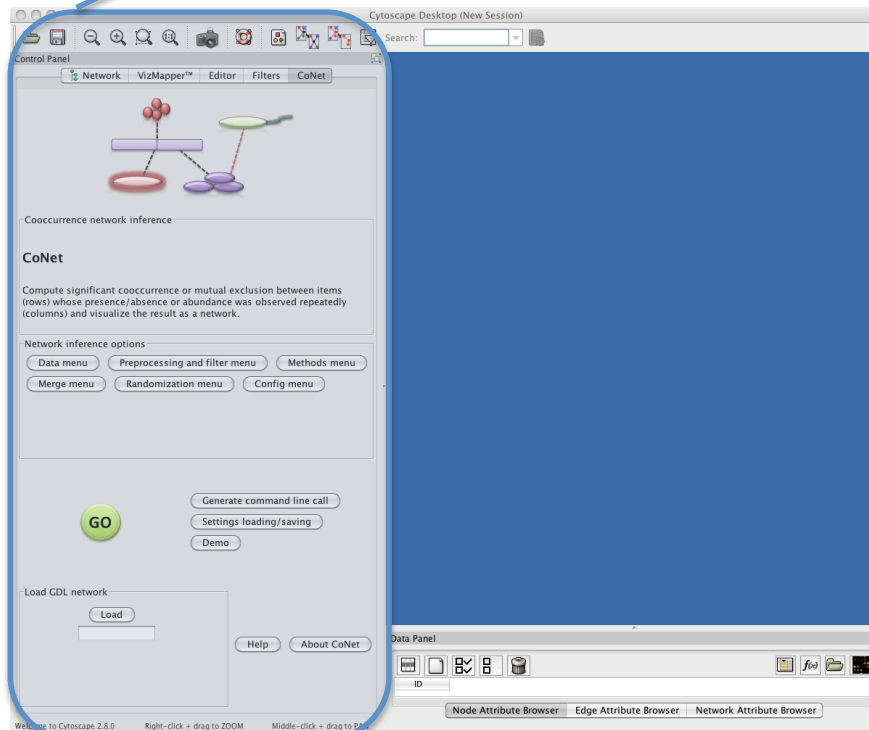
- few cross-body-area relationships: different body areas harbor distinct microbiota
- body sites can be grouped based on cross-links between their microbiota): oral, skin and vaginal sites form separate clusters, airways and stool separated from the oral cavity: clusters can be interpreted as different 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

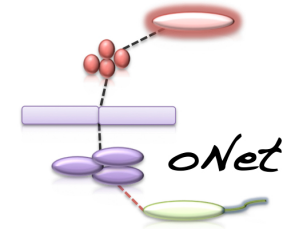


Cytoscape main window



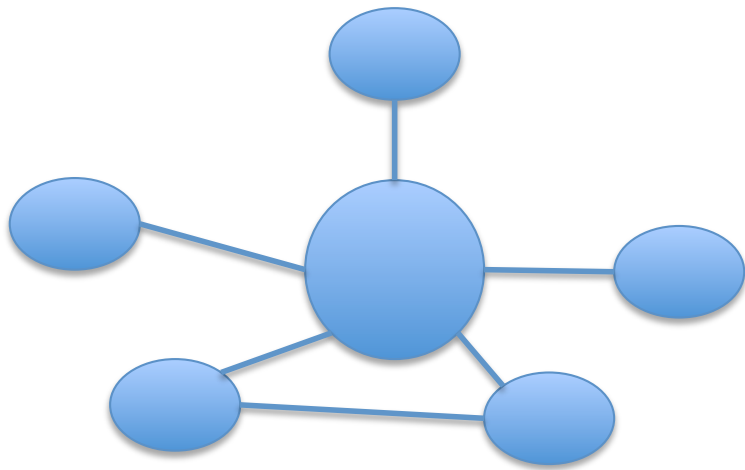
CoNet – Features

<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)

Association networks and biodiversity



Network topology and impact of species removal

- Idea: edge number (“connectedness”) in an association network as predictor for the importance of a species in an ecosystem: removal of highly connected hub species changes system stronger than removal of species with few connections (Steele et al.)
- hub species (= **keystone** species) and system stability: network robust against random node removal, but not against targeted (hub) node removal (Jeong et al.)

removal of the orange bridging node or the red or blue hub node disrupts network much more than removal of black or green nodes

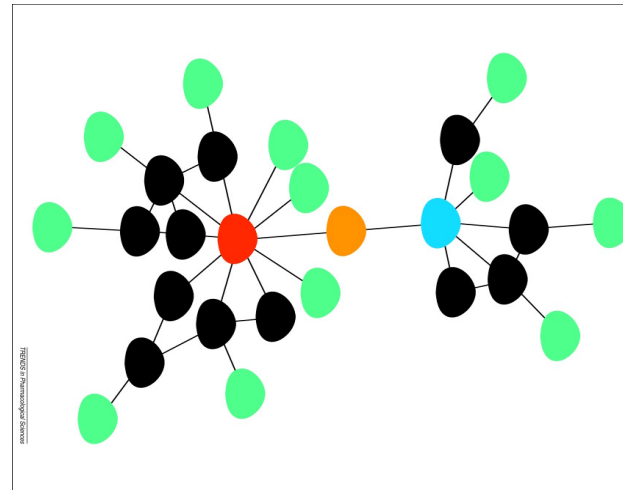
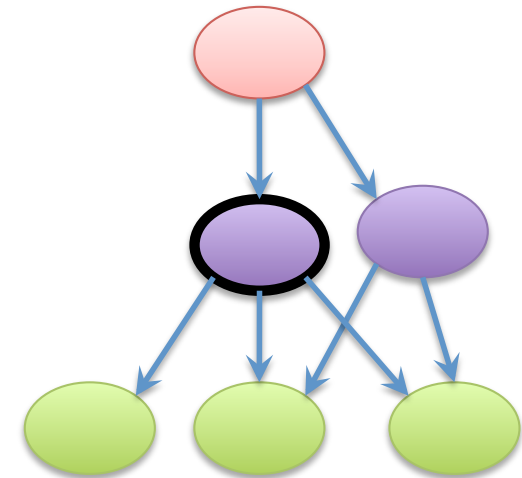


Image taken from: Penrod et al. Trends in Pharmacological Sciences, Vol. 32 , pp. 623-630 2011.

Steele et al. (2011) “Marine bacterial, archaeal and protistan association networks reveal ecological linkages”, ISME 5, pp. 1414-1425.
Jeong et al. (2001) “Lethality and centrality in protein networks”, Nature 411, 41-42.

Network topology and impact of species removal - cascades

- ambiguity: keystone species originally does not refer to highly connected species in association networks, but to **top predators** in food webs (Paine)
- removal of **top predator** launches a **top-down cascade** (higher abundance of **intermediate predators**, lower abundance of their **preys**)
- **bottom-up cascades**: starting from primary producers or important service providers (e.g. nitrogen fixation)
- species whose removal causes strong cascading effects are not necessarily hub species
- need directed networks to identify likely sources of cascades (as in food webs), but association networks are usually undirected

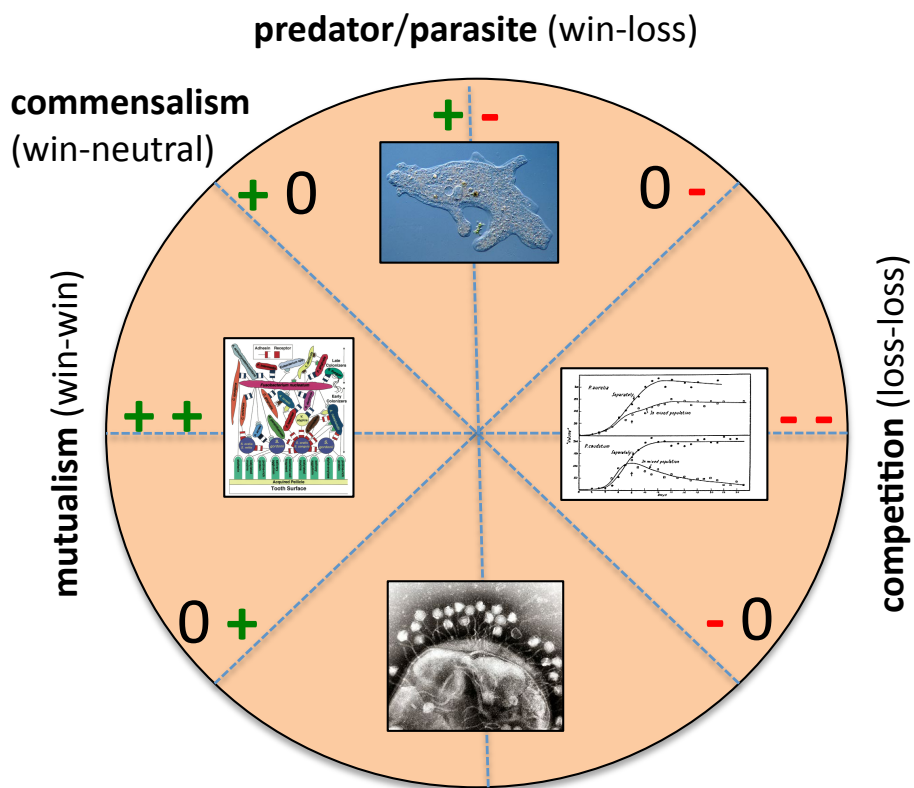


top predator is less connected than the intermediate predators, but its removal will likely affect the system much stronger

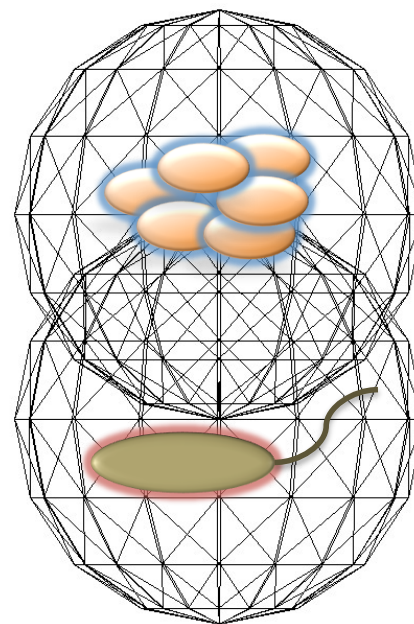
Network topology and impact of species removal – edge ambiguity

the interpretation of the edges in co-occurrence networks is ambiguous!

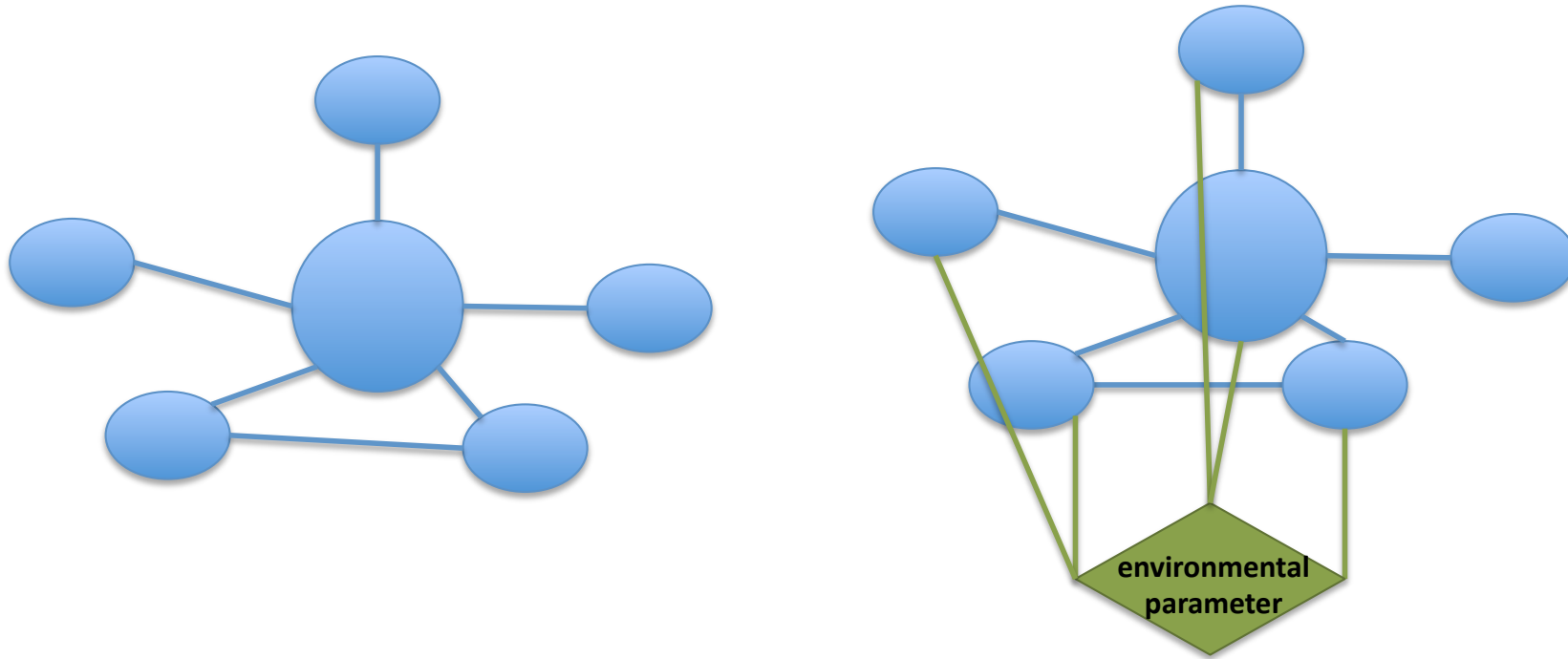
ecological relationships



niche overlap



Network topology and impact of species removal – environmental parameters



- integrating environmental parameters (pH, temperature, salinity etc.) in the network analysis can help to disambiguate edge interpretation
- considering environmental parameters may also pinpoint species that are sensitive to changes in these parameters (regardless of their number of links to other species)

Networks as input of community models

- network inference techniques can help to determine interaction strengths between species from time series data
- interaction strengths needed by some community models
- community models can predict the outcome of species removal

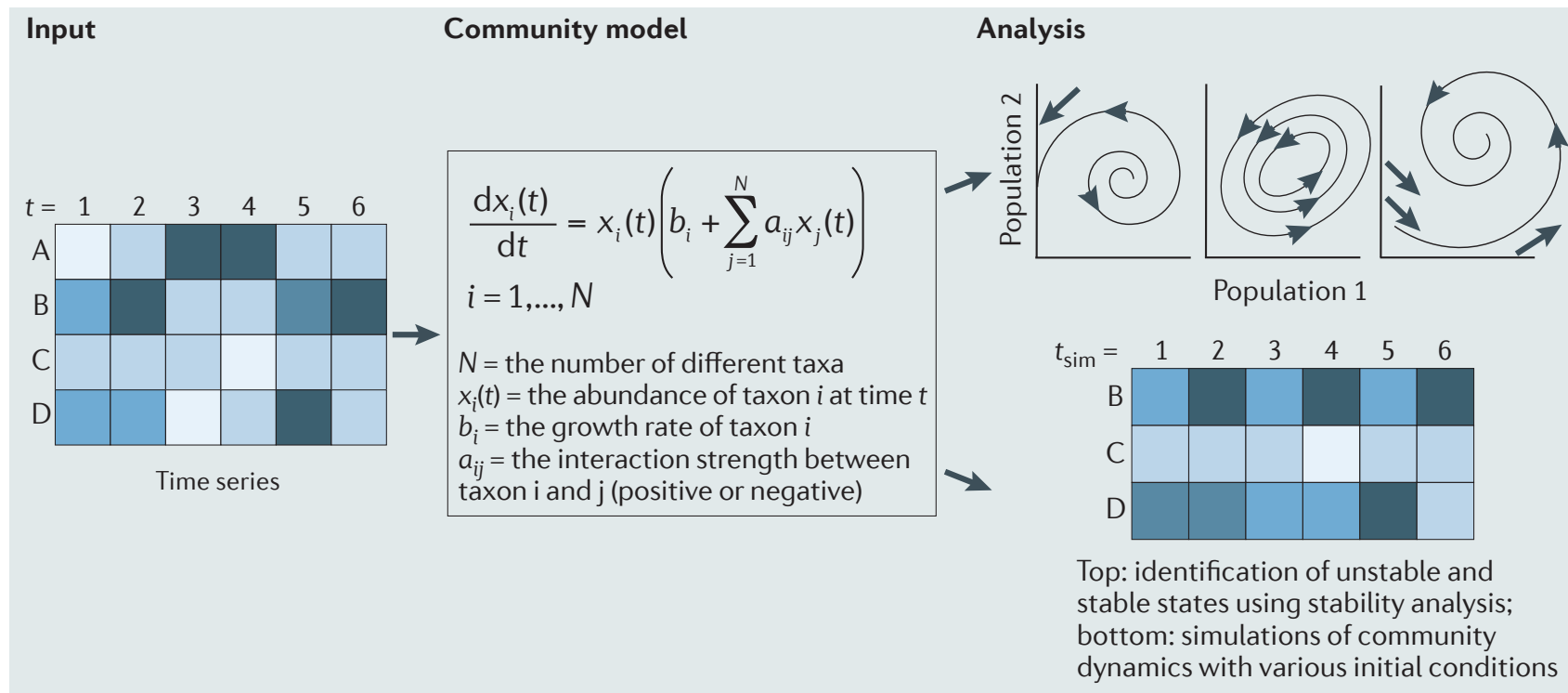


Image taken from: Faust et al. Nature Review Microbiology, Vol. 10 , pp. 538-550 2012.

Acknowledgement

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



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



Curtis Huttenhower



Fah Sathirapongsasuti



Nicola Segata



HMP Consortium for data access

Funding



Bacterial abundances from 16S reads

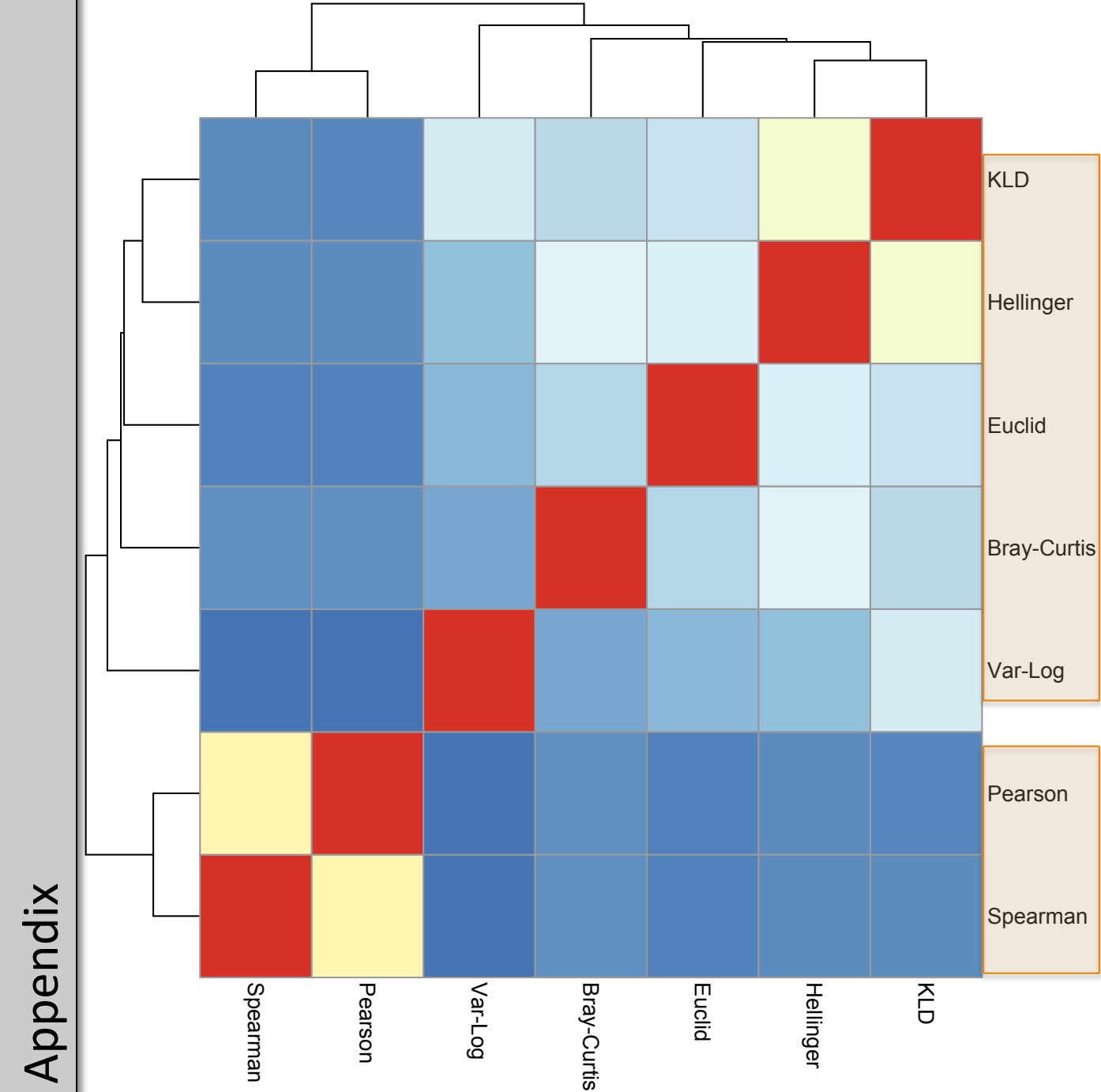
- raw 16S rRNA reads were processed by Pat Schloss with his **mothur** pipeline
- processing steps included sequence trimming (primers and barcodes removal), filtering (of ambiguous bases, homo-polymers and redundant sequences) and chimera removal (with ChimeraSlayer)
- mothur assigned reads to ~730 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)

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)

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

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

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

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

Euclidean distance

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

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

Require pseudo-counts or smoothing because $\log(0) = -\text{Inf}$

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

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

Definition of measures continued

Variance of log-ratios $d(x, y) = \text{var}\left(\log\left(\frac{x_i}{y_i}\right)\right)$

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 $\log(0) = -\text{Inf}$

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

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.

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).

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

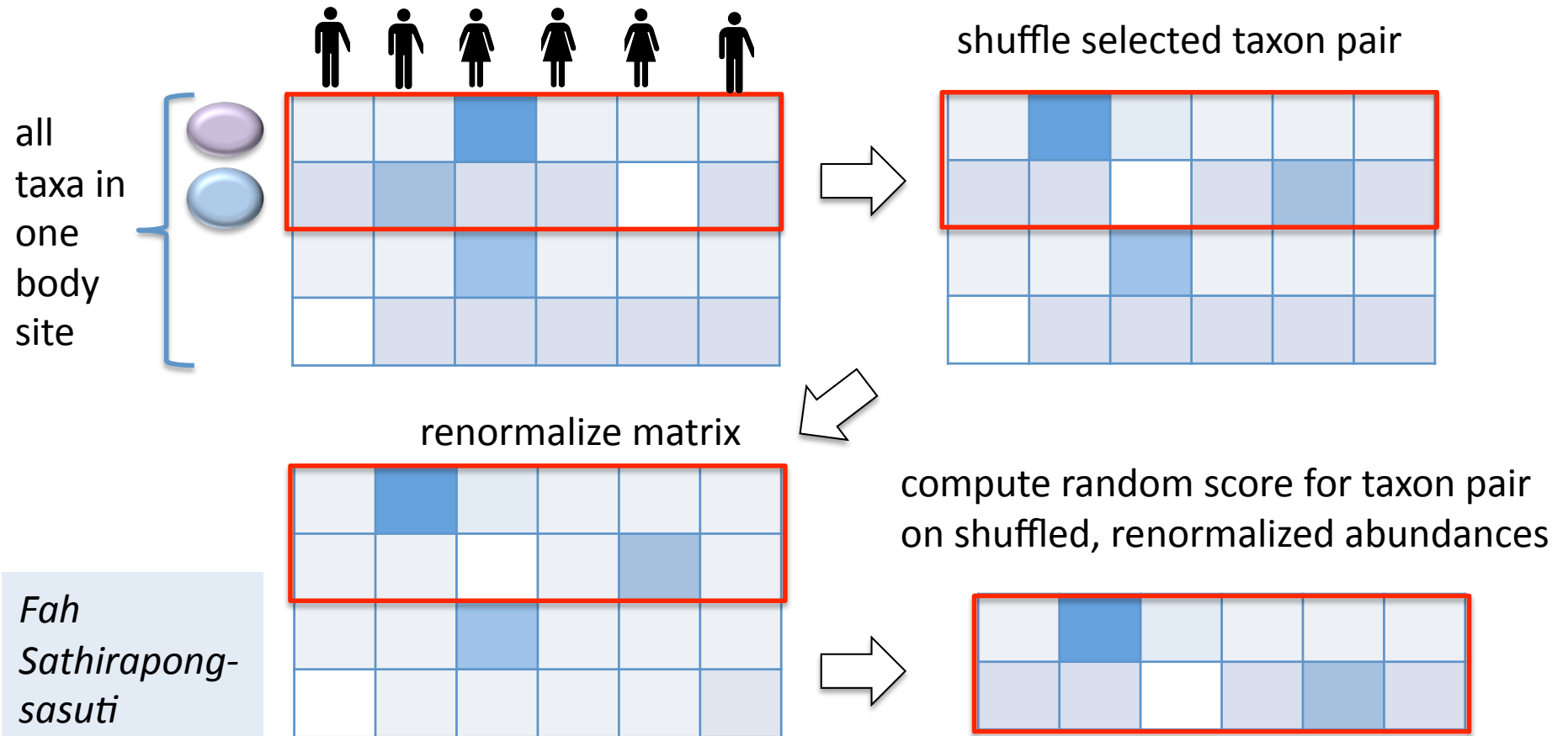
$$AR^2 = 1 - (1 - R^2) \frac{n - 1}{n - p - 1}$$

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

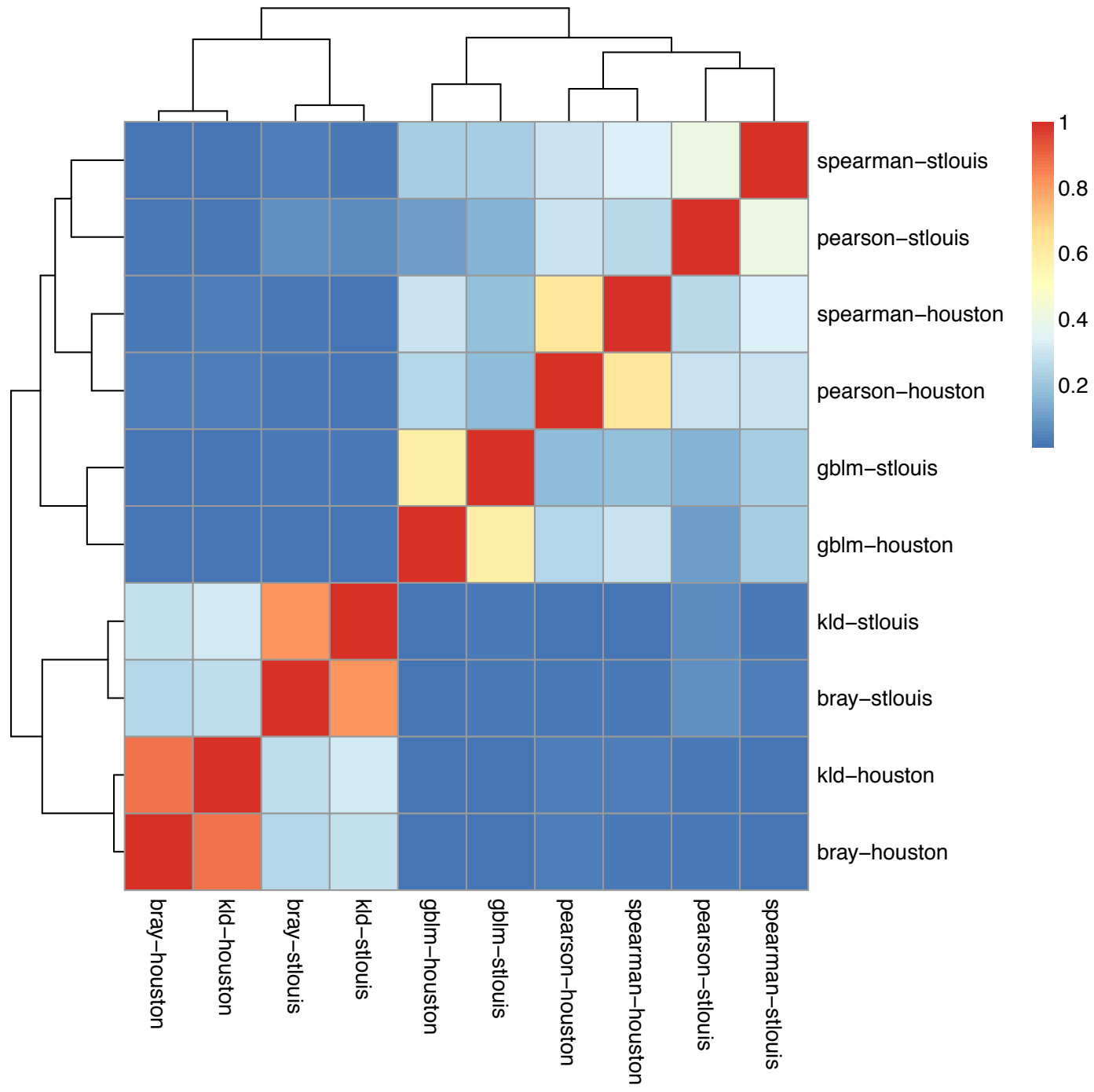
Cross-validation

- 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

ReBoot: Adjust null distribution to mitigate the compositionality bias



Agreement between data and methods



Appendix