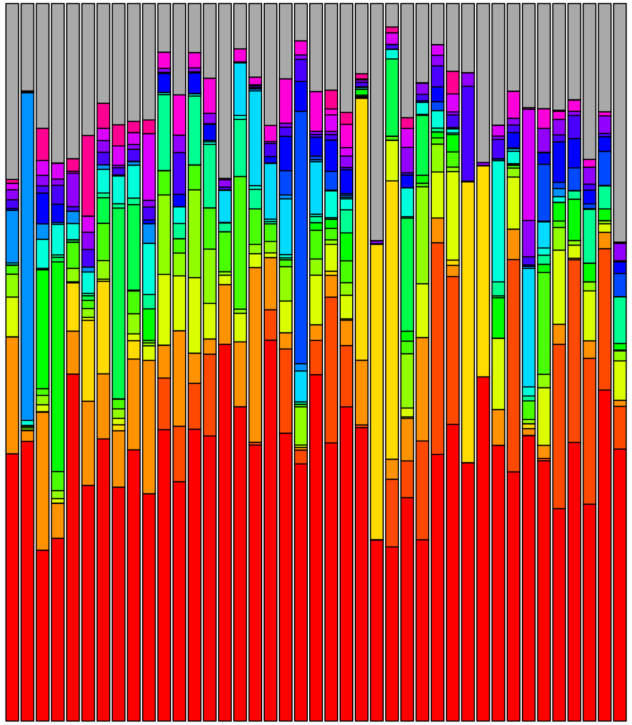
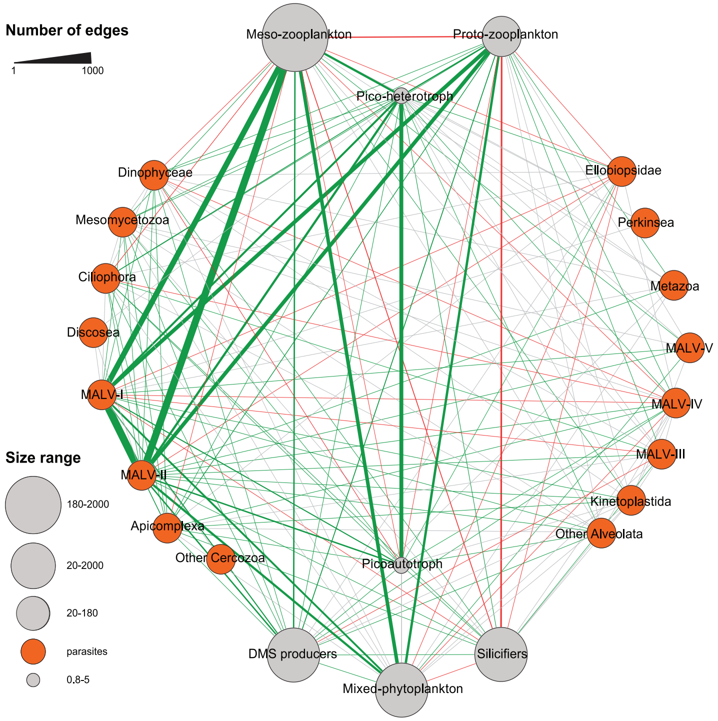
Karoline Faust
msysbiology.com

KU LEUVEN

Inference and analysis of microbial networks from sequencing data

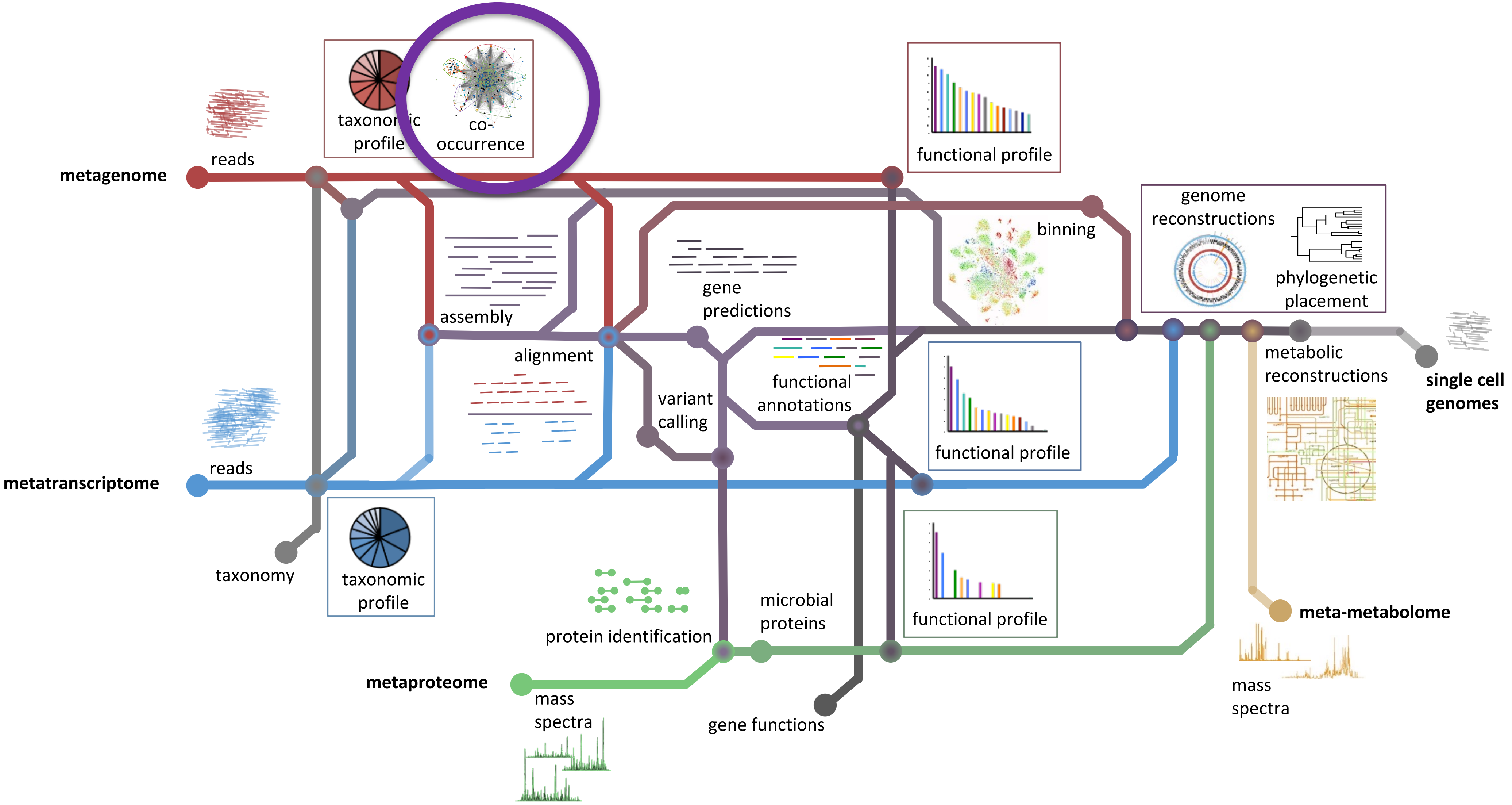
Integrated multi-omic analyses of microbial communities
23-29 April 2022
Luxembourg



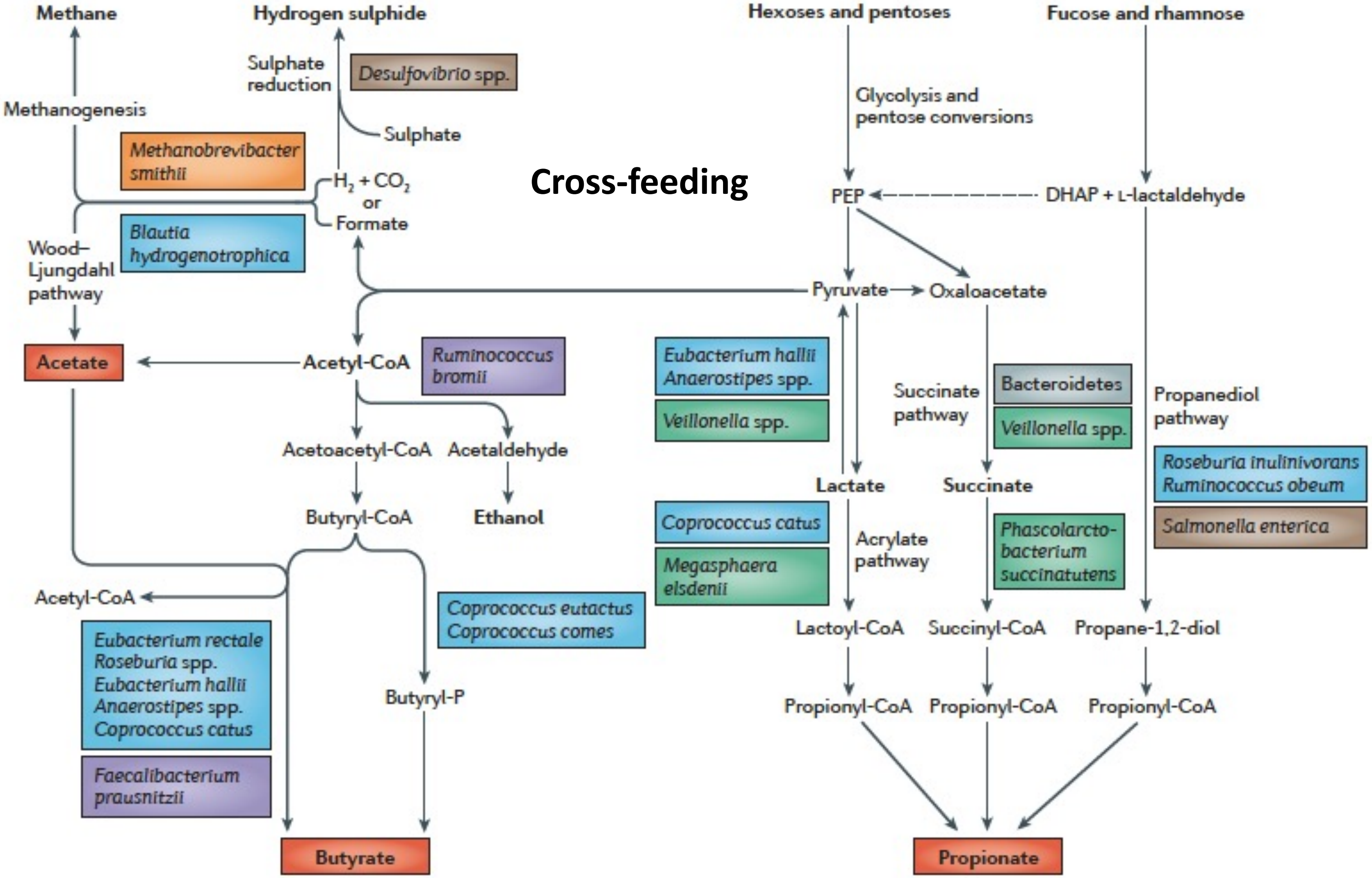


EMBO
excellence in life sciences

29th April 2022


reads
taxonomic profile
co-occurrence
functional profile
metagenome
binning
genome reconstructions
phylogenetic placement
assembly
gene predictions
alignment
variant calling
functional annotations
functional profile
metabolic reconstructions
single cell genomes
reads
metatranscriptome
taxonomy
taxonomic profile
functional profile
microbial proteins
protein identification
meta-metabolome
metaproteome
mass spectra
gene functions
mass spectra

Microbes interact

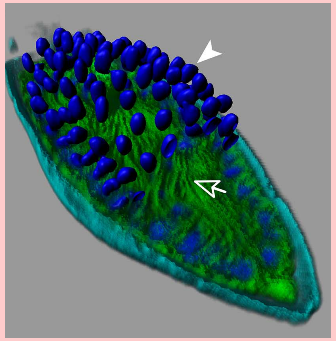


Cross-feeding

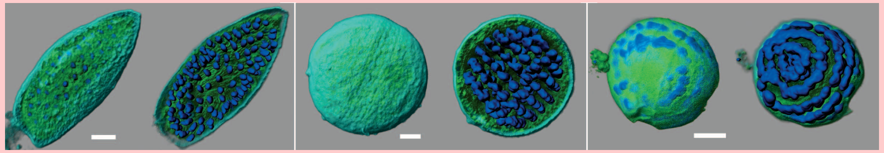
Louis et al. *The gut microbiota, bacterial metabolites and colorectal cancer.* Nature Reviews Microbiology 12:661-672 (2014).

Microbes interact

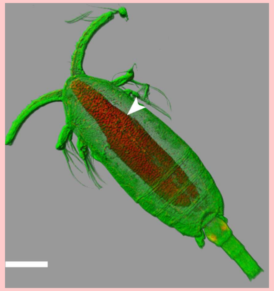
Parasitism



Dinoflagellate infected by Amoebophrya (MALV-II)

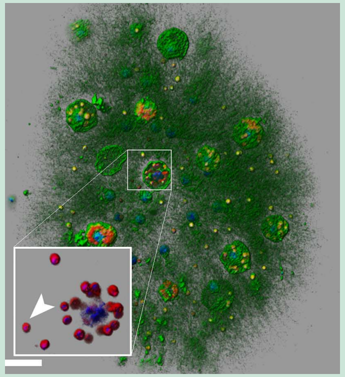


Different dinoflagellate species infected with syndiniales parasites

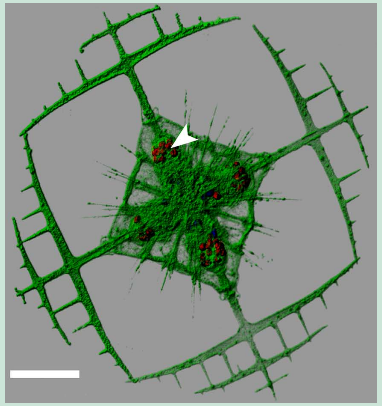


Copepod with parasitic dinoflagellates

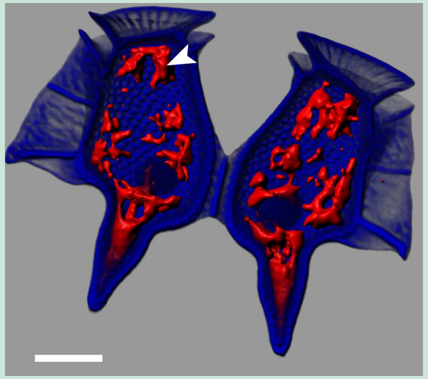
Endosymbiosis



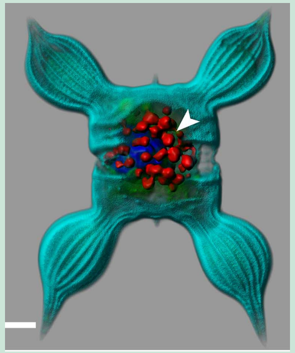
Collodaria colony with dinoflagelllate symbionts



Acantharian with endo-symbiotic Phaeocystis



Dinoflagellate with kleptoplasts



Diatom with chloroplasts

Images taken from de Vargas et al. Science 348, 1261605 (2015).

Introduction

Ecological interactions

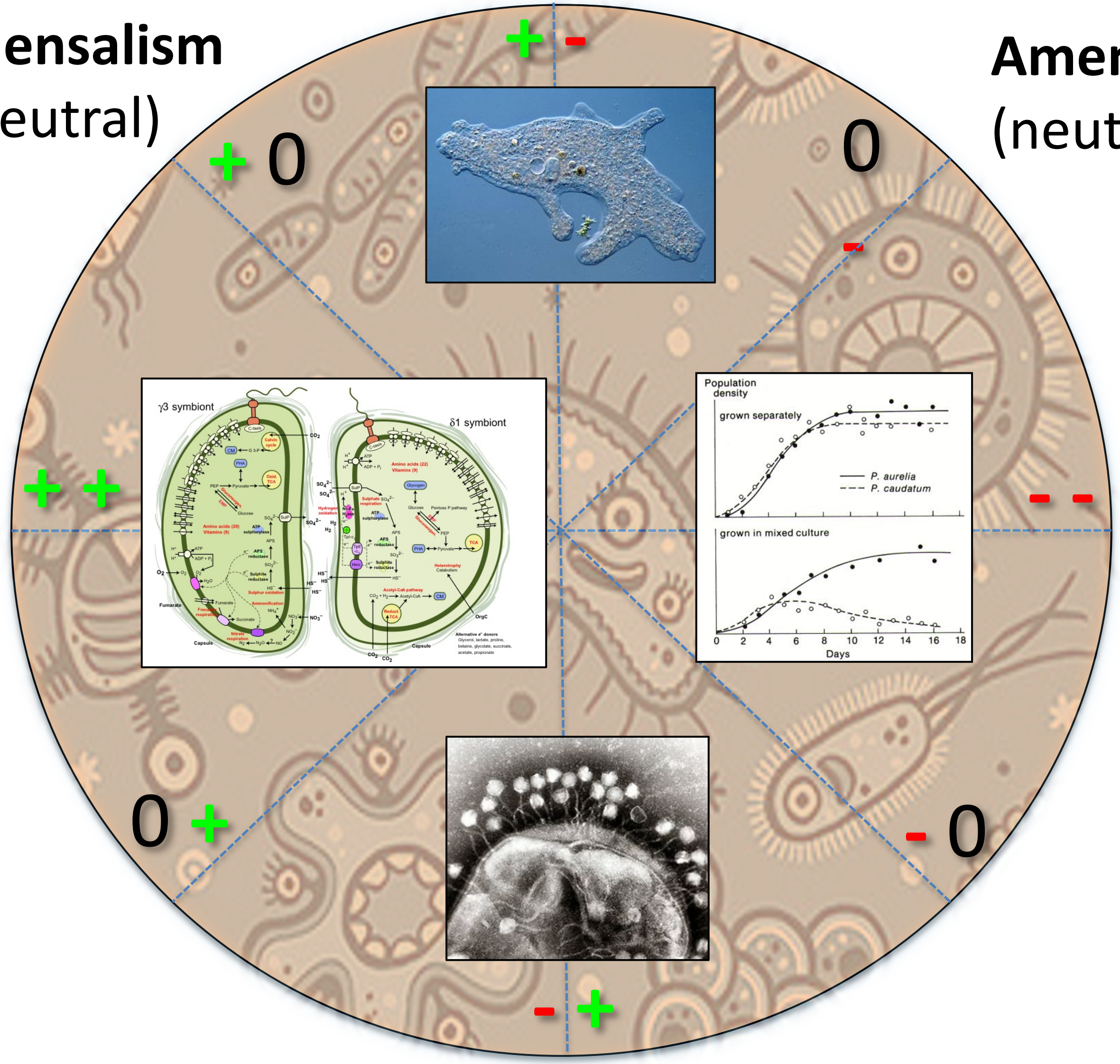
Predator/parasite (win-loss)

Commensalism (win-neutral)

Amensalism (neutral-loss)

Mutualism (win-win)

Competition (loss-loss)

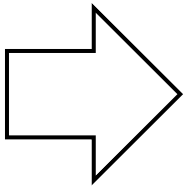


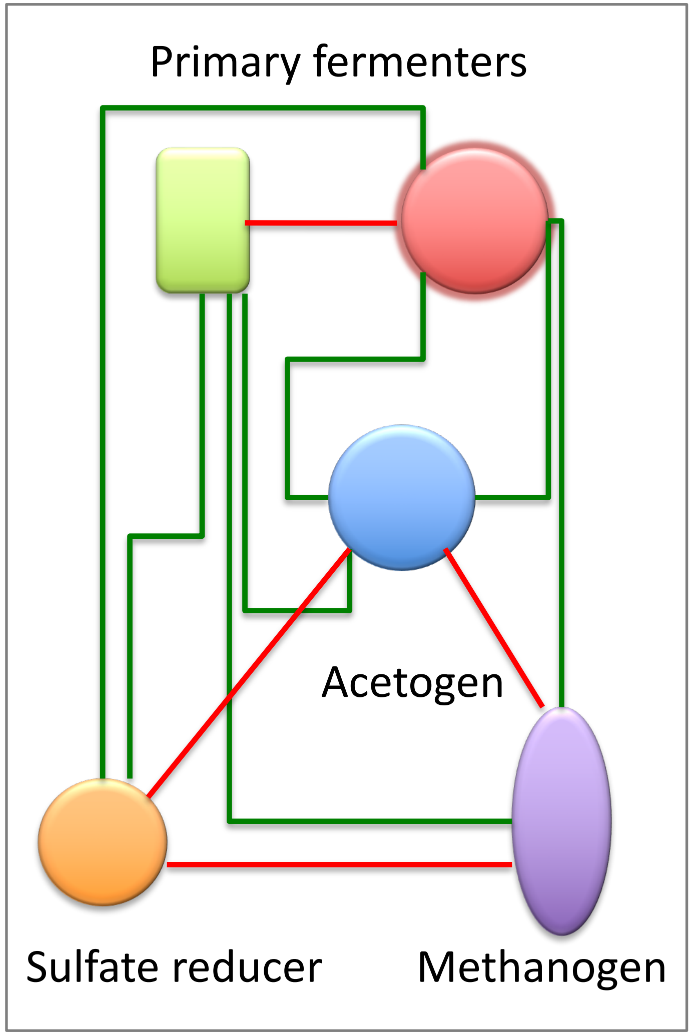
Prey/host (loss-win)

Adapted from Lidicker, W.Z. BioScience 29, 475-477, 1979.

Network representation of microbial communities

Who is there and with which abundance?





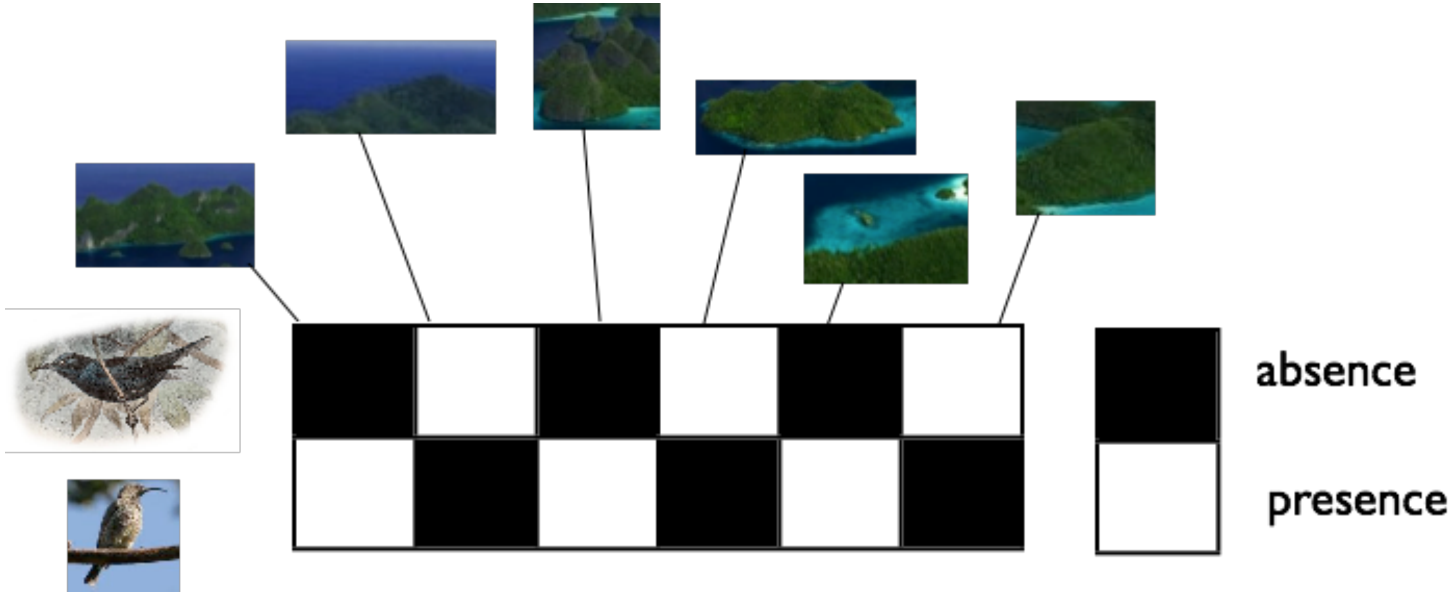
Who interacts with whom?

Co-occurrence analysis to the rescue?

Introduction

History of co-occurrence analysis in ecology

- Jared Diamond suggested assembly rules:
- Rule e: “*Some pairs of species never coexist, either by themselves or as part of a larger combination.*”
- Competition between species can be inferred from their presences/absences across habitats (checkerboard pattern)



Diamond, J. (1975) “Assembly of species communities”, pp. 342-444 in “Ecology and evolution of communities” edited by Cody and Diamond, Harvard University Press.

History of co-occurrence analysis in ecology cont'd

- Connor & Simberloff: "*We challenge Diamond's idea that island species distributions are determined by competition [...]. In order to demonstrate that competition is responsible for the joint distributions of species, one would have to falsify a null hypothesis stating that the distributions are generated by the species randomly [...]*"
- Importance of a **null model**

Connor & Simberloff (1979) "The Assembly Of Species Communities: Chance or Competition", Ecology, 6061, 1132-1140.

Microbial network inference

Co-occurrence analysis is network inference

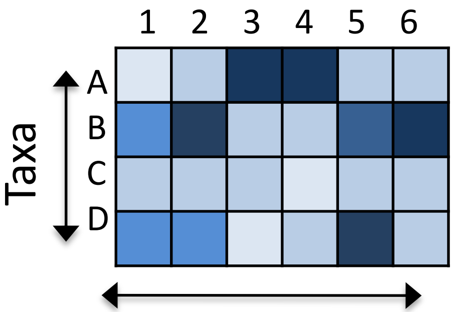
INPUT

Presences/absences (incidences)

Taxa	1	2	3	4	5	6
A	□	□	■	■	□	□
B	■	■	□	□	■	■
C	□	□	□	□	□	□
D	■	■	□	□	■	□

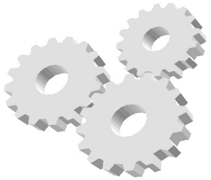
Samples (location or time)

Abundances



Samples (location or time)

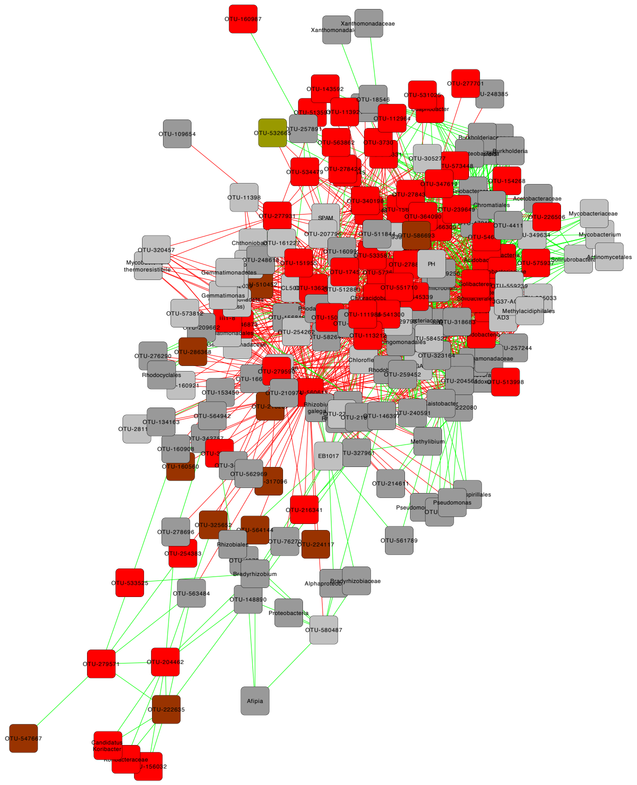
NETWORK INFERENCE



- FlashWeave
- LSA local similarity analysis
- CoNet
- CCLasso
- MDSINE
- SPIEC-EASI

OUTPUT

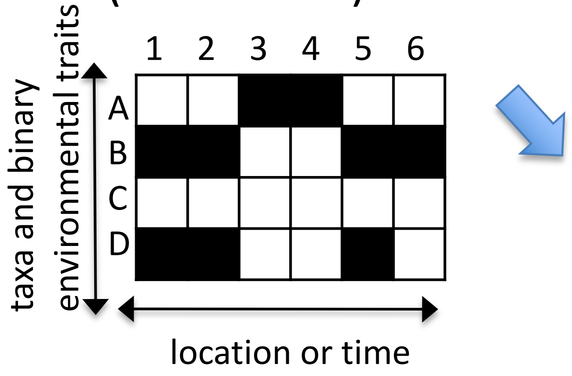
Microbial association network
Nodes: taxa (OTUs, genera, …) or metadata (pH, temperature, …)
Edges: significant associations



Principle of similarity-based network inference

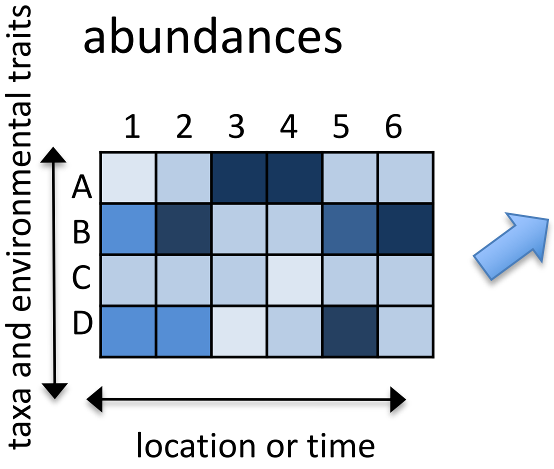
INPUT

presences/absences
(incidences)

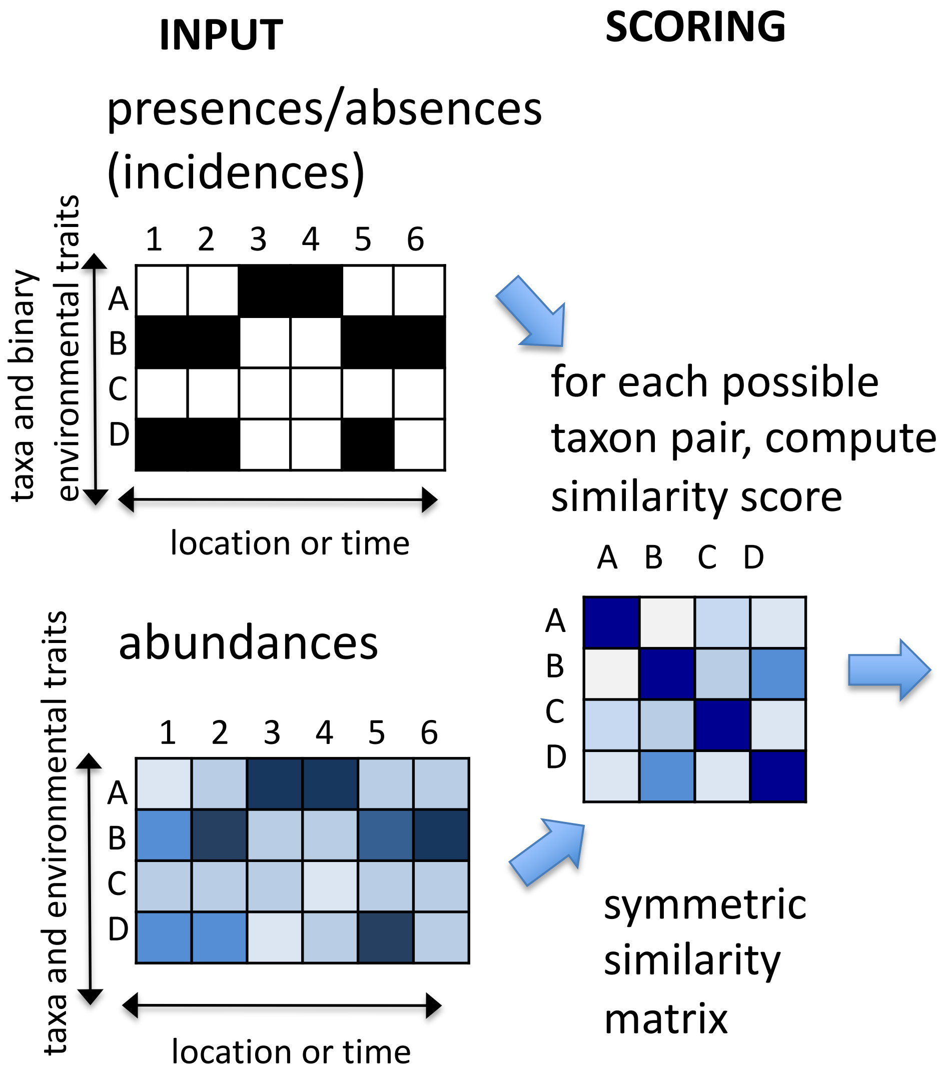


	1	2	3	4	5	6
A	□	□	■	■	□	□
B	■	■	□	□	■	■
C	□	□	□	□	□	□
D	■	■	□	□	■	□

abundances



Principle of similarity-based network inference



INPUT

SCORING

presences/absences (incidences)

abundances

for each possible taxon pair, compute similarity score

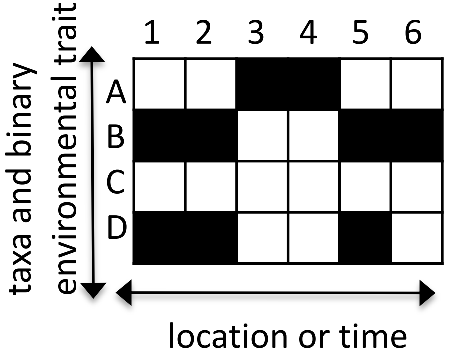
symmetric similarity matrix

Microbial network inference

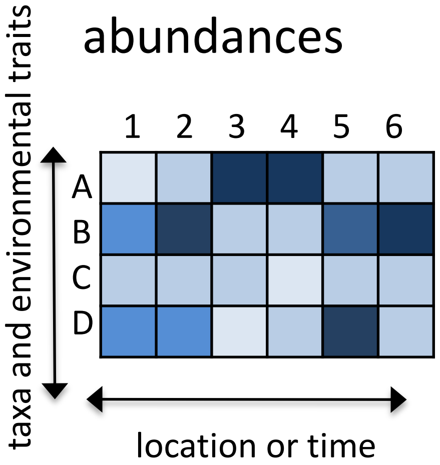
Principle of similarity-based network inference

INPUT

presences/absences (incidences)

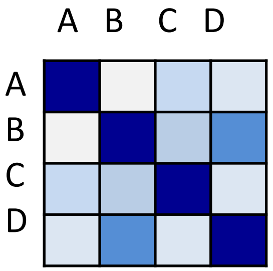


abundances



SCORING

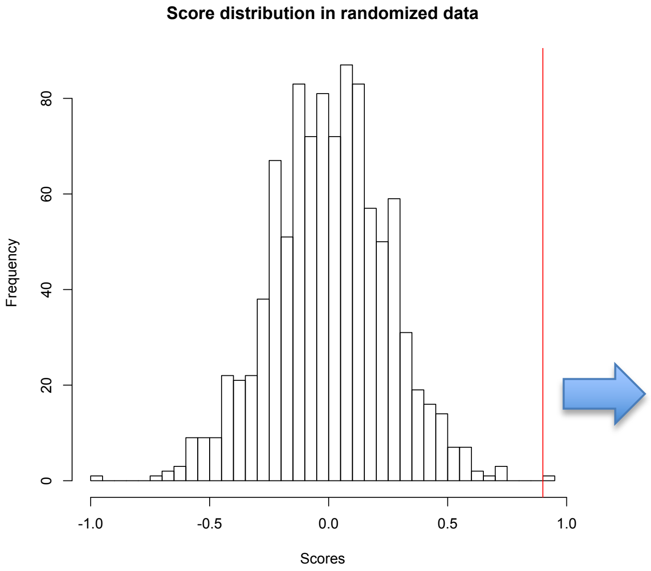
for each possible taxon pair, compute similarity score



symmetric similarity matrix

ASSESSMENT OF SIGNIFICANCE (Null model)

repeat scoring step many times with randomized data



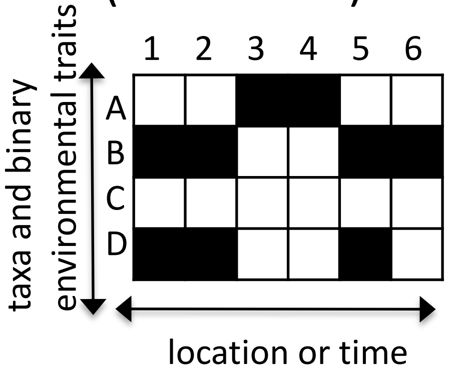
calculate p-values from the random score distribution, correct for multiple testing and discard relationships with p-values above a specified threshold

Microbial network inference

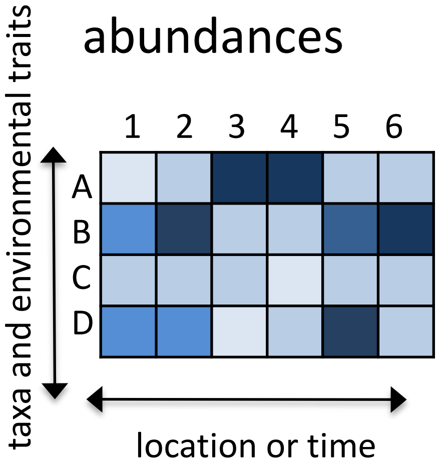
Principle of similarity-based network inference

INPUT

presences/absences (incidences)

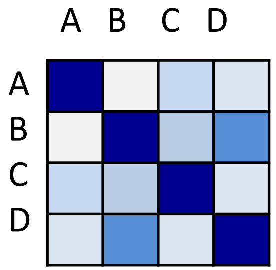


abundances



SCORING

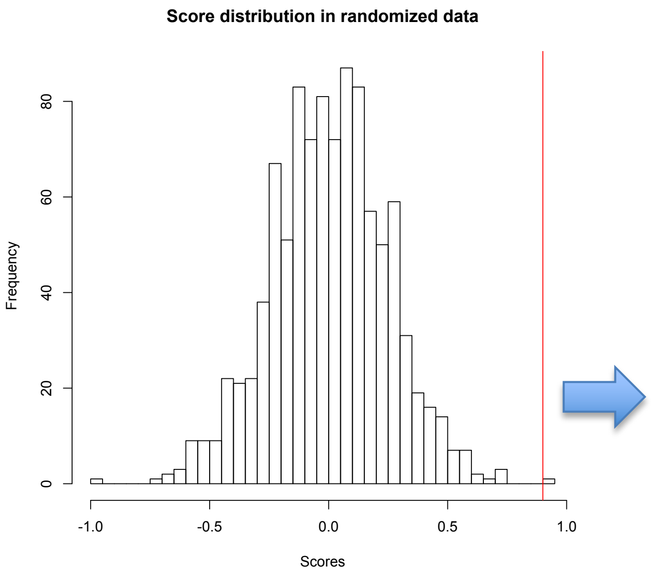
for each possible taxon pair, compute similarity score



symmetric similarity matrix

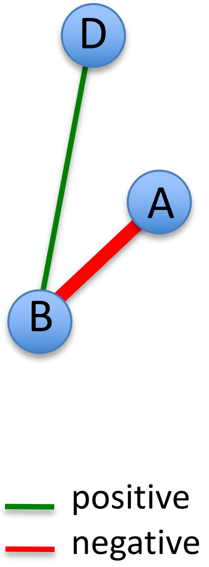
ASSESSMENT OF SIGNIFICANCE (Null model)

repeat scoring step many times with randomized data



calculate p-values from the random score distribution, correct for multiple testing and discard relationships with p-values above a specified threshold

VISUALIZATION



visualize taxon pairs with significant scores as a network

Challenges

- What are the challenges of microbial network inference?

Microbial network inference: Challenges

Problem 1: Varying sequencing depth

Shallowly sequenced sample

Deeply sequenced sample

Genus 1

Genus 2

Genus 3

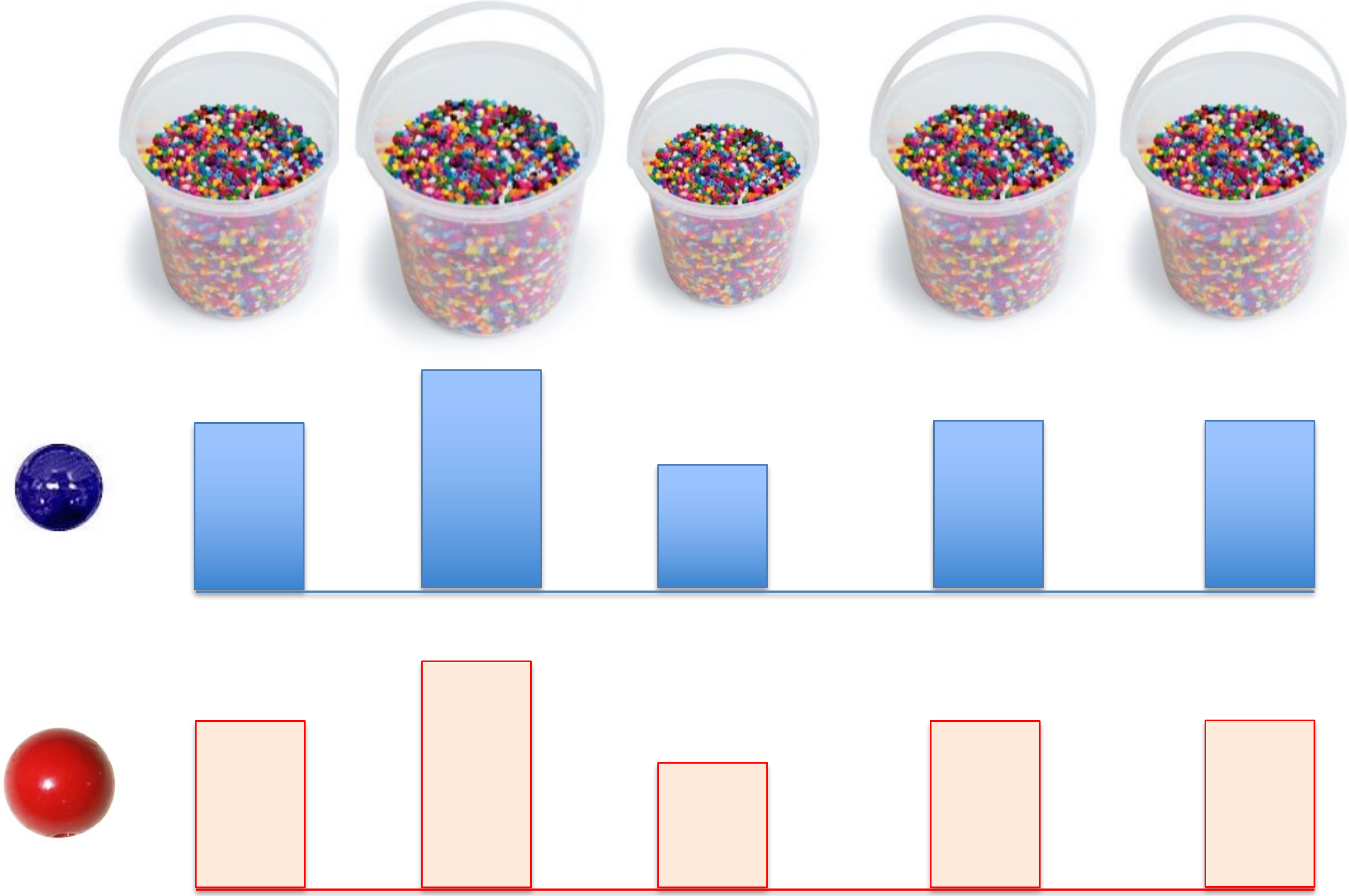
Genus 4

Technical variability

Read count != cell count

Microbial network inference: Challenges

Varying sequencing depth leads to spurious correlations

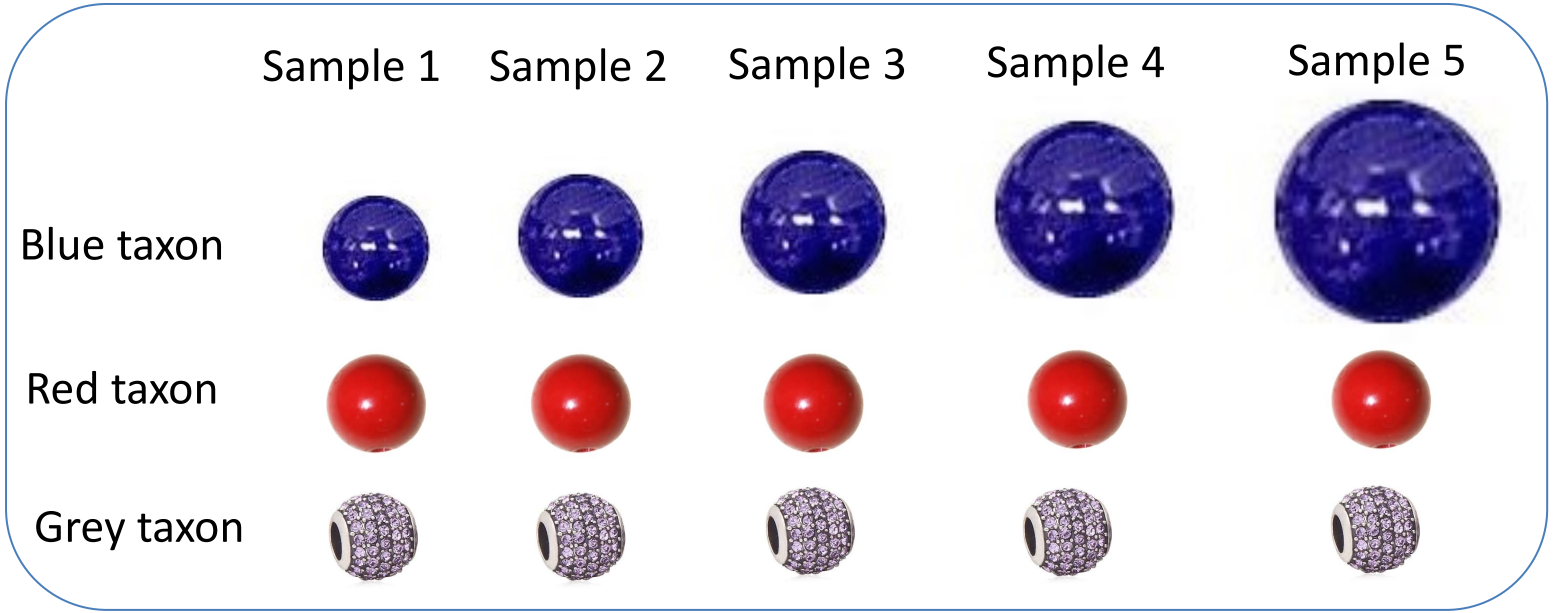


Removal of sequencing depth bias

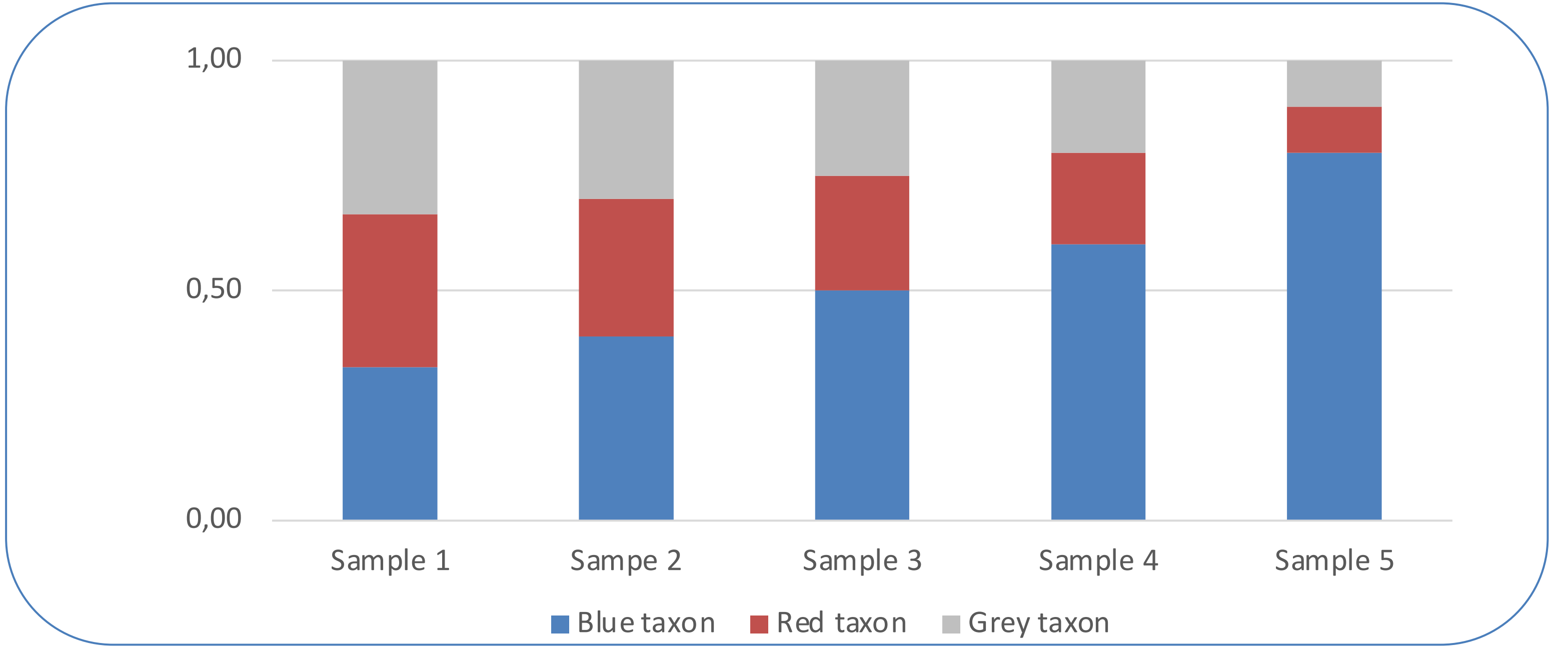
Rarefaction: Select beads from the big bucket with a probability equal to their proportion, until selected bead number is the same as in the small bucket
=> Additional zeros can be introduced, counts are preserved

Normalization: Convert counts into relative abundances (proportions) => counts are lost, no additional zeros

Rarefaction/normalization: compositionality



Absolute abundances



Relative abundances

How to deal with compositionality

- Transform data e.g. using centered log ratio transformation
- Use compositional-robust measures such as Aitchison distance or Bray Curtis dissimilarity

OR

- Quantify total cell counts, e.g. using flow cytometry

⇒ Working with log ratios poses a zero treatment problem

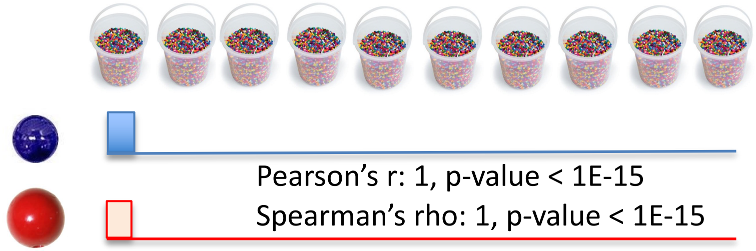
⇒ Relative abundances can be multiplied with cell counts, solving compositionality issue experimentally

⇒ But: if cell counts do not depend on microbial interactions, they are a confounder driving associations, so it depends

Problem 2: The challenge of rare taxa

- Microbial abundance tables are zero-rich
- Ambiguity of the zero: taxon may be absent or present below detection level (sampling and sequencing depth)
- Ignoring zeros (e.g. in log-ratio transformations) is a loss of information but not treating zeros can lead to spurious associations
- Currently: ad-hoc filters on taxon absences (prevalence filter) or taxon pairs (number of matching zeros); upper bound on zero number above which statistical tests are no longer meaningful (Cougoul et al.)

The problem of co-absences



Pearson's r: 1, p-value < 1E-15
Spearman's rho: 1, p-value < 1E-15

Ad-hoc solution: Prevalence filter

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
▬	**0**	5	2	**0**	3	**0**
✹	**0**	**0**	**0**	2	**0**	**0**
✺	58	56	45	129	81	18
◡	**0**	2	**0**	175	**0**	**0**

Presence in at least 50% of samples

Garbage taxon

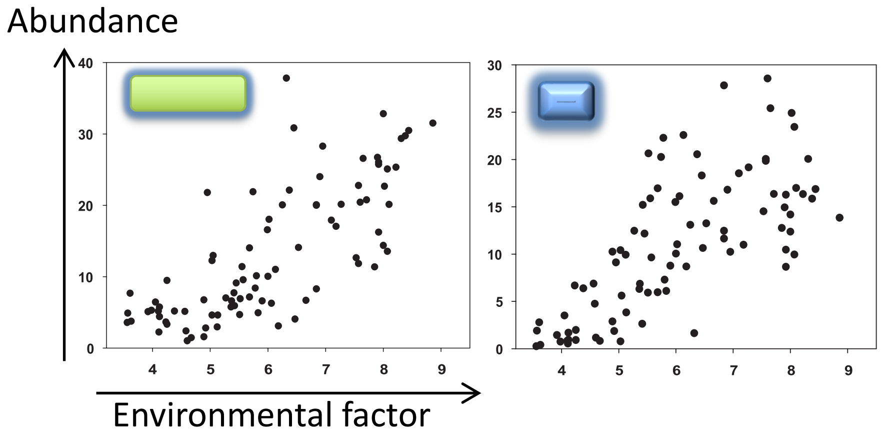
	S1	S2	S3	S4	S5	S6
▬	**0**	5	2	**0**	3	**0**
✺	58	56	45	129	81	18
🗑	**0**	2	**0**	177	**0**	**0**

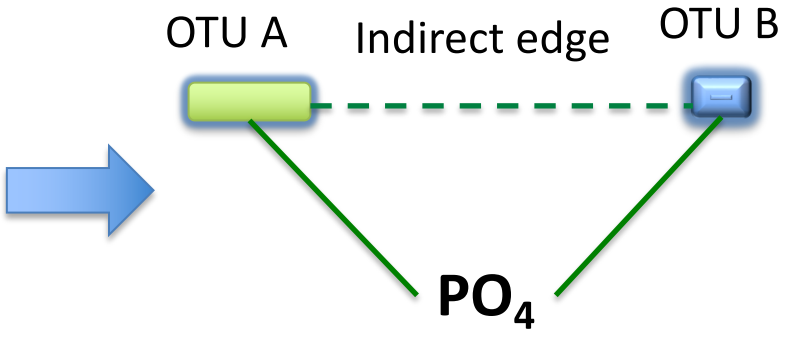
Cougoul et al. (2019) "Rarity of microbial species: in search of reliable associations." PLoS ONE 14, e0200458.

Problem 3: Indirect edges

- **Indirect edge**: a spurious edge introduced by the response of two taxa to a third factor (another **taxon** or an **environmental factor**)

- “correlation is not causation”

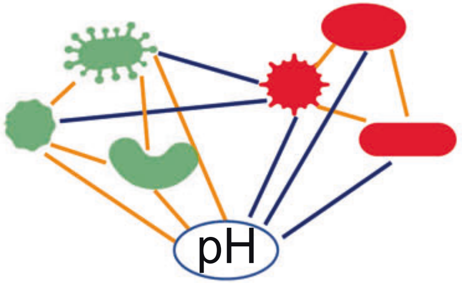




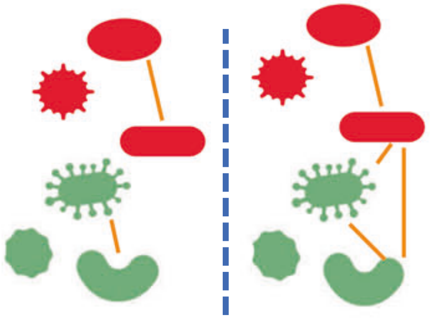
Microbial network inference: Challenges

Environmentally induced indirect edges: Possible solutions

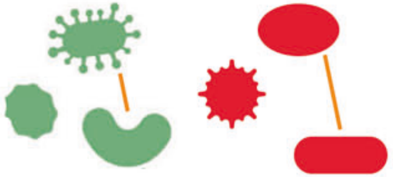
1) Include environmental factors in network



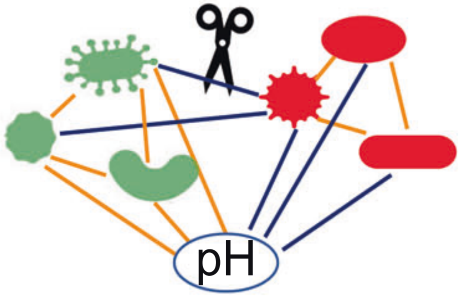
2) Stratify samples and compute a network per sample group



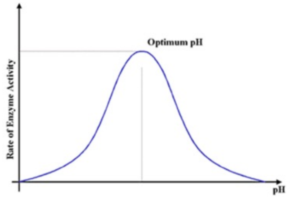
3) Regress out environmental factors before network construction



4) Filter network



Problem with third solution: non-linear dependencies of taxon abundance on environmental factors (optima)

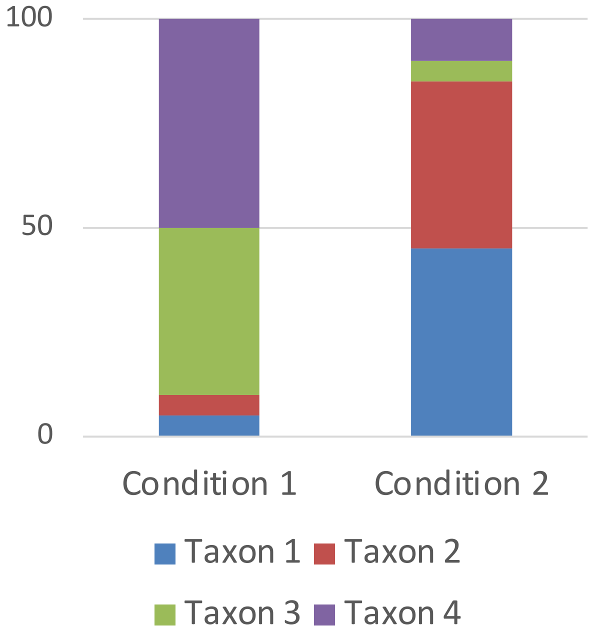


Faust (2021) “Open challenges for microbial network construction and analysis.” The ISME Journal 15, 3111-3118.

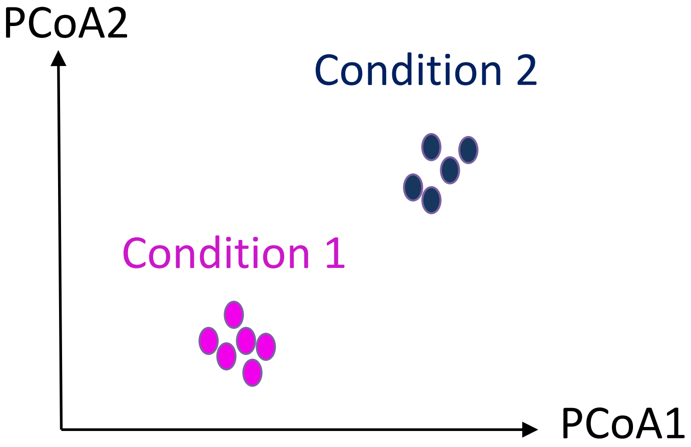
Microbial network inference: Challenges

Environmentally induced indirect edges continued

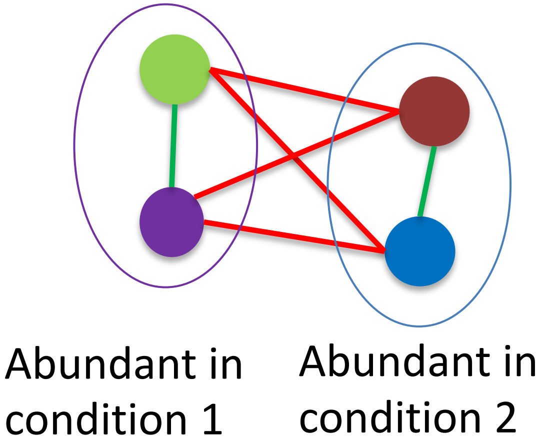
Sample heterogeneity



Clusters in sample-wise PCoA



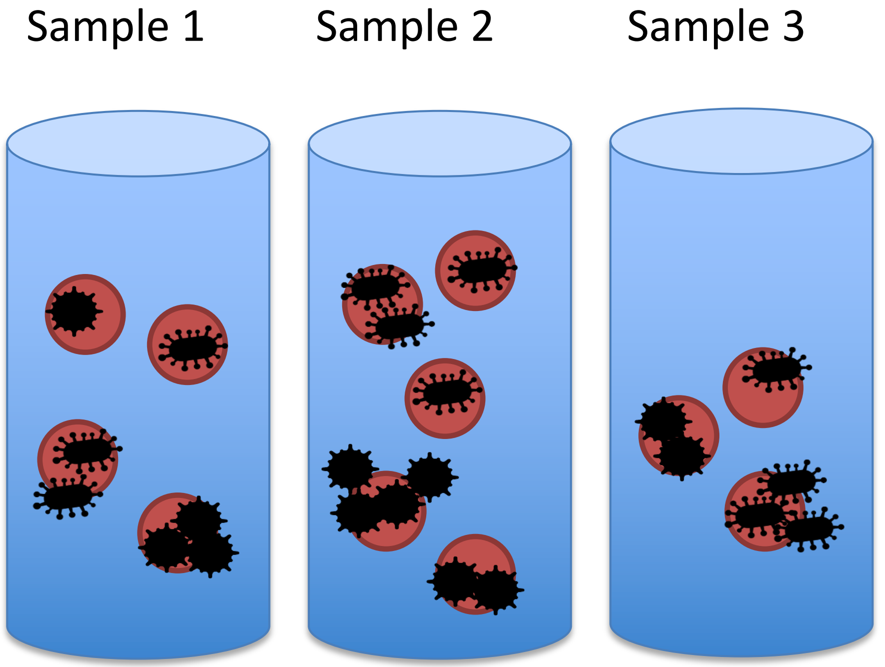
Indirect edges in inferred network



Microbial network inference: Challenges

Problem 4: Sampling scale

- Edges may differ depending on sampling scale
- Storage effect: variability at the microscale allows survival of competing species (one or the other dominates locally, but both co-occur globally)
- Problem of experimental design

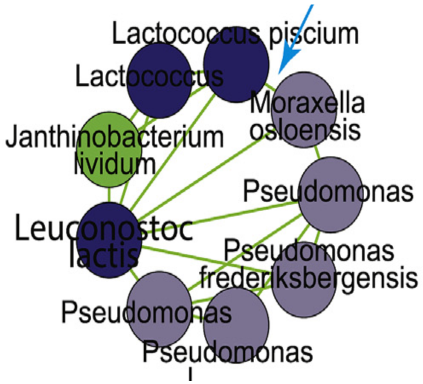


Inferred edge is negative within a sample but positive across the samples

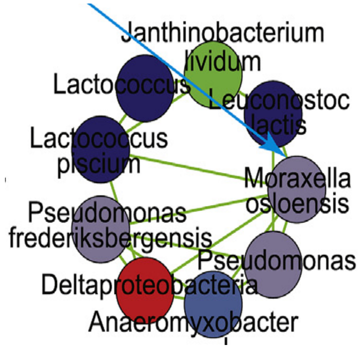
Chesson (2000) “Mechanisms of Maintenance of Species Diversity” 31, 343-366.
Armitage & Jones (2019) “How sample heterogeneity can obscure the signal of microbial interactions” The ISME Journal 13, 2639-2646.

Sampling scale: example

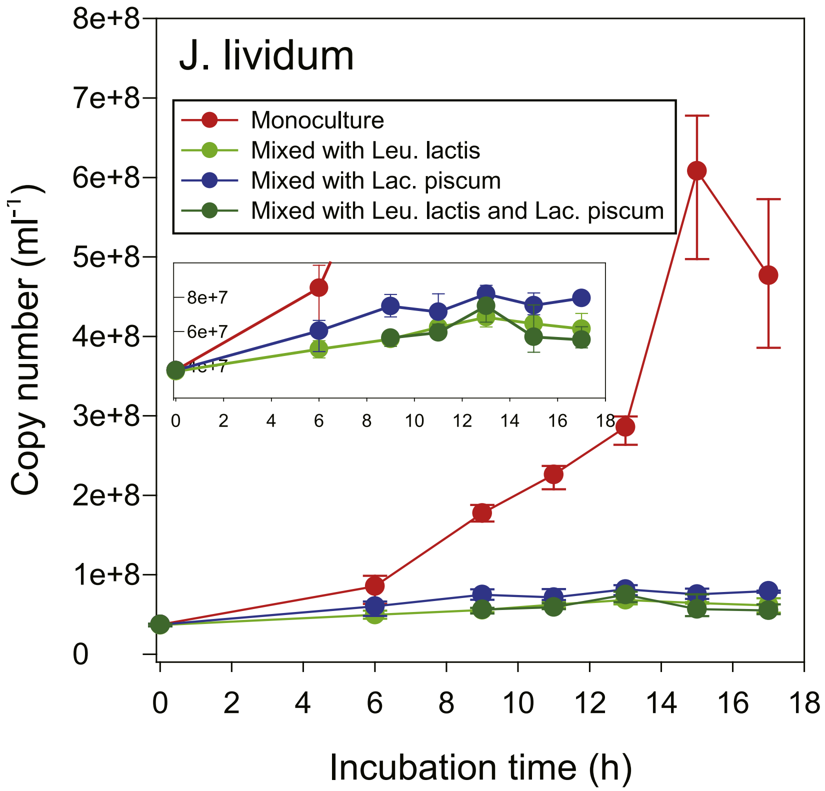
- Local competition hidden by shared niche preference



Tool: CoNet



Tool: MENA



Positive association predicted

Negative interaction found

Wang et al. (2017) “Combined use of network inference tools identifies ecologically meaningful bacterial associations in a paddy soil” Soil Biology & Biochemistry 105, 227-235.

Tools

Which microbial network inference tools are available and how do they work?

LSA
local similarity analysis

oNet

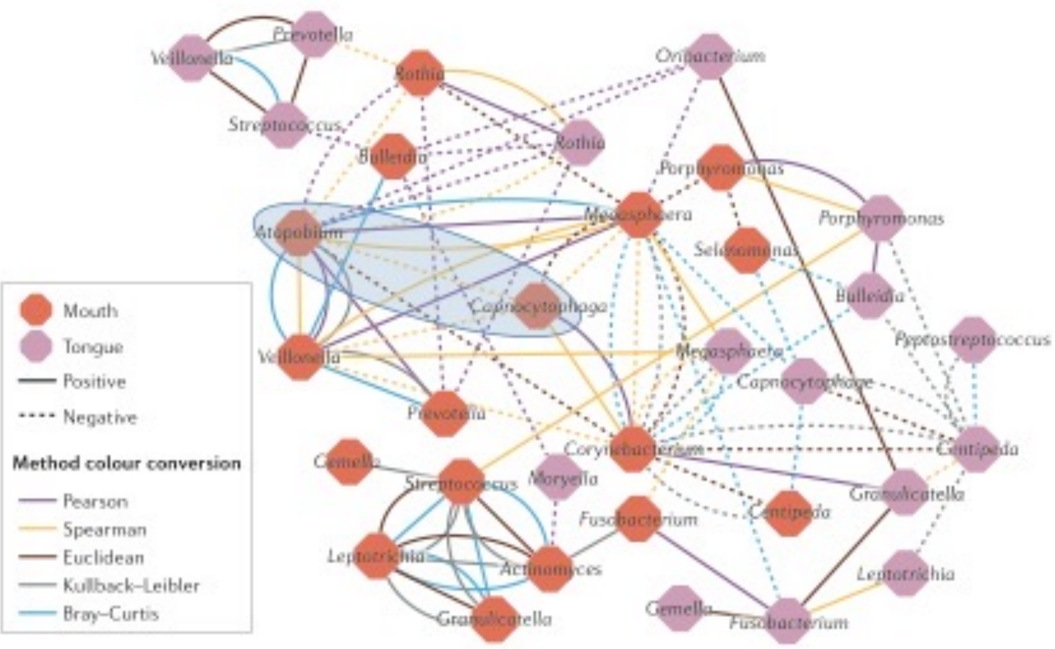
FlashWeave

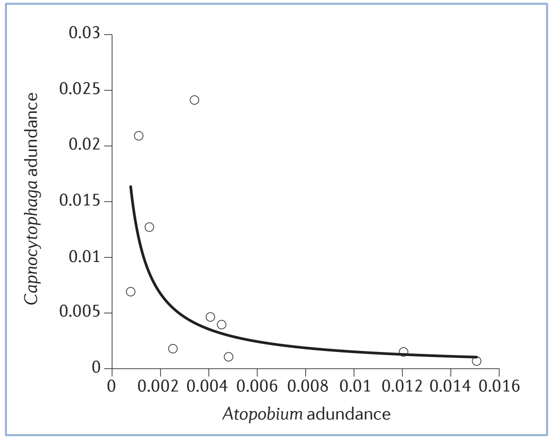
</> MDSINE

SPIEC-EASI

CoNet

- Different measures (Pearson, Spearman, Bray Curtis, ...) capture different types of relationships, but they converge when thresholds are increased
- Ensemble: measures make different mistakes, but tend to agree on correct result, so combine them



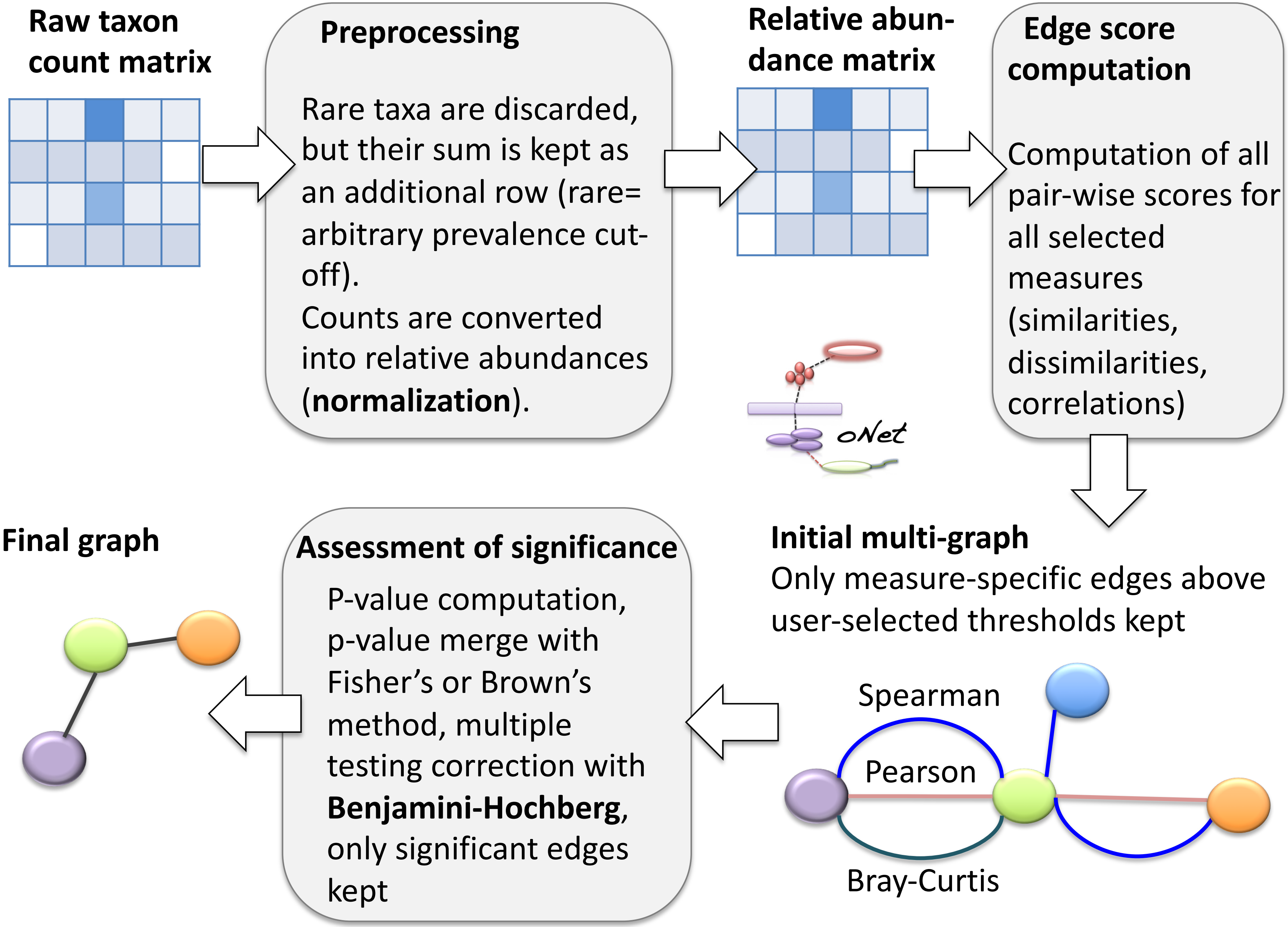


non-linear relationship is missed by Pearson

Faust & Raes (2012) “Microbial interactions: from networks to models.” Nature Reviews Microbiology 10 (8), 538-550.

Microbial network inference: Tools

CoNet: Overview



Raw taxon count matrix

Preprocessing

Rare taxa are discarded, but their sum is kept as an additional row (rare= arbitrary prevalence cut-off).
Counts are converted into relative abundances (**normalization**).

Relative abundance matrix

Edge score computation

Computation of all pair-wise scores for all selected measures (similarities, dissimilarities, correlations)

Initial multi-graph
Only measure-specific edges above user-selected thresholds kept

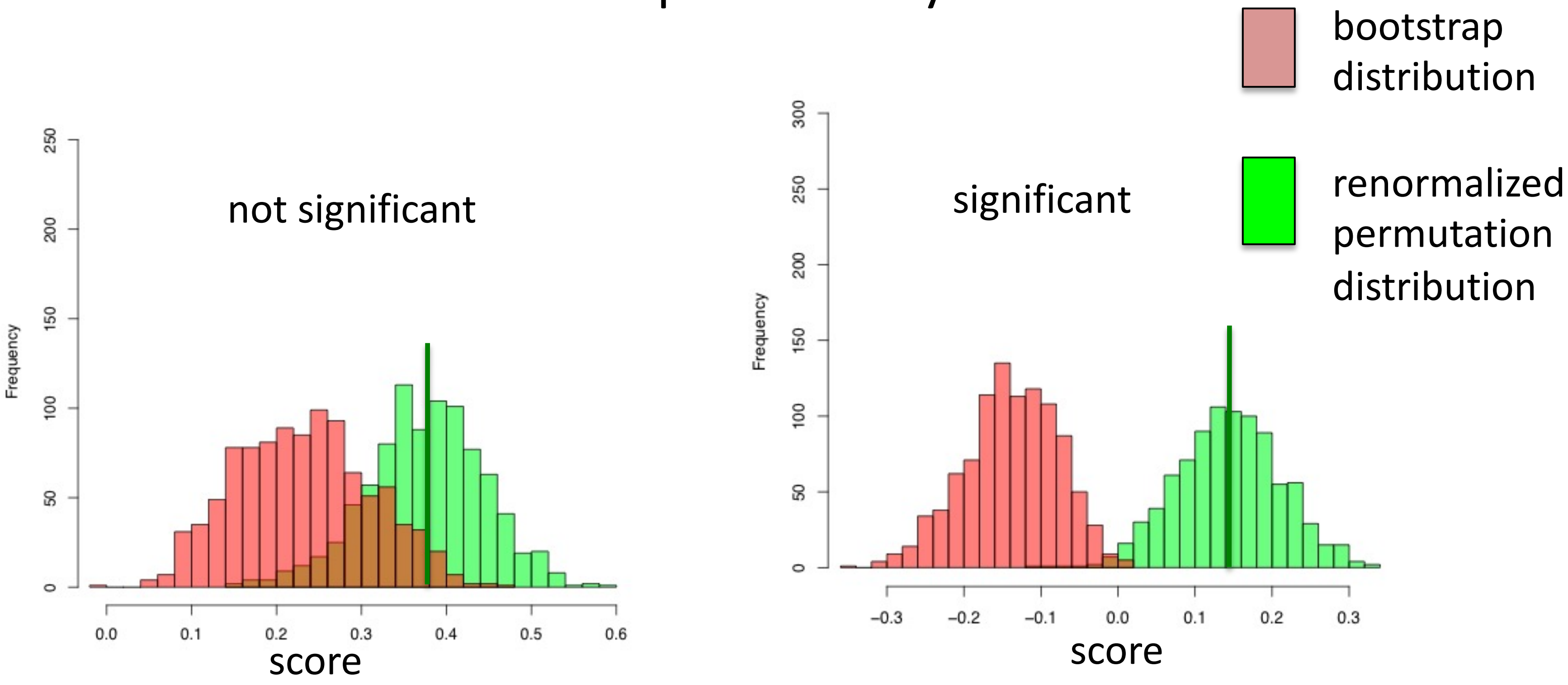
Assessment of significance

P-value computation, p-value merge with Fisher’s or Brown’s method, multiple testing correction with **Benjamini-Hochberg**, only significant edges kept

Final graph

CoNet: P-value computation (CCREPE)

- Edge- and measure-specific p-value is computed with a **Z-test**: probability of the permutation distribution mean given the (normally distributed) bootstrap distribution
- Renormalization to reduce compositionality bias



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

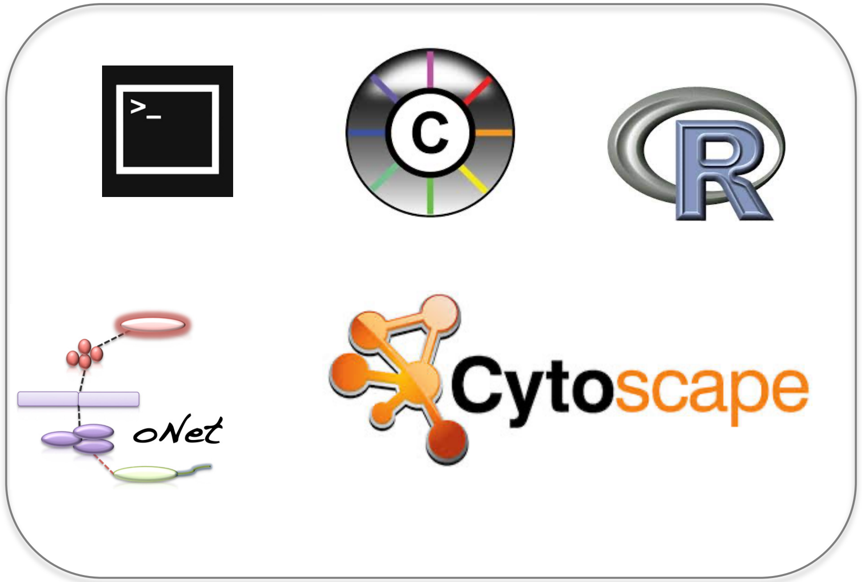
Microbial network inference: Tools

CoNet: Implementation

- CoNet is available on command line and as a Cytoscape app (versions 2.X and 3.X)
- CoNet page: **http://msysbiology.com/conet**
- Cytoscape app: **http://apps.cytoscape.org/apps/conet**
- R implementation of core functions: **https://hallucigenia-sparsa.github.io/seqgroup**

Co-developers & contributors
Fah Sathirapongsasuti
Jean-Sébastien Lerat
Gipsi Lima-Mendez
Jeroen Raes

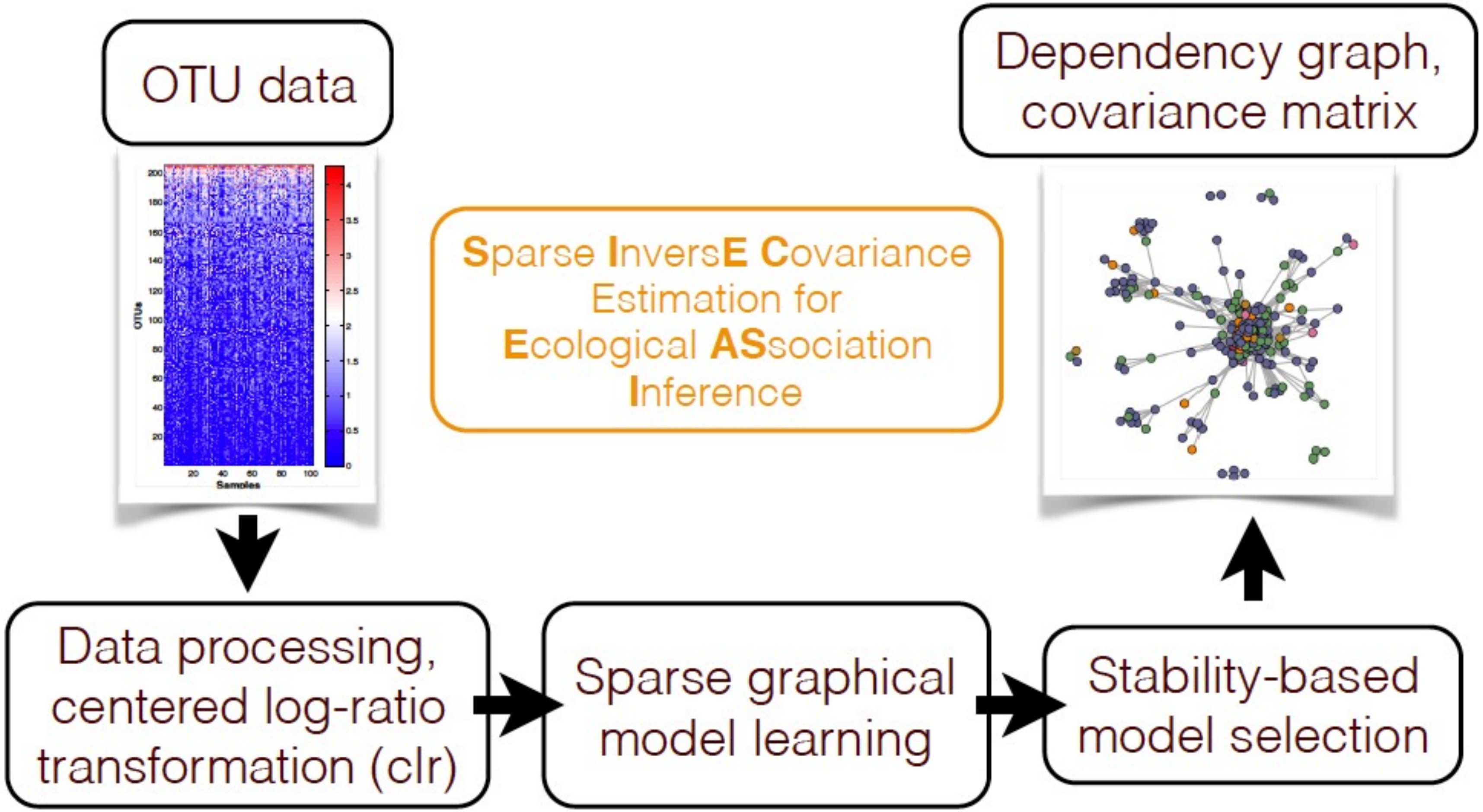
> 29,700 downloads from Cytoscape app store



Faust & Raes (2016). “CoNet app: inference of biological association networks using Cytoscape” F1000Research 5:1519

Microbial network inference: Tools

SPIEC-EASI



OTU data

Sparse InversE Covariance Estimation for Ecological ASsociation Inference

Dependency graph, covariance matrix

Data processing, centered log-ratio transformation (clr)

Sparse graphical model learning

Stability-based model selection

Image taken from: https://stamps.mbl.edu/images/f/f0/STAMPS_Network_1.pdf (Christian Mueller)

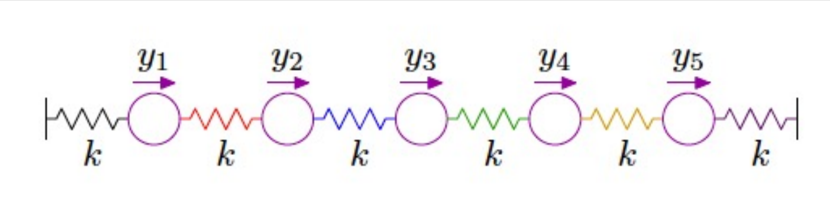
Kurtz et al. (2015) “Sparse and Compositionally Robust Inference of Microbial Ecological Networks” PLoS Computational Biology 11(5), e1004226.

Microbial network inference: Tools

SPIEC-EASI: Sparse graphical models

- SPIEC-EASI estimates the **inverse covariance matrix**, such that resulting **network has fewer indirect edges**
- Zero in the inverse covariance matrix: conditional independence
- **Assumptions**: data are multivariate normally distributed and **all relevant variables are taken into consideration**

Intuitive example by David MacKay:



Weights (nodes) connected by springs (edges)

Covariance matrix

$$\mathbf{K} = \frac{T}{k}\begin{bmatrix} 0.83 & 0.67 & 0.50 & 0.33 & 0.17 \\ 0.67 & 1.33 & 1.00 & 0.67 & 0.33 \\ 0.50 & 1.00 & 1.50 & 1.00 & 0.50 \\ 0.33 & 0.67 & 1.00 & 1.33 & 0.67 \\ 0.17 & 0.33 & 0.50 & 0.67 & 0.83 \end{bmatrix}$$

Inverse covariance matrix

$$\mathbf{K}^{-1} = \frac{k}{T}\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

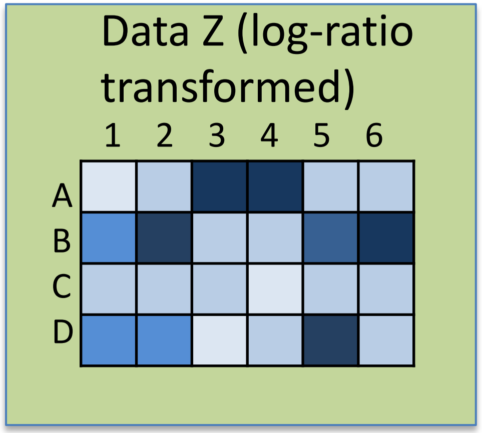
Source: http://www.inference.org.uk/mackay/humble.pdf

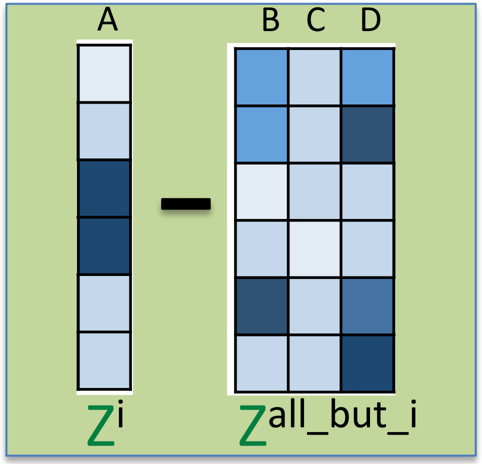
Sparse = few non-zero entries in the inverse covariance matrix = few edges

Microbial network inference: Tools

SPIEC-EASI: Meinshausen & Bühlmann

One of SPIEC-EASI's methods to infer the inverse covariance matrix:
Meinshausen & Bühlmann method (neighborhood selection)





For each taxon i, do:

$$\hat{\beta}^{i,\lambda} = \underset{\beta\in\mathbb{R}^{p-1}}{\arg\min}\left(\frac{1}{n}\left\|Z^{i} - Z^{\neg i}\beta\right\|^{2} + \lambda\left\|\beta\right\|_{1}\right)$$

Penalty parameter

Species number

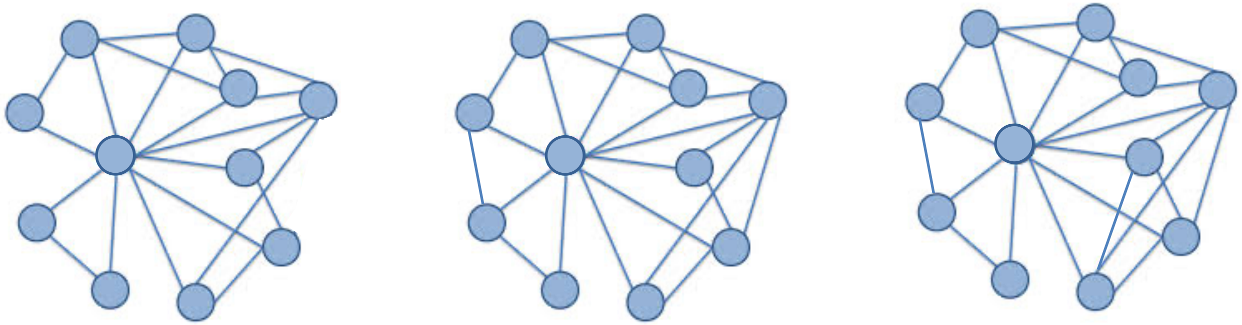
Sample number

Regression coefficients

Result: matrix of regression coefficients; is symmetrised
Edge = non-zero regression coefficient

SPIEC-EASI: Stability-based model selection

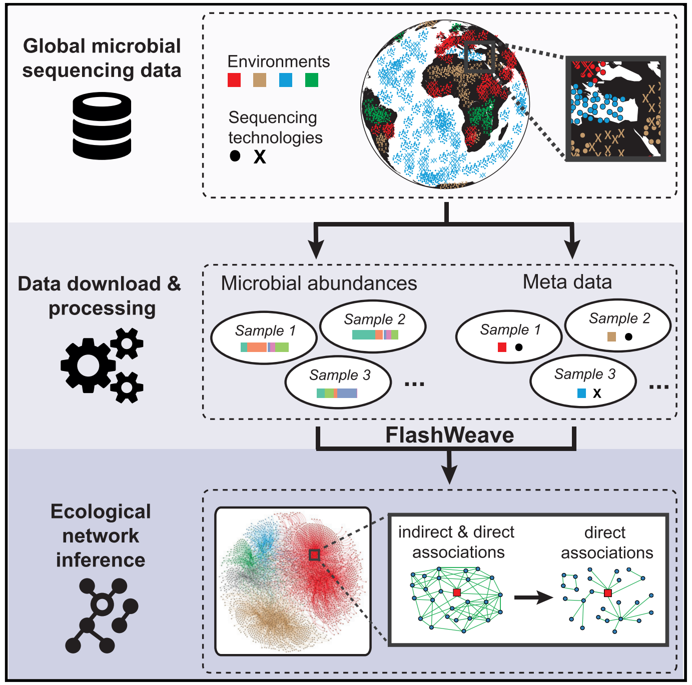
- **StARS**: **St**ability **A**pproach to **R**egularization **S**election
- Bootstrap technique: Repeat network construction a number of times with 80% of the samples (bootstrap iteration number = rep.num parameter)
- Purpose: select **penalty parameter λ** such that the number of edges present across bootstrap iterations is maximized
- Stability means here: stable with respect to small changes in the data



Microbial network inference: Tools

FlashWeave

- SPIEC-EASI’s main weakness: does not take environmental data into account
- "FlashWeave = SPIEC-EASI + metadata”: clr-transforms data and exploits conditional independence to reduce indirect edge number, taking metadata into account (algorithm: si-HITON-PC)
- Implemented in Julia



julia

Tackmann, Rodrigues and van Mering (2019): “Rapid Inference of Direct Interactions in Large-Scale Ecological Networks from Heterogeneous Microbial Sequencing Data” Cell Systems 2019.08.002.

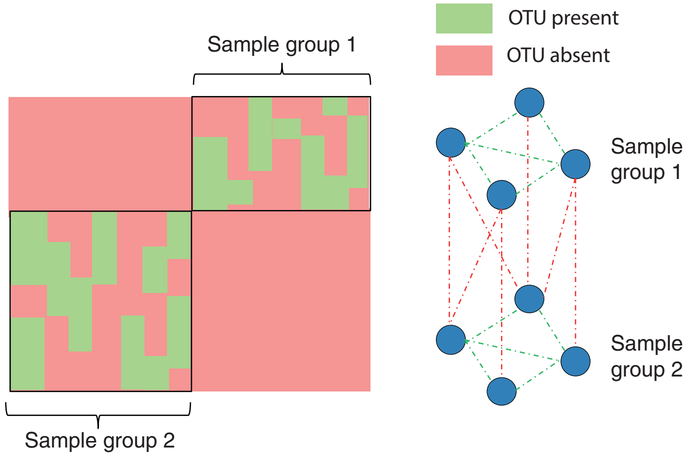
FlashWeave: modes

Sensitive vs fast mode

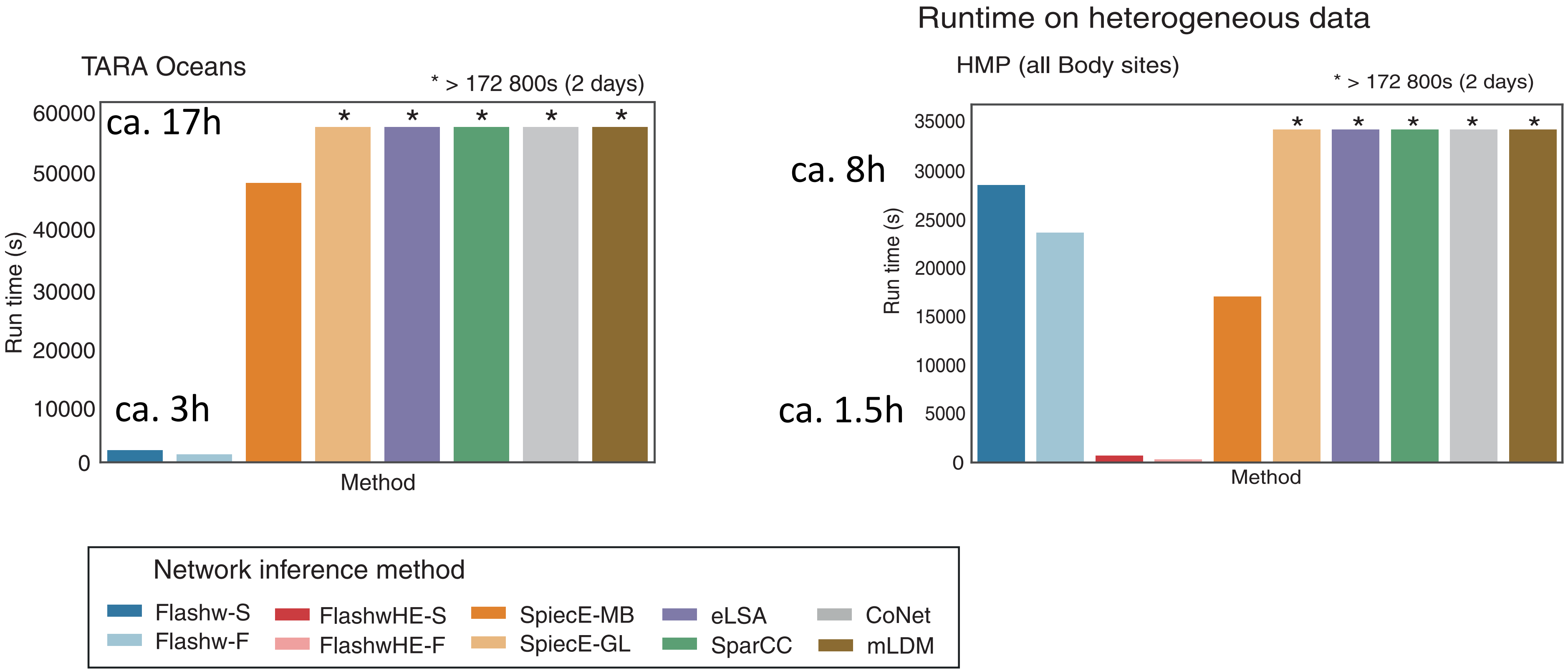
- Implementation of conditional independence:
  - Sensitive mode: partial correlations on abundances, assumes multivariate normal distribution (weak assumption)
  - Fast mode: mutual information on presence/absences

HE mode

- FlashWeave can optionally ignore zeros (‘structural zeros’) to deal with heterogeneous samples



FlashWeave: Run time



=> My current choice for best cross-sectional microbial network inference tool; pending independent evaluation

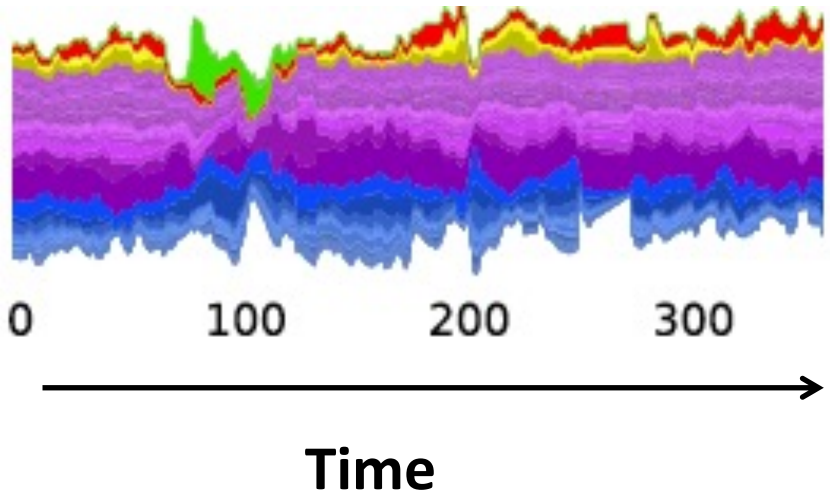
Other microbial network inference tools

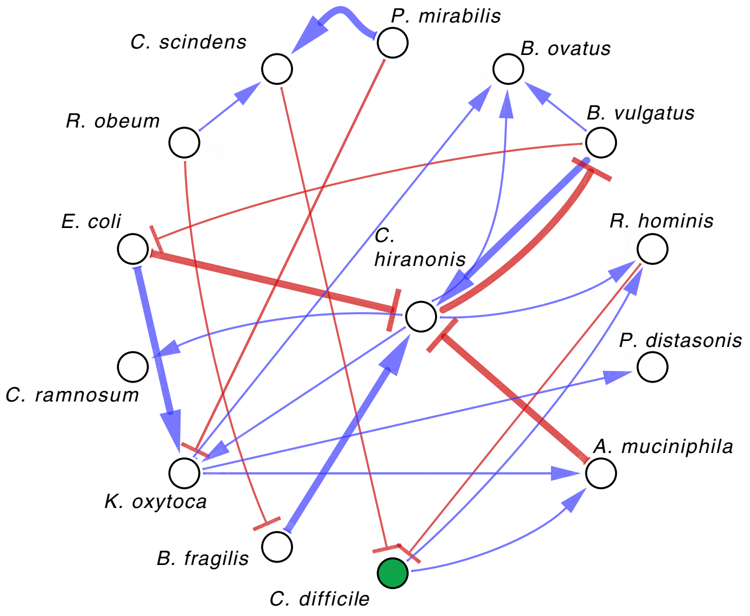
- **MENAP** (Molecular Ecological Network Analyses Pipeline): exploits random matrix theory to threshold similarity matrix
- **SparCC**: sparse correlations robust to compositionality
- **REBACCA/CCLasso**: sparse compositionality-robust correlations
- **MInt:** Takes environmental factors into account through hierarchical regression
- **gCoda:** estimates inverse covariance like SPIEC-EASI, but deals differently with compositionality
- **NetCoMi**: correlation networks with comparison functions

List is not complete

MENAP: Zhou et al. (2010) "Functional Molecular Ecological Networks" mBio 1 (4), e00169-10.
SparCC: Friedman & Alm (2012) "Inferring Correlation Networks from Genomic Survey Data." PLoS Comp Bio 8 (9), e1002687.
REBACCA: Ban et al. (2015) "Investigating microbial co-occurrence patterns based on metagenomic compositional data" Bioinformatics 31(20):3322-3329.
CCLasso: Fang et al. (2015) "CCLasso: correlation inference for compositional data through Lasso" Bioinformatics 31(19):3172-3180.
MInt: Biswas et al. (2015) "Learning Microbial Interaction Networks from Metagenomic Count Data" RECOMB, Research in Computational Molecular Biology, 32-43 (Lecture Notes in Computer Science).
gCoda: Huaying et al. (2017) "gCoda: Conditional Dependence Network Inference for Compositional Data" Journal of Computational Biology 24(7): 699-708.
NetCoMi: Peschel et al. (2021) "NetCoMi: network construction and comparison for microbiome data in R" Briefings in Bioinformatics 22(4), 1-18.

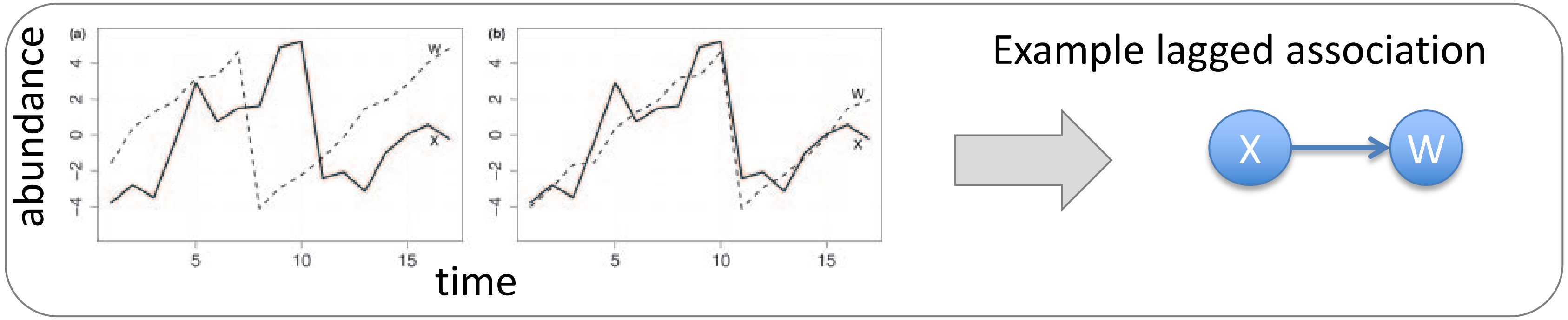
Tools exploiting time series information





Inference of **directed** networks

Local Similarity Analysis (LSA)

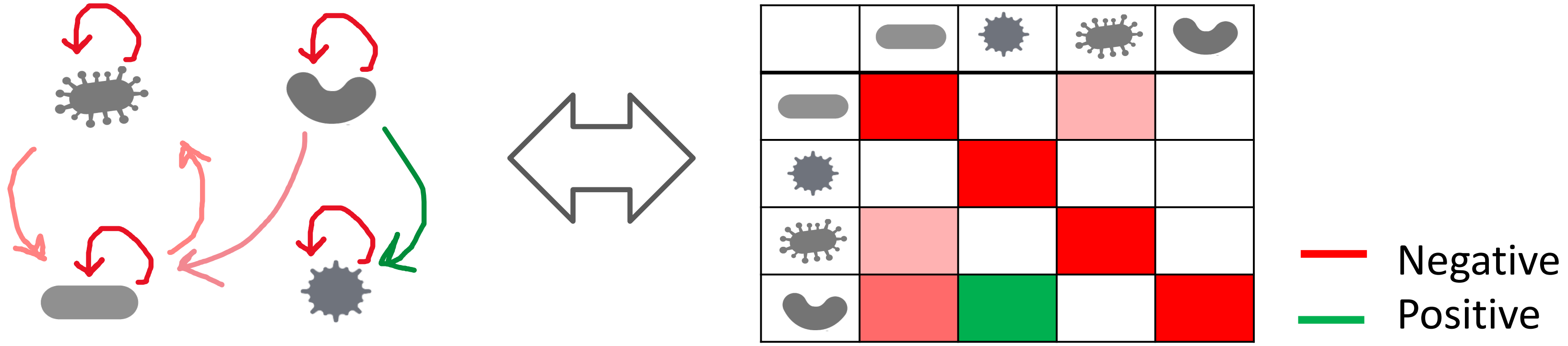


- LSA uses dynamic programming to find local associations and lagged associations
- Can be applied to cross-sectional and time series data
- P-values computed through permutation or formula
- Command line tool:
  - **https://bitbucket.org/charade/elsa/wiki/Home**

Xia et al. (2013) “Efficient statistical significance approximation for local similarity analysis of high-throughput time series data” Bioinformatics 29 (2), 230-237.
Durno et al. (2013) “Expanding the boundaries of local similarity analysis” BMC Genomics 14 (1), S3.
Xia et al. (2011) “Extended local similarity analysis (eLSA) of microbial community and other time series data with replicates.” BMC Systems Biology 5 (2), S15.
Ruan et el. (2006) “Local similarity analysis reveals unique associations among marine bacterioplankton species and environmental factors” Bioinformatics 22 (20), 2532-2538.

Generalized Lotka-Volterra (gLV)

- The species network can be represented by the **directed** interaction matrix A; entries represent interaction strengths



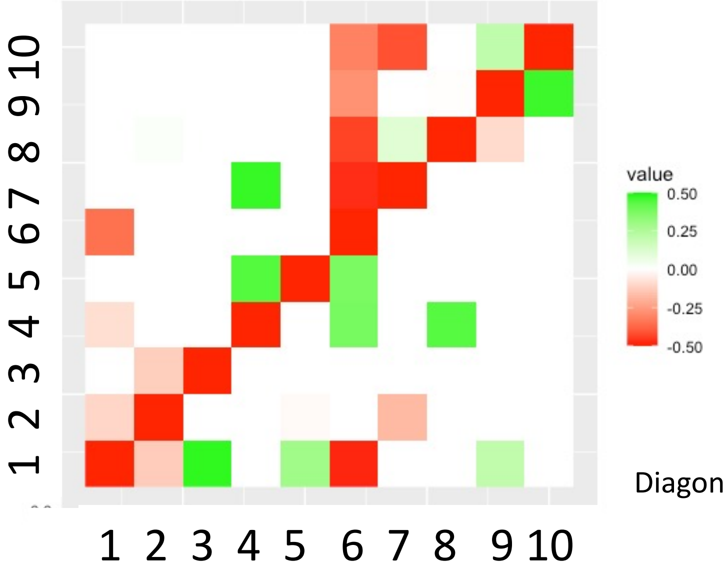
- Species abundance vector X changes as a function of species initial abundance, growth rates *B* and its interactions A

$$\boxed{\frac{dX(t)}{dt} = X(t)\left(B + AX(t)\right)}$$

Generalized Lotka Volterra (gLV)

Simulation with gLV

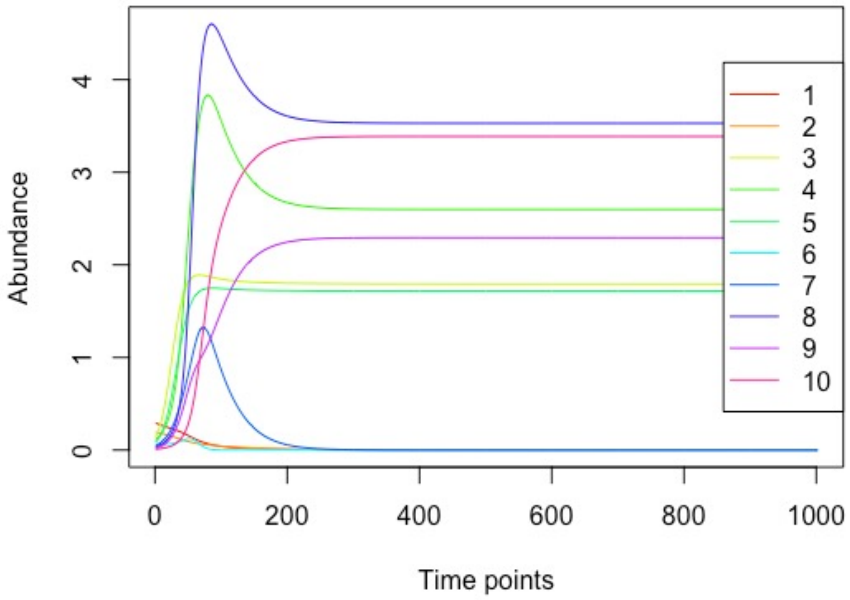
Community matrix

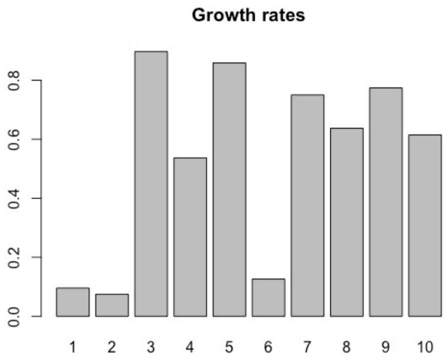


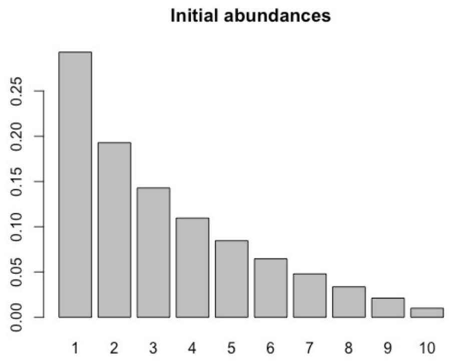
Diagonal: -0.5

Transient

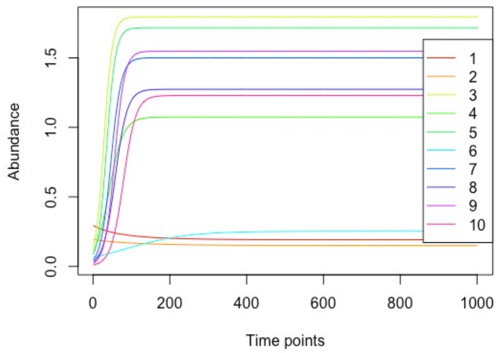
Steady state







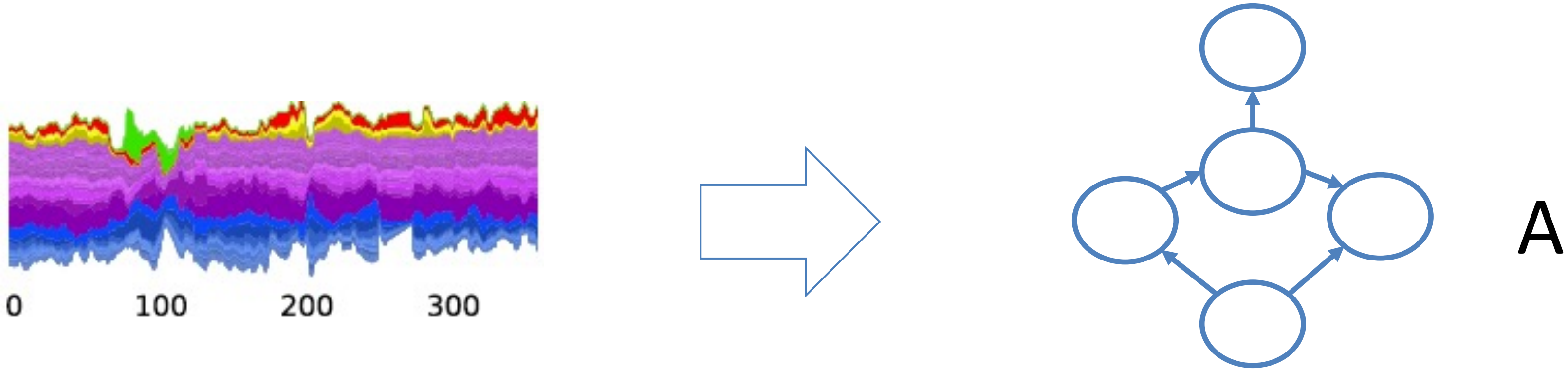
No inter-species interactions:



Microbial network inference using time

GLV parameterization

- Idea: infer gLV parameters from time series
- gLV parameters include species interaction matrix A
- gLV parameterization is a type of network inference



$$\frac{dX(t)}{dt} = X(t)\left(B + AX(t)\right)$$

GLV parameterization tools

- Tools parameterizing gLV equation:
  - **LIMITS**: step-wise forward regression plus bootstrap
  - **MDSINE**: parameterizes gLV with maximum likelihood and Bayesian algorithms
  - **SgLV-EKF**: parameterizes a stochastic gLV model with an extended Kalman Filter
  - **MetaMIS**: parameterizes gLV with partial least square regression

List is not complete

LIMITS: Fisher and Mehta (2014). "Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression." *PLoS one* 9**,** e102451.
MDSINE: Bucci et al. (2016) "Microbial Dynamical Systems INference Engine for microbiome time-series analyses" Genome Biology 17:121.
Alshawaqfeh et al. (2017) "Inferring microbial interaction networks from metagenomic data using **SgLV-EKF** algorithm" BMC Genomics 18:228.
Shaw et al. (2016): "**MetaMIS**: a metagenomic microbial interaction simulator based on microbial community profiles" BMC Bioinformatics 17:488.

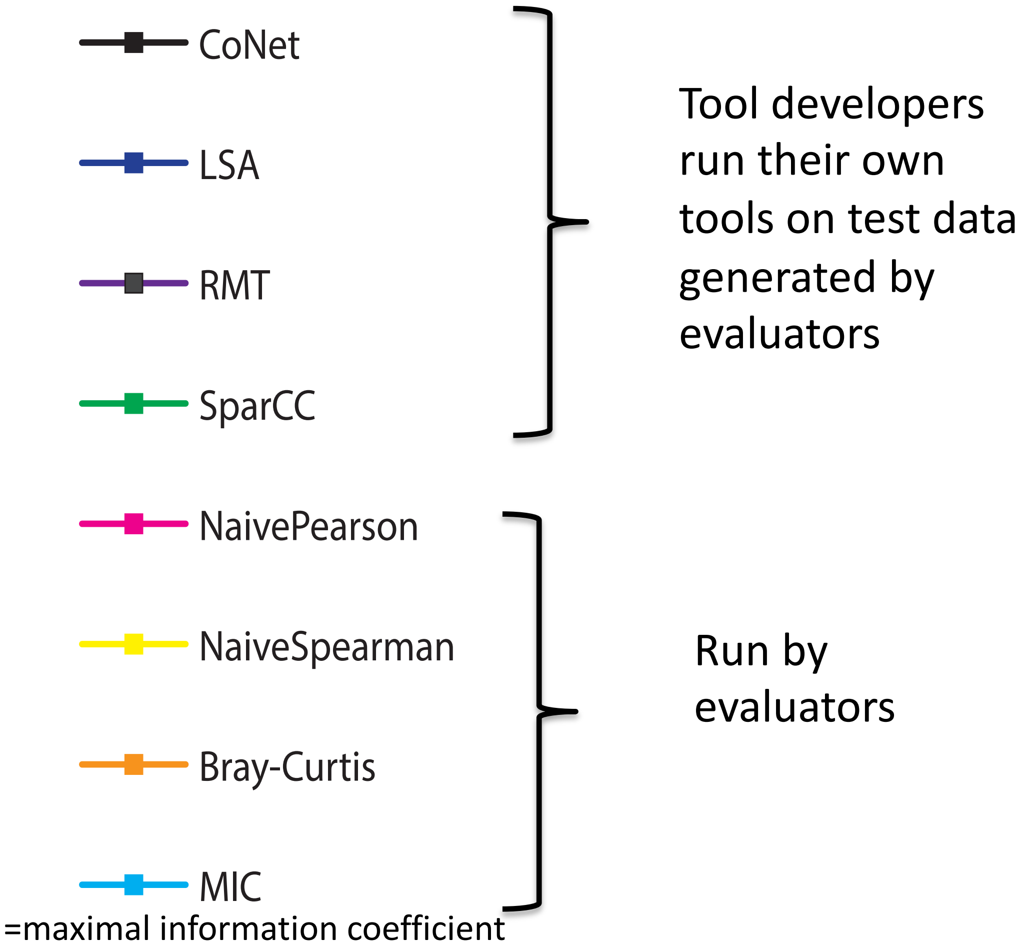
Evaluation

- How well do microbial network inference tools perform?

Warning:
If not stated otherwise, everything in this section is on synthetic data only

Microbial network inference evaluation

Tool Evaluation I



CoNet

LSA

RMT

SparCC

Tool developers run their own tools on test data generated by evaluators

NaivePearson

NaiveSpearman

Bray-Curtis

MIC

=maximal information coefficient

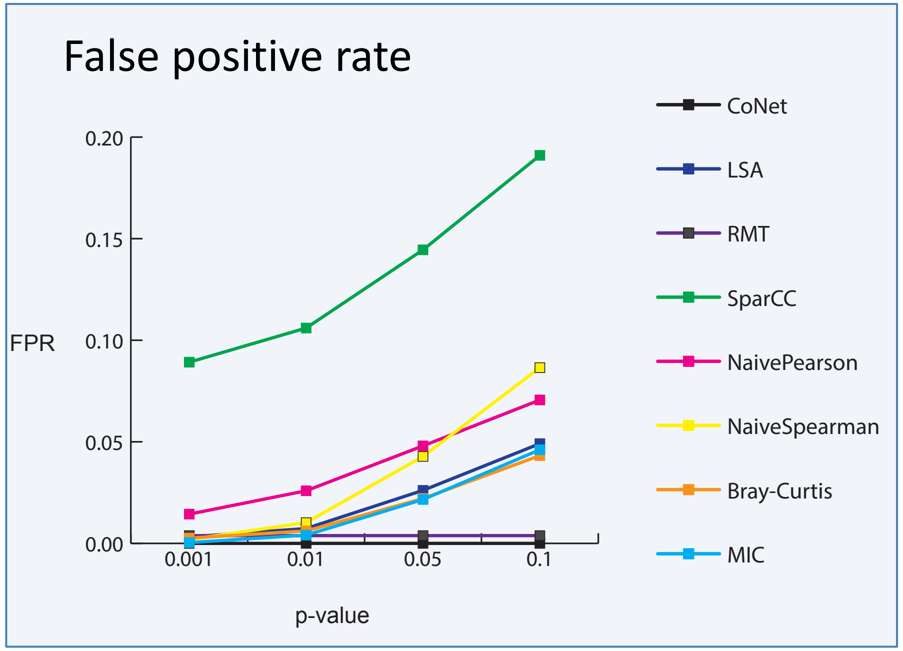
Run by evaluators

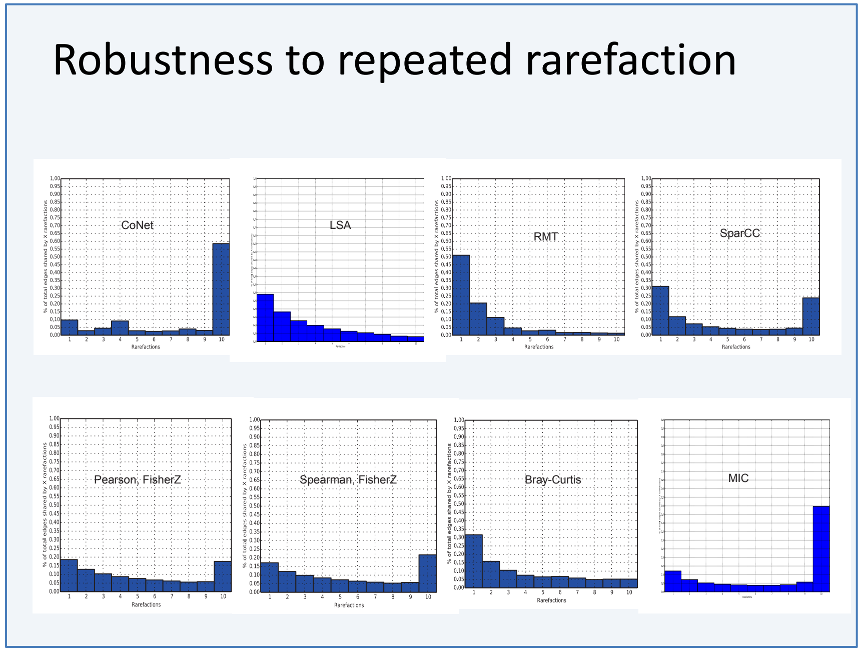
Evaluation: Weiss, Van Treuren, Lozupone, Faust et al. The ISME Journal 10, 1669-1681, 2016.
MIC: Reshef et al. Science 334, 1518-1524, 2011.

Microbial network inference evaluation

Tool Evaluation I: False positives and noise

- Most tools predict low number of false positives in data simulated without interactions (Dirichlet-Multinomial)
- CoNet and MIC are robust to noise (similar networks after repeated rarefactions)

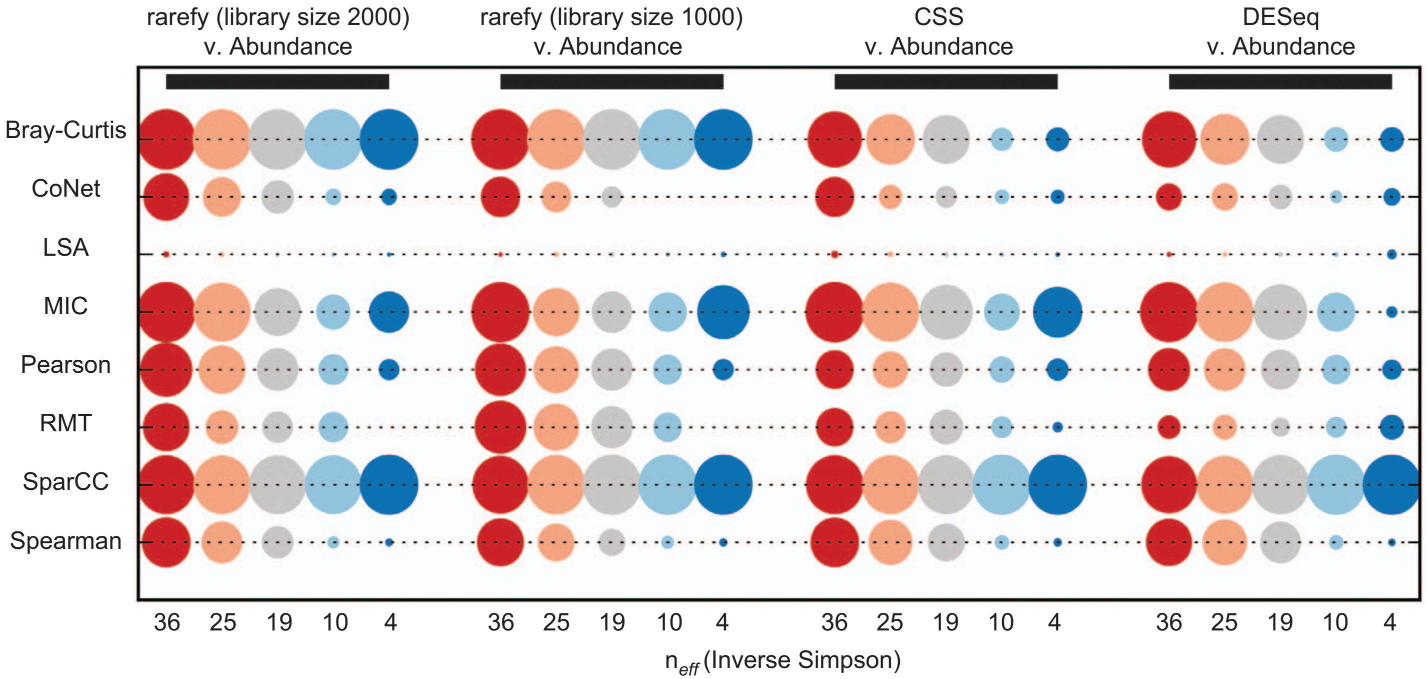

False positive rate


Robustness to repeated rarefaction

Microbial network inference evaluation

Tool Evaluation I: Effect of compositionality

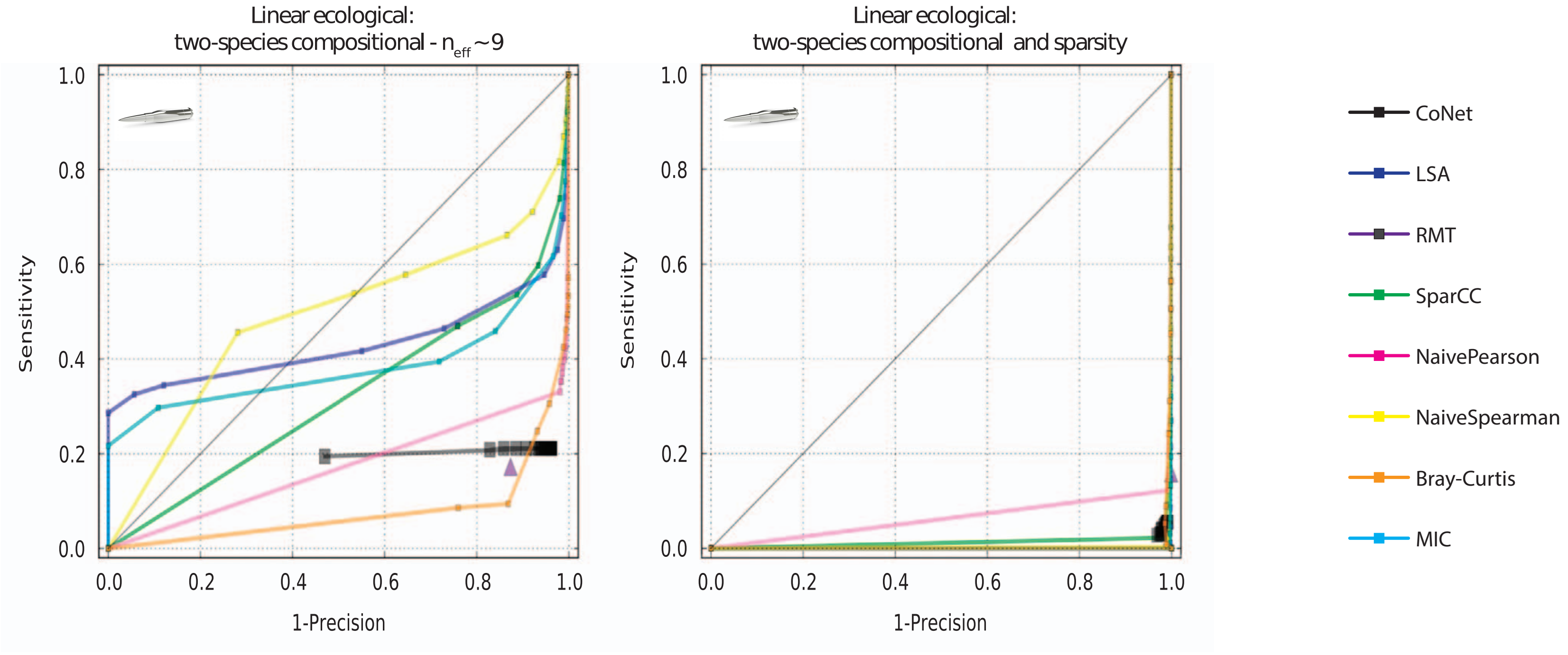
- Compositionality effect is stronger for lower evenness (n_{eff})
- Bray-Curtis and SparCC are compositionally robust (absolute versus relative abundance does not alter results)
- Alternative normalization techniques (CSS/DESeq) do not outperform rarefaction



Abundance of selected taxon pair multiplied with a constant to decrease evenness

Tool Evaluation I: Interactions

- Interaction detection accuracy in zero-rich, compositional data is low for all tools

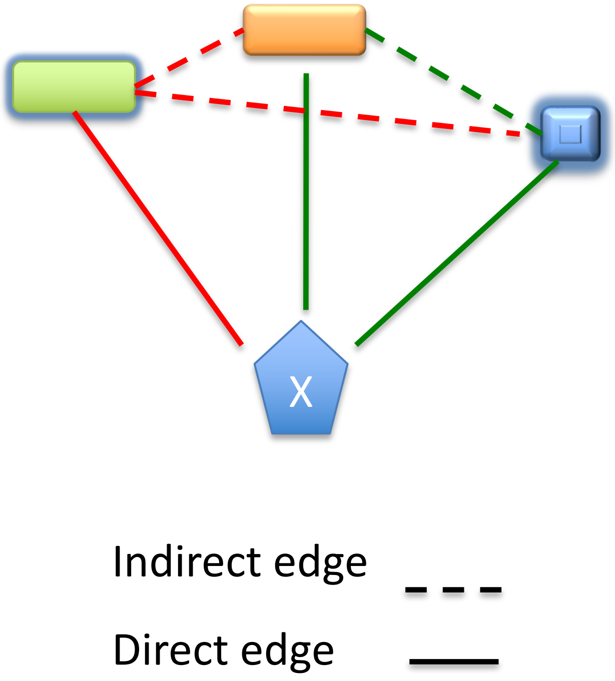


Sensitivity: TP/(TP+FN)
Precision (positive predictive value): TP/(TP+FP)

Microbial network inference evaluation

Inverse covariance to the rescue?

- One source of error: indirect edges
- Tools based on inverse covariance take them out
- Are these new tools (SPIEC-EASI, gCoda) more accurate than previous ones?
- FlashWeave not included (published afterwards)



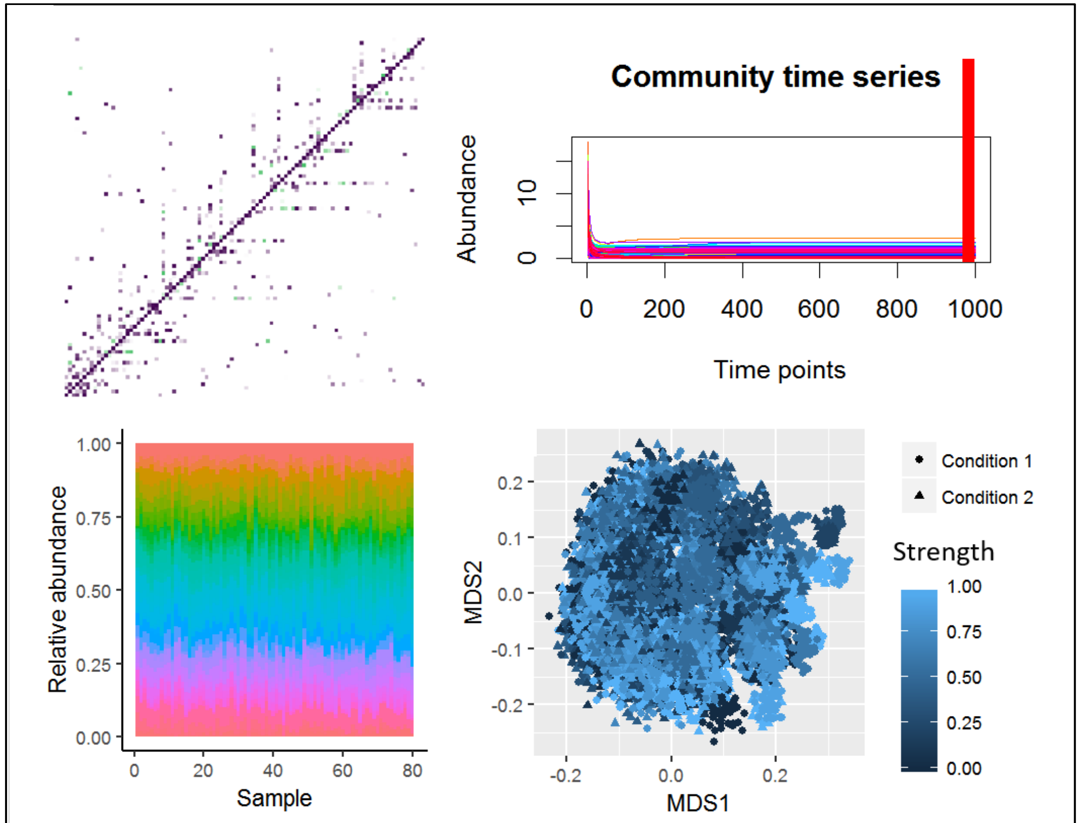
SPIEC-EASI: Kurtz et al. PLoS Computational Biology 11(5), e1004226, 2015.
gCoda: J. Comput. Biol. 24(7), 699-708, 2017.

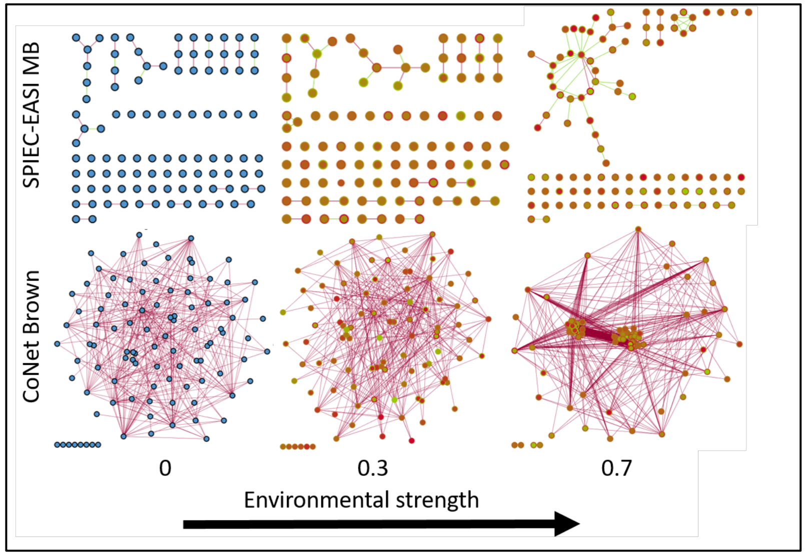
Tool evaluation II (with environment)

Data generation:

- Modular and scale-free interaction matrix (Klemm-Eguiluz)
- Simulations with generalized Lotka-Volterra including environmental effects
- Cross-sectional microbiome abundances generated

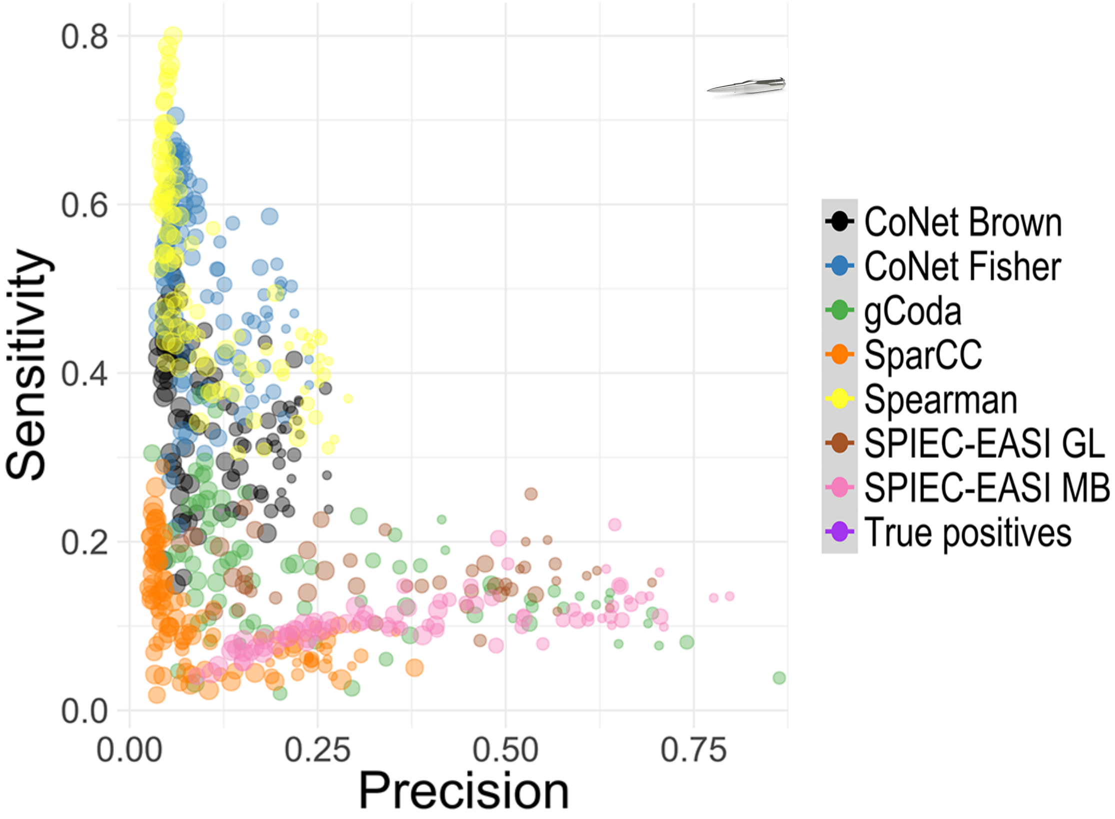
Sam Röttjers





With increasing environmental impact in the simulations, clusters form in the networks.

Tool evaluation II (with environment)



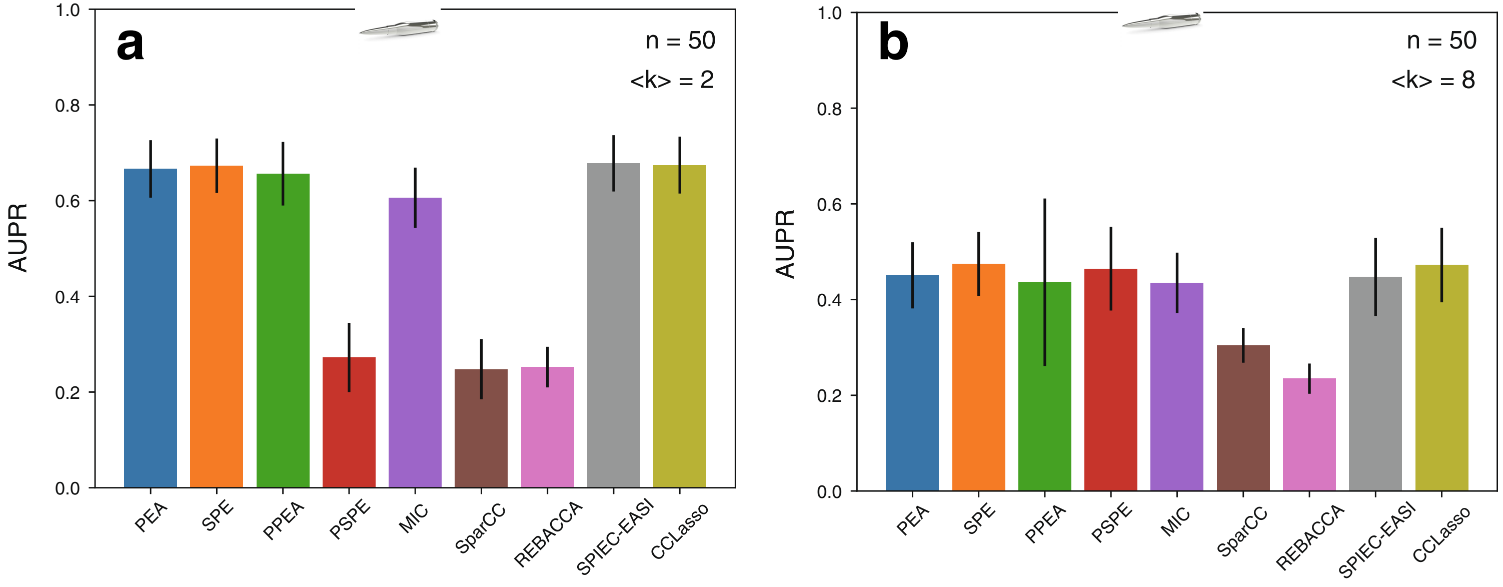
Node size scales with strength of environmental effect

- Tools based on inverse covariance (SPIEC-EASI, gCoda) are more precise, but less sensitive than other tools
- Increasing environmental effect tends to lower precision, especially in tools based on inverse covariance
- There is no silver bullet tool

Röttjers & Faust (2018) “From hairballs to hypotheses - biological insights from microbial networks” FEMS Microbiology Reviews 42, 761-780.

Tool evaluation III

- Data generated with gLV simulations
- Note the good performance of standard correlation methods
- FlashWeave not included



AUPR: area under the precision/recall curve

PEA: Pearson

PPEA: Pearson’s partial correlation

SPE: Spearman

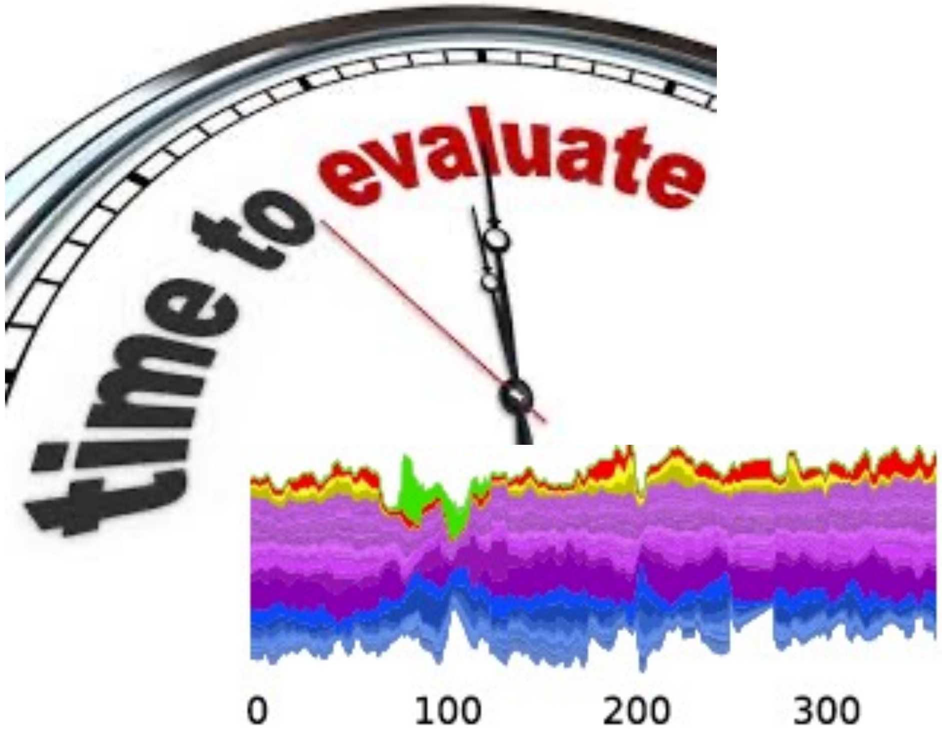
PSPE: Spearman’s partial correlation

Hirano & Takemoto (2019) ”Difficulty in inferring microbial community structure based on co-occurrence network approaches” BMC Bioinformatics 20, 329.

Microbial network inference evaluation

Evaluation of microbial network inference from time series

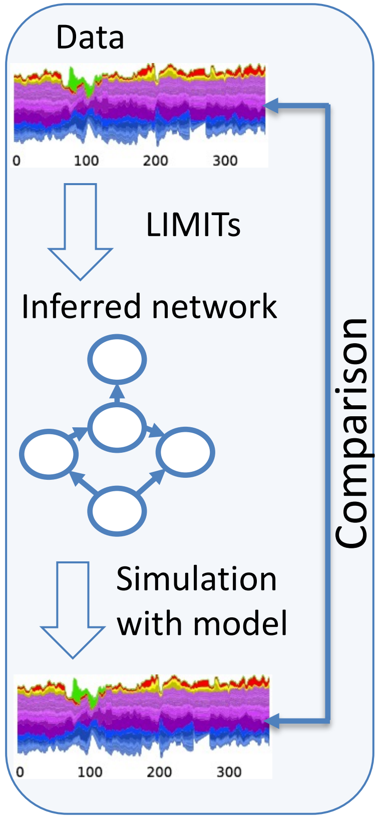
- How well do time series inference tools perform?



Tool evaluation for time series

- Time series generated with different population models (including gLV)
- Parameters (that is networks) known
- Networks inferred from simulated time series with LIMITS (the only tool evaluated)
- Two comparisons:
  - Known network directly compared to inferred network (accuracy of inference)
  - Original time series compared to time series generated with model parameterized with inferred network (goodness of fit)

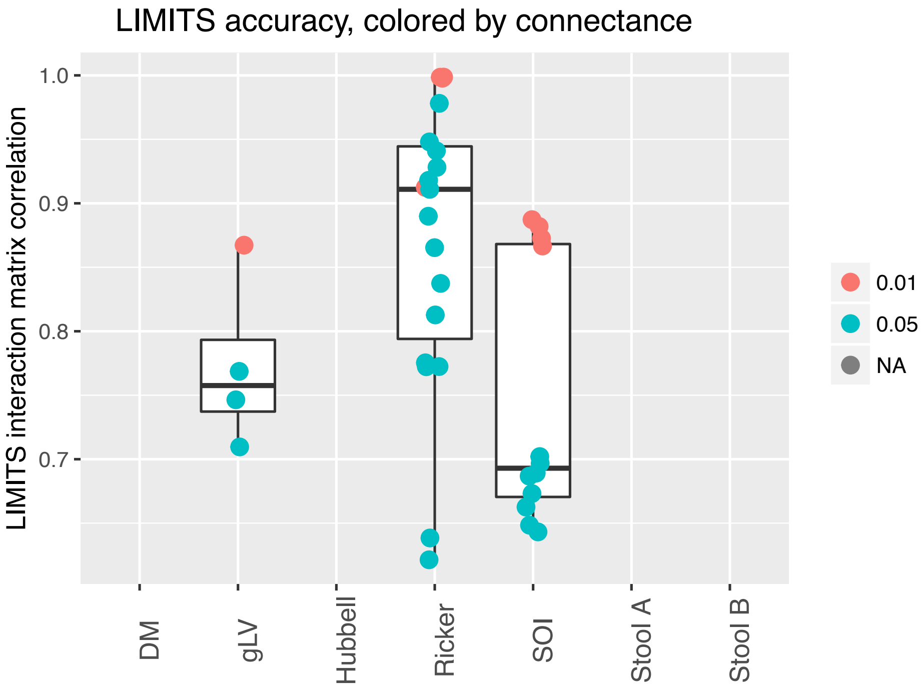
Goodness of fit

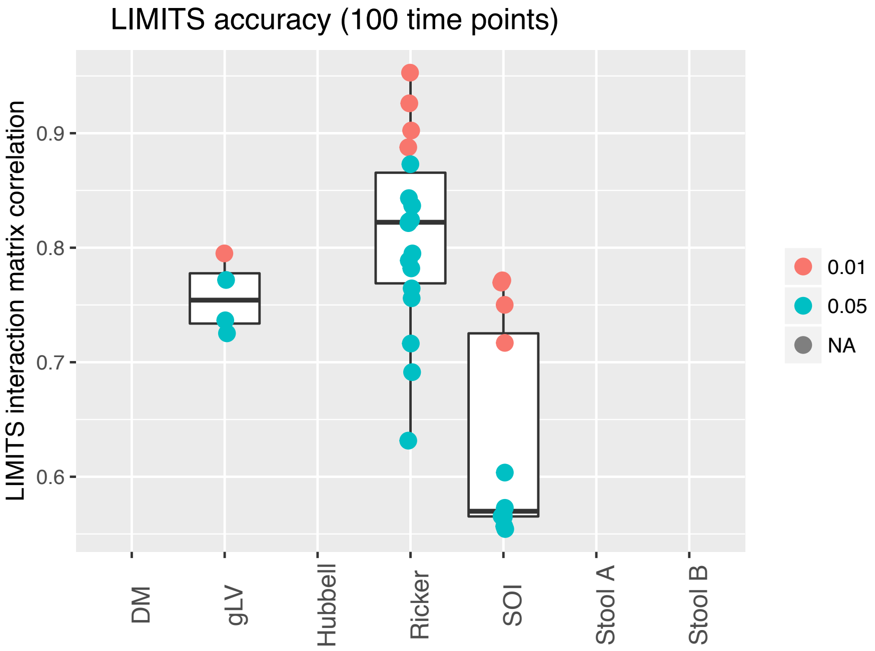


LIMITS: Fisher and Mehta (2014). “Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression.” *PLoS one* 9**,** e102451.
Evaluation: Faust et al. (2018) “Signatures of ecological processes in microbial community time series”, Microbiome 6, 120.

Tool evaluation for time series

- Interaction matrix known: compare inferred to known interaction matrix -> accuracy of inference
- The more links to infer, the lower the accuracy of LIMITS
- Accuracy for shorter time series is lower, but still reasonable
- Type of interaction model (gLV, Ricker, SOI) does not matter much





DM: Dirichlet-Multinomial

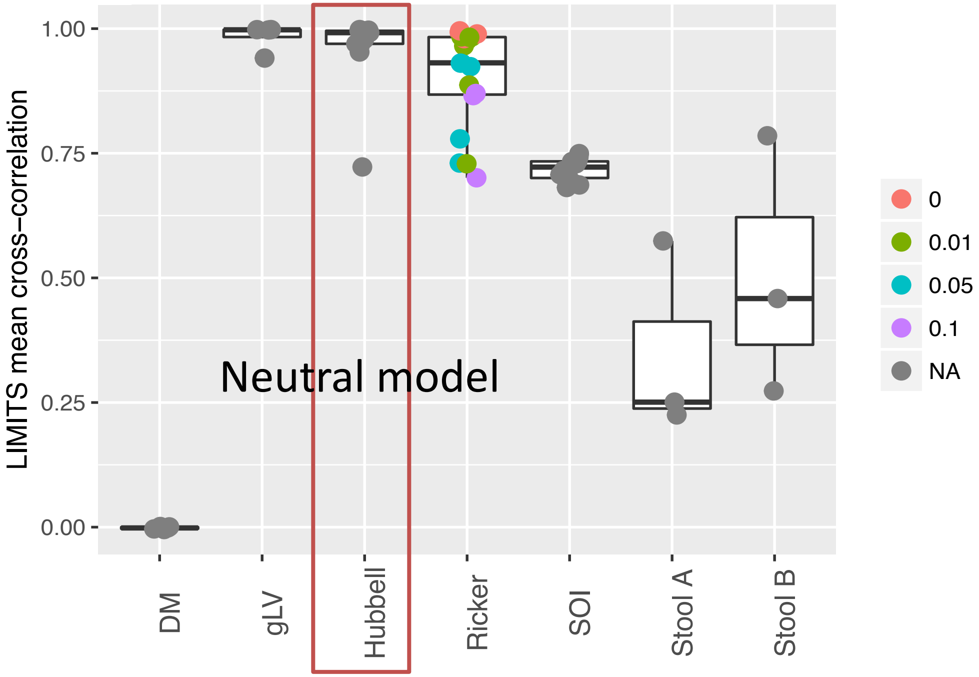
100 species, 3000 time points

100 species, 100 time points

Tool evaluation for time series

- Interaction matrix unknown: compare observed time series to those generated with parameterized interaction model -> goodness of fit
- **Goodness of fit can be misleading**: it is high even for a neutral model that does not take interactions into account explicitly (over-fitting)

LIMITS goodness of fit, colored by sigma

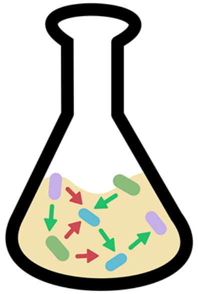


100 species,
3000 time points
Sigma: level of internal noise

See also: Cao et al. (2017) “Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons” Bioessays 39(2).

Evaluation of microbial network inference on biological data

- How well do the tools perform on biological data?



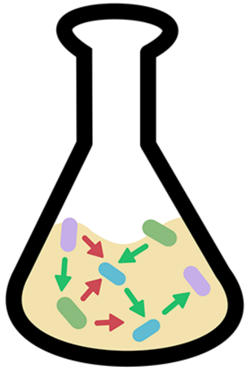
Estimate performance on synthetic community

Compare interactions to reference database

Microbial network inference evaluation

The challenges of biological validation

- With complex ecosystems in situ, it is hard to know whether two species do not interact (confirming the negative is harder than confirming the positive)
- That means that we can only assess sensitiviy in situ
- In vitro, we can measure all pair-wise interactions comprehensively and thus can assess accuracy
- But we don’t know whether interactions in vitro also happen in situ
- And there may be higher-order interactions (roughly: an interaction between two species that is modified by the presence of additional species)



Biological validation of interaction prediction: mixed results

Arabidopsis root (Durán et al., Cell 2018)
Tool: Spearman/SparCC
Data: 16S on 144 plant samples
Validation data: high-throughput screen of 2,862 antagonistic bacterial-fungal interactions
Result: predictions for ca. 24 out of 32 tested bacterial OTUs confirmed

Artificial community (Biswas et al., Lecture Notes in Bioinformatics 2015)
Tool: MInt
Data: 16S on synthetic 9-species community
Validation data: co-growth on plates
Result: 2 out of 2 edges confirmed
100% accuracy (no false negatives)

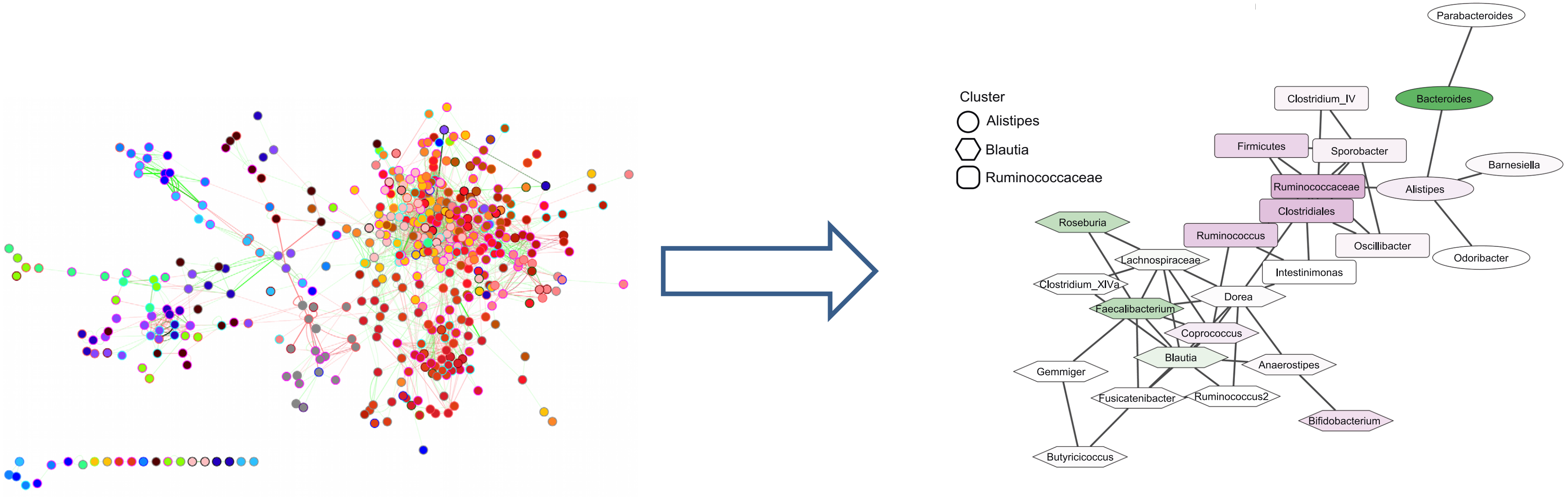
TARA Oceans (Lima-Mendez et al., Science 2015)
Tool: CoNet
Data: 16S/18S on 313 open-ocean samples
Validation data: genus-level eukaryotic interactions from the literature (mostly endosymbiosis)
Known pairs: 43
Sensitivity: 42% (18 found)
Precision: uncertain
Note: one novel interaction confirmed using microscopy

Phage-Host (Edwards et al., FEMS 2015)
Tool: Pearson
Data: 3025 global metagenomic samples
Validation data: Known hosts for 820 phages
Sensitivity: hosts correctly predicted for 9.5% of the phages (78 found)
Precision: uncertain

Note that global metagenomic data set filters for cosmopolitan phages, however tested phages may not be cosmopolitan

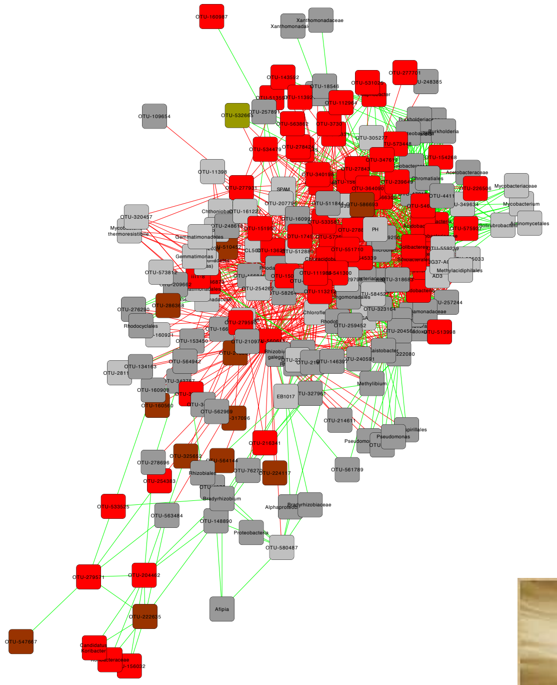
Questions so far?

Part II: Microbial network analysis



From hairballs to biological hypotheses

Low-accuracy hairball



Hypotheses

- Who interacts with whom
- Which taxa respond to which environmental factors
- Which taxa cluster together and why
- ….

Microbial network analysis examples

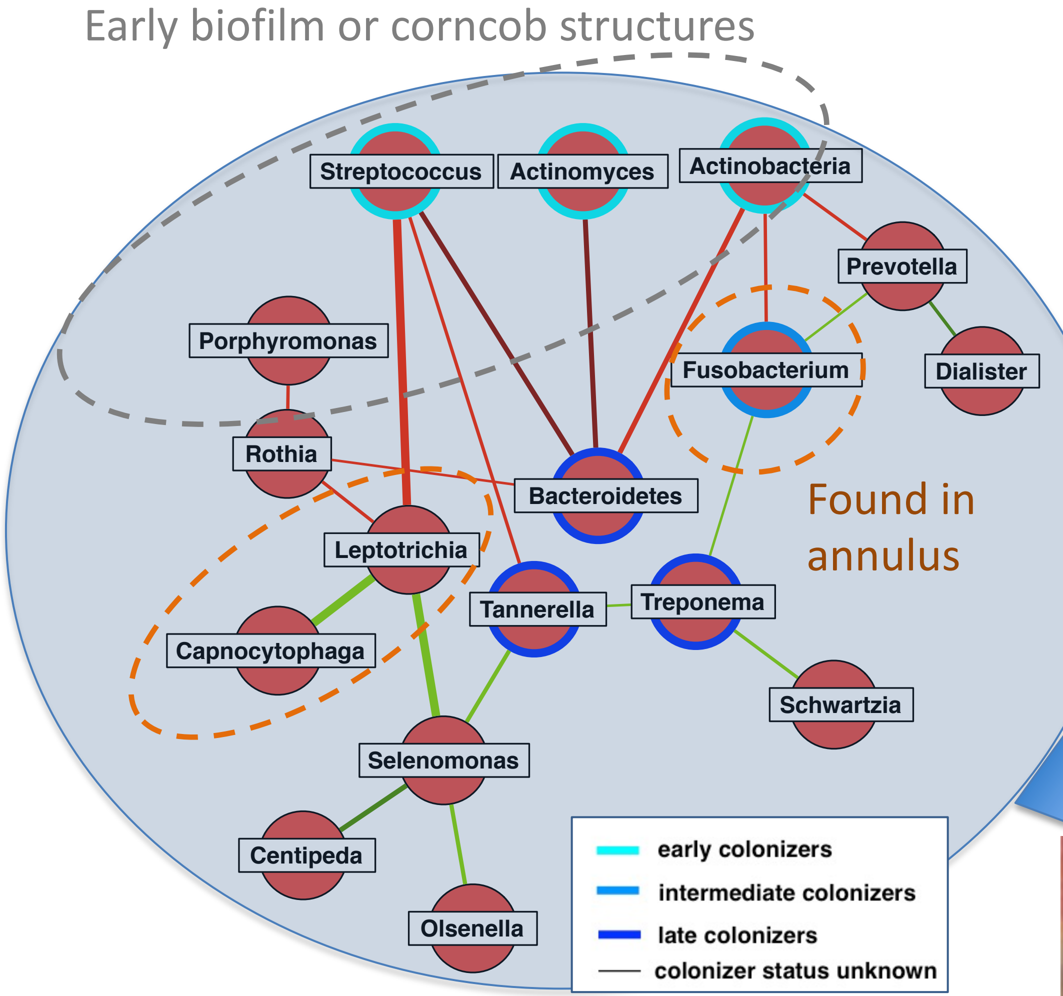
Example 1: Human Microbiome Project data

- 242 healthy individuals
- sampled in up to 18 body sites
- 16S & metagenomic sequencing
- Metadata

The Human Microbiome Project Consortium.
Nature 486, 207-214 (2012).
Nature 486, 215-221 (2012).

HMP data: Niche structure

Dental plaque subnetwork:


Early biofilm or corncob structures

Found in annulus

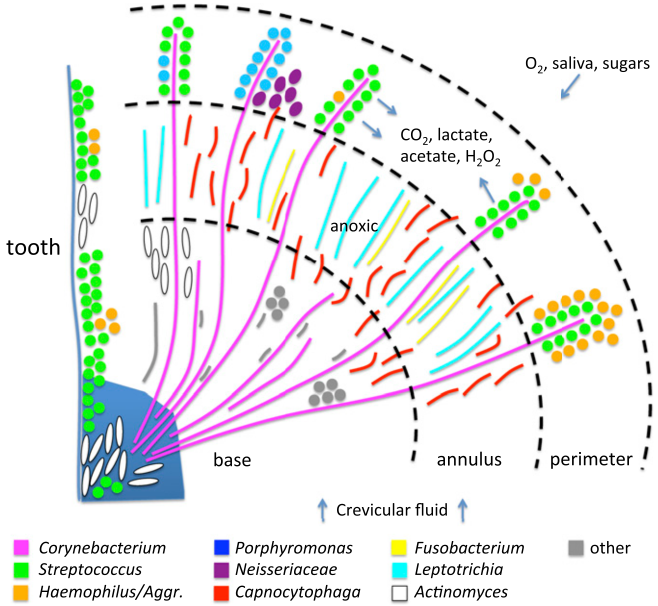


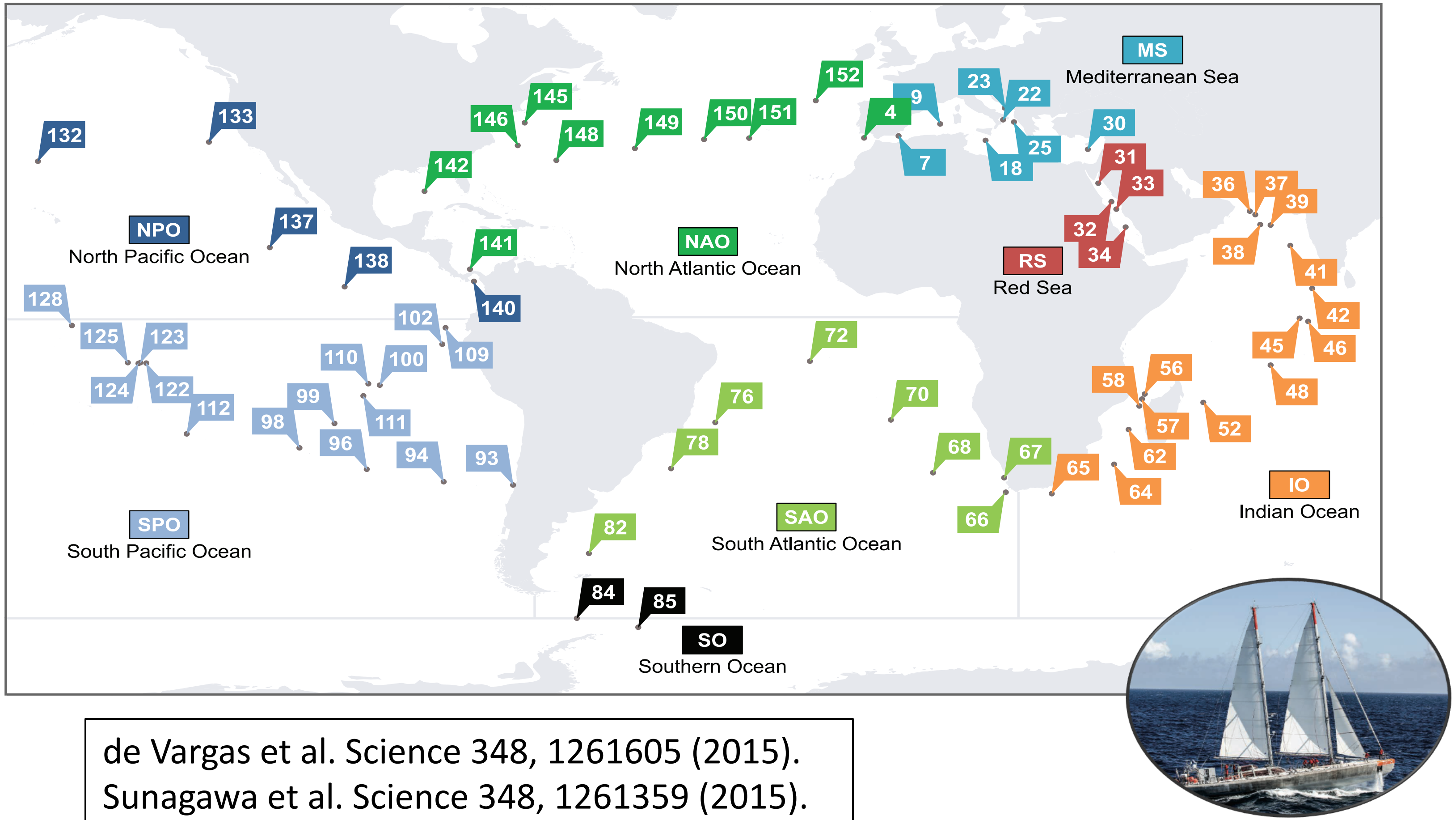
Image taken from de Welch et al. PNAS, E791-E800 (2016).

Faust et al. (2012) PLoS Computational Biology 8 (7) e1002606.

Microbial network analysis examples

Example 2: TARA Oceans data

- Global marine expedition, >200 stations spanning 8 oceanic regions, sampled at 2-3 depths
- 18S (4 cell size fractions), 16S, viral contigs



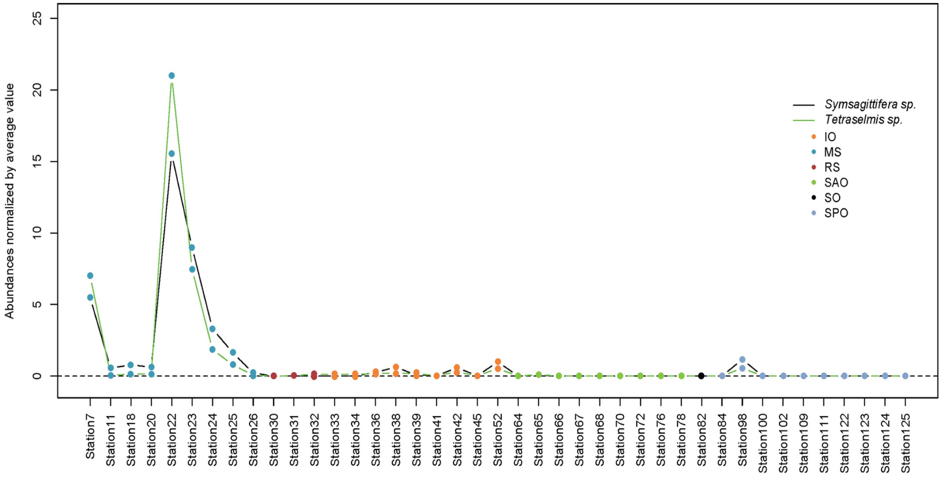
de Vargas et al. Science 348, 1261605 (2015).
Sunagawa et al. Science 348, 1261359 (2015).
Pesant et al. Scientific Data 2, 150023 (2015).

Microbial network analysis examples

TARA: Interaction prediction

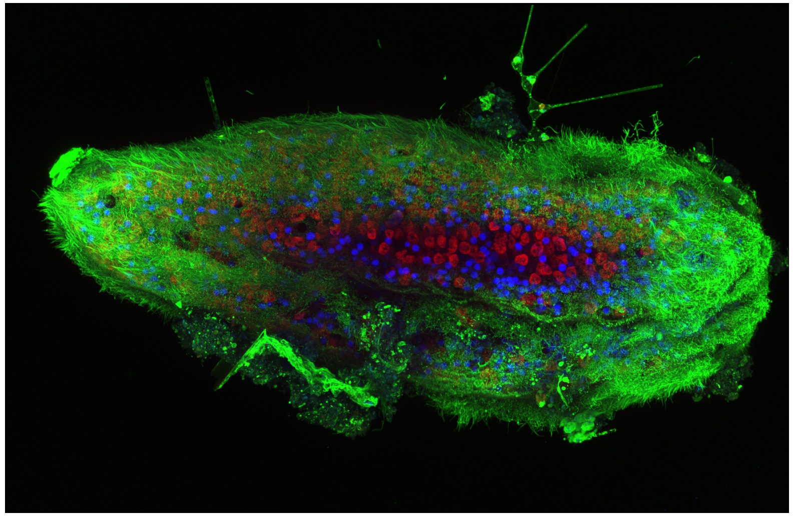
- Interaction candidate in TARA Ocean data

Prediction



Abundance profiles from 18S marine phytoplankton data

Experimental validation (microscopy)



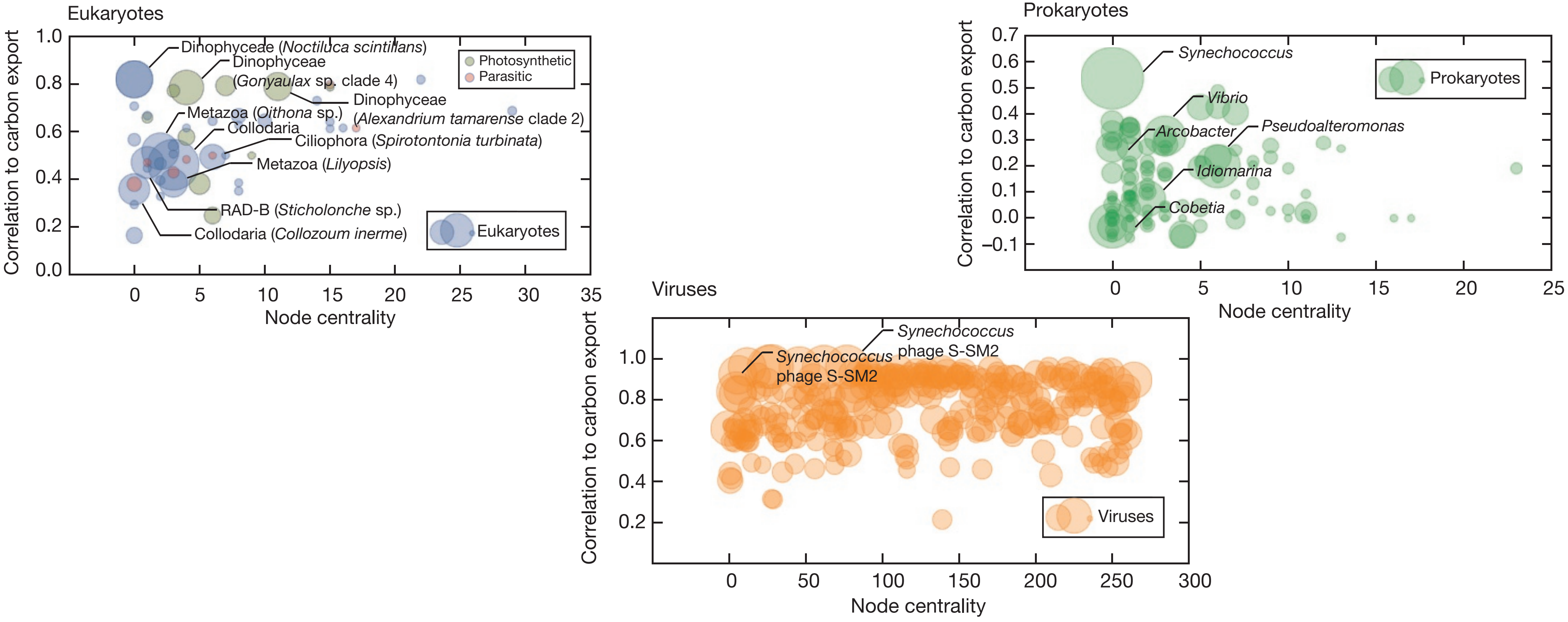
Flatworm with photosynthetic microalgal endosymbionts

Lima-Mendez*, Faust*, Henry* et al. (2015) “Determinants of community structure in the global plankton interactome” Science 348, 1262073.

Microbial network analysis examples

TARA: Linking taxa to function

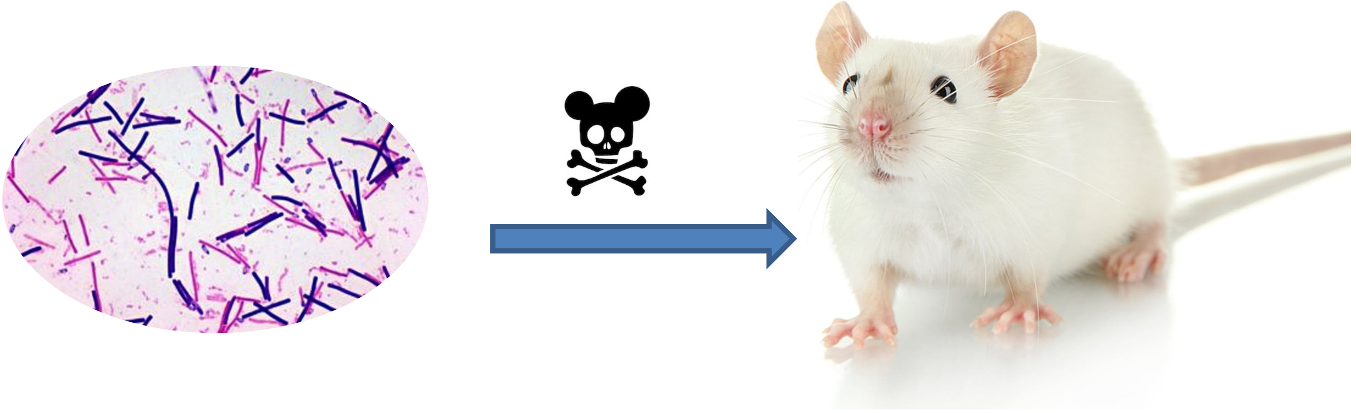
- Clustering of TARA oceans microbial network using WGCNA
- Cluster representatives screened for strong association to carbon export: Synechococcus (cyanobacterium) identified



Guidi et al. (2016) “Plankton networks driving carbon export in the oligotrophic ocean” Nature 532, 465-470.

Example 3: Network inference from time series

- *Clostridium difficile* is an intestinal pathogen in mammals
- It can thrive when killing gut microbiota with antibiotics
- Experiment: Mice infected with *C. difficile* after exposure to different antibiotics

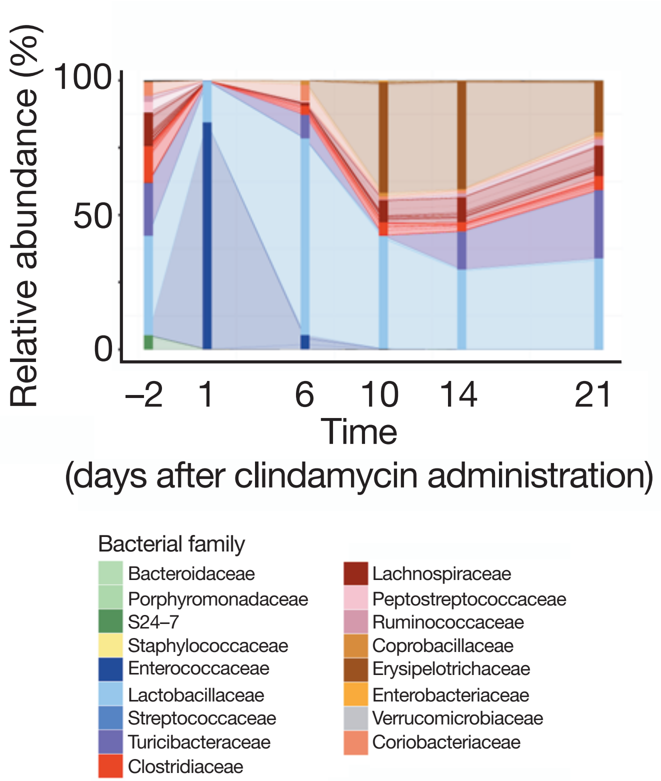


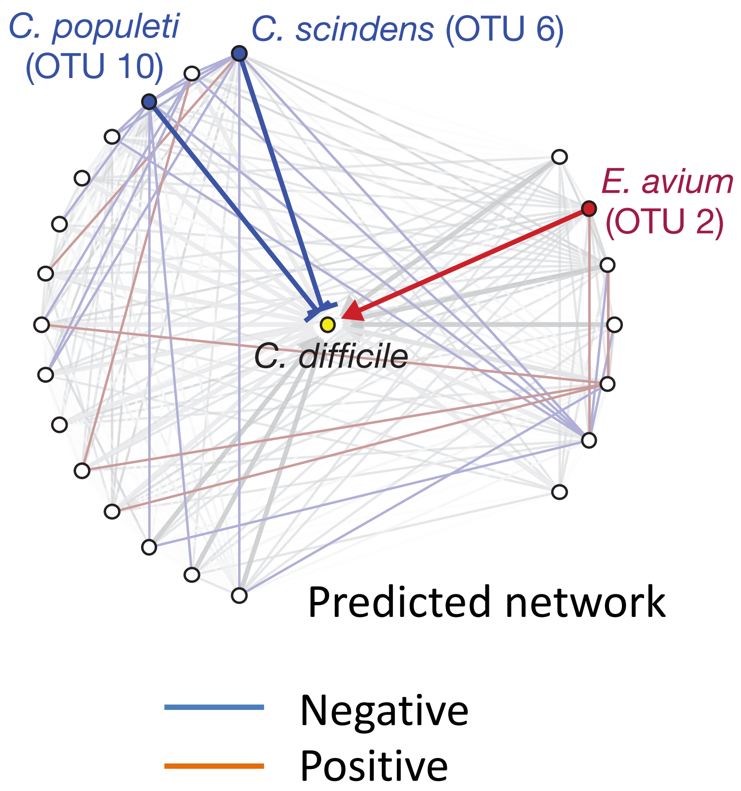
Buffie et al. (2014) “Precision microbiome reconstitution restores bile acid mediated resistance to Clostridium difficile” Nature 517, 205-208.

Microbial network analysis examples

Example 3: Network inference from time series

- Bacterial interaction network predicted from fecal 16S time series of mice by parameterising gLV model



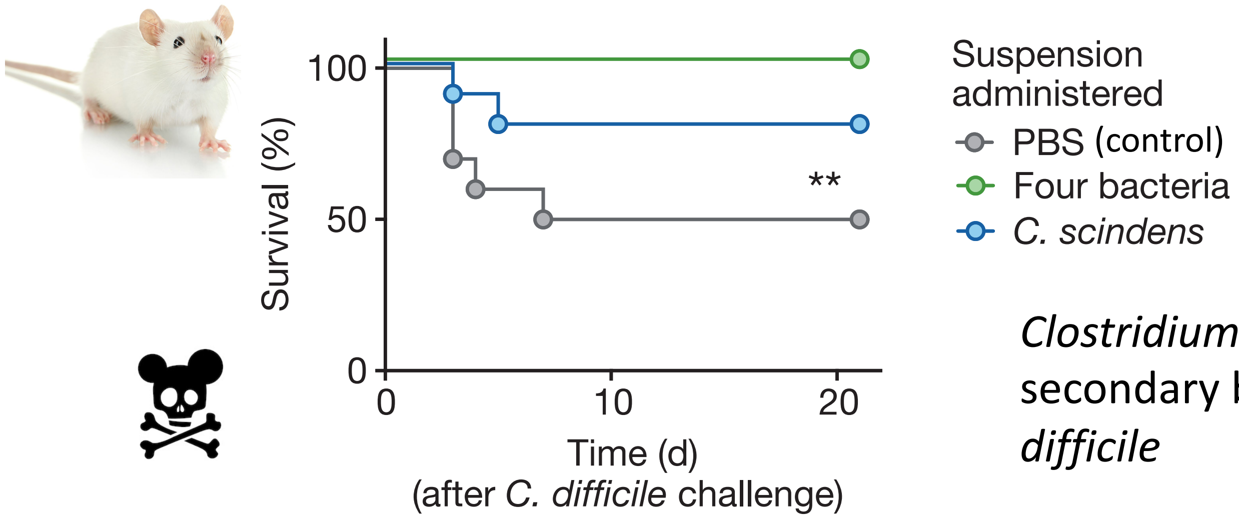


Predicted network

Buffie et al. Nature 517, 205-208 (2014).

Example 3: Network inference from time series

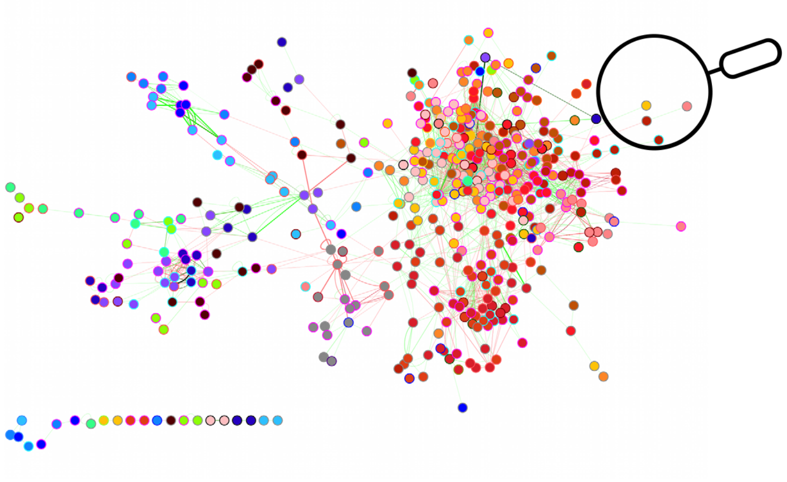
- Treating mice with bacteria that interact negatively with *C. difficile* increases their survival rate



Clostridium scindens produces secondary bile acids that inhibit *C. difficile*

Buffie et al. Nature 517, 205-208 (2014).

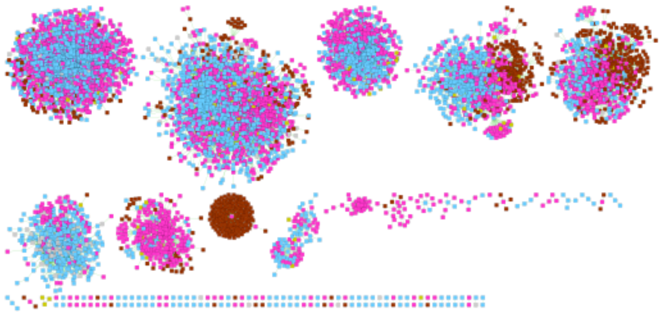
Microbial network analysis tools



Manta: Microbial network clustering

- Challenges:
  - Most existing cluster algorithms (e.g. MCL) do not exploit information given in negative edges
  - Microbial networks have a low accuracy

- Manta addresses these challenges

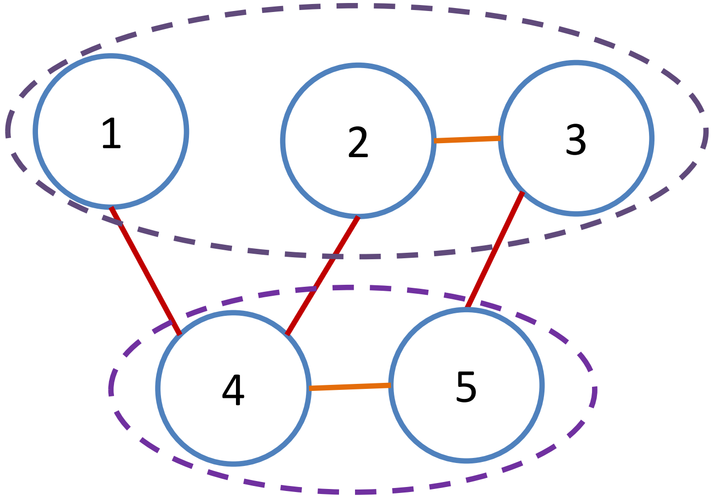


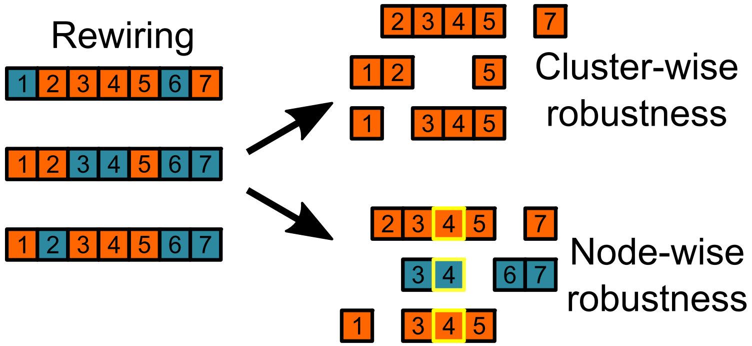
Sam Röttjers

Microbial network analysis tools

Manta – key ideas

- Use the principle: “an enemy of an enemy is a friend” to group taxa that share “enemy nodes” (nodes linked with negative edges)
- Weak node assignments (nodes that cannot be clustered)
- Repeat clustering on partly rewired networks to assess robustness of clusters and cluster memberships





Röttjers & Faust (2020) “manta-a clustering algorithm for weighted ecological networks” mSystems 5 e00903-19.

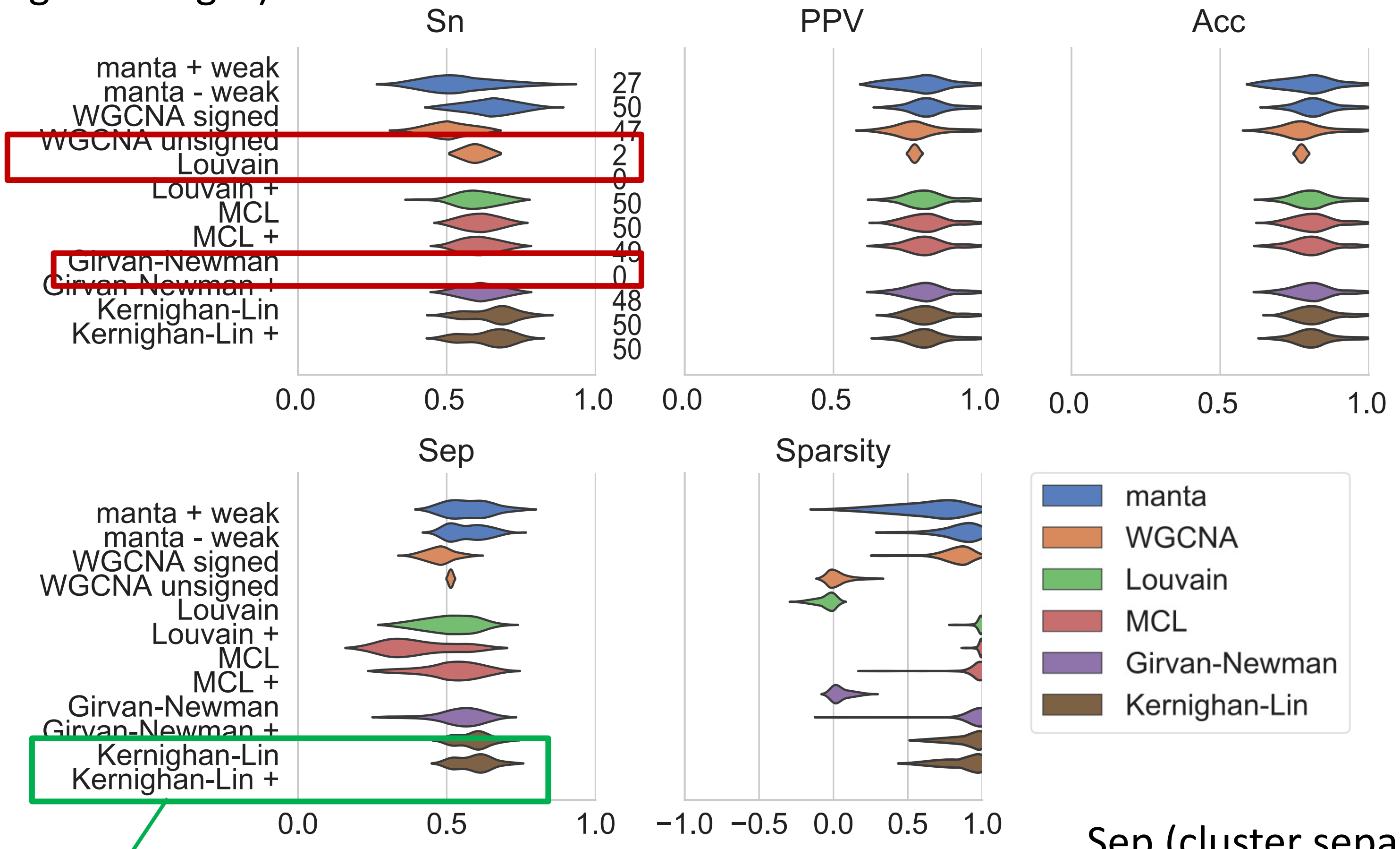
Does manta work? How to evaluate a network cluster algorithm?

- Need a data set with known clusters to check whether the tool finds them back
- Microbiome data with known clusters are hard to find
- Generate synthetic microbial abundances with known clusters for manta:
  - 1. Population model simulating different environmental effects on predefined groups of taxa
  - 2. Bicluster generation with FABIA
- Choice of data generation process often biases tool evaluation (two processes better than one)

Hochreiter et al. (2010) "FABIA: factor analysis for bicluster acquisition." Bioinformatics 26, 1520-1527.

Evaluation on population model

Louvain, GN and WGCNA unsigned failed (negative edges)



KL performs well but only allows 2 clusters.

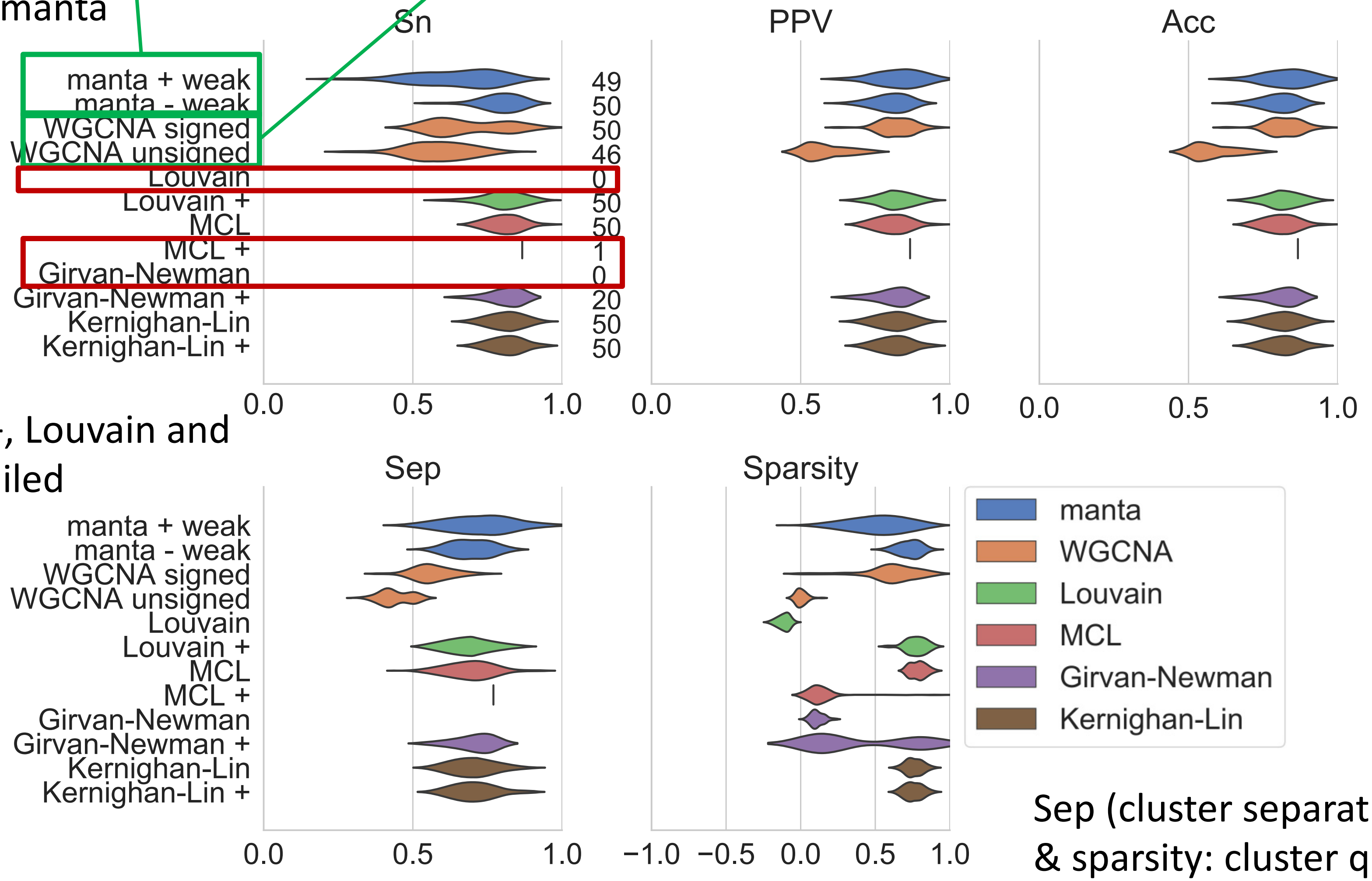
Sep (cluster separation) & sparsity: cluster quality scores

Evaluation on biclusters

Weak node assignments help manta

Keeping the edge signs improves WGCNA

MCL+, Louvain and GN failed



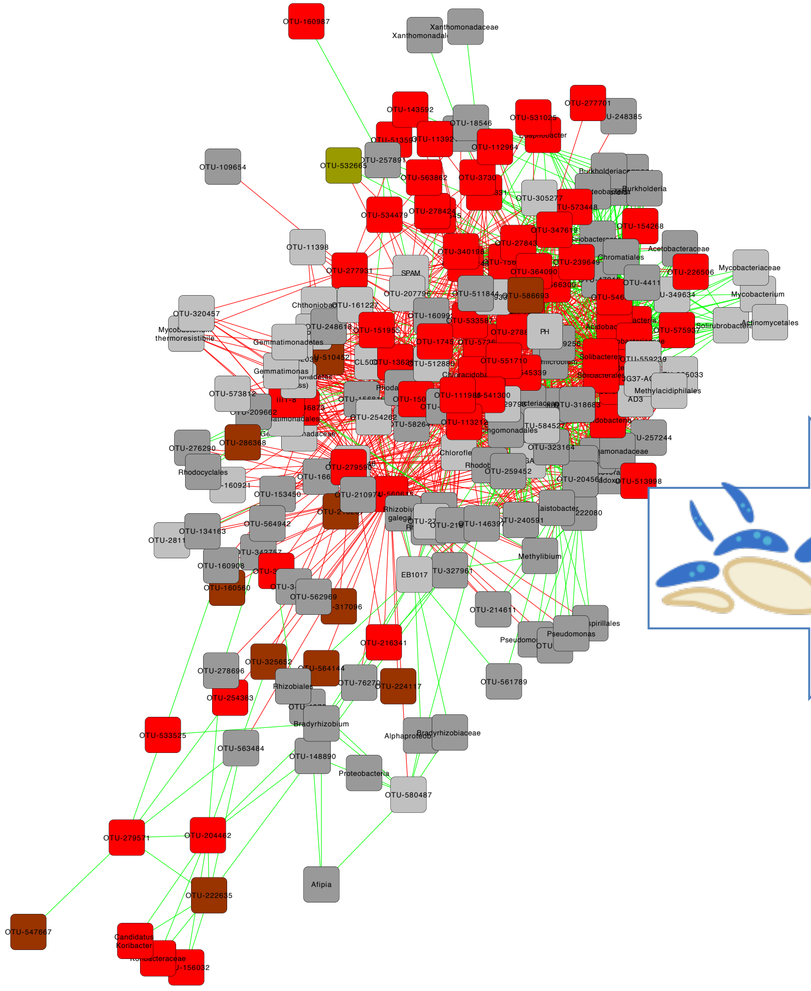
Sep (cluster separation) & sparsity: cluster quality scores

Microbial network clustering: lessons

- If your network contains negative edges, use tools that support them (signed WGCNA, Kernighan Lin, manta)
- If you think that there are only 2 clusters, run Kernighan Lin
- WGCNA makes an assumption about the network structure (i.e. that it is scale-free), which may not be true

Manta in action: Tundra soil

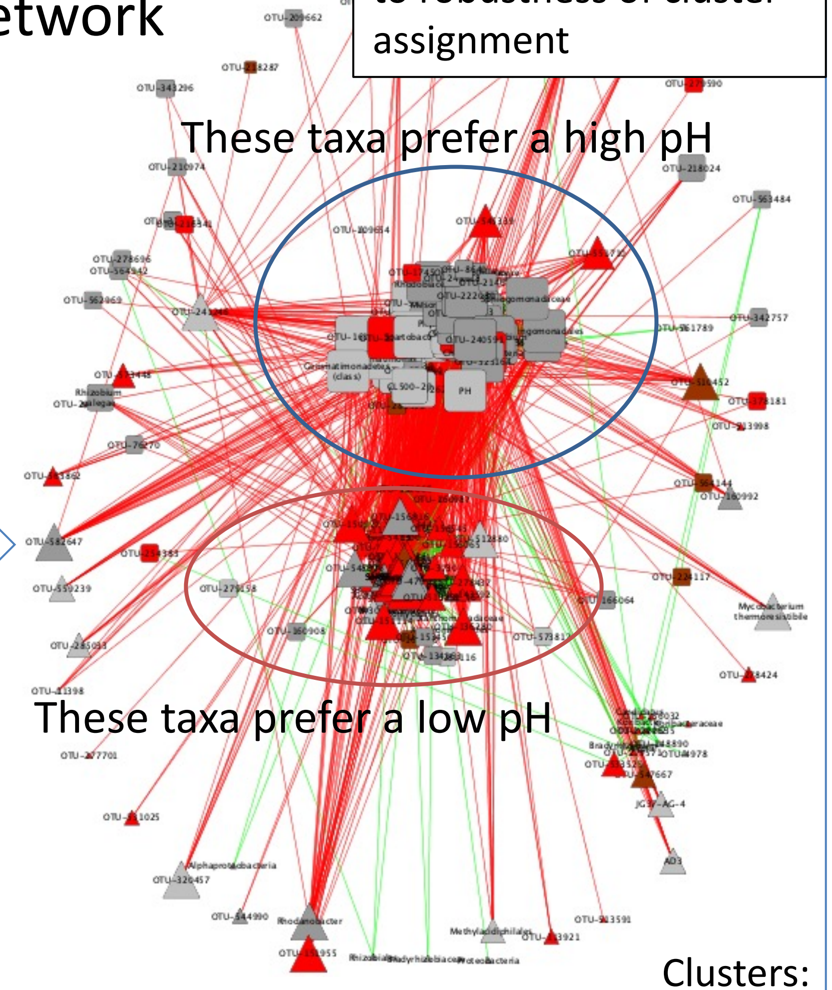
Hairball



Tundra soil network
Layout: organic

Informative network

Node size is proportional to robustness of cluster assignment



These taxa prefer a high pH

These taxa prefer a low pH

Clusters:

🟦

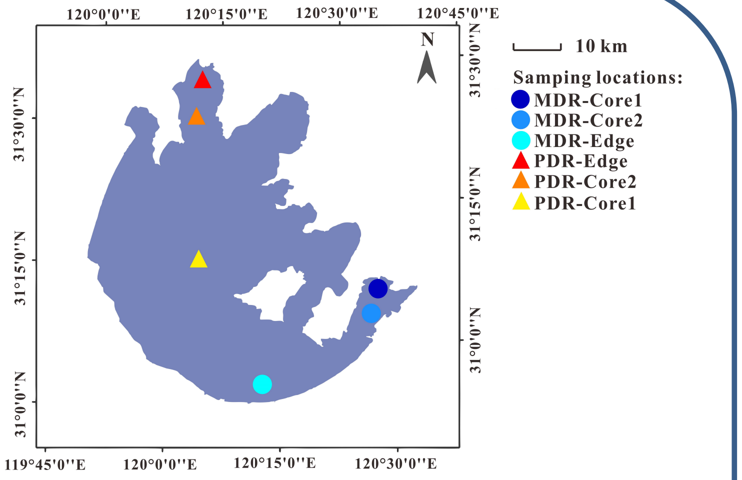
▲

Tundra soil network after clustering & layout with manta

Microbial network analysis tools

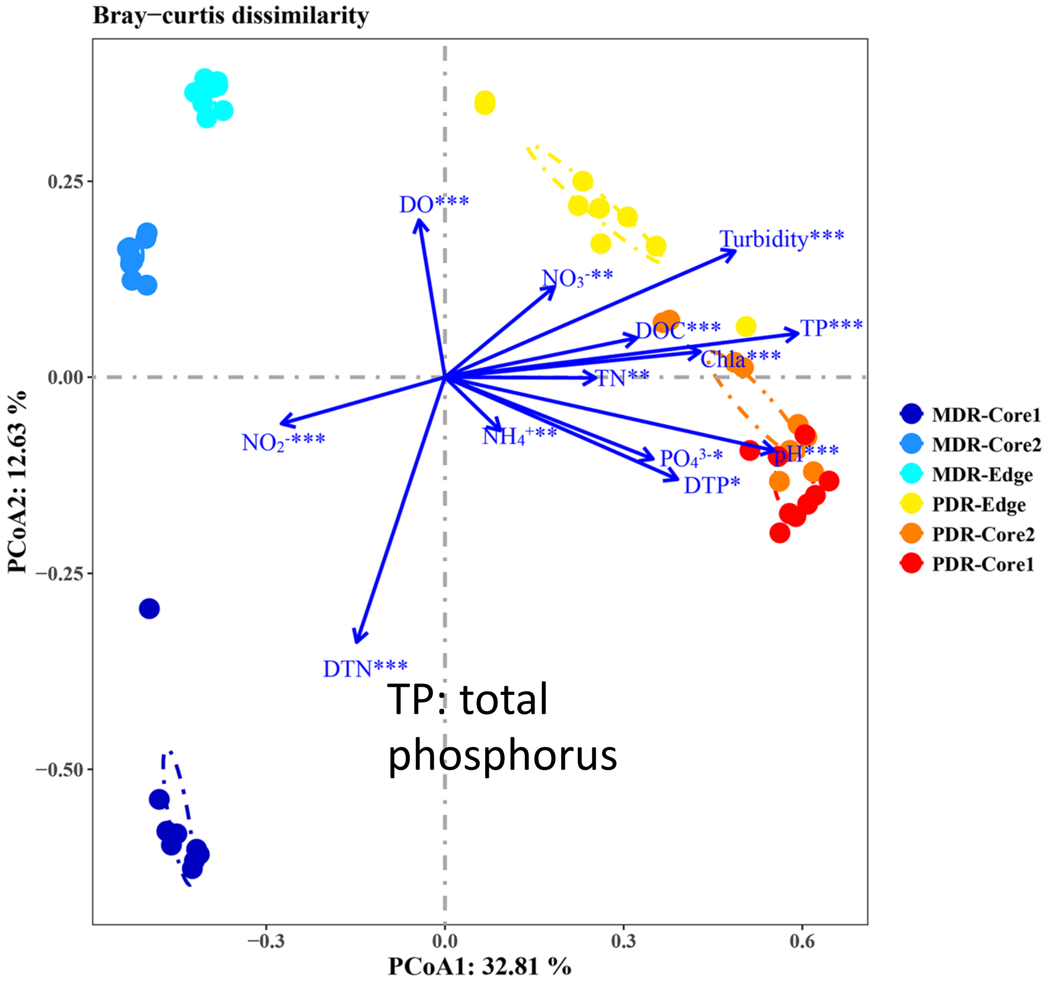
Manta in action: Lake Taihu

Eutrophication: two regimes driven by nutrient concentration



Macrophyte-dominated regime (MDR)

Phytoplankton-dominated regime (PDR)

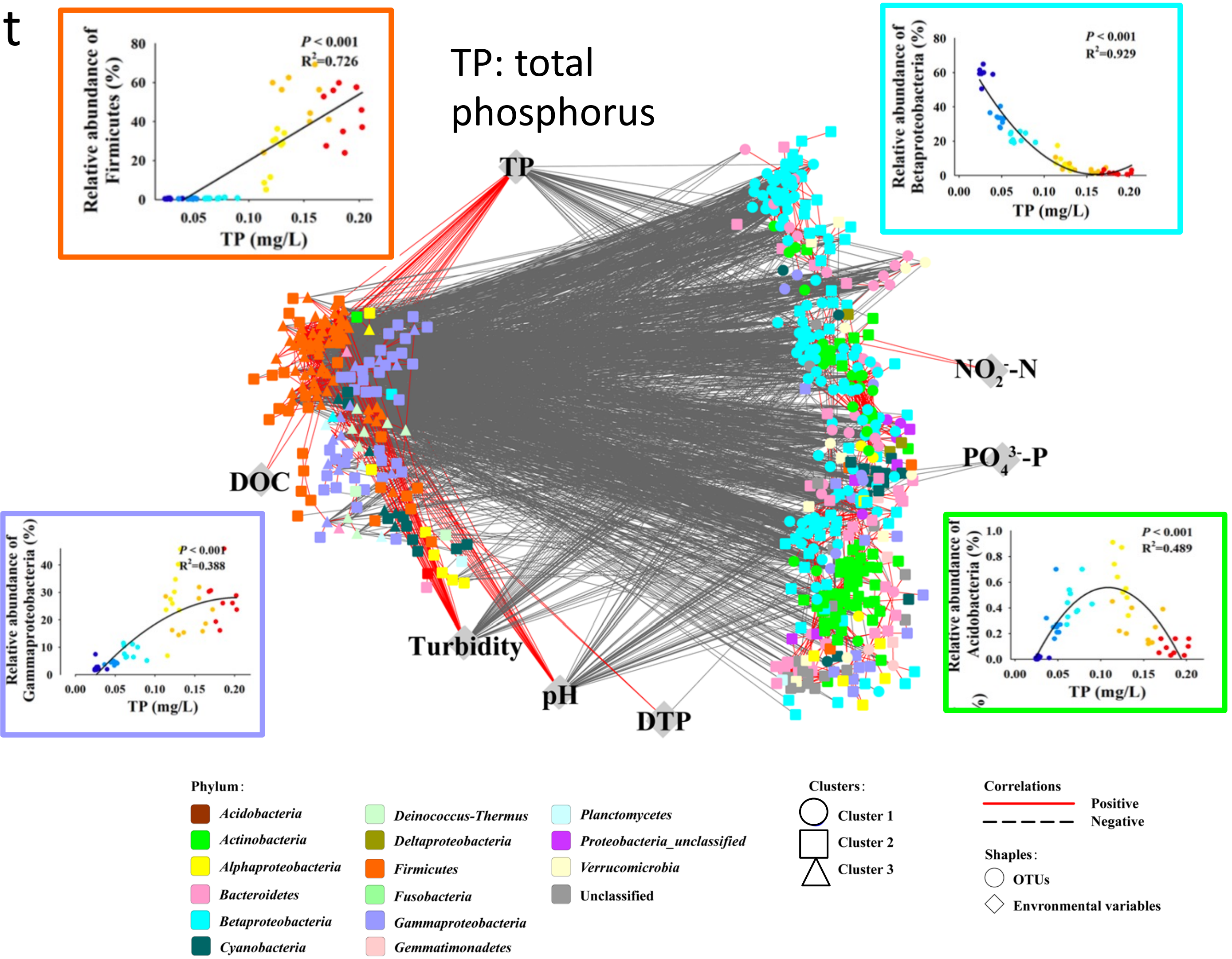


TP: total phosphorus

Microbial network analysis tools

Manta in action: Lake Taihu

- Network constructed with CoNet on all 54 samples and clustered with manta
- Cluster 1 mostly consists of Betaproteobacteria, which decline with nutrient levels
- Cluster 3 is dominated by Firmicutes, which tend to increase with nutrient levels
- Cluster 2: contains phyla with non-linear responses: Gammaproteobacteria (saturation) and Actinobacteria (optimum)

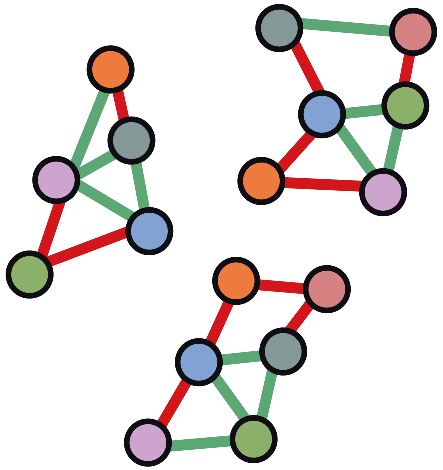

TP: total phosphorus

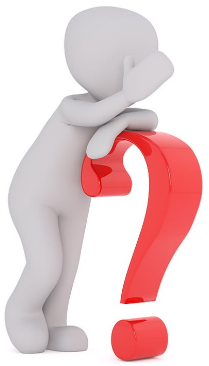
Cao, Zhao, Li, Röttjers, Faust and Zhang (2022) Microbial Ecology Accepted.

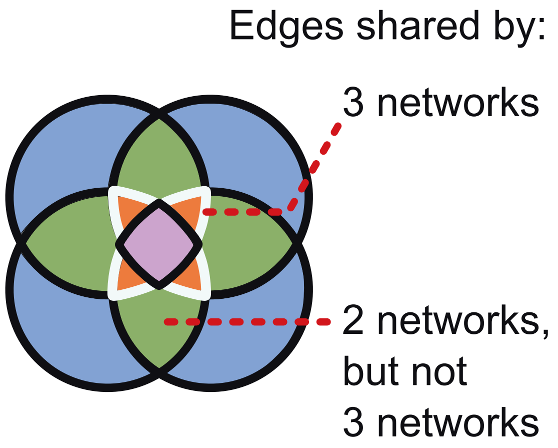
Microbial network comparison

- We can construct a set of networks e.g., one gut microbial network per person
- Is a network core present or do networks overlap not more than expected by chance?

Sam Röttjers







Edges shared by:

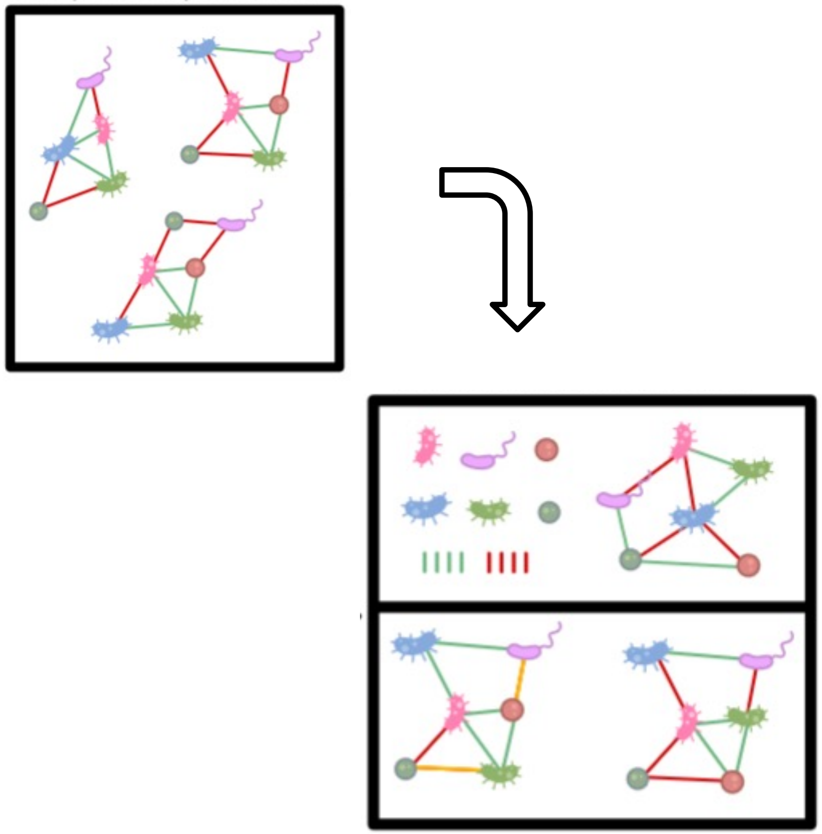
3 networks

2 networks, but not 3 networks

Microbial network analysis tools

Anuran: a toolbox for comparing noisy microbial networks

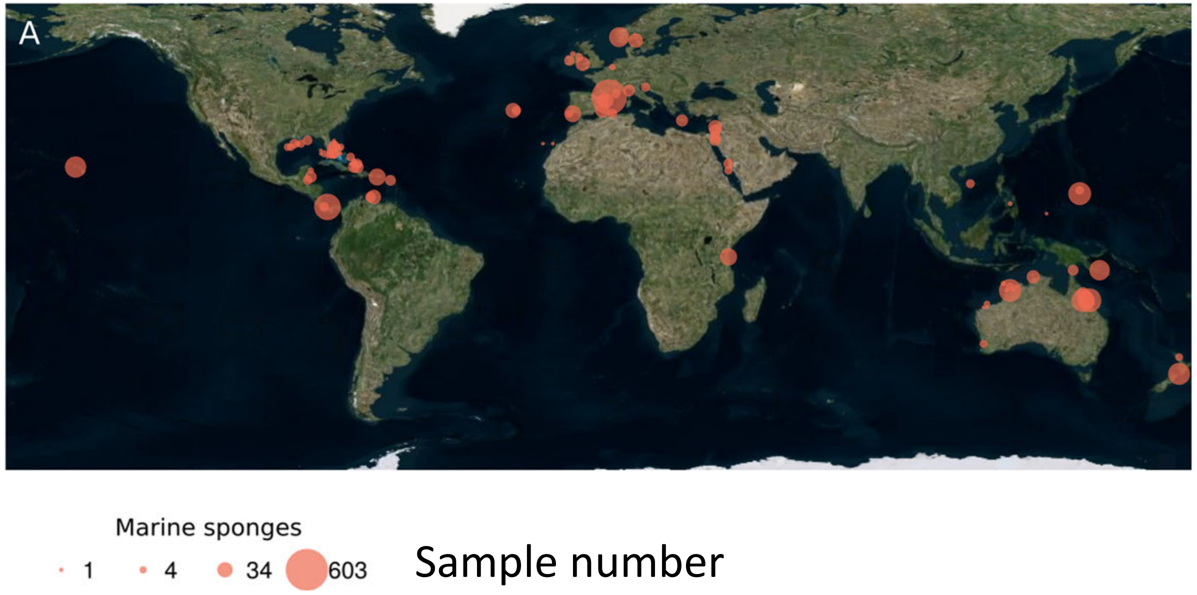
- Implements 2 types of null models: network randomization with and without preserving node degree distribution
- Tests whether a network property or a core network is significant given randomized networks



Röttjers S, Vandeputte D, Raes J and Faust K (2021). “Null-model-based network comparison reveals core associations” ISME Communications 1, 36

Anuran in action: Sponge microbiome

- Sponge microbiome project: microbiota of 268 different sponge host species collected around the globe
- > 3000 samples from ten sponge orders sequenced



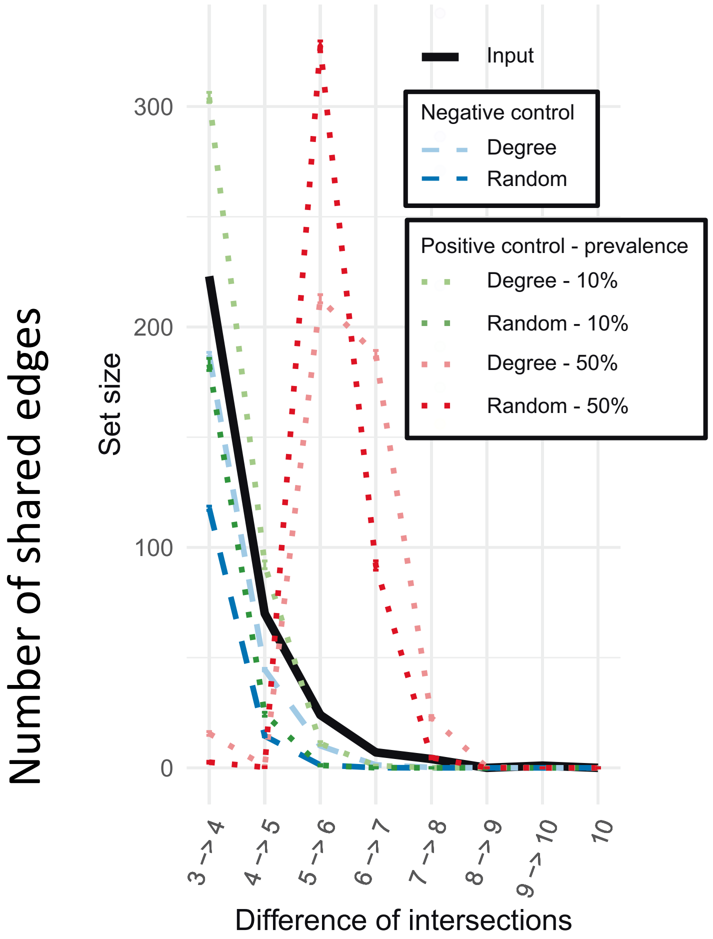
Sample number

Moitinho-Silva, L., Nielsen, S., Amir, A., Gonzalez, A., Ackermann, G. L., Cerrano, C., ... & Steinert, G. (2017). The sponge microbiome project. *GigaScience*, *6*(10), gix077.

Microbial network analysis tools

Anuran in action: Sponge microbiome

- Ten sponge-order-specific networks constructed with CoNet
- Number of shared edges is significant for edges conserved in three networks -> core network

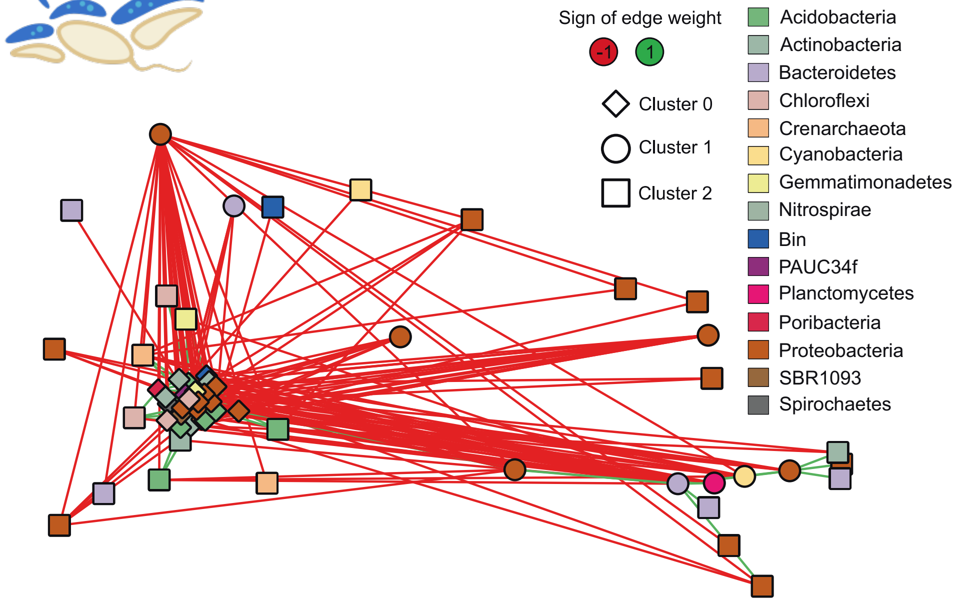


= Fraction of networks in which shared edges occur

Microbial network analysis tools

Anuran in action: Sponge microbiome

- Core network clustered with manta
- Clusters contain indicator taxa for high versus low microbial abundance sponges (HMA vs LMA)
- HMA vs LMA classification traverses sponge orders
- No 100% core expected (there are no more highly preserved edges than expected at random)

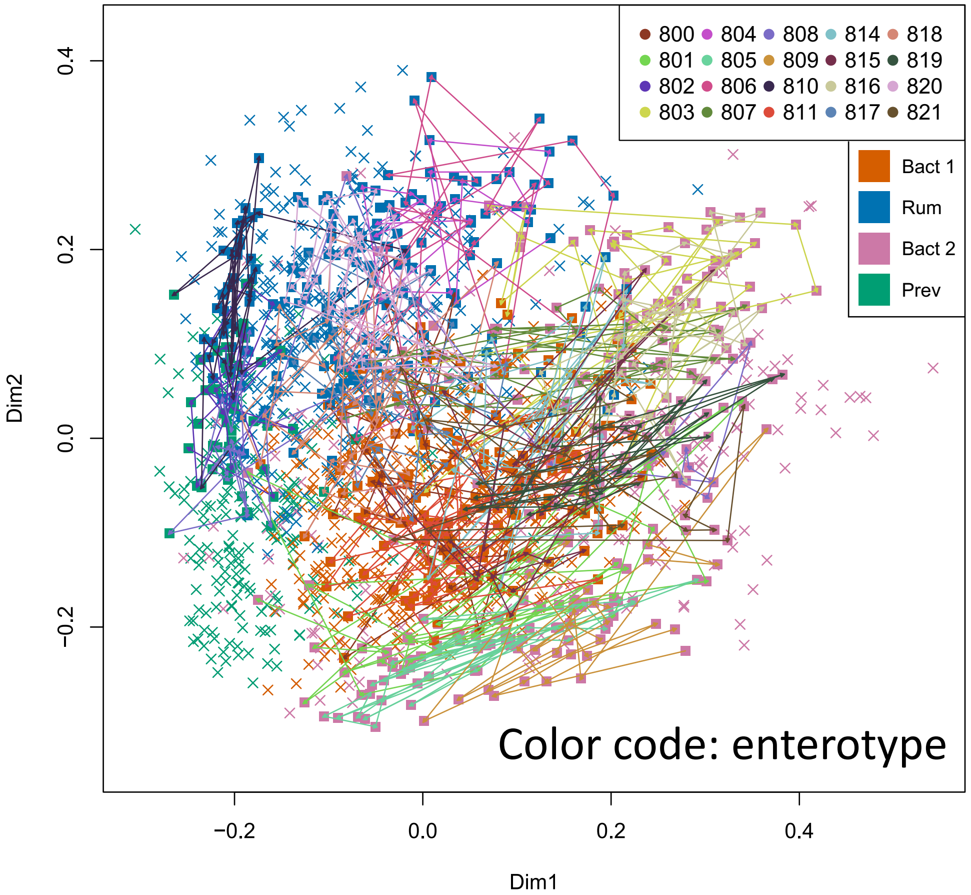


Cluster 0: enriched in HMA phyla
Cluster 1: only LMA phyla

Moitinho-Silva, L., Steinert, G., Nielsen, S., Hardoim, C. C., Wu, Y. C., McCormack, G. P., ... & Hentschel, U. (2017). Predicting the HMA-LMA status in marine sponges by machine learning. *Frontiers in microbiology*, *8*, 752.

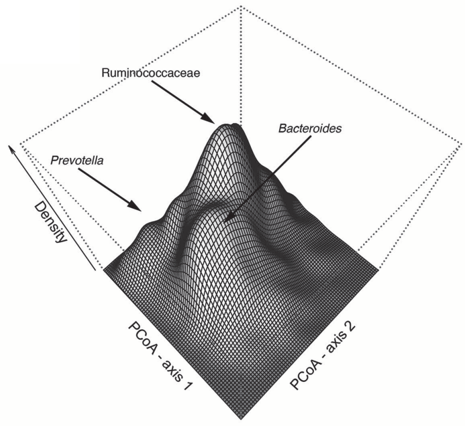
Anuran in action: Human gut microbiome

- Fecal samples collected for 20 women over six weeks (713 samples) and sequenced (16S rRNA)
- 20 microbial networks constructed with fastLSA



Color code: enterotype

PCoA of time series data collected from 20 women, mapped onto the Flemish Gut Flora Data set (1104 samples)

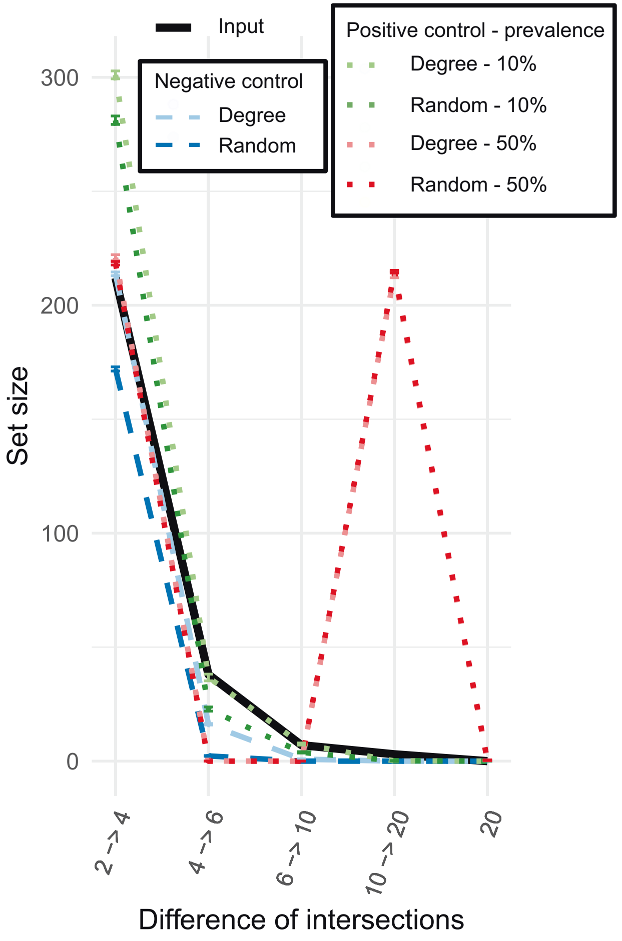


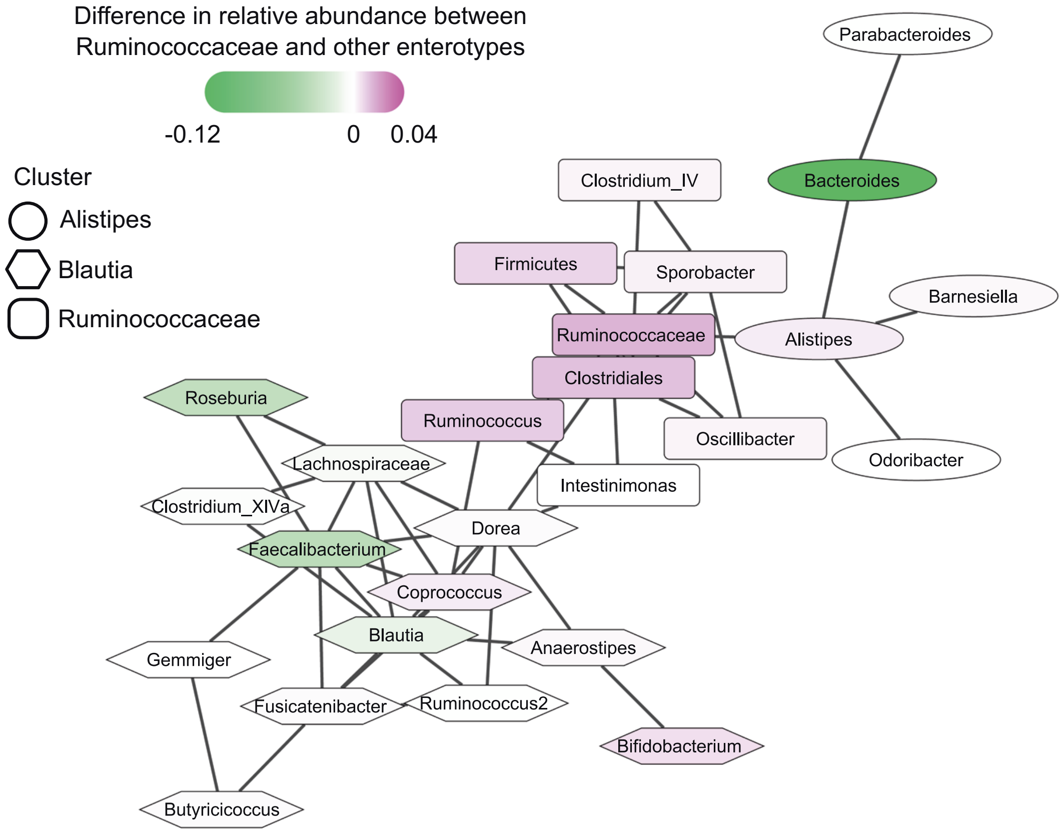
Enterotypes: gut microbial compositions dominated by different genera (Rum = Ruminococcus, Bact = Bacteroides, Prev = Prevotella)

Vandeputte,..., Faust, Raes (2021) Nature Communications 12:6740.

Anuran in action: Human gut microbiome

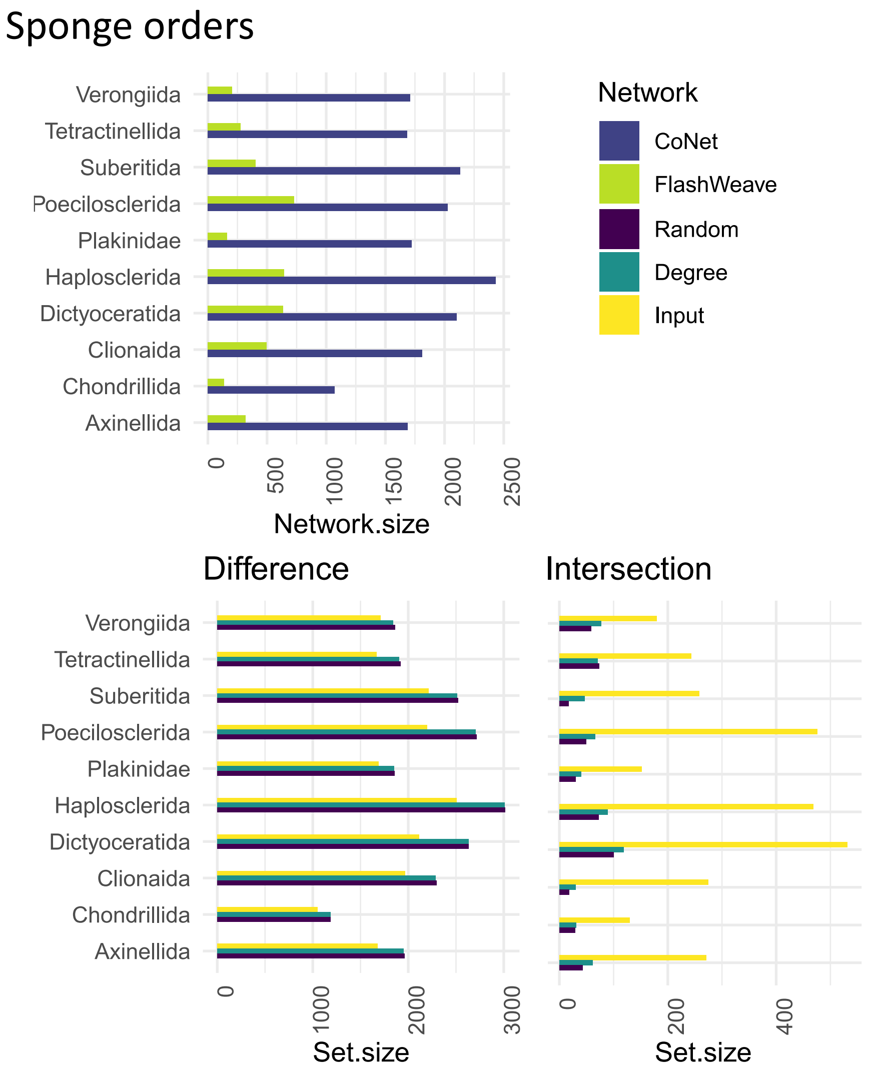
- Significant core network for edges in four or five networks
- Network clusters correspond to enterotypes





Anuran in action: tool comparison

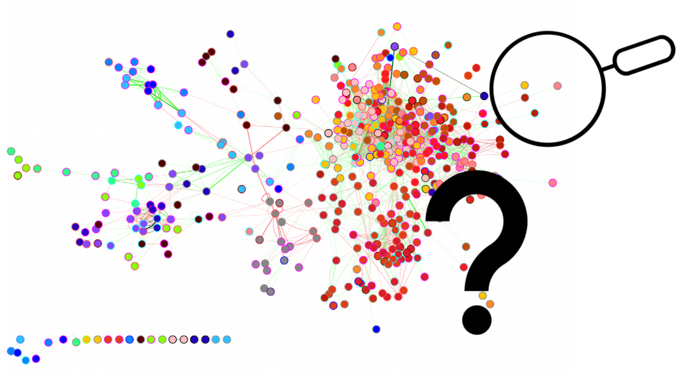
- Sponge-order specific networks constructed with CoNet and FlashWeave
- CoNet networks are systematically larger
- Tool-specific network intersection is highly significant
- Tools pick up the same associations, but CoNet reports many additional edges (indirect edges)



Microbial network analysis tools

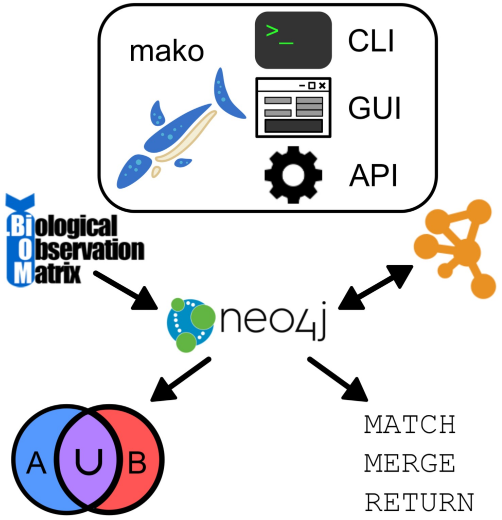
Querying microbial networks

- A database to query microbial networks would be useful
- Existing one by Hu et al.: http://www.microbialnet.org/mind_home.html
- There are dozens of tools to construct microbial networks
- Each tool comes with a range of settings
- Need for a flexible & local solution

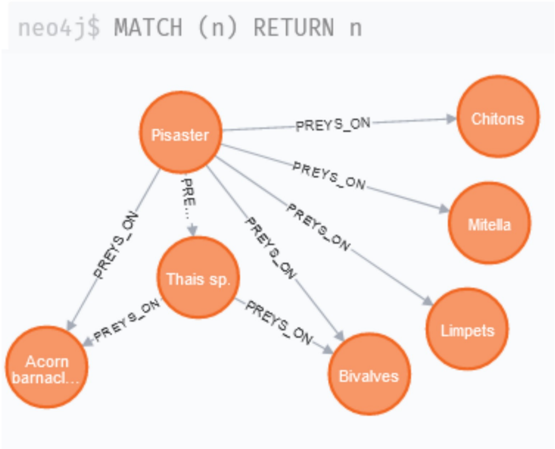


Mako – key ideas

- Use neo4j network database and CYPHER network query language
- Avoid a centralised database; create network database on the fly from user networks instead



Neo4j Browser



```
neo4j$ MATCH (n) RETURN n
```

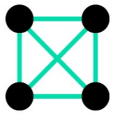
Sam Röttjers

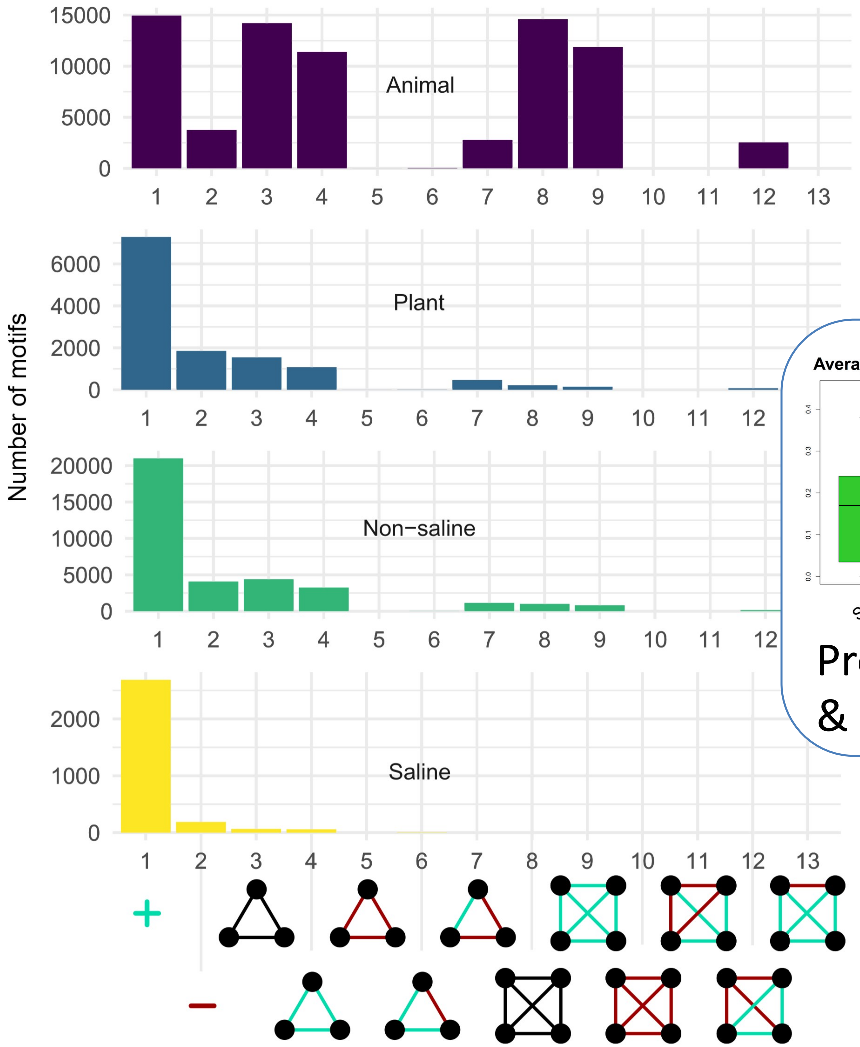
Röttjers & Faust "Fast and flexible analysis of linked microbiome data with mako" (2022) Nature Methods 19, 51-54.

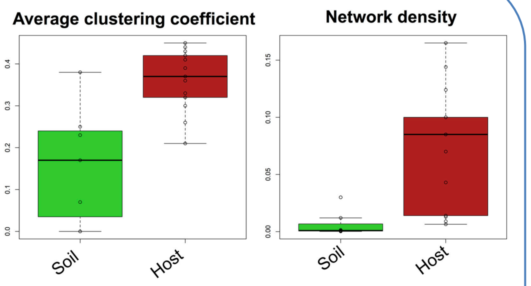
Microbial network analysis tools

Mako in action

- Networks constructed for 60 microbiome data sets taken from QIITA (tool: FlashWeave)
- Networks queried for specific motifs
- More positive than negative edges found
- Animal-derived microbial data sets enriched in positive 4-node clique







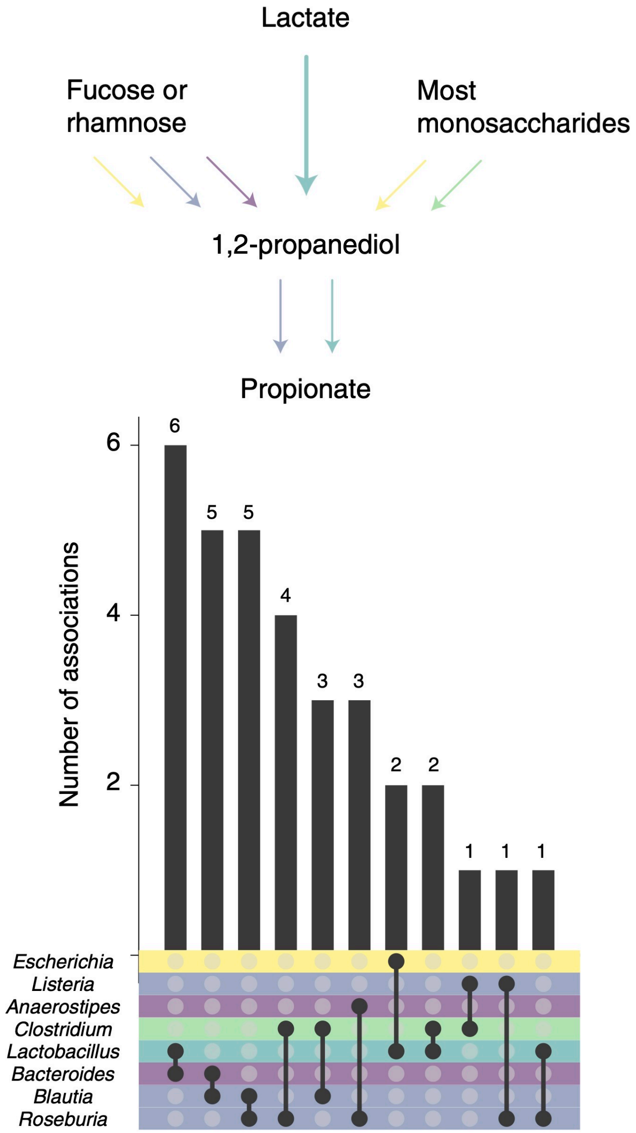
Previous findings on Qiita & EMP data with CoNet

Gonzalez et al. (2018) “Qiita: rapid, web-enabled microbiome meta-analysis” Nature Methods 15, 796–798.
Faust et al. (2015) “Cross-biome comparison of microbial association networks” Frontiers in Microbiology 6, 1200.

Microbial network analysis tools

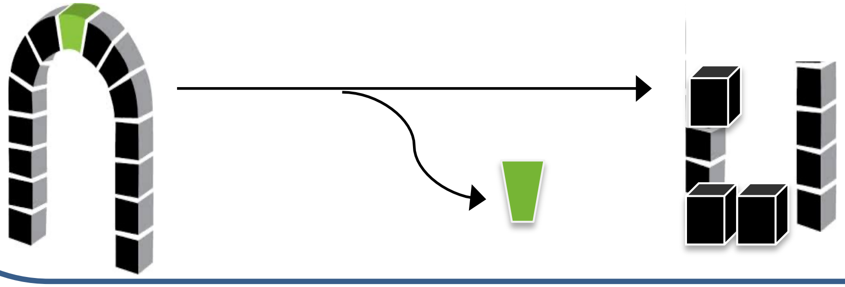
Mako in action

- Task: find associations between groups of gut bacteria able to synthesize propionate
- Mako applied to collection of 60 microbial networks constructed from QIITA to screen for associations
- Most frequent: Bacteroides and Lactobacillus

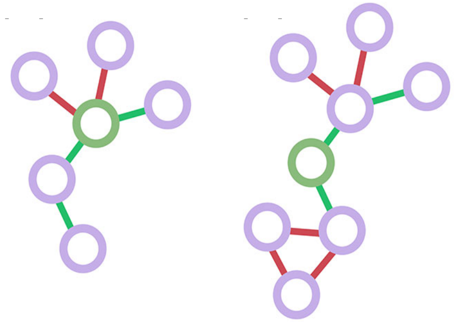


What about network properties?

Paine on **keystone species**: “These individual populations are the keystone of the community’s structure, and the integrity of the community and its unaltered persistence through time, that is, stability are determined by their activities and abundances”



Can we identify keystone species with networks?



Hub taxon: high degree

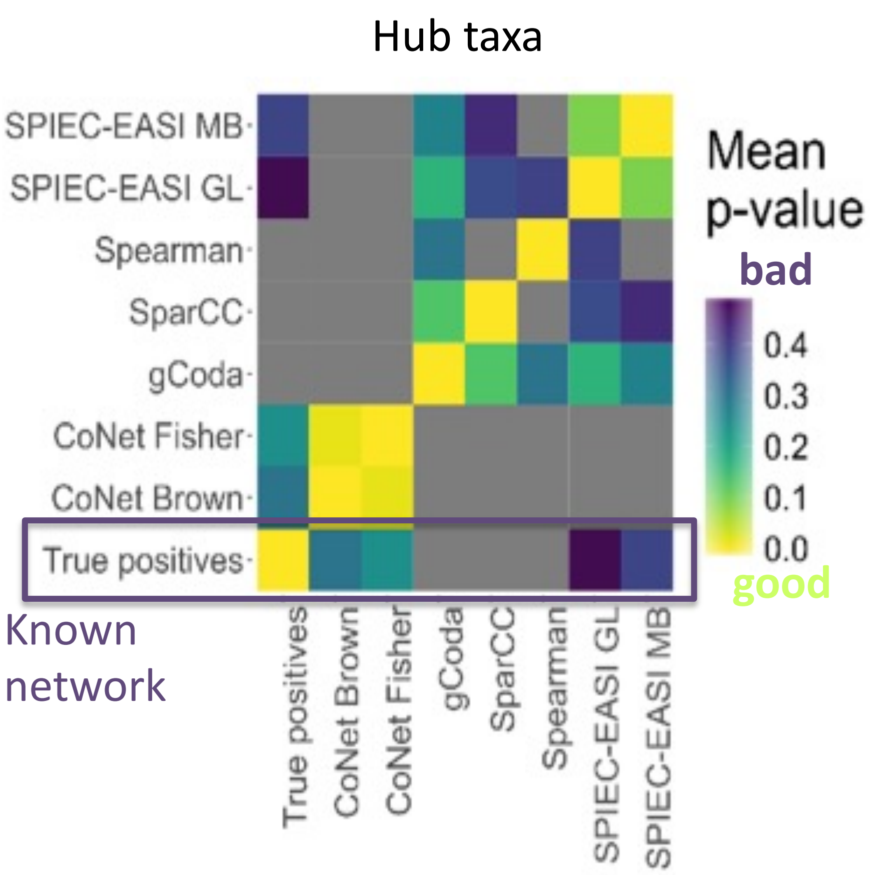
Connector taxon: high betweenness

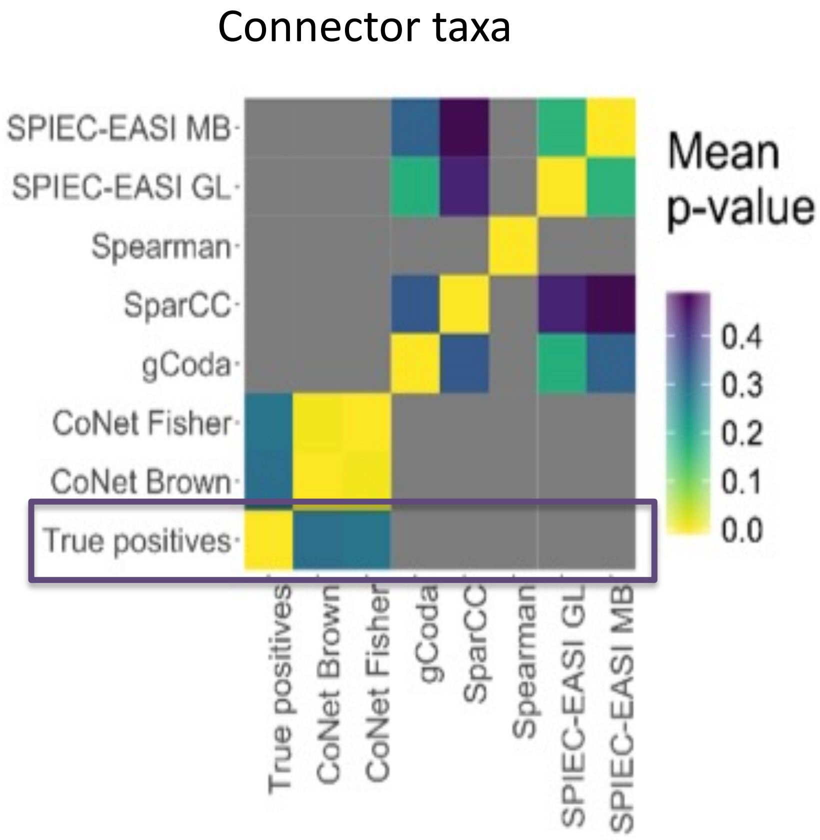
R.T. Paine (1969). A Note on Trophic Complexity and Community Stability. *The American Naturalist* 103, 91-93.

Microbial network properties

Can tools predict hub taxa?

- Not well (in synthetic data)



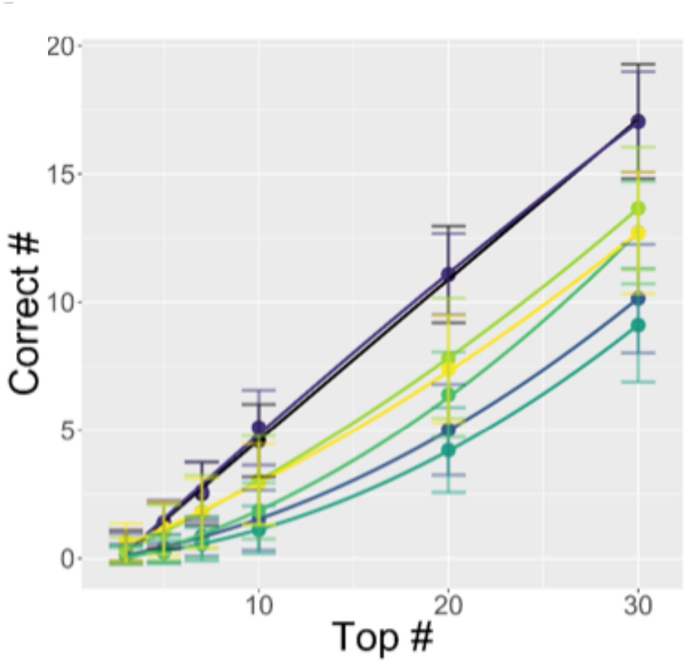


Röttjers & Faust (2018) FEMS Microbiology Reviews 42, 761-780.

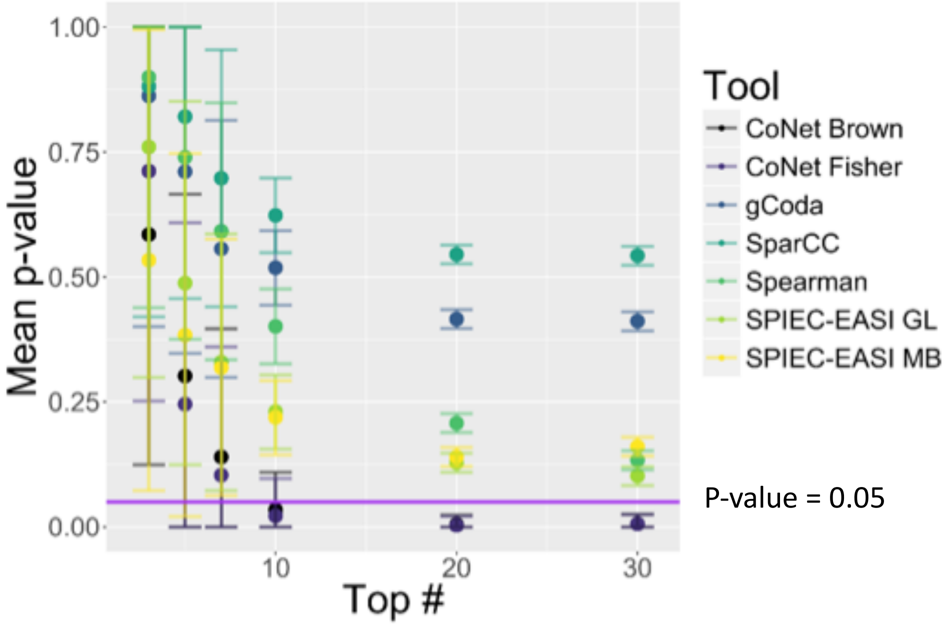
Can tools predict hub taxa?

- When a larger number of predicted top hub nodes is considered, CoNet significantly enriches for correct hubs (indirect edges may help with this – they are not always bad)

Matching fraction of hub nodes



P-value of matching fraction

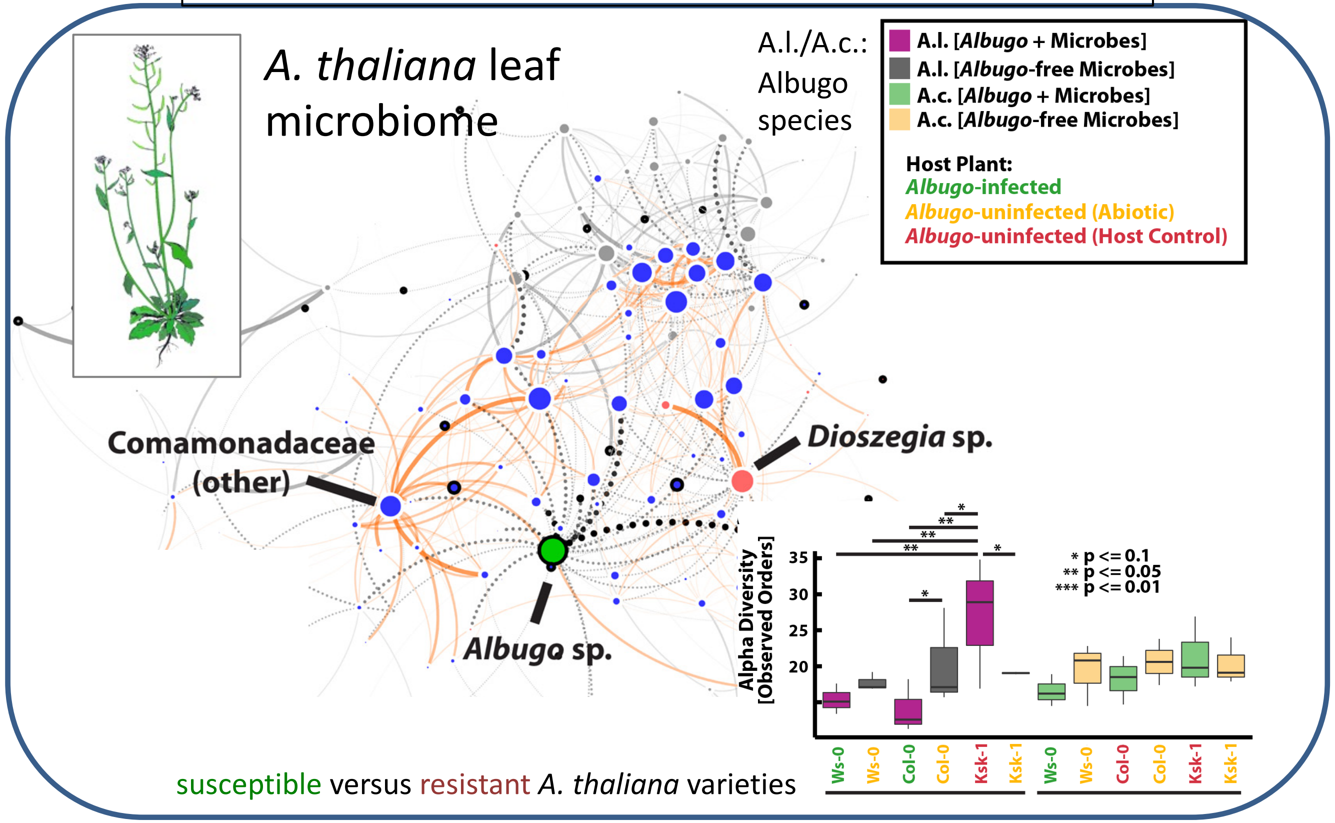


P-value = 0.05

Are correctly identified hub taxa keystones?

- Open question – perhaps in some cases

Example of a validated hub species (a parasite):



Agler et al. (2012) “Microbial Hub Taxa Link Host and Abiotic Factors to Plant Microbiome Variation” PLoS Biology 14 (1) e1002352.
Röttjers & Faust (2018) “Can we predict keystones?” (Comment) Nature Reviews Microbiology 17, 193

Summary: what to do when you want to build a microbial network

- **Data preprocessing**: split samples into groups if they are strongly heterogeneous -> filter and sum rare taxa -> normalise abundances -> if appropriate and available multiply with total counts
- **Network inference**: take metadata into account and reduce indirect edges (currently only FlashWeave supports both)
- **Analyse the hairball**: map external data onto nodes if available, check for enrichment of particular taxa or functions in clusters, compare with other networks and known interactions, experimentally validate interaction candidates

Microbial network construction and analysis tutorials

- SparCC, MENA, LSA, CoNet and SPIEC-EASI: **http://msysbiology.com/microbialnetworks/**
- FlashWeave and network analysis (manta & anuran): **https://rutjers.science/teaching/**
- Mako tutorials: **https://ramellose.github.io/mako_docs/**

Next steps in microbial network analysis

- Network annotation: link taxa to known physiological properties such as pH optima
  - Tool development ongoing (microbetag)
- Experimentally resolve microbial interaction networks to benchmark inference tools on biological data
- Explore whether microbial network properties reflect ecosystem properties

Acknowledgements

KU LEUVEN

Lab of Microbial Systems Biology
Sam Röttjers (now @BiomeMakers)

Raes lab
Jeroen Raes
Gipsi Lima Mendez (now @Namur)

河海大学 HOHAI UNIVERSITY

Xinyi Cao

Ecology of Marine Microbes

Ramiro Logares
Ina Deutschmann

VUB

Applied Physics
Sophie de Buyl

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

Curtis Huttenhower
Fah Sathirapongsasuti

Turun yliopisto
University of Turku

Open Research Labs
Leo Lahti

UNIVERSITY OF VIENNA

Stefanie Widder
Franziska Bauchinger

INRA
SCIENCE & IMPACT

Béatrice Laroche

ULB

Machine Learning Group
Gianluca Bontempi
Jean-Sébastian Lerat

Theoretical Chronobiology
Didier Gonze

Duke UNIVERSITY

Department of Biology
Alex Washburne

fwo

erc
European Research Council
Established by the European Commission

Acknowledgements

Integrated multi-omic analyses of microbial communities
23-29 April 2022
Luxembourg

- EMBO Practical Course
  funding

- ELIXIR Luxembourg
  support

- HPC team, Moodle team, Comms team, Finance team of uni.lu | LCSB
  organization

- Slack
  extended free workspace license

Thank you

Supplement

Tackling compositionality

- The ratio trick: since total abundance T cancels out in a ratio, the ratio removes dependency on total abundance in a composition

$$\frac{X_i}{X_j} = \frac{\frac{X_i}{T}}{\frac{X_j}{T}}$$

Xi, Xj: abundances of taxa i and j

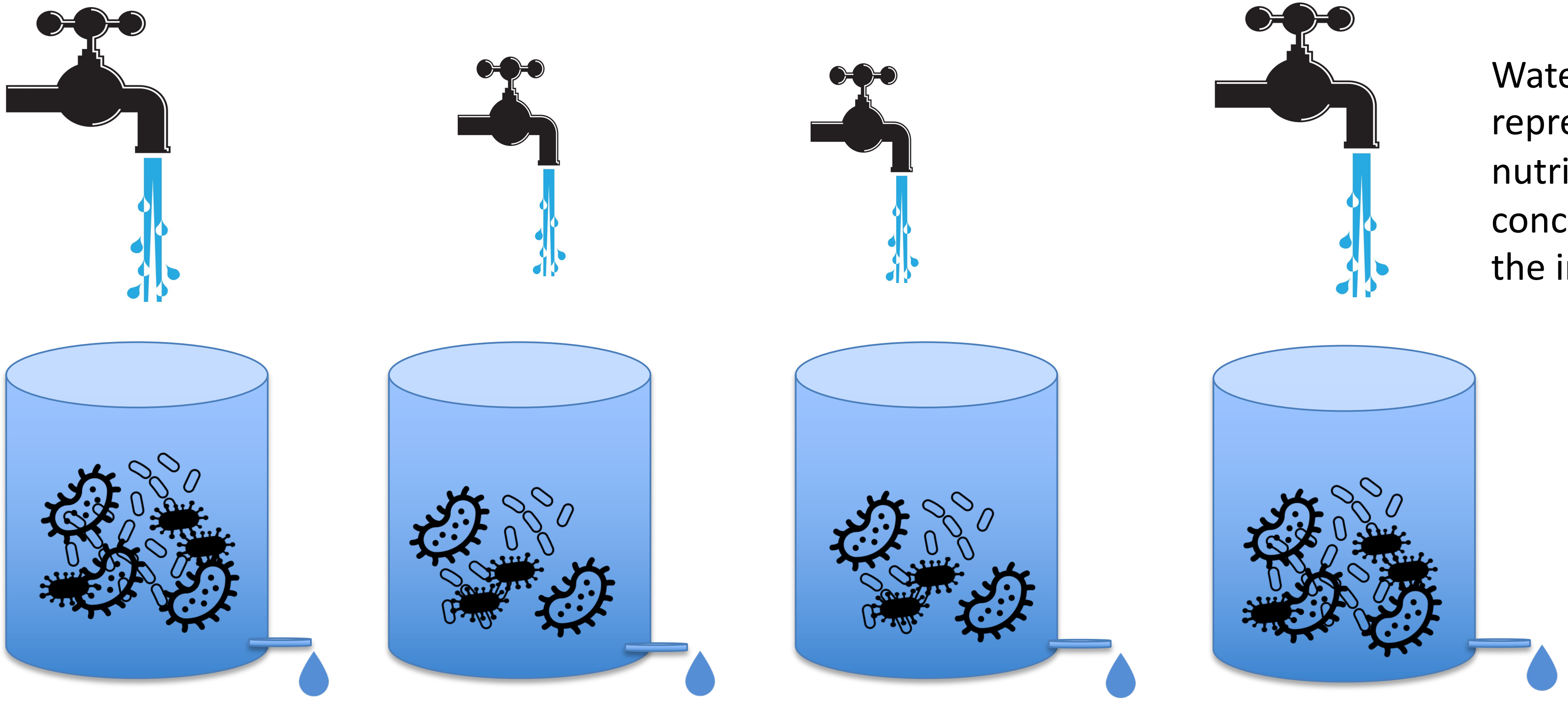
- CLR transform (introduces neg values):

$$clr(X_i) = \log(\frac{X_i}{\left(\prod_j^n X_j\right)^{1/n}})$$

Divide abundance of taxon i by the geometric mean of the abundances in its sample and take the log

Supplement

Relative vs absolute abundances in network inference



Water tap size represents nutrient concentration in the inflow

Chemostats with different nutrient concentrations in the inflow. In case density differences are solely determined through nutrient concentration, total counts are a confounder to be removed.

Supplement

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

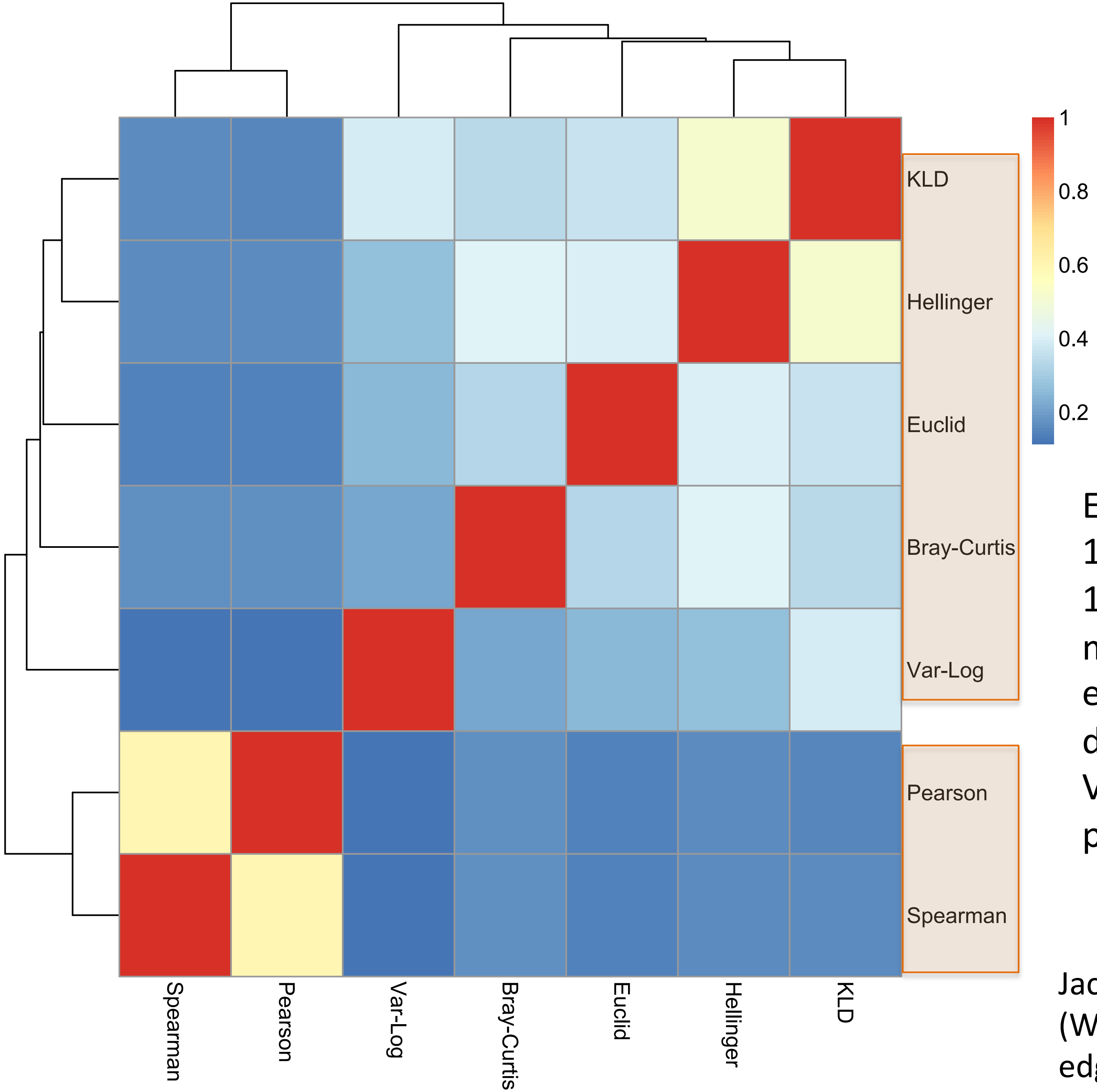
Pearson $$d(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

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. This also means that correlations are scale-invariant, so do not change when multiplied with a constant.

Supplement

Comparison of measures



Experiment: Select 1,000 top-ranked and 1,000 bottom-ranked measure-specific edges in Houston data subset of HMP V35 mothur-processed 16S data

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

Supplement

Fisher’s method of p-value merging

$$X^2_{2k} \sim -2\sum_{i=1}^{k}\ln(p_i)$$

k: number of association measures
p_i: p-value of the *i*th association measure
X^2_{2k}: Value is chi-square distributed with 2k degrees of freedom

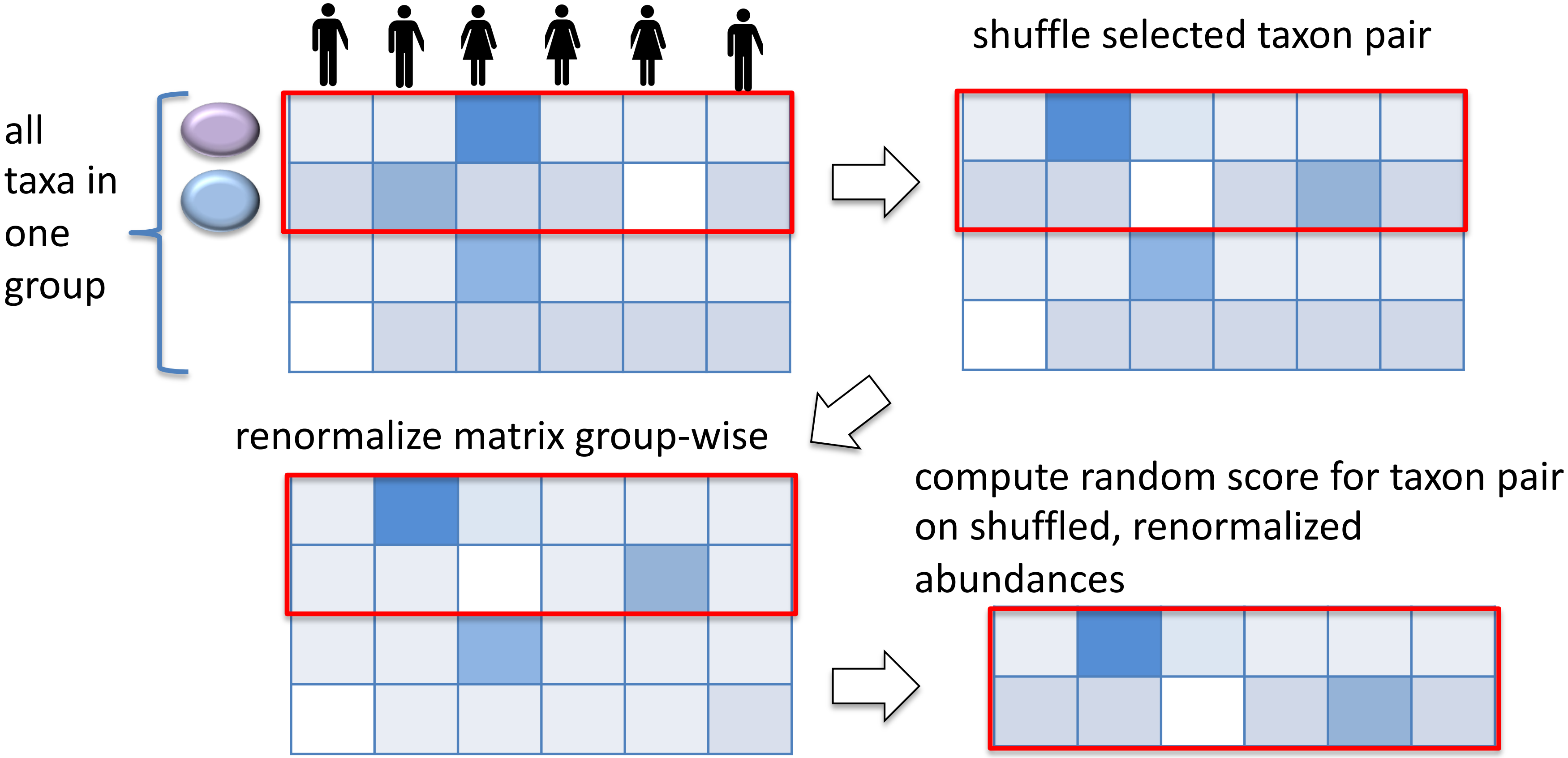
The resulting p-value is the p-value of the Chi-square value.

Fisher’s method is biased by correlated association measures. This bias is taken out by Brown’s p-value merging method.

CoNet: ReBoot

Supplement

permutation with renormalization (**ReBoot**)



all taxa in one group

shuffle selected taxon pair

renormalize matrix group-wise

compute random score for taxon pair on shuffled, renormalized abundances

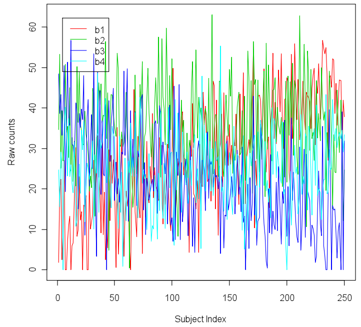
Fah Sathirapongsasuti

Supplement

CoNet: ReBoot II

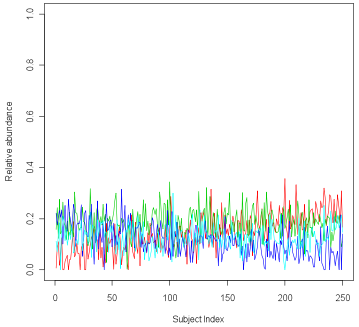
- Permutation test: removes correlation, but also any bias due to compositionality
- Permutation with **renormalization**: shifts null distribution

raw data



true anti-correlation between b1 and b3

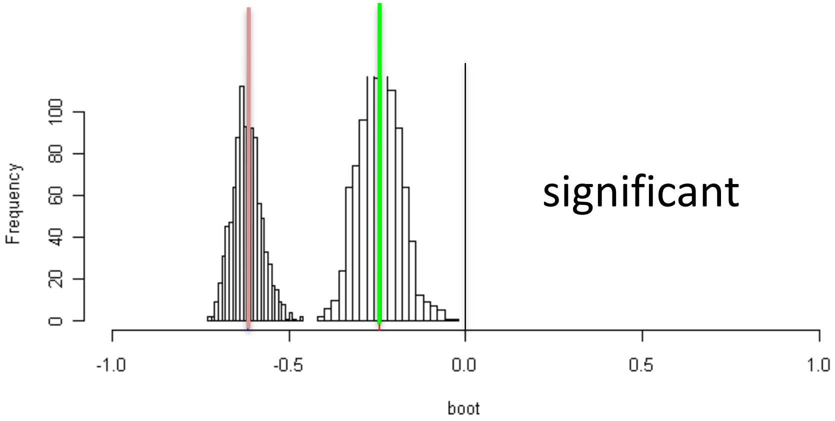
normalized data



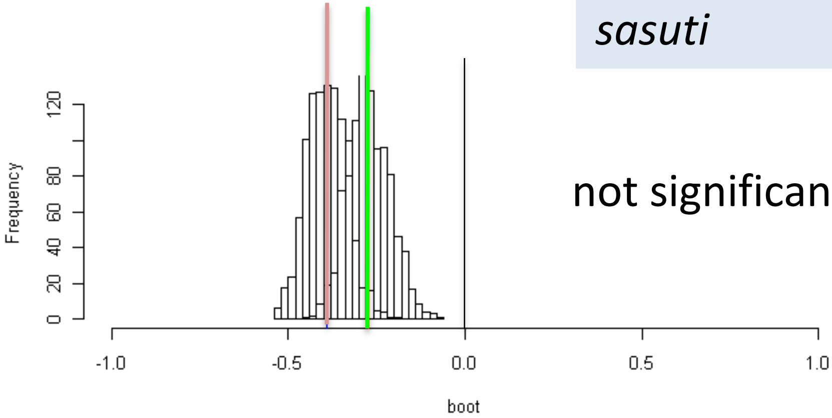
spurious correlation between b2 and b4 introduced by normalization

Fah Sathirapong-sasuti

▬ bootstrap distribution mean
▬ renormalized permutation distribution mean



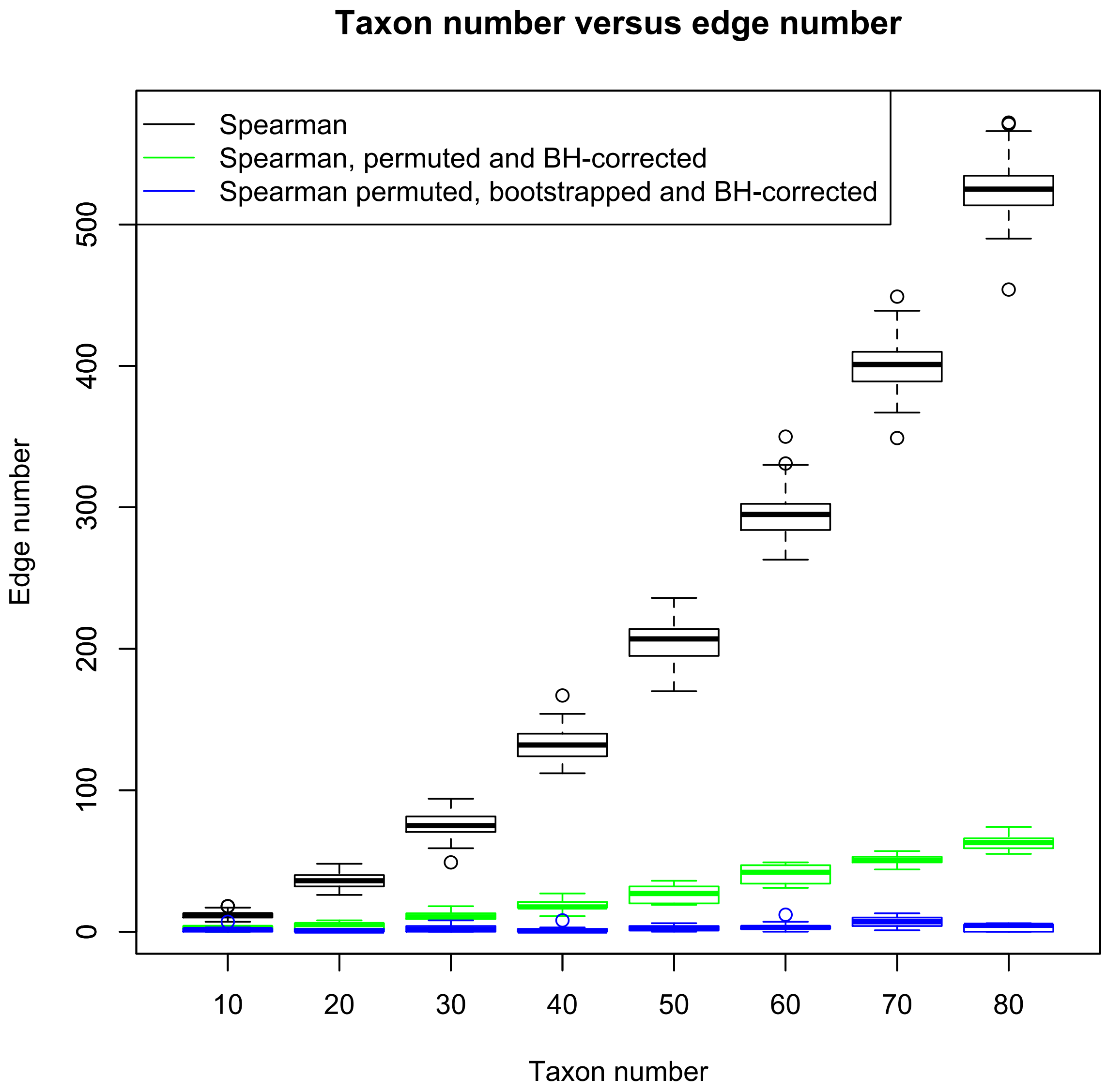
significant



not significant

Supplement

CoNet's assessment of significance reduces number of false positives



(matrix not normalized, permutation carried out without renormalization)

Simulations with Dirichlet-Multinomial

Simulation parameters:
samples = 50
p_i=1/S (max. even)
sequencing depth = 1000
θ = 0.002
repetitions = 100 (black)
repetitions = 10 (blue, green)
permutations: 100
bootstraps: 100
BH = Benjamini-Hochberg

Supplement

SparCC

- basic idea: use the variance of log ratios (a distance measure robust to compositionality bias, Aitchison 2003)

$$D(x_i, x_j) = \mathrm{var}\left(\log\left(\frac{x_i}{x_j}\right)\right)$$

$x_{i,}$ x_j are taxon abundance vectors

- the variance of log-ratios is not scaled, i.e. its maximum value is unknown
- starting from the variance of log ratios, an approximation is developed to estimate correlations robustly

$$D(x_i, x_j) = \omega_i^{\,2} - \omega_j^{\,2} - 2\rho_{ij}\omega_i\omega_j$$

where ω is the variance of the (log-transformed) abundance vector of taxon i and ρ the covariance between taxa i and j

- SparCC estimates covariance ρ for all taxon pairs, assuming that most pairs are only weakly correlated

Friedman & Alm (2012) "Inferring Correlation Networks from Genomic Survey Data." PLoS Comp Bio 8 (9), e1002687.
Aitchison (2003) "A concise guide to compositional data analysis" In: 2nd Compositional Data Analysis Workshop, Girona, Italy.

Supplement

SparCC Parameters

Iterations

- SparCC fits a Dirichlet distribution to the counts and samples from this distribution to estimate counts
- final correlation is reported as the median over several sampling rounds

P-values

- Bootstraps generated by sampling with replacement
- P-values computed from bootstrap distribution as the proportion of bootstrapped correlations that are at least as large as the original correlation value

Implementations

- **https://bitbucket.org/yonatanf/sparcc** (original in Python)
- Part of the SPIEC-EASI R package

Supplement

Discrete version of GLV: Ricker model

$$x_i(t+\delta t)=\eta_i(t)x_i(t)\exp(\delta t\sum_j a_{ij}(x_j(t)-\langle x_j\rangle))$$

δt: discrete time step

X_i(t): abundance of target species *i* at time point *t*

<x_j>: steady state abundance of species *j* (carrying capacity)

η_i(t): log-normal noise

a_{ij}: interaction coefficient between taxa i and j

For η_i(t) = 1 (no noise) and δt -> 0, Ricker model reduces to generalized Lotka-Volterra in continuous form.

Fisher and Mehta (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression. *PLoS one* 9**,** e102451.

Supplement

LIMITS - principle

- LIMITS: **L**earning **I**nteractions from **MI**crobial **T**ime **S**eries
- Principle: select interaction coefficients such that change between consecutive time points in one species is well predicted from the other species

$$y_1 = \log x_i(2) - \log x_i(1) = \sum_j a_{ij}(x_j(1) - \langle x_j \rangle)$$

$$y_2 = \log x_i(3) - \log x_i(2) = \sum_j a_{ij}(x_j(2) - \langle x_j \rangle)$$

$$y_3 = \log x_i(4) - \log x_i(3) = \sum_j a_{ij}(x_j(3) - \langle x_j \rangle)$$

↓

$$y_t = \log x_i(t+1) - \log x_i(t) = \sum_j a_{ij}(x_j(t) - \langle x_j \rangle)$$

⇨

Vector of log abundance differences for species i for all time point pairs (t+1,t)

$$a_{i*} = yX^{-1}$$

interaction matrix row (interactions between species i and selected predictor species)

Pseudo-inverse of species abundance matrix of selected predictor species

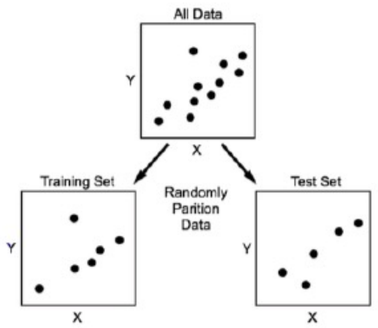
$\eta_i(t)$: 1, δt: 1 (no noise, smallest possible time step)

Fisher and Mehta (2014). Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression. *PLoS one* 9, e102451.

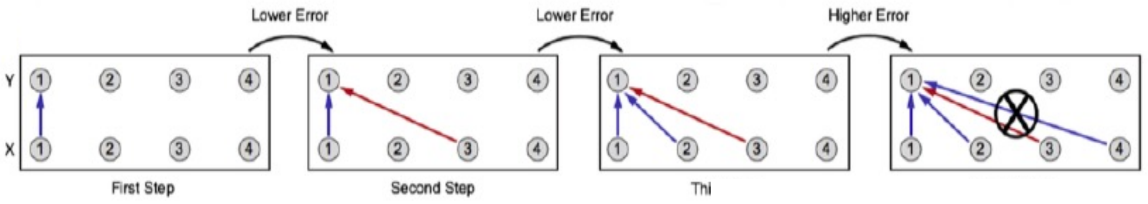
Supplement

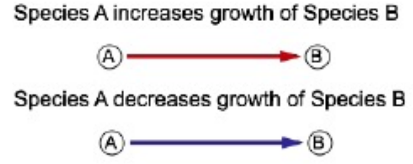
LIMITS – workflow

- Data is split into **training and test set**. Inference is done on training set, prediction error is calculated on test set.



- Interaction matrix inference: For each species i, select the set of predictor species j that minimise the error on the test set via **step-wise forward regression**



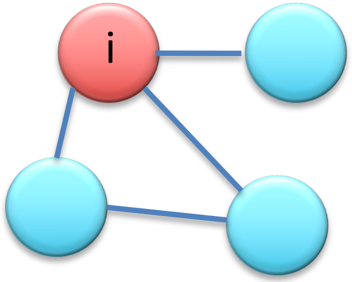

Species A increases growth of Species B
A
B
Species A decreases growth of Species B
A
B

- Repeat data splitting and interaction matrix inference a number of times and report the median (**bootstrap**)

Error measurements:

- Difference between y and X in the test set, with interaction coefficients inferred from the training set (reported by LIMITs)
- Difference between observed time series and time series predicted with Ricker (simulation with inferred interaction coefficients)
- Difference between known and inferred interaction matrices

Examples of network properties



k=3

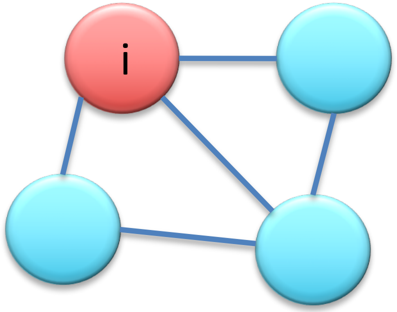
n=1

C_i = 1/3

E=4

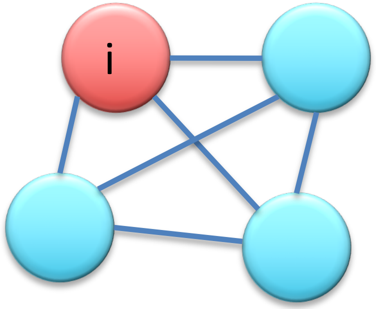
S=4

D = 2*4/(4*3)=2/3



C_i = 2/3

D = 5/6



C_i = 1

D = 1

fully connected clique

Clustering coefficient of node i

$$C_i = \frac{2 \cdot n}{k_i \cdot (k_i - 1)}$$

k = number of neighbors of node *i*

n = number of edges between the neighbors of node *i*

Average clustering coefficient

$$C = \frac{1}{S} \cdot \sum_{i=1}^{S} C_i$$

Network density (connectance)

$$D = \frac{2 \cdot E}{S \cdot (S - 1)}$$

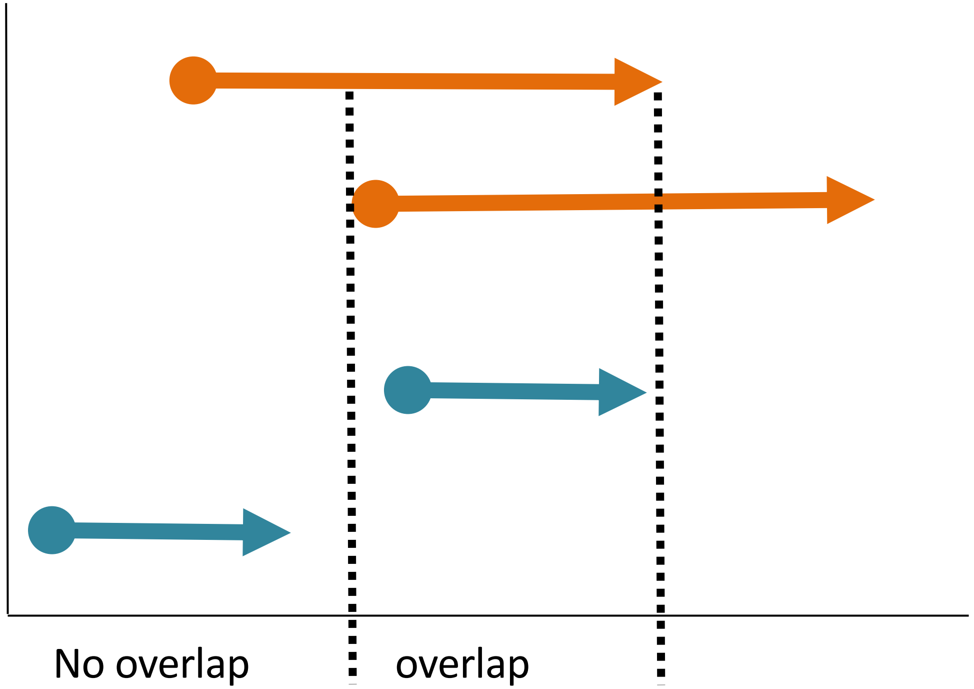
E = number of edges in the network

S = number of taxa in the matrix

Supplement

Indirect edge removal: Overlap

Overlap uses the start and length of co-occurrence in time



No overlap
→ not indirect

overlap
→ indirect

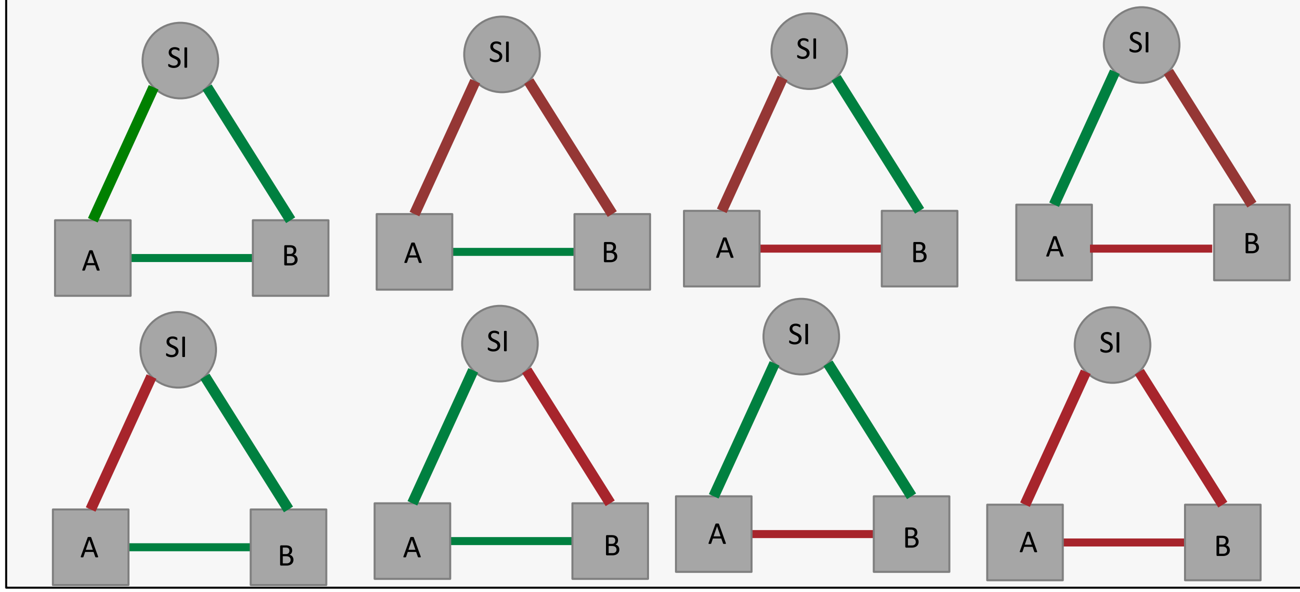
Ina Deutschmann

Supplement

Indirect edge removal: sign patterns

- ■ Taxon
- ● Env. factor
- 🟩 Positive link
- 🟥 Negative link

8 possible sign patterns for two taxa and one environmental factor in a triplet



Sign patterns indicating an **indirect** taxon relationship

Sign patterns indicating a **direct** taxon relationship

Supplement

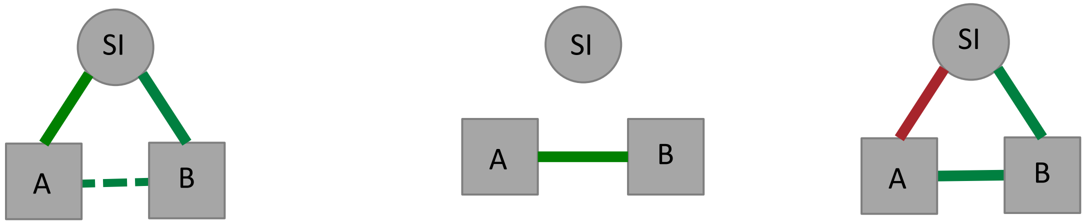
Indirect edge removal: interaction information

- Interaction information indicates whether a triplet contains an indirect edge
- It is an assumption that the indirect edge is the taxon-taxon edge (this is a good assumption for environmental factors that cannot be quickly influenced by taxa, such as temperature)

negative: redundancy $CI(X,Y|Z) < MI(X,Y)$

zero: no interaction $CI(X,Y|Z) = MI(X,Y)$

positive: synergy $CI(X,Y|Z) > MI(X,Y)$



■ Taxon
● Env. factor
🟩 Positive link
🟥 Negative link

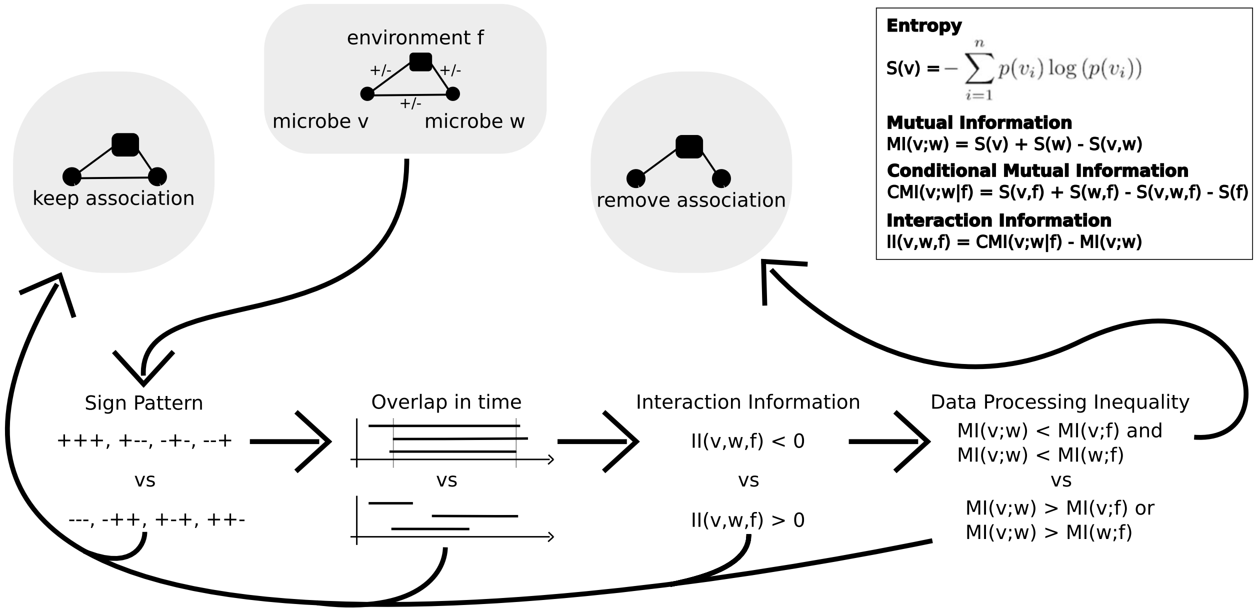
CI = conditional mutual information
MI = mutual information
II = interaction information

$$\boxed{II = CI(X,Y \mid Z) - MI(X,Y)}$$

Supplement

EnDED: shrink the hairball

- Problem: edges in microbial networks are often driven by environmental factors
- EnDED combines several methods to remove indirect edges



environment f

+/-

+/-

+/-

microbe v

microbe w

keep association

remove association

Entropy

$S(v) = -\sum_{i=1}^{n} p(v_i) \log(p(v_i))$

Mutual Information

MI(v;w) = S(v) + S(w) - S(v,w)

Conditional Mutual Information

CMI(v;w|f) = S(v,f) + S(w,f) - S(v,w,f) - S(f)

Interaction Information

II(v,w,f) = CMI(v;w|f) - MI(v;w)

Sign Pattern

+++, +--, -+-, --+

vs

---, -++, +-+, ++-

Overlap in time

vs

Interaction Information

II(v,w,f) < 0

vs

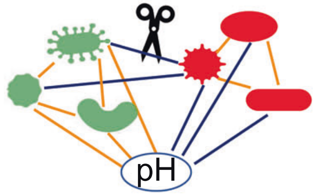
II(v,w,f) > 0

Data Processing Inequality

MI(v;w) < MI(v;f) and
MI(v;w) < MI(w;f)

vs

MI(v;w) > MI(v;f) or
MI(v;w) > MI(w;f)



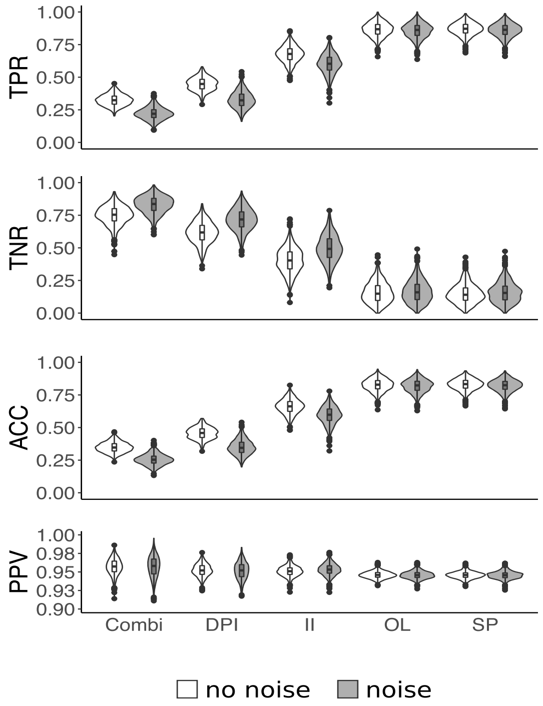
Ina Deutschmann

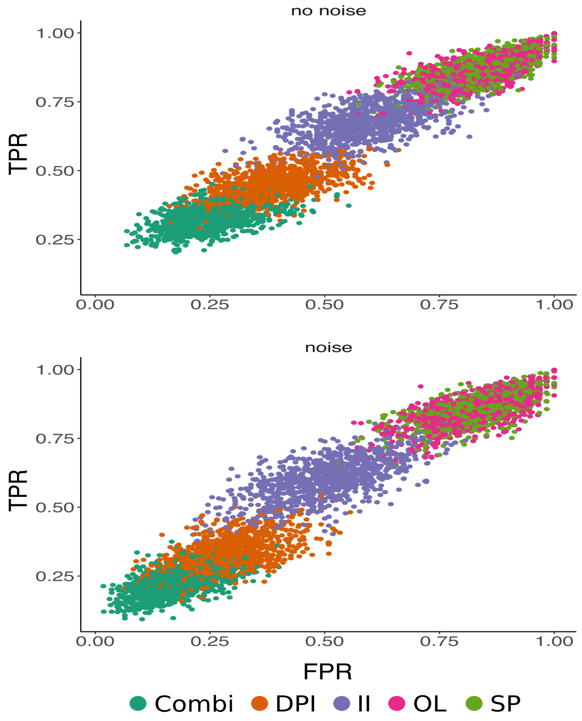
Deutschmann et al. (2021) "Disentangling environmental effects in microbial association networks" Microbiome 9, 232.

Supplement

EnDED performance

- Accuracy assessed on simulated data (extended Lotka Volterra model)
- Method combination: lowest accuracy but highest positive predictive value (removes fewer true edges at cost of keeping false ones)





DPI: data-processing inequality (edge with smallest mutual information in triplet is removed)

II: Interaction information (indicates redundancy in triplets)

OL: overlap (time series)

SP: Sign pattern

Combi: Combination

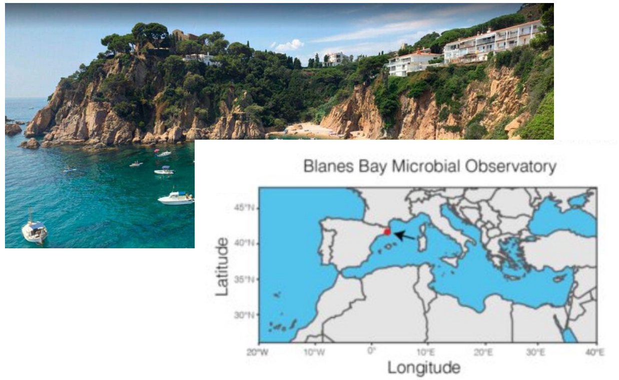
True positive: false edge correctly removed
False negative: false edge not removed

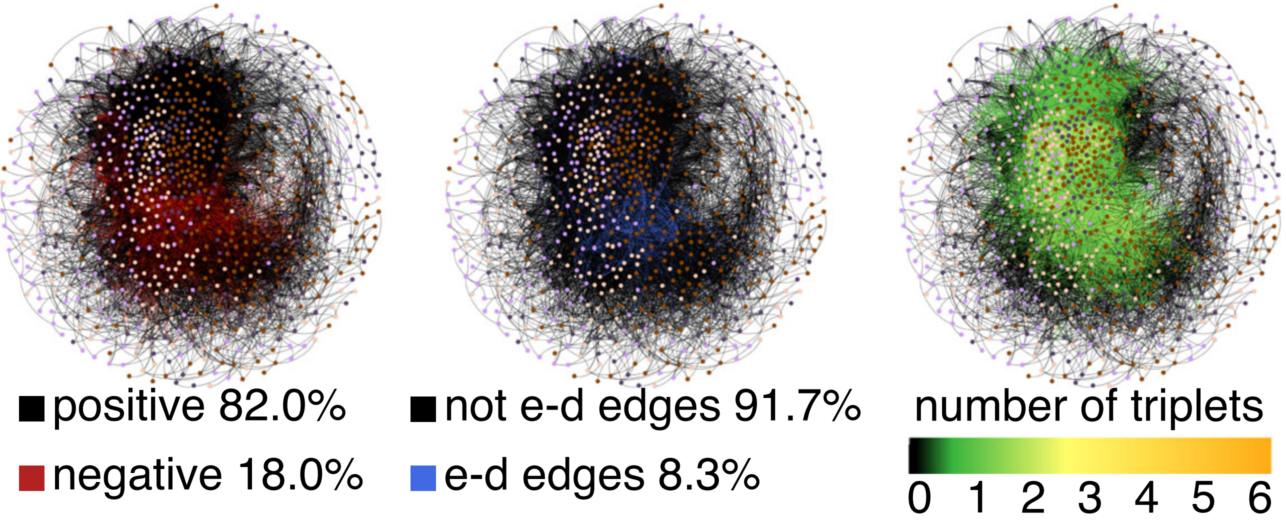
False positive: correct edge falsely removed
True negative: correct edge not removed

Supplement

EnDED in action

- Blanes Bay Microbial Observatory (BBMO) data: Mediterranean Sea sampled monthly from Jan 2004 to Dec 2013
- Environmental factors measured
- 18S & 16S V4 region sequenced
- Network constructed with eLSA
- EnDED (combi) removed 8.3% of the edges





29,820 edges: ■ positive 82.0% ■ not e-d edges 91.7%

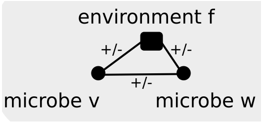
■ negative 18.0% ■ e-d edges 8.3%

754 nodes: ■ nB 37.0% ■ pB 22.4% ■ nE 20.7% ■ pE 19.9%

Taxon size fraction: nano and pico, kingdom: B=Bacteria, E=Eukaryotes

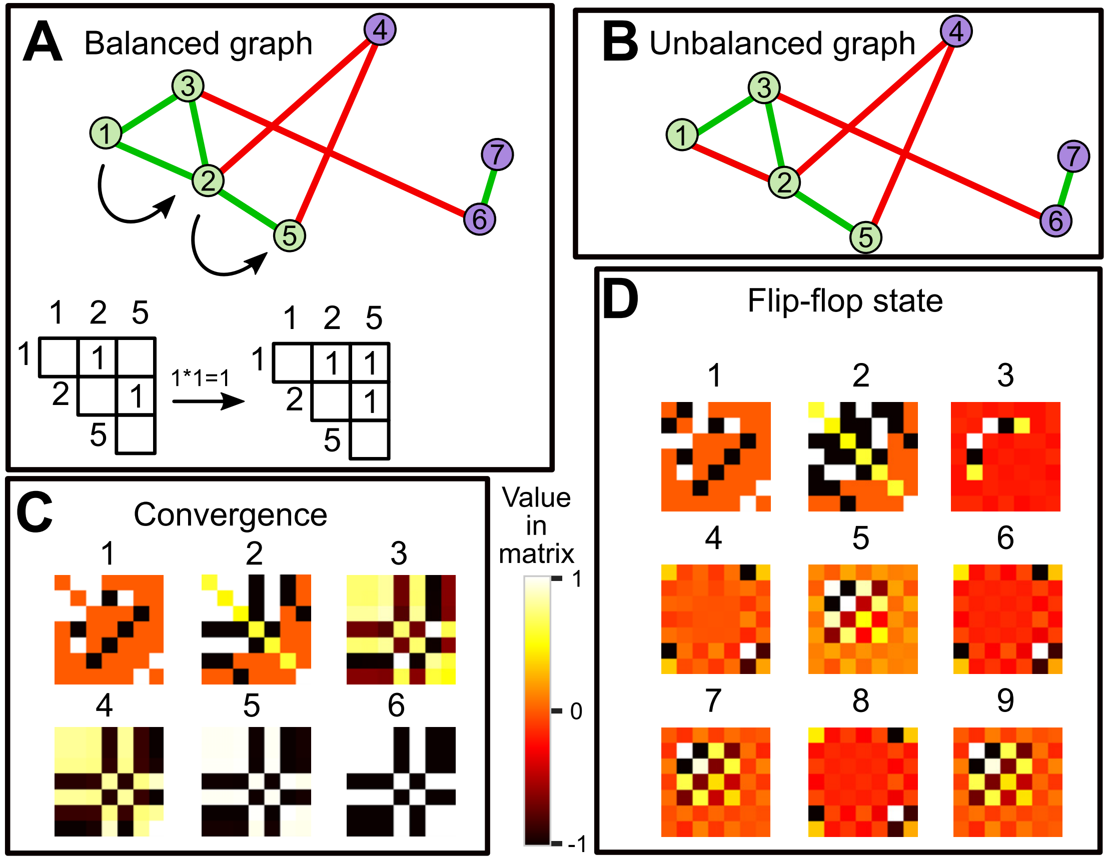
5 out of 29 known interactions lost

Triplet:



Supplement

Manta internals



B, D) Unbalanced graphs: no convergence (flip-flop state)

Trick: use sub-sets of the network to generate scoring matrix. For balanced subsets, follow procedure for balanced graphs. For unbalanced subsets, carry out only one iteration. When merging, only use nodes with signs consistent across most subsets.

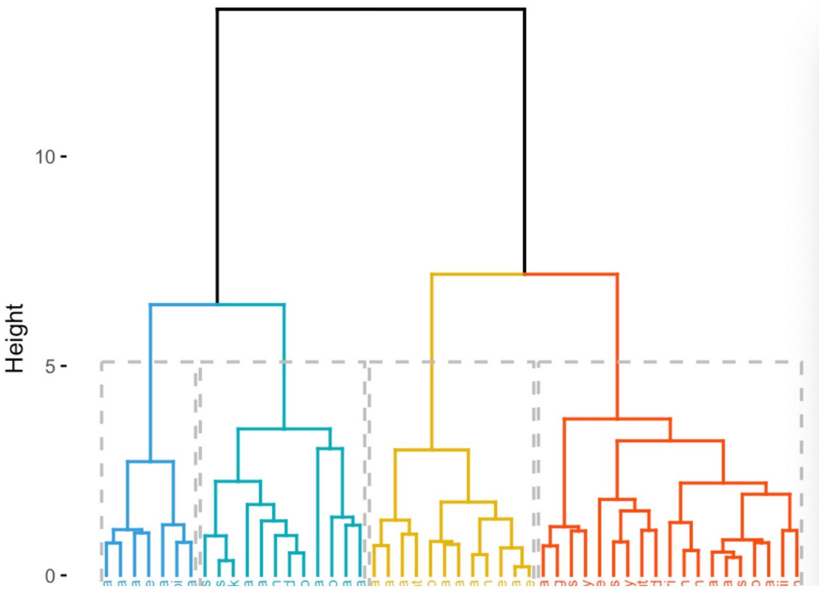
A, C) Balanced graphs: Multiplication and transformation (retaining signs) of weighted adjacency matrix MCL-style until convergence (scoring matrix)

=> Scoring matrix

Supplement

Manta internals

- Clusters derived from scoring matrix through agglomerative clustering with Euclidean distances
- Optimal cluster number determined with sparsity score S



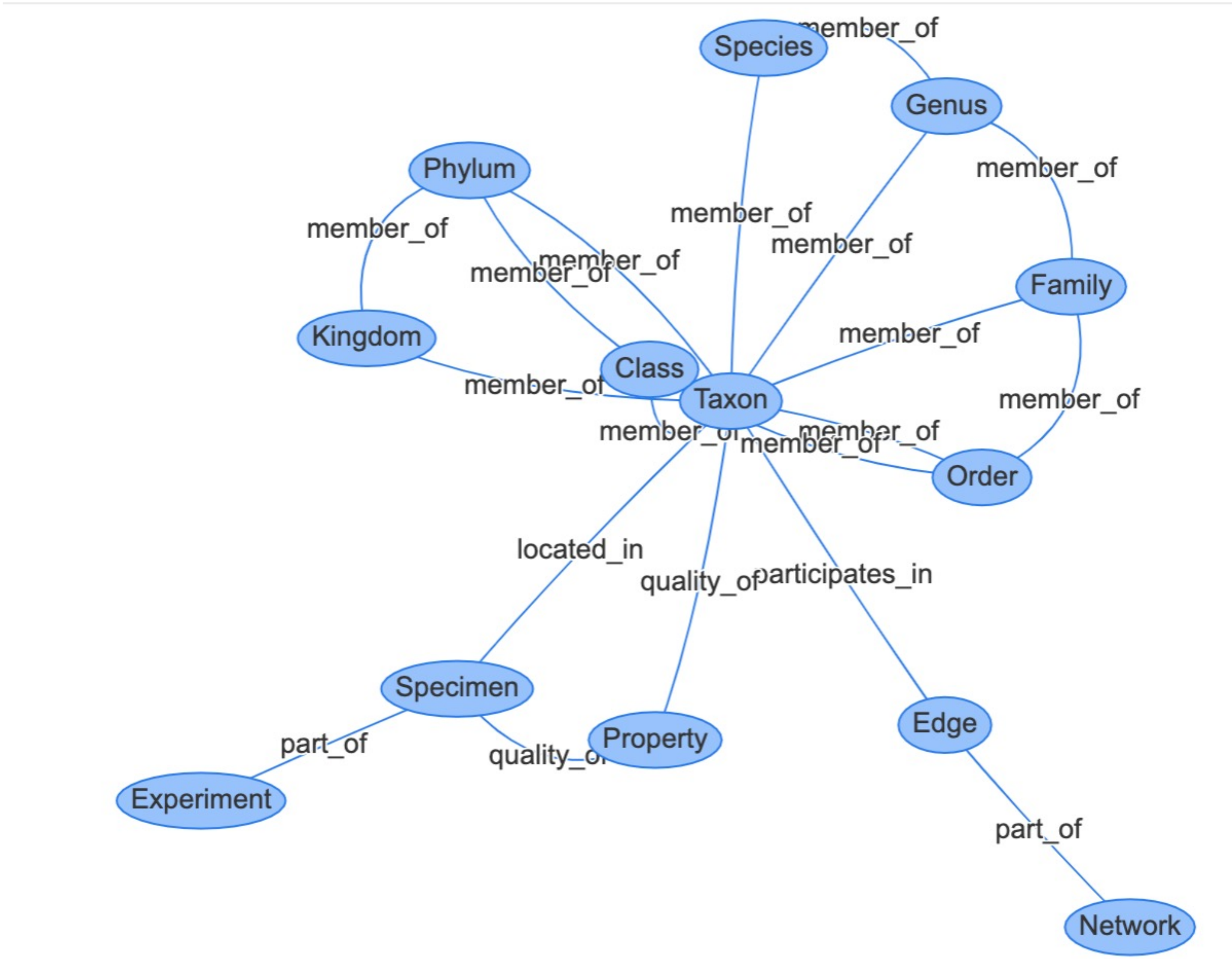
S = 1/E*(number of negative edges not in clusters + number of positive edges in clusters – number of negative edges in clusters – number of positive edges not in clusters)

E: total edge number

S ranges from -1 to 1, with -1 being the worst sore

Supplement

Mako’s data scheme



Supplement

Manta internals

- Weak cluster assignments:
  - Find nodes that fluctuate strongly in flip-flop iterations: oscillators (some diagonal values of scoring matrix)
  - Compute weight of shortest path from each node to closest oscillator
  - Negative weight or weight below user-defined threshold: node has a weak cluster assignment
- Treatment of small clusters: clusters in size below threshold are removed from scoring matrix and cluster membership of their nodes is assigned based on average shortest path weight to cluster members