Metabarcoding Pipeline Building

An introduction with hands on exercises

CUSO – DPEE Activity

Gerhard Thallinger, PhD, Graz University of Technology

Rachel Korn, PhD, Université de Fribourg

Magdalena Steiner, PhD, Agroscope Wädenswil

CUSO
CONFÉRENCE UNIVERSITAIRE
DE SUISSE OCCIDENTALE

TU Graz

UNI FR

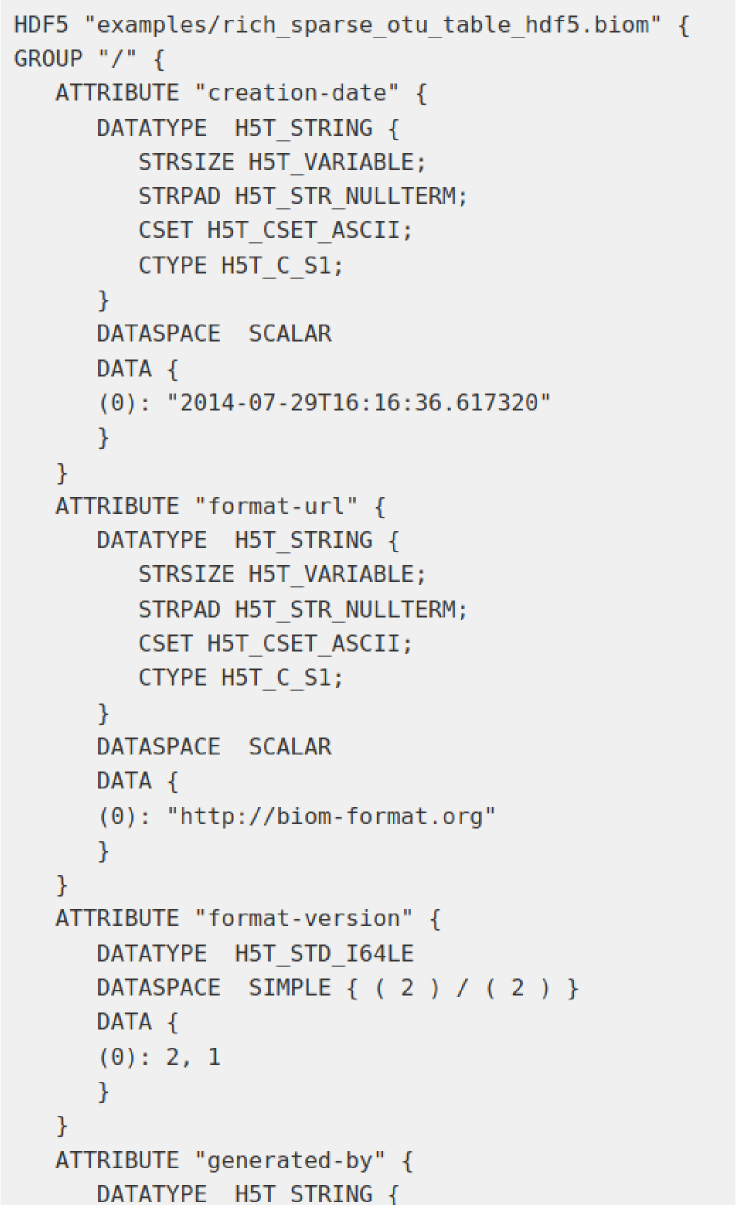
BIOM format 2.1

- BIOM = Biological Observation Matrix [‘baɪoʊm]
- By Earth Microbiome Project
- Based on HDF5 (widely supported binary format with native parsers available within many programming languages)

https://biom-format.org/

BIOM format 2.1

- Header before conversion to binary
- Readable by most common software (R, Python…)



```
HDF5 "examples/rich_sparse_otu_table_hdf5.biom" {
GROUP "/" {
   ATTRIBUTE "creation-date" {
      DATATYPE  H5T_STRING {
         STRSIZE H5T_VARIABLE;
         STRPAD H5T_STR_NULLTERM;
         CSET H5T_CSET_ASCII;
         CTYPE H5T_C_S1;
      }
      DATASPACE  SCALAR
      DATA {
      (0): "2014-07-29T16:16:36.617320"
      }
   }
   ATTRIBUTE "format-url" {
      DATATYPE  H5T_STRING {
         STRSIZE H5T_VARIABLE;
         STRPAD H5T_STR_NULLTERM;
         CSET H5T_CSET_ASCII;
         CTYPE H5T_C_S1;
      }
      DATASPACE  SCALAR
      DATA {
      (0): "http://biom-format.org"
      }
   }
   ATTRIBUTE "format-version" {
      DATATYPE  H5T_STD_I64LE
      DATASPACE  SIMPLE { ( 2 ) / ( 2 ) }
      DATA {
      (0): 2, 1
      }
   }
   ATTRIBUTE "generated-by" {
      DATATYPE  H5T_STRING {
```

https://biom-format.org/

phyloseq

- R/Bioconductor package
- Import, store, analyze and visualize sequencing data
- S4 object containing OTU table + metadata (sample data, taxonomy table, phylogenetic trees etc.)

phyloseq on GitHub

phyloseq

- Flexible imports (BIOM, mothur, DADA2, QIIME etc.)
- Bridges to ggplot2 and numerous ecological R packages (e.g. vegan, ade4, ape)

phyloseq on GitHub

Analysis approaches: OTU/ASV vs. phylogeny

- Analysis pipelines create an OTU/ASV table with read abundance per sample
- OTUs/ASVs are taxonomically classified
- Two approaches:
  - Analyse OTU/ASV table directly
  - Aggregate (collapse) OTUs/ASVs with the same level of taxonomic classification → reduced table containing phylotypes
- Downstream analysis is the same for both table types

Preprocessing & visualization

- Look at your data (which taxa, reads, OTUs…)
- Check for spurious taxa and remove
- Data normalization
- Biodiversity measures
  - Number of OTUs
  - Diversity indices (Shannon, Chao, …)

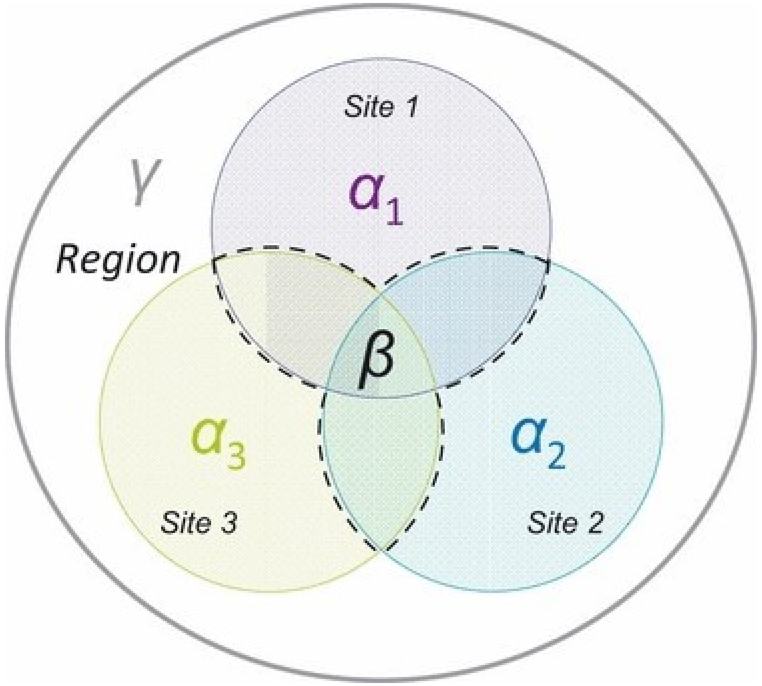
Normalization for community composition

- Different abundances due to different sequencing depths among your samples → reflect differential efficiency of the sequencing process (rather than true biological variation)
  - Rarefy or not rarefy (controversial)
    - Random subsampling of reads in your sample to equal numbers of reads (default is lowest sample)
    - Loose OTUs (Check for low-read samples before!)
  - Transform abundance of OTUs
    - Relative abundance (do not loose OTUs)
    - Scale
    - Rank abundance

McMurdie et al. 2014, Weiss et al. 2017

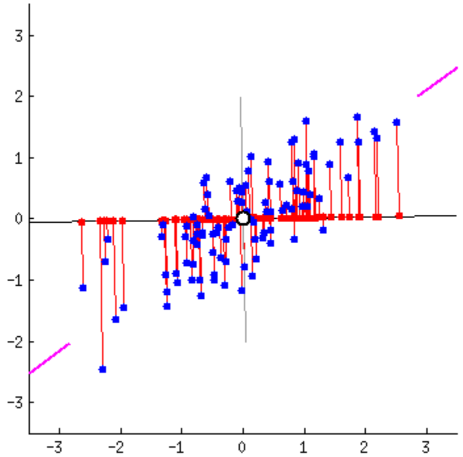
Biodiversity measures

- α-diversity: within one habitat, location or sample
  - Richness or diversity indices: Shannon, Evenness, Chao etc.
- β-diversity: difference in diversity or community composition between two or more habitats, locations or samples
  - Similarity indices – Jaccard or Sorensen (presence/absence)
  - Dissimilarity indices (relative abundances) – Bray-Curtis or Morisita-Horn
- ɣ-diversity: across all habitats of a region or all samples



Zinger et al. 2011

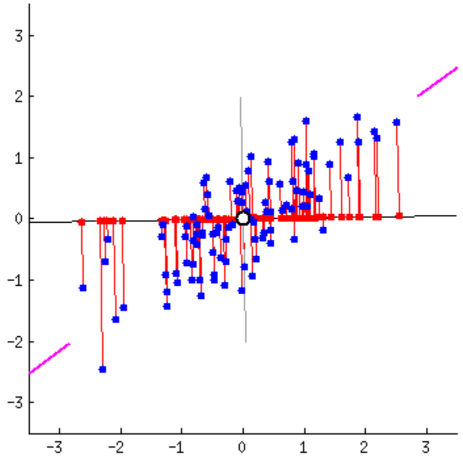
Ordination



- Latin *ordinatio* – the action of setting in order
- Exploratory analysis
- Orders multidimensional objects by similarity in a low(er)-dimensional space Legendre et al. 2012

StackExchange amoeba 2015

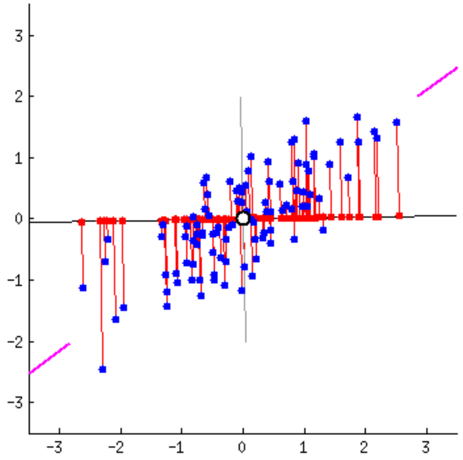
Ordination vs. clustering



- Clustering
  - Pairwise distances among objects → fine relationships
- Ordination
  - Variability of the whole association matrix → general patterns (gradients) Legendre et al. 2012

StackExchange amoeba 2015

Ordination



- Problem: high-dimensional data
- Solution: projection into low-dimensional space = dimension reduction Legendre et al. 2012

StackExchange amoeba 2015

Ordination

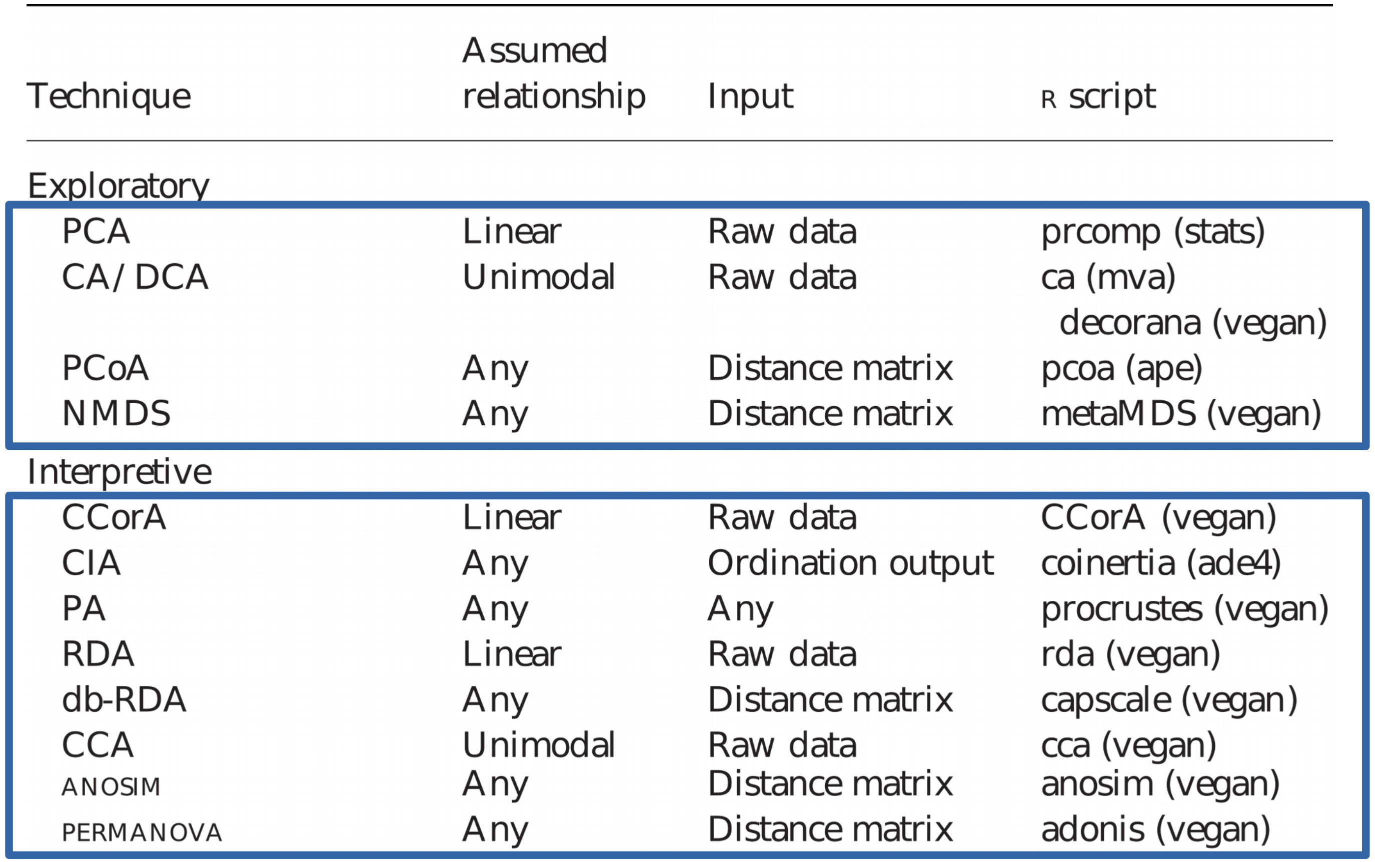
- Unconstrained ordination = indirect gradient analysis
  - Community composition
- Constrained ordination = direct gradient analysis
  - Community composition + environmental factors

Ordination techniques

- Unconstrained
  - Principal component analyis (PCA, ***inadequate*** for community composition!)
  - Principal Coordinate analysis (PCoA) = multidimensional scaling (MDS)
  - Nonmetric multidimensional scaling (nMDS)
- Constrained
  - Redundancy analysis (RDA)
  - Canonical correlation analysis (CCA)
  - And many more!

Ordination overview

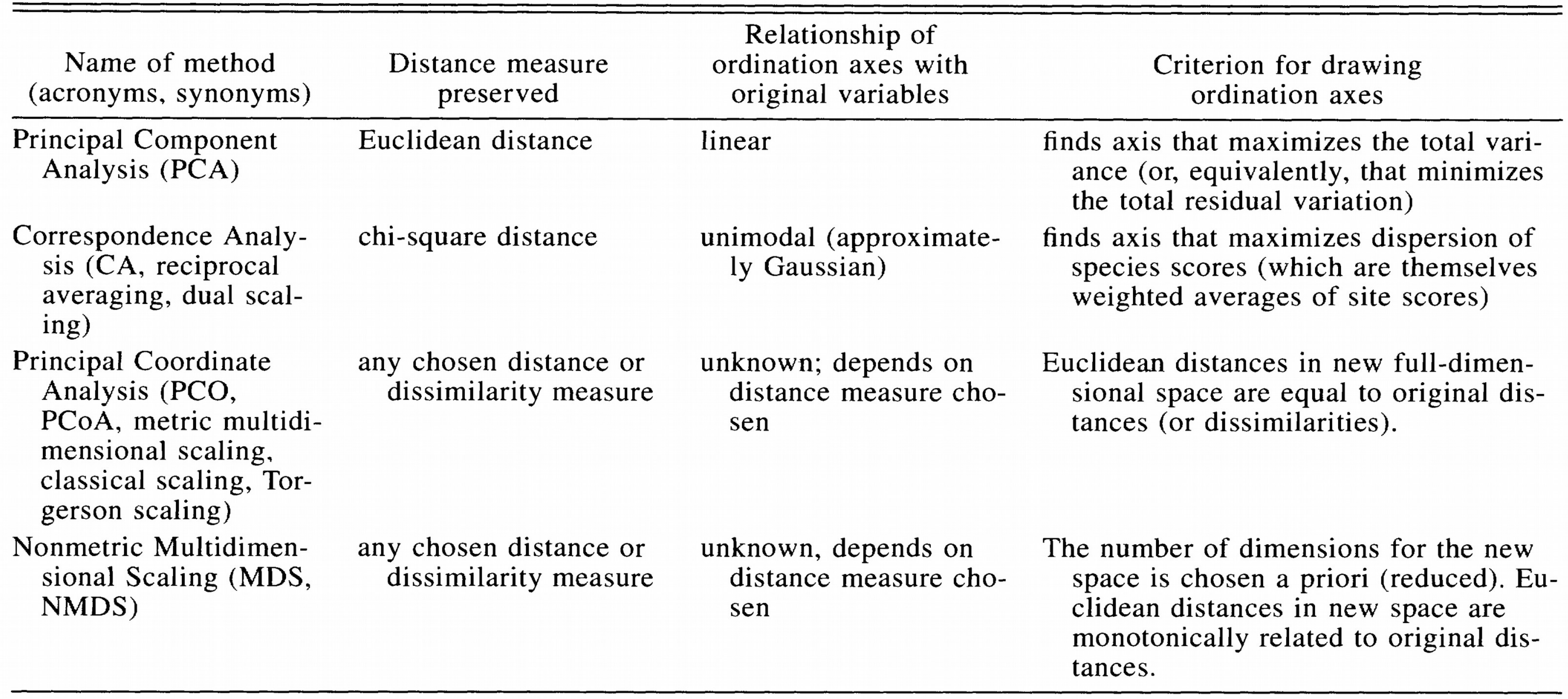
- Unconstrained ordination
- Constrained ordination



Technique	Assumed relationship	Input	R script
Exploratory			
PCA	Linear	Raw data	prcomp (stats)
CA/ DCA	Unimodal	Raw data	ca (mva) decorana (vegan)
PCoA	Any	Distance matrix	pcoa (ape)
NMDS	Any	Distance matrix	metaMDS (vegan)
Interpretive			
CCorA	Linear	Raw data	CCorA (vegan)
CIA	Any	Ordination output	coinertia (ade4)
PA	Any	Any	procrustes (vegan)
RDA	Linear	Raw data	rda (vegan)
db-RDA	Any	Distance matrix	capscale (vegan)
CCA	Unimodal	Raw data	cca (vegan)
ANOSIM	Any	Distance matrix	anosim (vegan)
PERMANOVA	Any	Distance matrix	adonis (vegan)

Paliy et al. 2016

Unconstrained ordination

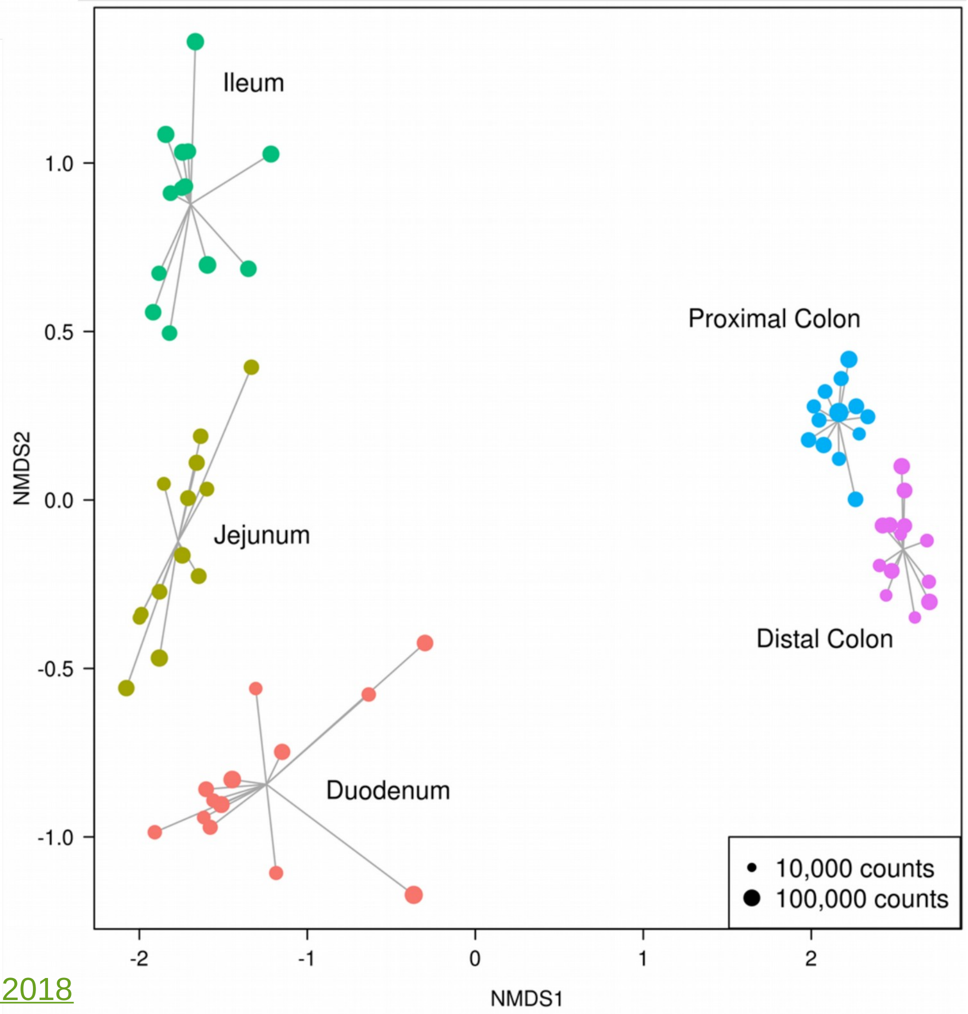


Name of method (acronyms, synonyms)	Distance measure preserved	Relationship of ordination axes with original variables	Criterion for drawing ordination axes
Principal Component Analysis (PCA)	Euclidean distance	linear	finds axis that maximizes the total variance (or, equivalently, that minimizes the total residual variation)
Correspondence Analysis (CA, reciprocal averaging, dual scaling)	chi-square distance	unimodal (approximately Gaussian)	finds axis that maximizes dispersion of species scores (which are themselves weighted averages of site scores)
Principal Coordinate Analysis (PCO, PCoA, metric multidimensional scaling, classical scaling, Torgerson scaling)	any chosen distance or dissimilarity measure	unknown; depends on distance measure chosen	Euclidean distances in new full-dimensional space are equal to original distances (or dissimilarities).
Nonmetric Multidimensional Scaling (MDS, NMDS)	any chosen distance or dissimilarity measure	unknown, depends on distance measure chosen	The number of dimensions for the new space is chosen a priori (reduced). Euclidean distances in new space are monotonically related to original distances.

Anderson et al. 2003

Non-metric multidimensional scaling (nMDS)

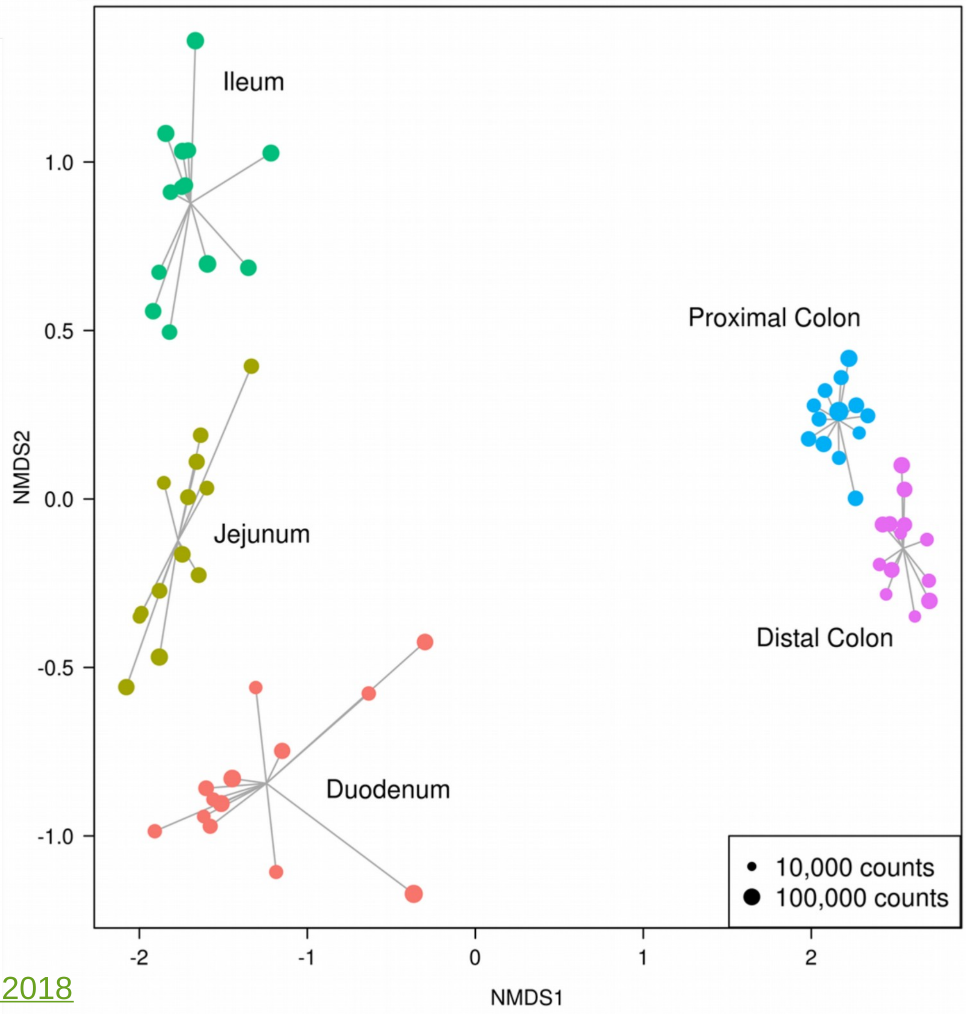
- Non-linear compression of the distances
- Preserves order of relationships, based on ranks
- Input: any distance matrix



Crespo-Piazuelo et al. 2018

Non-metric multidimensional scaling (nMDS)

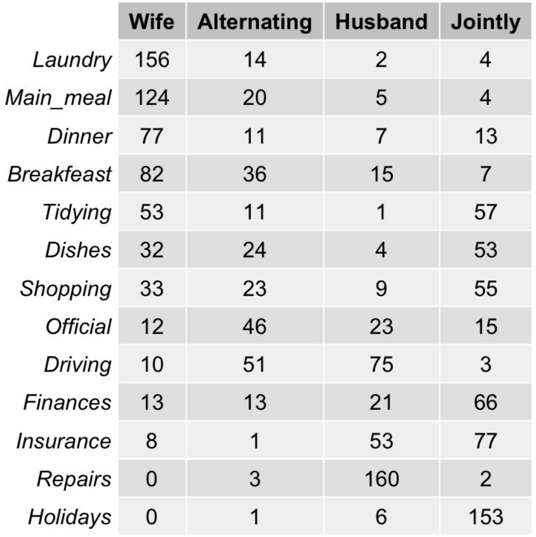
- Axes are not necessarily ordered
- Number of axes defined a priori
- Iterative process (local maxima trap!)
- No unique solution
- Stress value
  - < 0.2 = okay
  - < 0.1 = good
  - < 0.05 excellent



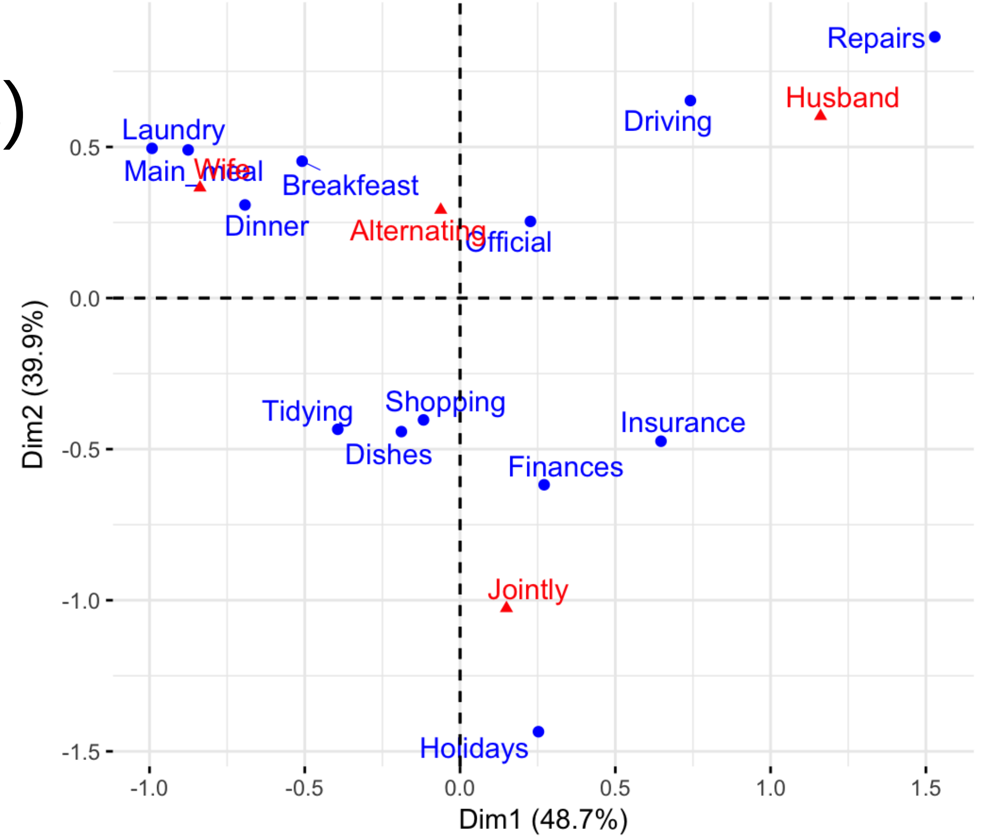
Crespo-Piazuelo et al. 2018

Correspondence analysis (CA)

- χ^2 distance (excludes double-zeros)
- Eigenvector-based
- Community composition at sampling sites



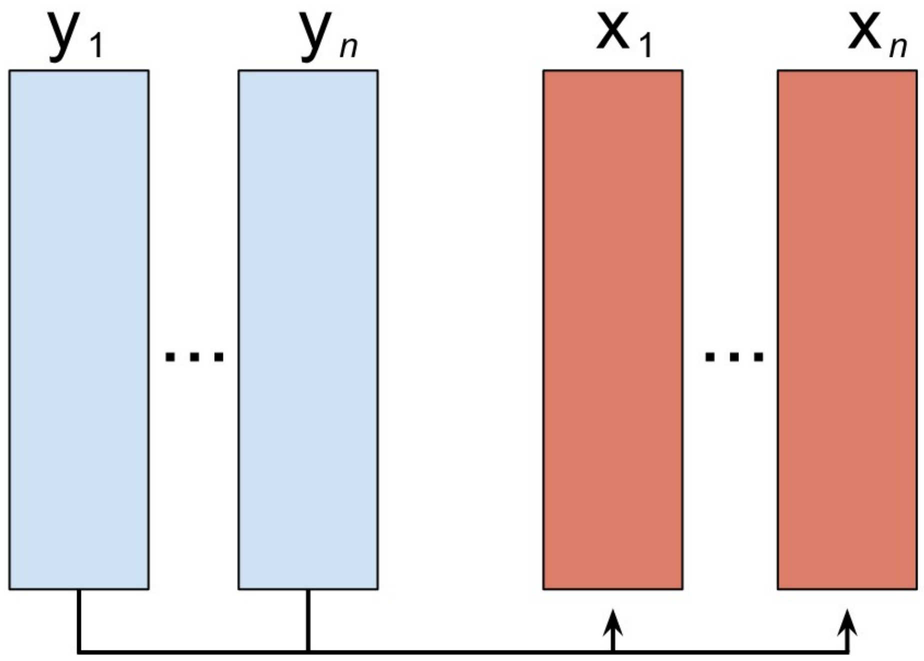
	Wife	Alternating	Husband	Jointly
Laundry	156	14	2	4
Main_meal	124	20	5	4
Dinner	77	11	7	13
Breakfeast	82	36	15	7
Tidying	53	11	1	57
Dishes	32	24	4	53
Shopping	33	23	9	55
Official	12	46	23	15
Driving	10	51	75	3
Finances	13	13	21	66
Insurance	8	1	53	77
Repairs	0	3	160	2
Holidays	0	1	6	153

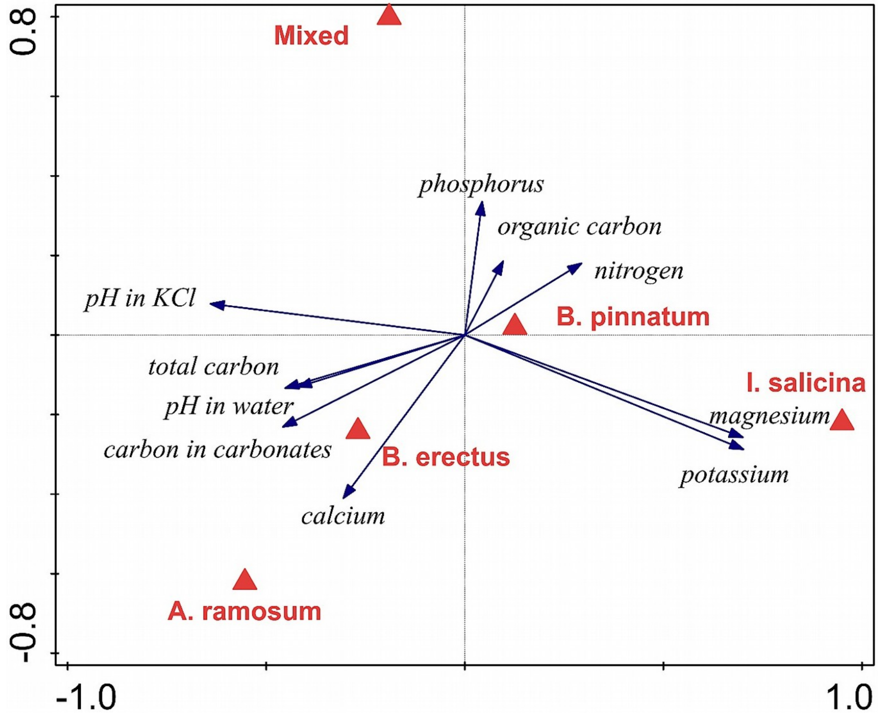


STHDA kassambara 2017

Redundancy analysis (RDA)

- Extension of multiple linear regression





GUSTA ME

Merová et al. 2016

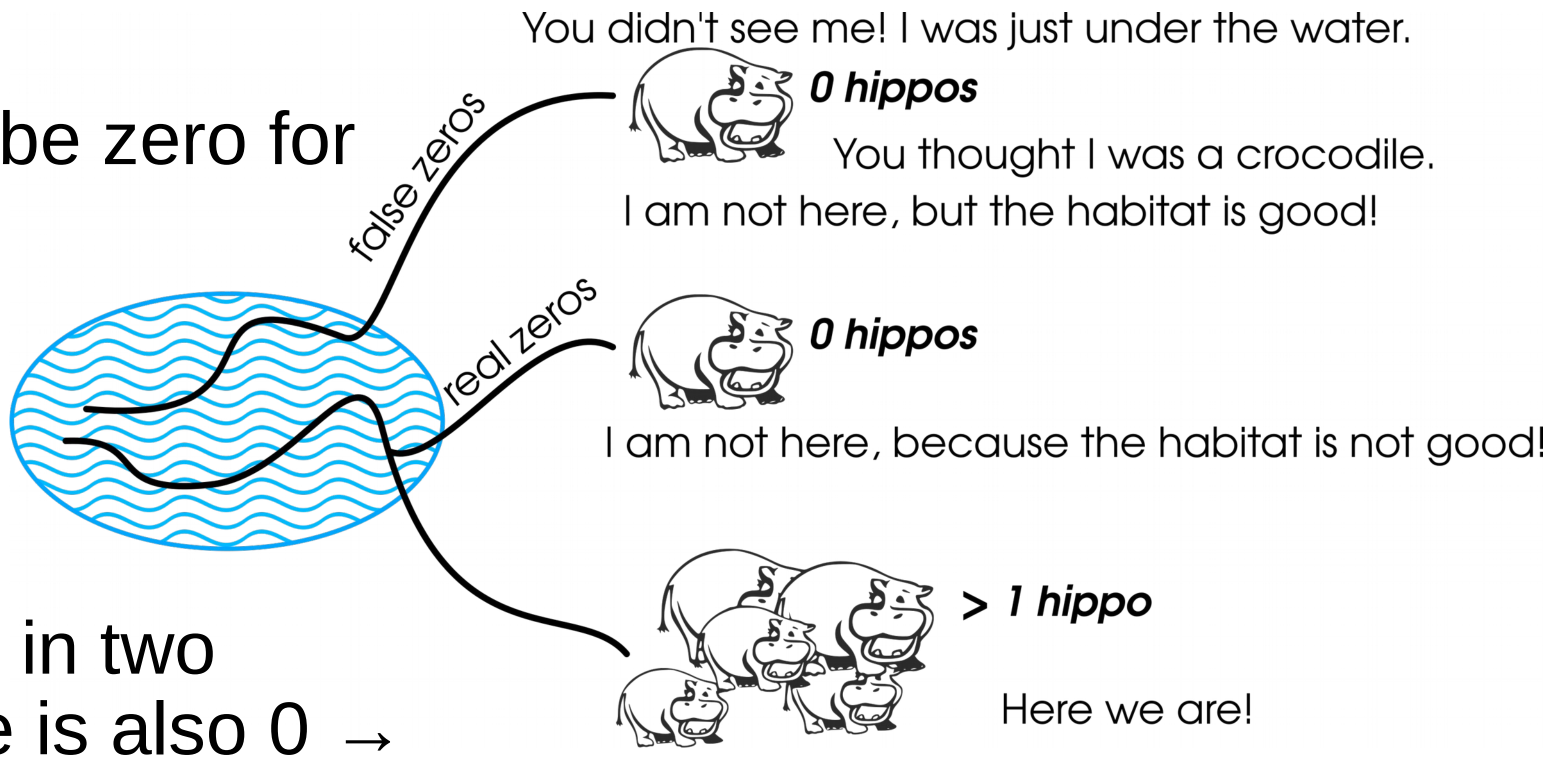
(Dis)Similarity indices

- Data type
  - Presence/absence
  - Count data, relative abundance
- Interpretation of zeros
  - Symmetric indices: zeros (absences) have same meaning as presences
  - Asymmetric indices: ignore double zeros (uncertain if a species was overlooked or is actually absent)

Legendre et al. 2012

Double zero problem

- A species count may be zero for different reasons
  - Actually absent
  - Overlook
  - Misclassified
- If a species count = 0 in two samples, the distance is also 0 → may lead to wrong conclusions



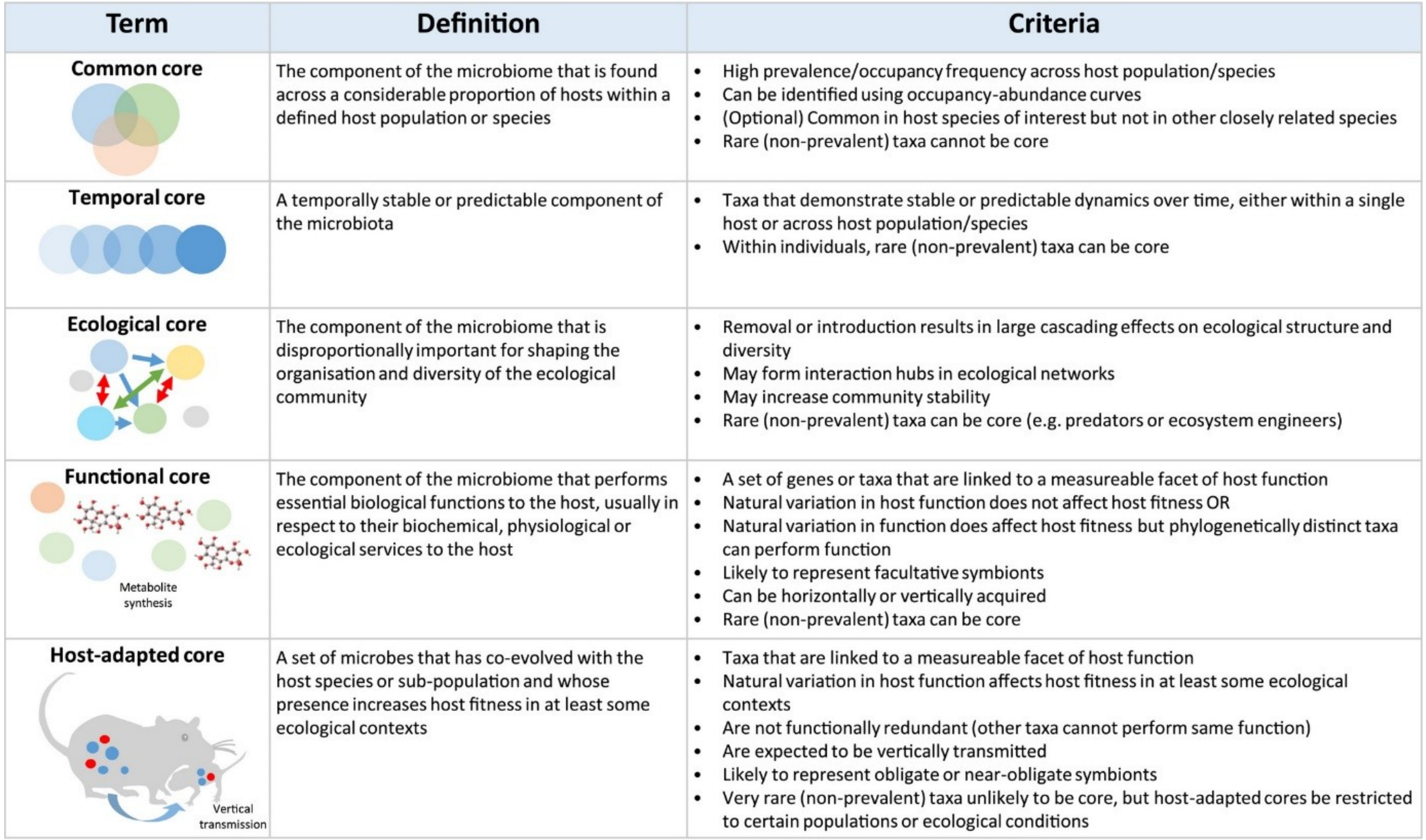
You didn't see me! I was just under the water.
false zeros
0 hippos
You thought I was a crocodile.
I am not here, but the habitat is good!
real zeros
0 hippos
I am not here, because the habitat is not good!
> 1 hippo
Here we are!

Zuur et al. 2009

Core microbiome

- Aim to identify groups of microbes that are particularly widespread across the sample population
- Define a component of the microbiome that may be particularly important
- However increasing evidence that rare taxa are likely to be just as important as widespread taxa

Risely 2020



Term	Definition	Criteria
Common core	The component of the microbiome that is found across a considerable proportion of hosts within a defined host population or species	High prevalence/occupancy frequency across host population/species Can be identified using occupancy-abundance curves (Optional) Common in host species of interest but not in other closely related species Rare (non-prevalent) taxa cannot be core
Temporal core	A temporally stable or predictable component of the microbiota	Taxa that demonstrate stable or predictable dynamics over time, either within a single host or across host population/species Within individuals, rare (non-prevalent) taxa can be core
Ecological core	The component of the microbiome that is disproportionally important for shaping the organisation and diversity of the ecological community	Removal or introduction results in large cascading effects on ecological structure and diversity May form interaction hubs in ecological networks May increase community stability Rare (non-prevalent) taxa can be core (e.g. predators or ecosystem engineers)
Functional core	The component of the microbiome that performs essential biological functions to the host, usually in respect to their biochemical, physiological or ecological services to the host	A set of genes or taxa that are linked to a measureable facet of host function Natural variation in host function does not affect host fitness OR Natural variation in function does affect host fitness but phylogenetically distinct taxa can perform function Likely to represent facultative symbionts Can be horizontally or vertically acquired Rare (non-prevalent) taxa can be core
Host-adapted core	A set of microbes that has co-evolved with the host species or sub-population and whose presence increases host fitness in at least some ecological contexts	Taxa that are linked to a measureable facet of host function Natural variation in host function affects host fitness in at least some ecological contexts Are not functionally redundant (other taxa cannot perform same function) Are expected to be vertically transmitted Likely to represent obligate or near-obligate symbionts Very rare (non-prevalent) taxa unlikely to be core, but host-adapted cores be restricted to certain populations or ecological conditions

Differentially abundant taxa

- Which specific taxa (OTUs/ASVs) are significantly differentially abundant between groups?

- Typical workflow:
  - Estimate differential abundance of taxa independently
  - Perform some test using the estimate
  - Perform correction for multiple testing

Differentially abundant taxa

- Challenges with microbiome data:
  - Large number of taxa
  - Sparse data matrix (a lot of zeros)
  - Count data / instead of continuous data
  - Number of taxa >> number of samples
  - High variance within a single taxon

McMurdie et al. 2014, Weiss et al. 2017

Differentially abundant taxa

- Standard approaches (*t*-test, ANOVA) not applicable
  - Data not normally distributed
  - False negatives due to high variance within taxa
  - False positives due to large number of zeros in taxa
- Use approaches from RNA-seq analysis
  - Variance “shrinkage” due to estimates from all taxa
  - R packages: edgeR, DESeq2

Robinson et al. 2010, Love et al. 2014

Further packages & webservers

- R packages:
    - BiomMiner
    - MicrobiomeExplorer
- Webservers
    - Calypso

Shamsaddini et al. 2020 , Reeder et al. 2021, Zakrzewski et al. 2017