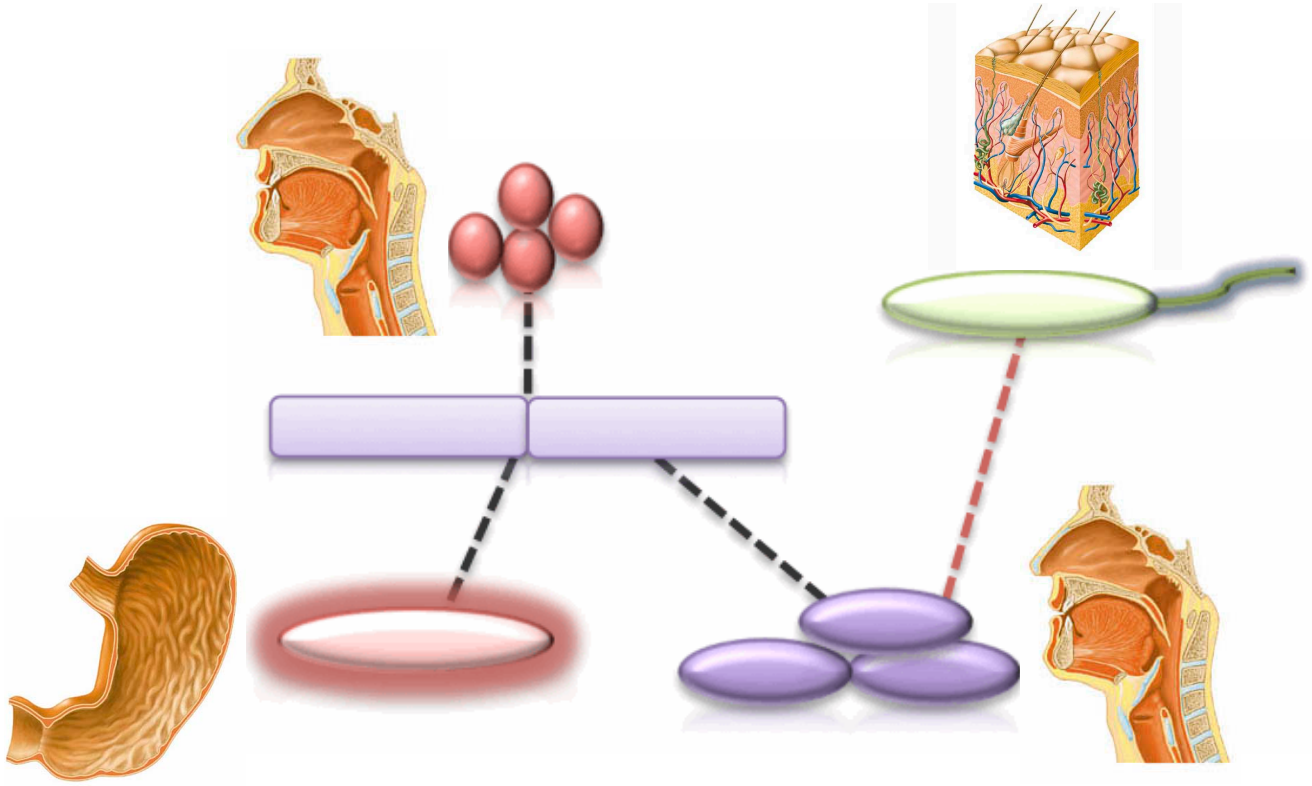
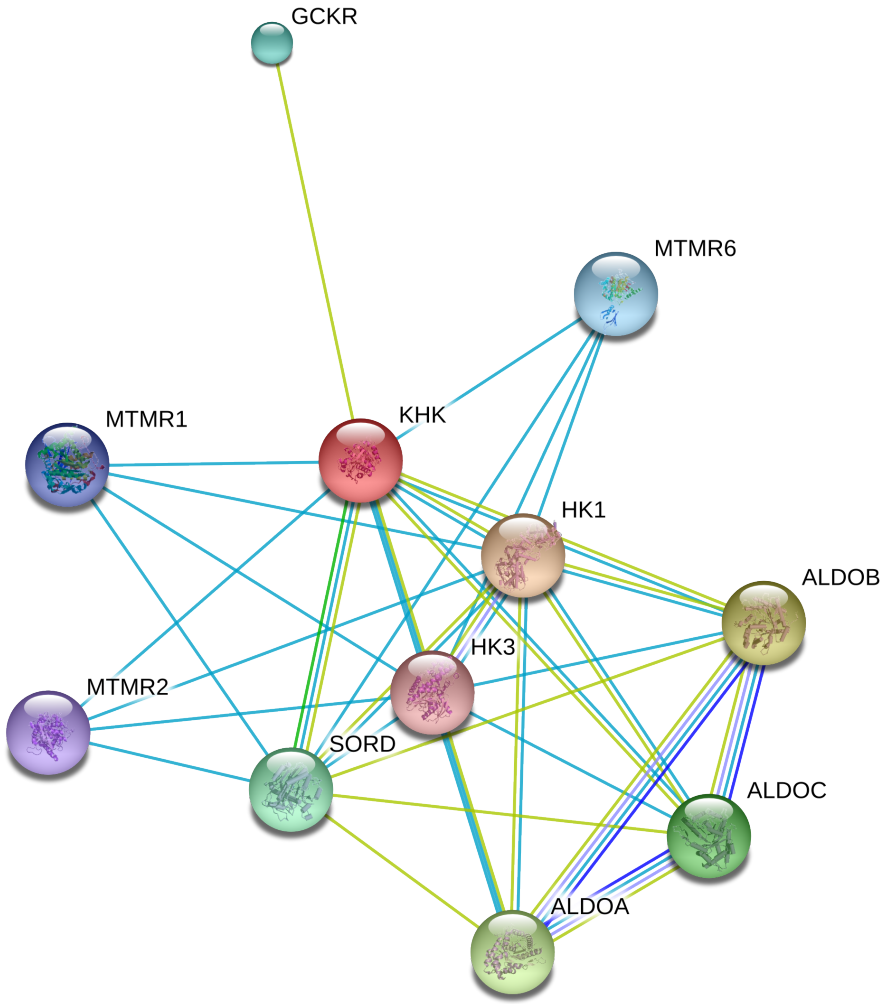
Karoline Faust
PI: Jeroen Raes

20 April 2012
VIB Seminar

Microbial co-occurrence relationships in the human microbiome



Network representation in bioinformatics

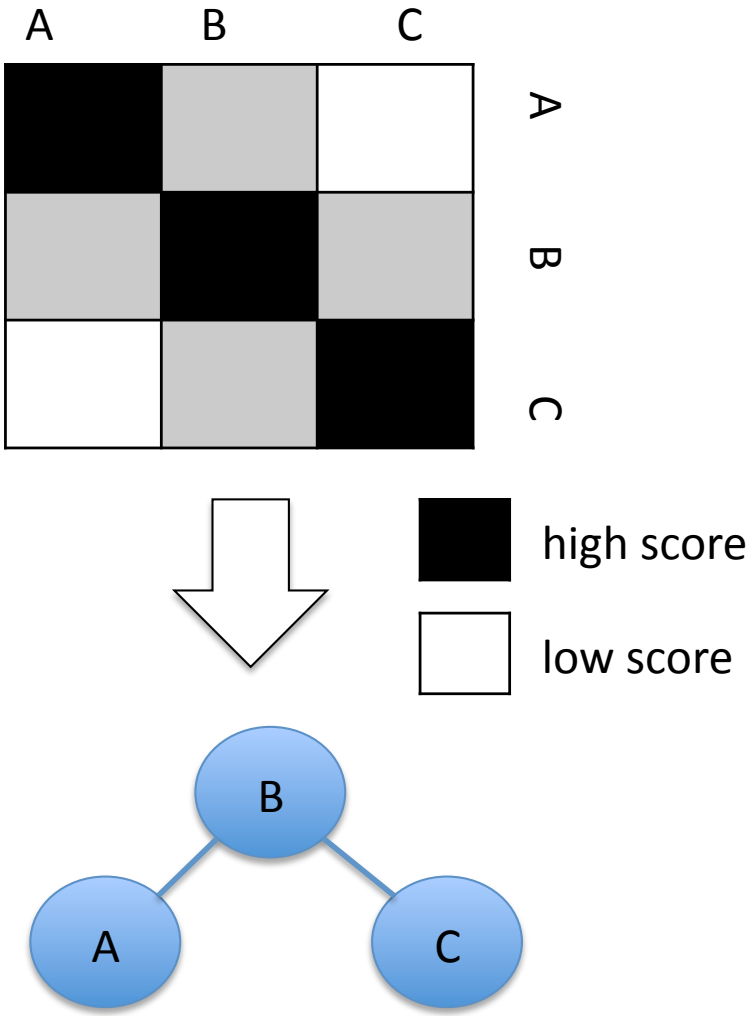


interactions (derived from various data sources) of human fructokinase (KHK=ketohexokinase) with other proteins, obtained with STRING

- nodes represent biological objects (genes, proteins, metabolites...)
- edges represent relationships between objects and may be weighted (according to the strength of the relationship)
- edges may be of different types (according to source that supports relationship)

Szklarcyk et al. (2011) "The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.", Nucleic Acid Research, vol. 39, pp. D561-D568.

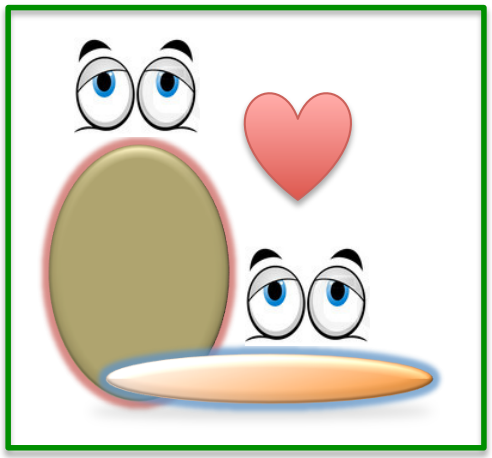
Network inference in a nut shell

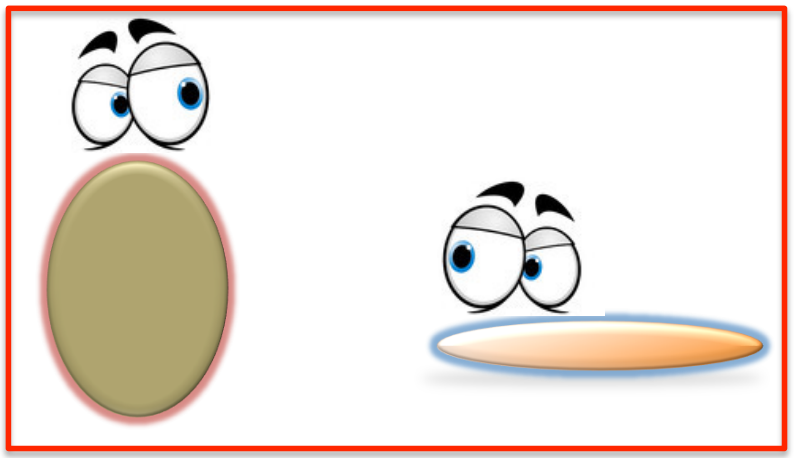


- biological network inference: the problem to find relationships between biological objects (genes, proteins, metabolites, species...)
- a network is built from a similarity matrix that describes all pair-wise relationships between objects
- the inferred network is a representation of the filtered similarity matrix

Goal: Infer network of microbial relationships

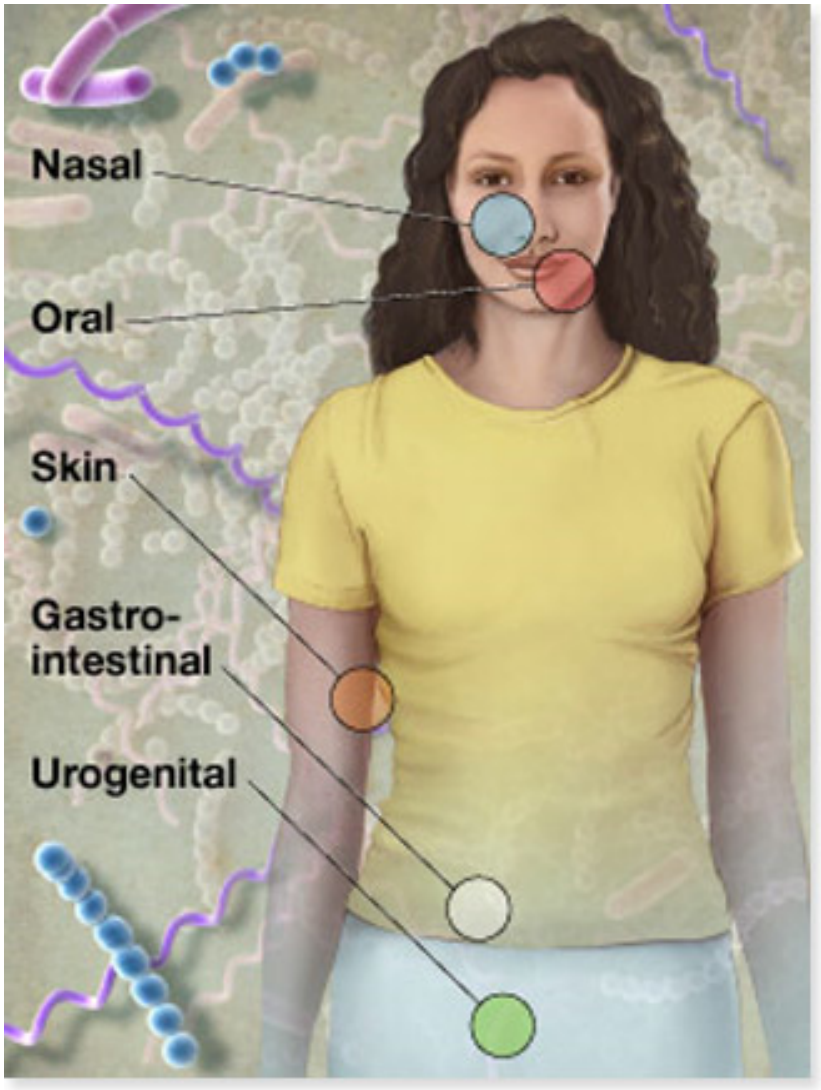
- several recent metagenomic data sets measure microbial abundance across a large number of samples
- network inference techniques can identify significant relationships between microorganisms from these data
- significant co-presence (co-occurrence of two microbes across samples) can be interpreted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





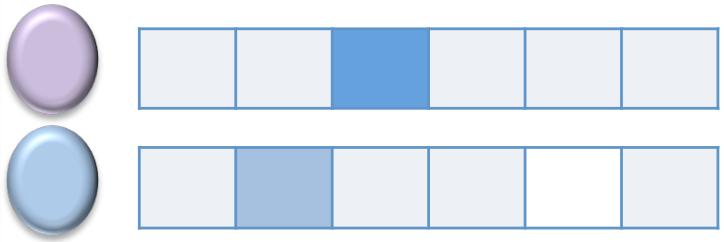
The Human Microbiome Project

- 18 body sites (15 male sites)
- 239 healthy individuals sampled multiple times
- 16S rRNA: 5,366 samples were 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
- whole genome shotgun: 736 samples were illumina-sequenced (Illumina HiSeq 1000)
- phylotypes (with resolution down to genus-level) obtained from 16S data with mothur pipeline (Pat Schloss)

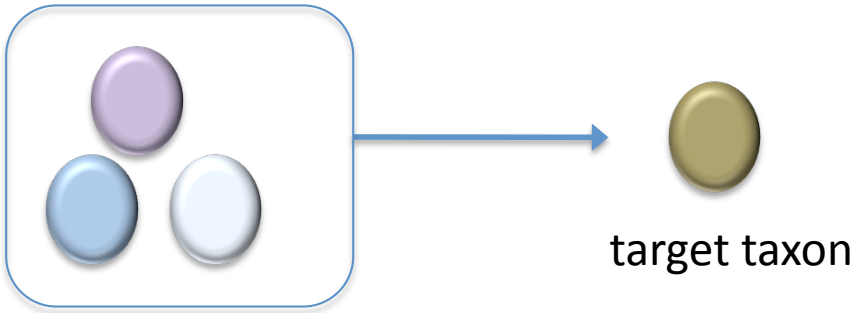


The NIH HMP Working Group et al. "The NIH Human Microbiome Project." Genome Research 2009.

Assessing strength of relationships between microorganisms



abundance profiles across samples



source taxa

Pair-wise relationships

- Pearson correlation
- Spearman correlation
- Kullback-Leibler dissimilarity
- Bray Curtis dissimilarity

Complex relationships

- 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)

J. Fah Sathira-pongsasuti, Curtis Huttenhower

4. Methods

Assessing significance of relationships and building the network

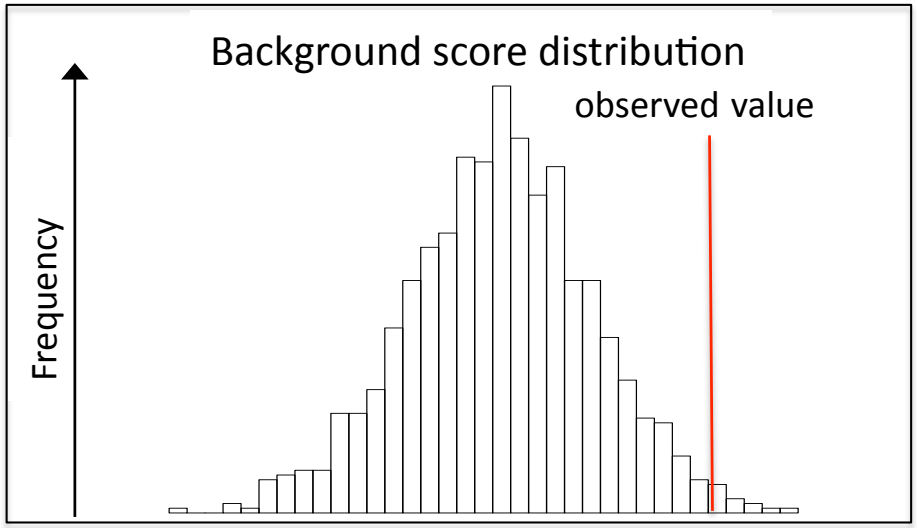
Repeat score computation for each measure and each relationship 1,000 times on permuted data (background score distributions)

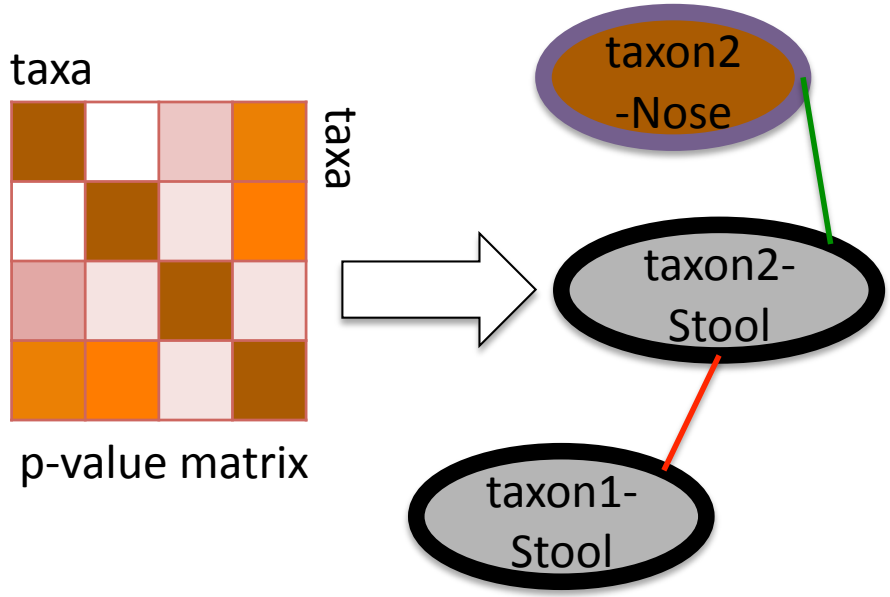
Compute p-values from background score distributions

Merge measure-specific p-values using Simes' method

Multiple-test-correct p-values (using Benjamini-Hochberg-Yekutieli) and discard all relationships with final p-values above selected significance level (0.05)

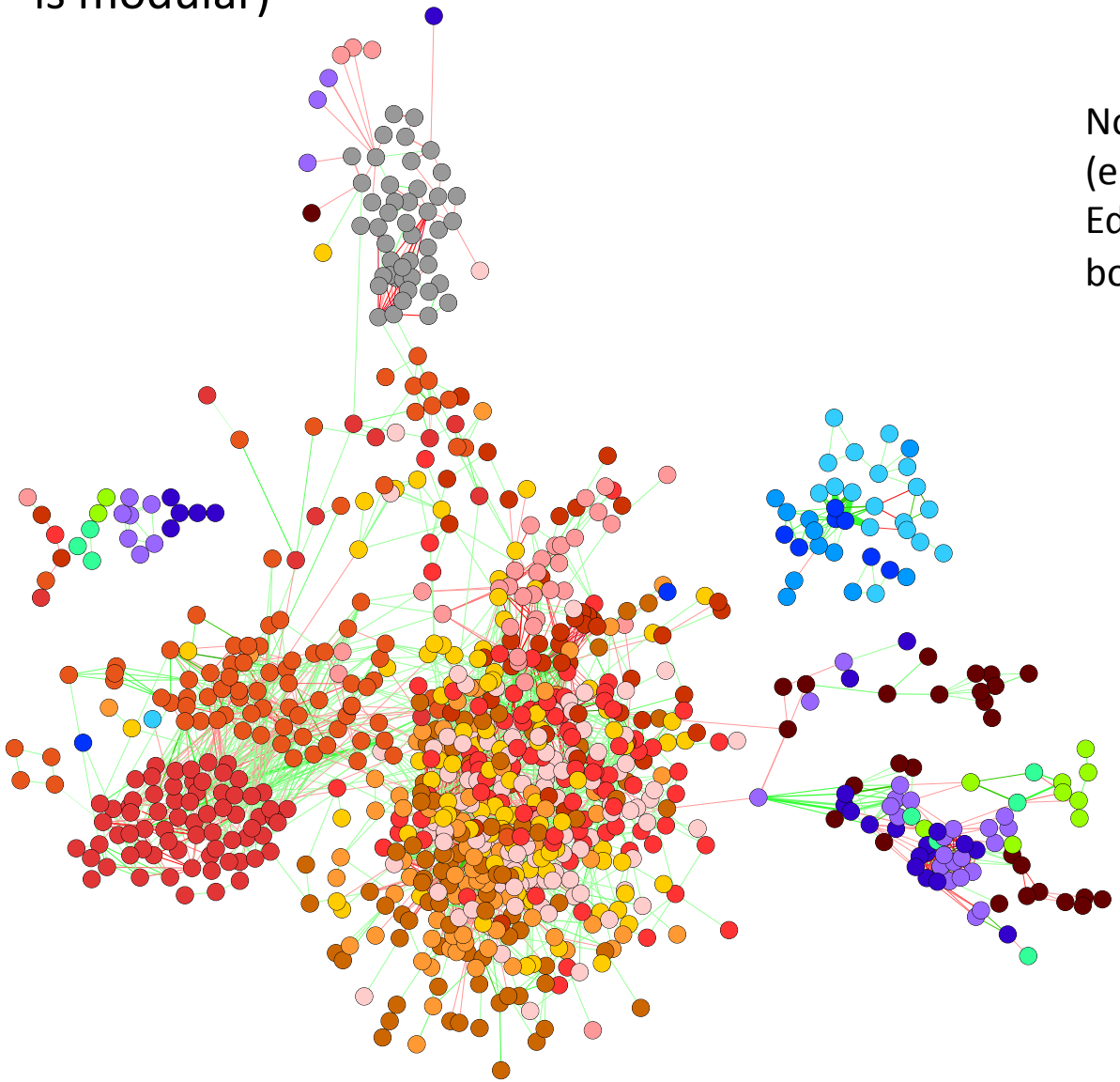
Draw remaining relationships as a network





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

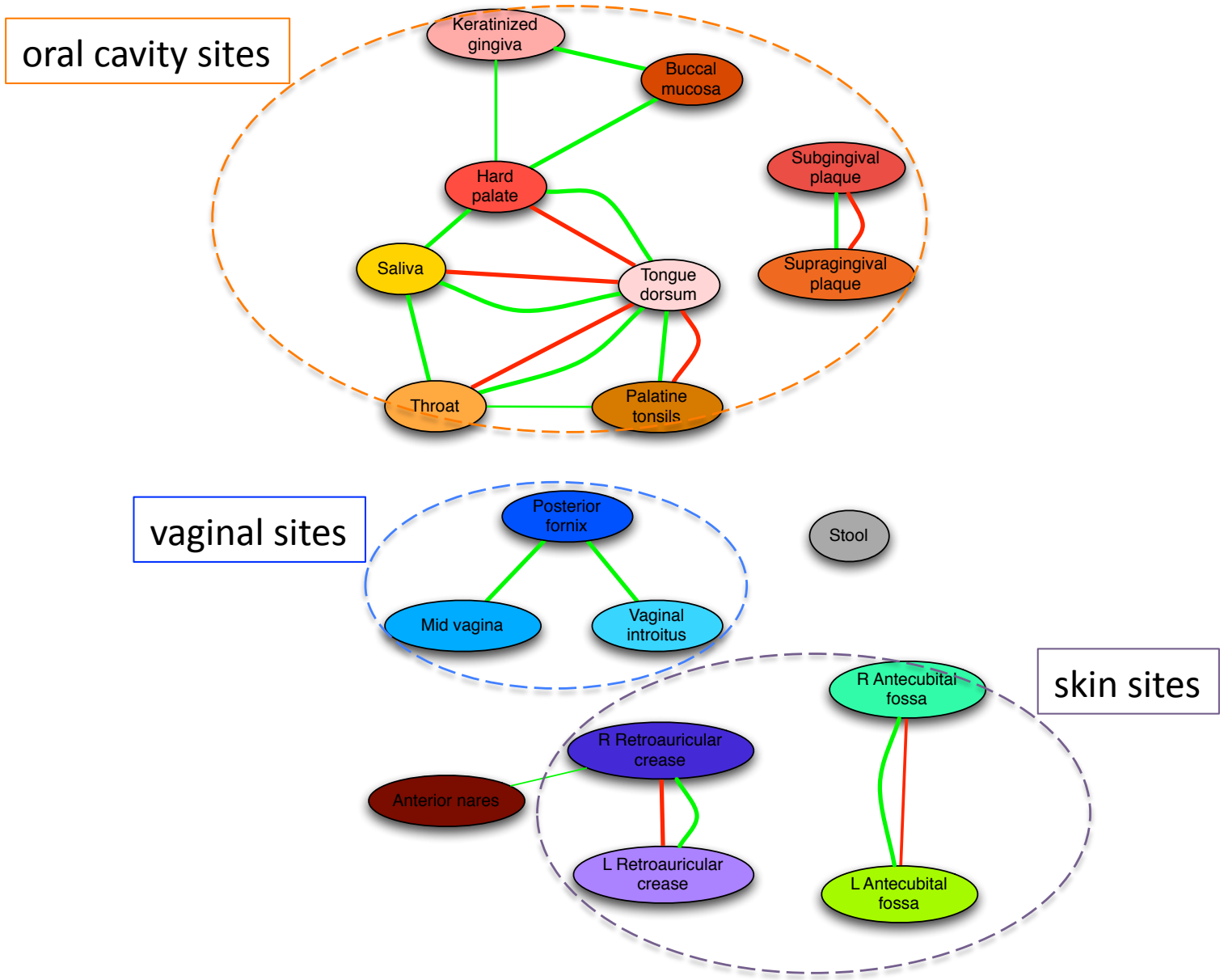
Mid vagina
Posterior fornix
Vaginal introitus

Edge color code

positive

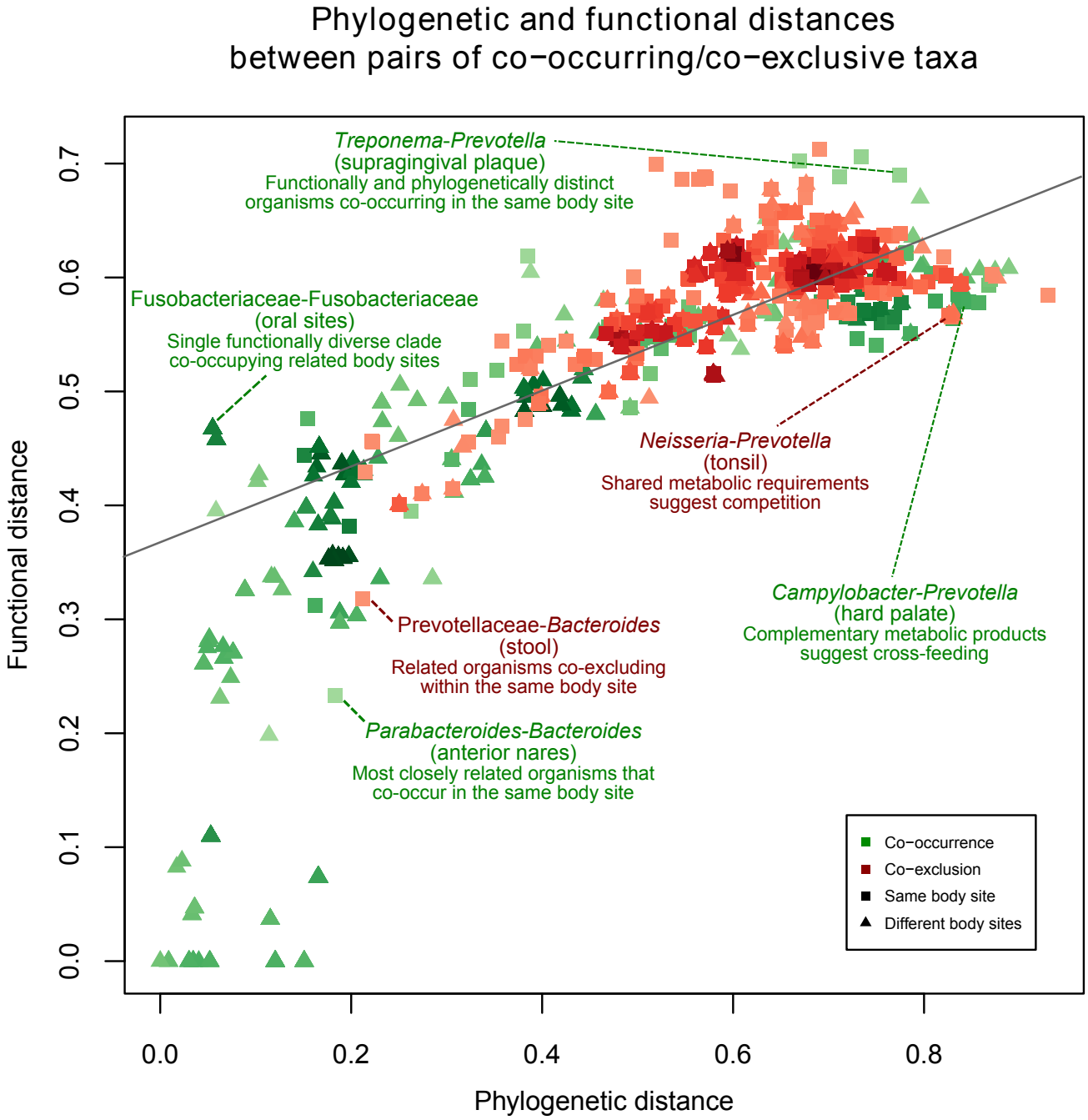
negative

HMP 16S phylotypes network – body-site relationships



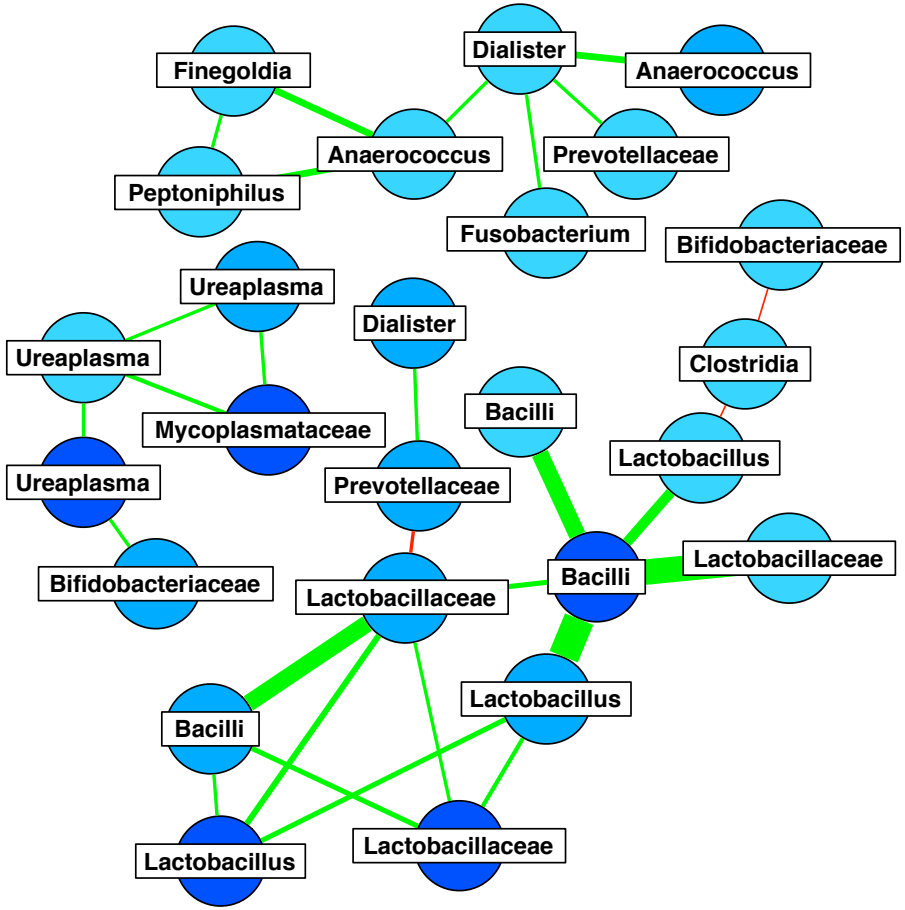
5. Results

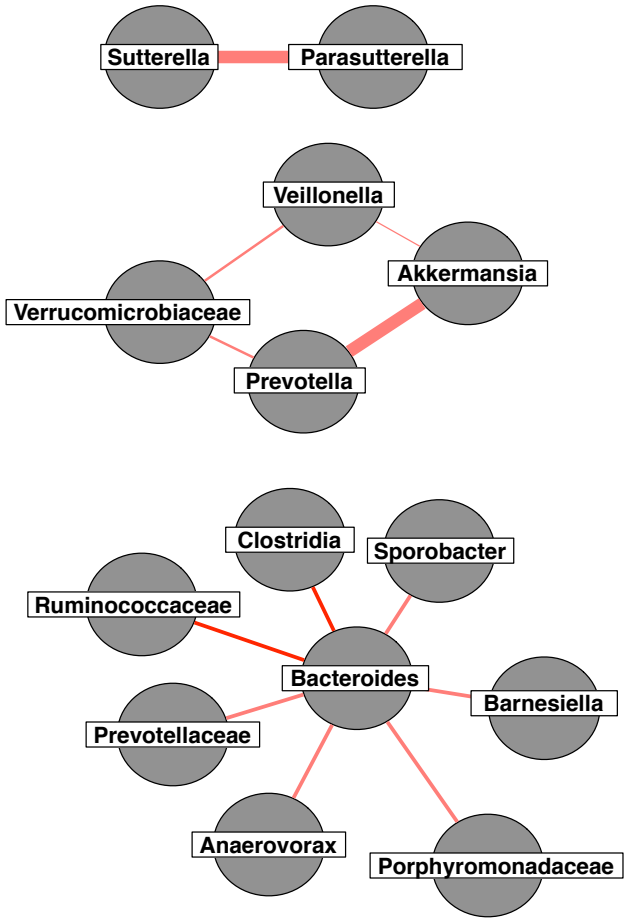
HMP 16S phylotypes functional analysis



J. Fah Sathira-pongsasuti, Nicola Segata

Known alternative communities captured



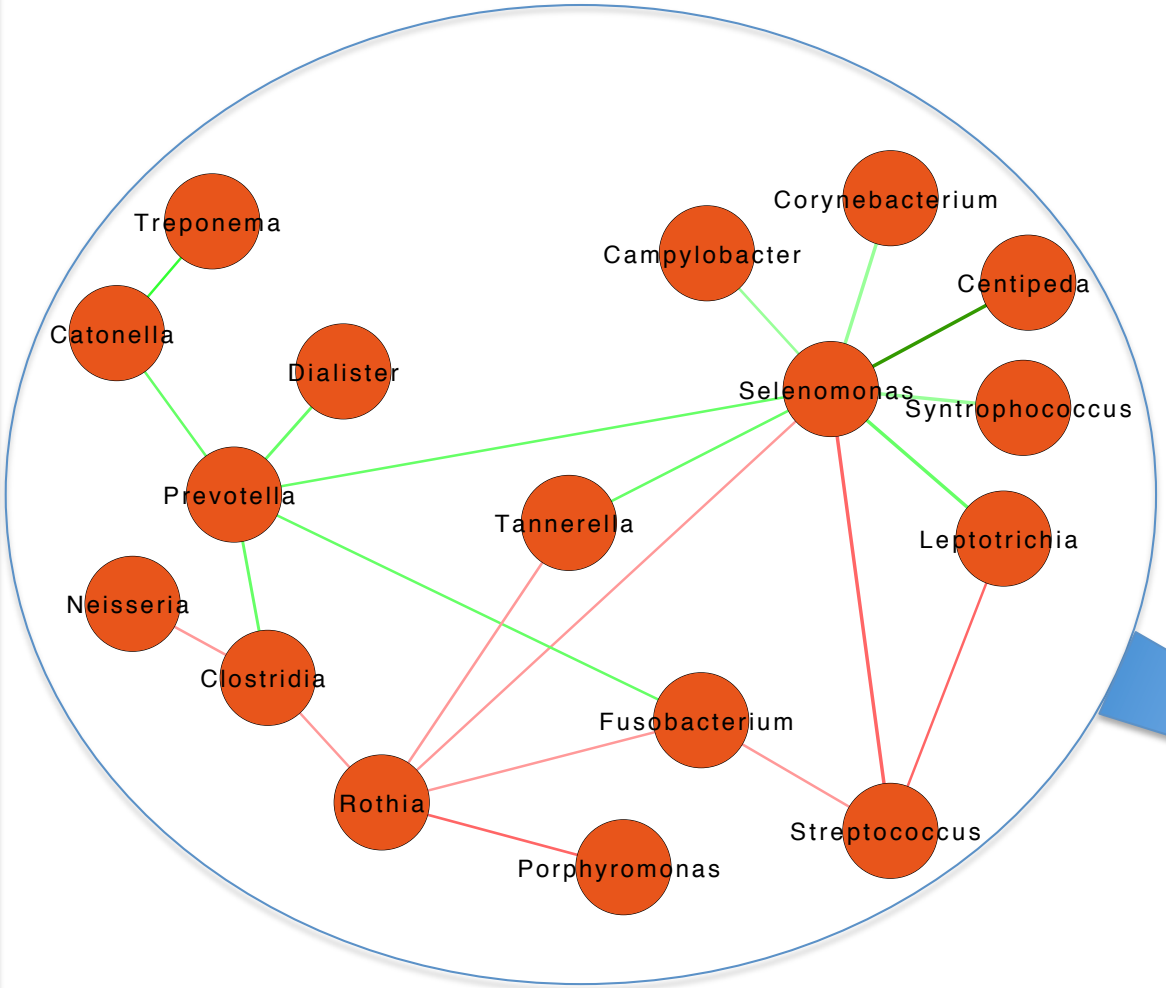


Vagina (Ravel et al.): 5 community types, 4 dominated by different *Lactobacillus* species, one diverse

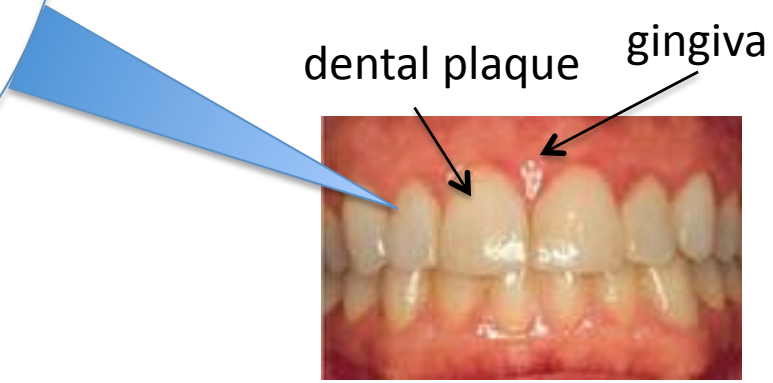
Gut (Arumugam, Raes et al.): 3 enterotypes, driven by *Ruminococcus*, *Bacteroides* and *Prevotella*

Ravel, J. et al. (2011) "Vaginal microbiome of reproductive-age women", PNAS, vol. 108, pp. 4680-4687.
Arumugam, M., Raes, J. et al. (2011) "Enterotypes of the human gut microbiome", Nature, vol. 473, pp. 174-180.

Stages of dental plaque formation captured



early colonizers (*Streptococcus*) have negative relationships with intermediate (*Fusobacterium*) and late colonizers (*Selenomonas*)



Kolenbrander, P.E. et al. "Communication among Oral Bacteria". *MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS* 66, 486-505 (2002).

5. Conclusions

Conclusions

- few cross-body-area relationships (modular network): different body areas harbor distinct microbiota
- body sites can be classified into different microbial niches based on cross-links between their microbiota: oral, skin and vaginal sites form separate clusters, airways and stool separated from the oral cavity
- alternative microbial community configurations previously observed in the vagina and the gut detected as mutual exclusions
- successional stages in dental plaque formation captured as mutual exclusions
- closely related microbes tend to form positive relationships (mostly between related body sites), whereas most negative relationships occur between more distantly related microbes

Acknowledgement

Acknowledgement

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

Curtis Hutten-hower

J. Fah Sathira-pongsasuti

Nicola Segata

Dirk Gevers, Broad institute

Jacques Izard, Forsyth institute

HMP

HMP Consortium for data access

...and Alvin Lo for his comments on dental plaque formation and Dominique Maes for discussions on normalization

Appendix

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 (genus-level) using the Ribosomal Database Project (RDP) reference 16S rRNA sequences and the RDP phylogenetic tree
- mothur also assigned reads to ~9450 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

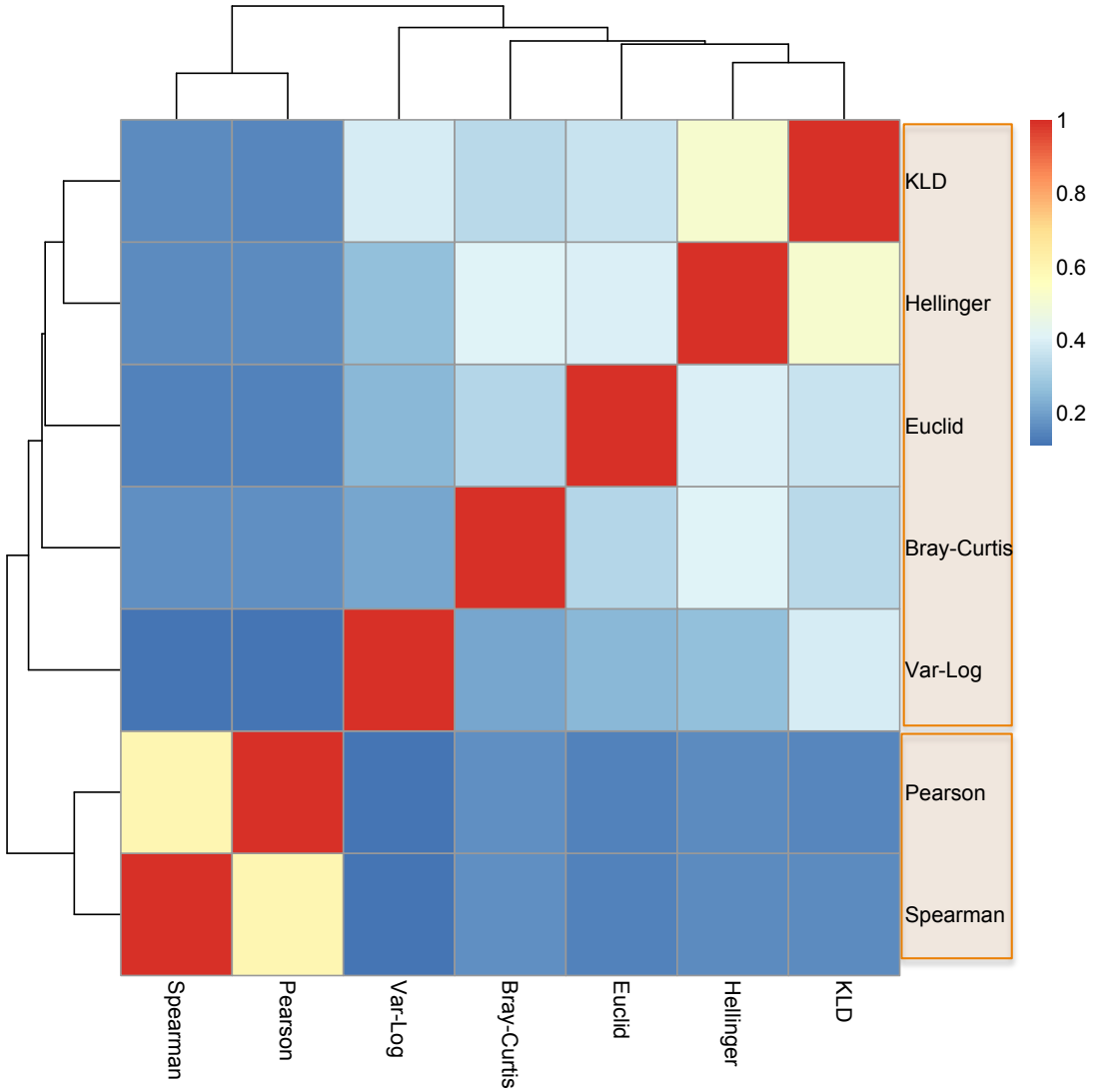
Cole, J.R. et al. (2009) "The Ribosomal Database Project: improved alignments and new tools for rRNA analysis", Nucleic Acid Research, vol. 37, pp. D141-D145

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

Appendix

Selection of score functions

Experiment: Compute the top 1,000 and bottom 1,000 relationships for several measures in the 16S HMP Houston data set



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

Appendix

Definition of score functions

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

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

Kullback-Leibler
(*x* and *y* 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

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

Variance of log ratios

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

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

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

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

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

Appendix

Definition of score functions

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.

Mutual information

$$I(x,y) = \sum\sum p(x,y)\log\left(\frac{p(x,y)}{p_1(x)p_2(y)}\right)$$

Measures (potentially non-linear) dependency between two vectors ("generalized correlation")

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
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

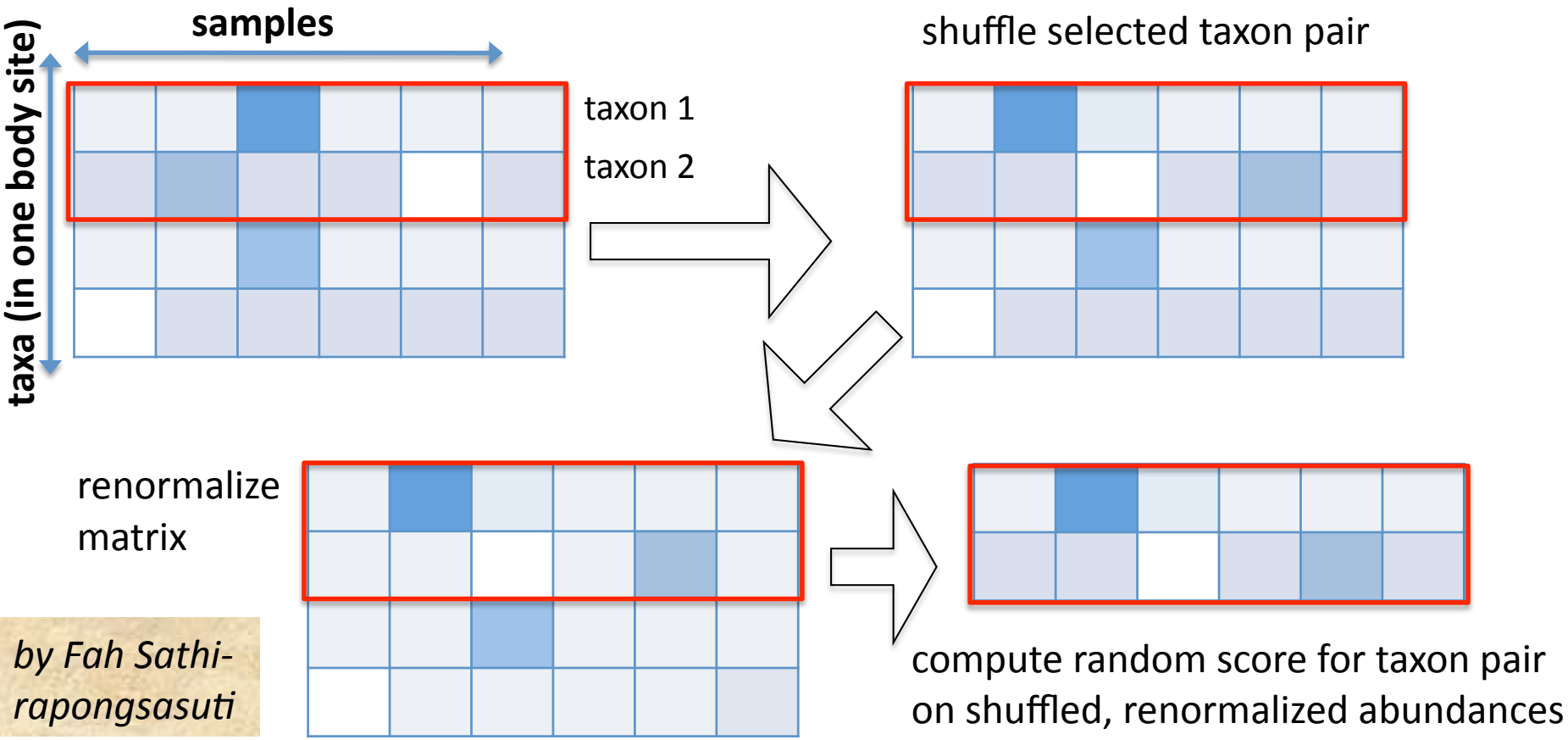
Appendix

Work-around the compositionality bias

Idea: capture impact of compositionality bias when computing edge-specific null distribution

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)



by Fah Sathi-rapongsasuti

Appendix

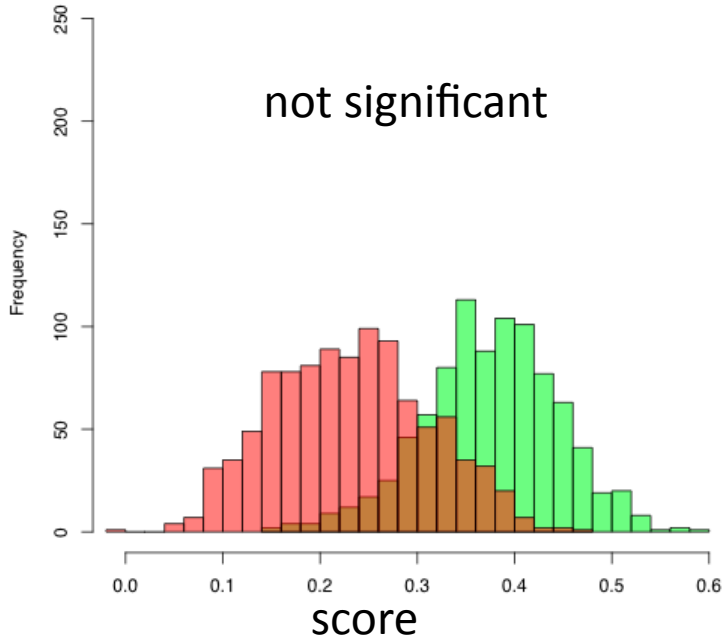
Combining null and bootstrap distributions to compute p-values

Bootstrap distribution gives the confidence interval of the observed score.

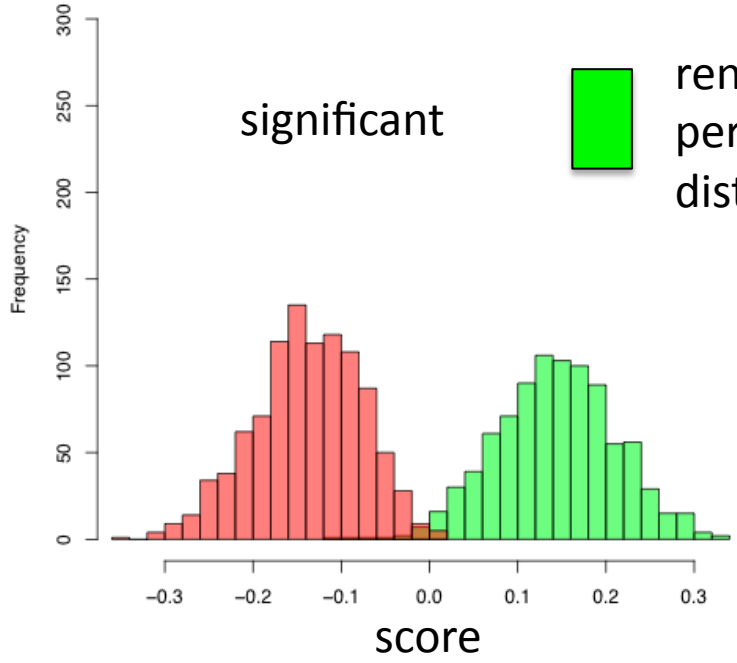
Edge-specific p-value is computed with a **Z-test** (p-value of the bootstrap mean given the null distribution, assuming normality for the null distribution)

bootstrap distribution

renormalized permutation distribution



Fusobacteriales versus Streptococcaceae in buccal mucosa (Pearson)



Actinobacteria versus Bacteroidetes in subgingival plaque (Spearman)

Appendix

Agreement between data and methods