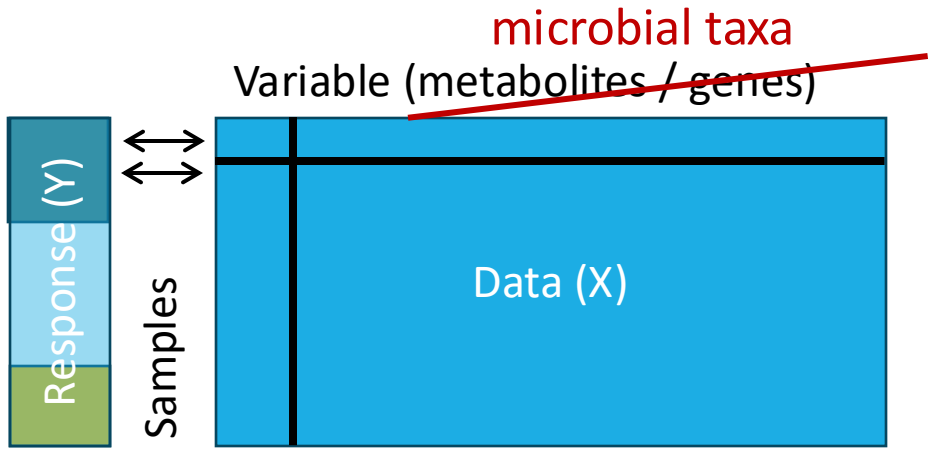
Microbiome Data Analysis

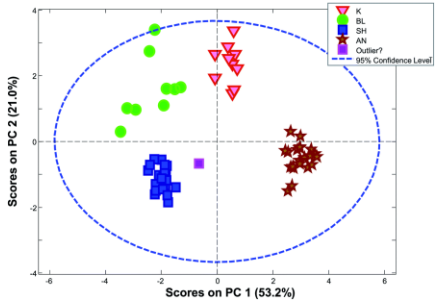
Anna Heintz-Buschart

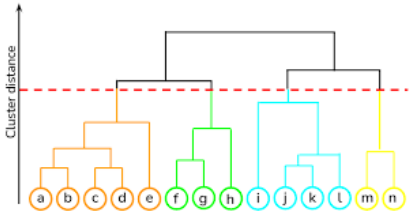
28 January 2024

Methods



- Univariate analysis (Gene expression data)
  - Comparability
  - Error structure
- Multivariate data analysis methods
  - Gene expression and metabolomics data
- Microbiome data
  - Sparse and compositional data requires specific multivariate data analysis tools based on between sample distances





- Exploration (X)
  - PCA
  - Clustering
  - Y information is not used in calculations, but only to color samples
- Supervised analysis
  - Use Y to calculate model (Y = Xb + E)
  - Classification: Y defines groups
  - High dimensional ANOVA: Y defines experimental design information
  - Regression: Y defines quantitative information
- Machine learning for nonlinear models
  - Random Forest
  - Neural Nets
- Validation:
  - Test whether model does not overfit / effects are different from zero

Overview – lecture

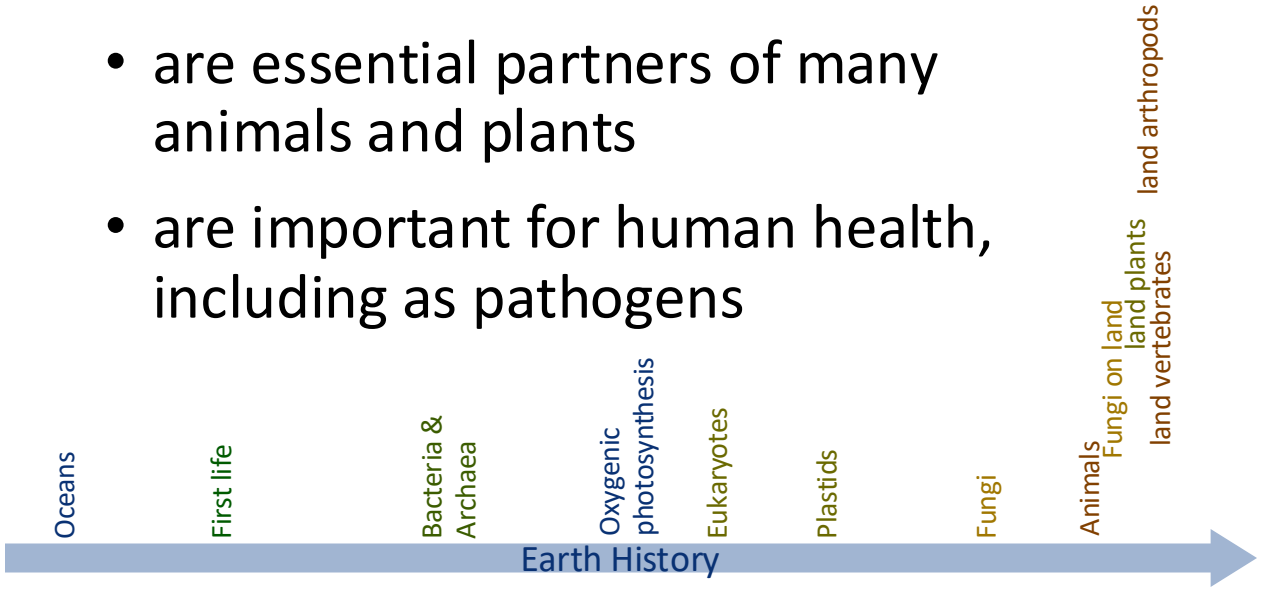
- What are microbiomes and how do we study them?
- Challenges in data analysis
  - What causes them
  - How to recognize common problems
  - Data transformation / normalization
- Microbial ecology data analysis methods for distances between samples

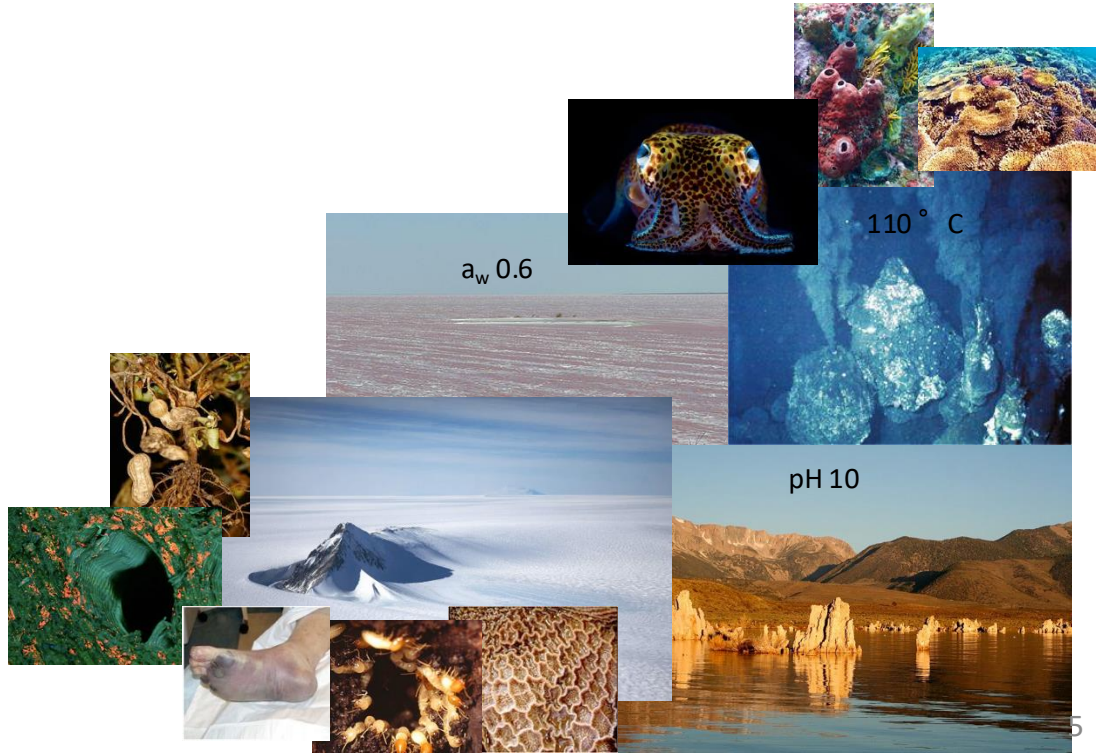
Why do we study microbiomes?

Microbiomes

Microbes:

- are the oldest form of life today
- have shaped our world
- can and do live nearly everywhere on Earth
- are essential partners of many animals and plants
- are important for human health, including as pathogens





Microbiomes

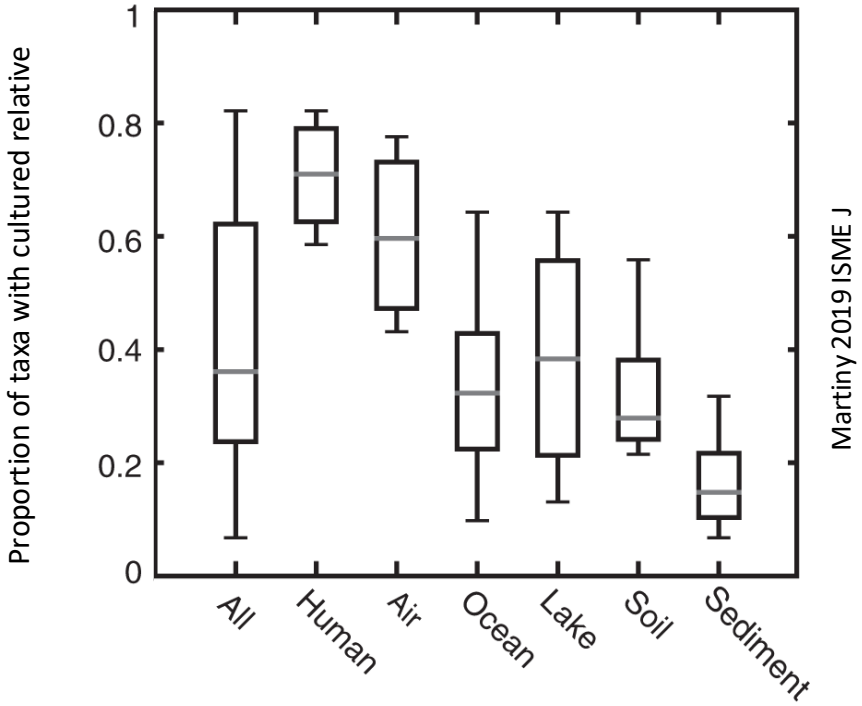
Microbes:

- nearly always live in **mixed communities**

  (= microbiomes)

- Most microbes are difficult or impossible to isolate and to culture

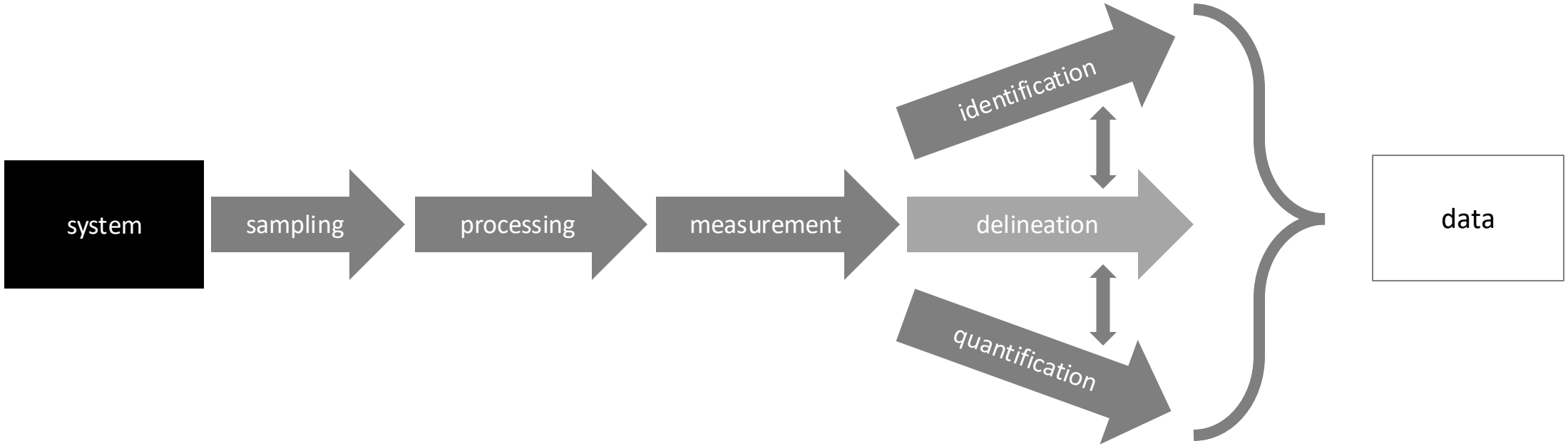
- feed of each other's products
- shape each other's environment



How do we study microbiomes?

Microbiomes are studied by *in-situ* multiplex methods, such as “meta-omics” or “meta-barcoding”

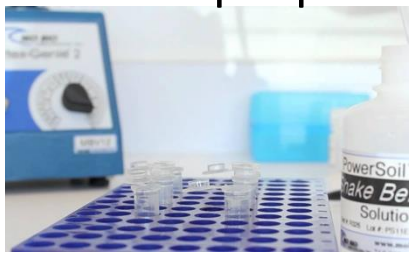
Measuring microbiomes: omics paradigm

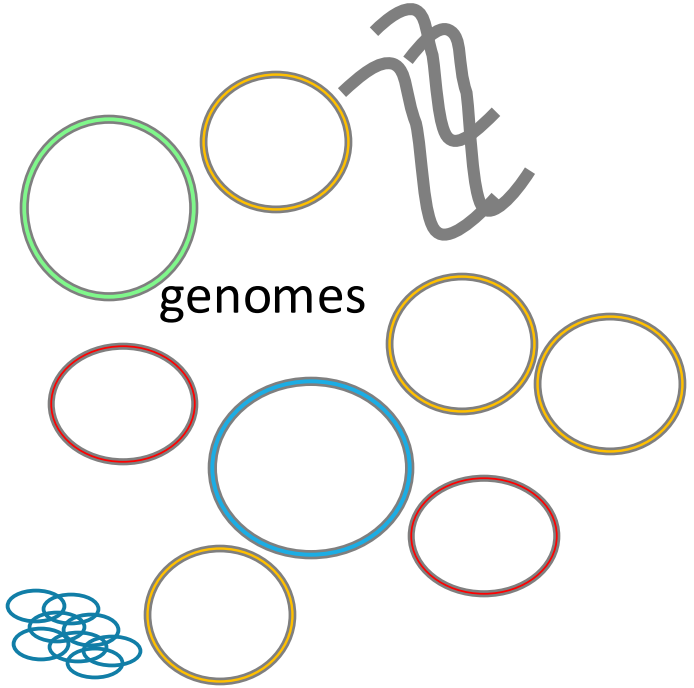


Measuring microbiomes: who is there?

sampling + preservation

sample processing



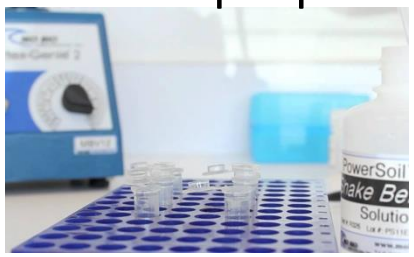


“taxon” (pl. taxa): group of organisms that form a unit (e.g. a species, a genus, a family etc.)

Measuring microbiomes: who is there?

sampling + preservation

sample processing



sequencing & identification

Bacterium 1

Bacterium 2

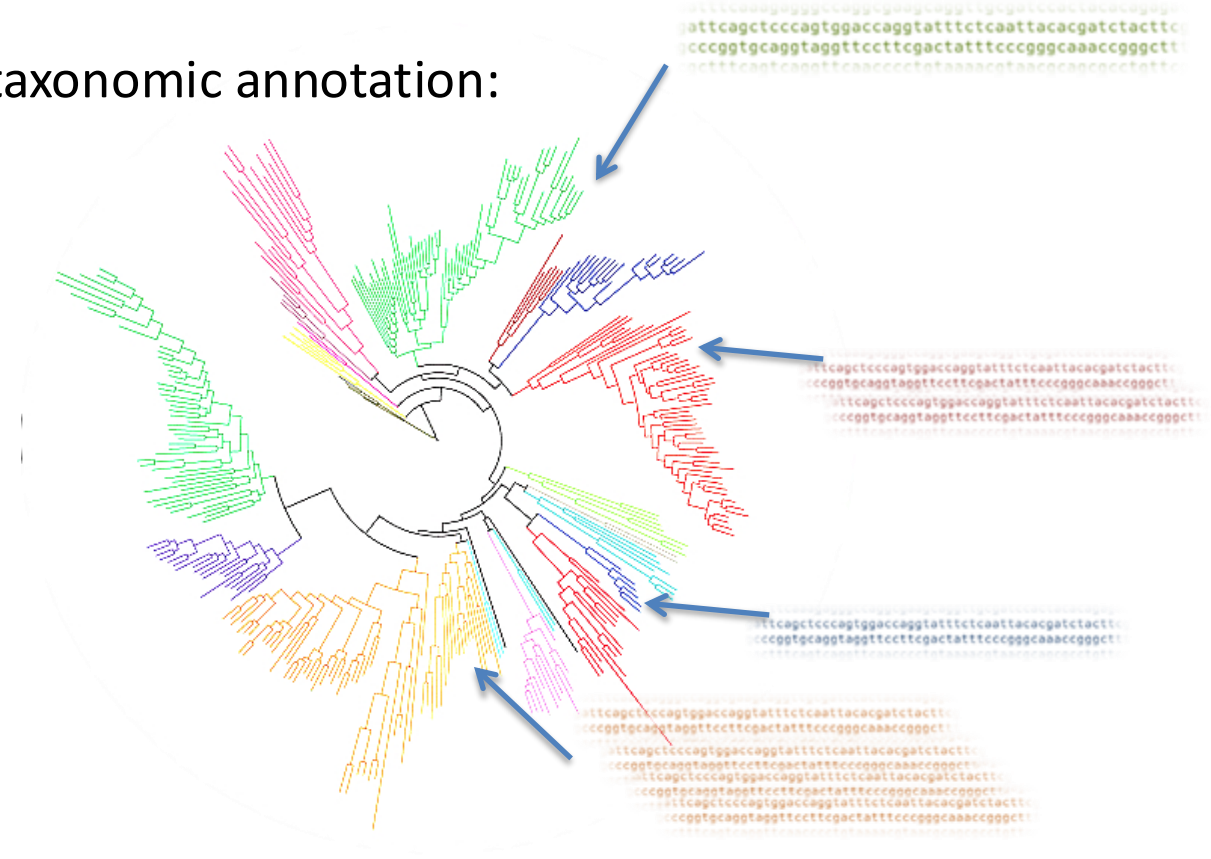
Bacterium 3

Bacterium 4

Measuring microbiomes: who is there?

sequences:

taxonomic annotation:



Wu *et al.* 2009 Nature

Measuring microbiomes: who is there? and how much?

sequencing & identification

Bacterium 1

Bacterium 2

Bacterium 3

Bacterium 4

counting:

taxon	Bact. 1	Bact. 2	Bact. 3	Bact. 4
Sample 1	1	2	1	4
Sample 2	2	2	3	2
Sample 3	1	0	0	1
Sample …	…	…	…	…
Sample N	4	1	7	0

Microbiome research questions

- **What kind of microbes form a microbiome?**

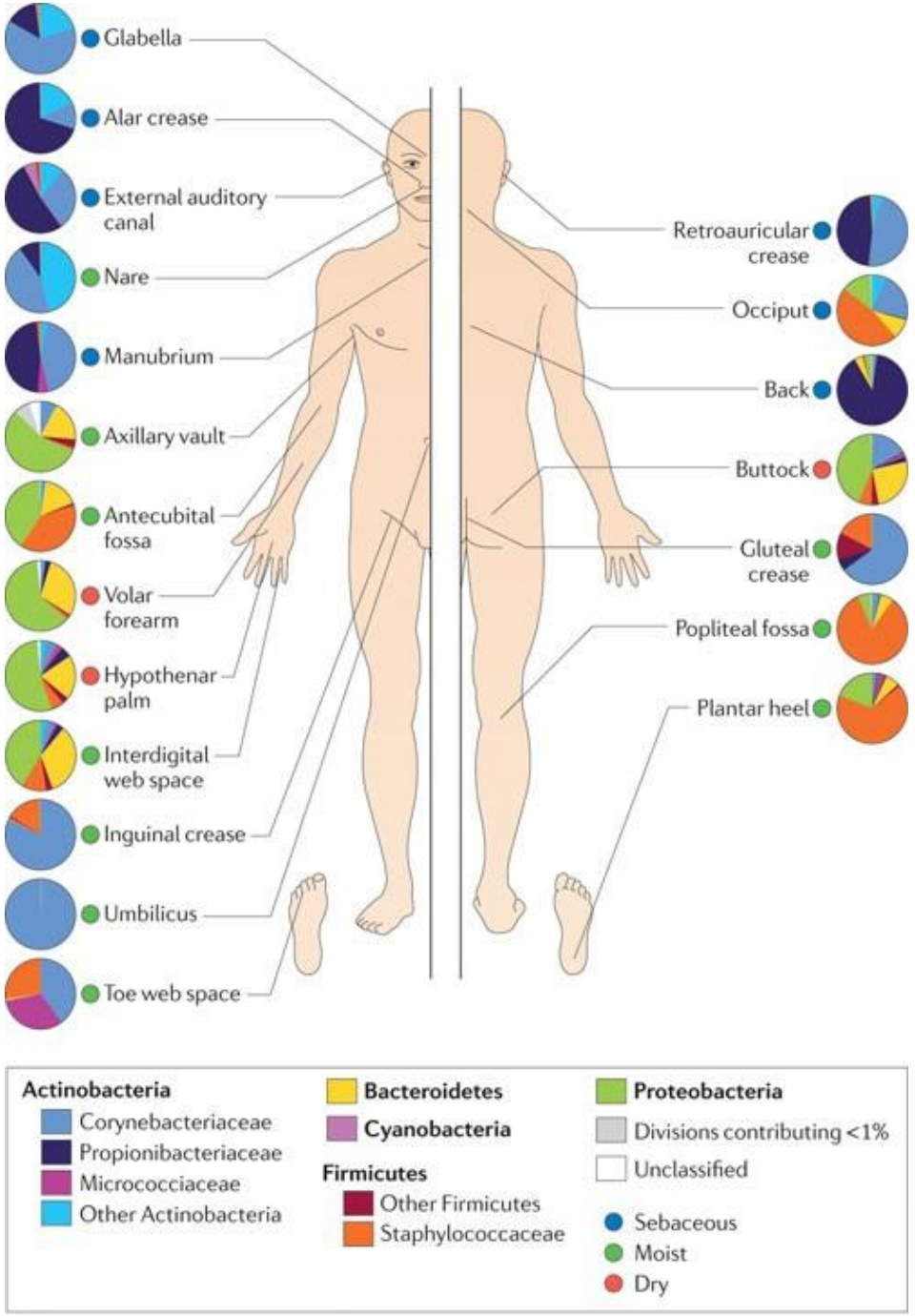
- **How (dis-)similar are microbiomes?** what do they have in common / what sets them apart?

  – cross-sectionally or over time

How would you analyse this question?

You have measured microbiome taxonomic profiles from different places on someone's body

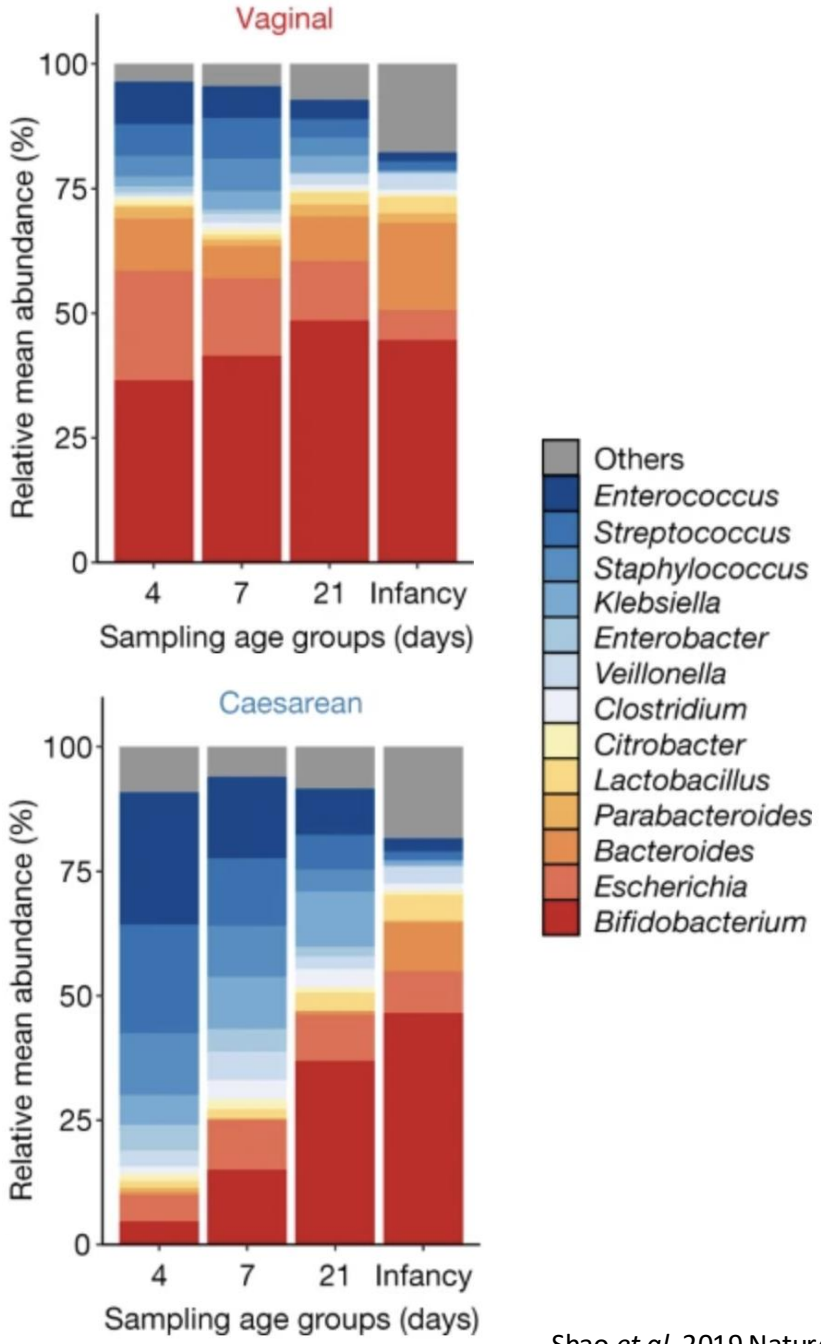
Which microbiomes are composed of similar bacteria?



Grice & Segre 2011
Nature Reviews | Microbiology

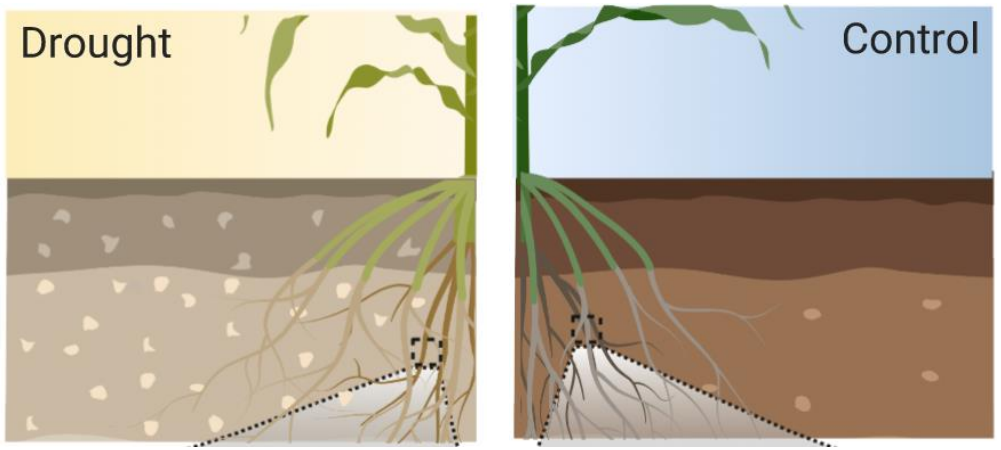
You have measured faecal microbiome samples from 300 random babies and their mothers

Can I see if the babies were born by caesarian section or vaginally?



Shao *et al.* 2019 Nature

You have performed an experiment with sorghum grown under drought and control conditions. You've measured root-associated microbiomes in both conditions.

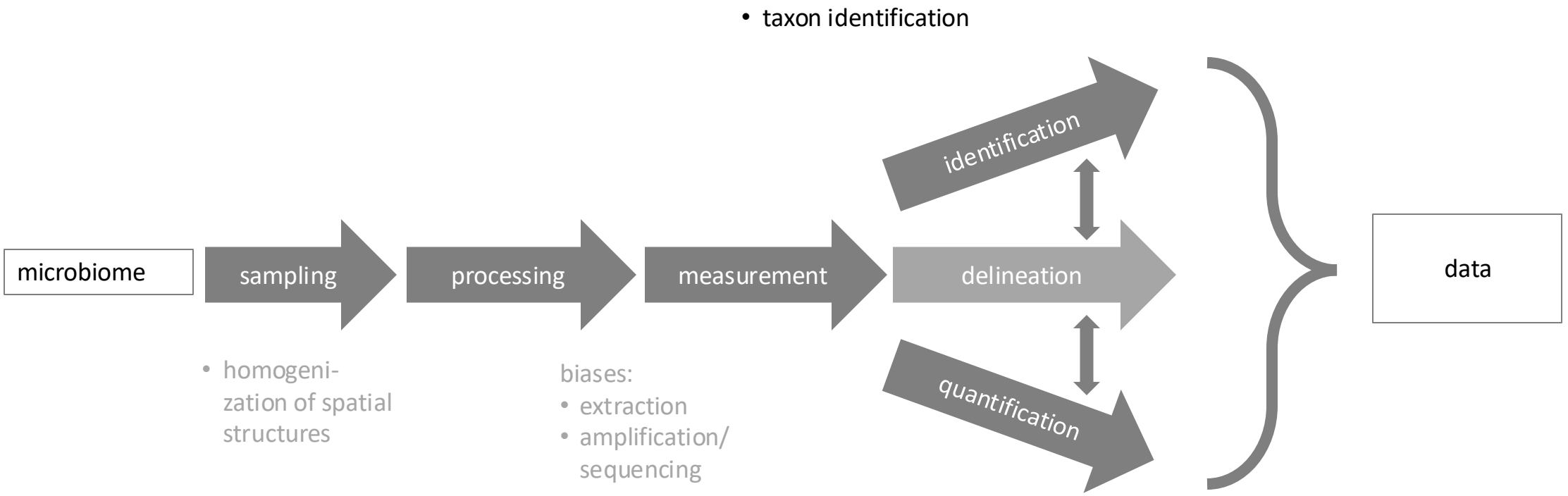


Which bacteria are more abundant in stressed plants?

Xu *et al.* 2018 PNAS
Xu *et al.* 2021 Nature Communications

What are the challenges?

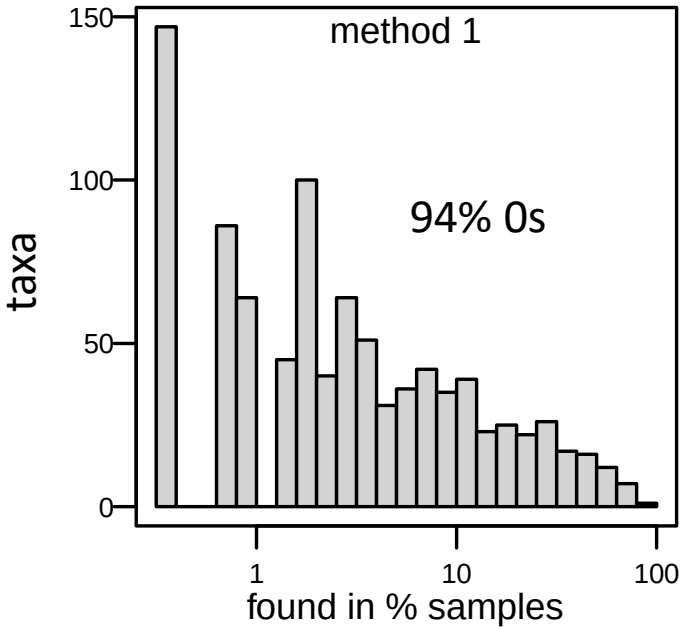
Measuring microbiomes: challenges



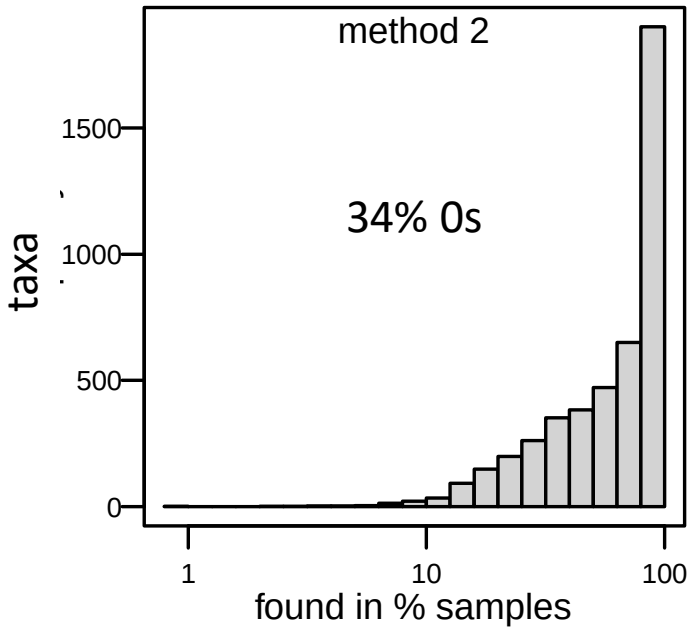
Effect of feature identification on data

Example – from the practical data set:

specific method

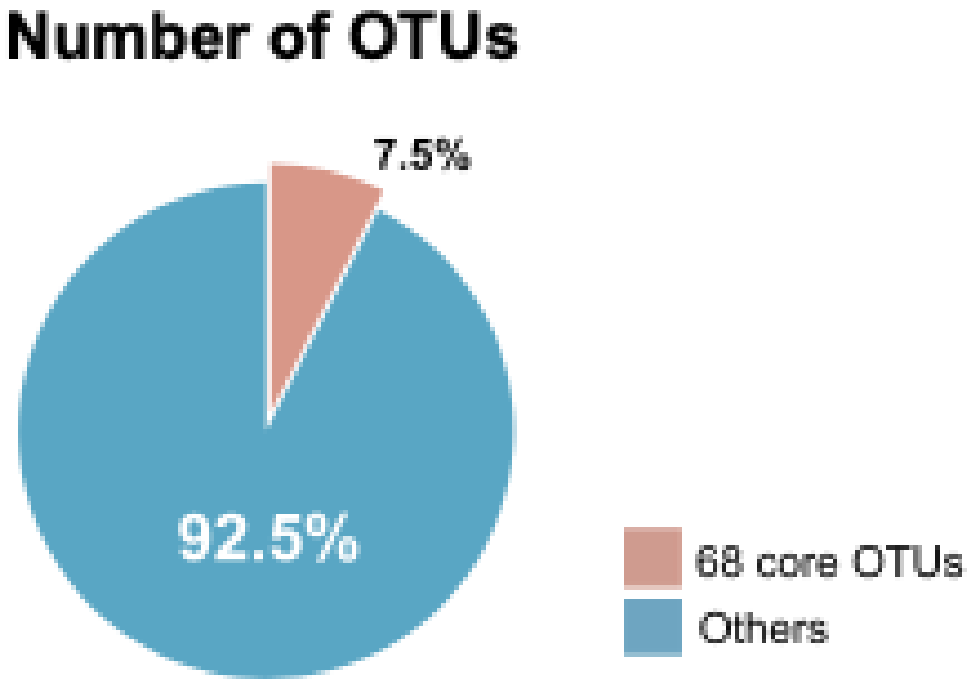


method with many false positives



The many 0s in microbiome studies

- microbiome taxonomic data matrices are full of 0s
  - ➢ becau
    - ≠ tra
    - ther
  - ➢ micro
    - e.g



y the genome)

e in one place

ferent ecosystems

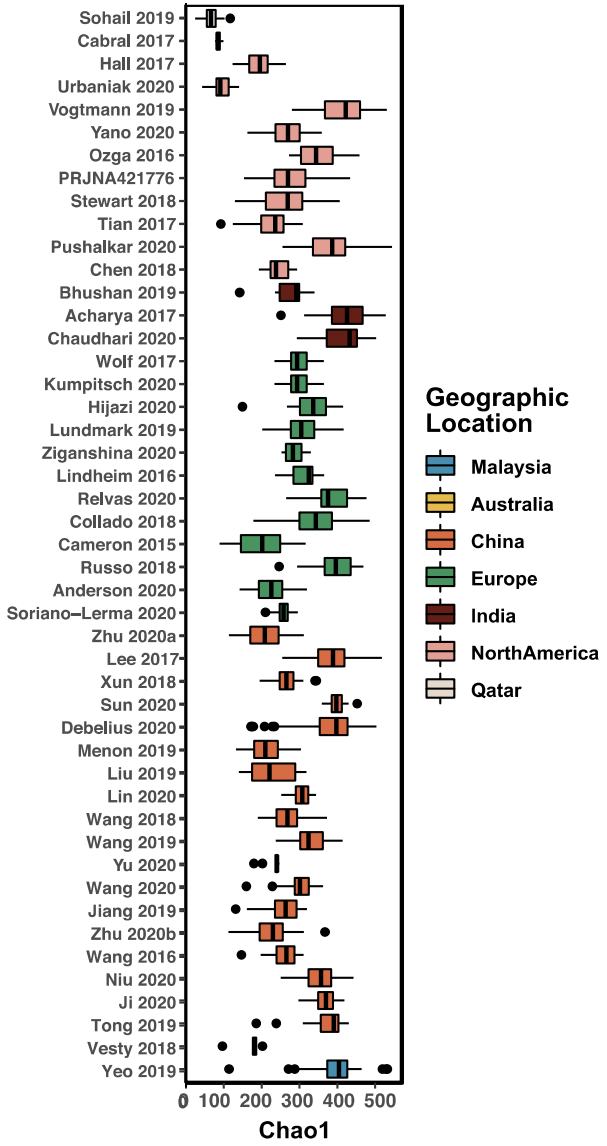
ic environment

Ruan *et al.* 2022
npj biofilms & microbiomes

non-0s: richness & α-diversity

microbiome method

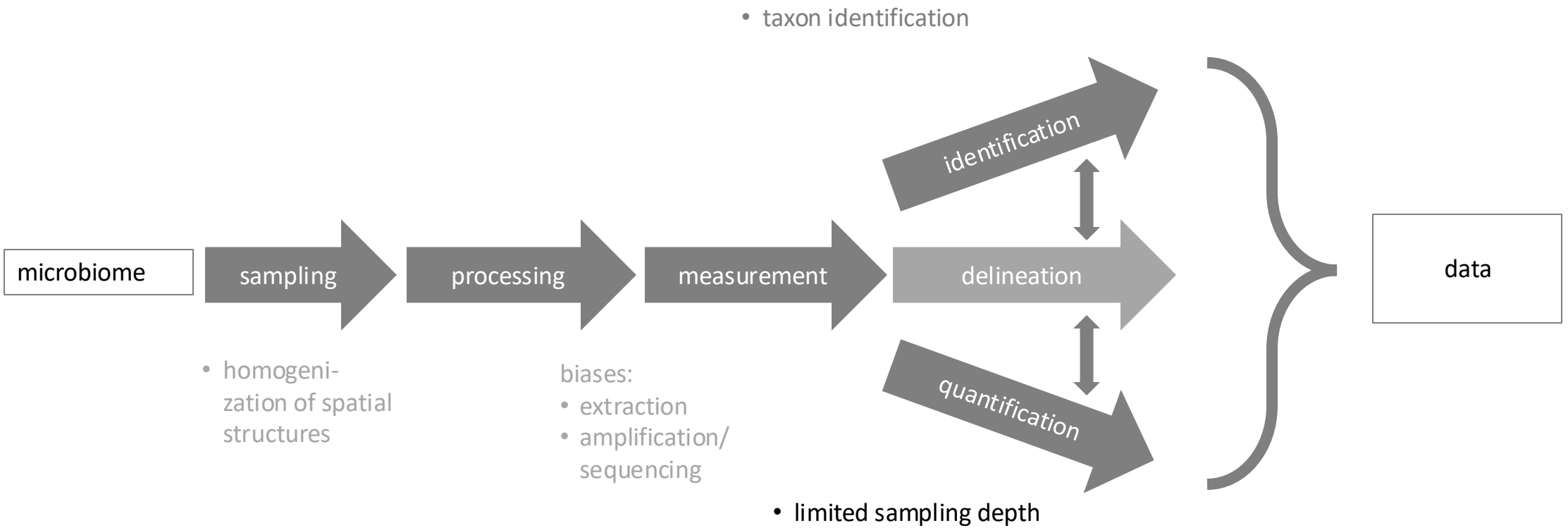
- the number of taxa in a sample (=richness) can have ecological meaning



adapted from Ruan *et al.* 2022 npj biofilms & microbiomes

"α-diversity": diversity of taxa within a sample

Measuring microbiomes: challenges

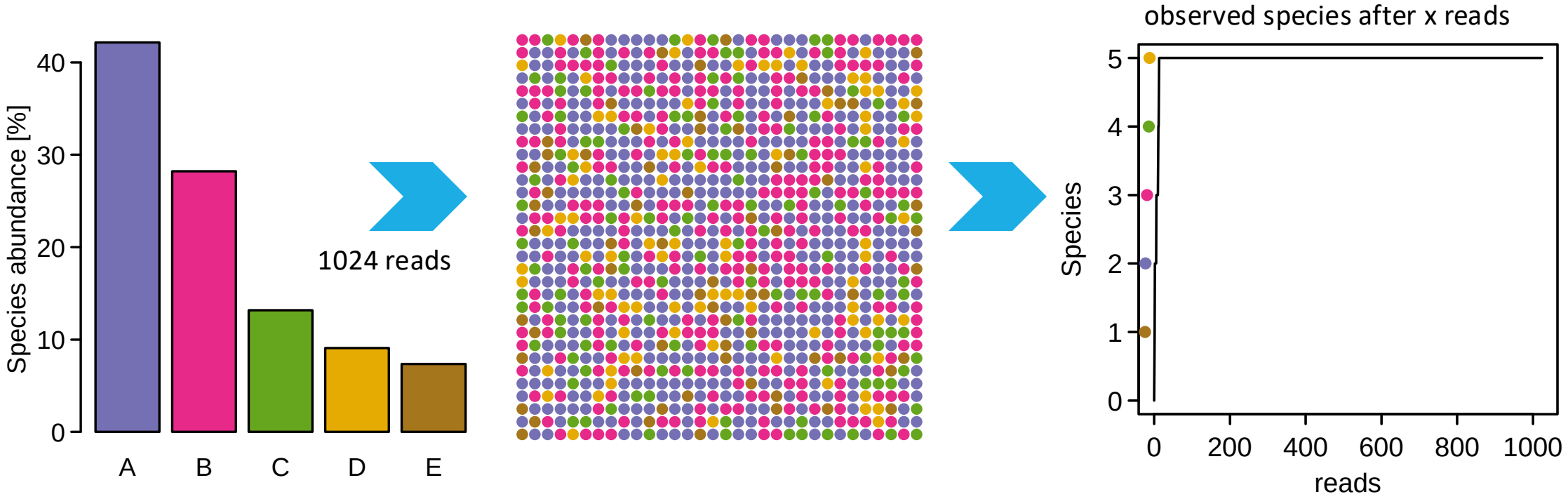


Sampling depth:
Rarefaction curves & analysis

- Did I see all the taxa in my sample?

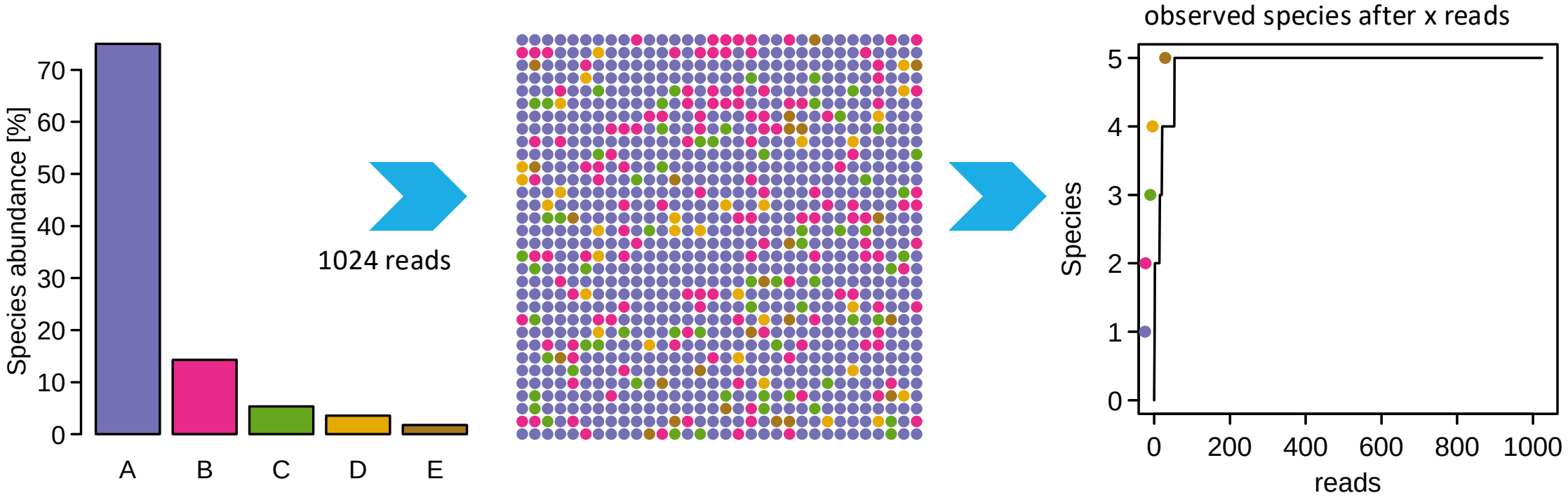
Rarefaction curves & analysis

- number of new species in increasing subsamples



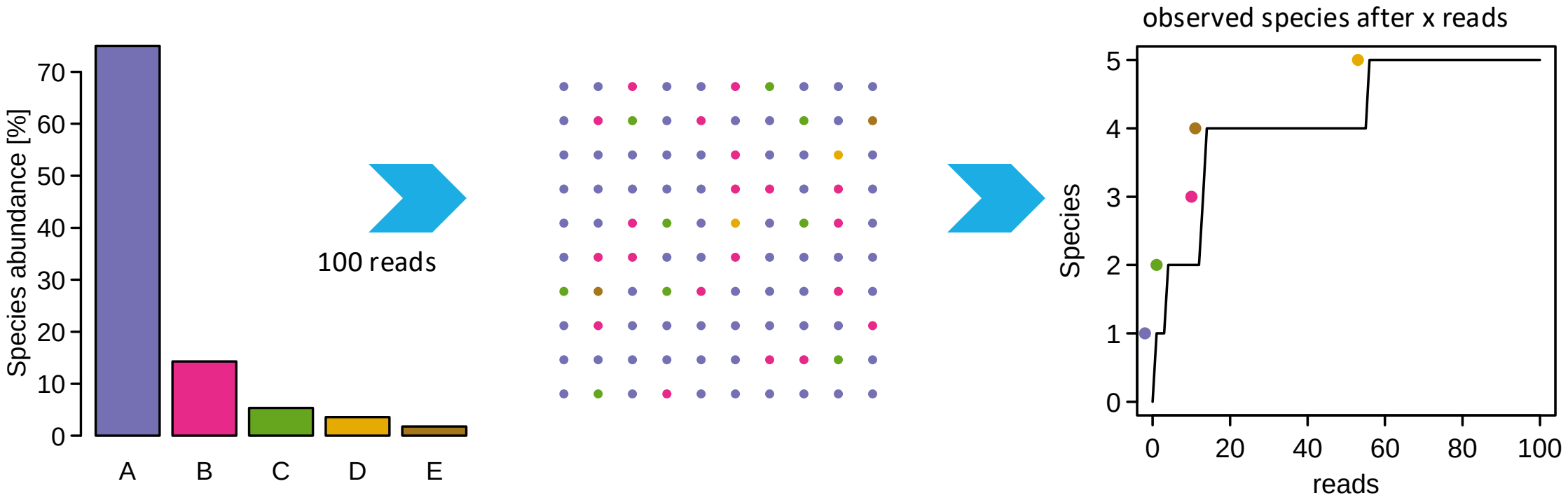
Rarefaction curves & analysis

- number of new species in increasing subsamples



Rarefaction curves & analysis

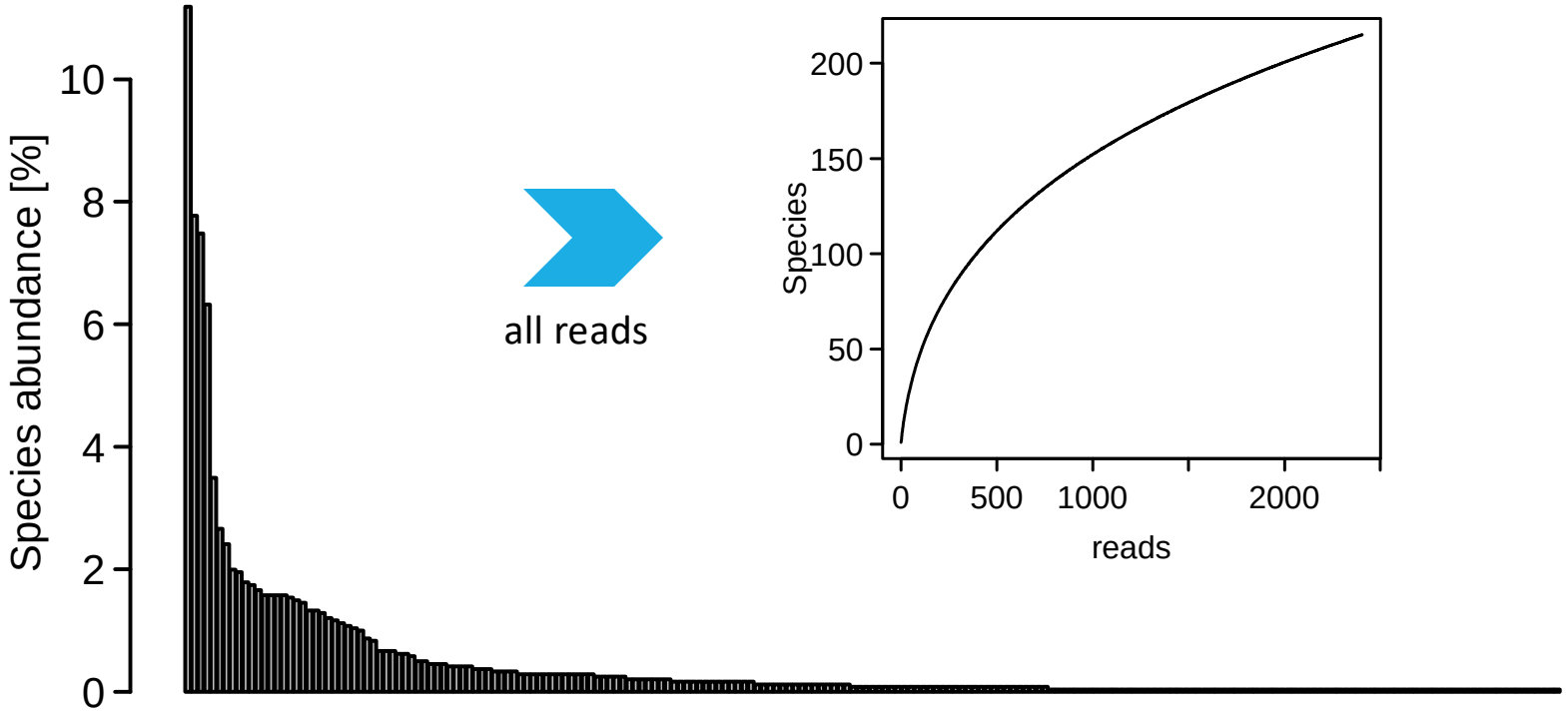
- number of new species in increasing subsamples



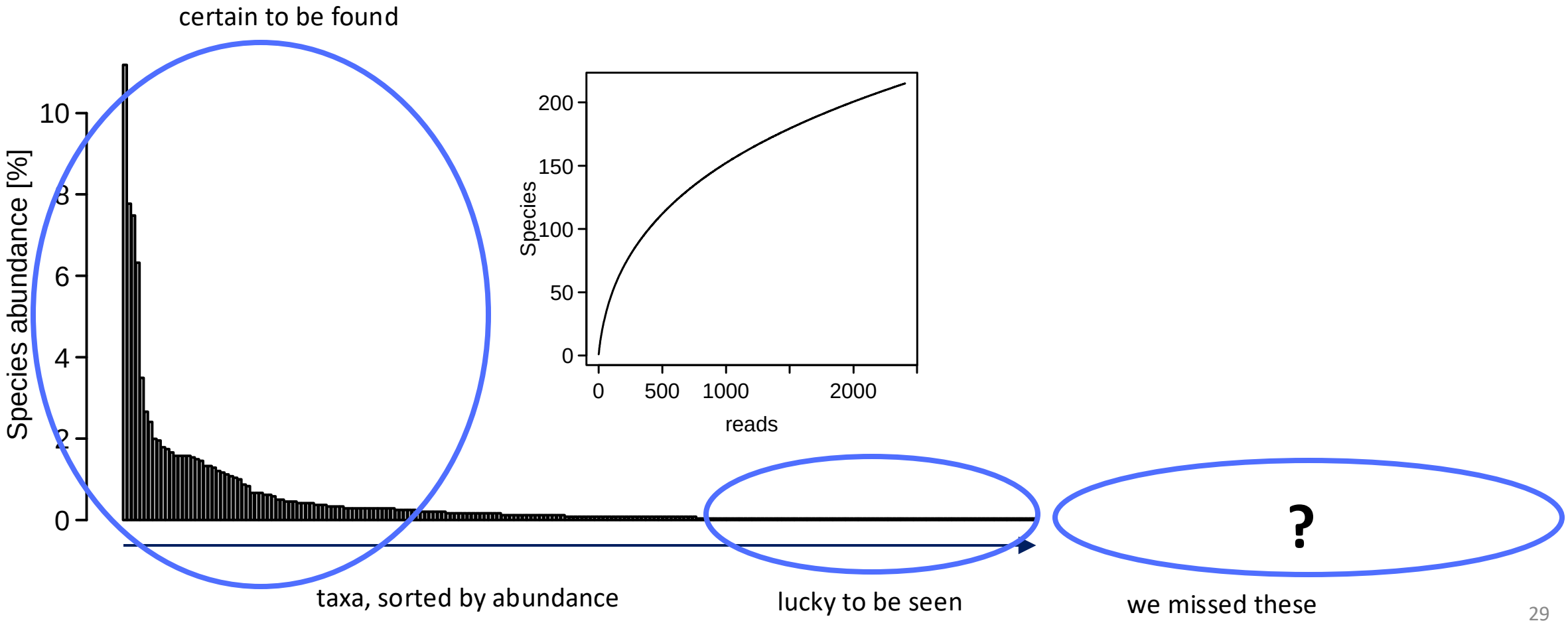
microbiome method

Rarefaction curves & analysis

- number of new species in increasing subsamples

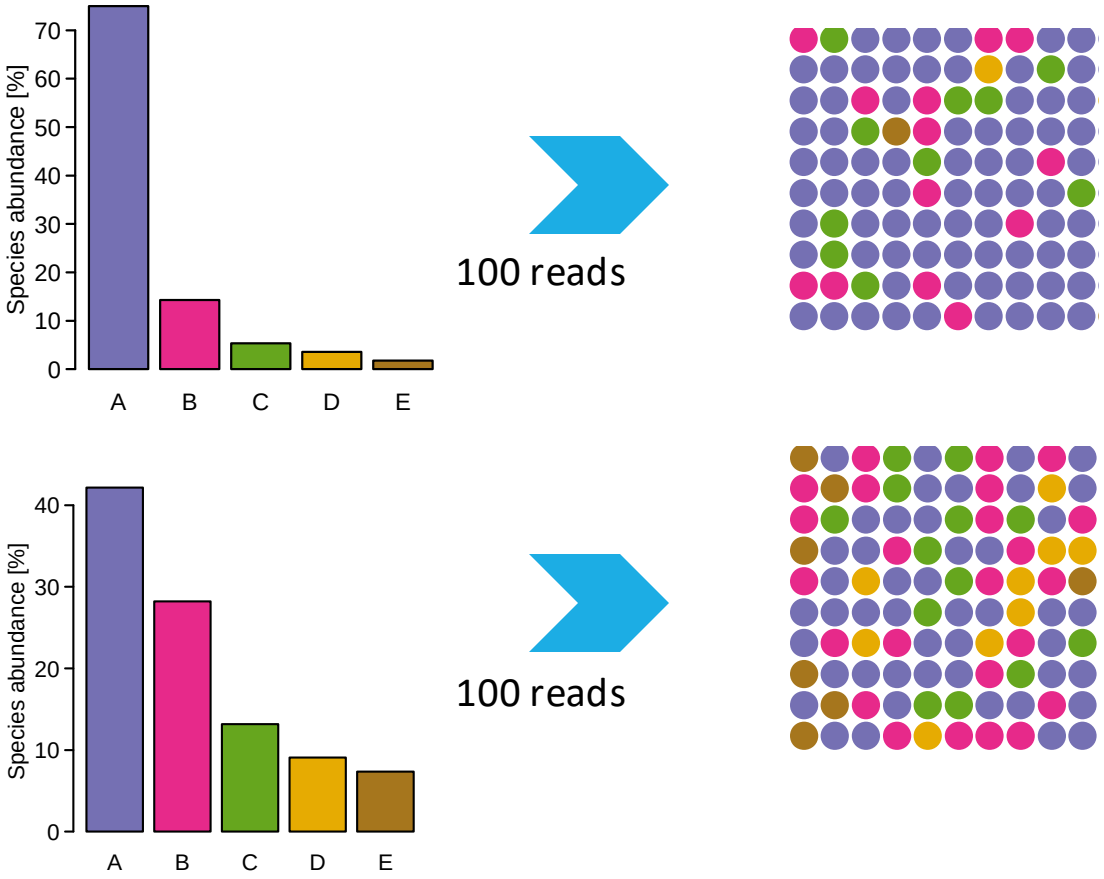


Sampling depth:
Rarefaction curves & analysis



Rarefaction (normalisation)

- subsample reads to keep equal numbers per sample



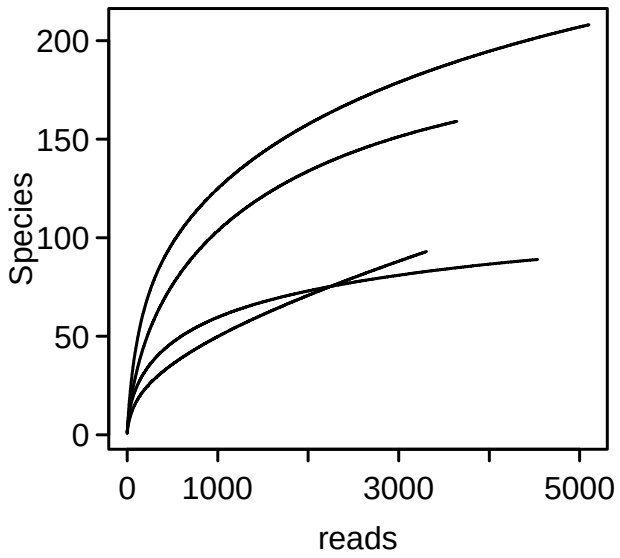
- all samples have the same sum of reads

microbiome method

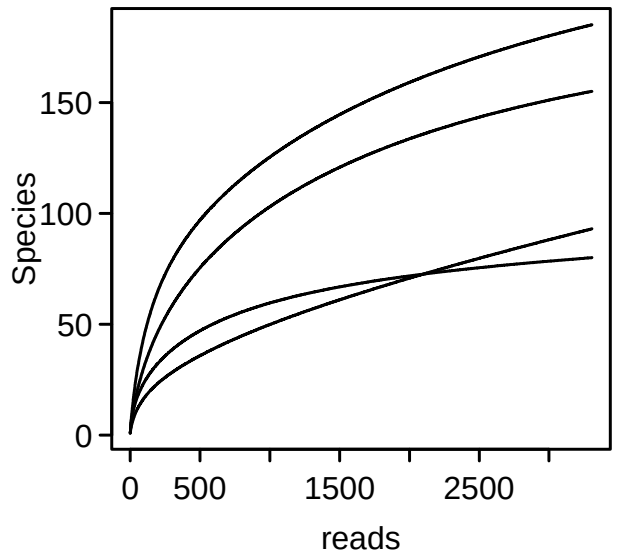
Rarefaction (normalisation)

- subsample reads to keep equal numbers per sample
- you can re-do this many times to estimate what error it introduces

without:

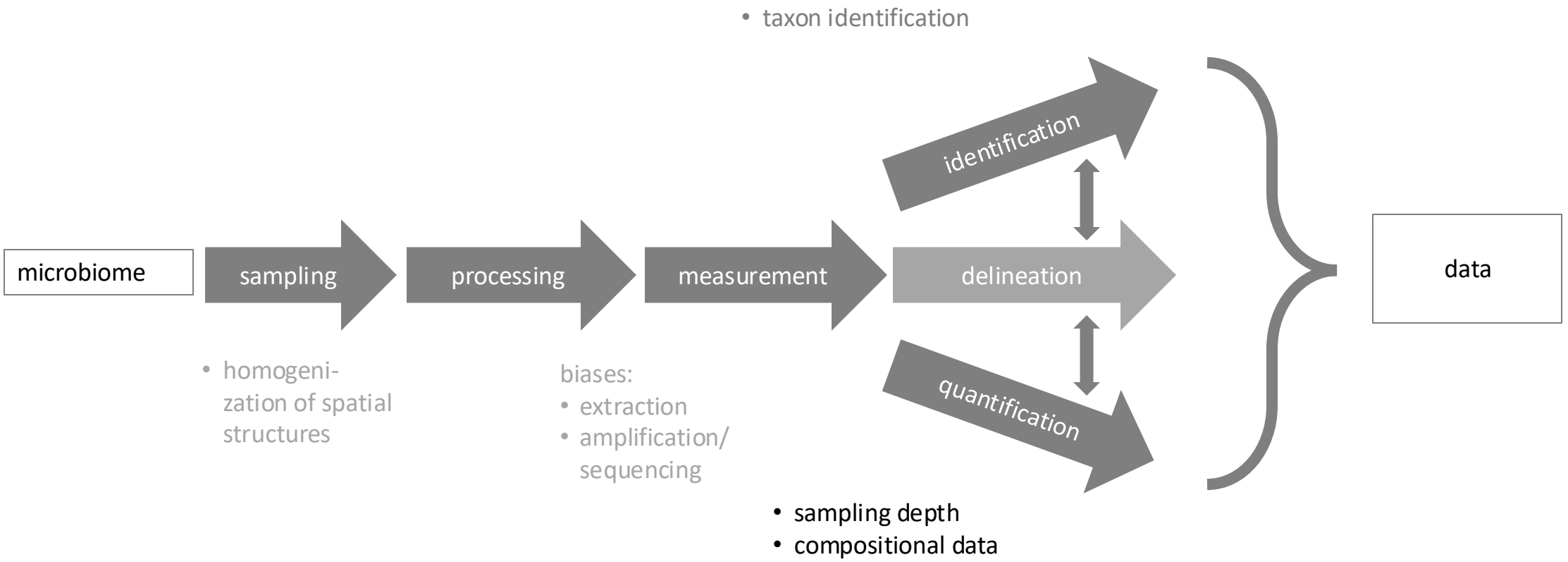


with rarefaction:



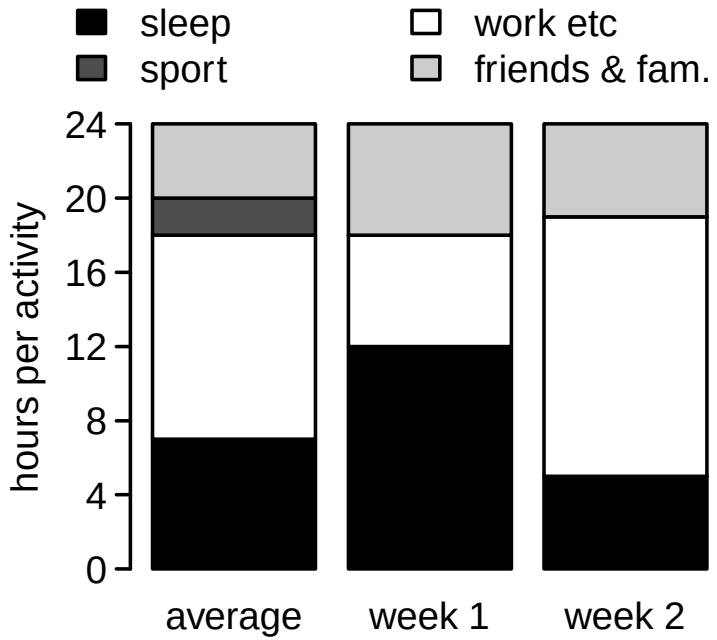
- all samples have the same sum of reads

Measuring microbiomes: challenges



Compositionality

- compositional data consist of vectors whose components are the proportion of some whole



Compositionality

- compositional data consist of vectors whose components are the proportion of some whole

$$\mathbf{x} = (x_1, x_2, ...x_s, ..., x_D)$$

- the proportions are constrained to a constant (unit-sum-constraint)

$$x_1 + x_2 + ... + x_s + ... + x_D = 1$$

- the sample space of the proportions vector is not Euclidean

Compositionality

microbiome method

- **use ratios** between variables instead of proportions of the whole

- ratio to the geometric mean

  = centered log ratio:

$$\mathrm{clr}(\mathbf{x}_j) = \left[\ln \frac{x_{1j}}{g(\mathbf{x}_j)}, \ldots, \ln \frac{x_{Dj}}{g(\mathbf{x}_j)}\right]$$

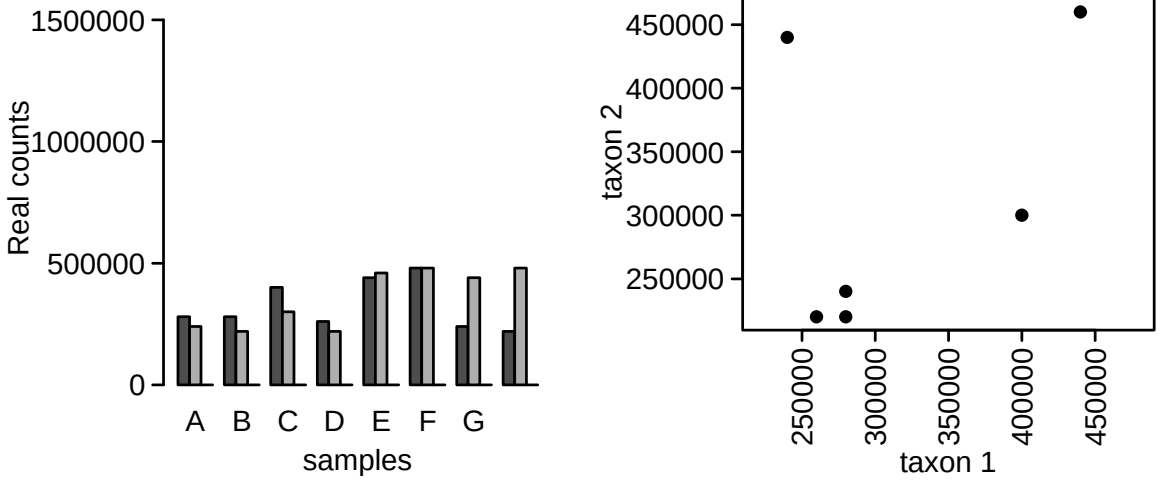
- **transform ratios to log-scale for better handling**

- handle 0s by replacing them with a low value

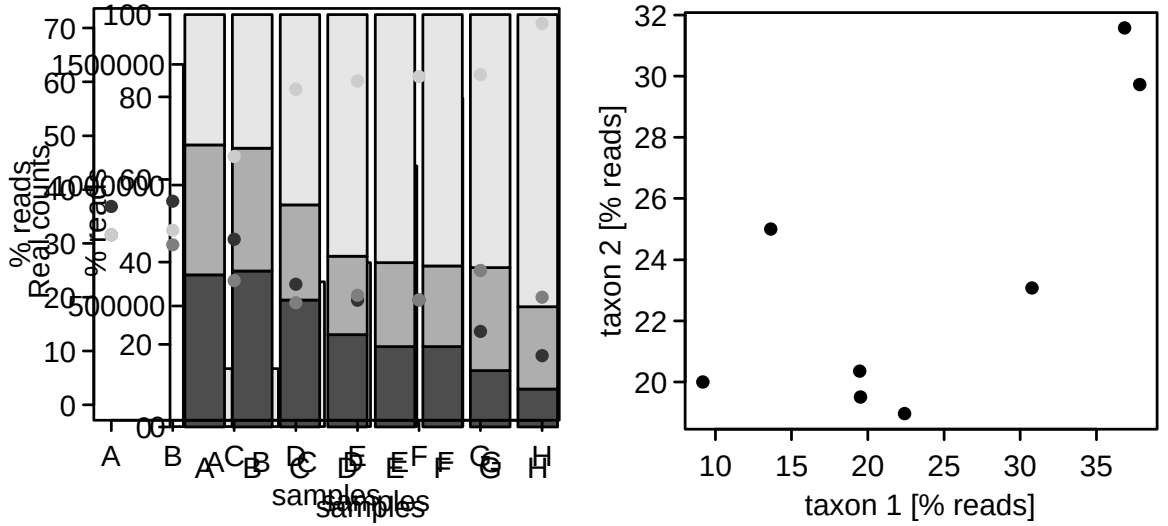
- replacing 0s by some value is always problematic

Compositionality - example

- taxa 1 and 2 have no special relationship

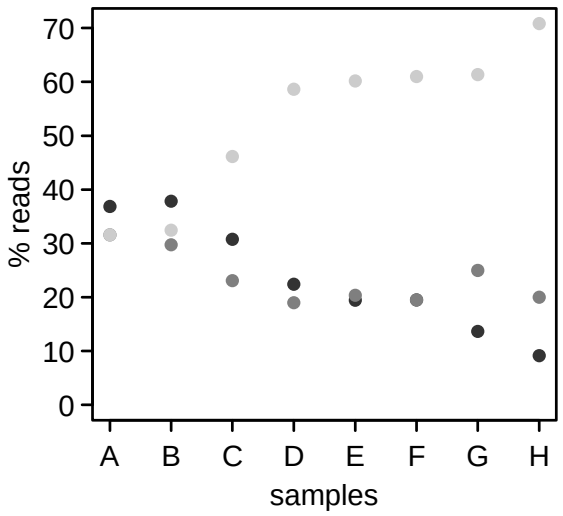


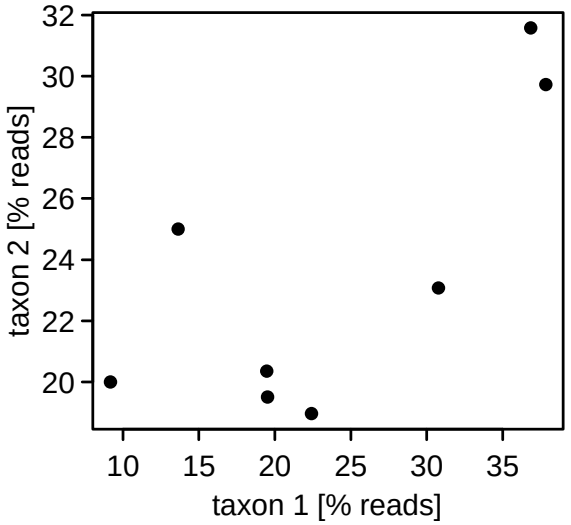
- taxon 3 **introduces a positive correlation**



Compositionality - example

- taxon 3 introduces a positive correlation

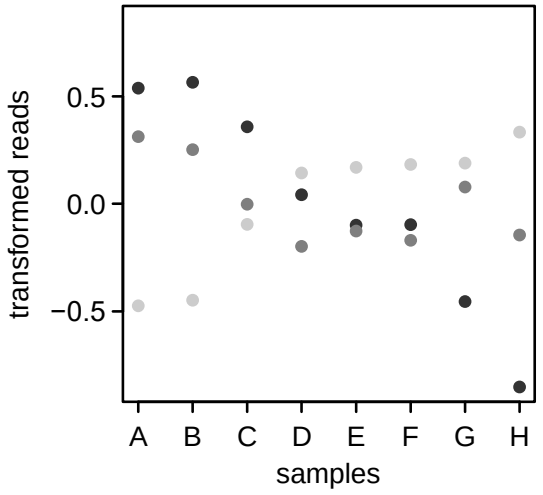


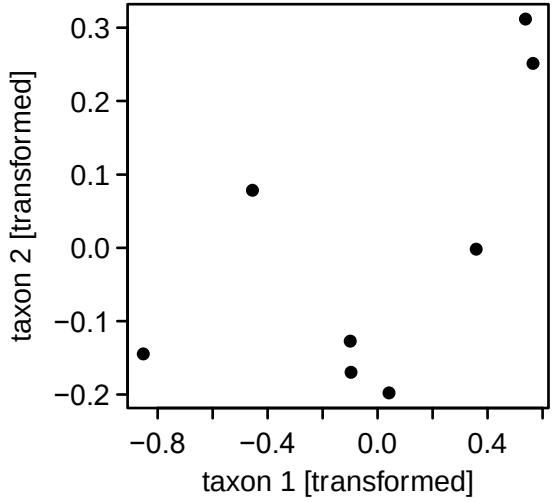


- transformation:

  centered log ratio

$$\mathrm{clr}(\mathbf{x}_j) = \left[\ln \frac{x_{1j}}{g(\mathbf{x}_j)}, ..., \ln \frac{x_{Dj}}{g(\mathbf{x}_j)}\right]$$





- transformation has removed/weakened the spurious correlation

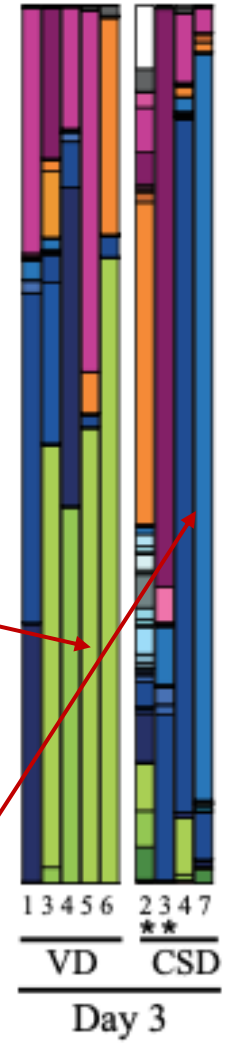
Compositionality – transformation can't do magic

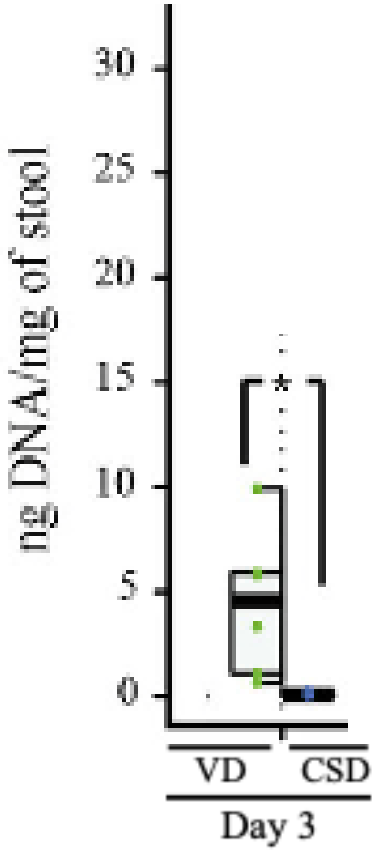
- in microbiome data, the constrained sum is often not representative of anything we know
- we need to measure the total to be sure

are these missing in the other samples?

OR

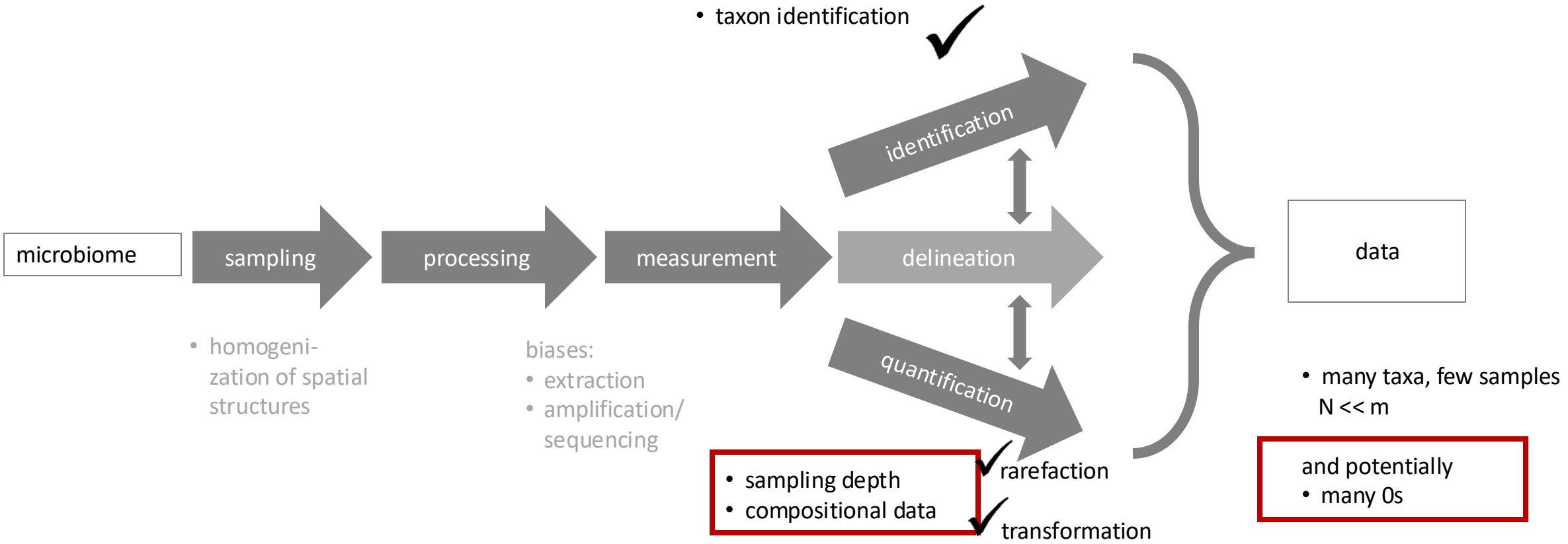
are these a lot more abundant in here?





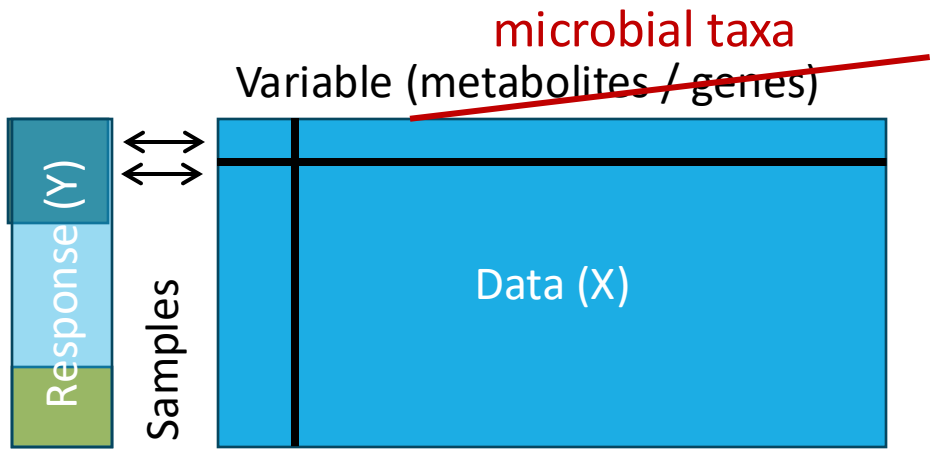
Wampach *et al.* 2017
Frontiers in Microbiol

Measuring microbiomes: challenges



Measuring microbiomes: taxon abundance profiles

taxon	Bact. 1	Bact. 2	Bact. 3	Bact. 4	Bact. …	Bact. m
Sample 1	1	2	1	4	…	0
Sample 2	2	2	3	2	…	0
Sample 3	1	0	0	1	…	0
Sample …	…	…	…	…	…	…
Sample N	4	1	7	0	…	1

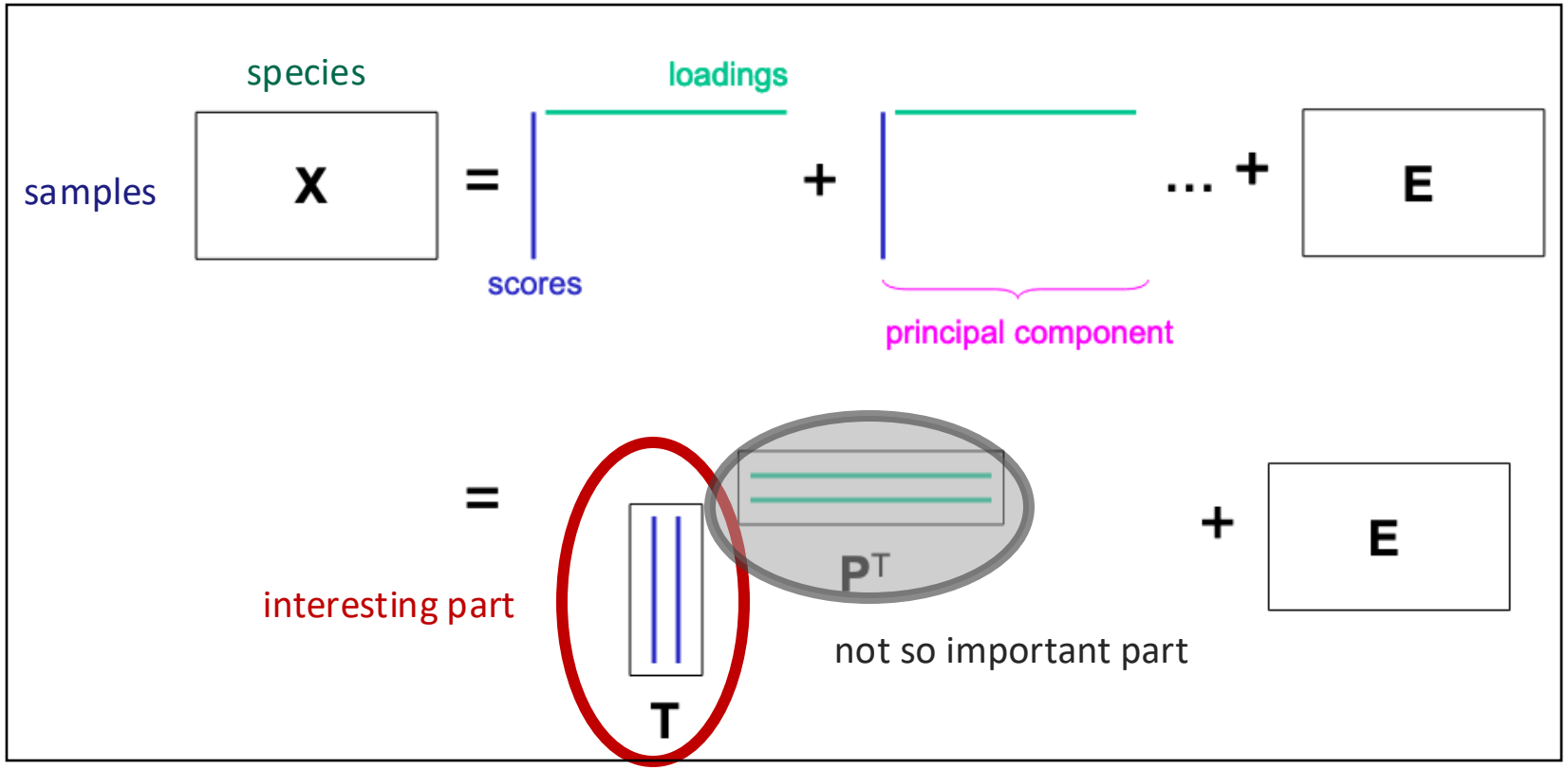


Why don’t most microbiome researchers use many of the methods you’ve learned in the last weeks?

- because they ask different questions
- because of different interpretations of 0s

Ecological questions

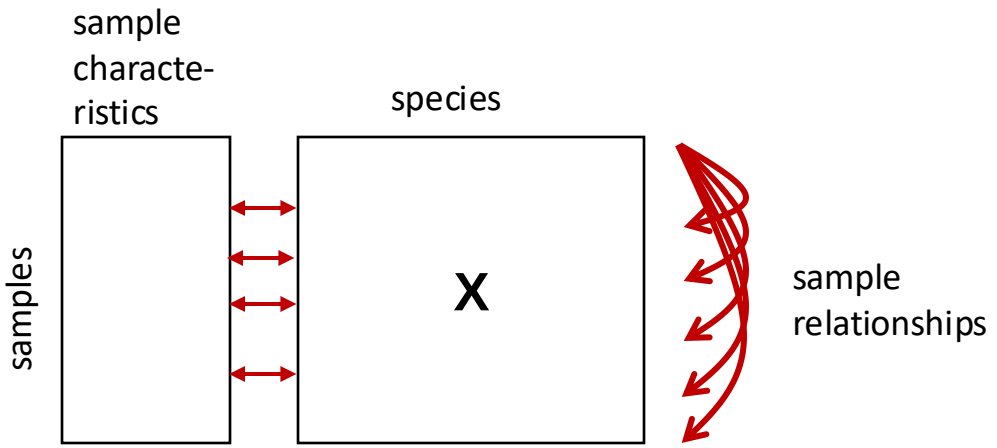
- traditionally, ecology focuses on the **relationship between samples** and not on which species differ between samples



from lecture 3

Ecological questions

- traditionally, ecology does not focus on which species differ between samples
- ask instead how the **relationship between samples relates to environmental factors**



Ecology and 0s

- there are several issues around 0s:

1) why do we think we observe 0s ?
2) what is the meaning / interpretation of *double 0s ?*
3) how dissimilar are two samples that have *no taxa in common ?*

“double 0”: a taxon that is absent in two samples

Double 0s in ecology

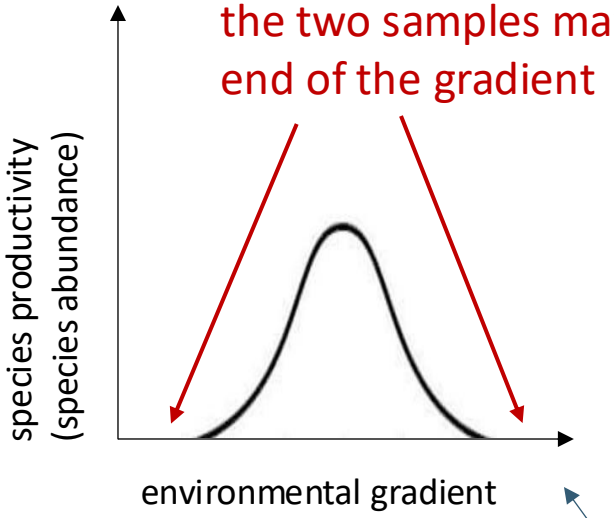
- double 0s are the fields in the matrix where one species is not observed in two samples

taxon	Bact. 1	Bact. 2	Bact. 3
Sample A	**0**	1	1
Sample B	**0**	4	8
Sample C	1	0	0

Double 0s in ecology

- what do double 0s say about sample (dis-)similarity?

taxon	Bact. 1	Bact. 2	Bact. 3
Sample A	**0**	1	1
Sample B	**0**	4	8
Sample C	1	0	0



other reasons include:
dispersal limits, priority effects,
recent extinction by virus/predator

e.g. pH, temperature, nutrient availability….

Recap

observed 0s can mean:

- that the conditions mean too little or too much for a species
- that a species has not (yet) happened to encounter a place
- that a different species has already occupied the niche

- that the species was there, but we missed to observe it

double 0s contain no ecological information

After the break:
Microbiome research has its own set of analysis methods

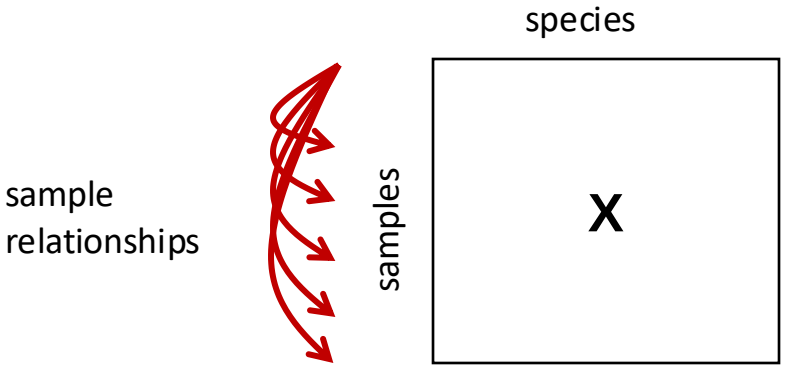
Microbiome research has its own set of analysis methods

Microbiome analysis methods

- ask microbiome questions
- have microbiome interpretations of 0s

Microbiome analysis methods

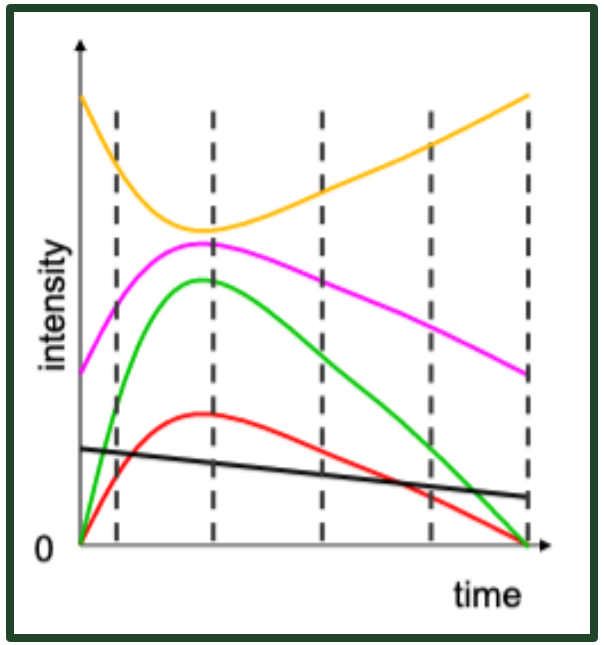
"β-diversity": distance/dissimilarity between samples



β-diversity measures – comparing **pairs of samples**

- what do you consider dissimilar?

- there is not one correct measure,
- **it depends on your question**



from lecture 4

More 0s: Samples without common taxa

- can samples without common taxa be compared?

taxon	Bact. 1	Bact. 2	Bact. 3
Sample A	**0**	1	1
Sample B	**0**	4	8
Sample C	1	0	0

- which samples are more similar?
  - A & B
  - A & C

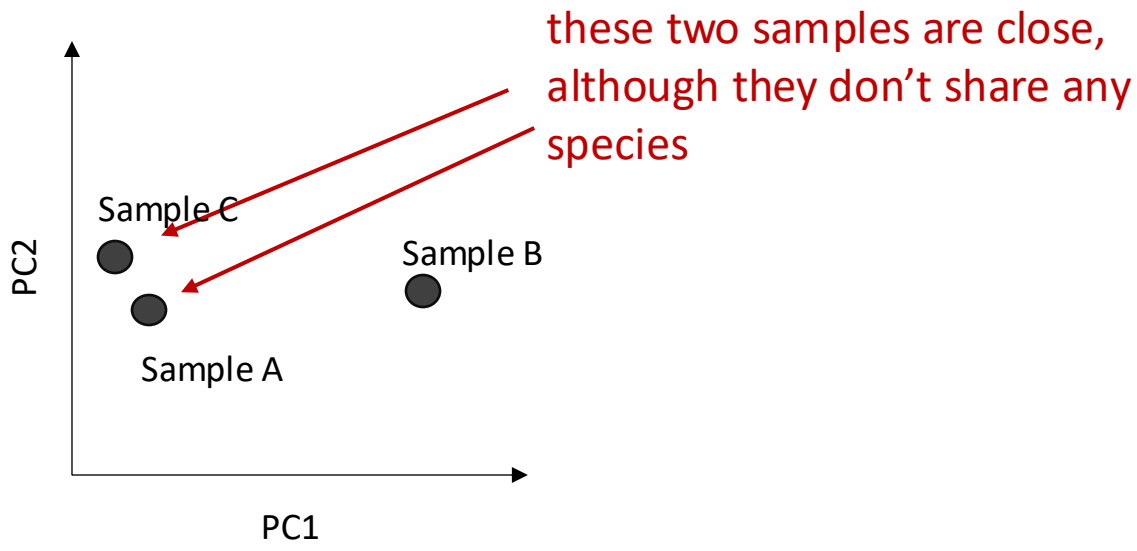
- the ecological interpretation is that C is maximally different from A and B, because they have no common species

β-diversity measures: why not Euclidean distance?

according to the ecological interpretation, C is maximally different from A and B, because they have no common species

taxon	Bact. 1	Bact. 2	Bact. 3
Sample A	**0**	1	1
Sample B	**0**	4	8
Sample C	1	0	0

however, in PCA:



example based on Orlóci, 1978

β-diversity measures: why not Euclidean distance?

- this is not just a normalization issue

taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Σ
Sample A	0	1	16	0	0	0	0	0	0	0	0	0	0	0	0	**17**
Sample B	0	10	5	0	0	0	0	0	0	0	0	0	0	0	0	**15**
Sample C	5	0	0	1	1	1	1	1	1	1	1	1	1	1	1	**18**

β-diversity measures: why not Euclidean distance?

- this is not just a normalization issue

% taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	Σ
Sample A	0	6	94	0	0	0	0	0	0	0	0	0	0	0	0	**100**
Sample B	0	67	33	0	0	0	0	0	0	0	0	0	0	0	0	**100**
Sample C	28	0	0	6	6	6	6	6	6	6	6	6	6	6	6	**100**

$$d_{Euc}(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{s=1}^{D} (x_{si} - x_{sj})^2}$$

D_{Euc}	Sample A	Sample B	Sample C
Sample A	0	0.86	1.009
Sample B	0.86	0	0.827
Sample C	1.009	0.827	0

β-diversity measures: the Aitchison distance

microbiome method

$$d_A(\mathbf{x}_i, \mathbf{x}_j) = \sqrt{\sum_{s=1}^{D} \left[\ln \frac{x_{si}}{g(\mathbf{x}_i)} - \ln \frac{x_{sj}}{g(\mathbf{x}_j)} \right]^2}$$

$$g(\mathbf{x}) = \sqrt[D]{x_1 x_2 \ldots x_D} \text{ is the geometric mean}$$

- this is the Euclidean distance of the centered log ratio transformed data
- we need to add pseudo-counts

taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sample A	1	1	16	1	1	1	1	1	1	1	1	1	1	1	1
Sample B	1	10	5	1	1	1	1	1	1	1	1	1	1	1	1
Sample C	5	1	1	1	1	1	1	1	1	1	1	1	1	1	1

D_A	Sample A	Sample B	Sample C
Sample A	0	1.99	3.82
Sample B	1.99	0	3.95
Sample C	3.82	3.95	0

β-diversity measures: the Bray-Curtis dissimilarity

microbiome method

$$d_{BC}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{s=1}^{D} |x_{si} - x_{sj}|}{\sum_{s=1}^{D} (x_{si} + x_{sj})}$$

% taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Sample A	0	6	94	0	0	0	0	0	0	0	0	0	0	0	0	
Sample B	0	67	33	0	0	0	0	0	0	0	0	0	0	0	0	
Sample C	28	0	0	6	6	6	6	6	6	6	6	6	6	6	6	Σ
numerator	0	61	61	0	0	0	0	0	0	0	0	0	0	0	0	122
denominator	0	73	127	0	0	0	0	0	0	0	0	0	0	0	0	200

$d_{BC}(\mathbf{x}_A, \mathbf{x}_B)$

D_{BC}	Sample A	Sample B	Sample C
Sample A	0	0.61	
Sample B	0.61	0	
Sample C			0

β-diversity measures: the Bray-Curtis dissimilarity

microbiome method

$$d_{BC}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\sum_{s=1}^{D} |x_{si} - x_{sj}|}{\sum_{s=1}^{D} (x_{si} + x_{sj})}$$

- numerator and denominator are equal, if no taxa are in common

% taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	
Sample A	0	6	94	0	0	0	0	0	0	0	0	0	0	0	0	
Sample B	0	67	33	0	0	0	0	0	0	0	0	0	0	0	0	
Sample C	28	0	0	6	6	6	6	6	6	6	6	6	6	6	6	Σ
numerator	28	6	94	6	6	6	6	6	6	6	6	6	6	6	6	200
denominator	28	6	94	6	6	6	6	6	6	6	6	6	6	6	6	200

$d_{BC}(\mathbf{x}_A, \mathbf{x}_C)$

D_{BC}	Sample A	Sample B	Sample C
Sample A	0	0.61	1
Sample B	0.61	0	1
Sample C	1	1	0

β-diversity measures: the binary Sørensen dissimilarity

microbiome method

$$d_{Sor}(\mathbf{x}_i, \mathbf{x}_j) = 1 - \frac{2a}{2a + b + c}$$

coding	x_j present	x_j absent
x_i present	*a*	*b*
x_i absent	*c*	*d*

- we're only counting whether samples have taxa in common

taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sample A	0	1	16	0	0	0	0	0	0	0	0	0	0	0	0
Sample B	0	10	5	0	0	0	0	0	0	0	0	0	0	0	0
Sample C	5	0	0	1	1	1	1	1	1	1	1	1	1	1	1

	a	*b*	c
A,B	2	0	0
A,C	0	13	13
B,C	0	13	13

$$d_{Sor}(\mathbf{x}_A, \mathbf{x}_B) = 1 - \frac{2 \cdot 2}{2 \cdot 2 + 0 + 0}$$

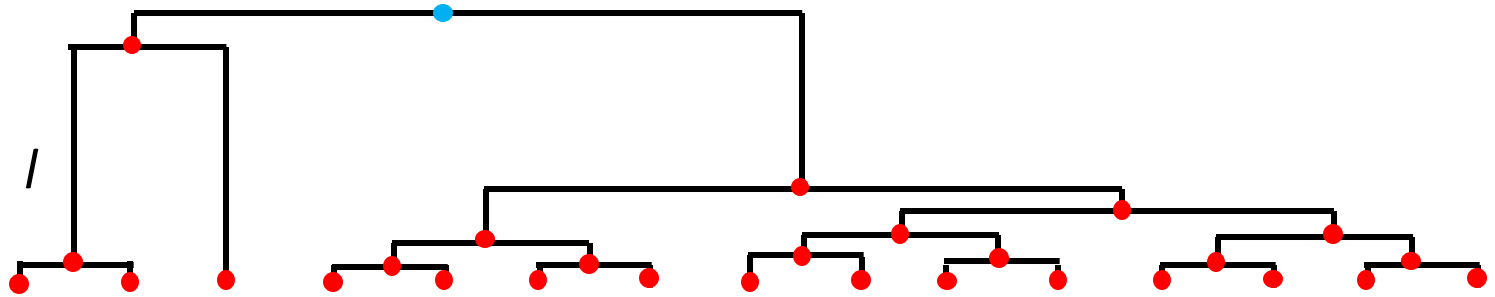
$$d_{Sor}(\mathbf{x}_A, \mathbf{x}_C) = 1 - \frac{2 \cdot 0}{2 \cdot 0 + 13 + 13}$$

D_{Sor}	Sample A	Sample B	Sample C
Sample A	0	0	1
Sample B	0	0	1
Sample C	1	1	0

β-diversity measures: UniFrac distances - 1: unweighted

microbiome method

$$u = \frac{\sum_{n=1}^{N} l_n \, |A_n - B_n|}{\sum_{n=1}^{N} l_n \, max(A_n, B_n)}$$



indicator	present	absent
A_n	1	0
B_n	1	0

taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sample A	0	1	16	0	0	0	0	0	0	0	0	0	0	0	0
Sample B	*1*	10	5	0	0	0	0	0	0	0	0	0	0	0	0
Sample C	5	0	0	1	1	1	1	1	1	1	1	1	1	1	1

- use information on relatedness

  l_n = branchlength

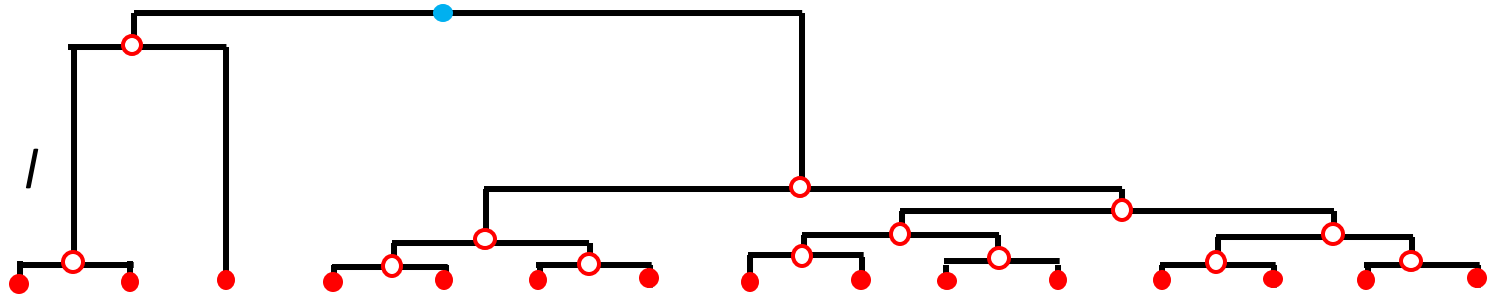
  n…N : nodes

u	Sample A	Sample B	Sample C
Sample A	0	0.05	0.98
Sample B	0.05	0	0.97
Sample C	0.98	0.97	0

β-diversity measures: UniFrac distances - 2: weighted

microbiome method

$$w = \frac{\sum_{n=1}^{N} l_n \left| \frac{A_n}{A_T} - \frac{B_n}{B_T} \right|}{\sum_{s=1}^{S} L_s}$$



taxon	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Sample A	0	1	16	0	0	0	0	0	0	0	0	0	0	0	0
Sample B	1	10	5	0	0	0	0	0	0	0	0	0	0	0	0
Sample C	5	0	0	1	1	1	1	1	1	1	1	1	1	1	1

- use information on relatedness and abundance

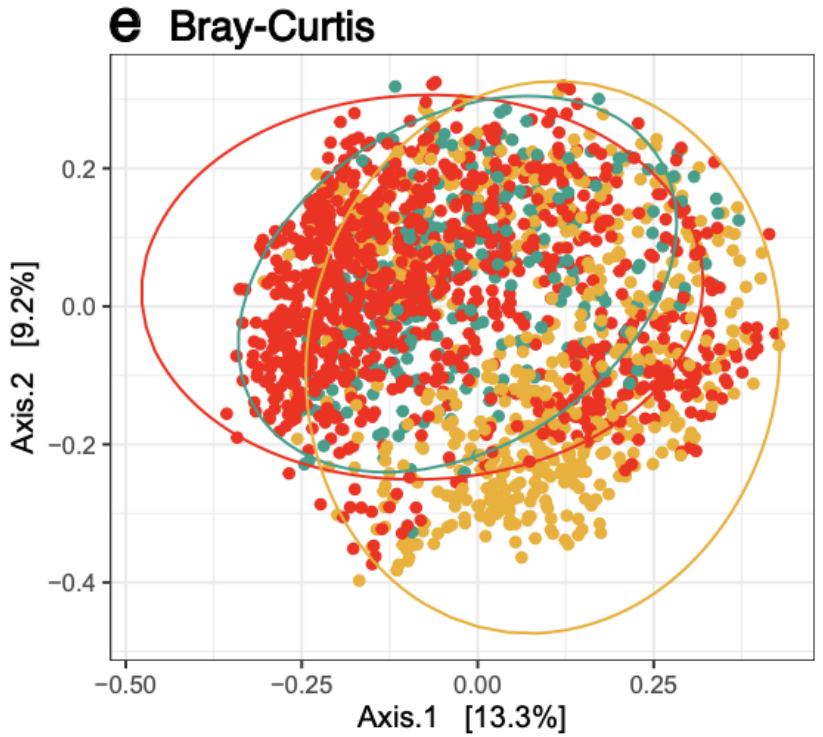
  l_n = branchlength (*n…N*: nodes)

  L_s = length root to tip (*s…S*: tips)

  $\frac{A_n}{A_T}$ $\frac{B_n}{B_T}$ = relative abundances

u	Sample A	Sample B	Sample C
Sample A	0	0.005	1
Sample B	0.005	0	0.98
Sample C	1	0.98	0

And can we represent the whole β-diversity matrix?



Ruan *et al.* 2022 npj biofilms & microbiomes

Representing the whole β-diversity matrix: Principal coordinate analysis

- represents any distance/dissimilarity matrix in Euclidean space

  = “ordination”

- because it starts from the β-diversity/distance/dissimilarity matrix, it does not use (retain) the original variable information

PCoA of β-diversity

microbiome method

- represents the distance matrix in Euclidean space

Δ_1	Sample A	Sample B	Sample C
Sample A	0.1932	0.0085	-0.2017
Sample B	0.0085	0.1932	-0.2017
Sample C	-0.2017	-0.2017	0.4033

eigenvectors	*v1*	*v2*	*v3*
Sample A	-0.4082	-0.7071	0.5774
Sample B	-0.4082	0.7071	0.5774
Sample C	0.8165	0	0.5774

eigenvalues λ	0.6051	0.1847	0

- transform matrix

$$\mathbf{A} = -\frac{1}{2}\mathbf{D}^2$$

- centre rows and columns

$$\mathbf{\Delta}_1 = \left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)\mathbf{A}\left(\mathbf{I} - \frac{\mathbf{1}\mathbf{1}^T}{n}\right)$$

- calculate eigenvectors & eigenvalues

"PCoA": Principal coordinate analysis

microbiome method

PCoA of β-diversity

- represents the distance matrix in Euclidean space

Δ_1	Sample A	Sample B	Sample C
Sample A	0.1932	0.0085	-0.2017
Sample B	0.0085	0.1932	-0.2017
Sample C	-0.2017	-0.2017	0.4033

- transform matrix
- centre rows and columns

eigenvectors	*v1*	*v2*	*v3*
Sample A	-0.4082	-0.7071	0.5774
Sample B	-0.4082	0.7071	0.5774
Sample C	0.8165	0	0.5774
eigenvalues λ	**0.6051**	**0.1847**	**0**
square root √λ	**0. 7779**	**0.4298**	**0**

- calculate eigenvectors & eigenvalues
- multiply eigenvectors by squareroot of eigenvalues

"PCoA": Principal coordinate analysis

microbiome method

PCoA of β-diversity

- represents the distance matrix in Euclidean space

Δ_1	Sample A	Sample B	Sample C
Sample A	0.1932	0.0085	-0.2017
Sample B	0.0085	0.1932	-0.2017
Sample C	-0.2017	-0.2017	0.4033

U norm. eigenvectors	*u*1	*u*2
Sample A	-0.3176	-0.3039
Sample B	-0.3176	0.3039
Sample C	0.6351	0

eigenvalues λ	0.6051	0.1847	0
square root √λ	0. 7779	0.4298	0

- transform matrix
- centre rows and columns

- calculate eigenvectors & eigenvalues
- multiply eigenvectors by squareroot of eigenvalues

"PCoA": Principal coordinate analysis

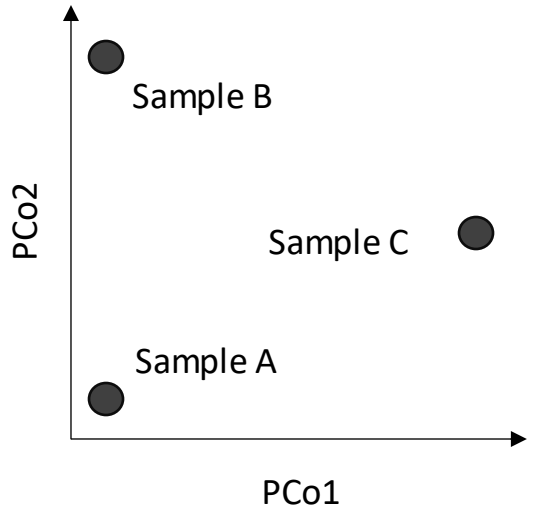
microbiome method

PCoA of β-diversity

- represents the distance matrix in Euclidean space

D_{BC}	Sample A	Sample B	Sample C
Sample A	0	0.61	1
Sample B	0.61	0	1
Sample C	1	1	0

U norm. eigenvectors	*u*1	*u*2
Sample A	-0.3176	-0.3039
Sample B	-0.3176	0.3039
Sample C	0.6351	0



"PCoA": Principal coordinate analysis

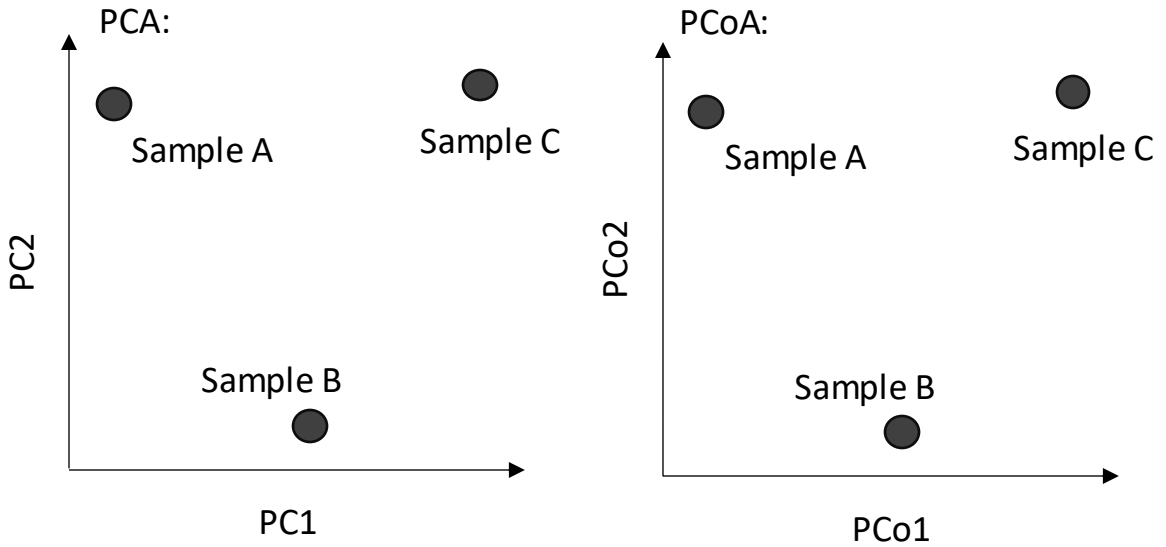
PCoA of β-diversity

microbiome method

- represents the distance matrix in Euclidean space

D_{Euc}	Sample A	Sample B	Sample C
Sample A	0	0.86	1.009
Sample B	0.86	0	0.827
Sample C	1.009	0.827	0

- if the matrix already holds Euclidean distances, the normalised eigenvectors are the same as the PCA-scores on the original dataset



"PCoA": Principal coordinate analysis – for a Euclidean distance matrix, the objects find the same place as in PCA

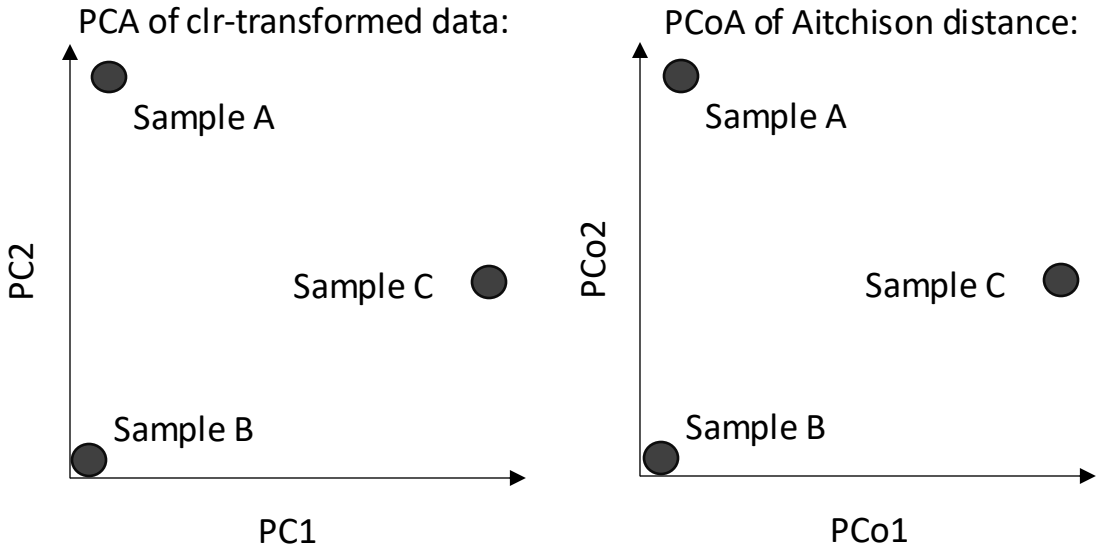
PCoA of β-diversity

microbiome method

- represents the distance matrix in Euclidean space

D_A	Sample A	Sample B	Sample C
Sample A	0	1.99	3.82
Sample B	1.99	0	3.95
Sample C	3.82	3.95	0

- if the matrix already holds Euclidean distances, the normalised eigenvectors are the same as the PCA-scores on the original dataset



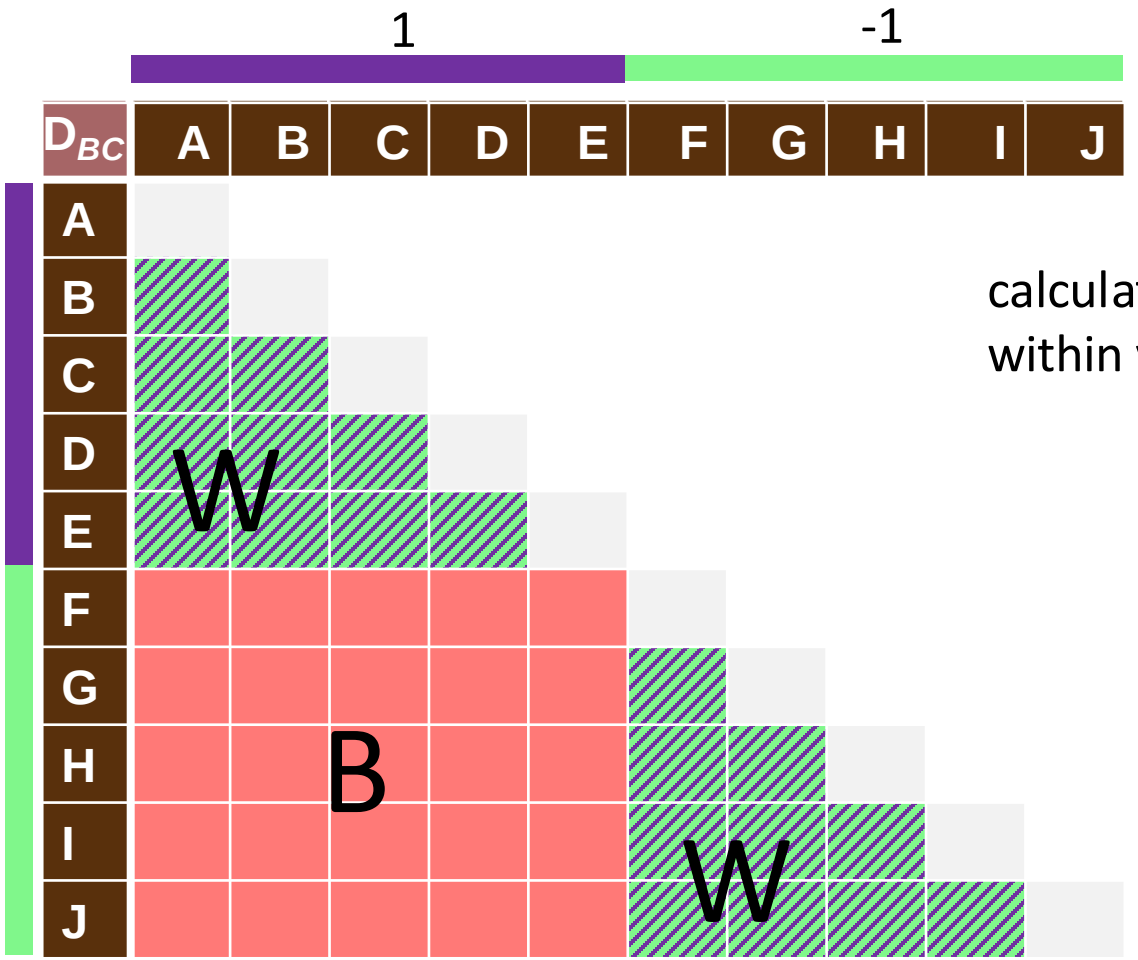
"PCoA": Principal coordinate analysis – for a Euclidean distance matrix, the objects find the same place as in PCA

PCoA of β-diversity - recap

- represents any distance/dissimilarity matrix in Euclidean space
  = “ordination”
- if the matrix holds Euclidean distances, the ordination is the same as the PCA-scores
- does not use information on the original variables
- does not return information on the original variable

Testing effects on β-diversity: PERMANOVA

microbiome method



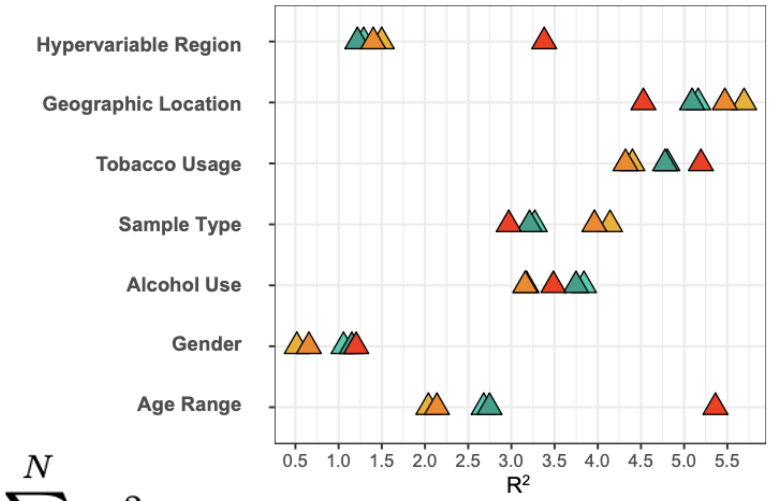
calculate between vs. within variation

$$SS_T = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2$$

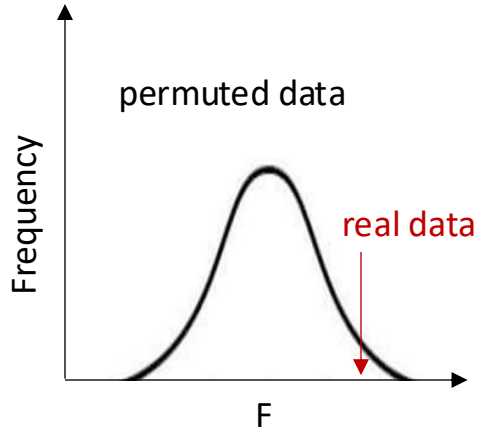
$$SS_W = \frac{1}{n}\sum_{i=1}^{N-1}\sum_{j=i+1}^{N} d_{ij}^2 \in_{ij}$$

$$SS_B = SS_T - SS_W$$

$$F = \frac{\frac{SS_B}{groups-1}}{\frac{SS_W}{N-groups}}$$



Ruan *et al.* 2022 npj biofilms & microbiomes



adapted from Anderson 2001 Austral Ecology

Recap

- Microbiome composition is studied using omics methods
- Unequal sampling depth and compositionality may be addressed by normalization/ transformation
- Microbiome data contains many 0s due to microbiology/ecology
- β-diversity measures emphasize different aspects of sample (dis-)similarity
- Principal coordinate analysis can represent β-diversity matrices in Euclidean space

Practical

- Dataset from the human gut microbiome: inflammatory bowel disease
- 43 individuals in 3 groups
- 7 time points / individual

- questions:
  - alpha-diversity
  - 0s
  - compositionality
  - comparison to previous methods
  - ordination of beta-diversity
  - comparison of beta-diversity
  - differential abundance analysis

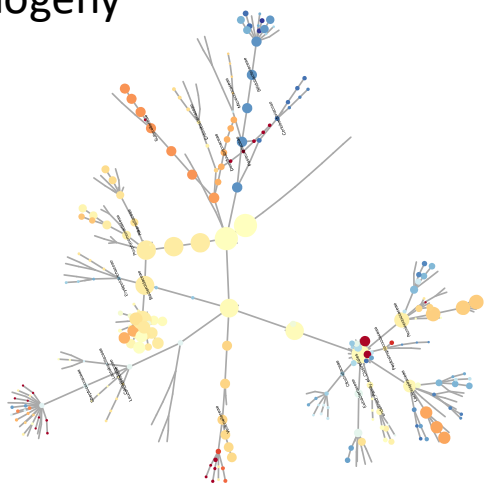
End of BDA lectures

please do the evaluation!

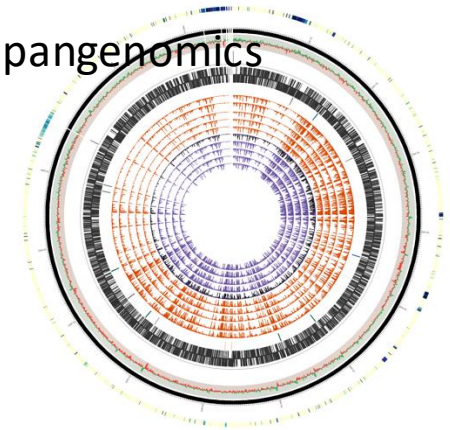
Microbiomes: other analyses

- Meta-omics measurements can be used for multiple purposes
- Here are some examples that we did not dicuss today

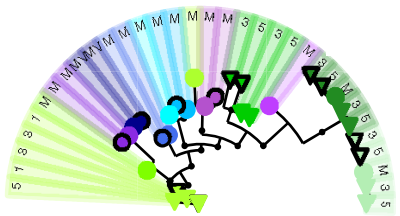
phylogeny



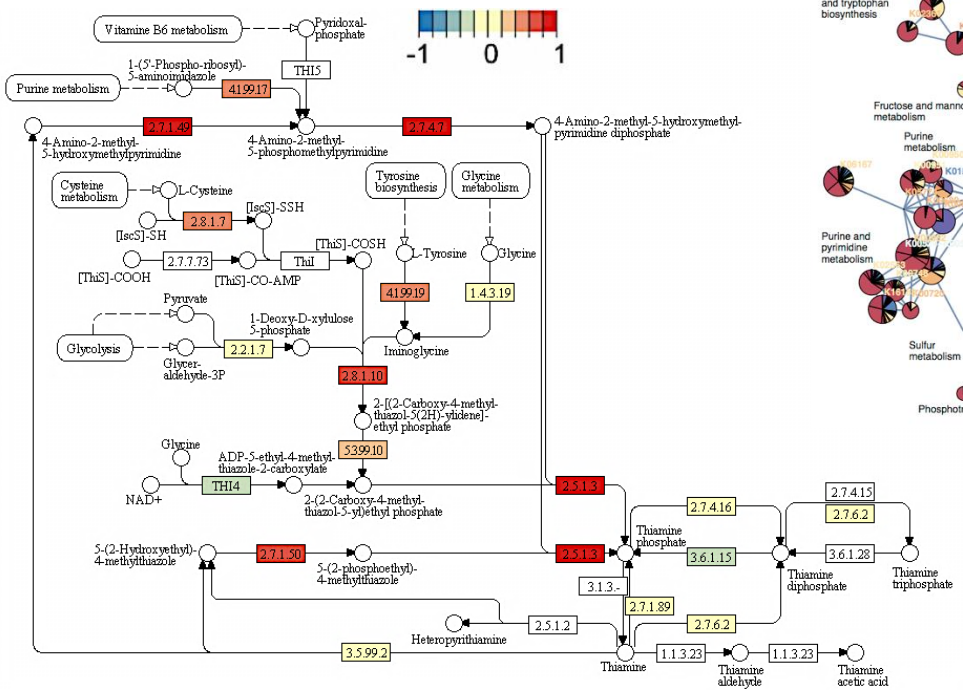
genomics and pangenomics



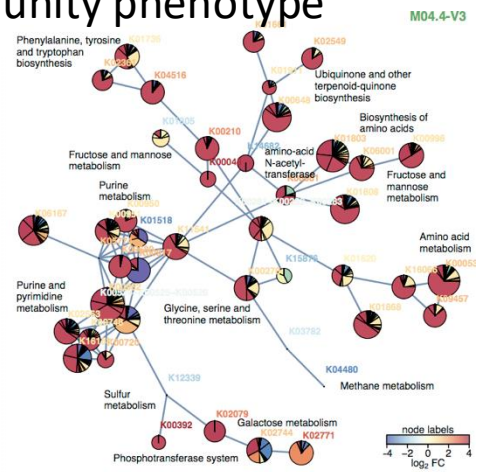
strain-tracking



metabolic activity



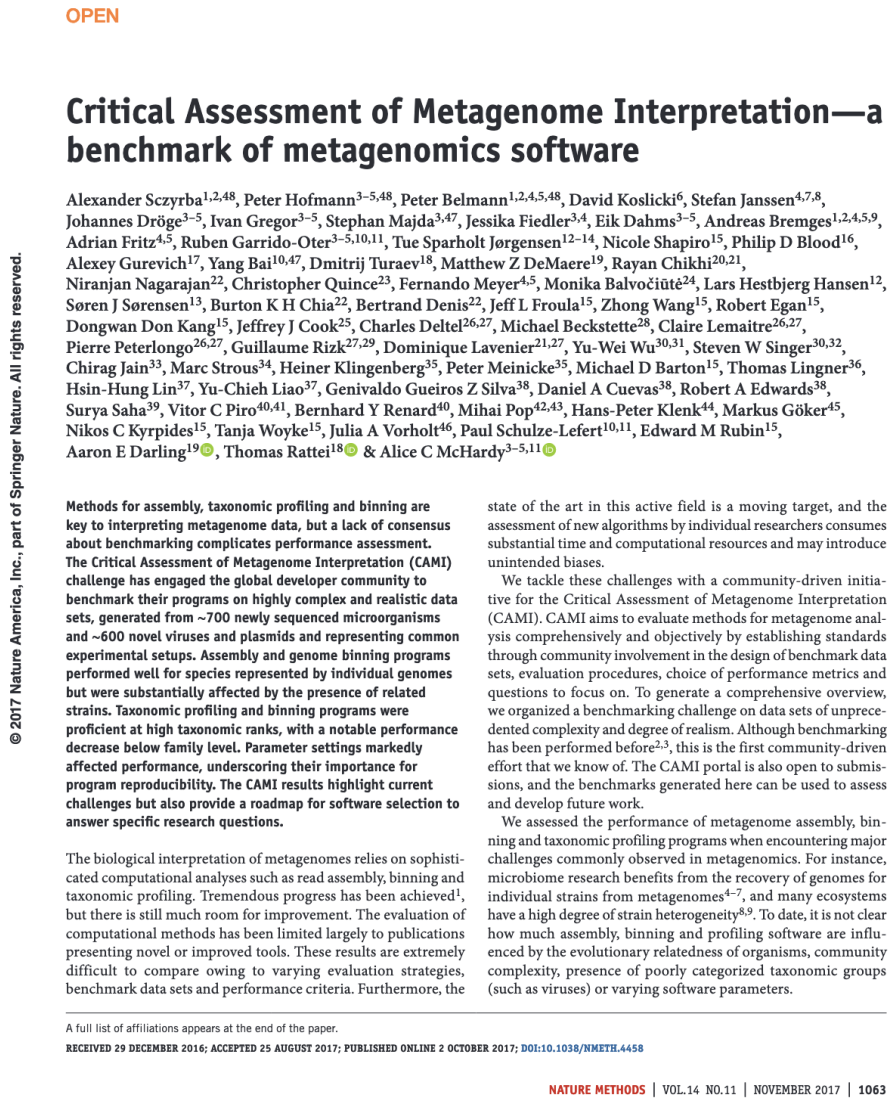
contribution of taxa to community phenotype



Choosing identification methods

"benchmarking":

- testing different algorithms on the same data set
- data sets are simulated, so we know what would be a perfect outcome "ground truth"



OPEN

Critical Assessment of Metagenome Interpretation—a benchmark of metagenomics software

Alexander Sczyrba[1,2,48], Peter Hofmann[3–5,48], Peter Belmann[1,2,4,5,48], David Koslicki[6], Stefan Janssen[4,7,8], Johannes Dröge[3–5], Ivan Gregor[3–5], Stephan Majda[3,47], Jessika Fiedler[3,4], Eik Dahms[3–5], Andreas Bremges[1,2,4,5,9], Adrian Fritz[4,5], Ruben Garrido-Oter[3–5,10,11], Tue Sparholt Jørgensen[12–14], Nicole Shapiro[15], Philip D Blood[16], Alexey Gurevich[17], Yang Bai[10,47], Dmitrij Turaev[18], Matthew Z DeMaere[19], Rayan Chikhi[20,21], Niranjan Nagarajan[22], Christopher Quince[23], Fernando Meyer[4,5], Monika Balvočiūtė[24], Lars Hestbjerg Hansen[12], Søren J Sørensen[13], Burton K H Chia[22], Bertrand Denis[22], Jeff L Froula[15], Zhong Wang[15], Robert Egan[15], Dongwan Don Kang[15], Jeffrey J Cook[25], Charles Deltel[26,27], Michael Beckstette[28], Claire Lemaitre[26,27], Pierre Peterlongo[26,27], Guillaume Rizk[27,29], Dominique Lavenier[21,27], Yu-Wei Wu[30,31], Steven W Singer[30,32], Chirag Jain[33], Marc Strous[34], Heiner Klingenberg[35], Peter Meinicke[35], Michael D Barton[15], Thomas Lingner[36], Hsin-Hung Lin[37], Yu-Chieh Liao[37], Genivaldo Gueiros Z Silva[38], Daniel A Cuevas[38], Robert A Edwards[38], Surya Saha[39], Vitor C Piro[40,41], Bernhard Y Renard[40], Mihai Pop[42,43], Hans-Peter Klenk[44], Markus Göker[45], Nikos C Kyrpides[15], Tanja Woyke[15], Julia A Vorholt[46], Paul Schulze-Lefert[10,11], Edward M Rubin[15], Aaron E Darling[19], Thomas Rattei[18] & Alice C McHardy[3–5,11]

Methods for assembly, taxonomic profiling and binning are key to interpreting metagenome data, but a lack of consensus about benchmarking complicates performance assessment. The Critical Assessment of Metagenome Interpretation (CAMI) challenge has engaged the global developer community to benchmark their programs on highly complex and realistic data sets, generated from ~700 newly sequenced microorganisms and ~600 novel viruses and plasmids and representing common experimental setups. Assembly and genome binning programs performed well for species represented by individual genomes but were substantially affected by the presence of related strains. Taxonomic profiling and binning programs were proficient at high taxonomic ranks, with a notable performance decrease below family level. Parameter settings markedly affected performance, underscoring their importance for program reproducibility. The CAMI results highlight current challenges but also provide a roadmap for software selection to answer specific research questions.

The biological interpretation of metagenomes relies on sophisticated computational analyses such as read assembly, binning and taxonomic profiling. Tremendous progress has been achieved[1], but there is still much room for improvement. The evaluation of computational methods has been limited largely to publications presenting novel or improved tools. These results are extremely difficult to compare owing to varying evaluation strategies, benchmark data sets and performance criteria. Furthermore, the state of the art in this active field is a moving target, and the assessment of new algorithms by individual researchers consumes substantial time and computational resources and may introduce unintended biases.

We tackle these challenges with a community-driven initiative for the Critical Assessment of Metagenome Interpretation (CAMI). CAMI aims to evaluate methods for metagenome analysis comprehensively and objectively by establishing standards through community involvement in the design of benchmark data sets, evaluation procedures, choice of performance metrics and questions to focus on. To generate a comprehensive overview, we organized a benchmarking challenge on data sets of unprecedented complexity and degree of realism. Although benchmarking has been performed before[2,3], this is the first community-driven effort that we know of. The CAMI portal is also open to submissions, and the benchmarks generated here can be used to assess and develop future work.

We assessed the performance of metagenome assembly, binning and taxonomic profiling programs when encountering major challenges commonly observed in metagenomics. For instance, microbiome research benefits from the recovery of genomes for individual strains from metagenomes[4–7], and many ecosystems have a high degree of strain heterogeneity[8,9]. To date, it is not clear how much assembly, binning and profiling software are influenced by the evolutionary relatedness of organisms, community complexity, presence of poorly categorized taxonomic groups (such as viruses) or varying software parameters.

A full list of affiliations appears at the end of the paper.

RECEIVED 29 DECEMBER 2016; ACCEPTED 25 AUGUST 2017; PUBLISHED ONLINE 2 OCTOBER 2017; DOI:10.1038/NMETH.4458