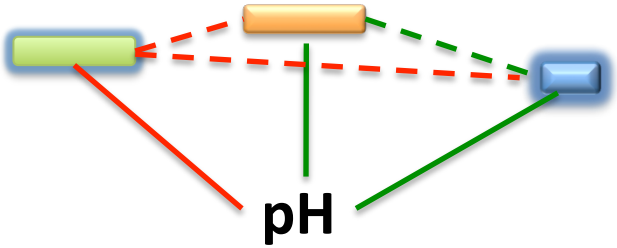
Karoline Faust
Raes lab

Challenges in Microbiome
Data Analysis Workshop
25th of February 2016

On the removal of environmentally driven microbial associations



Ecological interactions between microorganisms

Introduction

Predator/parasite (win-loss)

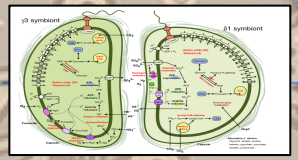
Commensalism (win-neutral)

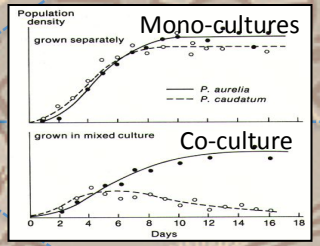
Amensalism (neutral-loss)

Mutualism (win-win)

Competition (loss-loss)

Prey/host (loss-win)



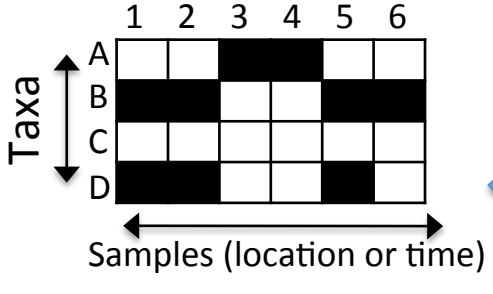


Adapted from Lidicker, W.Z. BioScience 29, 475-477, 1979.

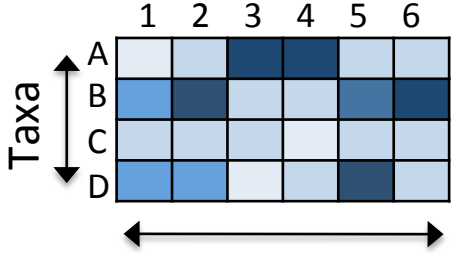
Microbial network inference

INPUT

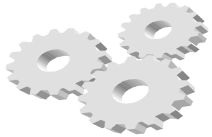
Presences/absences



Abundances



NETWORK INFERENCE

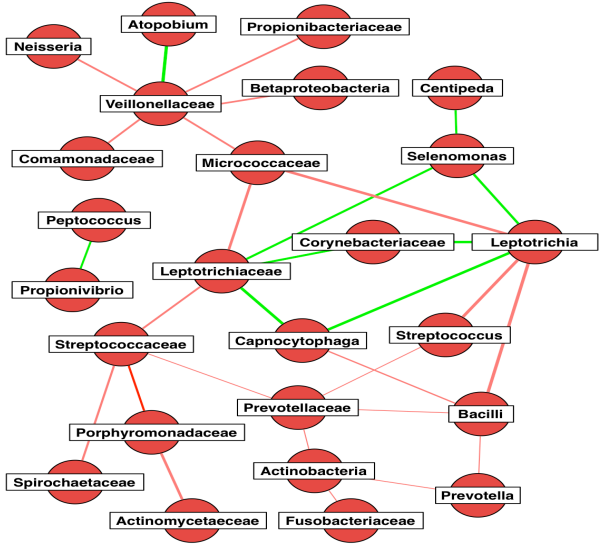


SparCC
CoNet
LSA
SPIEC-EASI
REBACCA
LIMITS
...

OUTPUT

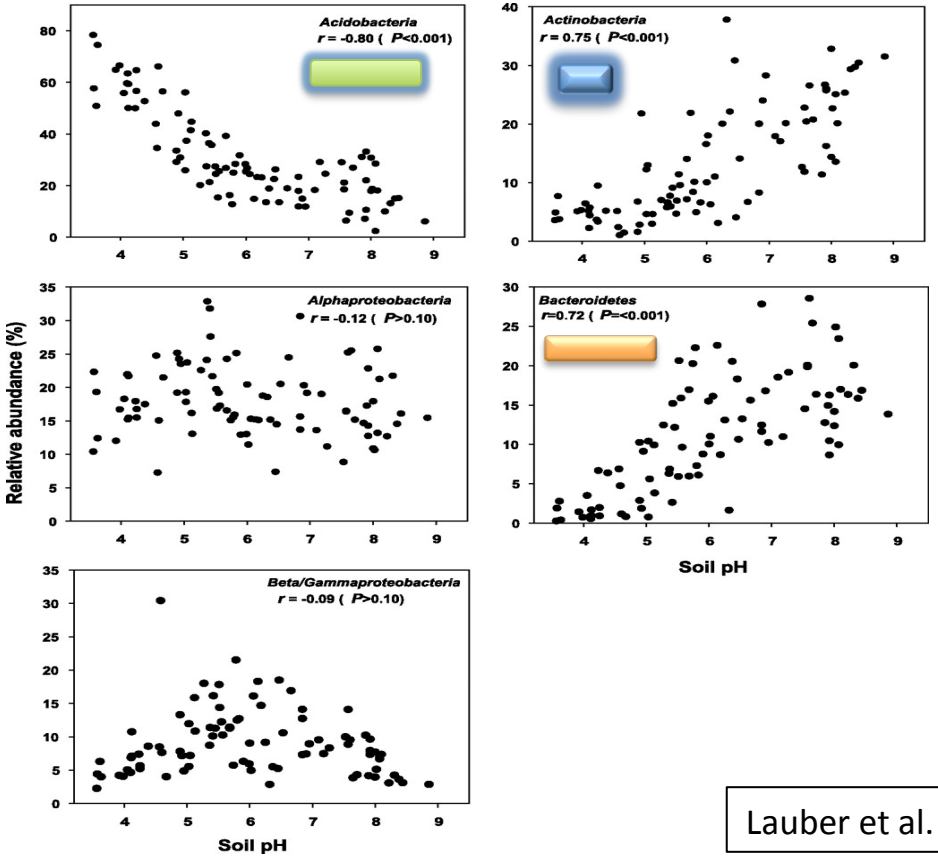
Microbial association network

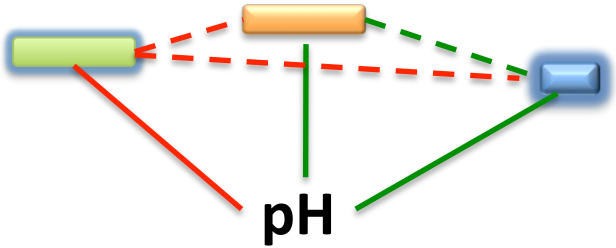
Nodes: taxa (OTUs, genera, ...)
Edges: associations between taxa



The problem: environmentally driven indirect taxon edges

Introduction



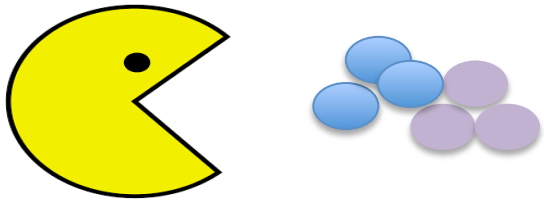


associations can be inferred due to a common response of two microbial taxa to an environmental factor

Lauber et al. Appl. Environ. Microbiol. 75, 5111-5120, 2009.

Taxon-driven indirect taxon edges

- A taxon can induce an association between two other taxa (e.g. a grazer feeding on two prey species)
- We assume that taxa alter the environment on a different time scale than taxa alter other taxa (approximation: environment influences taxa, but not vice versa)
- No such simplification possible for taxon-driven taxon edges



Strategies to remove environmentally driven edges

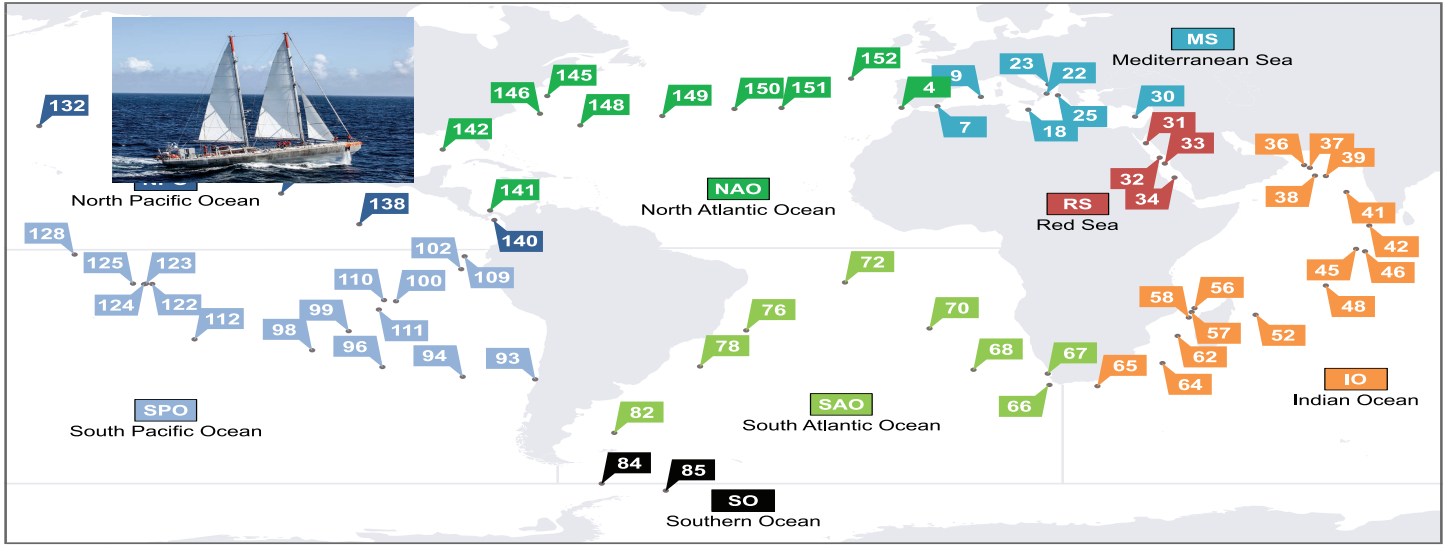
Introduction

- **"Associations" strategy**: compute and analyze associations between taxa and environmental factors
  - R package WGCNA (gene expression data): identify clusters in gene-wise correlation/dissimilarity matrices and check whether a representative (eigen-gene) is correlated to a trait
- **"Residuals" strategy**: regress out environmental factors and compute associations in the residuals

Langfelder & Horvath (2008) BMC Bioinformatics 9, 559.

Example: TARA Oceans

- global marine expedition, >200 stations spanning 8 oceanic regions



Brum et al. Science 348, 1261498 (2015).
de Vargas et al. Science 348, 1261605 (2015).
Lima-Mendez et al. Science 348, 1262073 (2015).
Sunagawa et al. Science 348, 1261359 (2015).
Villar et al. Science 348, 1261447 (2015).
Pesant et al. Scientific Data 2, 150023 (2015).

Main contributor in the Raes lab:
Dr. Lima-Mendez

TARA Oceans: Overview

Introduction

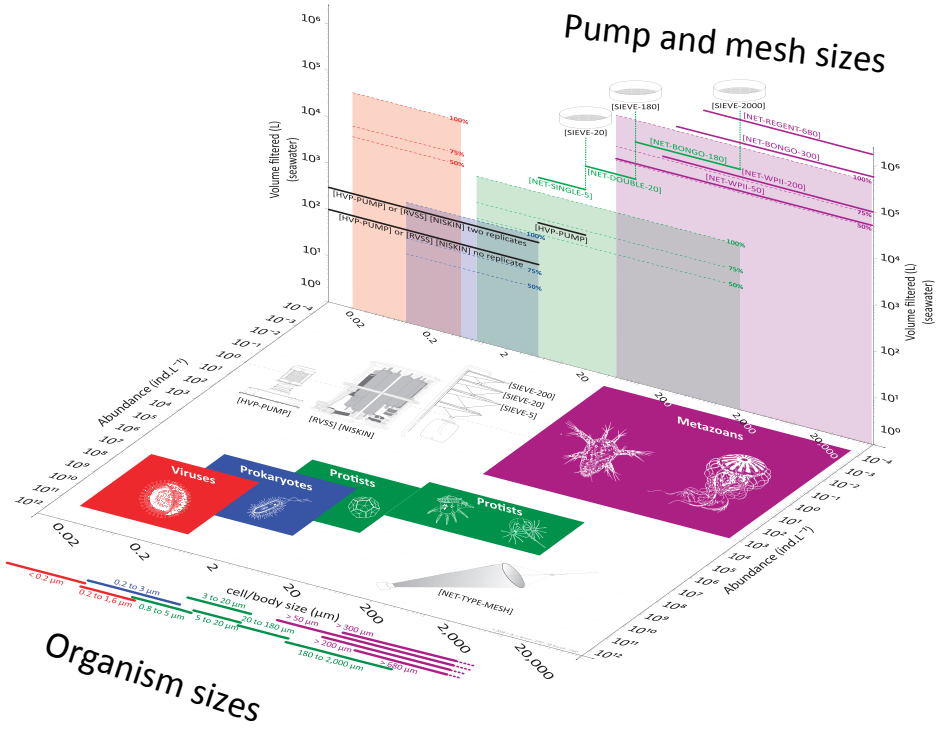


Image taken from Pesant et al. Scientific Data 2, 150023 (2015).

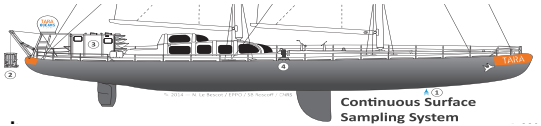
- 4 eukaryotic cell size fractions + bacteria, giruses and viruses
- 2 depths (SUR = surface, DCM = deep chlorophyll maximum)
- 16S bacterial OTU abundances (Illumina $_{mi}$tags)
- 18S V9 eukaryotic OTU abundances (Illumina)
- Viral contigs assembled from Illumina reads with SOAPdenovo and clustered

$_{mi}$tags: Logares et al. Environmental Microbiology 16(9), 2659-2671 (2014).
Viral contigs: Brum et al. Science 348, 1261498 (2015).

TARA Oceans: Environmental data

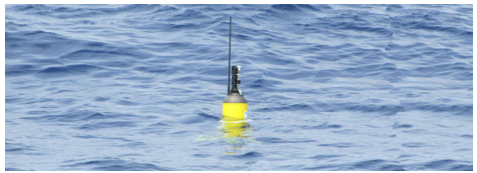
Introduction

Rosette vertical sampling system (Niskin bottles and sensors)



Continuous surface sampling system

Satellite data



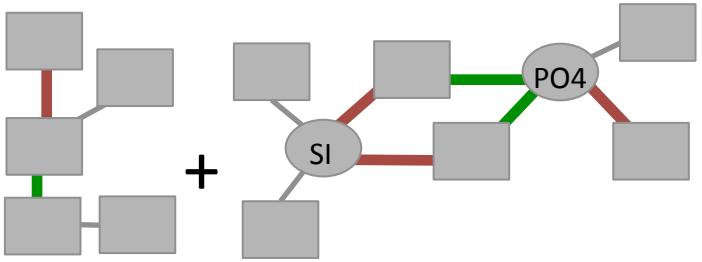
Argo profiling float network

Environmental factors in TARA:
- Temperature
- Depth
- Pressure
- Salinity
- Oxygen
- Nitrogen
- Silicon
- Phosphate
- Chlorophyll
- Particle abundance (beam attenuation)
- Mean depth of mixed layer (MLD)
…

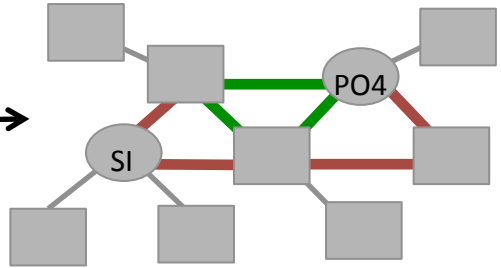
Remove environmentally driven taxon edges in triplets...

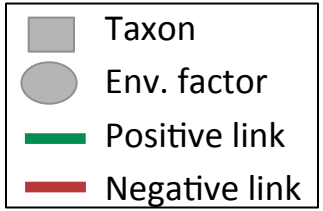
Methods - associations

Step 1: **Taxon-environment** and **taxon-taxon** network construction with CoNet

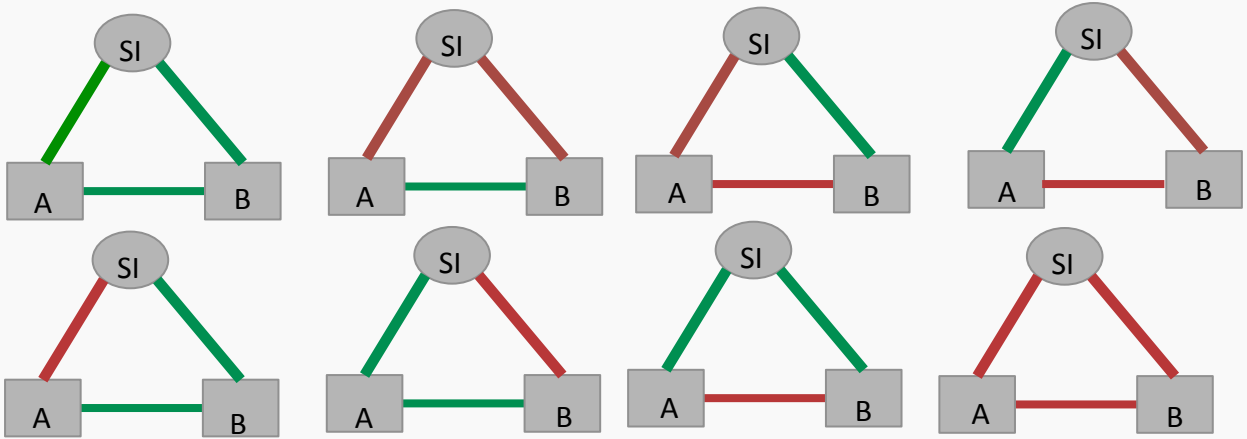


Step 2: Network merge and identification of **environment-taxon triplets**





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



Sign patterns consistent with **indirect** taxon edge

Sign patterns consistent with a **direct** taxon edge

... with the interaction information

Methods - associations

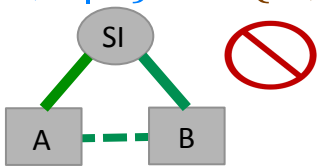
Step 3: Detection of environment-driven indirect edges with **interaction information (II)**

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

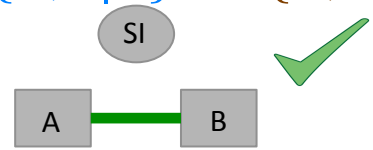
Conditional mutual information

Mutual information

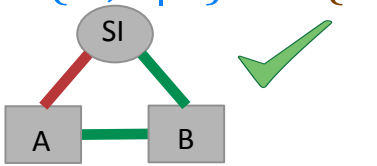
Negative II: redundancy $CI(X,Y \mid Z) < MI(X,Y)$



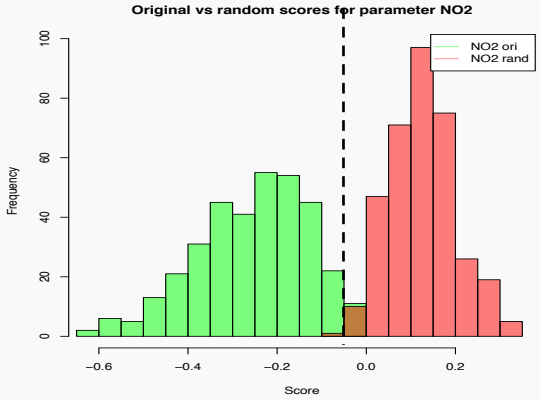
Zero II: no interaction $CI(X,Y \mid Z) = MI(X,Y)$



Positive II: synergy $CI(X,Y \mid Z) > MI(X,Y)$



Significance of interaction information for specific environmental factor:



Distribution of interaction information in triplets involving environmental factor.

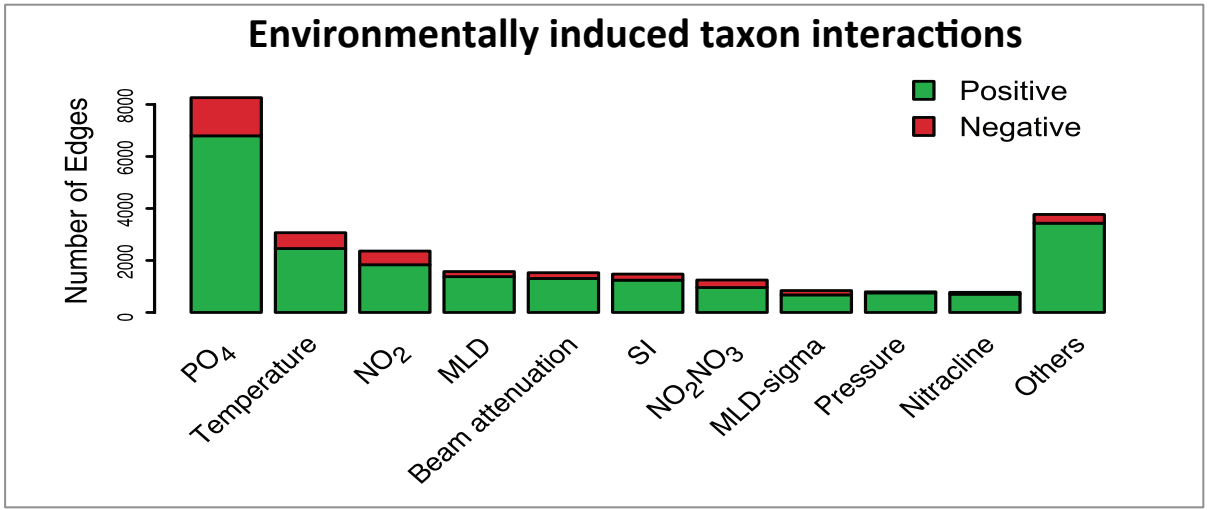
Distribution of interaction information in triplets with permuted environmental factor.

Factor-specific threshold for interaction information: Below 5% quantile of random interaction information distribution and below zero.

TARA Oceans: Environmentally driven taxon edge statistics

Results - associations

- Phosphate induces the largest number of indirect taxon interactions in marine plankton
- Higher percentage of environmentally driven indirect taxon interactions for eukaryotic than for prokaryotic phyla



Removing environmentally driven taxon edges: residuals

Methods - residuals

- **Multivariate regression**: regress out environmental factors that influence taxon abundances and look for taxon covariances/correlations in the residuals

$$Y = XB + E \qquad E \sim N(0, R)$$

- Can be refined by introducing latent variables that capture missing predictors and by using sparse regression (selection of the most relevant environmental factors)

Wisz et al. (2012). The role of biotic interactions in shaping distributions and realised assemblages of species: implications for species distribution modelling. *Biological Reviews Camb Philos Soc.* 88, 15-30.

Warton et al. (2015). So Many Variables: Joint Modeling in Community Ecology. *Trends in Ecology & Evolution* 30, 766-779.

Example: residuals in action

Methods - residuals

$$y_i = 1 \; for \; z_{ij} > 0$$

$$y_i = 0 \; \text{for} \; z_{ij} \leq 0$$

Y: Taxon presence/ absence (wood-decaying fungal species on tree logs)

$$\mu_i = \underbrace{\beta_0}_{\text{intercept}} + \underbrace{\beta_1 x_{1,i}}_{\text{coefficients}} + \ldots + \beta_k x_{k,i}$$

$$z_{ij} = \mu_{ij} + \text{logit}(F[e_{ij}])$$

$$e_i \sim \text{N}(0, R)$$

R: Correlation matrix (association network)

X: k Environmental factors (tree species, diameter, decay stadium, ground contact, fall type, bark cover etc.)

F[]: cumulative density function
Logit(x) = log(x/(1-x))

Ovaskainen et al. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91, 2514-2521.

Example: residuals in action, cont'd

- 3 null models tested (without environmental factors, without species interactions, without both)

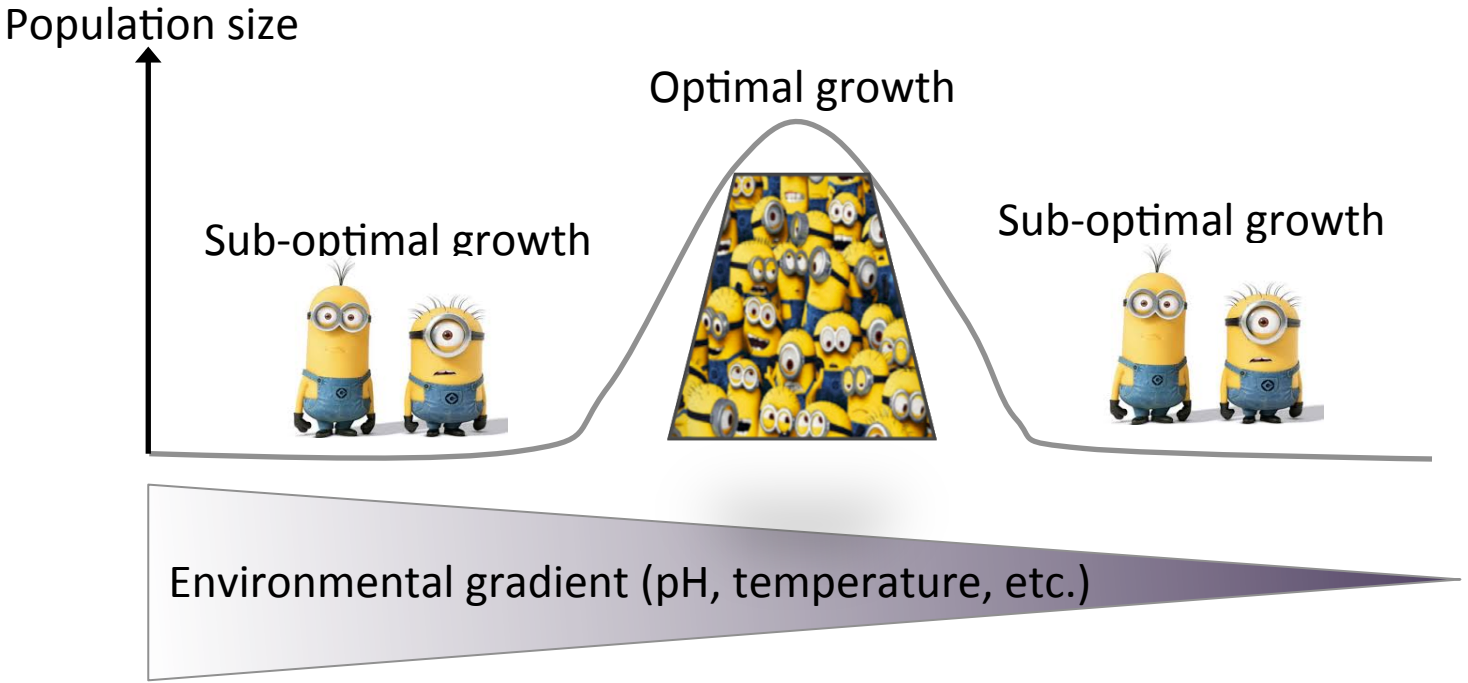
	correlation matrix set to identity	correlation matrix inferred
No environmental factors	$M1(I)$	$M1(R)$
environmental factors included	$M2(I)$	$M2(R)$

Considering environment and species interactions gives the best fit

Ovaskainen et al. (2010). Modeling species co-occurrence by multivariate logistic regression generates new hypotheses on fungal interactions. Ecology 91, 2514-2521.

The problem

- Organisms have growth optima, so they may not respond linearly to environmental factors



Discussion

How can we deal with non-linear responses to environmental factors?

Discussion

- Associations: use general measures of dependency such as mutual information (but mutual information needs a lot of data points and there is debate over its implementation)
- Residuals: measure response functions of species to environmental factors and select model accordingly

Fernandes & Gloor (2010). Mutual information is critically dependent on prior assumptions: would the correct estimate of mutual information please identify itself? *Bioinformatics* 26, 1135-1139.

Other considerations when removing environmentally driven indirect edges

Discussion

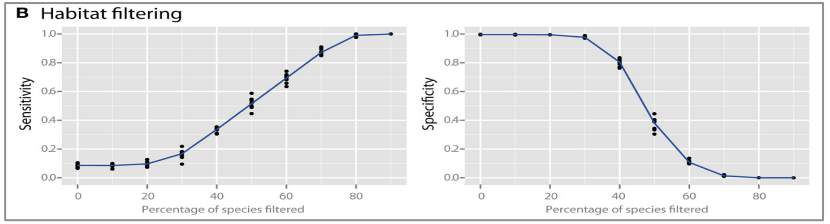
- Missing values
  - Associations: pairwise omission possible
- Accounting for sequencing depth differences
  - Associations: normalization/rarefaction, residuals: sequencing depth included in the regression model
- Compositionality
  - Associations: dedicated tools exist (REBACCA, SparCC)
  - Compositionality and multivariate regression?
- Over-fitting
  - With enough samples, cross-validation techniques can be applied to both strategies

Study design to the rescue: homogeneity

Discussion

- Berry & Widder: "[...]*when co-occurrence networks are used to infer putative interactions, samples should be drawn from similar environments in order to minimize the effects of habitat filtering*[...]."

  => Look at **only one habitat type**



Berry & Widder: Sensitivity and specificity versus number of habitat specialists

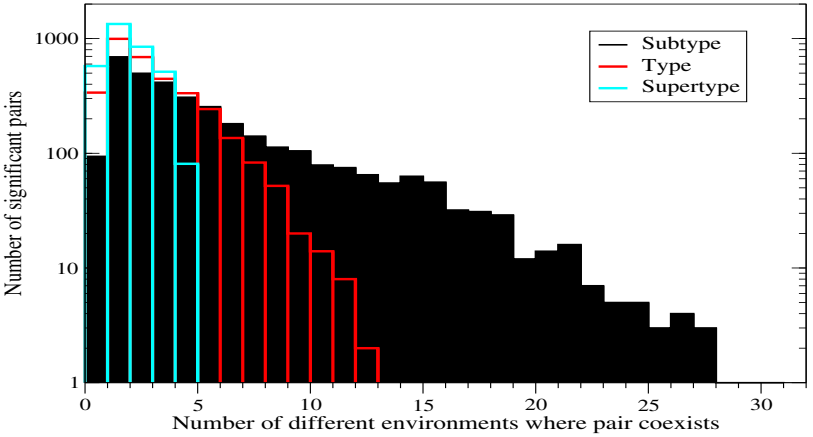
Berry & Widder (2014). Deciphering microbial interactions and detecting species with co-occurrence networks. Frontiers in Microbiology 5, 219.

Study design to the rescue: heterogeneity

Discussion

- Pascual-Garcia et al.: "*We found that 77% of the significantly aggregated pairs co- occur in samples from more than two different [habitat] subtypes, 60% from more than two types, and 57% from more than one supertype. These data support the view that most aggregations cannot be explained by habitat preferences.*"

  => Look at **cosmopolitans** occurring in **many different habitat types**



Pascual-Garcia et al.: Number of different habitats versus number of positive edges

Pascual-Garcia et al. (2014). Bacteria dialog with Santa Rosalia: Are aggregations of cosmopolitan bacteria mainly explained by habitat filtering or by ecological interactions? BMC Microbiology 14, 284.

Suggestions to reduce environmentally driven taxon edges in microbial networks

Discussion

- Collect many samples from a homogeneous environment or focus on cosmopolitans occurring in several different environments
- Cluster samples and check whether sample groups are driven by environmental factors
- Restrict network inference to samples in one group (stratify the data)

Validation of environmentally driven edge removal

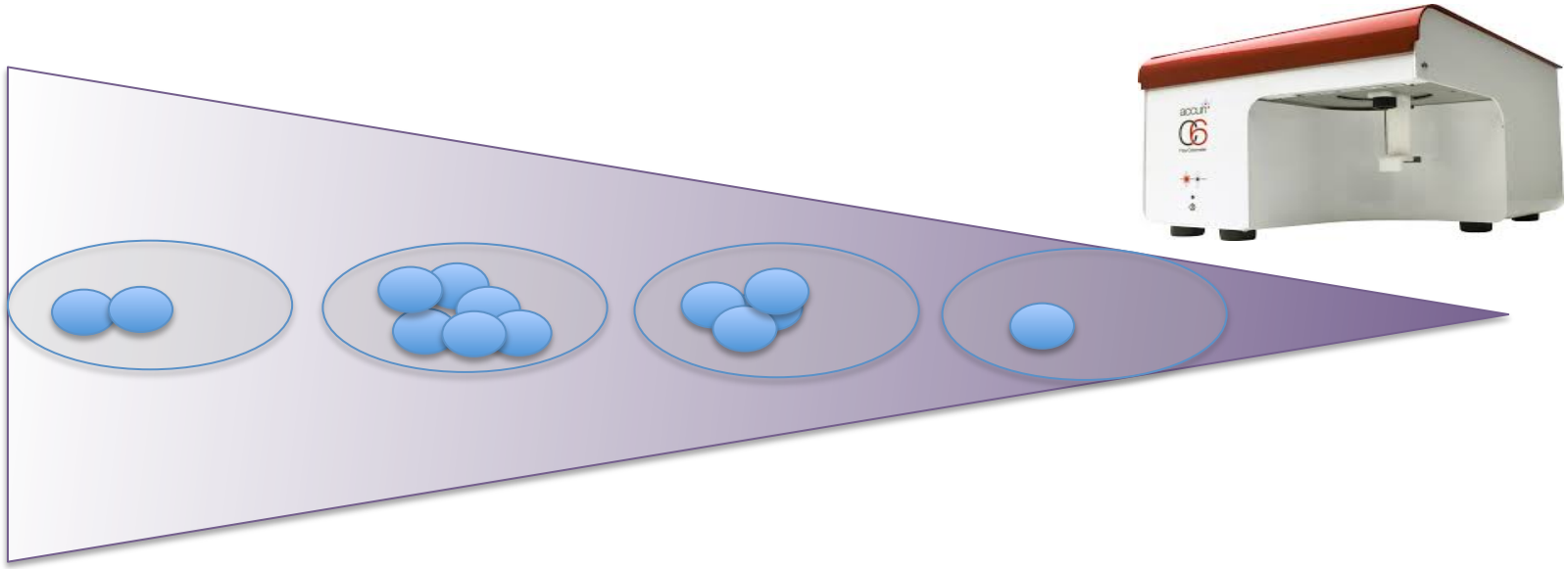
Outlook

- **In silico validation** using different models that integrate environment and species interactions as data generators
- Benchmark data set of **known microbial interactions** (to check whether ecological interactions are falsely removed)
- **In vitro experiments on synthetic microbial communities** exposed to varying environmental factors (time series with known ecological interactions for benchmarking)
- **Mesocosm experiments** exposing well-known microbial communities to varying environmental factors (time series with many known ecological interactions for benchmarking)

Large-scale measurement of response functions

Outlook

- Need large-scale cultivation studies that vary growth conditions (pH, temperature) in rich medium in mono-culture and measure impact on growth



Acknowledgments

TARA team

Raes lab
Gipsi Lima-Mendez
Samuel Chaffron
Youssef Darzi
Jun Wang

De Vargas Lab
Nicolas Henry
Johan Decelle
Sebastien Colin
Stephane Audic

Bork lab
Shinichi Sunagawa

Bowler Lab
Flora Vincent
Sabrina Speich
Lionel Guidi
Stephane Pesant
Lars Stemann
Gaby Gorsky

Sullivan lab
Simon Roux
J. Cesar Espinosa

Acinas Lab
Francisco Cornejo Castillo
Guillem Salazar
Marta Royo a.o.
Lucie Bittner
Fabrice Not
Patrick Wincker

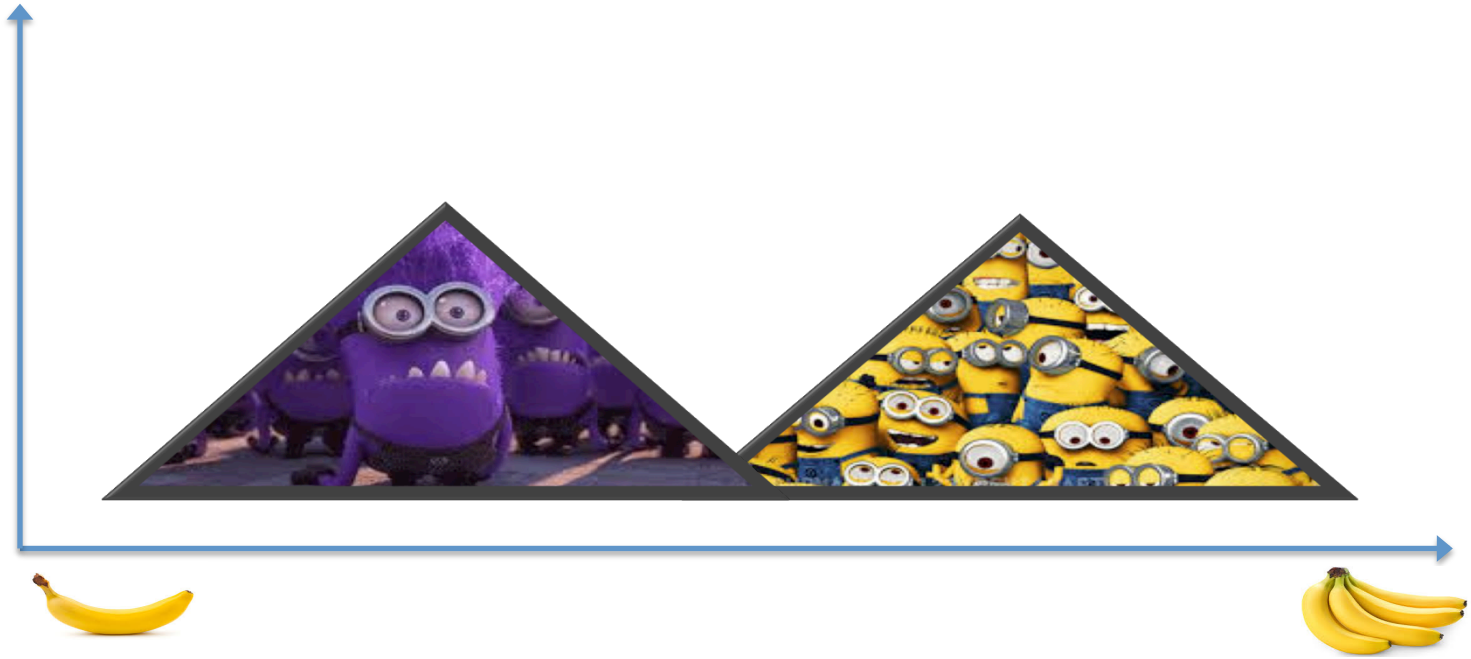
Tara SAB

KU Leuven/VUB
Caroline Souffreau
Luc De Meester
Fabrizio Carcillo
Gianluca Bontempi

Eric Karsenti & Steffi Kandels-Lewis
The Tara and Agnes B crew, coordinators

...and so many others!

Thank you



Appendix

- Pro residuals:
  - mathematical model describes how environmental factors influence taxon abundances
  - the combined effects of environmental factors are considered
- Contra residuals:
  - the mathematical model may be wrong (e.g. non-linear response functions to environmental factors modeled with linear regression)
  - Risk of over-fitting

Appendix

- TARA network construction settings for CoNet
  - Spearman and Kullback-Leibler dissimilarity (intersection enforced)
  - Permutation with renormalization (1000 iterations) and bootstrap (1000 iterations)
  - P-value per method and edge computed from both distributions
  - P-values of methods merged with Brown's method and multiple-testing corrected with Benjamini Hochberg (cut-off at 0.05)
- TARA false negatives due to removal of environmentally driven taxon edges: 1 out of 43 genus-level interactions whose partners are present in the input matrices

Appendix

- Edges linking taxa to phosphate: outlier for negative PO4-eukaryote edges at the surface – consequence of blooms?

Edge type (total node numbers in SUR and DCM taxon-environment union networks)	Positive SUR	Negative SUR	Positive DCM	Negative DCM
Prokaryotes (2,922 and 2,777)	186	286	98	191
Eukaryotes (4,334 and 3,502)	273	**1178**	383	411

Appendix

- Edges linking taxa to temperature: most positive edges to temperature at the surface

Edge type (total node numbers in SUR and DCM taxon-environment union networks)	Positive SUR	Negative SUR	Positive DCM	Negative DCM
Prokaryotes (2,922 and 2,777)	**422**	180	14	59
Eukaryotes (4,334 and 3,502)	**756**	99	29	94

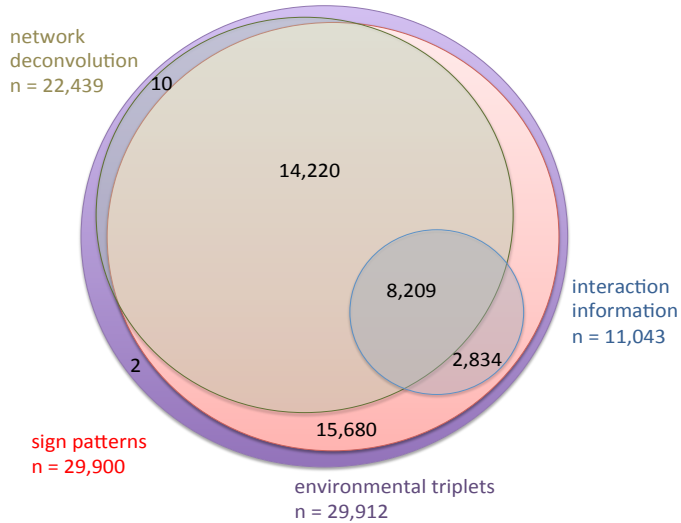
Appendix

- Edges linking taxa to NO2: more NO2-prokaryote edges at the surface than at DCM

Edge type (total node numbers in SUR and DCM taxon-environment union networks)	Positive SUR	Negative SUR	Positive DCM	Negative DCM
Prokaryotes (2,922 and 2,777)	**225**	**261**	20	62
Eukaryotes (4,334 and 3,502)	295	691	464	286

Appendix

- Comparison of environmentally-driven indirect taxon edge removal techniques applied to TARA data
  - Interaction information in full agreement with sign patterns indicative of an indirect edge and in partial agreement with network deconvolution



Feizi et al. (2013) Nature Biotechnology vol. 31, 726-731