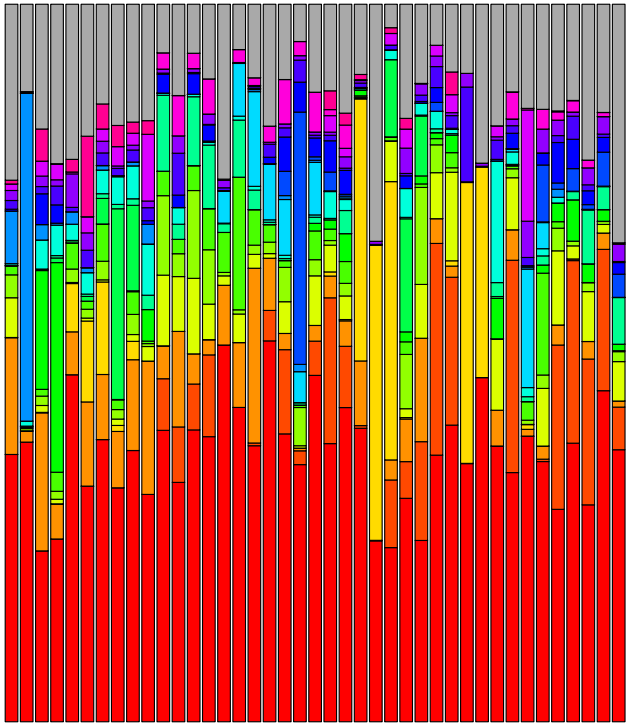
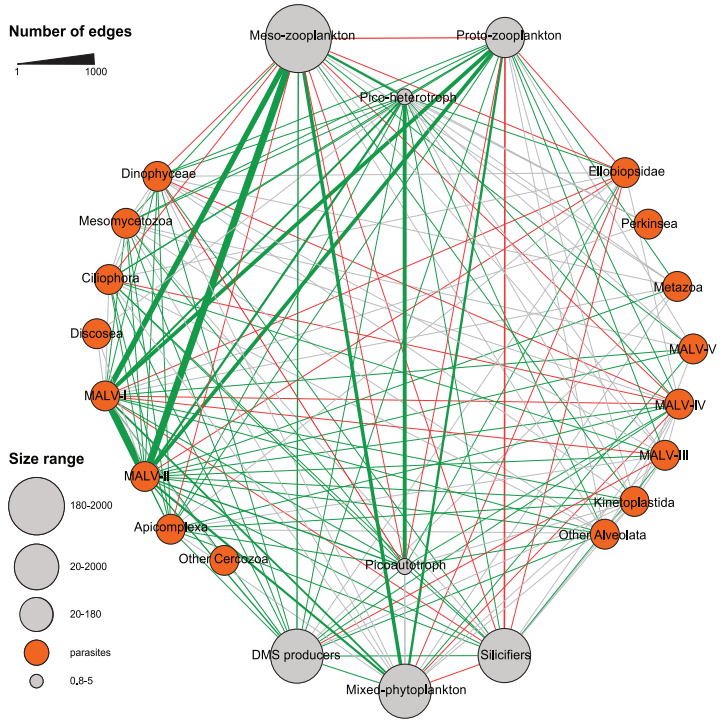
Karoline Faust
Lab of Microbial Systems Biology
KU Leuven

13th December 2019
Workshop II: Beyond Metagenomes

Microbial network inference in depth: challenges, tools and network comparison





Challenges

- What are the challenges of microbial network inference?

Problem 1: Varying sequencing depth

Shallowly sequenced sample

Deeply sequenced sample

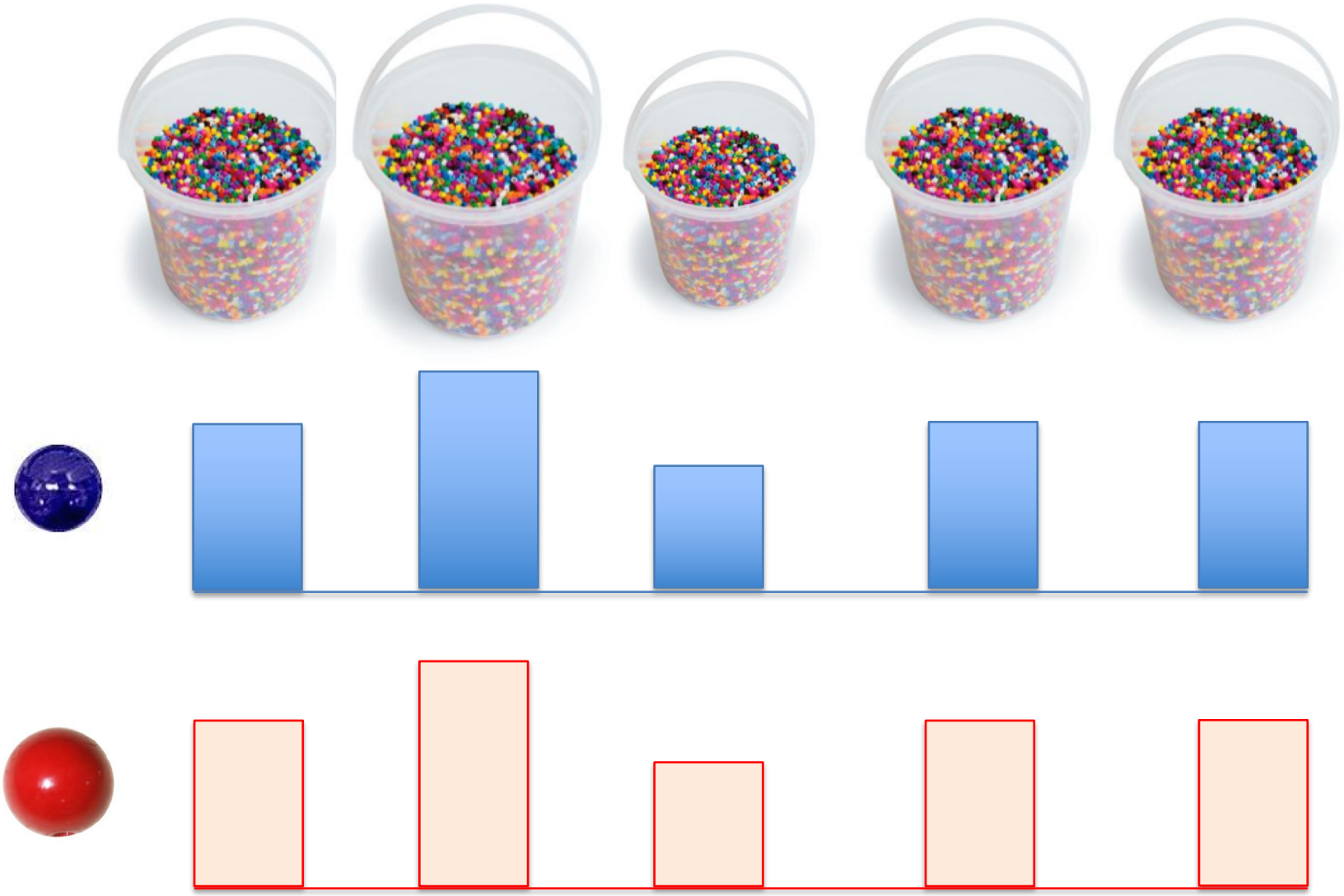
Genus 1

Genus 2

Genus 3

Genus 4

Varying sequencing depth leads to spurious correlations

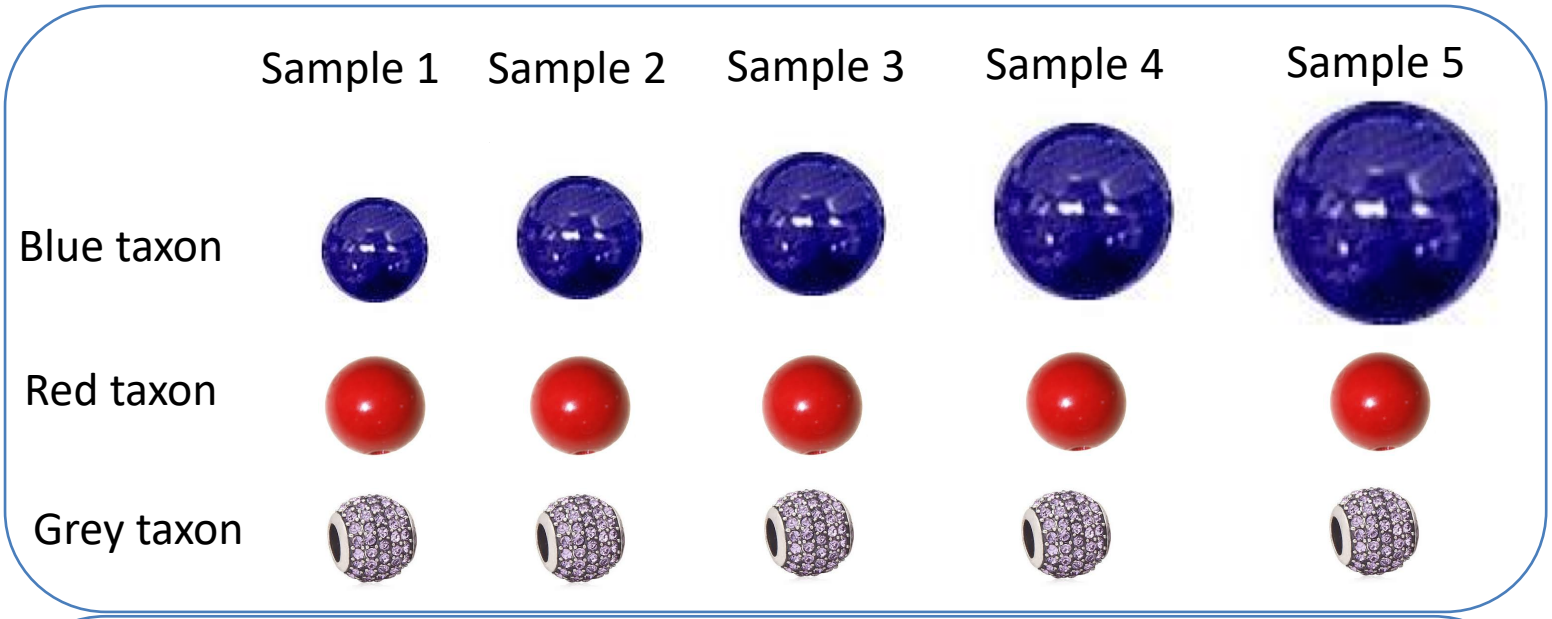


Removal of sequencing depth bias

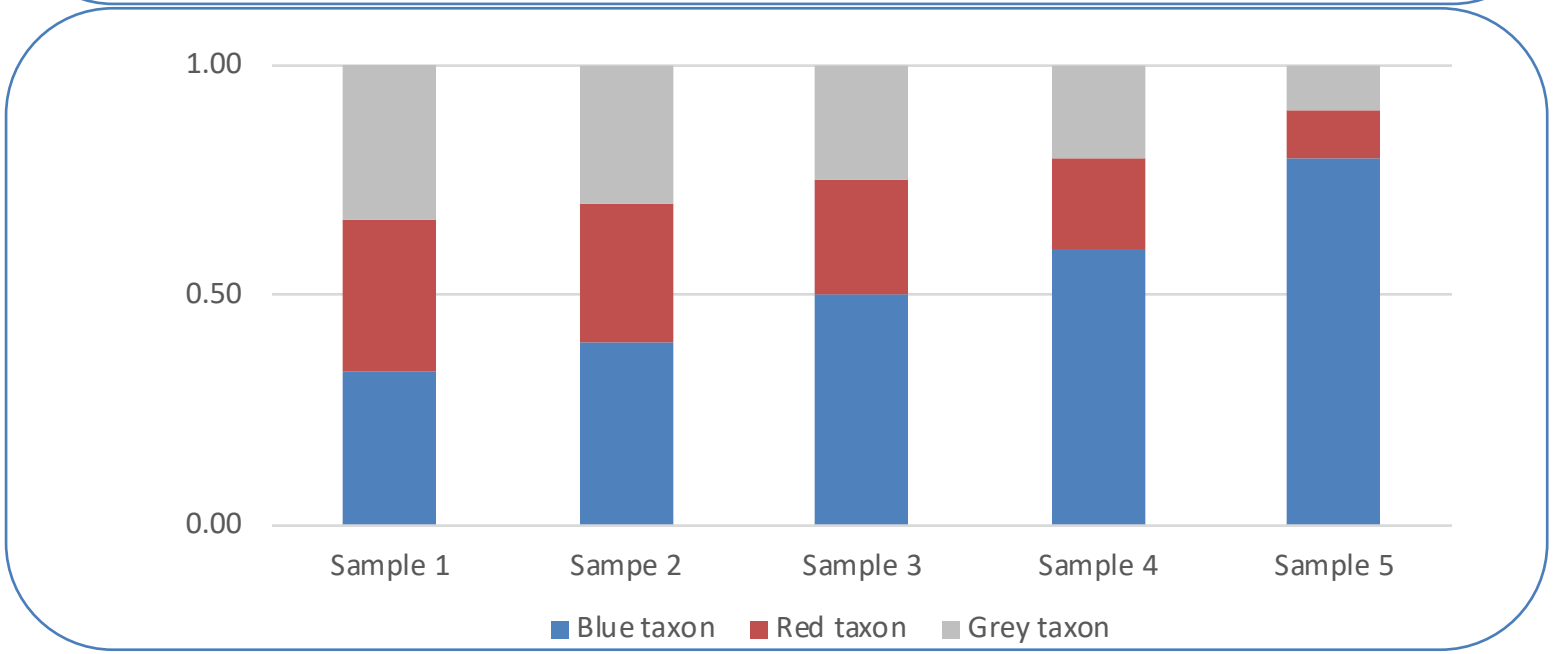
Rarefaction: Select beads from the big bucket with a probability equal to their proportion, until selected bead number is the same as in the small bucket
=> Additional zeros can be introduced, counts are preserved

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

Rarefaction/normalization: compositionality

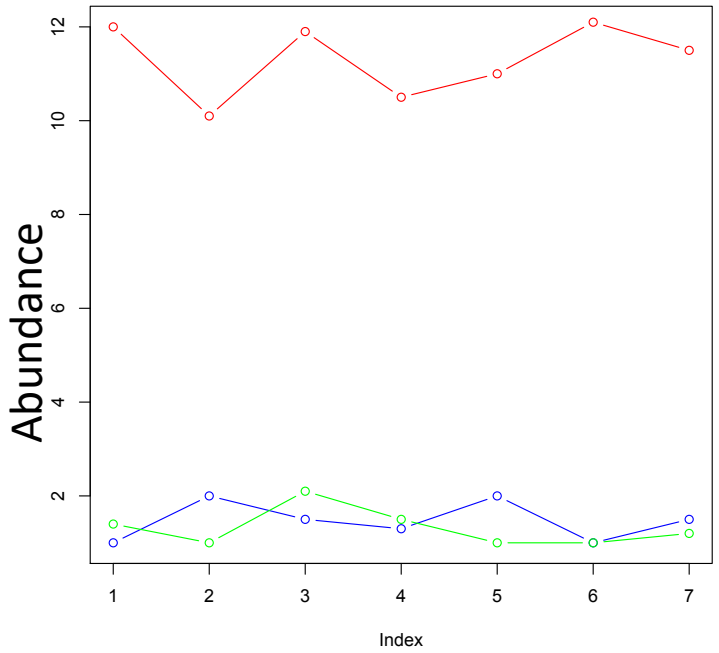


Absolute abundances

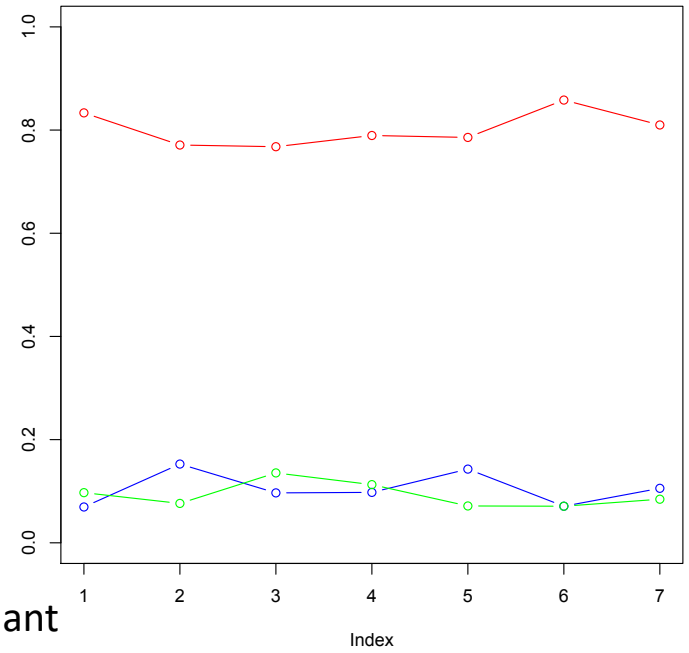


Relative abundances

Correlation distortion due to compositionality



normali-
zation



Abundant
Rare1
Rare2

	R1	R2	A
R1	1	-0.24	-0.69
R2		1	0.31
A			1

	R1	R2	A
R1	1	-0.32	-0.73
R2		1	-0.41
A			1

Pearson correlation

Tackling compositionality

- Compositionality distorts correlations when sample sum is constrained, which is the case after normalization or rarefaction
- Work-around 1: Transform data
  - **Centered log-ratio transform** (clr) is popular
  - Clr: divide each taxon by geometric mean of sample and take log
  - Problem: log(0) gives negative infinity: filter or **pseudo-counts**
- Work-around 2: Use compositionally robust measures based on ratios or log-ratios (e.g. Kullback-Leibler and Bray Curtis dissimilarities or variance of log ratios)
  - Problem: pseudo-counts & **lower sensitivity** than correlations
- Work-around 3: Convert to absolute counts using experimental measurements of total counts (e.g. flow cytometry)
  - Problem: **associations are driven by total counts** – not always desired (is difference in total count/carrying capacity of biological interest?)

Problem 2: Zeros

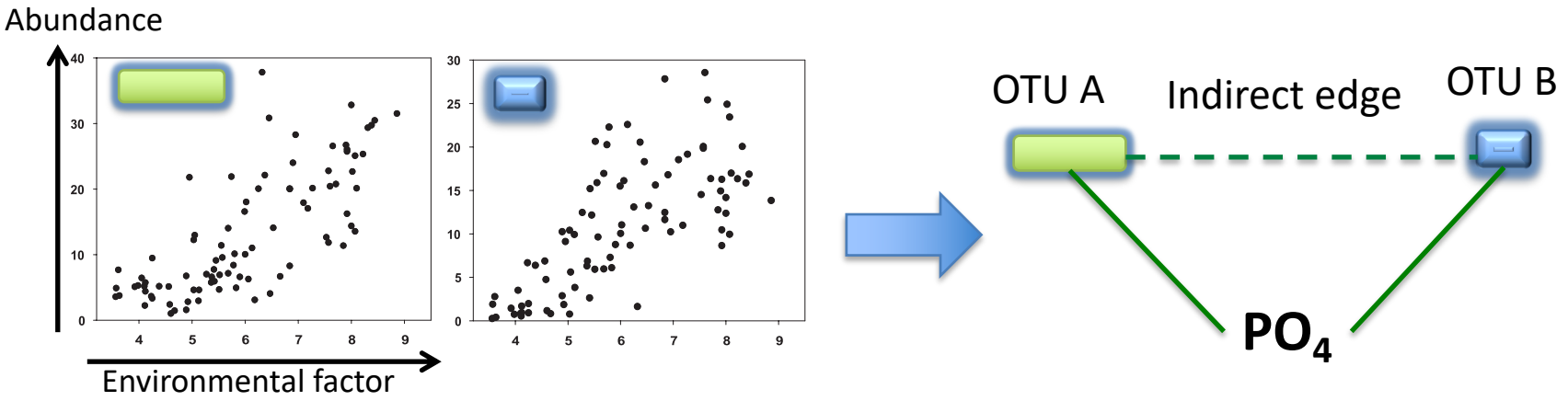
- Co-absence can yield a high correlation score, although it is not informative (a zero does not necessarily mean true absence, taxa could vary randomly below detection level)
- Measures unbiased by matching zeros, such as Bray Curtis and Kullback-Leibler dissimilarity, need sufficient number of informative data points

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

=> **Filter step required** to remove rare taxa (but keep their sum)

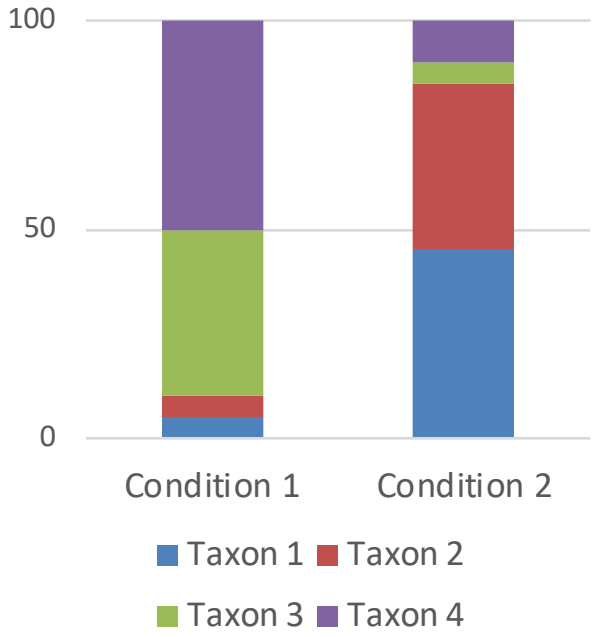
Problem 3: Indirect edges

- **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"
- Taking metadata into account during network construction can visualize indirect edges

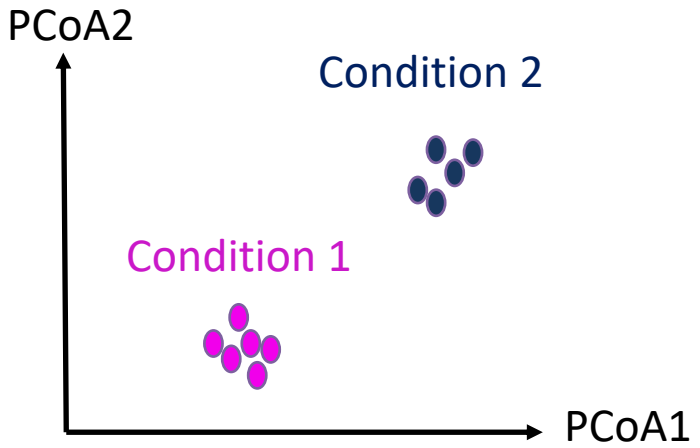


Problem 4: Heterogeneity

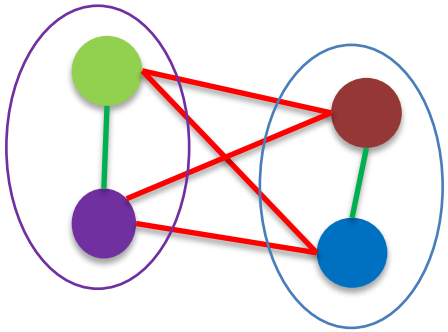
Sample heterogeneity



Clusters in sample-wise PCoA



Indirect edges in inferred network



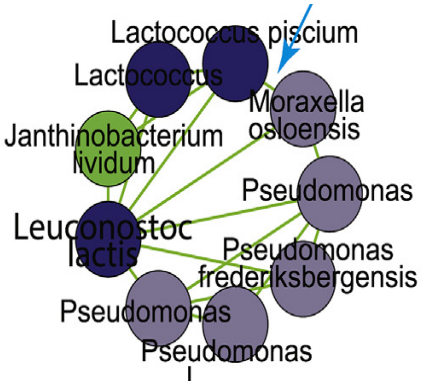
Abundant in condition 1

Abundant in condition 2

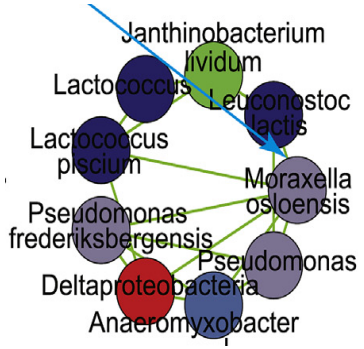
Solution: Split samples or include information on different conditions (e.g. temperature) in network inference

Problem 5: Resolution

- Local competition is hidden by shared niche preference
- Problem of experimental design

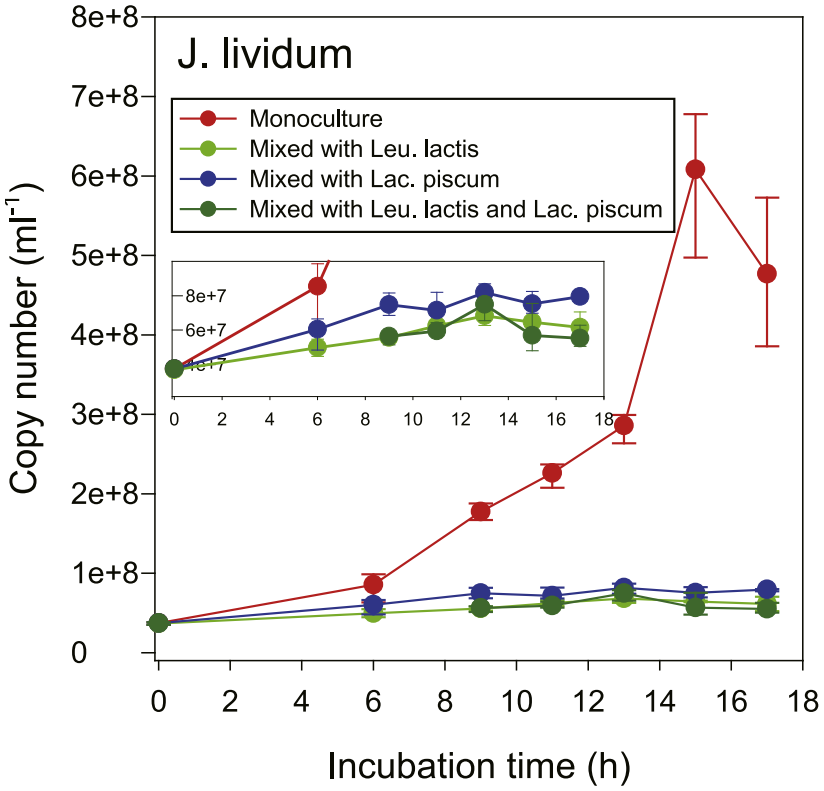


Tool: CoNet



Tool: MENA

Positive association predicted



Negative interaction found

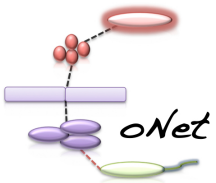
Wang et al. (2017) “Combined use of network inference tools identifies ecologically meaningful bacterial associations in a paddy soil” Soil Biology & Biochemistry 105, 227-235.

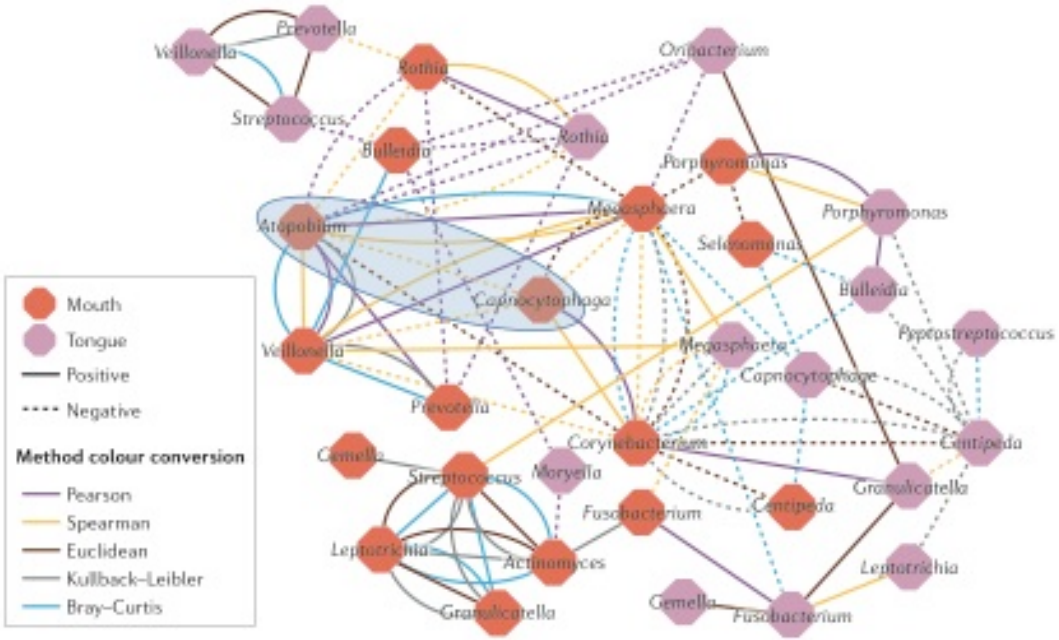
Tools

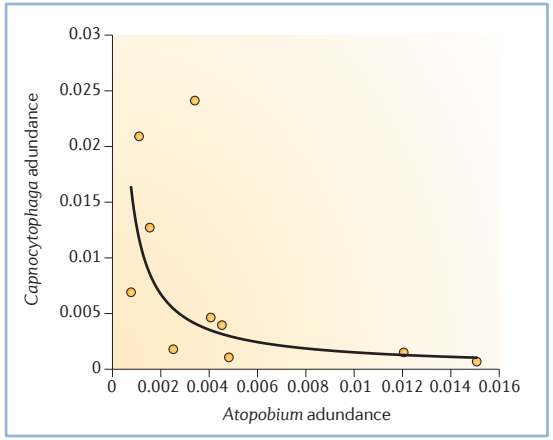
- Which are available microbial network inference tools and how do they work?

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





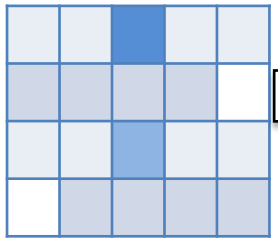


non-linear relationship is missed by Pearson

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

CoNet: Overview

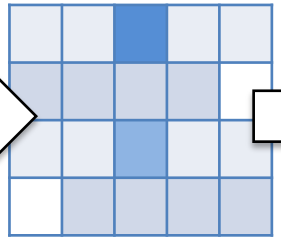
Raw taxon count matrix



Preprocessing

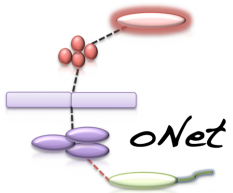
Rare taxa are discarded, but their sum is kept as an additional row (rare= arbitrary prevalence cut-off).
Counts are converted into relative abundances (**normalization**).

Relative abun-dance matrix



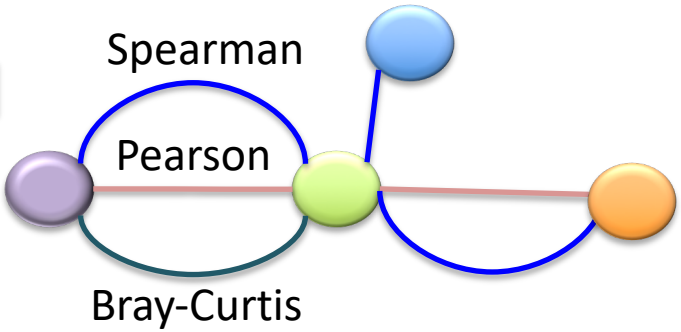
Edge score computation

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



Initial multi-graph

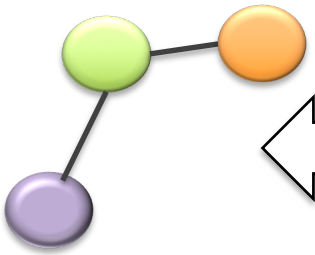
Only measure-specific edges above user-selected thresholds kept



Assessment of significance

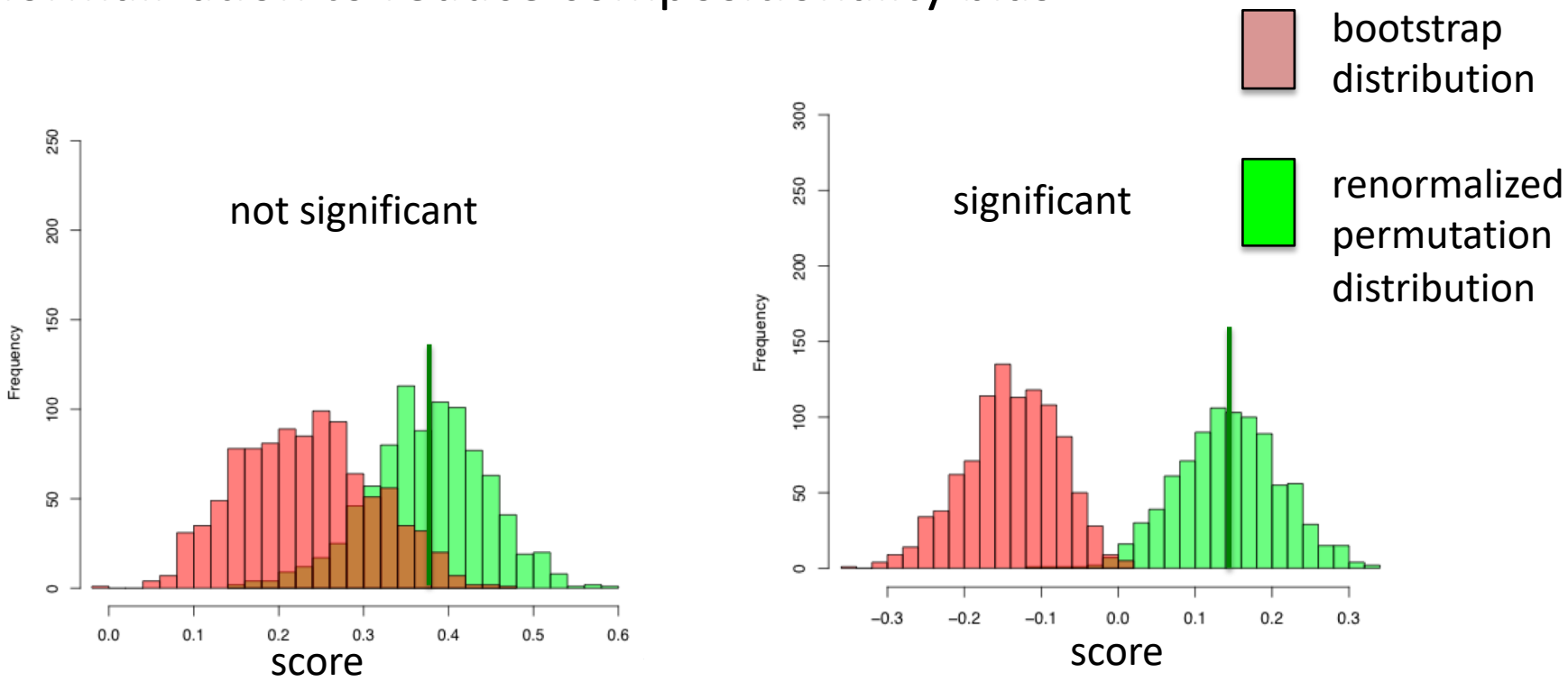
P-value computation, p-value merge with Brown's method, multiple testing correction with **Benjamini-Hochberg**, only significant edges kept

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



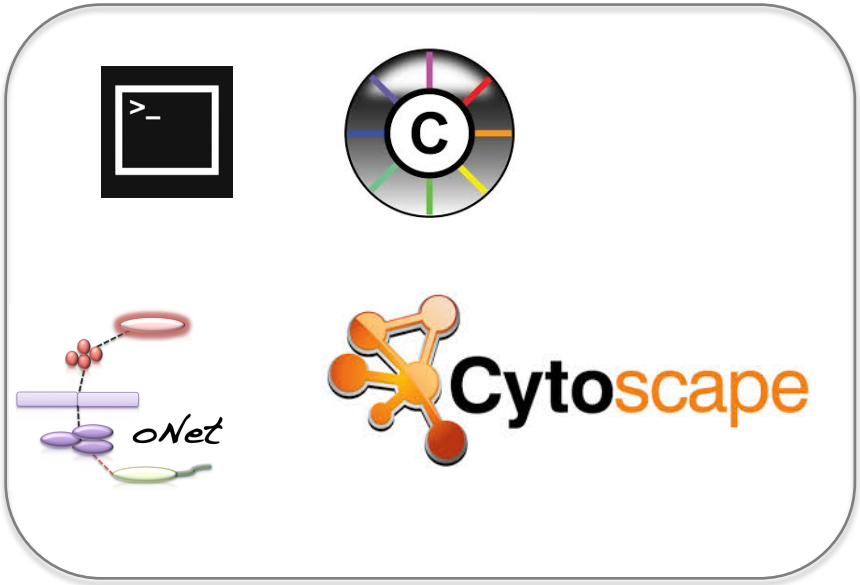
Faust*, Sathirapongsasuti* et al. “Microbial Co-occurrence Relationships in the Human Microbiome.” PLoS Computational Biology 8, e1002606, 2012.

CoNet Implementation

- CoNet is available on command line and as a Cytoscape app (versions 2.X and 3.X)
- CoNet page: **http://systemsbiology.vub.ac.be/conet**
- Cytoscape app: **http://apps.cytoscape.org/apps/conet**

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

> 19,000 downloads from Cytoscape app store



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

SPIEC-EASI

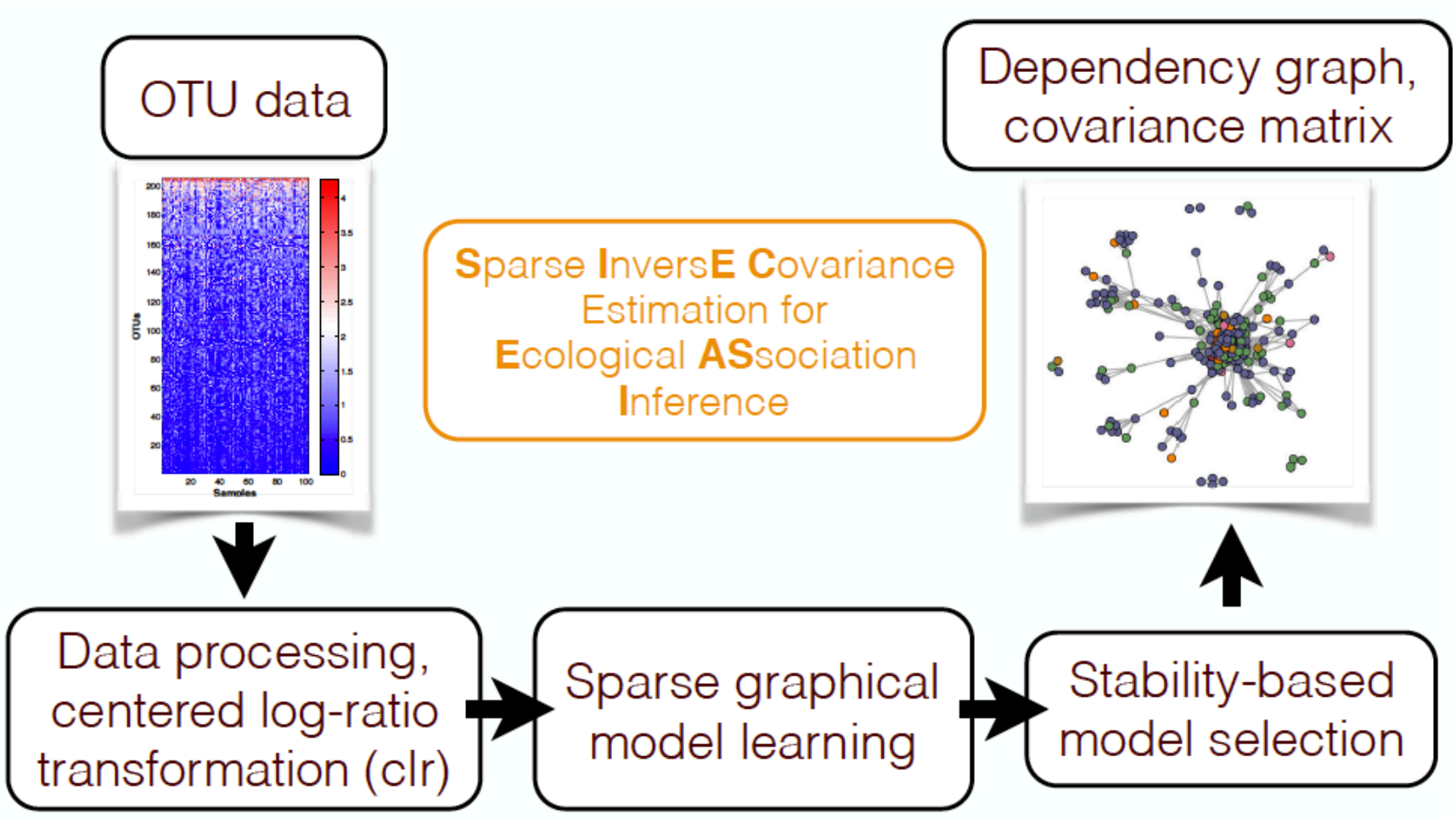


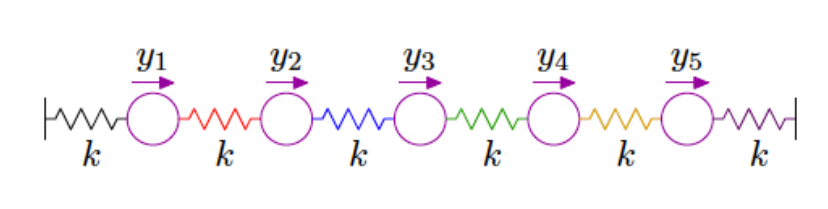
Image taken from: https://stamps.mbl.edu/images/f/f0/STAMPS_Network_1.pdf (Christian Mueller)

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

SPIEC-EASI: Sparse graphical models

- SPIEC-EASI estimates the **inverse covariance matrix**, such that resulting **network has fewer indirect edges**
- Assumptions: data are multivariate normally distributed and all relevant variables are taken into consideration
- A sparse inverse covariance matrix is estimated using penalty parameter λ (sparse: few non-zero entries)

Intuitive example by David MacKay:



Weights (nodes) connected by springs (edges)

Covariance matrix

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

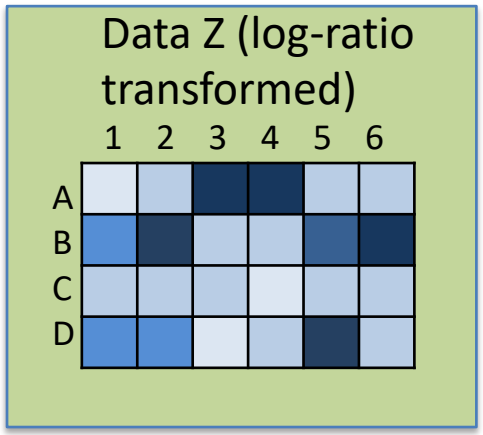
Inverse covariance matrix

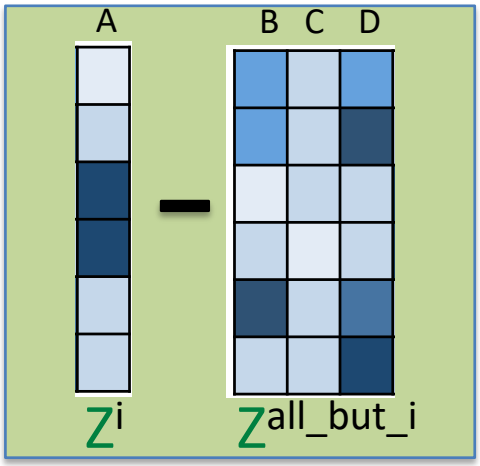
$$\mathbf{K}^{-1} = \frac{k}{T}\begin{bmatrix} 2 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 2 & -1 & 0 \\ 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & 2 \end{bmatrix}$$

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

SPIEC-EASI: Meinshausen & Bühlmann

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





$$\hat{\beta}^{i,\lambda} = \underset{\beta \in \mathbb{R}^{p-1}}{\arg\min}\left(\frac{1}{n}\left\|Z^i - Z^{\neg i}\beta\right\|^2 + \lambda\left\|\beta\right\|_1\right)$$

Penalty parameter

Species number

Sample number

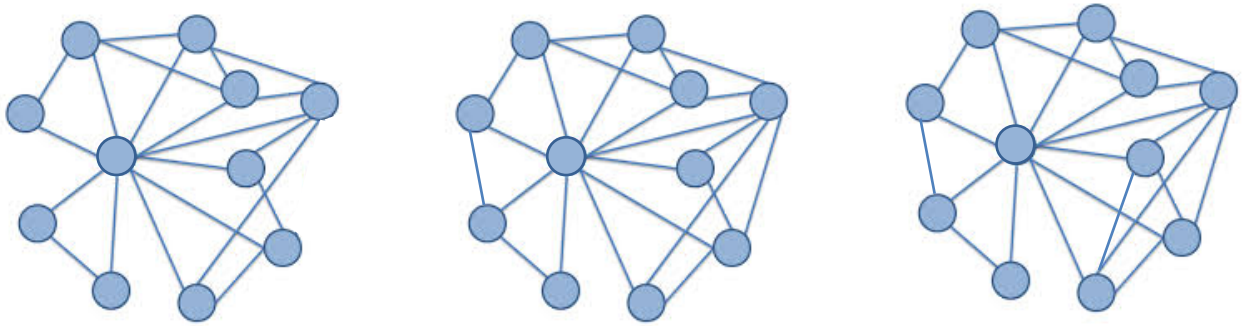
Regression coefficients

Result: matrix of regression coefficients; is symmetrised

Edge = non-zero regression coefficient

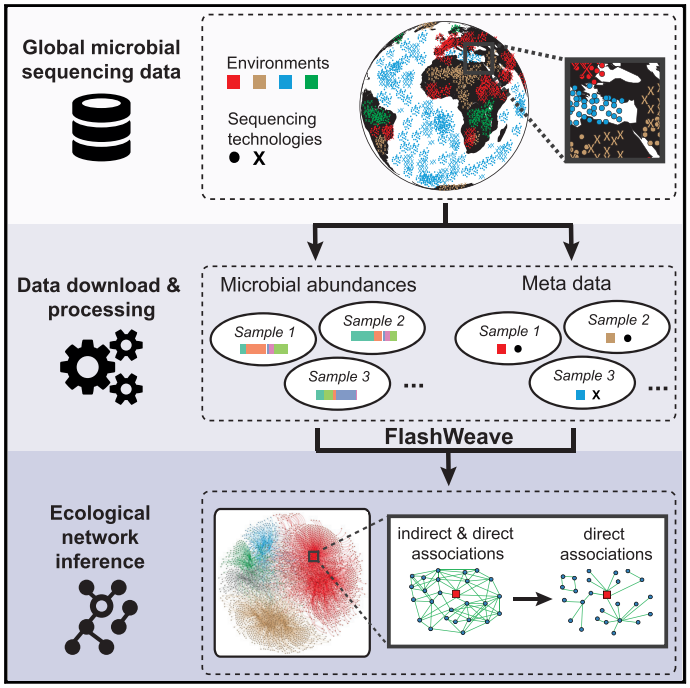
SPIEC-EASI: Stability-based model selection

- **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



The new kid on the block: FlashWeave

- SPIEC-EASI's main weakness: does not take environmental data into account
- "FlashWeave = SPIEC-EASI + metadata": exploits inverse covariance to reduce indirect edge number, taking metadata into account
- Optionally tackles heterogeneity by omitting zeros from the computation of associations ("structural zeros")



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.

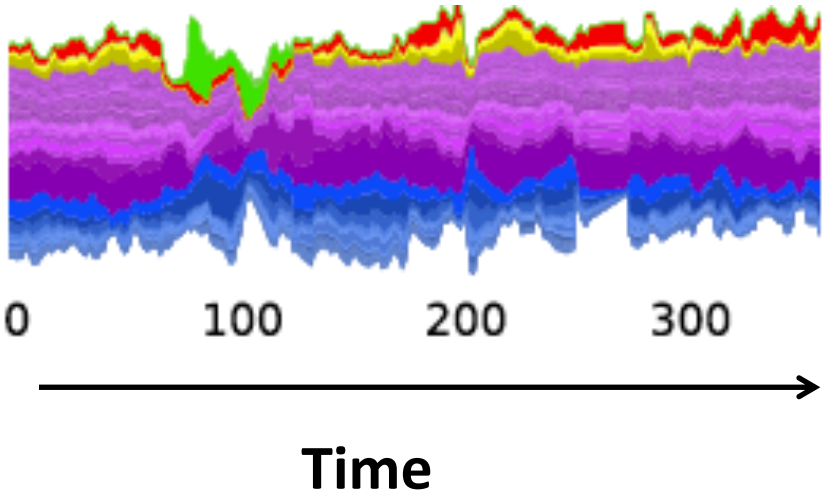
Other microbial network inference tools

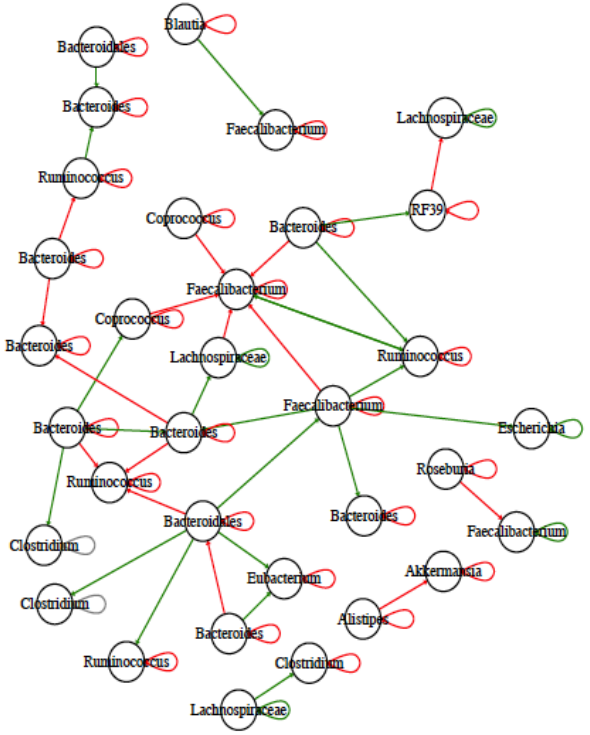
- **SparCC**: sparse correlations robust to compositionality
- **MENAP** (Molecular Ecological Network Analyses Pipeline): exploits random matrix theory to threshold similarity matrix
- **REBACCA/CCLasso**: sparse compositionality-robust correlations
- **MInt:** Takes environmental factors into account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

List is not complete

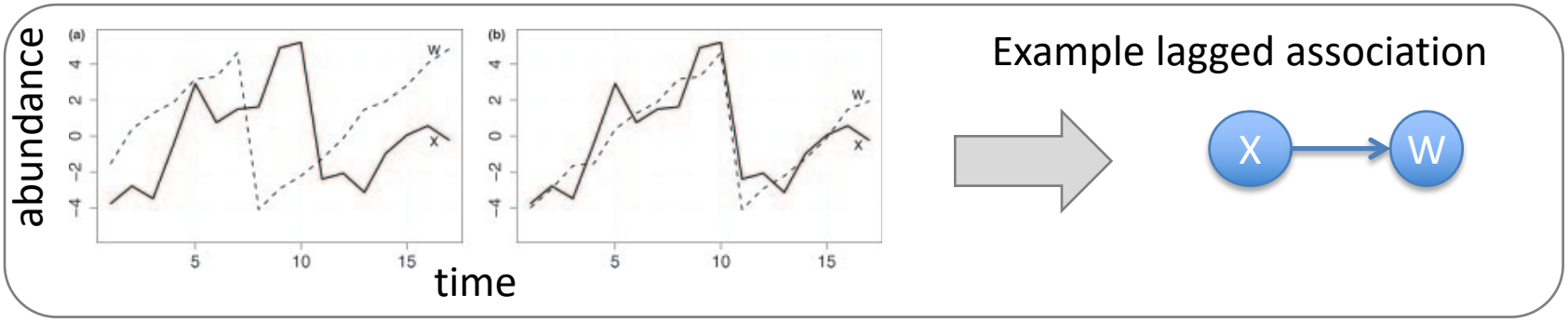
SparCC: Friedman & Alm (2012) "Inferring Correlation Networks from Genomic Survey Data." PLoS Comp Bio 8 (9), e1002687.
MENAP: Zhou et al. (2010) "Functional Molecular Ecological Networks" mBio 1 (4), e00169-10.
REBACCA: Ban et al. (2015) "Investigating microbial co-occurrence patterns based on metagenomic 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.

Tools exploiting time series information





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 tools:
  - **http://hallam.microbiology.ubc.ca/fastLSA/install/index.html**
  - **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.

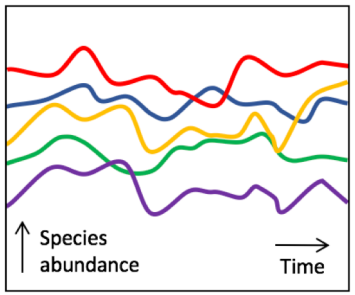
Generalized Lotka-Volterra (gLV)

- **Population model** describing community dynamics
- Species abundance x_i is modeled as a function of species *i's* initial abundance, its growth rate b_i and its interaction strengths a_{ij} with other species *j*

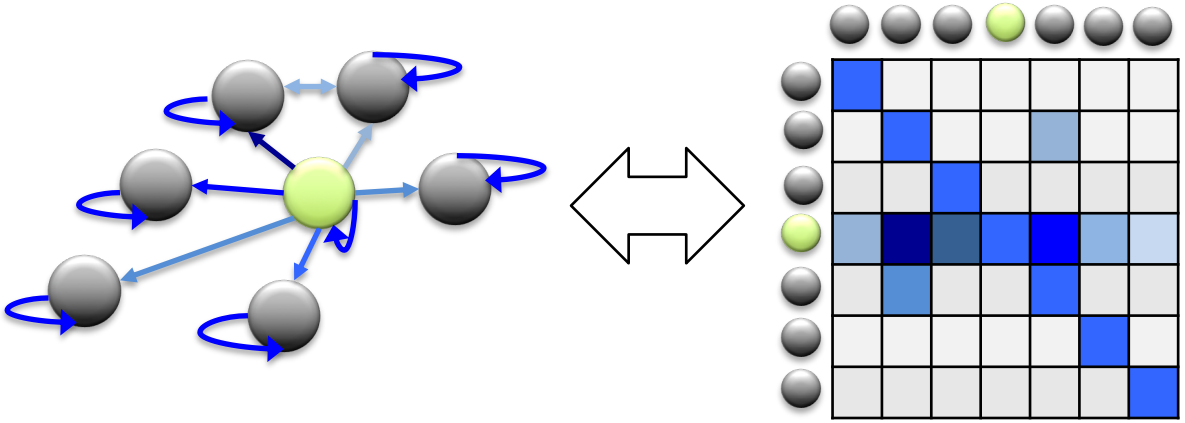
gLV equation

$$\frac{dx_i(t)}{dt} = x_i(t)\left(b_i + \sum_{j=1}^{N} a_{ij}x_j(t)\right)$$

Community time series



- Interaction matrix A (=community matrix) is a directed network, interaction strengths are edge weights



GLV parameterization tools

- Tools parameterizing gLV equation from a community time series (they infer a directed network):
  - **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

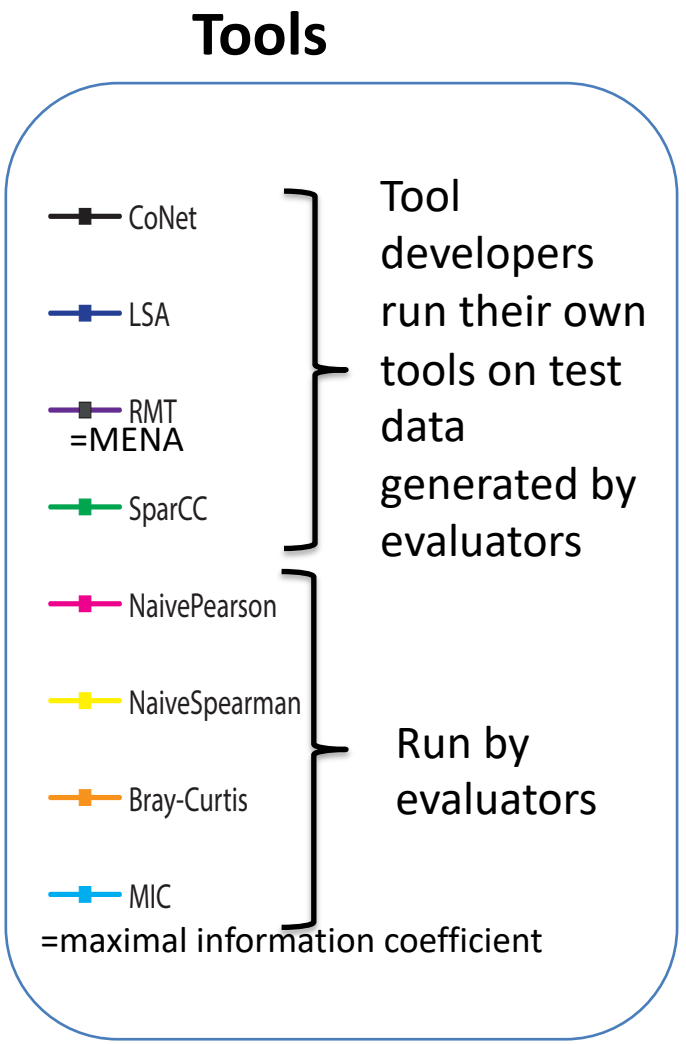
- How well do cross-sectional tools perform?

Warning:
Everything in this section is on synthetic data only

Tool Evaluation I

Data generation

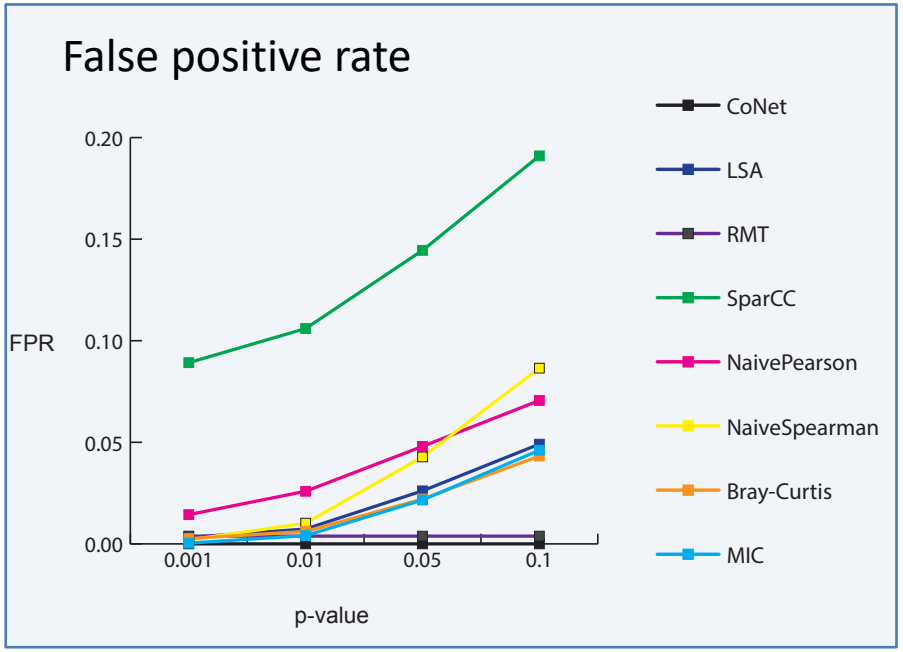
- False positive rate: Dirichlet Multinomial distribution (no interactions)
- Impact of repeated rarefaction: Sequencing data from a mouse study (Ridaura et al. 2013)
- Variation of compositionality: Multiplication of one OTU pair by a constant
- Ecological interactions: Random values + ecological interaction rules



Evaluation: Weiss, Van Treuren, Lozupone, Faust et al. The ISME Journal 10, 1669-1681, 2016.
Data: Ridaura et al. Science 341 (6150), 2013.
MIC: Reshef et al. Science 334, 1518-1524, 2011.

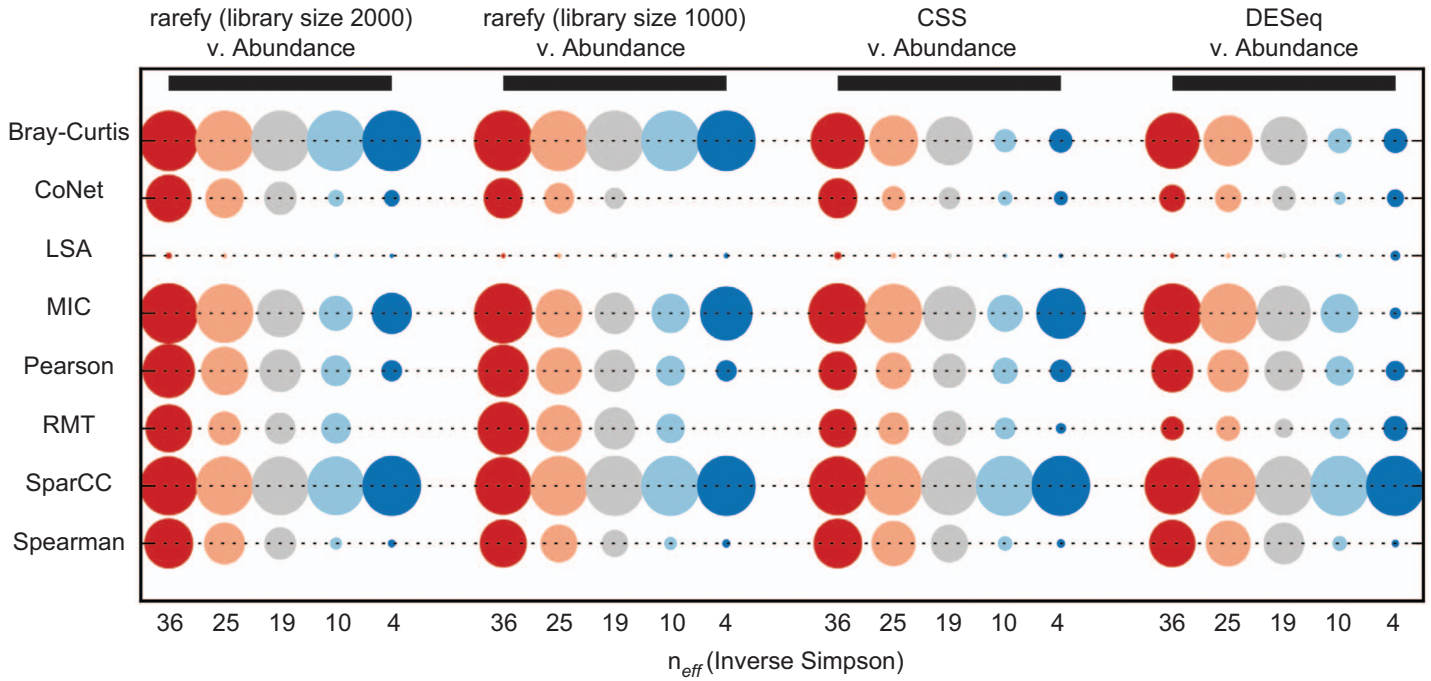
Tool Evaluation I: False positives and noise

- 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)



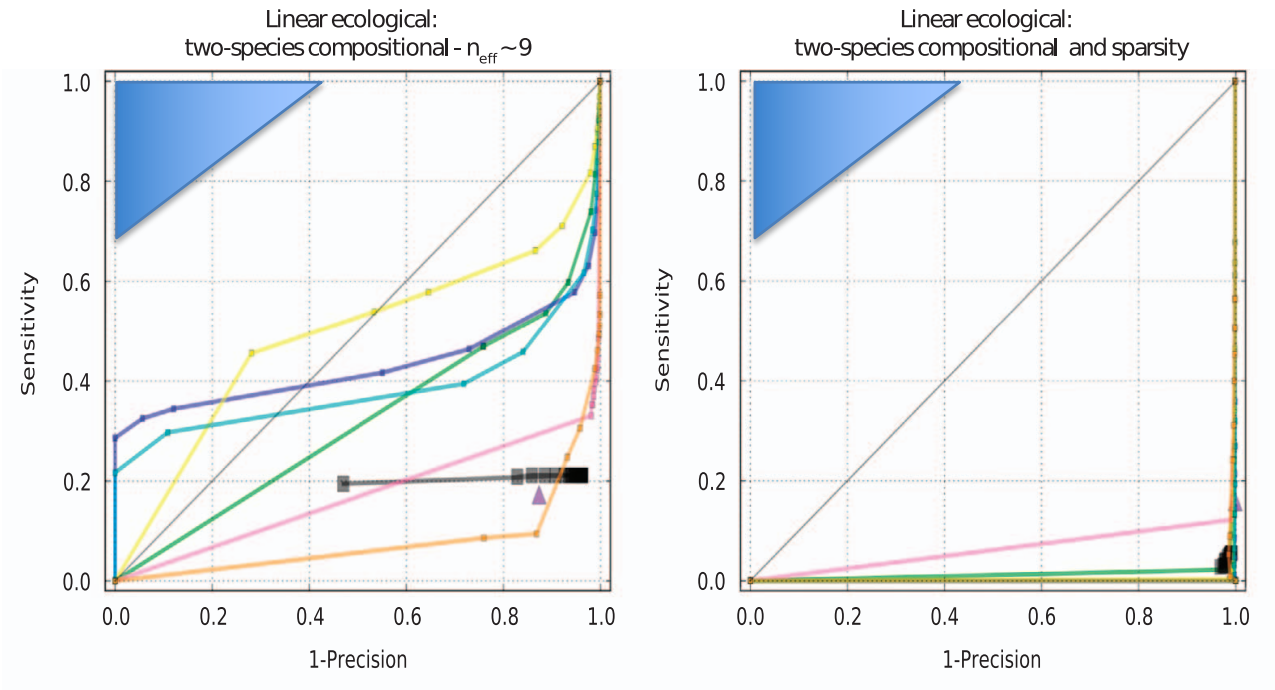
Tool Evaluation I: Effect of compositionality

- 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



Tool Evaluation I: Interactions

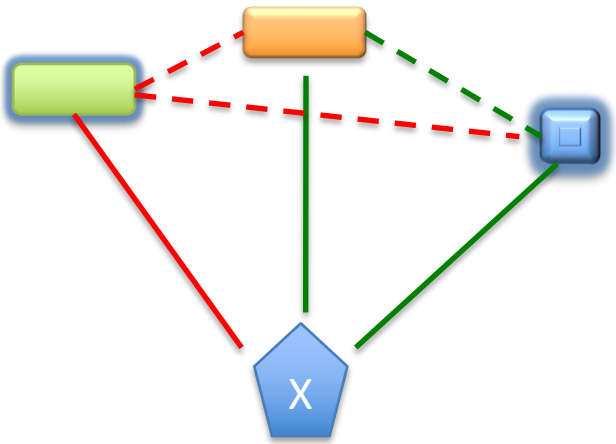
- Interaction detection accuracy in zero-rich, compositional data is low for all tools



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

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 yet included (too recent)



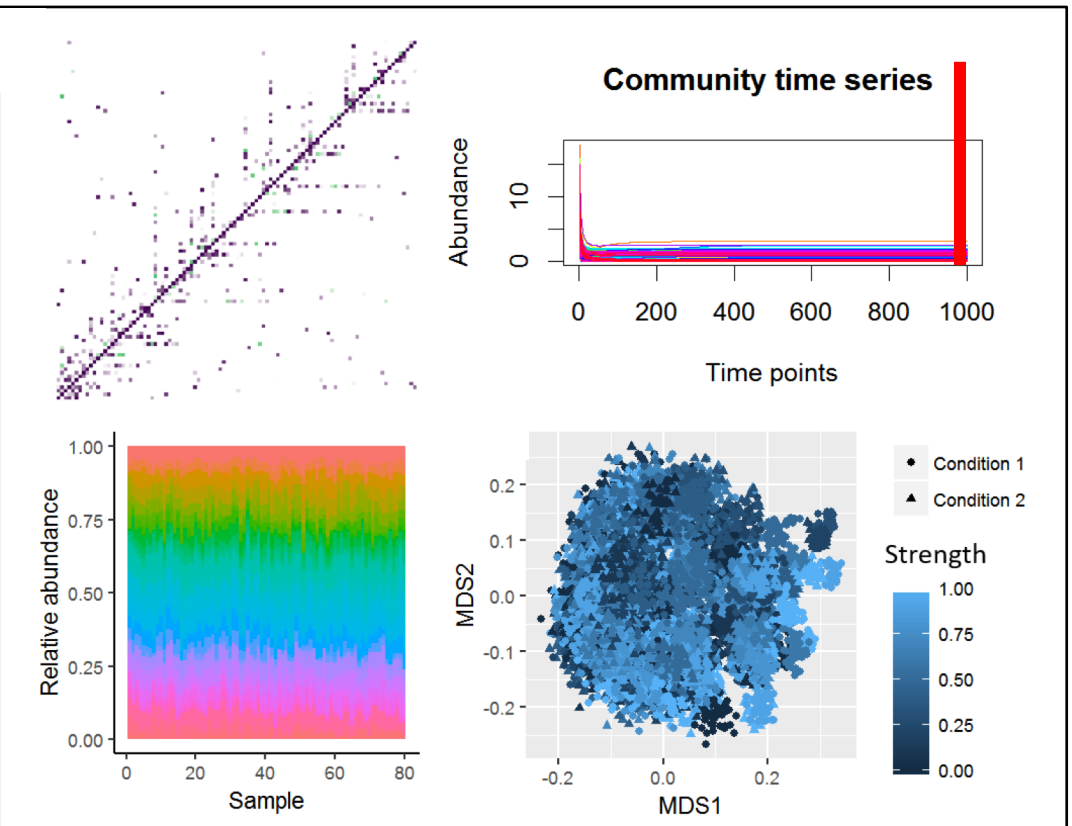
SPIEC-EASI: Kurtz et al. PLoS Computational Biology 11(5), e1004226, 2015.
gCoda: J. Comput. Biol. 24(7), 699-708, 2017.

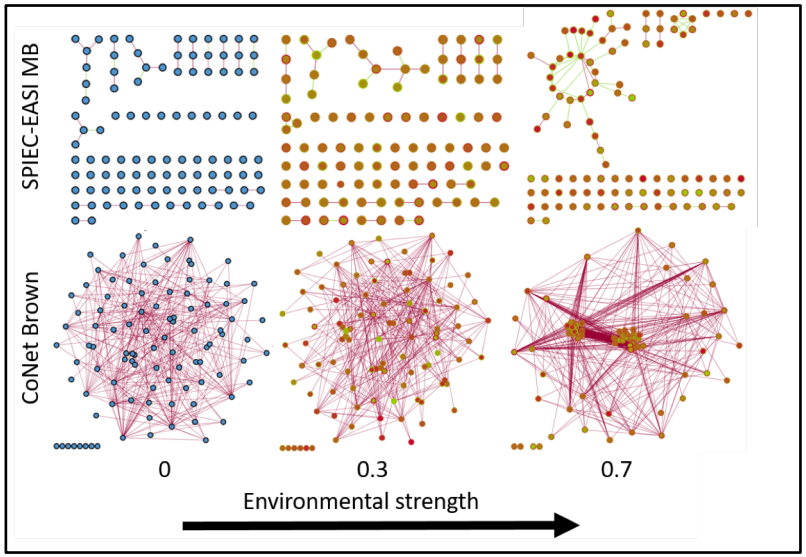
Tool evaluation II (with environment)

Data generation:

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

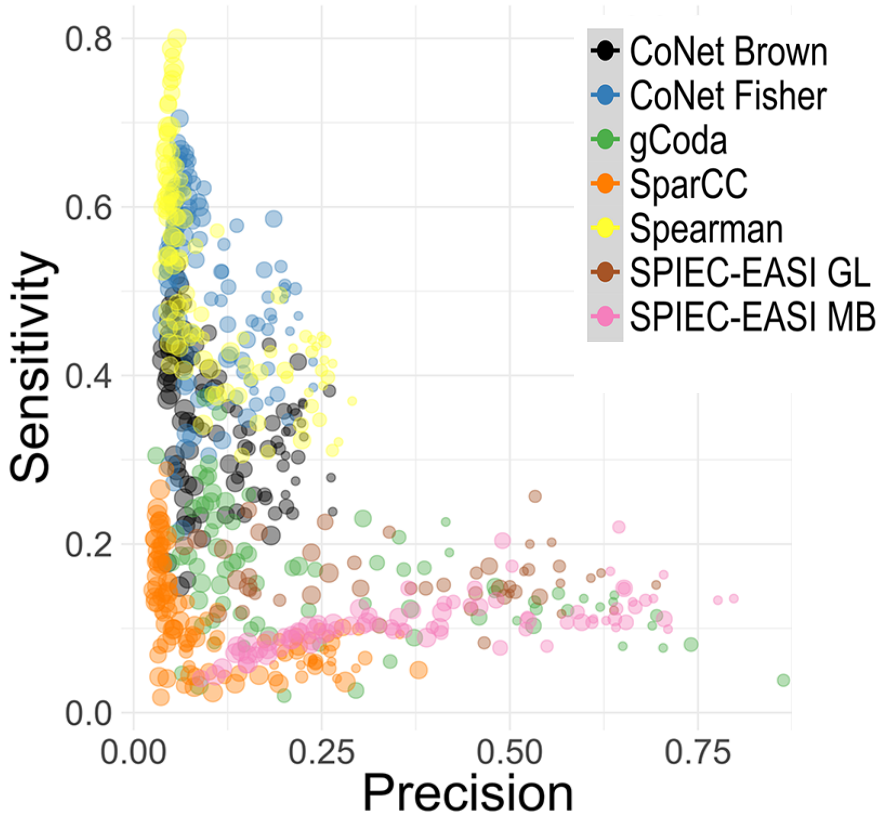
Lisa Röttjers





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 (yet)

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

Biological validation of interaction prediction in cross-sectional data: 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

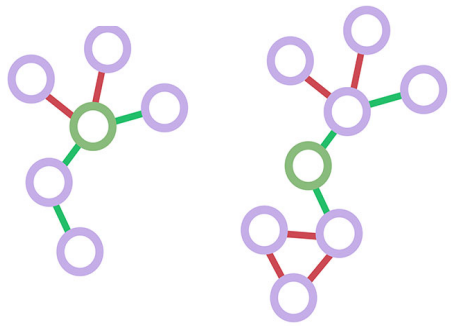
Phage-Host (Edwards et al., FEMS 2015)
Tool: Pearson
Data: 3025 global metagenomic samples
Validation data: Known hosts for 820 phages
Sensitivity: hosts correctly predicted for 9.5% of the phages (78 found)
Precision: uncertain

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

Can tools predict keystone species?

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"

- Keystone property is often assessed using network properties (hub or connector taxa)
- Are inferred networks sufficiently accurate to detect hub or connector taxa? Are hub or connector taxa keystones in the ecosystem?



Hub taxon: many neighbors

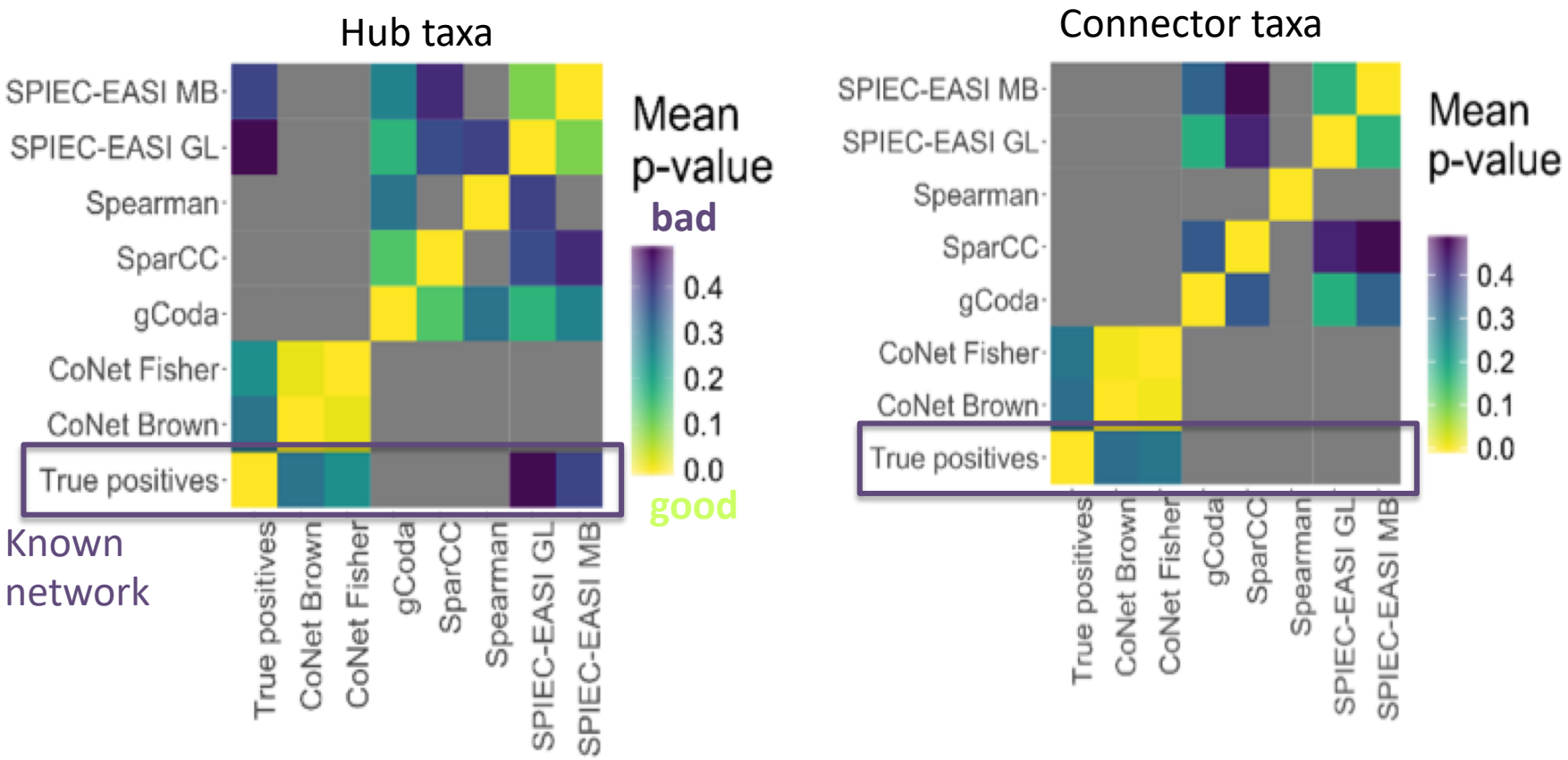
Connector taxon: crucial point for information flow

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

Can tools predict keystone species?

- Not really (simulations)

Significance of overlap between top 5 true and predicted hub/connector taxa:

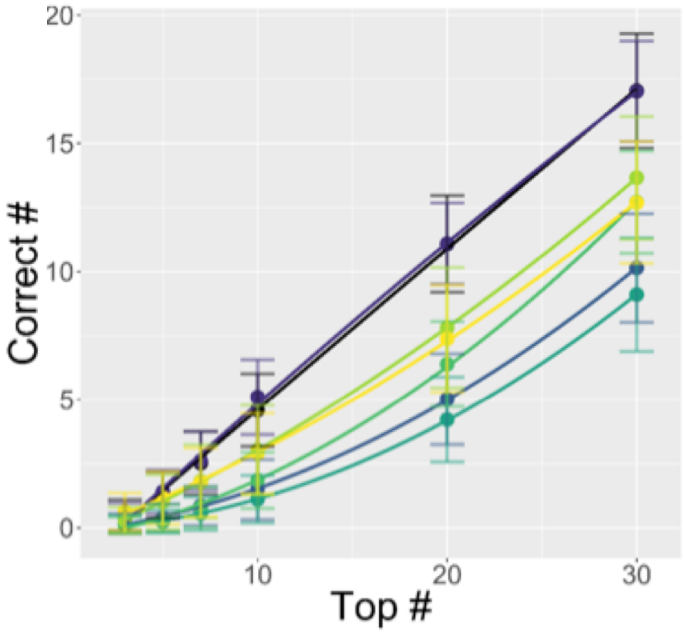


Röttjers & Faust, FEMS Microbiology Reviews 42, 761-780.

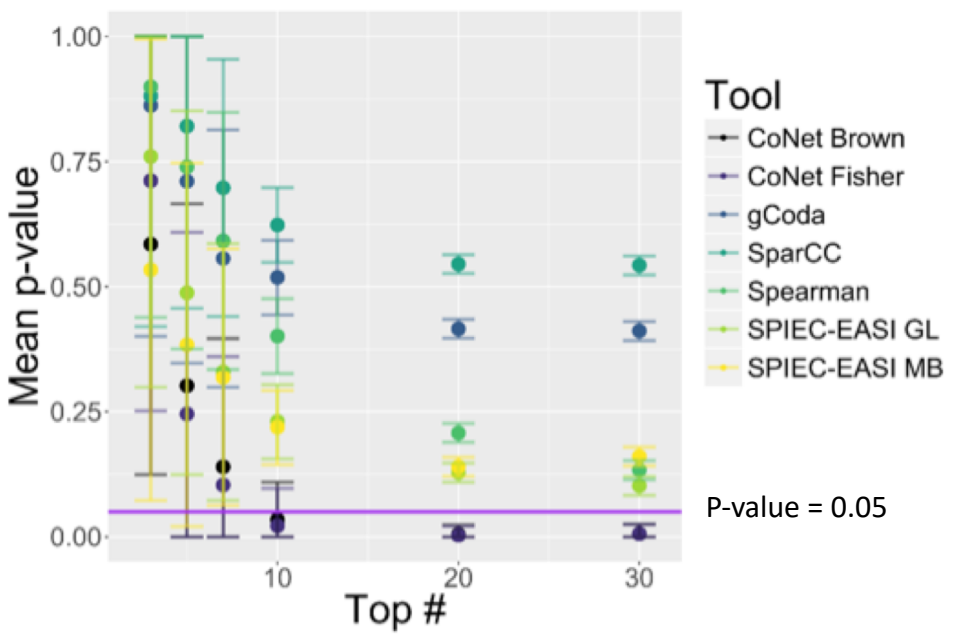
Can tools predict keystone species?

- When a larger number of predicted top hub nodes is considered, CoNet significantly enriches for correct hubs

Matching fraction of hub nodes

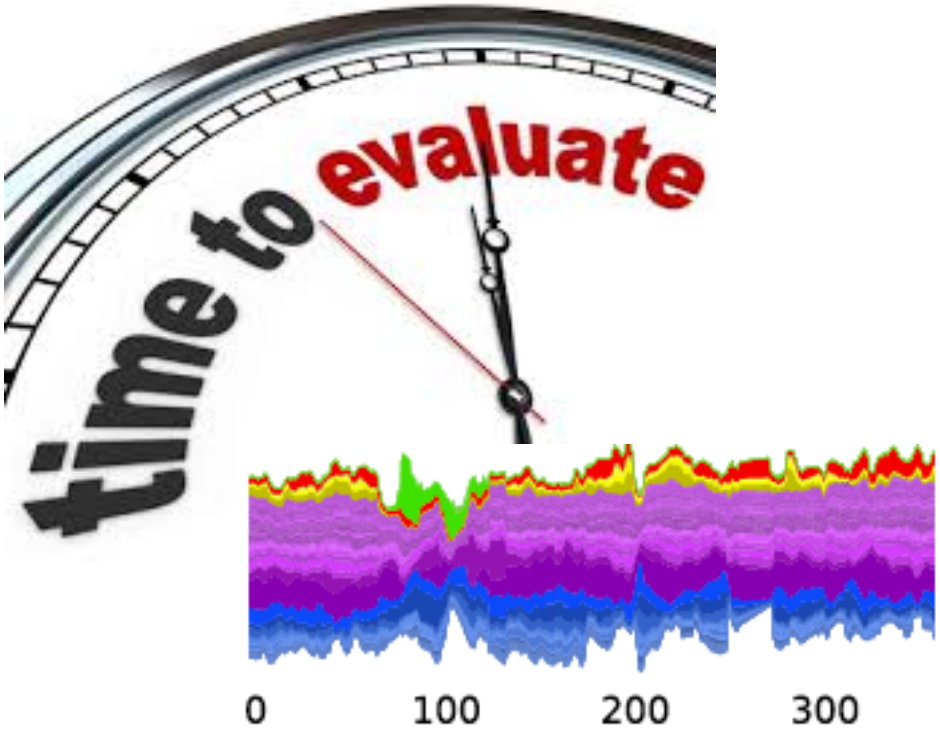


P-value of matching fraction



Evaluation of network inference from time series

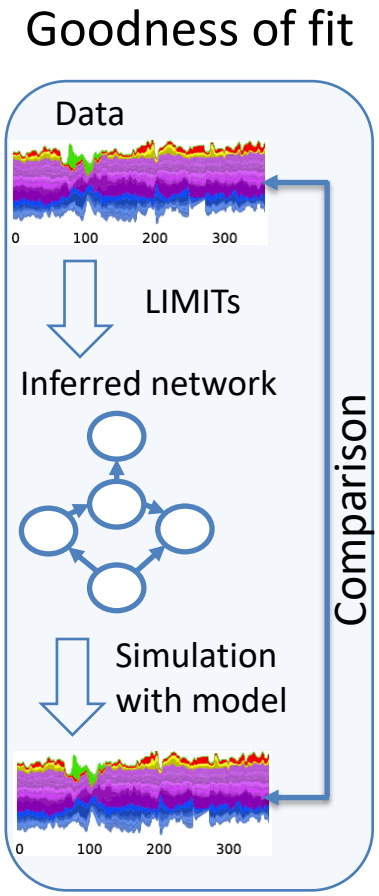
- How well do tools dedicated to time series perform?



Warning:
Again on synthetic data only

Tool evaluation III (time series)

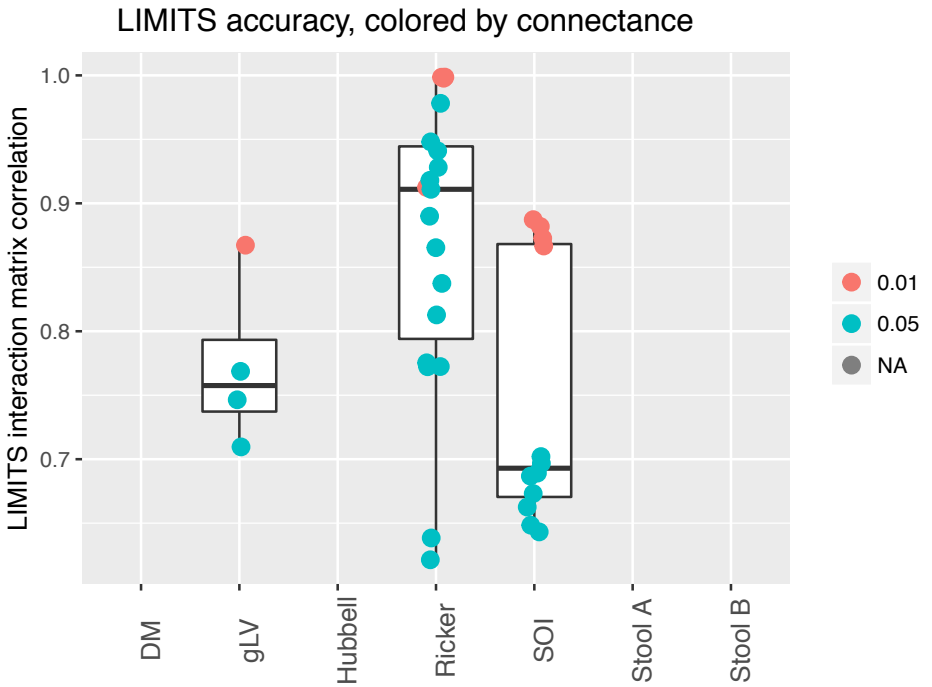
- Time series generated with different population models (including gLV)
- Parameters (that is networks) known
- Networks inferred from simulated time series with LIMITS (the only tool evaluated)
- Two comparisons: 1) Known network directly compared to inferred network (accuracy of inference), 2) Original time series compared to time series generated with model parameterized with inferred network (goodness of fit)



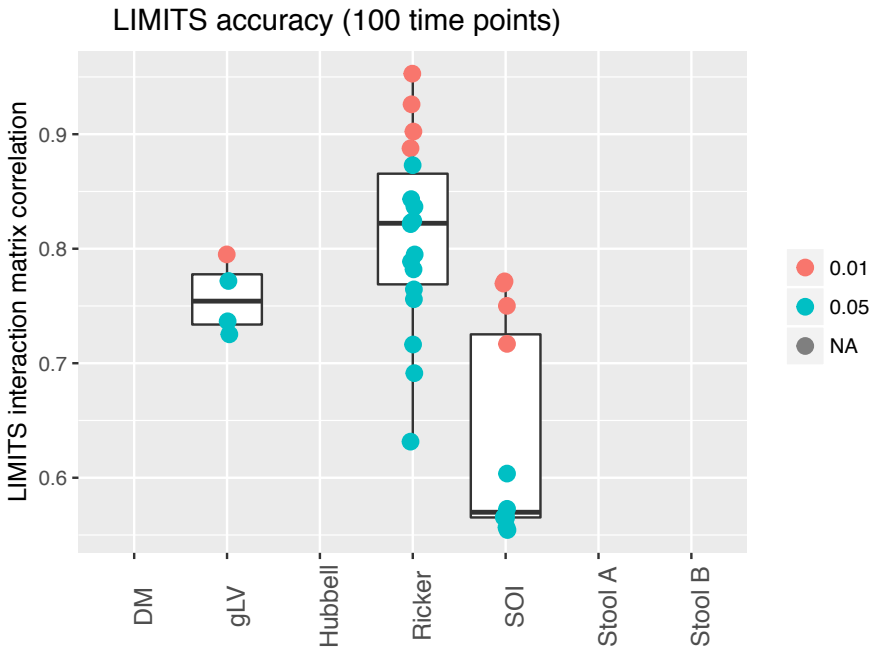
LIMITS: Fisher and Mehta (2014). "Identifying Keystone Species in the Human Gut Microbiome from Metagenomic Timeseries using Sparse Linear Regression." *PLoS one* 9, e102451.
Evaluation: Faust et al. (2018) "Signatures of ecological processes in microbial community time series", Microbiome 6, 120.

Tool evaluation III (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



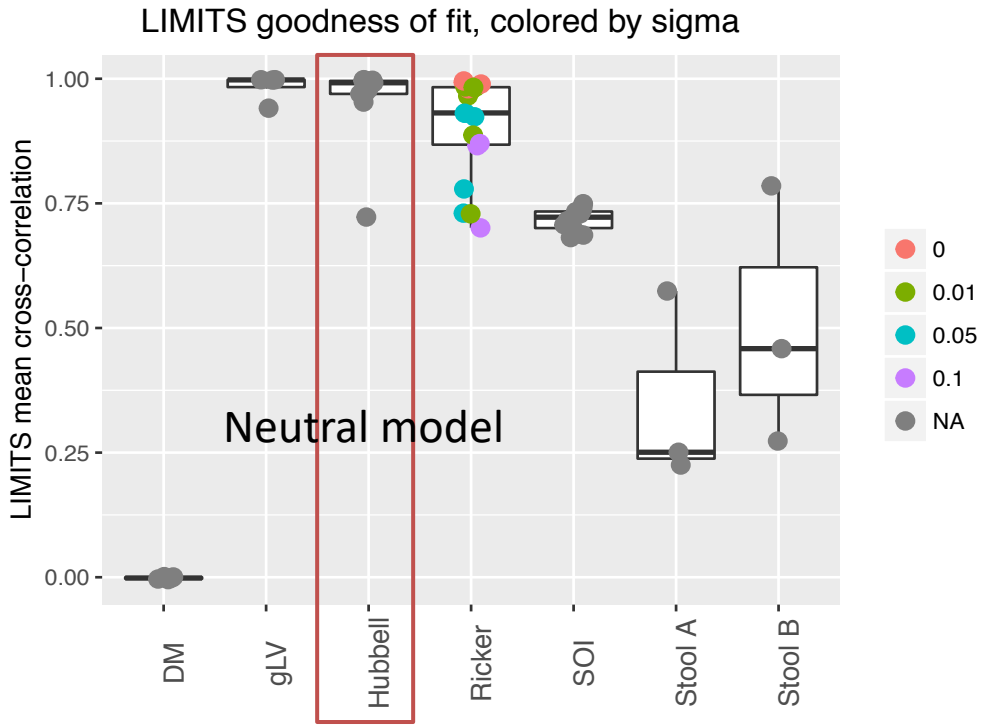
100 species, 3000 time points



100 species, 100 time points

Tool evaluation III (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)



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

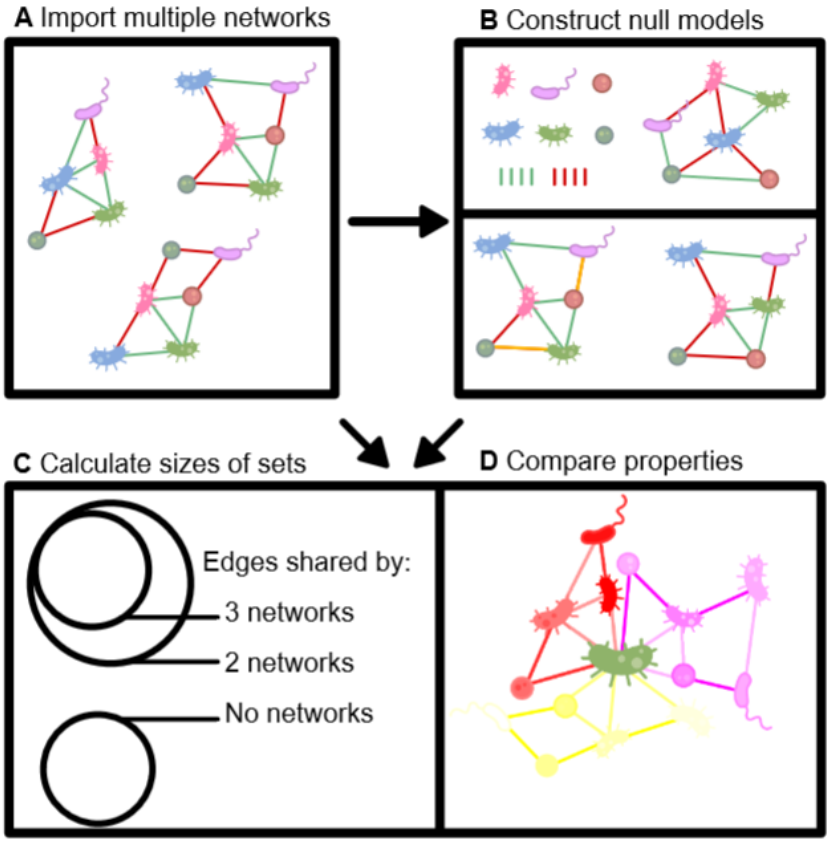
See also: Cao et al. (2017) "Inferring human microbial dynamics from temporal metagenomics data: Pitfalls and lessons" Bioessays 39(2).

Network comparison

- Problem 1: How many networks are needed to test whether their properties are significantly different?
- Problem 2: How many networks are needed to test whether a network property is significantly correlated to time or a gradient?
  - Example: Are five networks sufficient to conclude that complexity (=edge number) increases in a succession?
- **Problem 3: Is a network core present or do networks overlap not more than expected by chance?**

Anuran: a toolbox for comparing noisy networks

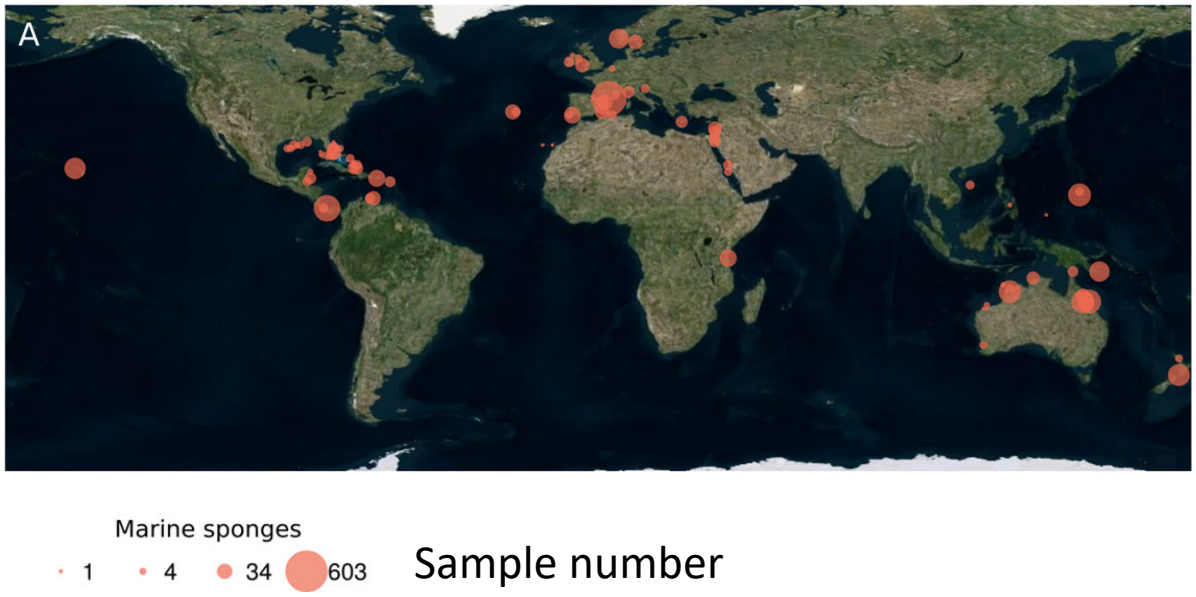
- Implements 2 types of null models: network randomization with and without preserving node degree distribution
- Tests whether a network property or a core network is significant given randomized networks
- Extension: test whether properties of a network set are significantly different from another set or whether a property in a network set significantly correlates to a vector
- To be released soon



Lisa Röttjers

Anuran in action: Sponge microbiome

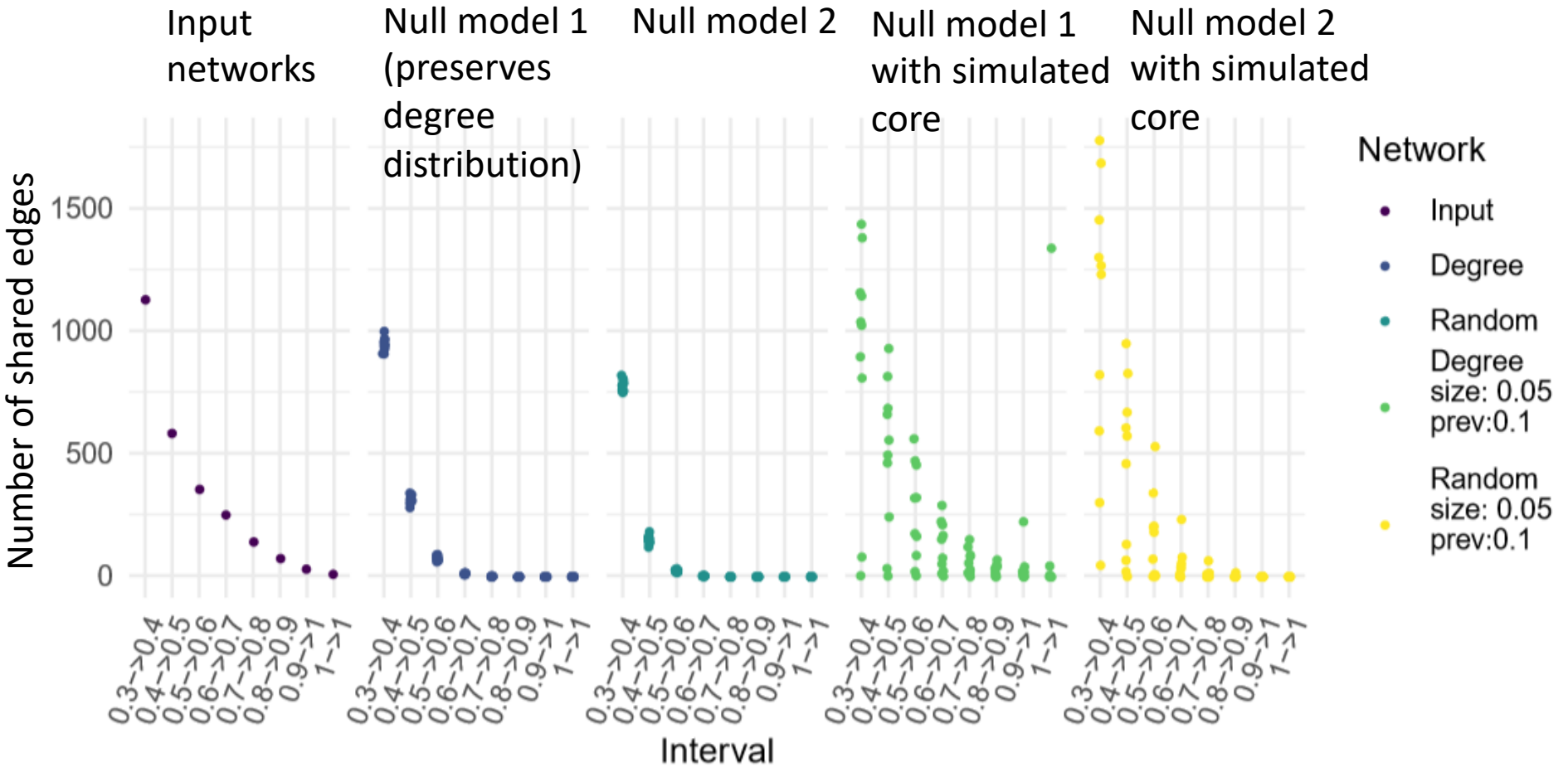
- Sponges microbiome project : microbiota of 268 different sponge host species collected around the globe
- > 3000 sponge specimens sequenced



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.

Anuran in action: Sponge microbiome

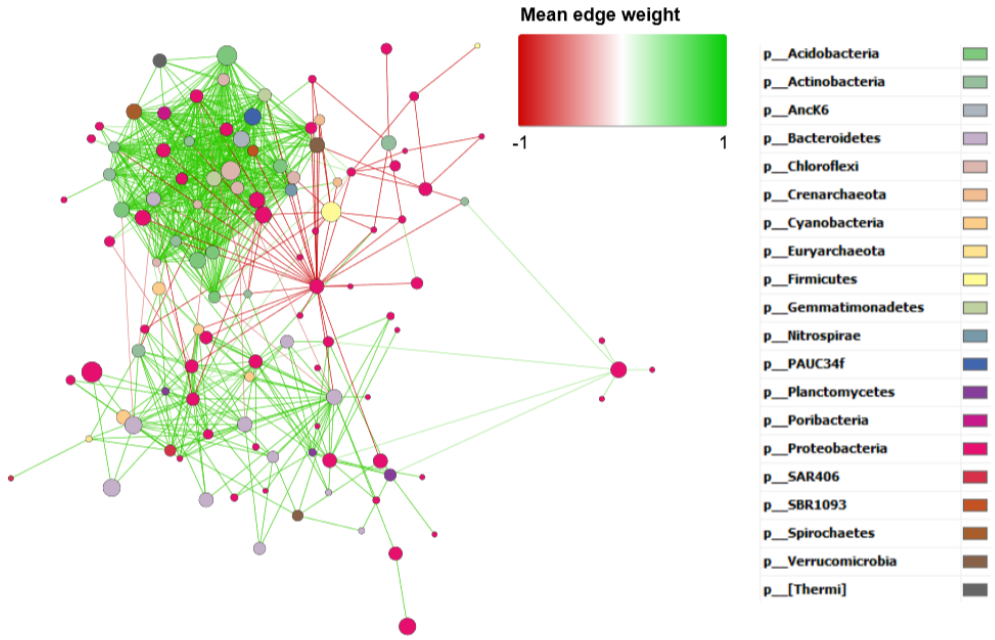
- Anuran run on CoNet networks specific to sponge orders
- Significant core network found for a fraction of networks



= Fraction of networks in which shared edges occur

Anuran in action: Sponge core network

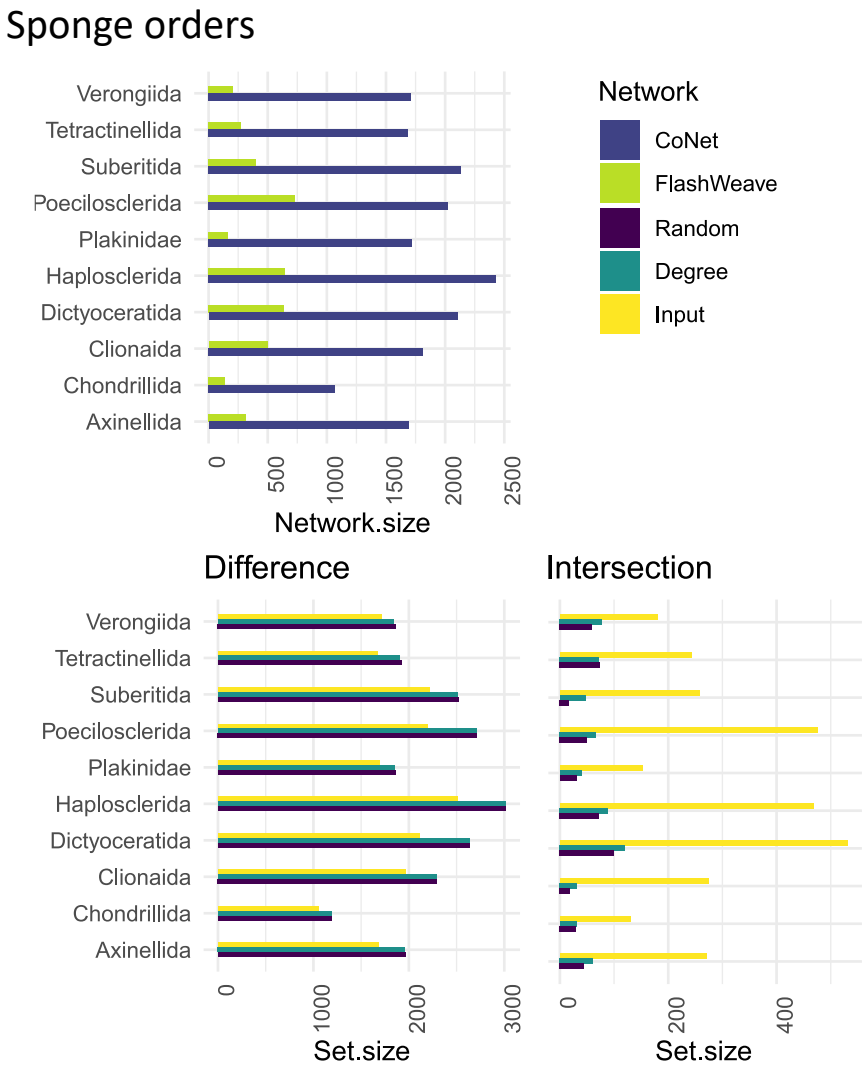
- Core network (edges in >= 4 networks) has two clusters
- Clusters contain indicator taxa for high versus low microbial abundance sponges (HMA vs LMA)
- HMA vs LMA classification traverses sponge orders
- No 100% core expected (there are no more highly preserved edges than expected at random)



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: 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)



Take-home messages

- Microbial network construction: split heterogeneous samples, filter rare taxa (but keep their sum), normalize and use a log transform or compositionality-robust association measure(s); consider conversion to absolute counts
- Be aware that the accuracy of ecological interaction prediction can be really low
- Properties of inferred networks do not necessarily reflect the properties of the true networks
- GLV parameterization: goodness of fit can be misleading
- Significance of network comparison can be assessed with null models


That's all Folks!

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):

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

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

Supplement

Definition of measures

Supplement

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)

Definition of measures continued

Supplement

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

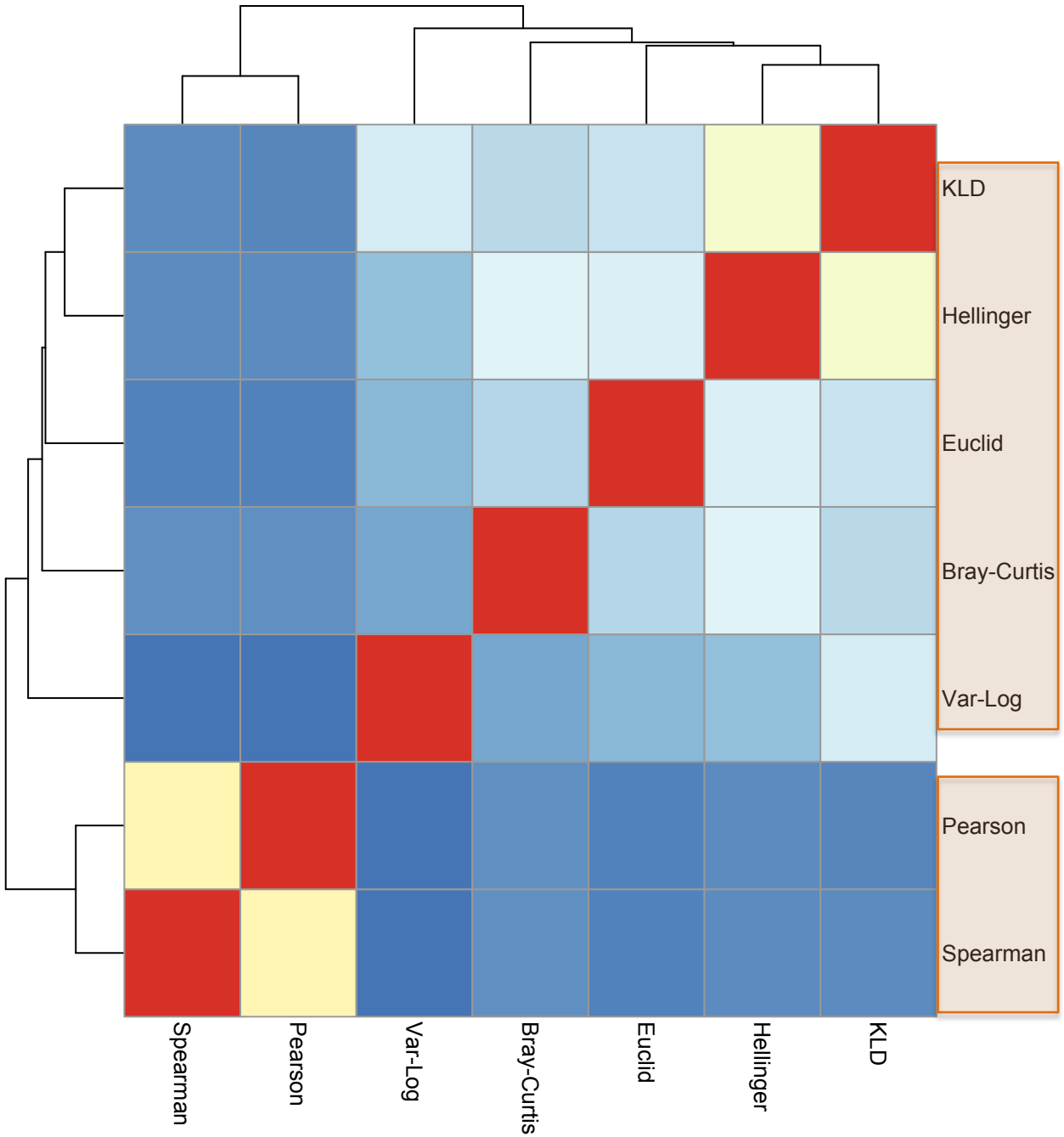
Pearson $$d(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}}$$

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.

Comparison of measures

Supplement



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

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.

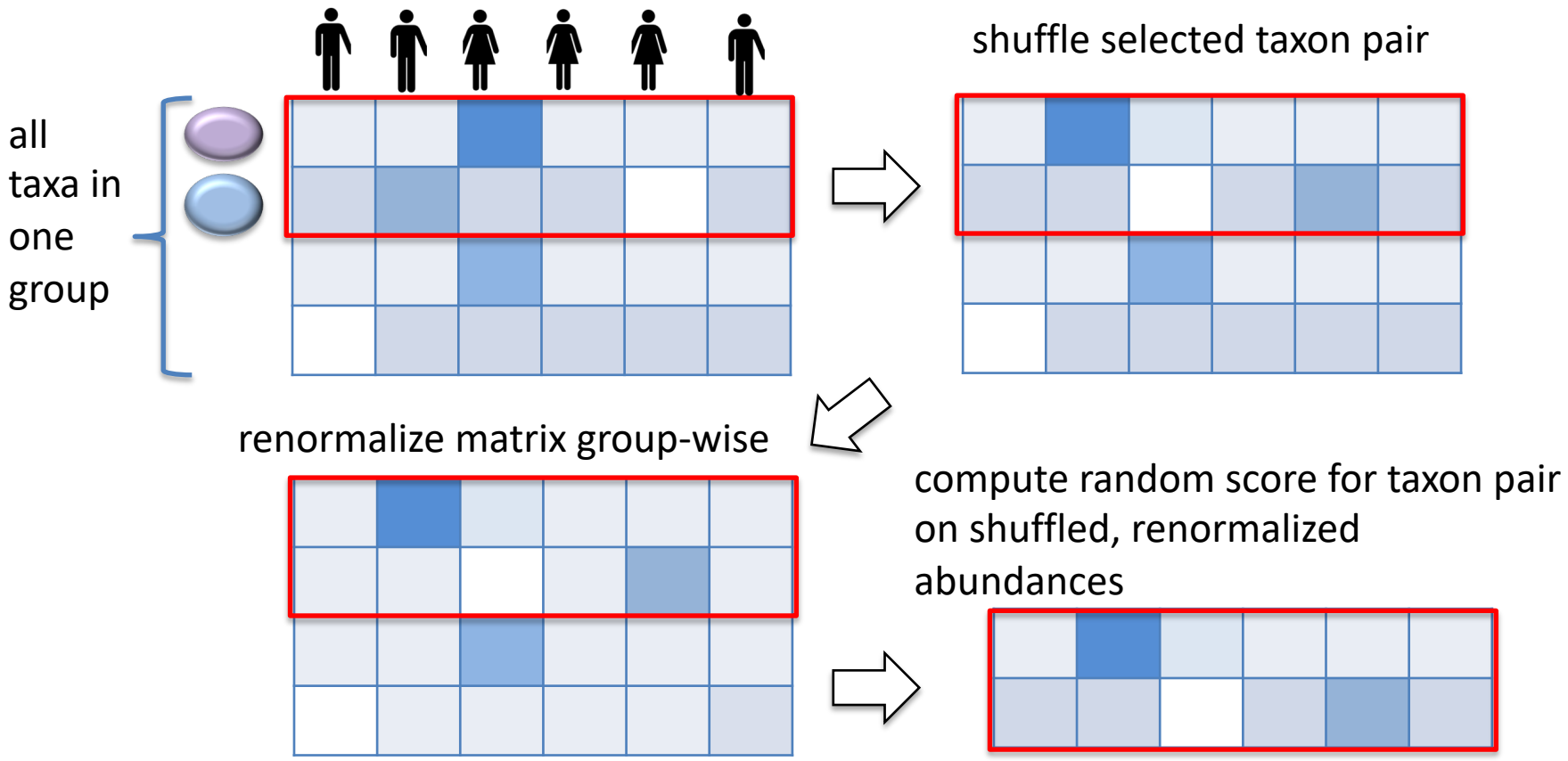
Fisher's method is biased by correlated association measures. This bias is taken out by Brown's p-value merging method.

Supplement

CoNet: ReBoot

Supplement

permutation with renormalization (**ReBoot**)

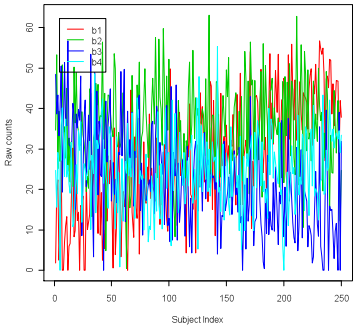


Fah Sathirapongsasuti

CoNet: ReBoot II

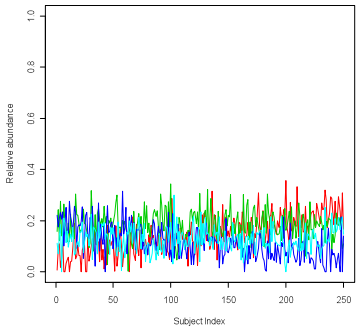
- Permutation test: removes correlation, but also any bias due to compositionality
- Permutation with **renormalization**: shifts null distribution

raw data



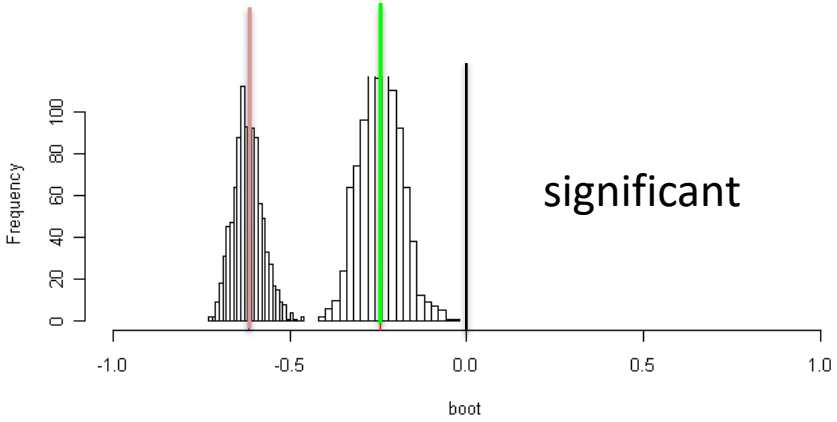
true anti-correlation between b1 and b3

normalized data

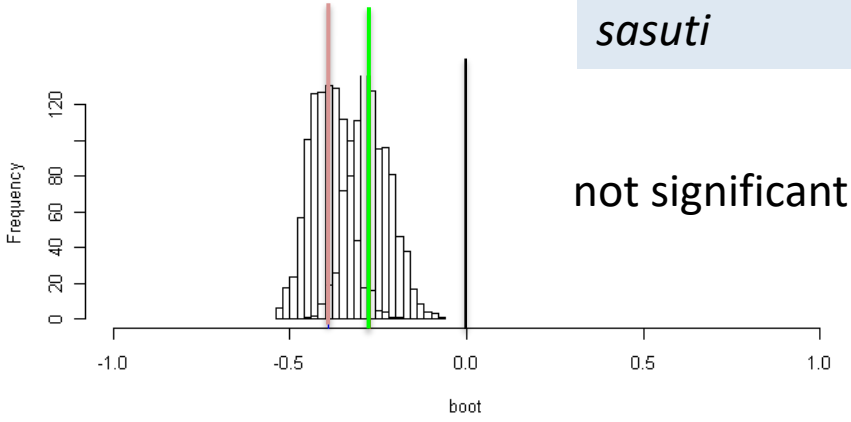


spurious correlation between b2 and b4 introduced by normalization

bootstrap distribution mean
renormalized permutation distribution mean



significant



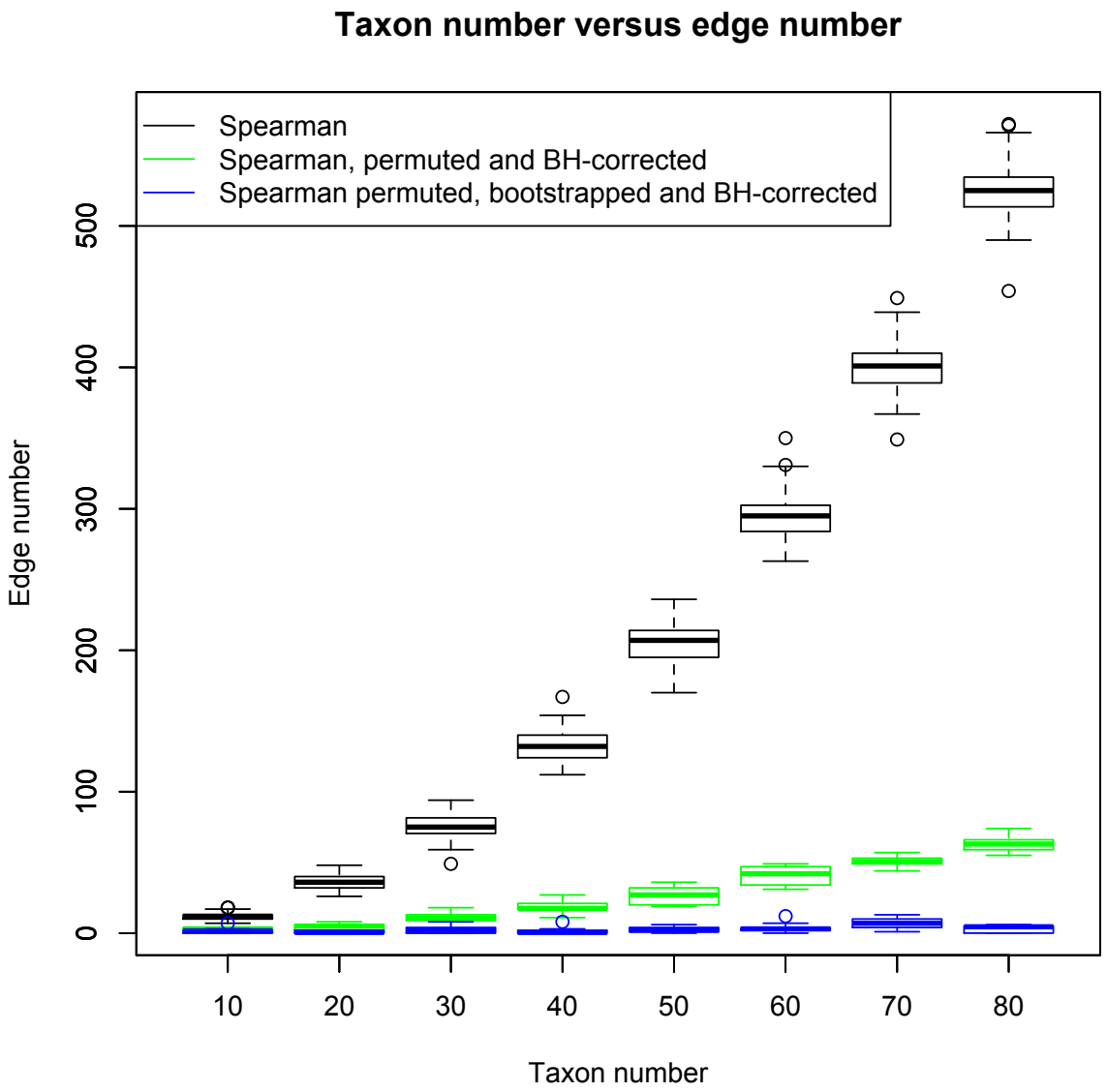
not significant

Fah Sathirapong-sasuti

Supplement

CoNet's assessment of significance reduces number of false positives

Supplement



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

(matrix not normalized, permutation carried out without renormalization)

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

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

δt: discrete time step
$X_i(t)$: abundance of target species *i* at time point *t*
$\langle x_j \rangle$: steady state abundance of species *j* (carrying capacity)
$\eta_i(t)$: log-normal noise
a_{ij}: interaction coefficient between taxa i and j

For $\eta_i(t) = 1$ (no noise) and $\delta t \to 0$, Ricker model reduces to generalized Lotka-Volterra in continuous form.

Supplement

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

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

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

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

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

$$\downarrow$$

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

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

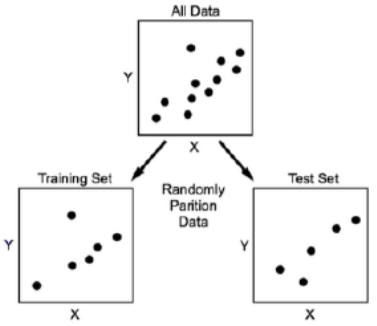
$\eta_i(t)$: 1, δt: 1 (no noise, smallest possible time step)

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

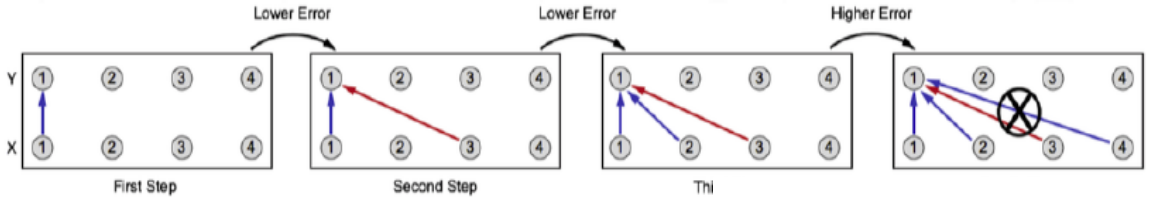
Supplement

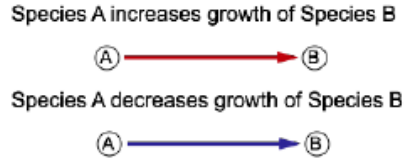
LIMITS – work flow

- 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**





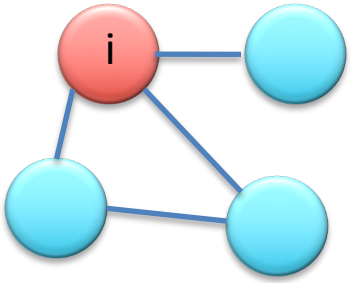
- Repeat data splitting and interaction matrix inference a number of times and report the median (**bootstrap**)

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

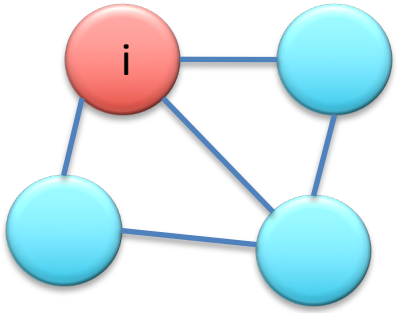
Supplement

Examples of network properties

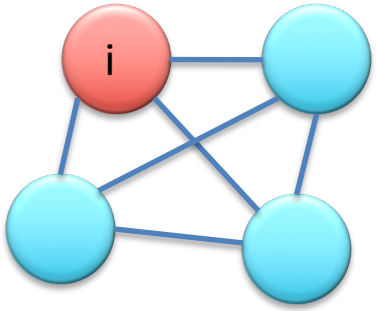


k=3
n=1
C_i = 1/3

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