Multivariate statistics and data visualization

Lecture 14.03.2019

Outline

- Diversity concepts
- Alpha diversity indices and rarefaction curves
- Rare biosphere
- Beta diversity and ordination methods
- Compositionality
- Hypothesis testing
- Co-occurrence networks
- Data visualization in R
    - Getting your data into shape
    - R data and object types
    - R errors

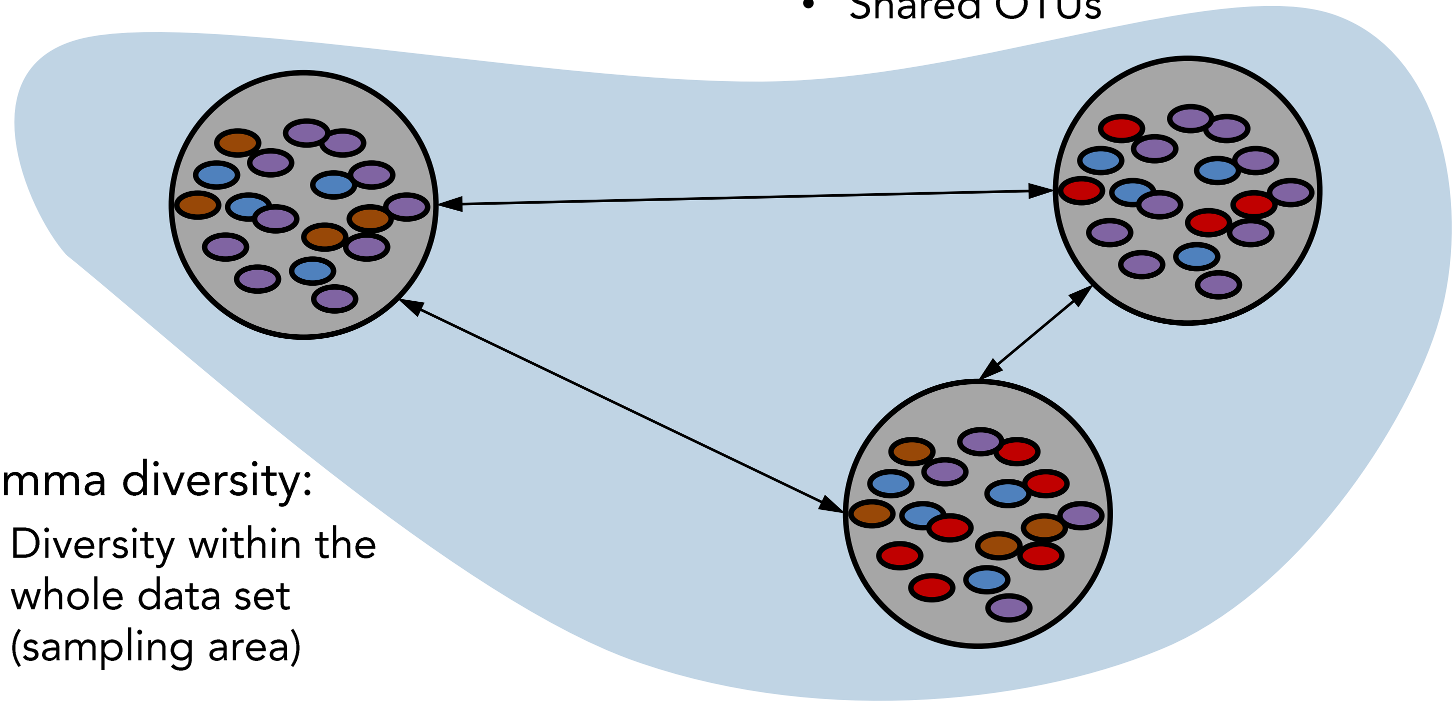
Diversity concepts

Alpha diversity:

- Diversity within one sample
- Richness
- Evenness

Beta diversity:

- Diversity between samples (comparison)
- Dissimilarity
- Shared OTUs

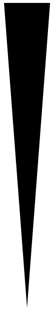


Gamma diversity:

- Diversity within the whole data set (sampling area)

Alpha diversity

- Alpha diversity indices:
  - Chao1
  - ACE
  - Richness
  - Shannon
  - Inverse Simpson



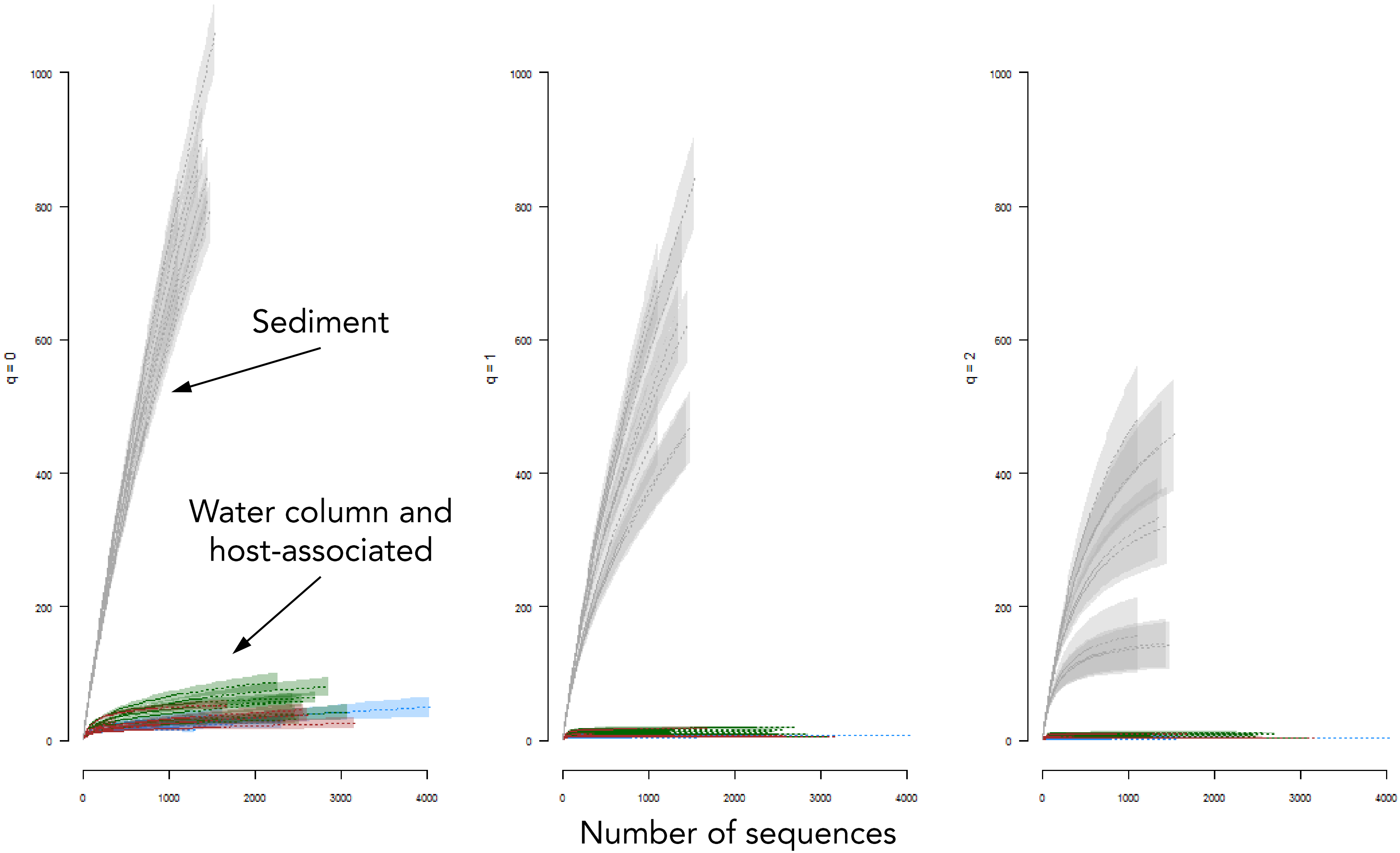
Influence of rare OTUs

- Unifying concept: Hill numbers
  - Richness (q = 0)
  - Exponential Shannon (q = 1)
  - Inverse Simpson (q = 2)

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^{q} \right)^{1/(1-q)} $$

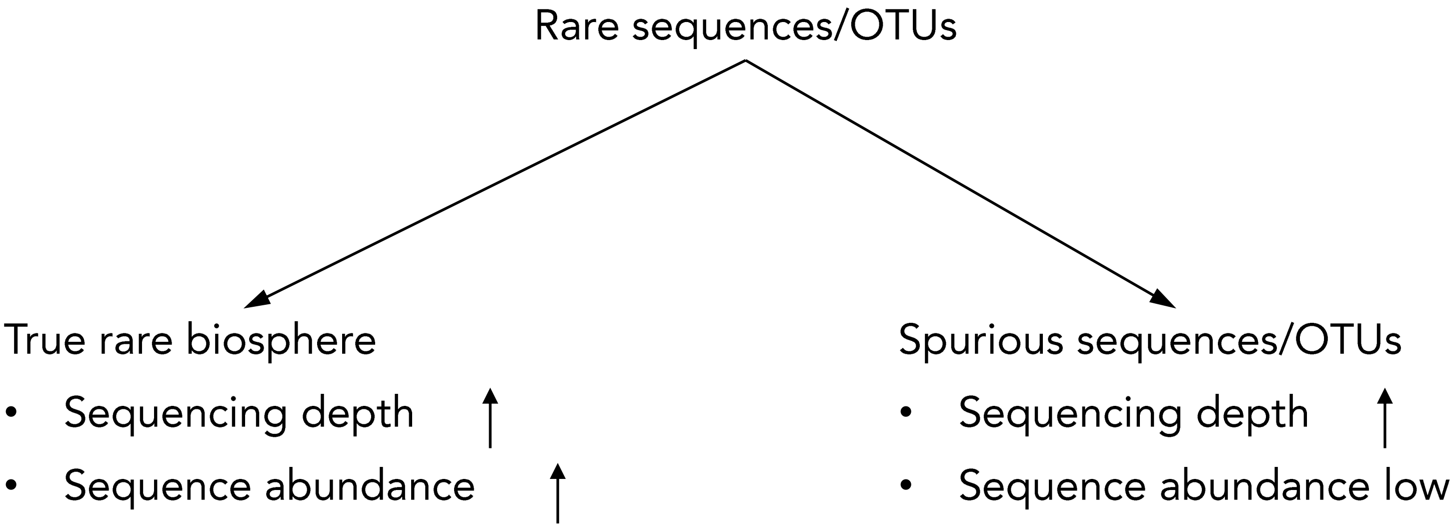
- Sequencing depth and rare biosphere?
  - Subsample sequences to equal library sizes
  - Correct the number of singletons per sample
  - Use rarefaction curves to estimate covered diversity by available sequencing depth

Rarefaction curves



Rare biosphere

- No technique can provide results beyond detection limit and measurement uncertainty
- Depending on error rate and sequencing depth



Rare sequences/OTUs

True rare biosphere

- Sequencing depth ↑
- Sequence abundance ↑

Spurious sequences/OTUs

- Sequencing depth ↑
- Sequence abundance low

- Remove rare OTUs (e.g. singletons, doubletons) from your data set

Beta diversity

Distance vs. dissimilarity

- Community (dis)similarity between samples

	OTU1	OTU2	OTU3	OTU4
S1	14	2	14	14
S2	10	14	0	8
S3	0	5	0	2
S4	0	0	1	0

presence/ → absence

	OTU1	OTU2	OTU3	OTU4
S1	1	1	1	1
S2	1	1	0	1
S3	0	1	0	1
S4	0	0	1	0

Asymmetrical vs. symmetrical

Bray-Curtis vs. euclidean

Jaccard

	S1	S2	S3	S4
S1	0			
S2	0.5	0		
S3	0.8	0.6	0	
S4	1.0	1	1	0

	S1	S2	S3	S4
S1	0			
S2	19.8	0		
S3	23.3	14.7	0	
S4	23.8	19	5.5	0

	S1	S2	S3	S4
S1	0			
S2	0.25	0		
S3	0.5	0.33	0	
S4	0.75	1	1	0

- Zeros in ecology: Is this species really not there or did we just not find it?

→ double zeros not relevant

Ordination

- Visualization of a multidimensional matrix in a reduced set of dimensions
- E.g.: PCA, PCoA, NMDS

Principal component analysis	Principle coordinate analysis	Non-metric multidimensional scaling
• Continuous environmental data	• Species abundance data	• Species abundance data
• Metric ordination based on euclidean distances	• Metric ordination based on any kind of distance/dissimilarity measure	• Non-metric ordination based on any kind of distance/dissimilarity measure
• Create new axes (principal components) along direction of highest variability (N_{PC} = $N_{variables}$)	• Create new axes (principal components) along direction of highest variability (N_{PC} = $N_{variables}$ - 1)	• Show maximum variation in 2 (or 3) dimensions

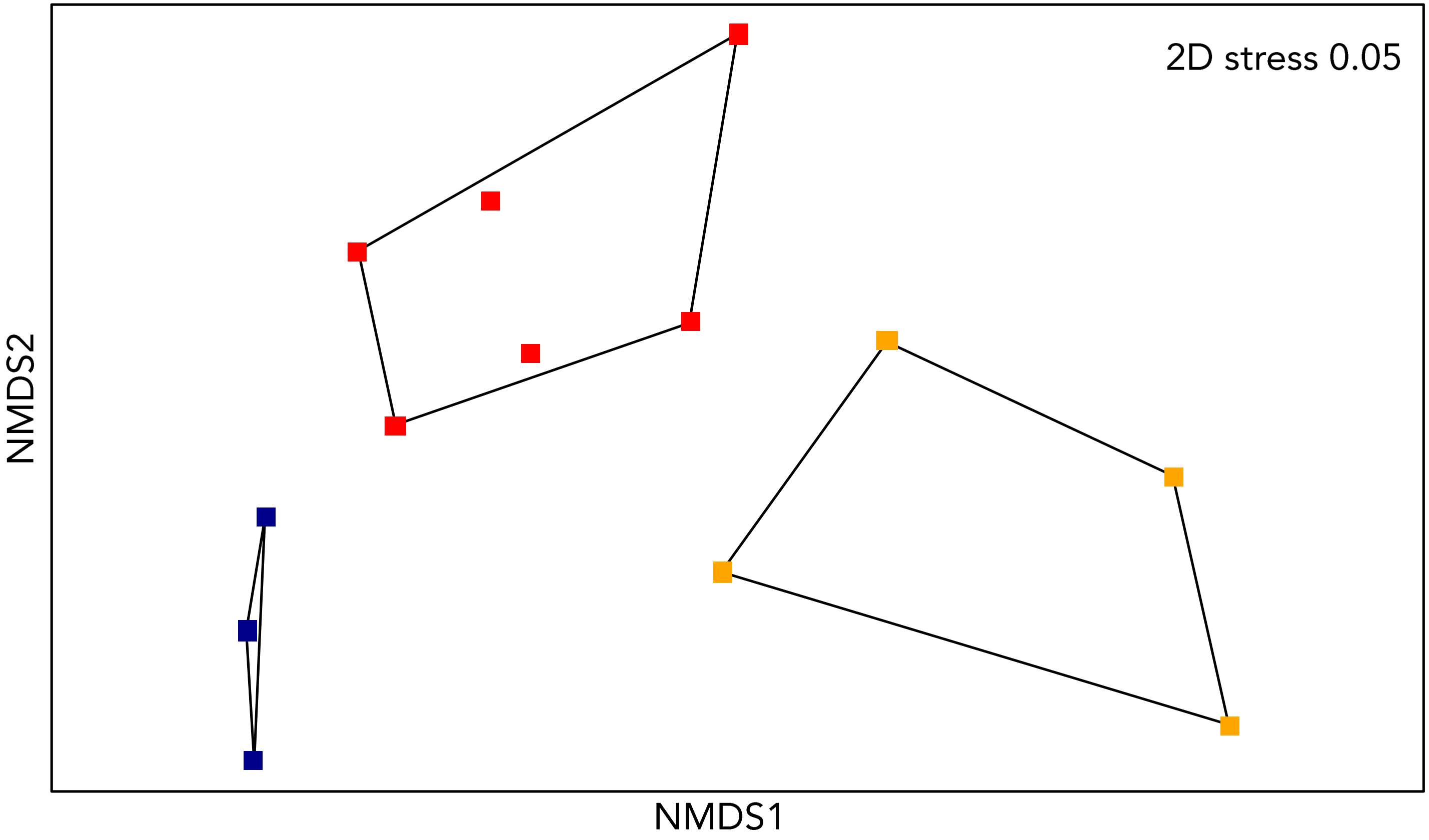
More information:

GUide to STatistical Analysis in Microbial Ecology

http://mb3is.megx.net/gustame

Ordination

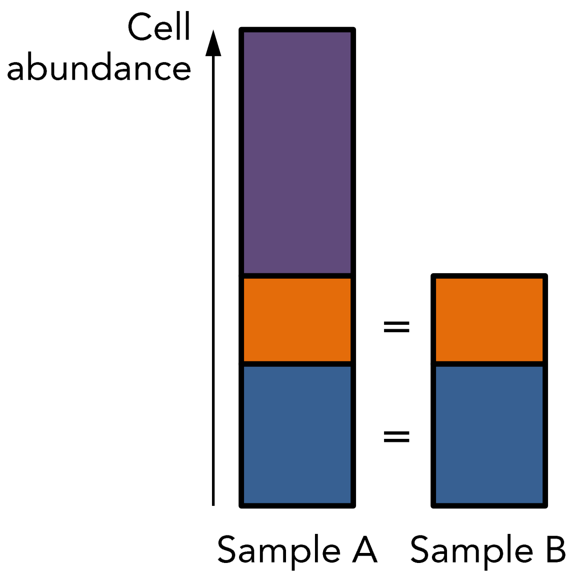
- Visualization of a multidimensional matrix in a reduced set of dimensions
- E.g.: PCA, PCoA, **NMDS**



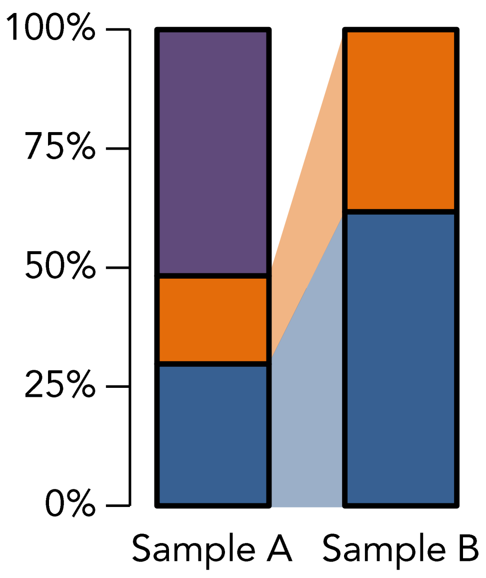
Compositionality

- OTU ‘abundances‘ are not independent, but proportions of a whole

Absolute abundance



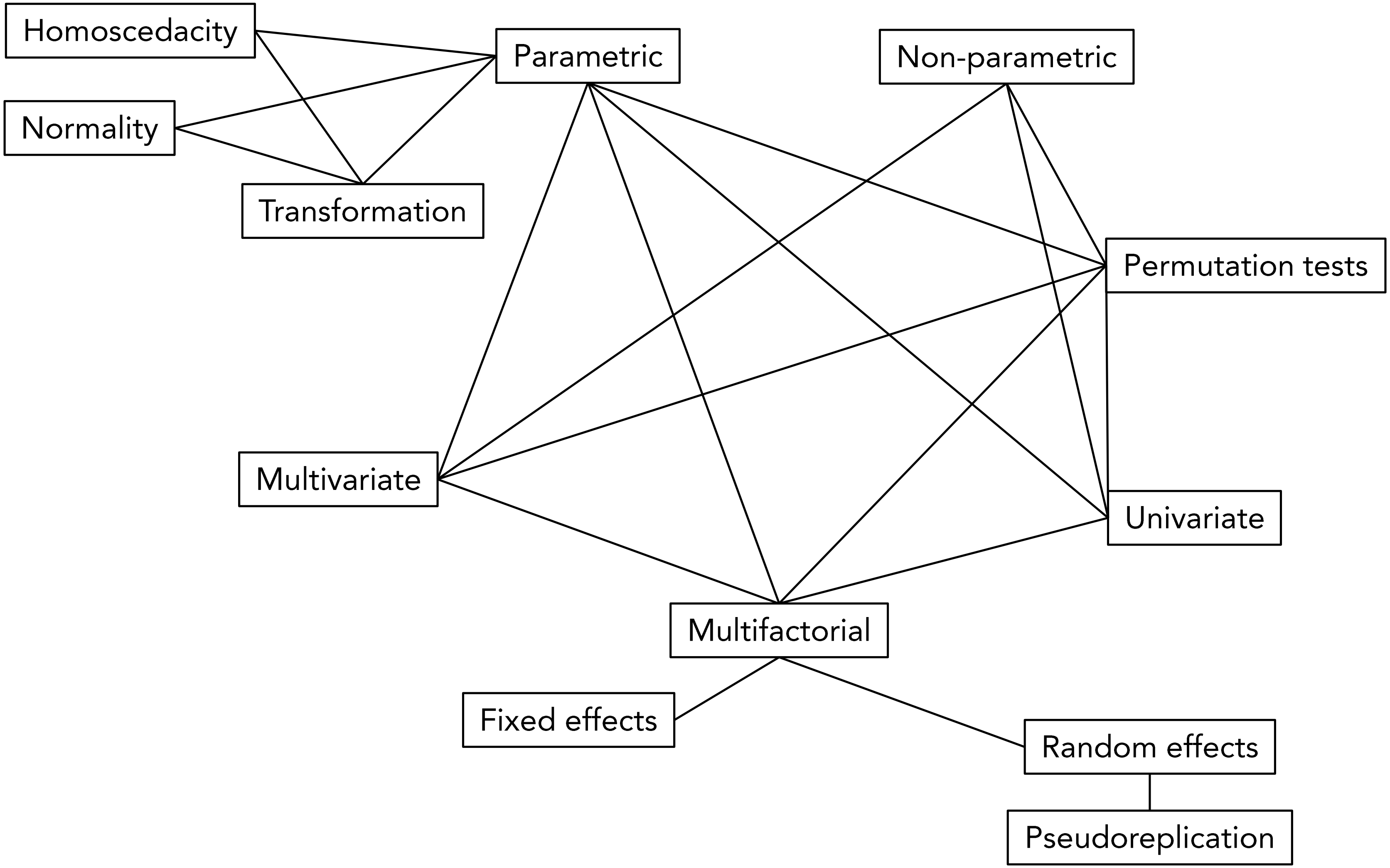
Sequence proportion



- Centered log-ratio transformation (clr): $\log(x_i) – \log(n\sqrt{product(x_1 \ldots xn)})$

Hypothesis testing

Which test to use…



Mixed effects models

- Extension of GLMs
- Additional feature: include random effects

- GLM: $$F = \frac{explained\ variation}{unexplained\ variation} = \frac{SS_{fixed}/dffix_{ed}}{SS_{error}/dferr_{or}}$$

- GLMM: $$F = \frac{explained\ variation}{unexplained\ variation} = \frac{SS_{fixed}/dffix_{ed}}{SS_{random}/dfran_{dom}}$$

- Example: repeated measurements
  - 3 treatments x 10 replicates x 3 measurements = 90 values
  - df_{fixed} ~ number of treatments
  - df_{error} ~ total sample size (without random factor) ←
  - df_{random} ~ number of levels in random factor

df_{error} ↑

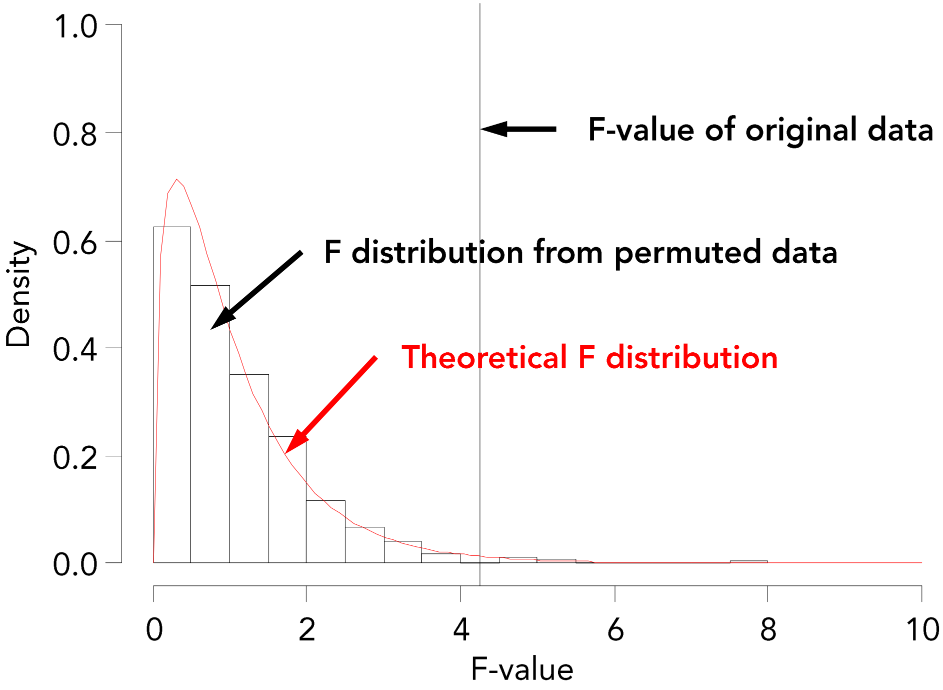
denominator ↓

F ↑

p ↓

Permutation tests

- Create your own theoretical distribution of the test statistic
- Randomly reshuffle the response variable



F-value of original data

F distribution from permuted data

Theoretical F distribution

- Implemented in R as default for several tests (mostly multivariate tests)

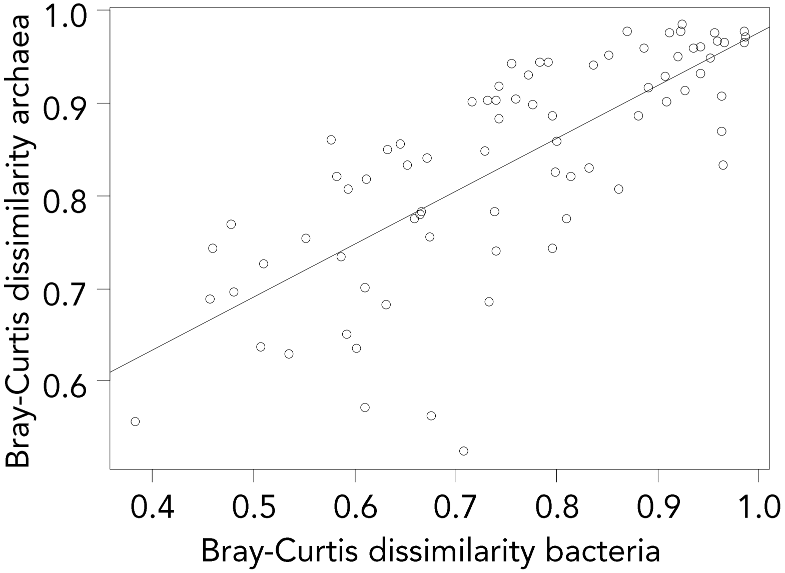
Testing for patterns in community composition

Analysis of similarity (ANOSIM)	Non-parametric multivariate ANOVA (PERMANOVA)	Redundancy analysis (RDA)
• Overlap/separation of communities • Unifactorial	• Indirect assessment of effects • Multifactorial	• Constrained ordination • Direct assessment of maringal effects
• Non-parametric • Based on ranked dissimilarities	• Non-parametric • Based on ranked dissimilarities	• Parametric • Linear technique
	• ANOVA-like output • Explained variation	• ANOVA-like output • Explained variation
• Based on permutation tests	Based on permutation tests	Based on permutation tests

Comparing patterns in community composition

Mantel test

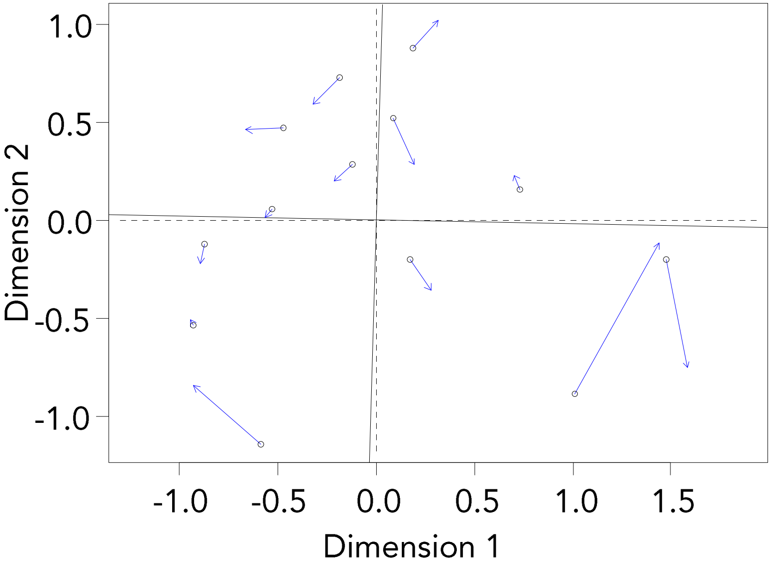
- Correlation of dissimilarity matrices
- Comparison based on all variation



Mantel statistic r: 0.7511
Significance: 0.001

Procrustes test

- Correlation of ordination objects
- Comparison based on the majority of the variation



Procrustes SS: 0.1266
Correlation (symmetric rotation): 0.9346
Significance: 0.001

Differential OTU proportions

- Testing statistical differences between environmental conditions **for each OTU**
- Compositionality correction: clr-transformation
- P-value correction:

$$FWER = 1 - (1 - alpha)^{n}$$

Number of comparisons

Family-wise error rate

Significance threshold per comparison

n	FWER
1	0.05
3	0.14
1000	~ 1

- Implementation: ALDEx2 (http://www.microbiomejournal.com/content/2/1/15, https://github.com/ggloor/ALDEx2)

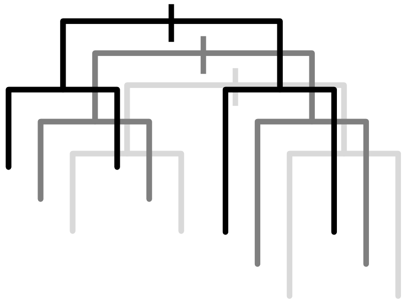
Exclude rare OTUs

Random forests

- Is it possible to predict a contextual parameter based on the community composition?
- Which OTUs are most important for a correct prediction?

Exclude rare OTUs

Random Forest model → Error rate (cross-validation) → Significance → Interpretation



Multiple decision trees calculated for a subset of the data (bootstrap)

Confusion matrix:

	group1	group2
group1	10	0
group2	1	9

Out-of-bag error:

- Internal validation based on samples outside bootstrap data sets

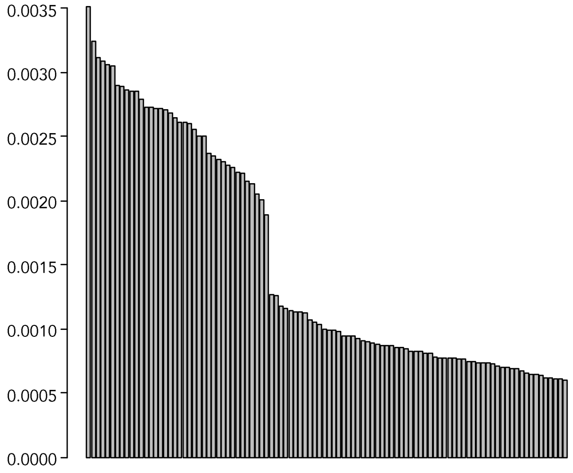
Leave-one-out cross-validation:

- Train model with n-1 sample n times
- Test model with left-out sample

Permutation tests

Importance scores

- Decrease in model accuracy if features (OTUs) are removed



Co-occurrence networks

Exclude rare OTUs

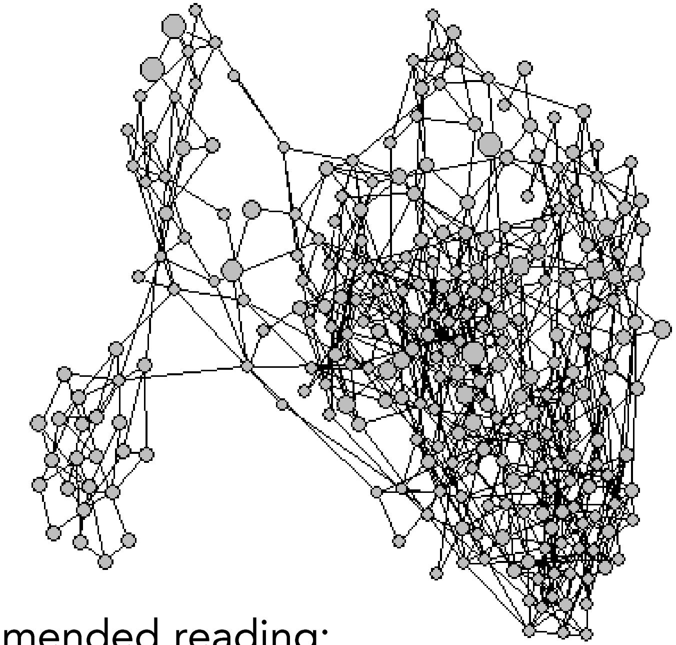
- Problems: sparcity and compositionality

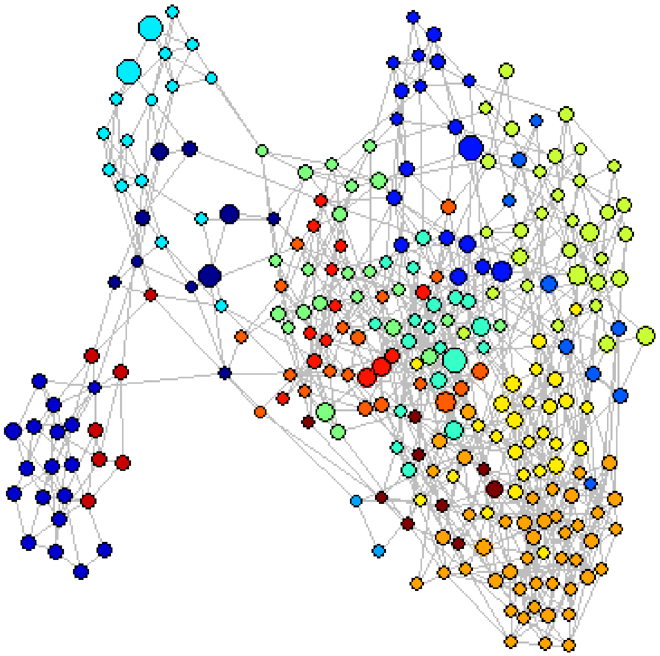
Network inference:

→ SPIEC-EASI: Sparse InversE Covariance estimation for Ecological Association and Statistical Inference

Module detection:

→ Louvain clustering





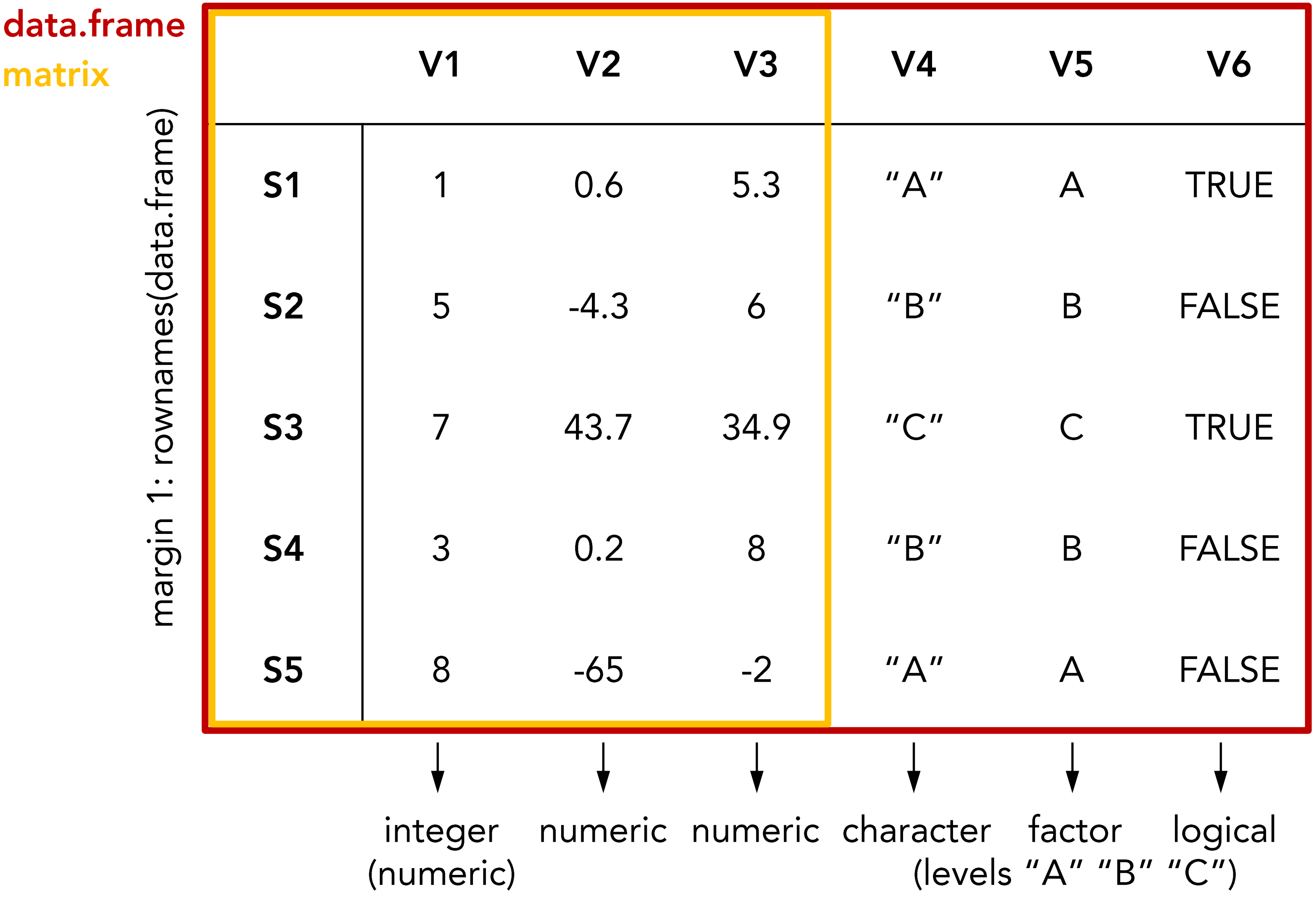
Recommended reading:

- Röttjers & Faust (2018) FEMS Microbiology Reviews 42:761-780. doi: 10.1093/femsre/fuy030
- Chafee et al. (2017) ISMEJ 12:237-252. doi: 10.1038/ismej.2017.165
- Guidi et al. (2016) Nature 532:465-470. doi: 10.1038/nature16942

Short introduction to R

Data and object types

margin 2: colnames(data.frame)



data.frame

matrix

margin 1: rownames(data.frame)

	V1	V2	V3	V4	V5	V6
S1	1	0.6	5.3	“A”	A	TRUE
S2	5	-4.3	6	“B”	B	FALSE
S3	7	43.7	34.9	“C”	C	TRUE
S4	3	0.2	8	“B”	B	FALSE
S5	8	-65	-2	“A”	A	FALSE
	integer (numeric)	numeric	numeric	character	factor (levels “A” “B” “C”)	logical

Short introduction to R

Getting your data into shape

Community matrix

- Numerical matrix
- Output from sequence processing software

Taxonomic assignment

- Character matrix or data.frame
- Output from sequence processing software

Contextual information and metadata

- Data.frame
- Provided by user for data analysis

- Most common input format for tabular data:
  - .txt
  - .csv
  - .tsv
- Include variable names in first row (header)
- Values usually tab, space, or comma separated
- Avoid special characters and spaces in data values, variable names, and file names

	Bad	Good
Variable name	mean temperature mean-temperature mean temperature [°C]	temperature.mean
Data value	day 1	day1 1 (variable name: day)

Short introduction to R

Getting your data into shape

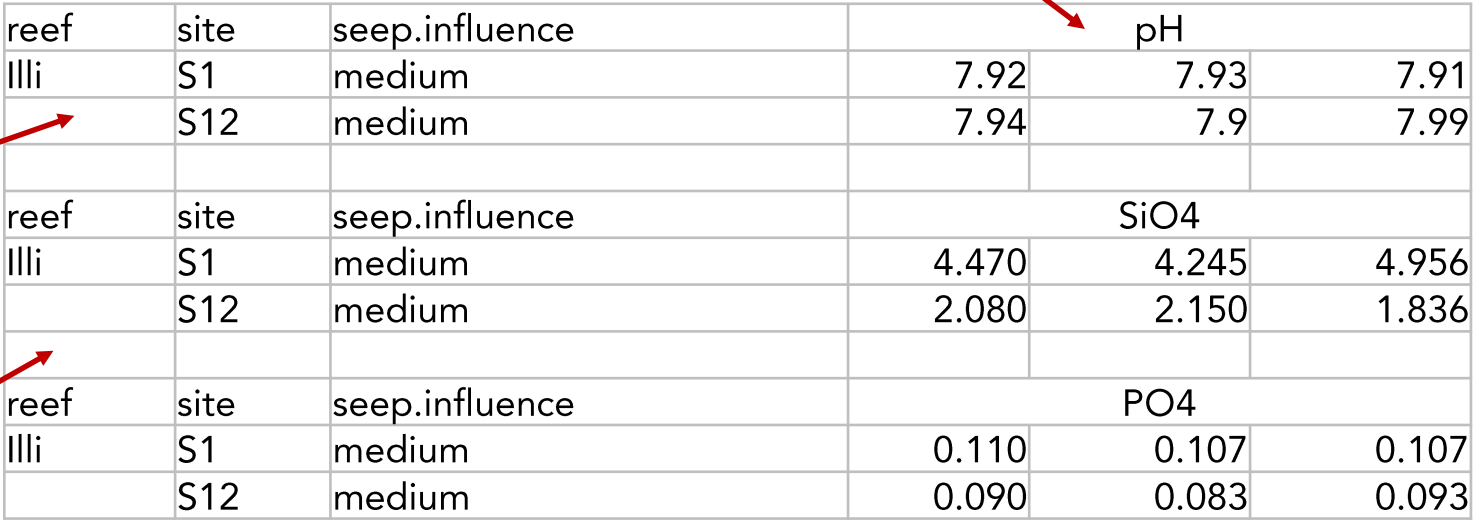
- **Bad:**

Merged cells

Empty cells

Interspersed header

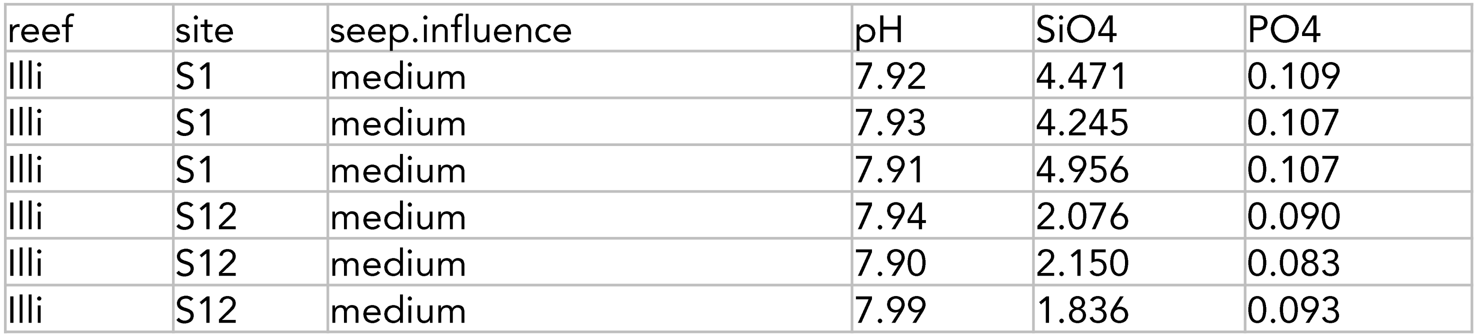
Empty rows



reef	site	seep.influence	pH		
Illi	S1	medium	7.92	7.93	7.91
	S12	medium	7.94	7.9	7.99
reef	site	seep.influence	SiO4		
Illi	S1	medium	4.470	4.245	4.956
	S12	medium	2.080	2.150	1.836
reef	site	seep.influence	PO4		
Illi	S1	medium	0.110	0.107	0.107
	S12	medium	0.090	0.083	0.093

- **Good:**

Wide data format



reef	site	seep.influence	pH	SiO4	PO4
Illi	S1	medium	7.92	4.471	0.109
Illi	S1	medium	7.93	4.245	0.107
Illi	S1	medium	7.91	4.956	0.107
Illi	S12	medium	7.94	2.076	0.090
Illi	S12	medium	7.90	2.150	0.083
Illi	S12	medium	7.99	1.836	0.093

Short introduction to R

Data formats

Long data format:

- One data value per line
- Additional comlums with contextual data (usually categories)

Wide data format:

- More easily readable
- Values either calculated based on or rearrangement of long data format

reef	site	seep.influence	measurement	value
Illi	1	medium	pH	7.92
Illi	1	medium	pH	7.93
Illi	1	medium	pH	7.91
Illi	12	medium	pH	7.94
Illi	12	medium	pH	7.90
Illi	12	medium	pH	7.99
Illi	1	medium	SiO4	4.471
Illi	1	medium	SiO4	4.245
Illi	1	medium	SiO4	4.956
Illi	12	medium	SiO4	2.076
Illi	12	medium	SiO4	2.150
Illi	12	medium	SiO4	1.836
Illi	1	medium	PO4	0.109
Illi	1	medium	PO4	0.107

Original data - rearranged

reef	site	seep.influence	pH	SiO4	PO4
Illi	S1	medium	7.92	4.471	0.109
Illi	S1	medium	7.93	4.245	0.107
Illi	S1	medium	7.91	4.956	0.107
Illi	S12	medium	7.94	2.076	0.090
Illi	S12	medium	7.90	2.150	0.083
Illi	S12	medium	7.99	1.836	0.093

Mean values

reef	site	seep.influence	pH	SiO4	PO4
Illi	S1	medium	7.92	4.539	0.108
Illi	S12	medium	7.94	2.021	0.089

Short introduction to R

Errors

Syntax errors

- When R doesn't understand you, because the command doesn't make sense…
- R returns an error message
- E.g.: Trying to calculate the mean of categorical data

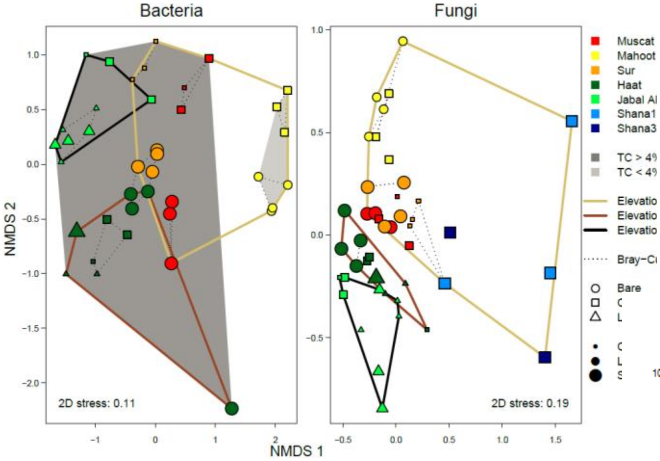
Semantic errors

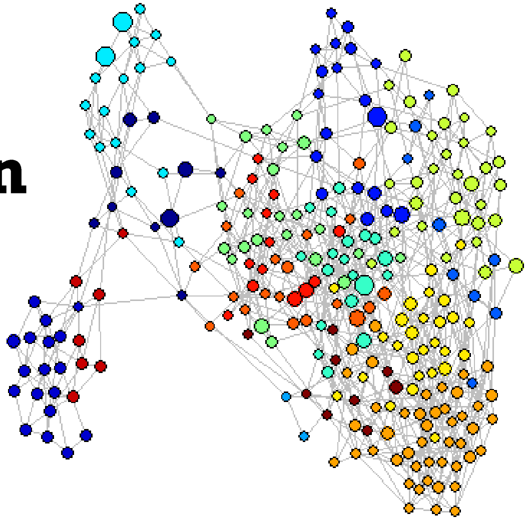
- When R doesn't do what you want, although the command makes sense…
- R will not return an error message, because the command is valid
- More dangerous errors
- E.g.: Calculating percentages over columns, and not rows

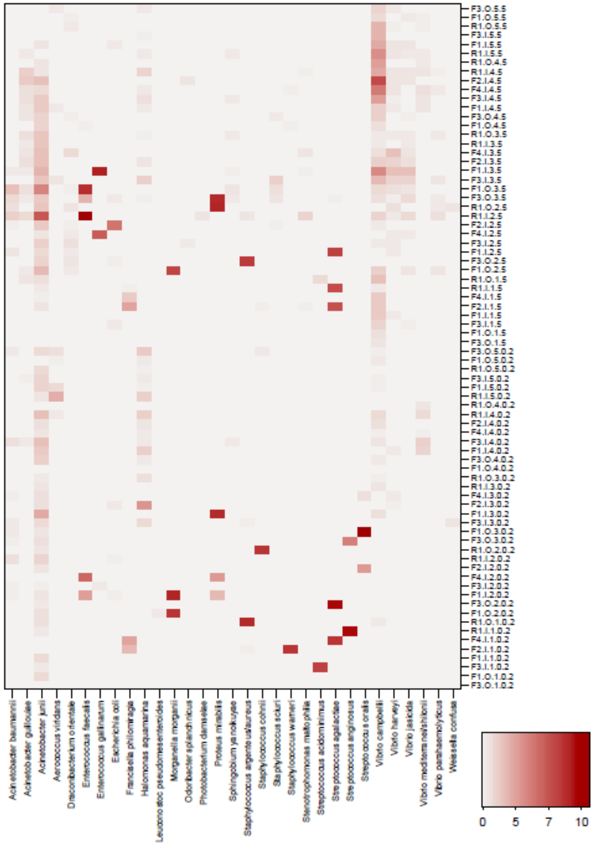
Google is your new best friend ☺

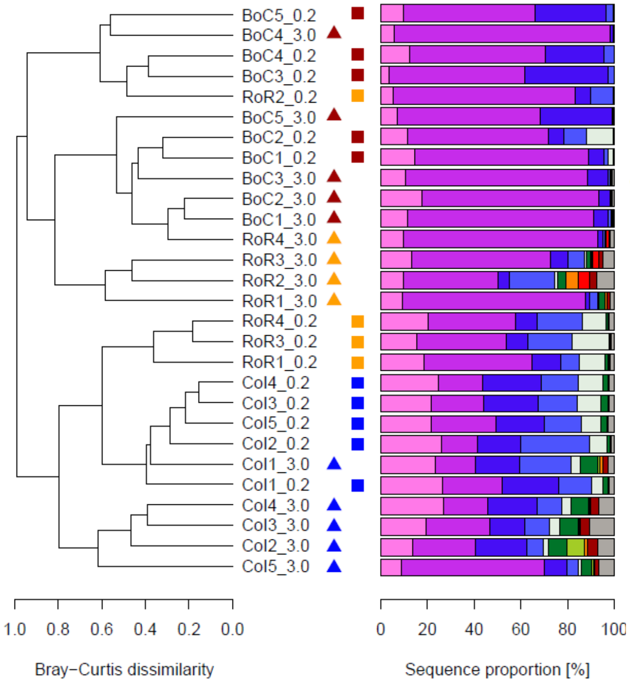
ZMT LEIBNIZ CENTRE for Tropical Marine Researc

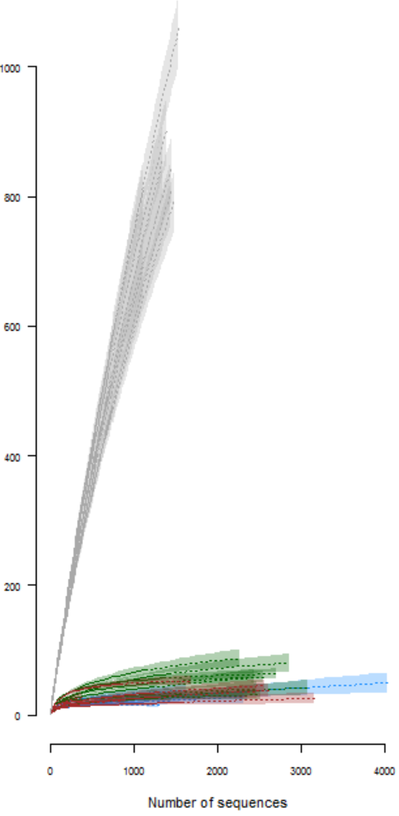
Data visualization

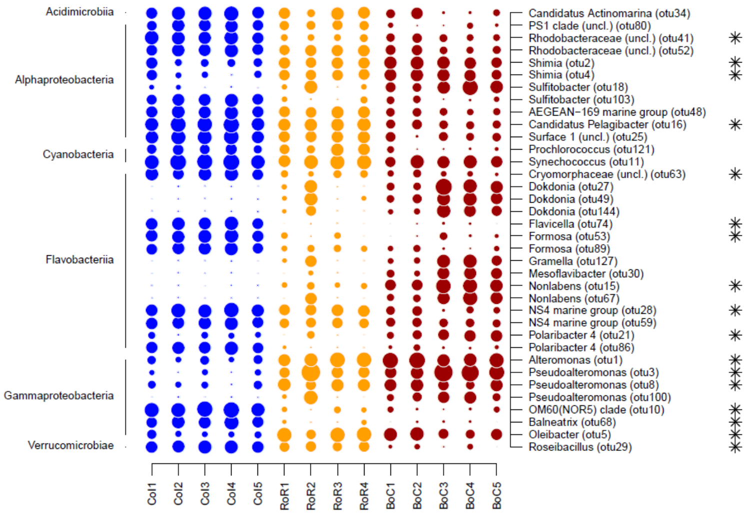












http://seqanswers.com/
https://peerj.com/articles/593/
http://onlinelibrary.wiley.com/doi/10.1111/1755-0998.12428/abstract
http://www.nature.com/ismej/journal/v9/n4/abs/ismej2014195a.html
http://www.nature.com/nmeth/journal/v13/n7/full/nmeth.3869.html
https://www.arb-silva.de/
https://www.ncbi.nlm.nih.gov/pubmed/24910773
https://sites.google.com/site/mb3gustame/
https://github.com/chassenr/Tutorials/tree/master/R_course_MPI
http://cc.oulu.fi/~jarioksa/opetus/metodi/vegantutor.pdf
http://onlinelibrary.wiley.com/doi/10.1890/13-0133.1/abstract
https://www.ncbi.nlm.nih.gov/pubmed/26855872
https://cran.r-project.org/web/packages/iNEXT/iNEXT.pdf
http://www.sciencedirect.com/science/article/pii/S1047279716300722
http://www.sciencedirect.com/science/article/pii/S1047279716300734
https://github.com/zdk123/SpiecEasi
https://rpubs.com/michberr/randomforestmicrobe
http://msystems.asm.org/content/2/1/e00162-16
https://github.com/LangilleLab/microbiome_helper/wiki/Random-Forest-Tutorial
https://www.nature.com/articles/ismej2016139