# Introducing Microbiome Bioinformatics

Part 9.

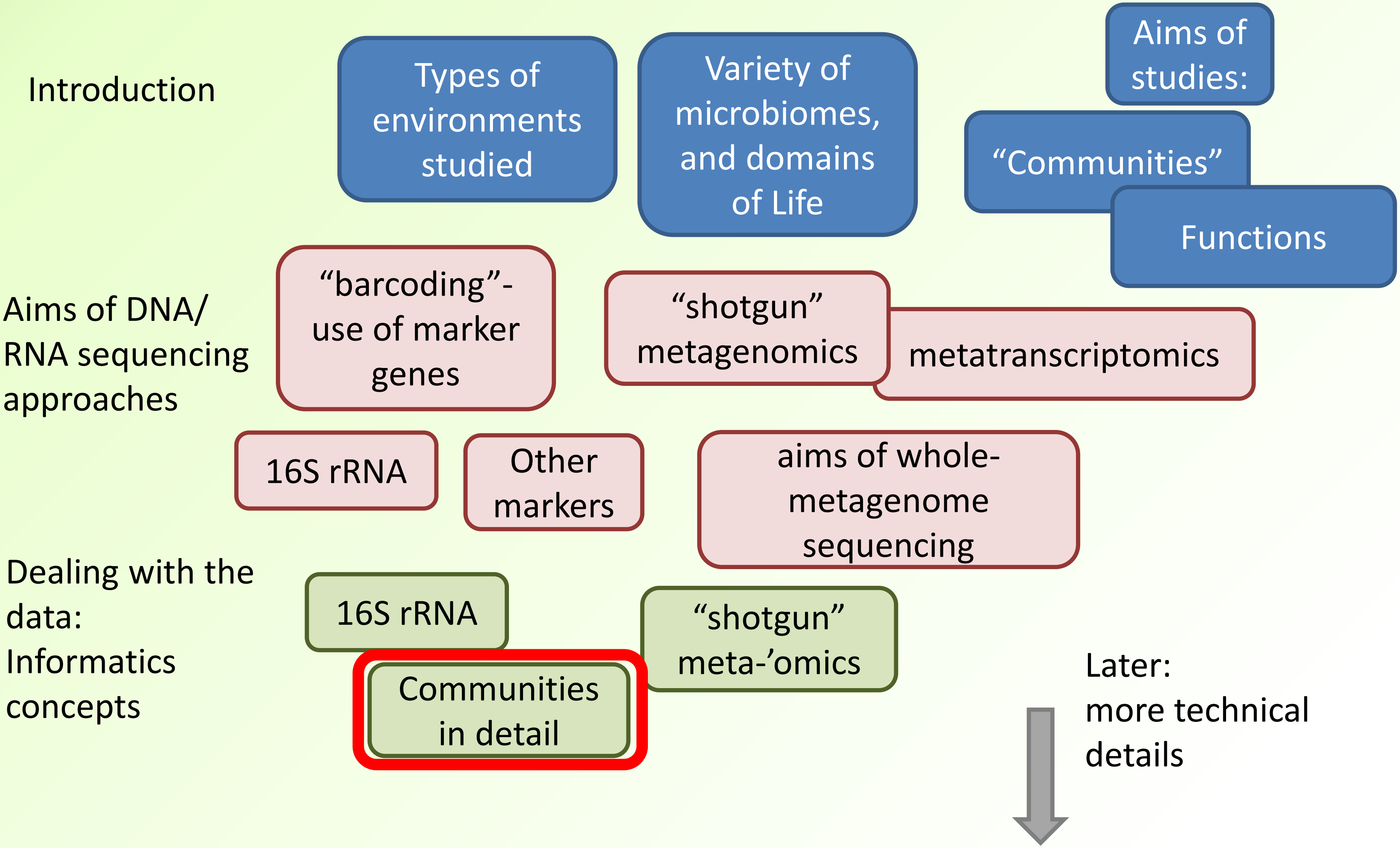
*Microbial ecology –*

*Diversity (part 2)*

# Recap: Aims

- **Microbiome analysis**
  - with particular regard to **sequence informatics concepts**
- “Top down” – putting analysis tools and resources in context
- No highly detailed technicalities (yet)
  - No instructions on how to run particular programs
- Why you are using the bioinformatics approaches you use; pros, cons; alternatives

# Topics, top-down



Introduction

Types of environments studied

Variety of microbiomes, and domains of Life

Aims of studies:

“Communities”

Functions

Aims of DNA/ RNA sequencing approaches

“barcoding”- use of marker genes

“shotgun” metagenomics

metatranscriptomics

16S rRNA

Other markers

aims of whole-metagenome sequencing

Dealing with the data: Informatics concepts

16S rRNA

Communities in detail

“shotgun” meta-’omics

Later: more technical details

# Series of talks

- 8 so far
- Open ended… as long there is demand
- Expected to be every 2 weeks
  - Notwithstanding some larger gaps for various reasons…
  - all dates will be confirmed in advance
  - *Please refer to:* ***Bite-size bioinformatics mailing list***
- Informal and flexible
  - Please interrupt and ask questions
  - **Suggestions for topics for further focus**

# Series of talks

- Part 1: 27/1/2017
  - “Biological and Experimental Stuff that a microbiome bioinformatician needs to know”
  - Overview of marker gene sequencing for community analysis
- Part 2: 10/2/2017
  - Overview of whole-metagenome sequencing
- Part 3: 24/2/2017
  - Focus on metatranscriptomics
- Part 4: 10/3/2017
  - Different bioinformatics approaches to processing 16S read data
- Part 5: 24/3/2017
  - *De novo* OTU clustering: sequence identities and how thresholds have been determined historically; relationships to taxonomic levels
- Part 6: 7/4/2017
  - The clustering problem: different approaches, and what can go wrong; the influence of amplification artefacts, sequencing errors and sequence lengths; computational OTUs versus species
- Part 7: 21/4/2017
  - Introducing microbial ecology: using observed abundances of OTUs (or species, or functions) to estimate the richness of the community (number of different OTUs, species etc)
- Part 8: 2/6/2017 – continuing microbial ecology: community diversity : diversity indices
- Part 9: today – continuing microbial ecology: community diversity : true diversity
- Slideshows - http://ghfs1.ifr.ac.uk/ghfs/

# Future talk(s)

- 30th June Barton
- TBC?
  - 14th July
  - 28th July
- None planned for August
- Topics?
- Format?

# Today

- Today and recent sessions:
- Measurements/estimations of richness and diversity of a microbiome
- (21st April) : Richness : number of species (or OTUs or functions etc)
- (2nd June) : Diversity indices
- (Today) :
  - True diversity
  - α-diversity, β-diversity, γ-diversity (…being optimistic?)
  - ~~Phylogenetic Diversity~~

# Recap

Measurement versus estimation

Richness

Diversity indices

# “Amounts of different things”

- “Things”: different –
  - Species
  - OTUs
  - Some other taxonomic unit
  - Phenotypes
  - Molecular functions
  - Pathways

*phylotypes*

*types of organism*

*types of gene*

- Whichever we are interested in, we will benefit from a **simple metric**, instead of a large table
- Enables easy and direct comparison between samples
  - Disease/health states
  - Genotypes
  - Different time points for the same subject

# You have a table like this:

***SAMPLES ….*** **…..**

***OTUs***

*or* ***species***

*…. or other ‘phylo-types’*

*…. or gene functions*

| | #1 | #2 | #3 | #4 | #5 | #6 | #7 | #8 |
|---|---|---|---|---|---|---|---|---|
| *a* | | | | | | | | |
| *b* | | | | | | | | |
| *c* | | | | | | | | |
| *d* | | | | | | | | |
| *e* | | | | | | | | |
| *f* | | | | | | | | |
| *g* | | | | | | | | |
| *h* | | | | | | | | |
| *i* | | | | | | | | |
| *j* | | | | | | | | |
| *k* | | | | | | | | |

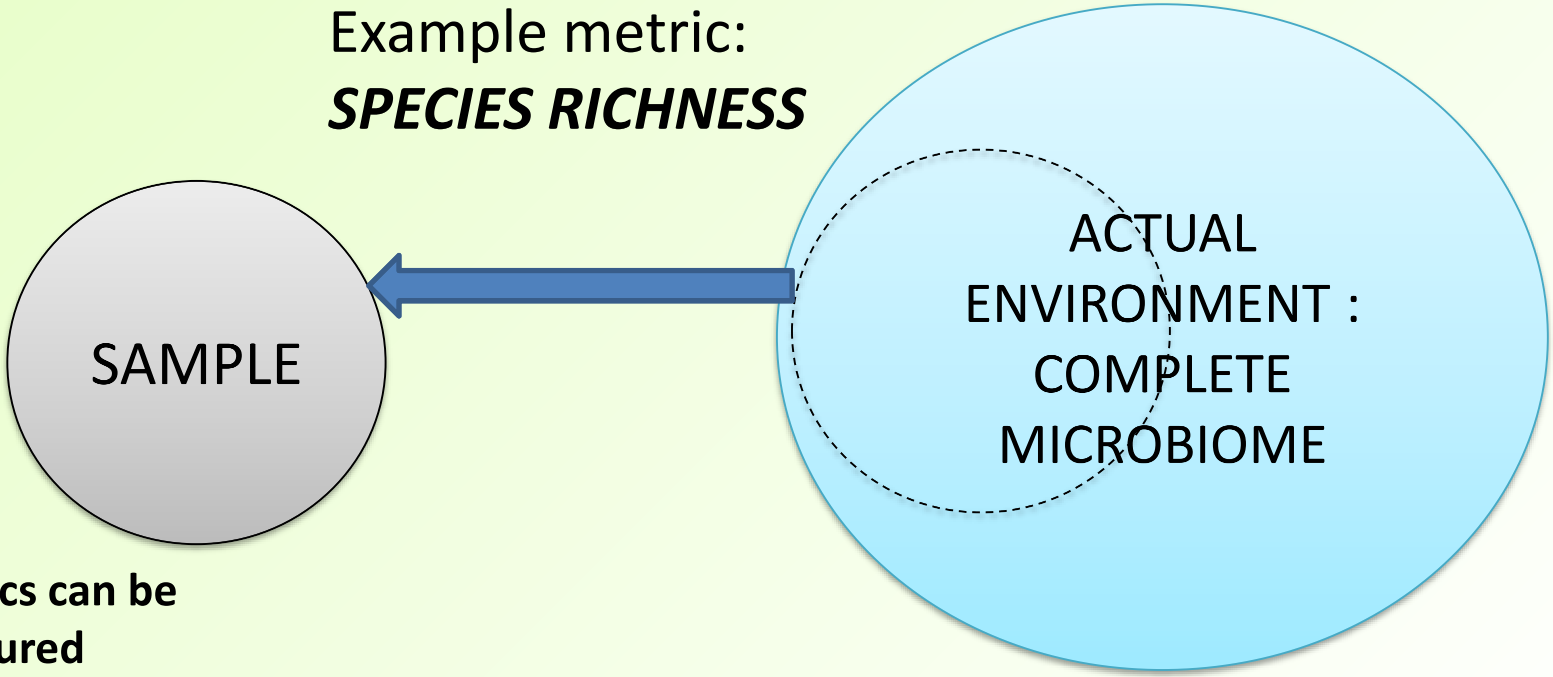
***(relative) frequencies….***

This could result from 16S rRNA gene sequence (16S rDNA) analysis, **or** metagenomics sequence analysis;

and from OTU-based approaches, and non-OTU based

# Measurement and estimation

Example metric:
***SPECIES RICHNESS***

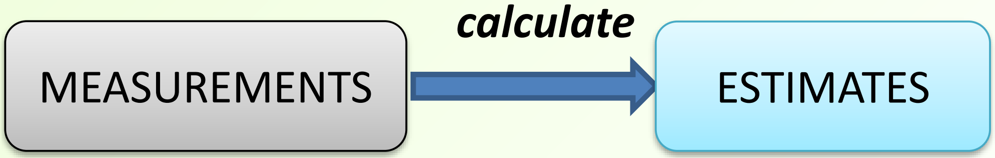


SAMPLE

ACTUAL ENVIRONMENT : COMPLETE MICROBIOME

**Metrics can be measured**
E.g. Richness: count the number of different species present

**Can't be 'measured'** – unless we are able to observe literally every organism present

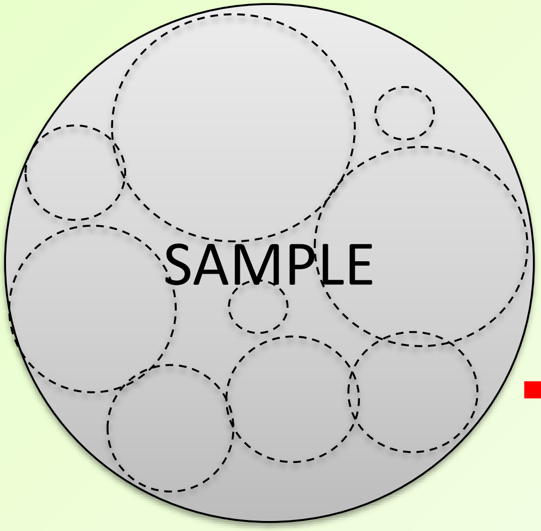


MEASUREMENTS

***calculate***

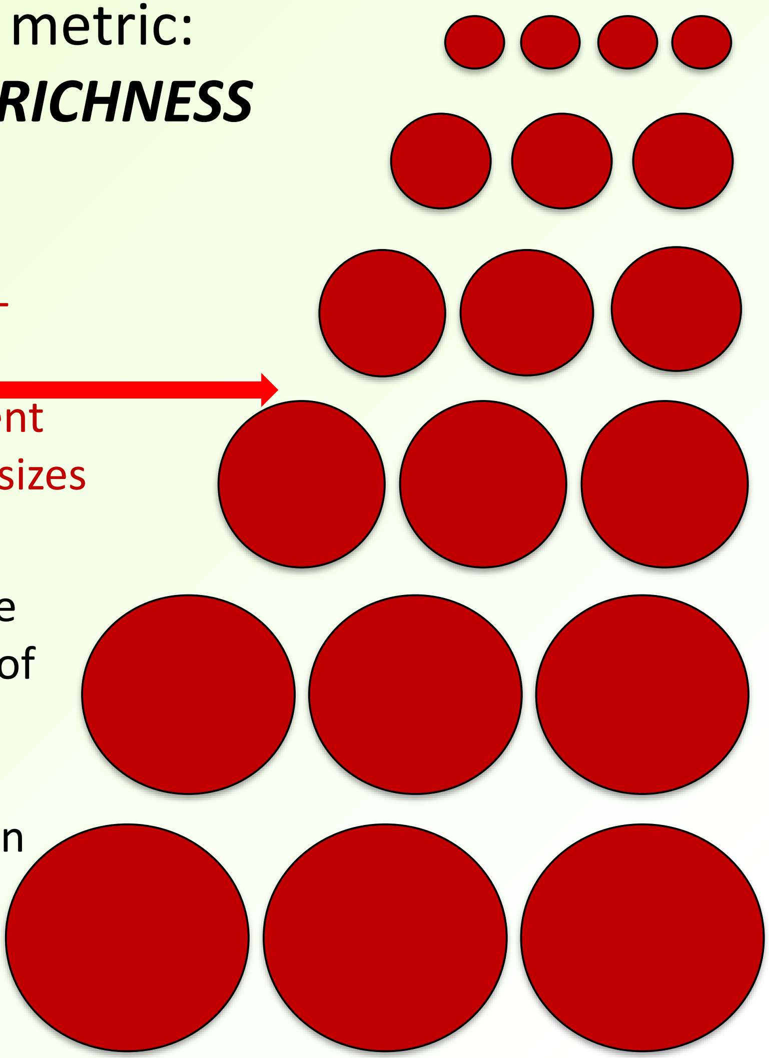
ESTIMATES

# Rarefaction: an aid to estimation

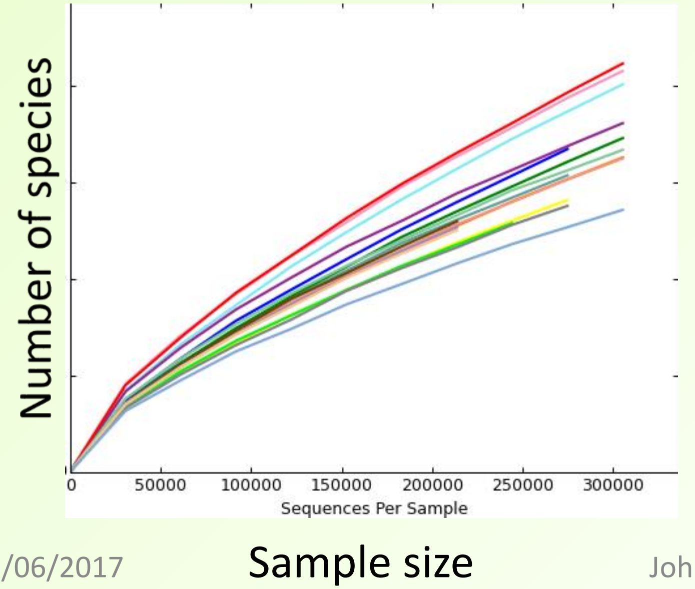


Example metric:
***SPECIES RICHNESS***

Systematic repeated random sub-sampling using different sub-sample sizes

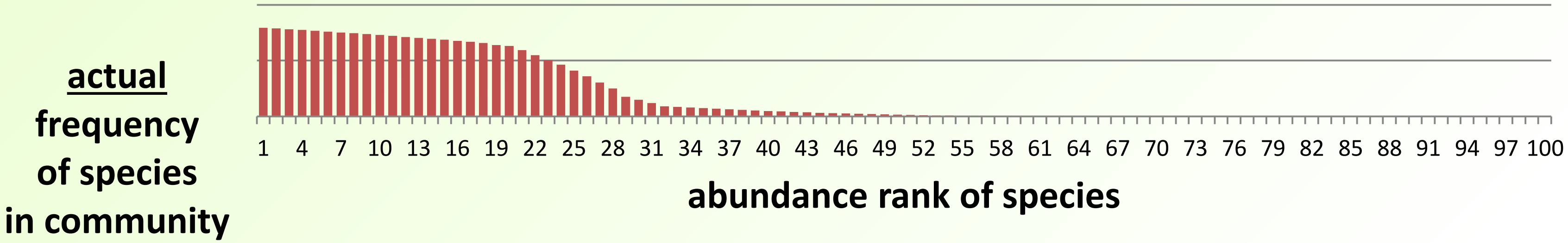


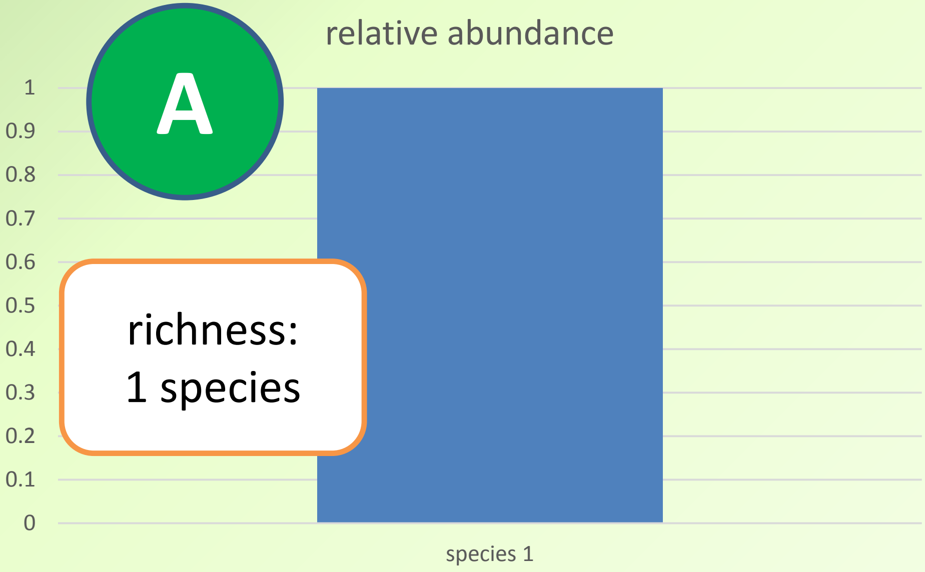
count the number of different species present in each sub-sample



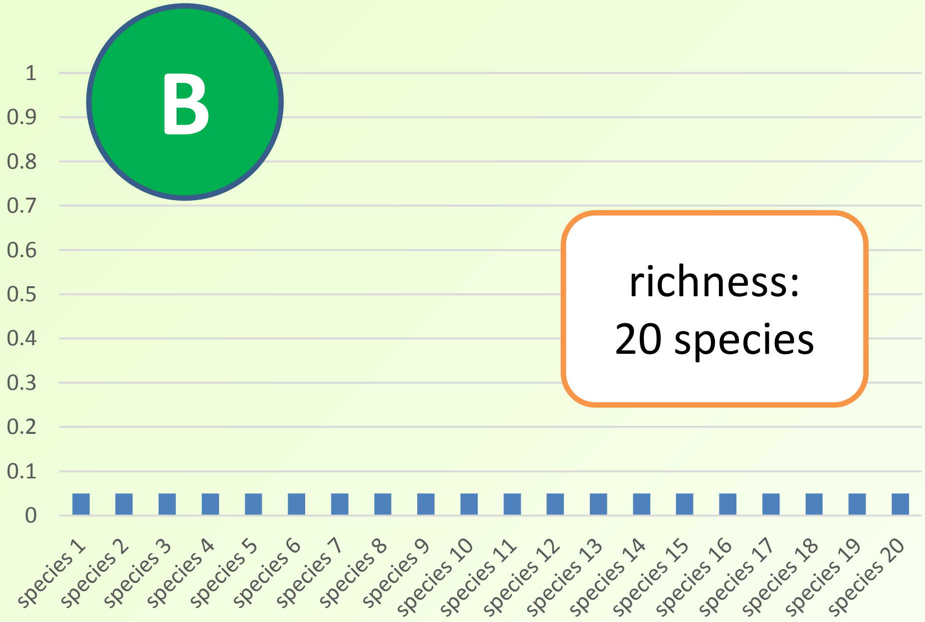
# Problems with Richness

- (21st April) Richness is difficult to estimate
  - Very thin tails (in distribution of abundance versus species)
  - Length of tail makes a big difference
  - Richness is very difficult to estimate reliably in the microbiome - e.g. Haegeman *et al.* (2013)
  - Rarefaction does not help with this
- How interested are we in the extremely low-abundant species?

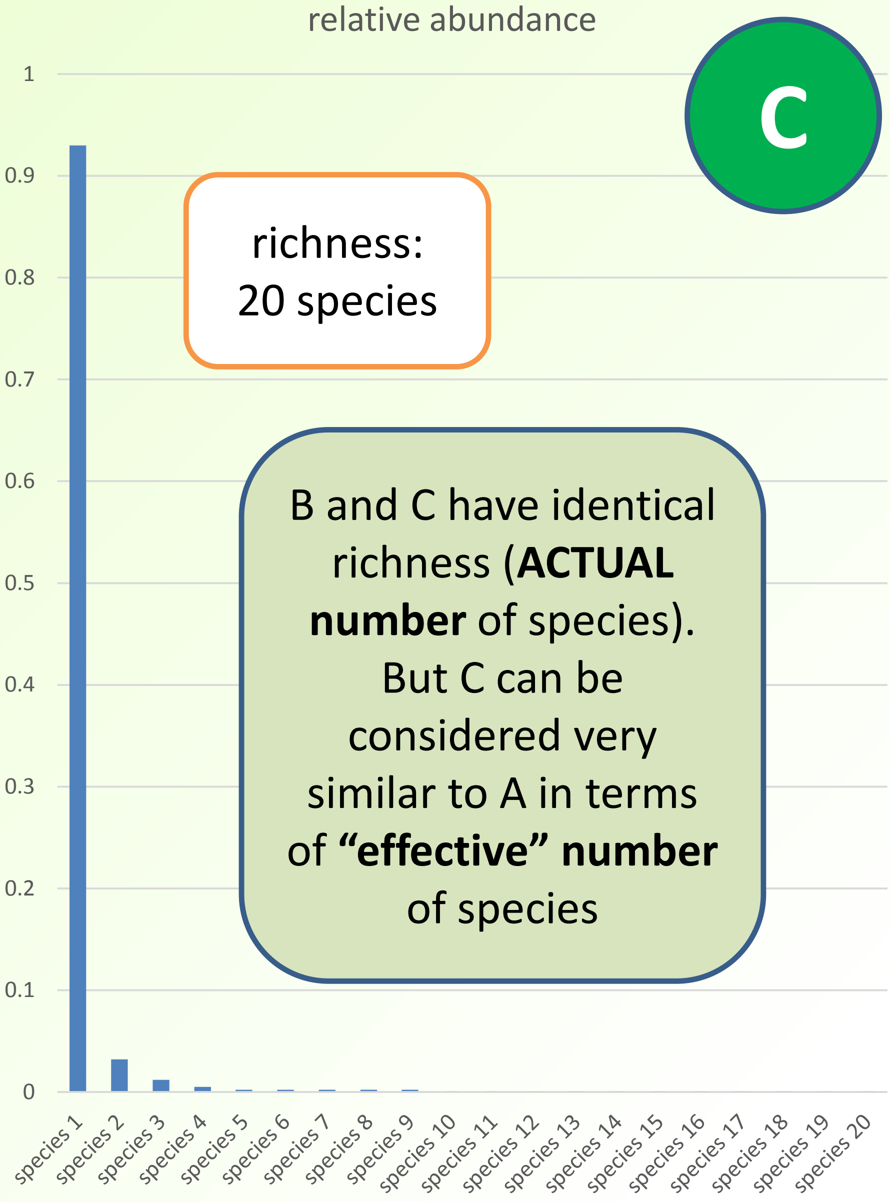




richness:
1 species

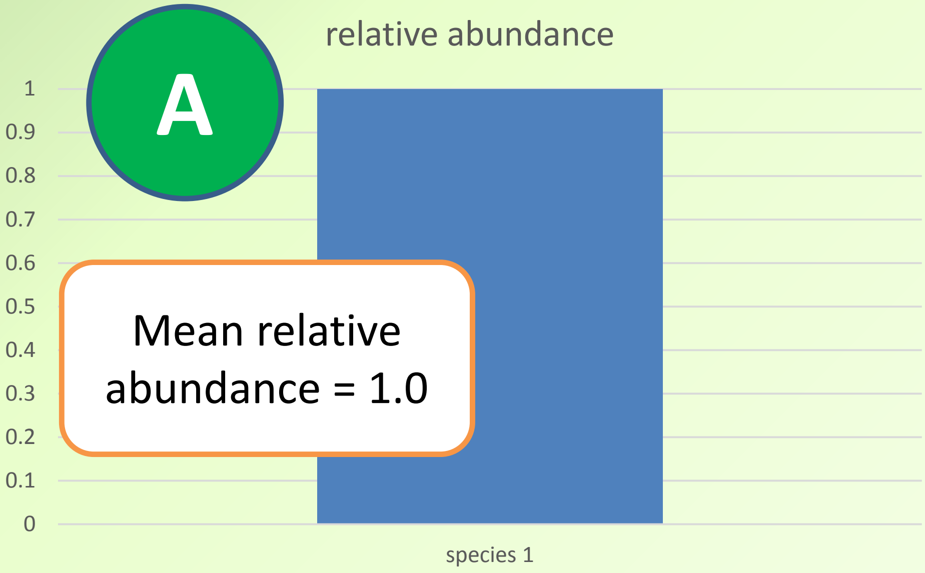


richness:
20 species

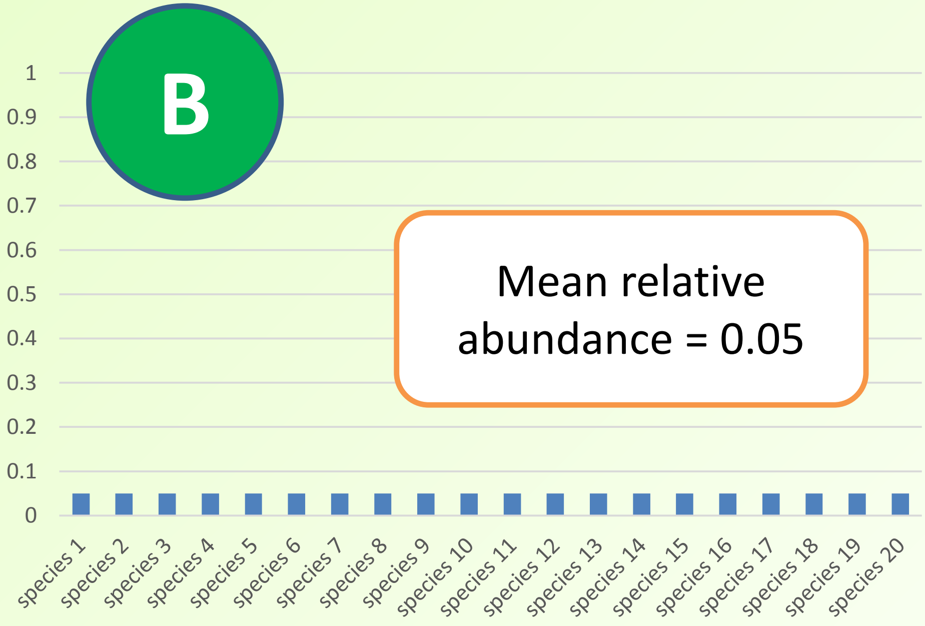


richness:
20 species

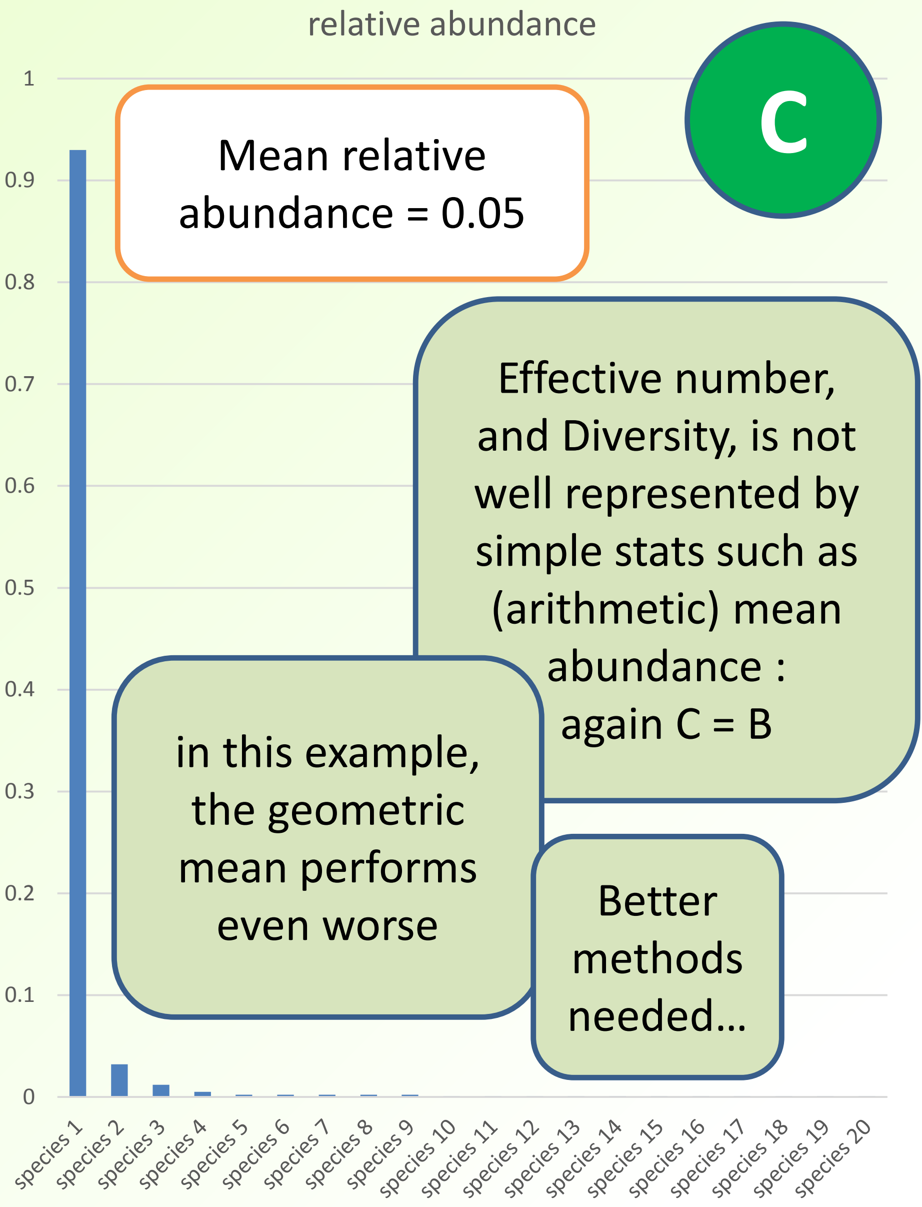
B and C have identical richness (**ACTUAL number** of species). But C can be considered very similar to A in terms of **“effective” number** of species



Mean relative abundance = 1.0



Mean relative abundance = 0.05

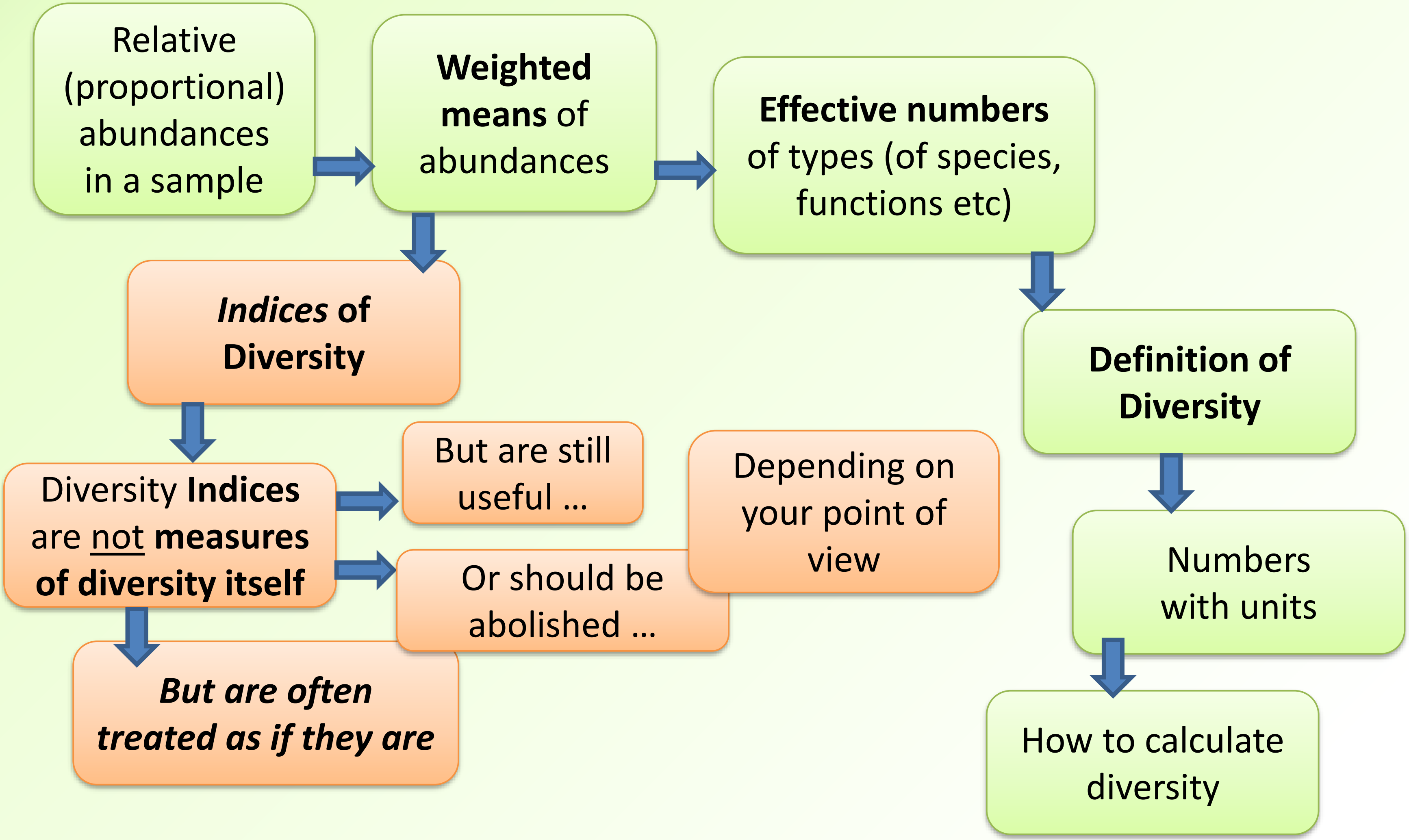


Mean relative abundance = 0.05

Effective number, and Diversity, is not well represented by simple stats such as (arithmetic) mean abundance : again C = B

in this example, the geometric mean performs even worse

Better methods needed…



Relative (proportional) abundances in a sample

**Weighted means** of abundances

**Effective numbers** of types (of species, functions etc)

***Indices of* Diversity**

**Definition of Diversity**

Diversity **Indices** are not **measures of diversity itself**

But are still useful …

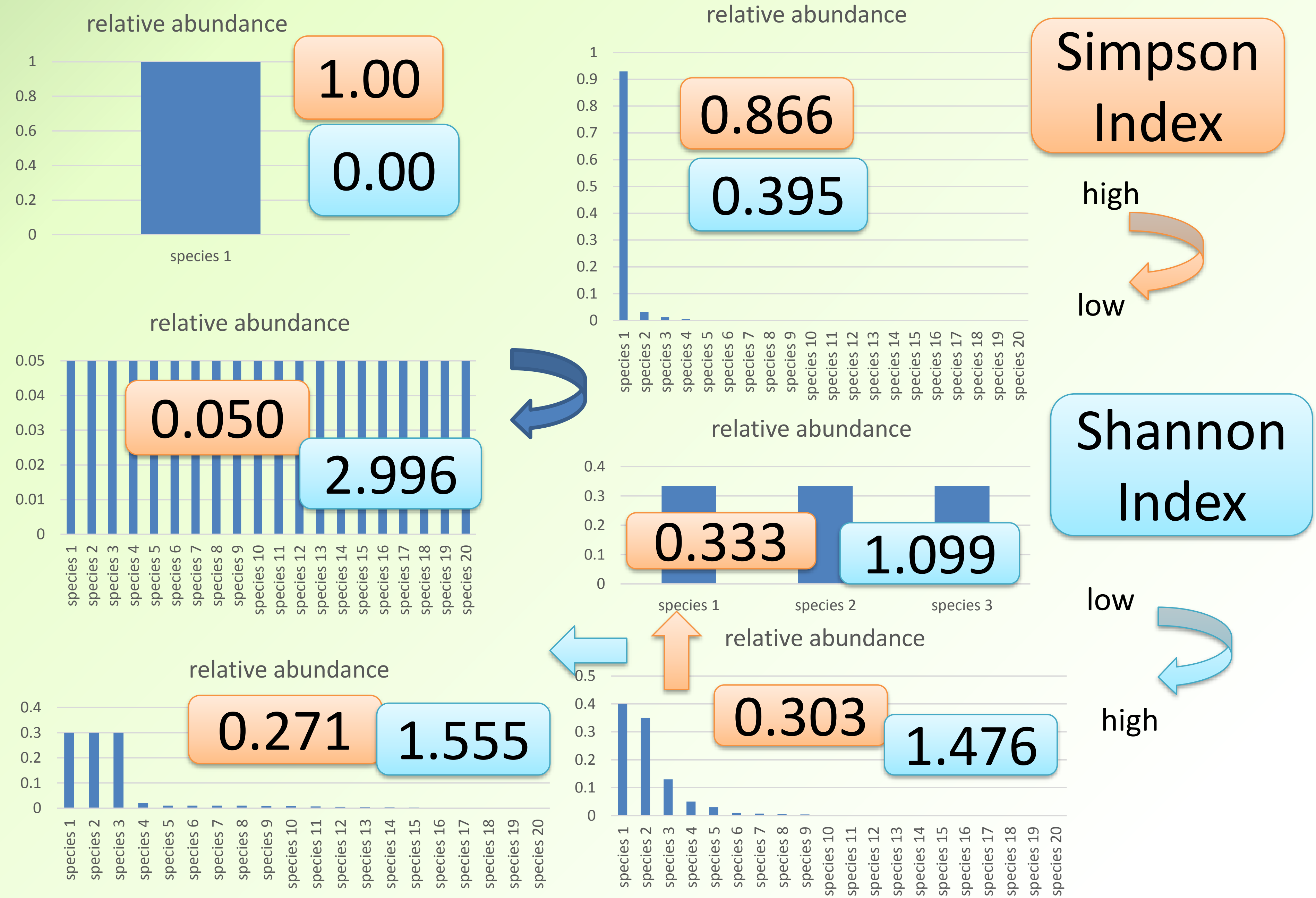
Or should be abolished …

Depending on your point of view

Numbers with units

***But are often treated as if they are***

How to calculate diversity



Simpson Index

high

low

Shannon Index

low

high

# Dealing with abundances

- We don't use absolute abundances
  - – i.e. counts of each species (or OTU, etc)
- We always deal with the **proportional abundance**
  - – often referred to here as "relative abundance"
  - – this is also effectively a **probability**
- *N* species: $x_1 , x_2 , x_3 , x_4 , x_5 \ldots , x_N$
  - – these will all be $0 < x_i \leq 1$
- $\sum x_i = x_1 + x_2 + x_3 + x_4 + x_5 \ldots + x_N = 1$

# Sums, Means and Weights

- Sum (unweighted sum):
  - $\sum = x_1 + x_2 + x_3 + x_4 + x_{5\,....} + x_N \;\; = 1$
- Unweighted arithmetic mean = $\boldsymbol{\sum / N}$ = 1 / N
- This arithmetic mean is simply a special case of a **weighted** mean
  - where each $x_i$ has a weighting $\boldsymbol{w_i}$
  - in this special case, all the weights ($w_i$)are the **same**
  - and all $w_i$ = **1/N**
  - Mean =
  - $\boldsymbol{w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + w_5 x_{5\,....} + w_N x_N}$
  - **= 1/N**

# Non-uniform weighting

- $\boldsymbol{w_1\, x_1 + w_2\, x_2 + w_3\, x_3 + w_4\, x_4 + w_5\, x_5 \ldots. + w_N\, x_N}$
- <u>Simpson index: </u>use $\boldsymbol{w_i = x_i}$
  - (weight each abundance by itself)
- $= x_1^2 + x_2^2 + x_3^2 + x_4^2 + x_5^2 \ldots. + x_N^2 = \sum x_i^2$
- <u>Shannon index: </u>use $\boldsymbol{w_i = -\ln(x_i)}$
- $= -x_1\ln(x_1) - x_2\ln(x_2) - x_3\ln(x_3) - \ldots. -x_N \ln(x_N)$
- $= -\sum x_i \ln(x_i)$ i.e. the Shannon entropy
- (often denoted ***H*** or ***H'***)

# An actual measurement of something
# is not the same as
# an ‘index of something’

units are different

# Example: **‘the size of a room’**

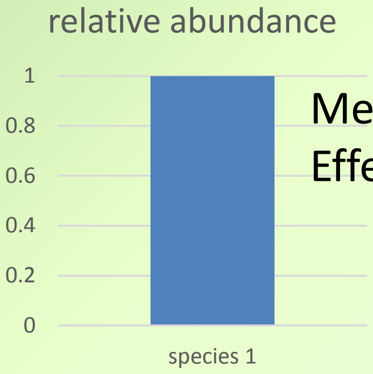
- Room sizes:

| | |
|---|---|
| Ormesby A | 9 |
| Ormesby B | 9 |
| Ranworth | 16 |
| Barton | 24 |
| Rollesby | 24 |
| Lecture Theatre | 126 |

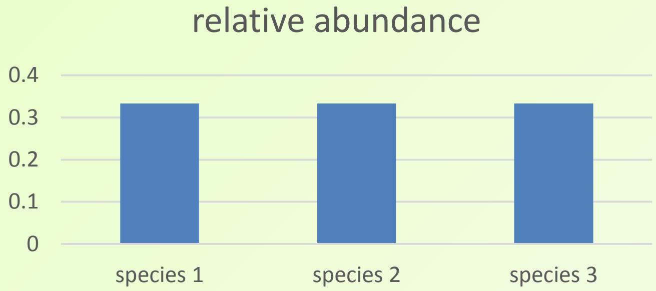
# Effective number of species

(or effective numbers of OTUs …
or of genera… function… etc)

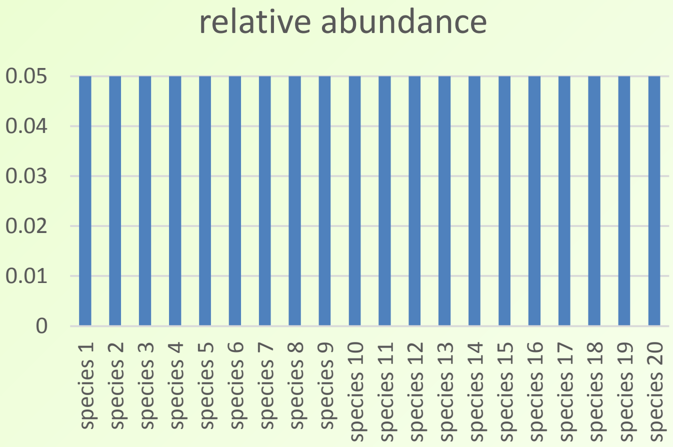
# More on effective number of species – the easy cases



Mean relative abundance = 1
Effective number = 1 / mean = **1**



*N* = 3 species only
All equally abundant
Therefore treated identically
Mean relative abundance = 0.333
Effective number = 1 / mean = **3**



*N* = 20 species: Mean relative abundance = 0.05
Effective number = 1 / mean = **20**

*N* = 1000 species:
Effective number = 1 / mean = **1000**

All obvious; but for any number *N,* the effective number will always = N
if the distribution is flat
irrespective of the weighting scheme

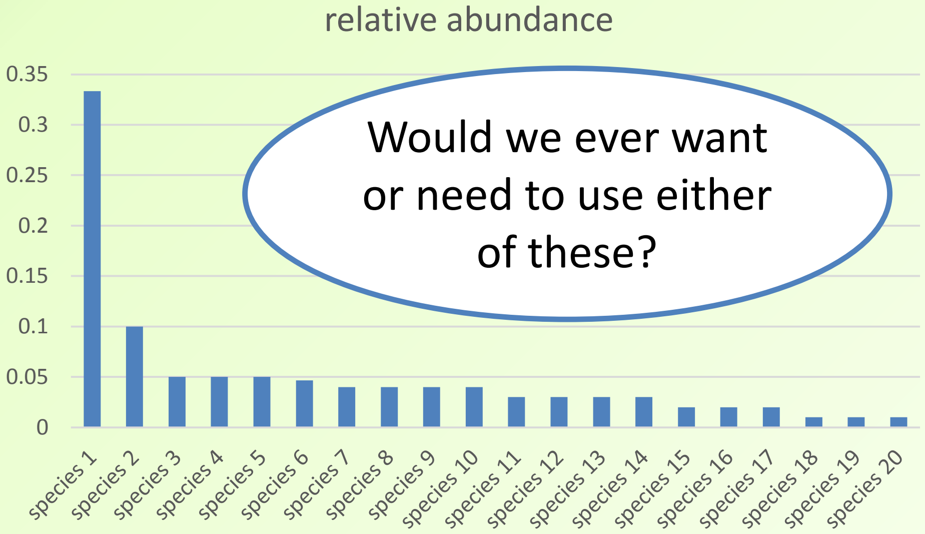
# Effective number of species

- Effective number of species is **always:**
- **the reciprocal of the mean of the relative abundances**
    - ***however that mean may be weighted***
- How do we weight that mean?
    - We have a choice
    - We have seen some approaches to this
    - **We can use more than one scheme at a time**
    - But it makes most sense to use a systematic, **generalised weighting scheme**
- For completely flat distributions, the weighted mean will always equal the unweighted mean, irrespective of the weighting applied

# Effective number of species

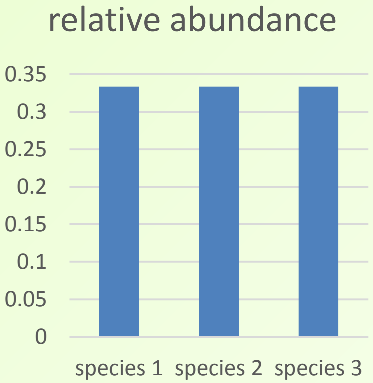
- We can choose to attach more or less importance to:
  - The more abundant species
  - The rarer species
- At the extremes, we can base the calculation entirely on the abundance of:
  - the single most abundant species
  - the single least abundant species

# Extreme weighting



Would we ever want or need to use either of these?

One extreme: we ignore all but the **most** abundant species
So, weighted mean = mean of most-abundant species = 0.33333
→ effective number of species
= 1 / 0.33333 = **3.00**



Either way
→ effective number of species
= 1 / 0.33333 = **3.00**

Other extreme: we ignore all but the **least** abundant species
So, weighted mean = mean of most-abundant species = 0.01
→ effective number of species
= 1 / 0.01 = **100.00**

# Effective number of species = Diversity

- The effective number of species
- i.e. the reciprocal of the weighted mean
    - **IS THE DEFINITION OF DIVERSITY**
- It has units: **species**
    - (or OTUs, or functions, or languages spoken by employees, or whatever it is you are assessing the diversity of)
- I.e. its units are identical to **richness**
- If we are using **more than one way** of calculating the weighted mean, in **systematic** approach:
- **Then we have a series of diversity values**

# The Hill Numbers

Mark Oliver Hill used a simple system of **weighted generalised means** of the relative abundances

Hill (1973)

(also known as “Hill Diversity” , “true diversity”)

# The Hill Numbers

- Consider a value *k*
- Sum each relative abundance $x_i$ raised to the power *k* and weighted by $w_i$
  - $\boldsymbol{w_1 x_1^k + w_2 x_2^k + w_3 x_3^k + w_4 x_4^k + w_5 x_5^k \ldots + w_N x_N^k}$
- i.e. $\sum (w_i x_i^k)$
- The weighted mean is the *k*th root of this sum
  - $\left( \sum (w_i x_i^k) \right)^{1/k}$
- So the diversity is the reciprocal of that weighted mean:
  - $1 / \left( \sum (w_i x_i^k) \right)^{1/k} \quad = \quad \left( \sum (w_i x_i^k) \right)^{1/-k}$

# The Hill Numbers

- This value $( \sum (w_i\, x_i^k) )^{1/-k}$
- is the **Hill Diversity of order *k*+1** (thus, $^{k+1}D$)
- - and is rarely (ever?) written like the above – but it makes more sense of the *k*th root (IMO)
- The Hill system in fact **always uses $w_i = x_i$**
- and is normally written
  - $^{q}D = ( \sum x_i^q )^{1/(1-q)}$
- i.e. **Hill Diversity of order *q*** (i.e., $q = k+1$)
- i.e. the weighted mean abundance is $( \sum x_i^q )^{1/(q-1)}$

# Some interesting properties of $^{q}D$

- $q$ can be a non-integer
- $q$ can be negative
- $^{q}D$ appears undefined for $q = 1$
- (so too does the reciprocal, i.e. the weighted mean abundance)

$$^{q}D = \left( \sum x_i^{q} \right)^{1/(1-q)}$$

- Weighted mean abundance
- = 1 / diversity

$= 1 / {}^{q}D = \left( \sum x_i^{q} \right)^{1/(q-1)}$

# $q = 0$

- $q = 0$
- ${}^{0}D =$
- $(x_1^{0} + x_2^{0} + x_3^{0} \dots x_N^{0})^{1}$
- $= N$
- i.e. the number of different species
- i.e. **RICHNESS**
- (and the weighted mean abundance is the simple arithmetic mean)

$${}^{q}D = ( \sum x_i^{q} )^{1/(1-q)}$$

$$1 / {}^{q}D = ( \sum x_i^{q} )^{1/(q-1)}$$

# $q = 2$

- $q = 2$
- ${}^2D =$
- $(x_1^2 + x_2^2 + x_3^2 \dots x_N^2)^{-1}$
- = the reciprocal of the Simpson index
- (and the weighted mean abundance *is* the Simpson index)

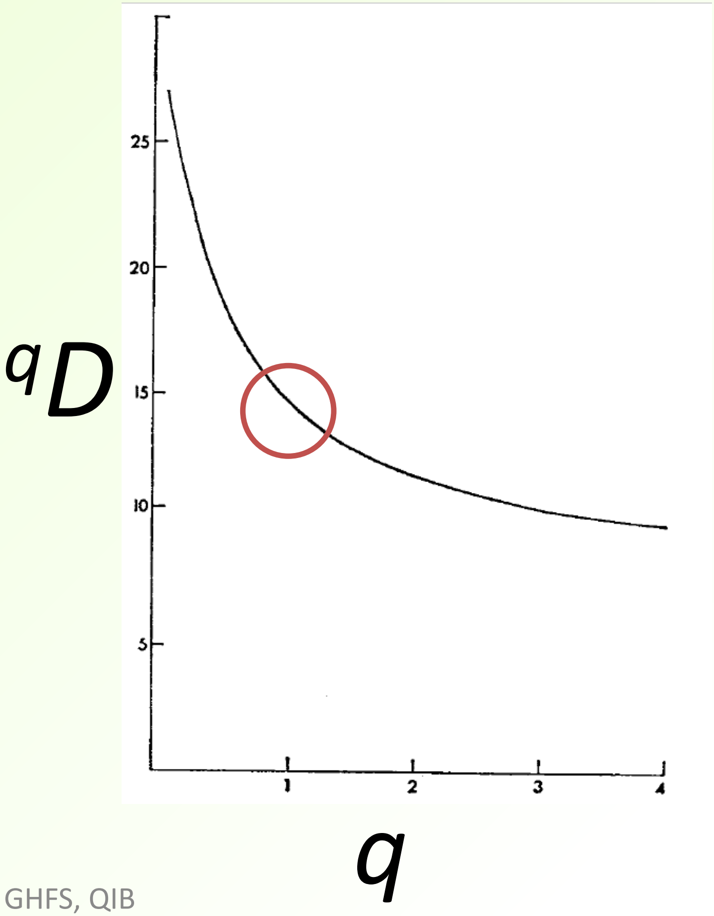
$$ {}^qD = ( \sum x_i^q )^{1/(1-q)} $$

- $1 / {}^qD = ( \sum x_i^q )^{1/(q-1)}$

# What about $q = 1$ ?

- $q = 1$
- ${}^{1}D$ appears to be
  - $( \sum x_i^q )^{\infty}$ i.e. $1^{\infty}$
- But do not assume ${}^{q}D = 1$
- By considering:
- what ${}^{q}D$ **tends to** as $q \rightarrow 1$
- It can be shown that ${}^{q}D$ is **continuous** at q = 1
- Hill included a proof of this
- Example data:

$${}^{q}D = ( \sum x_i^q )^{1/(1-q)}$$



# What about $q = 1$ ?

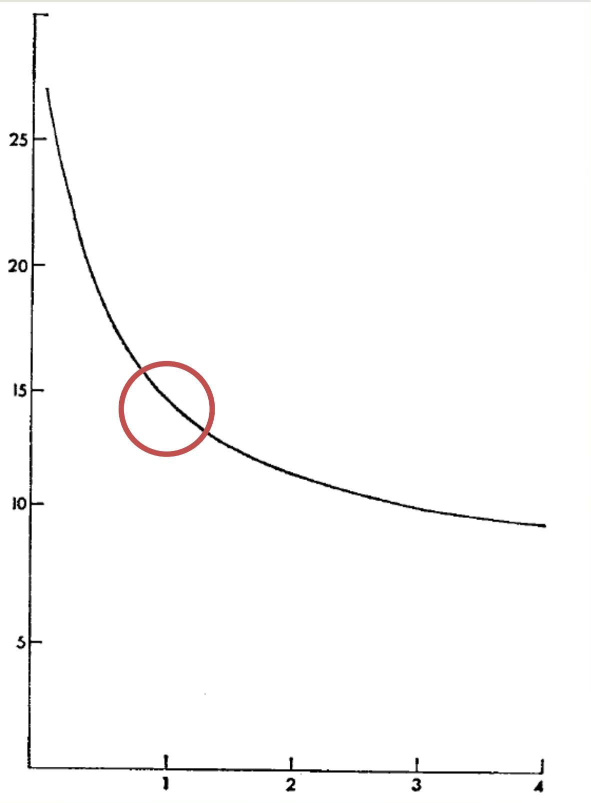
- $q = 1$
- $^{1}D = \lim_{q \to 1} ({}^{q}D)$
- It can be shown that:

$$^{1}D = e^{(-\sum x_i \ln(x_i))}$$

$$= e^{H}$$

- That is, ***e* to the power of the Shannon index**

$$^{q}D = (\sum x_i^{q})^{1/(1-q)}$$



# $q = \infty$

- $q = \infty$
- Can be shown that:
- $^{\infty}D =$
- the reciprocal of the proportional abundance of the commonest species
- and the weighted mean abundance

  = abundance of commonest species

$$^{q}D = ( \sum x_i^{q} )^{1/(1-q)}$$

$$1 / {}^{q}D = ( \sum x_i^{q} )^{1/(q-1)}$$

# $q = -\infty$

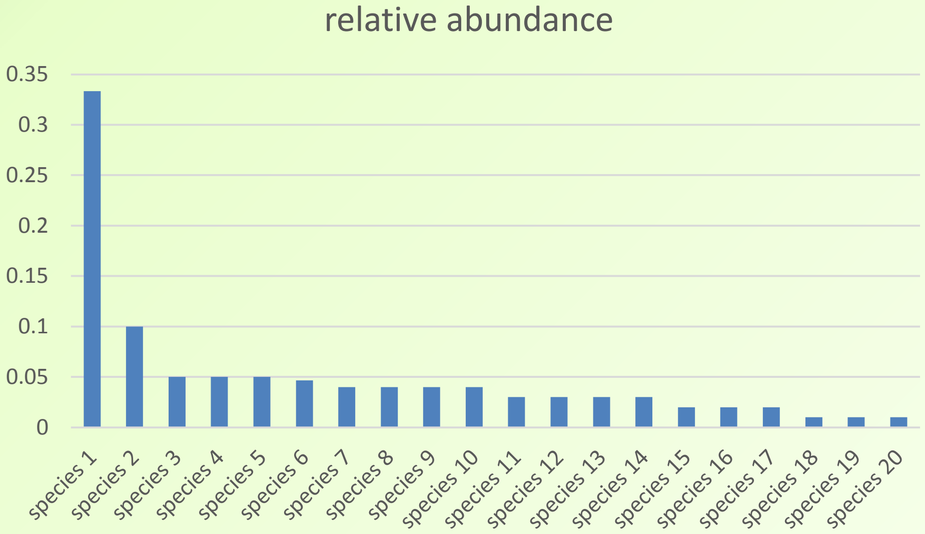
- $q = -\infty$
- Can be shown that:
- ${}^{-\infty}D =$
- the reciprocal of the proportional abundance of the rarest species
- and the weighted mean abundance

  = abundance of rarest species

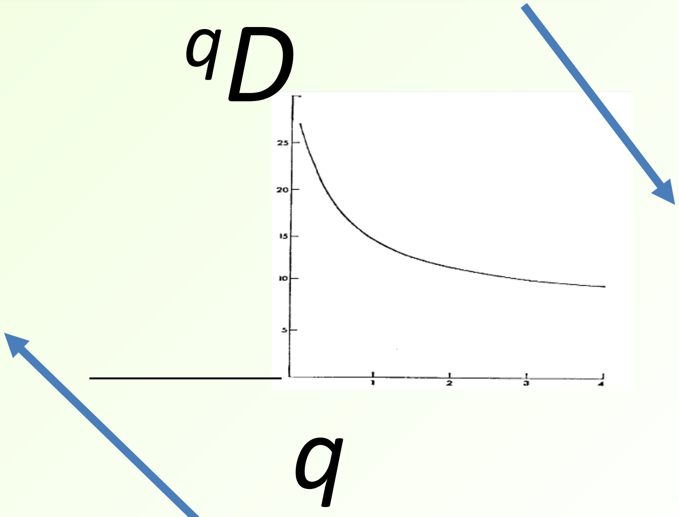
$$ {}^{q}D = (\sum x_i^{q})^{1/(1-q)} $$

$$ 1 / {}^{q}D = (\sum x_i^{q})^{1/(q-1)} $$

# $^{q}D$ : a sliding scale of weighting



One extreme: we ignore all but the **most** abundant species



Other extreme: we ignore all but the **least** abundant species

- As we vary *q* we attach more or less importance to:
  - The more abundant species
  - The rarer species
  - Some balance inbetween

# What about units?

- ${}^{q}D = ( \sum x_i^{q} )^{1/(1-q)}$ What are the units?
- Strictly speaking:
  - a proportional abundance ($x_i$) is unitless
- A mean proportional abundance has units **species$^{-1}$**
  - Because $N$ has units of: **species**
  - mean = 1 / $N$
- In the simple mean case, $\boldsymbol{w_1 x_1 + w_2 x_2 + w_3 x_3}$ **…**
  - Weightings also have units species$^{-1}$
  - Because $w_i = 1 / N$

# What about units?

- But the longer form of

$$^{q}D = ( \sum x_i^{q} )^{1/(1-q)}$$

- is:

$$^{q}D = ( \sum (w_i \, x_i^{q-1}) ) ^{1/(1-q)}$$

- and the units always work (i.e., $^{q}D$ has units: **species**) …
- …**as long as** $w_i$ is unitless, and $x_i$ has units: **species$^{-1}$**
- which is sort of the other way round to what might be expected
- So would it be better written like this? Discuss….

$$^{q}D = ( \sum (x_i \, w_i^{q-1}) ) ^{1/(1-q)}$$

# So how useful are these numbers?

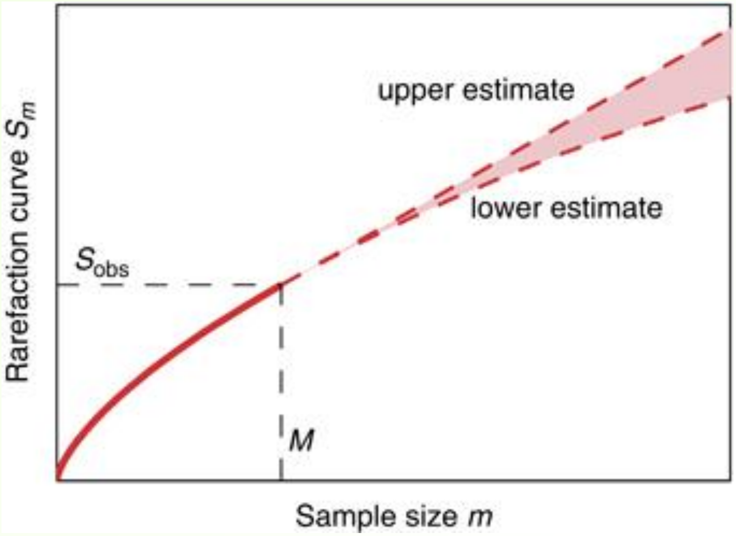
- For any sample, the series of Hill Diversities ${}^{q}D$ can be easily calculated
- Do you really need to calculate them for ‘all’

  $-\infty < q < +\infty$ ?
- Are even just a few integer values of *q* useful?
- When you **calculate** these Diversity numbers from the **measured** abundances in your sample –
- How well do they **estimate** the Diversity in the original environment?

# Using rarefaction with ${}^{q}D$

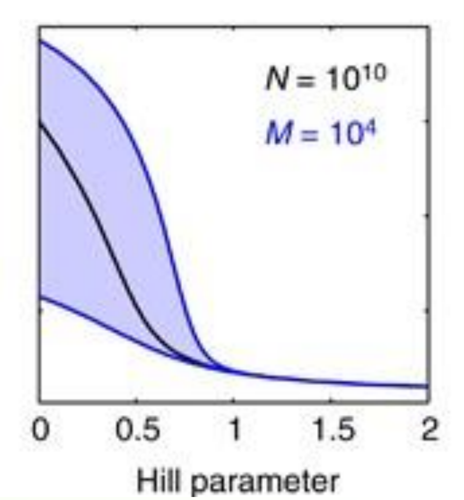
- As with any rarefaction, the curve can be extrapolated beyond the actual sample size
- The range of uncertainty in the extrapolated region can be calculated
- Haegeman *et al.* (2013)

  *Robust estimation of microbial diversity in theory and in practice*
- Contains equations
  - for lower and upper estimates

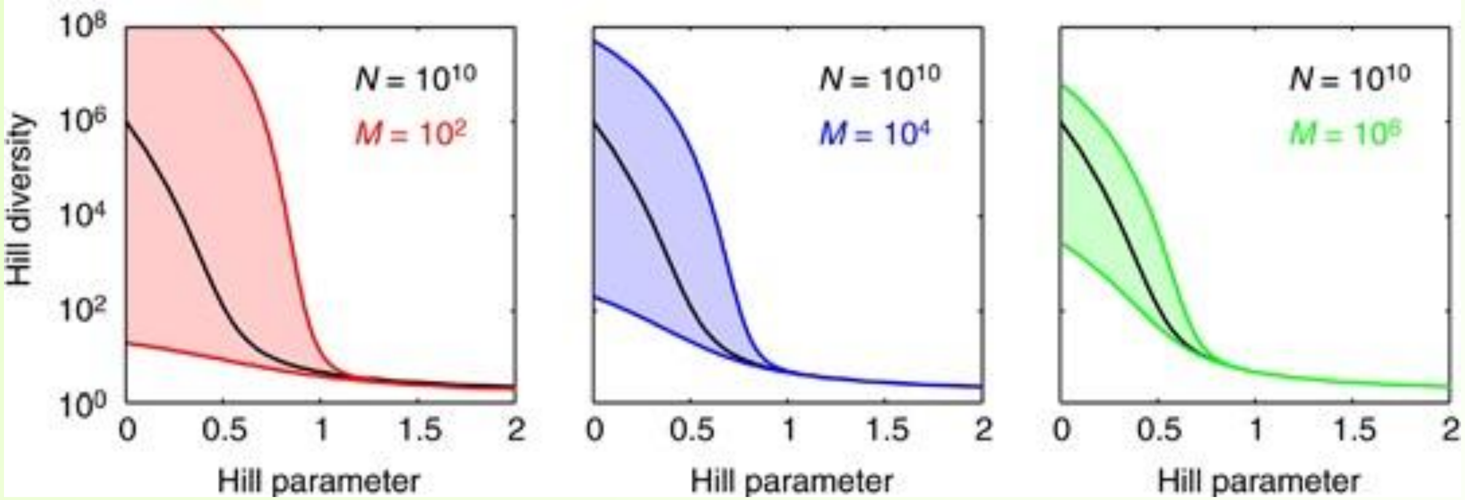


# *In silico* communities: we know the right answer

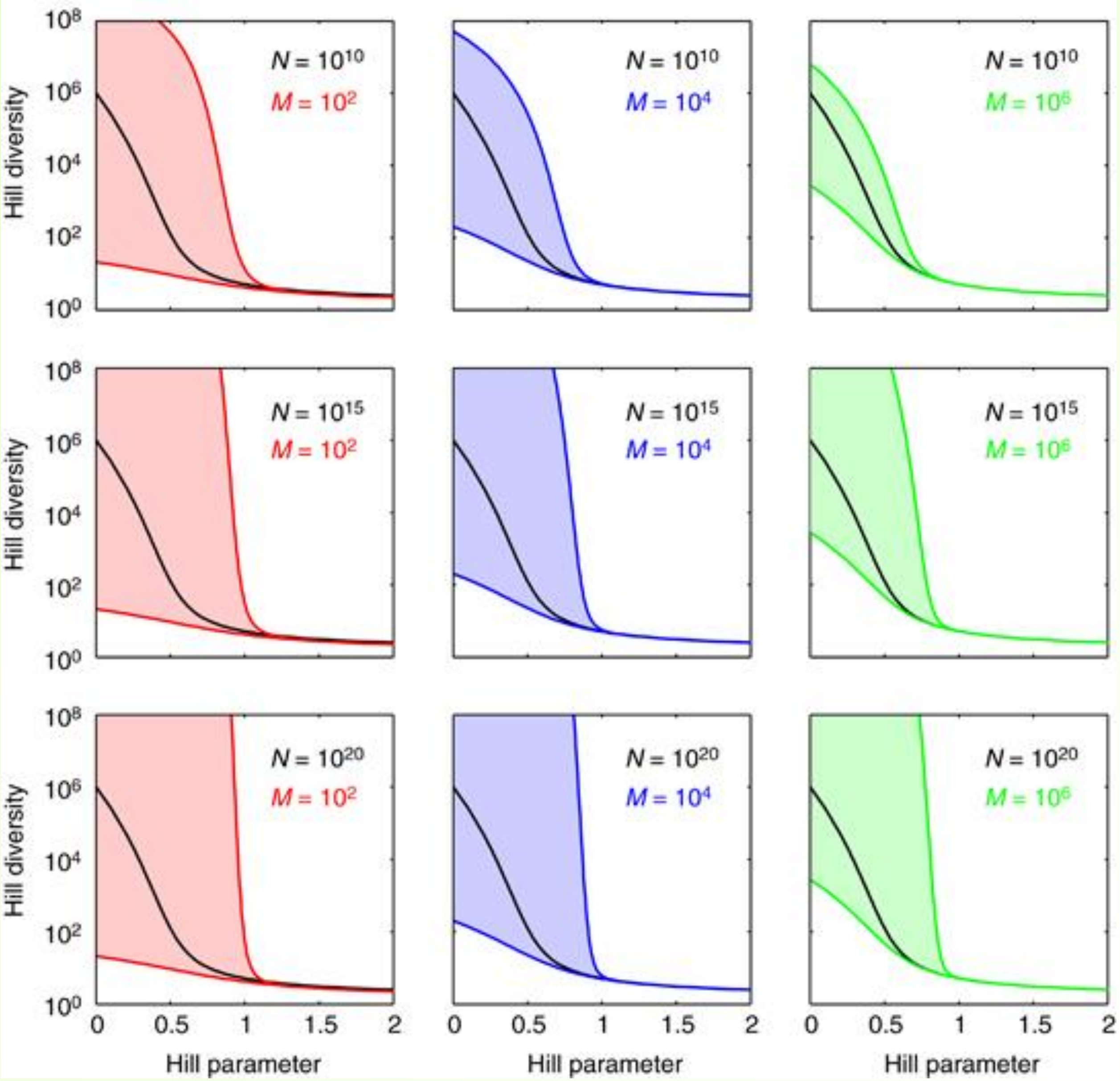
- A computer-generated community – i.e. true abundances of all organisms are known
- Here, *N* is the **total number of organisms** in the community
- Obtain a **sample of size *M***
- Perform rarefaction
- Can be done with any metric
  - – such as a ${}^{q}D$ : for a range of *q*



Haegeman *et al.* (2013)

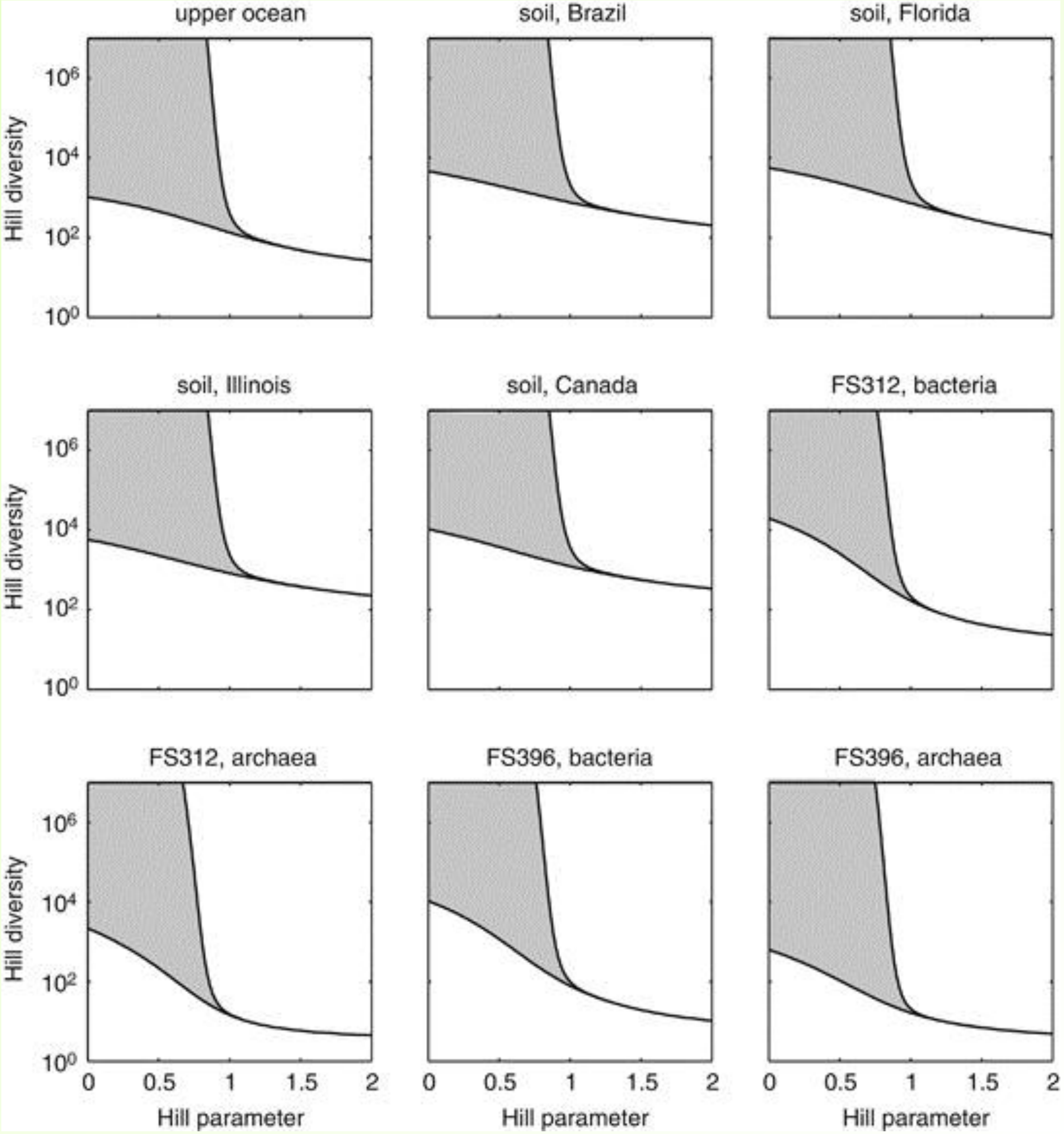


Haegeman *et al.* (2013)



Haegeman *et al.* (2013)

Using **real samples from real microbiomes** - We don’t know the right answer… but do we get estimates in a narrow uncertainty range?



Haegeman *et al.* (2013)

# α-diversity,
# β-diversity,
# γ-diversity

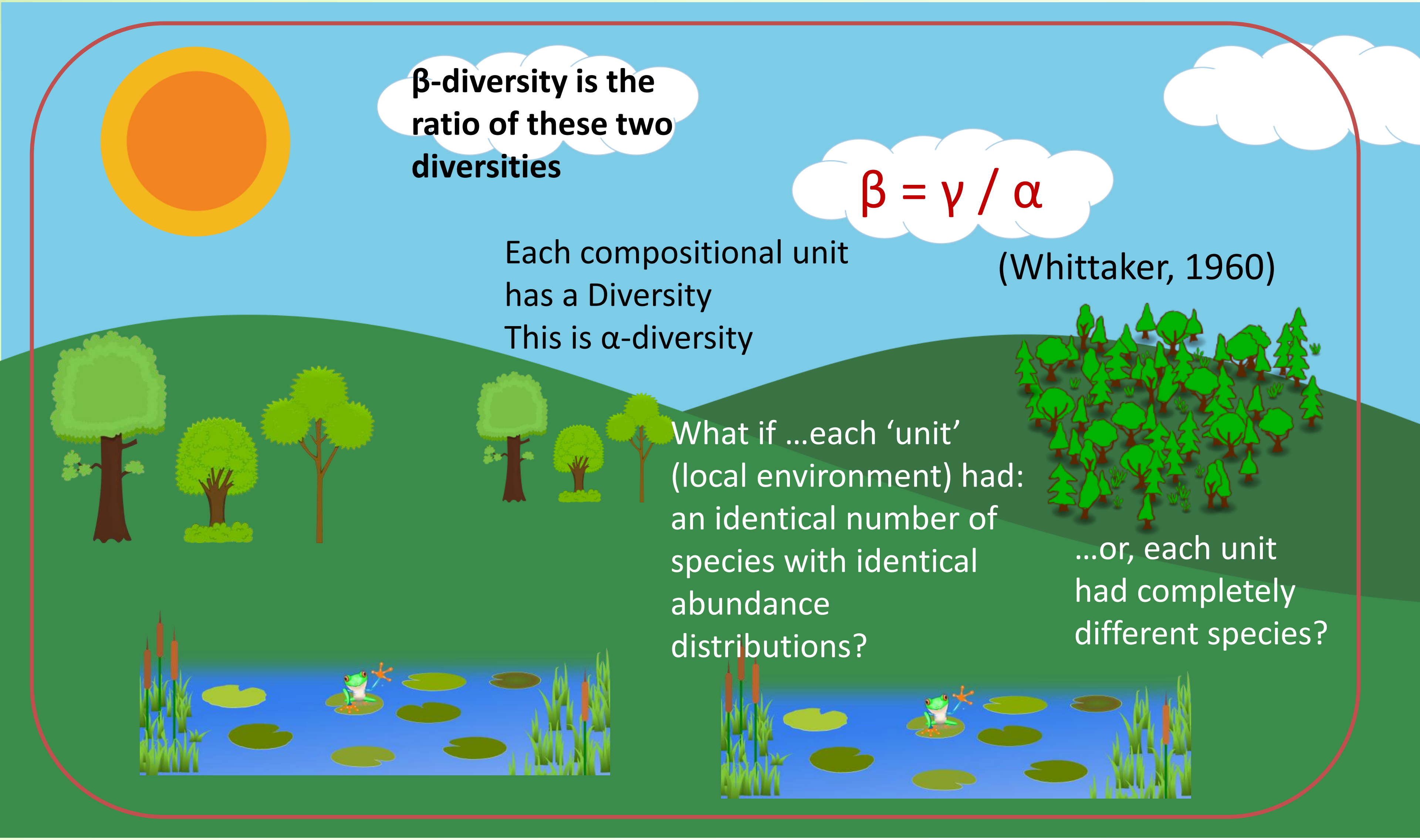
These are (or should be) related to each other in a straightforward way

# α-diversity

- α-diversity is the Diversity of a single "compositional unit"
- What you use as a measure of "Diversity" is your choice
- (but choose wisely)
- E.g. one (or more) of the Hill Diversities

"compositional unit":
represents a single "compartment"
Which could be:
a locality within a larger region
And also applies to a **sample**

# γ-diversity is the Diversity of the entire region



**β-diversity is the ratio of these two diversities**

$$\beta = \gamma / \alpha$$

(Whittaker, 1960)

Each compositional unit has a Diversity

This is α-diversity

What if …each ‘unit’ (local environment) had: an identical number of species with identical abundance distributions?

…or, each unit had completely different species?

# For a gentle introduction

- See ‘methods.blog’
  - Official blog of Methods in Ecology and Evolution
- https://methodsblog.wordpress.com/
- - see their most-accessed blog article ever
  - “What is beta diversity” Andrés Baselga
- Note that “constituent compositional unit”
  - (such as a localised ecosystem in a larger region, or a sample from it)
- ..is also equivalent to a constituent sampling unit in general
  - Such as multiple faecal samples from the same host

# Diversity of diversities

- “A consistent terminology for quantifying species diversity? Yes, it does exist” Tuomisto (2010) *Oecologia* **164** 853-860

“The term ‘diversity’ has been used in **at least four conceptually different ways** in the ecological literature,

primarily **because indices of diversity have been equated with diversity itself.**

Furthermore, an alpha component, or ‘alpha-diversity’, has been separated from total or ‘gamma diversity’ in **at least three different ways**.

The situation is even worse with ‘beta diversity’, which has been defined in **more than 30 different ways**;

Some of these are **not mathematically derived from alpha and gamma diversity in any way**, and the **values of many are uncorrelated** with each other”

- For more details of some of these, see Tuomisto H. (2010) “A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity”, *Ecography* **33:** 2-22

# Just two β-diversities for now

- Regional:local diversity ratio $\beta_{Mt}$
  - Used with one or more Hill diversities (*q* values)
  - Thus ${}^{q}\beta_{Mt}$
- “Quantifies how many times as rich in effective species an entire dataset is than its constituent sampling units are on average”
  - It is unitless
  - α-diversity involved ($\alpha_t$) is “mean species diversity within sampling units”
- True beta diversity $\beta_{Md}$
  - Put all your constituent units together –how many distinct units does it really look like, in a mathematical sense?
  - If all of your samples are replicates, you hope the answer is 1
  - Thus quantifies the number of composition units
  - $\beta_{Md}$ has units of: **species/compositional unit**

# Summary

- Use true diversities (Hill Diversities)
  - Even if just a small number of *q* values
- Straightforward to calculate
- Values of the Hill curve for $q \geq 1$ should be comparable
  - So should be similar for e.g. replicate samples
- Also they enable calculation of β-diversity
  - Which is a measure of **how many "distinct units" you <u>really have</u>** amongst your collection of samples

# References

- Haegeman B., Hamelin J., Moriarty J., Neal P., Dushoff J. and Weitz J.S. (2013) Robust estimation of microbial diversity in theory and in practice *ISME J.* **7:** 1092-1101
- Hill M.O. (1973) Diversity and evenness: A unifying notation and its consequences, *Ecology* **54:** 427-432
- Shannon C.E. (1948a) A Mathematical Theory of Communication *Bell System Technical Journal* **27** (3): 379-423
- Shannon C.E. (1948b) A Mathematical Theory of Communication *Bell System Technical Journal* **27** (4): 623-656
- Simpson E.H. (1949) Measurement of Diversity *Nature* **163:** 688
- Tuomisto H. (2010) A diversity of beta diversities: straightening up a concept gone awry. Part 1. Defining beta diversity as a function of alpha and gamma diversity, *Ecography* **33:** 2-22
- Tuomisto H. (2010) A consistent terminology for quantifying species diversity? Yes, it does exist, *Oecologia* **164:** 853-860
- Whittaker R.H. (1960) Vegetation of the Siskiyou Mountains, Oregon and California, *Ecol. Monogr.* **30** (3) 280-338