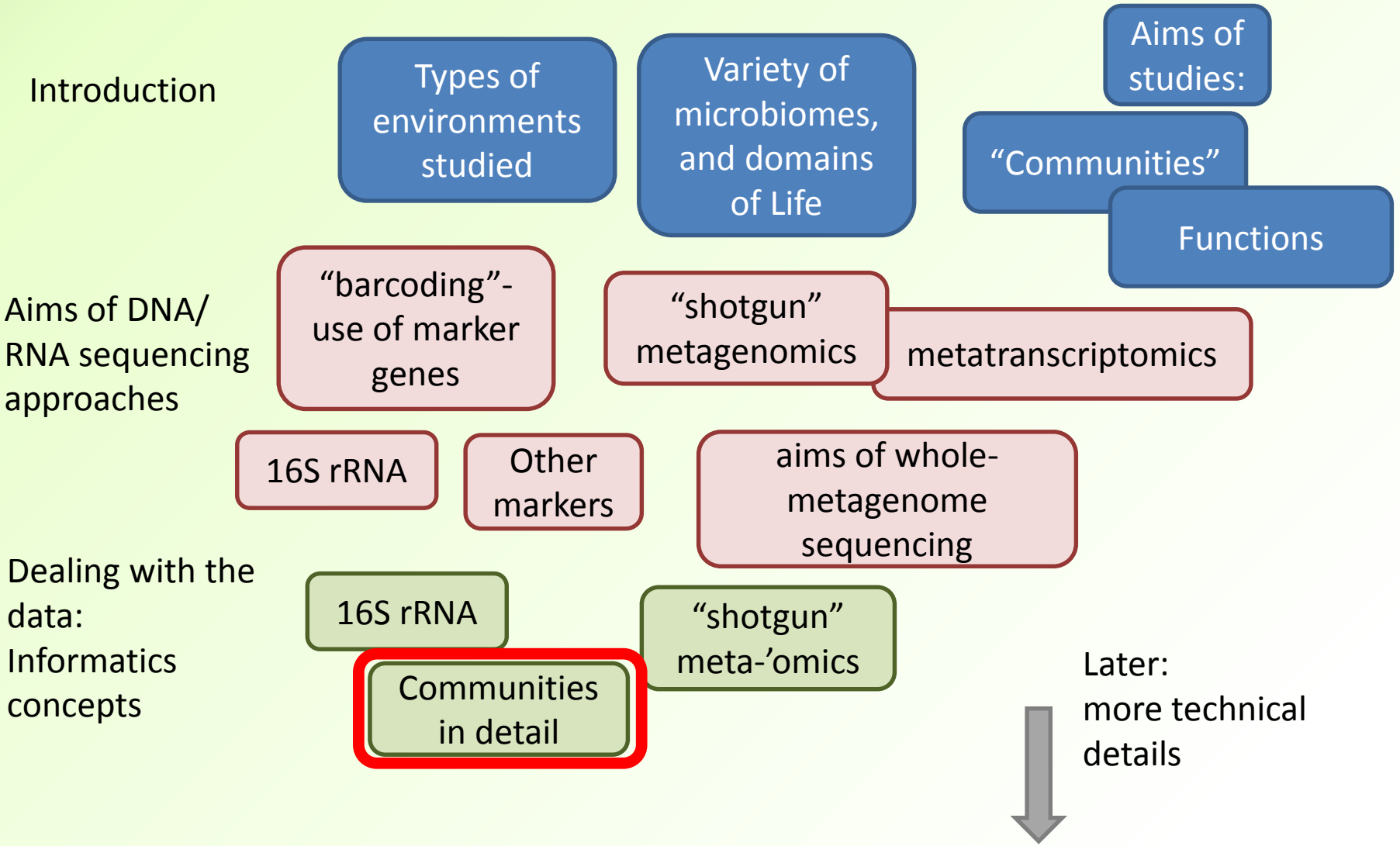
Introducing Microbiome Bioinformatics

Part 7.

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



Series of talks

- 6 so far
- Open ended… as long there is demand
- Expected to be every 2 weeks, but all dates will be confirmed in advance
  - *Bite-size bioinformatics mailing list*
- The next few will cover: (*not necessarily in this order…)*
  - 16S analysis for community profiling
  - Clustering and classification issues (taxonomies etc)
  - Analysing richness and diversity of those communities
  - Dealing with sequencing and other errors
- 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
- Slideshows
  - http://ghfs1.ifr.ac.uk/ghfs/

To be confirmed…

- **NO SESSION ON 5th MAY**
- **NO SESSION ON 19th MAY**

- 2nd June Barton
- 16th June Barton

Let’s take a break from Operational Taxonomic Unit assignment…

… what can you actually do with your OTU assignments?

(or any taxonomic assignments)

You have a table like this:

SAMPLES …. ***…..***

OTUs

or ***species***

…. or other 'phylo-types'

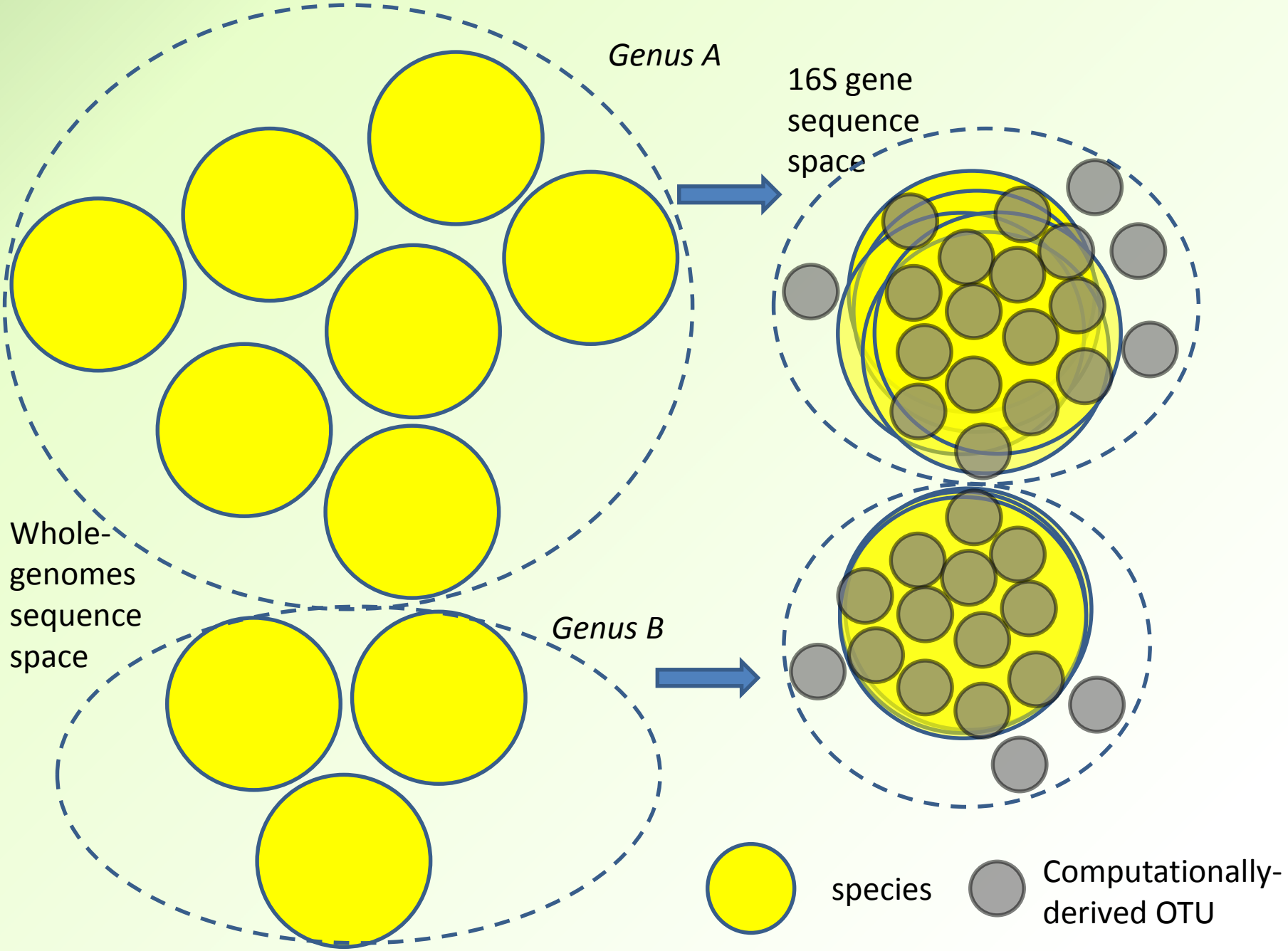
	#1	#2	#3	#4	#5	#6	#7	#8
a								
b								
c								
d								
e								
f								
g								
h								
i								
j								
k								

(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

A brief recap of genera, species and OTUs in 16S rRNA gene sequencing


Genus A
16S gene sequence space
Whole-genomes sequence space
Genus B
species
Computationally-derived OTU

What taxa to use in your frequency tables?

- With *16S rDNA*, using **species** as your 'taxonomic atom' is not really possible
- And **OTUs** may not be the best idea either
  - (may depend partly on how you arrived at those OTUs; another topic for later)
  - And you may not even have used an OTU-based approach in the first place
- So use the labels you have got
  - Which will almost certainly extend to different levels
- With shotgun *metagenomics*, **species**-level identification should be possible
  - with some but definitely not all reads
  - (another topic for later…)

Microbial ecology

- Generally, the same principles and metrics apply as in other ecological studies
- Estimates of
  - richness (numbers of different organisms)
  - diversity (frequency distributions of organisms)
  - ***Many*** **different ways** of calculating these
    - (strictly, *estimating* them)
  - **Within** and **between** communities/ecosystems

Ecology and taxa

- Usually, these methods used in ecology are applied to **species**
- But in principle, can be applied to other taxonomic levels
  - (such as genera, from 16S ; or OTUs)
- That is, the formulae can be applied to any category or 'class', in principle
- For simplicity, here we will refer to **'phylotypes'** as the category in question (usually….)
  - As that can refer to different levels of relatedness, as the case may be

Other uses of Metrics

- E.g. Richness and Diversity
- These are most commonly applied to richness and diversity of **phylotypic or taxonomic** groups
- They can also be applied to richness and diversity of other things
  - Such as phenotypes or molecular functions
- Diversity metrics of **functions** inferred from metagenome/ metatranscriptome sequencing are increasingly common in the literature

Sampling and estimation

- These methods for analysing communities / ecosystems necessarily use **sampling** in nearly all cases
- Studies where every single **individual** organism can be observed with certainty, are extremely unusual
  - And certainly do not include microbiome studies
- Many traditional ecological approaches involve **capture-release-recapture** sampling
  - Each individual might be observed once only
  - or more than once
  - or not at all

Sampling and estimation

- Always remember the distinction between:

a) The numbers **observed** in the **sample**

b) The **true numbers** in the **original community**

- (a) is used to **calculate** an **estimate** of (b)
- Some methods are based on the capture-release-recapture assumption, when performing these calculations
- Is this a sound assumption for sampling prokaryote cells by sequencing a piece of their DNA?
  - With shotgun metagenomics?
  - With amplified segments of 16S rRNA genes?
  - Discuss…?

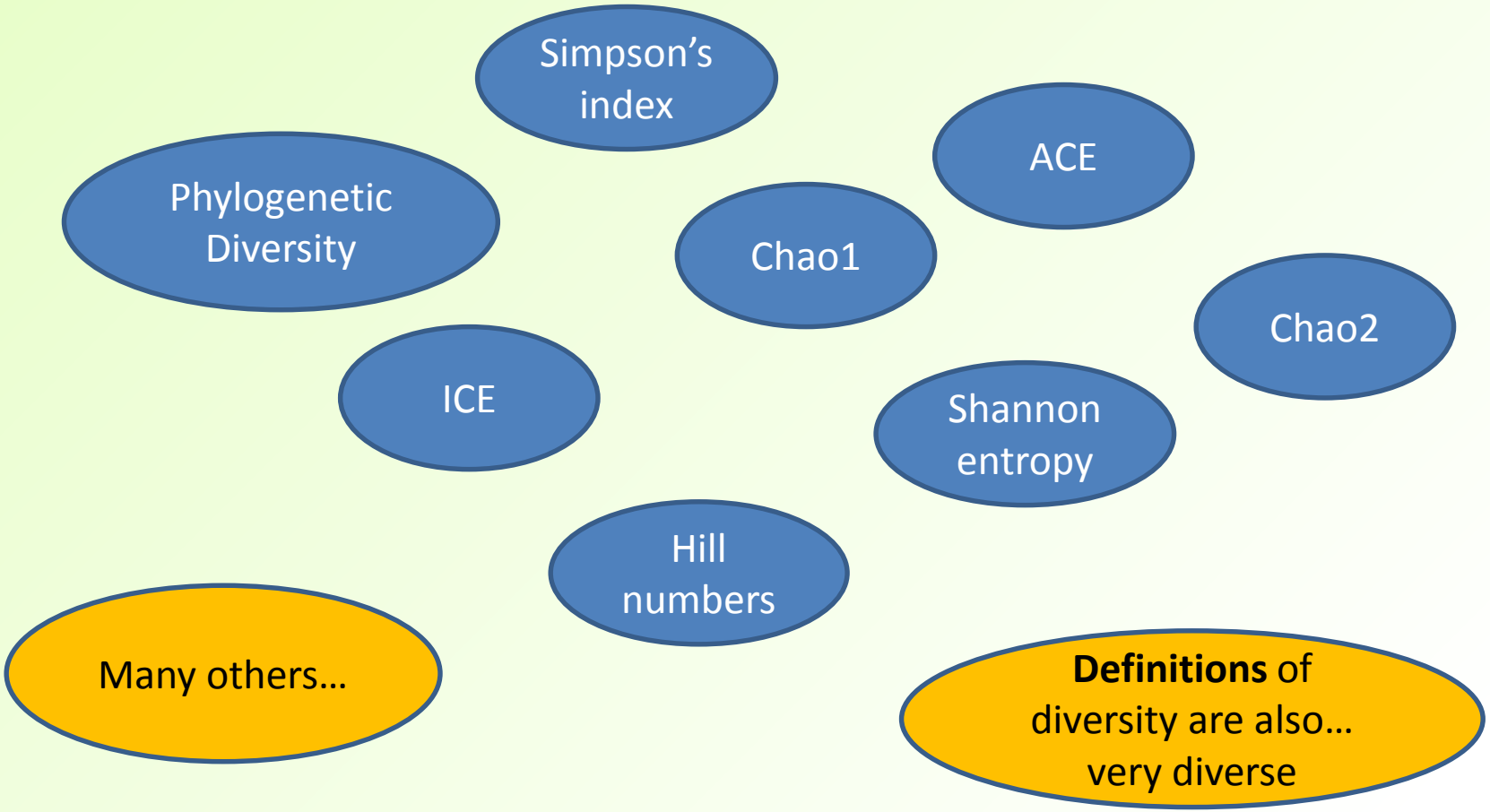
Frequencies (measure of abundance)

- Again, there are **actual frequencies** which we can try to estimate by **observed frequencies**
- **Observed frequencies:**
- Can be a **count** of number of times each species is observed
- Usually dealt with as a **proportion**
- In some ecological studies, **non-discrete** observations are more appropriate
  - E.g. dry mass

Richness and Diversity of organisms in ecosystems

(micro-organisms or otherwise)

Indices used for richness or diversity



Metrics of Richness and Diversity

- Strictly speaking, these are **estimates**, *not* measurements
- A useful way of describing a sample with a small amount of information (such as a single number)
- Enables assessment of differences between samples, and thus estimations of:
  - Differences between communities/ecosystems
  - Changes in a community/ecosystem over time
- Can be correlated with other aspects of the sample/ ecosystem e.g.
  - Levels of pollutants
  - Host phenotype
  - Host health/disease state

Richness

- Total number of organisms (species, OTUs or other phylotypes) present
- This can be very hard to estimate by sampling
- because in the general case, **we do not know what shape the distribution of frequencies** is
  - This can be tackled with non-parametric approaches
- It seems especially problematic if there are many species with a very low frequency
- Many of these could be missed in any given sample
- Also, some of the lowest-frequency organisms could be artefactual (undetected chimaeras; sequencing errors)

The simplest estimate of all

- How many different phylotypes have you observed in the sample?
- In general, likely to be a poor estimator for the actual number of phylotypes
- It is possible to evaluate* *whether* this number approximates a stabilised value
  - I.e. the maximum value you would ever get, with increasingly larger sample sizes
  - Which is hopefully a good indicator of the actual number
- **Estimated Rarefaction**: Repeatedly analyse subsets of your data, of all sub-sample sizes up to the actual size of the sample data set
  - (This is **individual-based** rarefaction)

Some example data

From a GHFS project

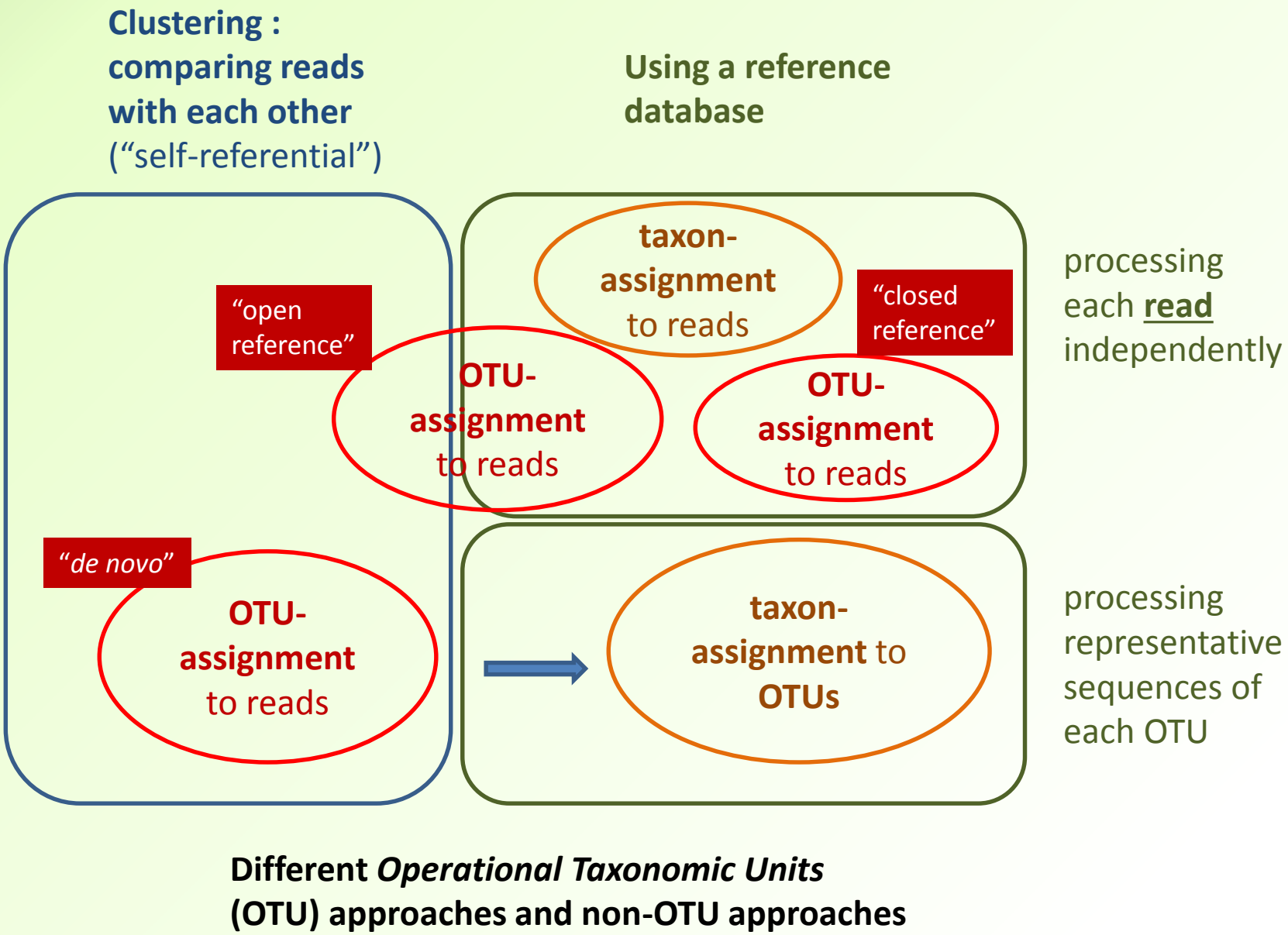
250 bp PE Illumina sequencing of 16S V4/5

Multiple samples, belonging to > 1 cohort

Reads from all samples will be considered collectively, for this illustration

Some example data

- An indication of numbers that might be expected if you use as phylotypes:
  - OTUs
  - Named taxa from reference taxonomies (assigned to those OTUs)
- Also the difference between two types of OTU-assignment
  - *De novo* clustering
  - Closed-reference assignment (use a reference OTU database)
- And between data which has been pre-screened for chimera sequences, and those which have not
- In all cases, the data has been pre-screened to discard low-quality sequences in the same way
- You might get very different numbers from your data of course!

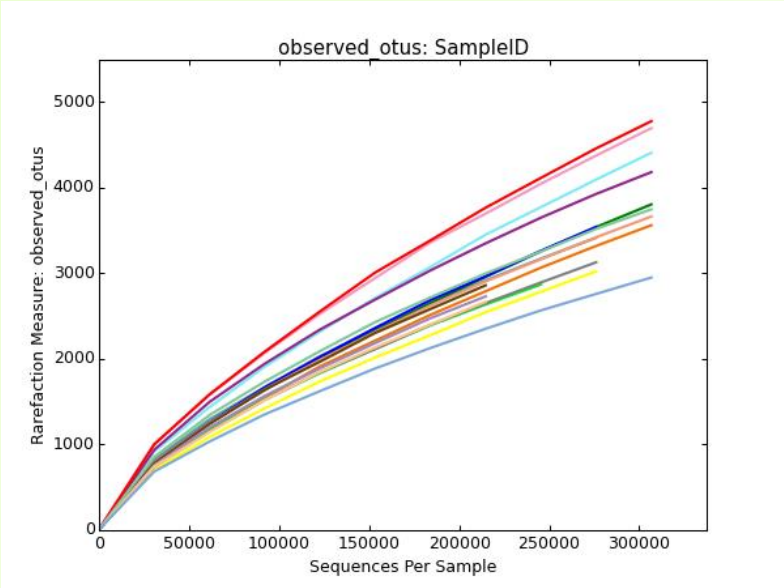


Different *Operational Taxonomic Units* (OTU) approaches and non-OTU approaches

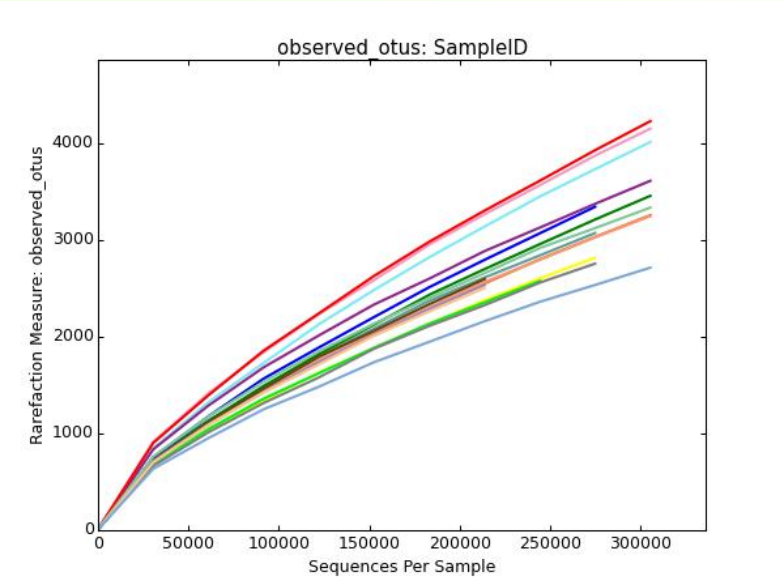
		De novo OTU clustering		Closed-reference OTU-assignment (uses ref. DB)	
		No chimera-screening	With chimera-screening	No pre chimera-screening	With pre chimera-screening
Total reads processed		5257222	5234178	5257222	5234178
reads assigned to OTUs		100%	100%	97%	97%
OTUs		29527	26306	2905	2884
OTUs assigned to	*named* genus	7229 (24%)	6217 (24%)	831 (29%)	826 (29%)
	named species	2328 (8%)	1862 (7%)	204 (7%)	203 (7%)
Unique taxa *names* assigned to OTUs		145	144	121	121
Unique taxa with	*named* genus	107 (74%)	107 (74%)	85 (70%)	85 (70%)
	named species	53 (37%)	53 (37%)	35 (29%)	35 (29%)

Rarefaction

- only *de novo* clustered OTUs shown here
- All samples considered **separately** here



No chimera-screening



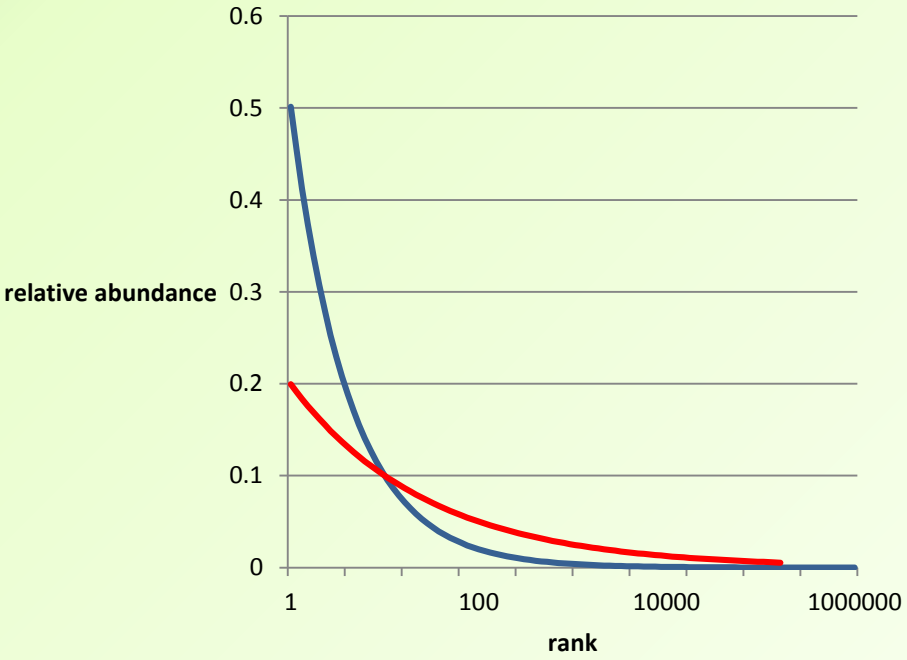
With pre chimera-screening

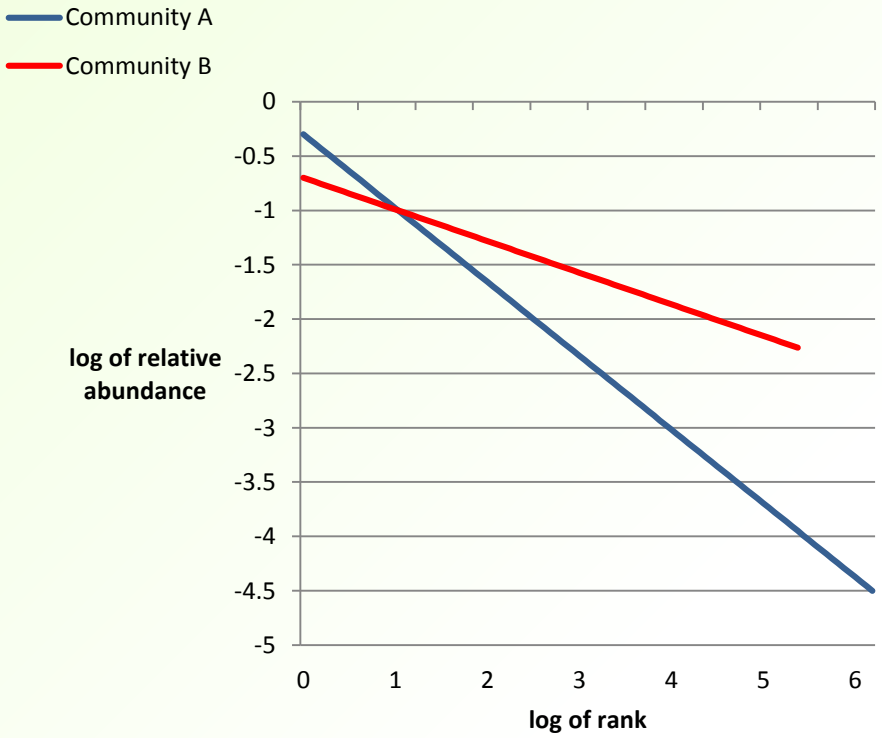
- It is also possible to extrapolate beyond the actual size
- Which you might be interested in doing, if your curve has not levelled off
- I.e. this is one way of calculating an estimate of richness (in the original community) from your observations (of the sample)
- *But* uncertainty rapidly increases as you extrapolate substantially beyond the sample size (Haegeman *et al.,* 2013)

- Rarefaction can also be performed on a per-sample basis
- E.g. 50 samples of the same thing
- Recalculate observed numbers by repeatedly analysing *n* samples of those 50

- The previous slides illustrated some differences resulting from different data-processing protocols
- For any given protocol, we would like to obtain the same results if we repeated the experiment again and again
  - But how likely is this, given the randomness of the experimental sampling process?
  - This is especially pertinent to the rarest phylotypes
  - And indeed features of the **abundance distribution in general**

Problems with rarefaction





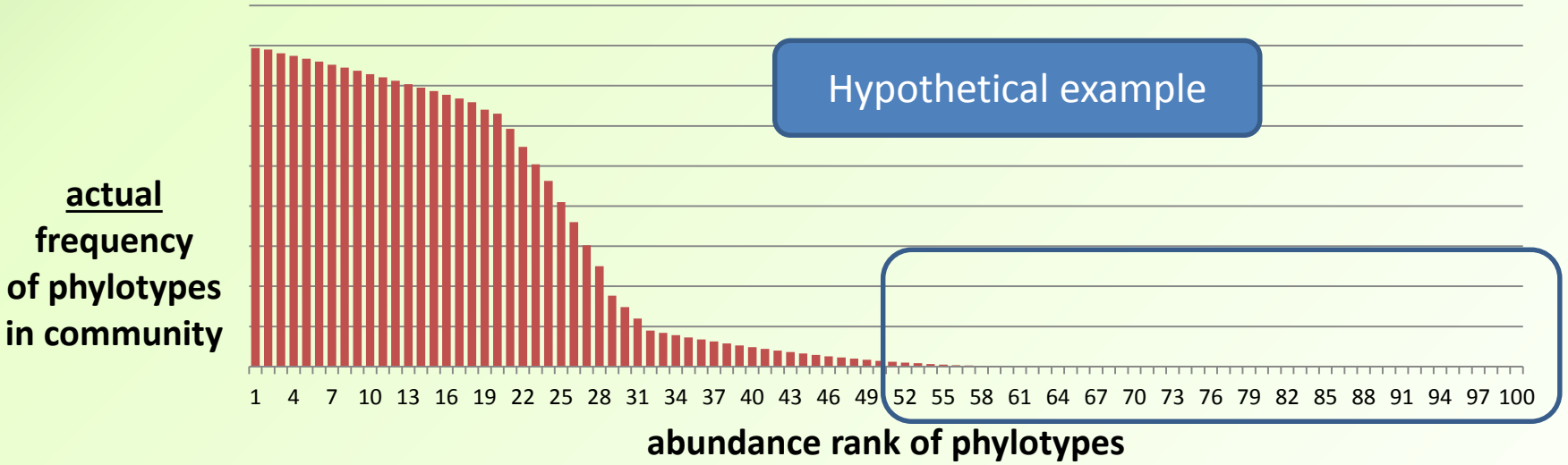
Haegeman *et al.* (2013)

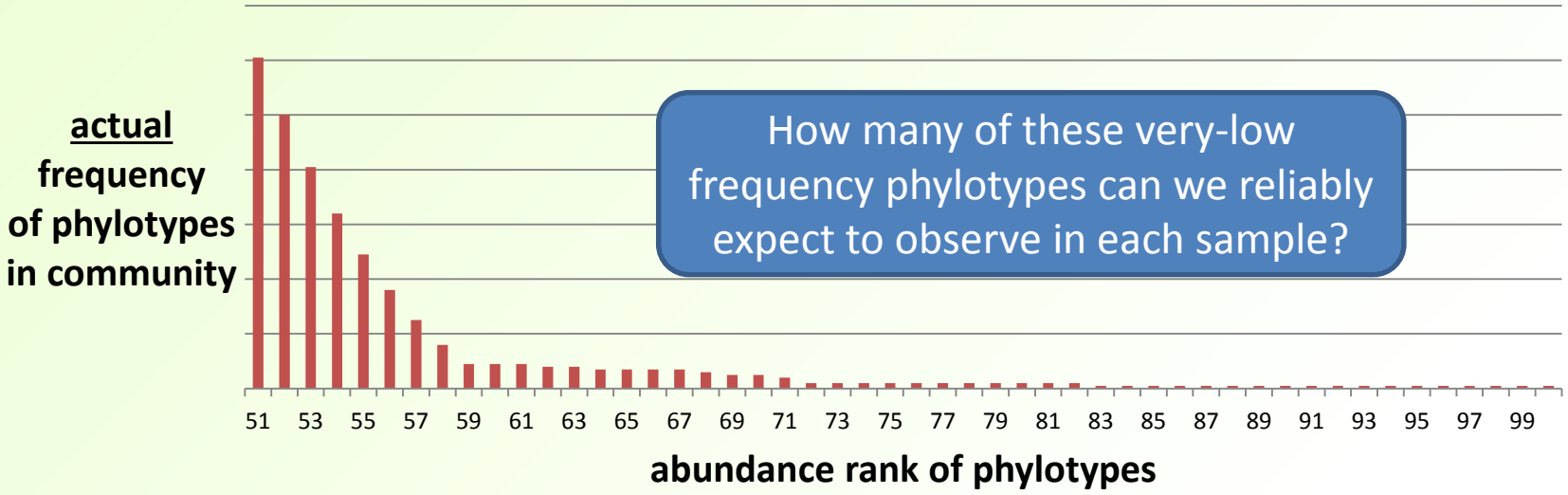
- [Figure 2 of Haegeman *et al.* (2013) https://www.nature.com/ismej/journal/v7/n6/fig_tab/ismej201310f2.html#figure-title]

- Three model communities
- S_n is the actual number of species in the community
- Sampling/rarefaction gives the reverse answer to the correct one
- With this sort of distribution, the problem does not get any better as the sample size increase
- How realistic are these model distributions? (Discuss…)

Long tails of rare types

Abundance versus rank: What shape is the tail? How long is it?





- We expect to ‘hit’ and ‘miss’ these very-low abundance phylotypes in a random way
- Can this expectation be used to estimate the true values in the community?
- The abundances of the most-rarely observed types can be used to estimate the number of types which were observed zero times by sampling
  - (but which are present)
- A principle first described by Good (1953)
  - “[Alan] Turing is acknowledged for the most interesting formula in this part of the work”

Traditional ecology versus DNA sequencing

- Good-Turing type estimators enable the estimation of the frequency of *events which have not yet happened*
- Such as, an estimation of the true frequency (abundance) of organisms which are in the community being studied – but which have not yet been observed
- But by using these techniques in DNA-sequencing, we will be estimating the occurrence of *rare DNA sequences* not yet observed
- Which will include:
  - **True DNA sequences not yet observed**
  - **Erroneous sequences caused by the sequencing platform, not yet observed**
  - **Chimera sequences not yet observed**
- So are these techniques less suitable for this situation, compared to, say, capturing invertebrates in pitfall traps?
  - Errors and misidentifications do also occur in traditional ecology sampling methods, so will also contribute to those stats
- In short: these types of estimators do not eliminate the effects of amplification/sequencing errors

A brief look at some of these types of estimators

“Abundance” versus “incidence”

- In this context, **abundance** means relative frequencies within a sample
  - How many times was each type observed?
- **Incidence** means the number of samples in which each type was observed
  - Irrespective of how often it was observed in each sample

Estimating richness from *abundance*

i.e. from relative frequencies of phylotypes in a sample

Chao1 (Chao, 1984)

$$\bar{\theta} = d + \frac{{n_1}^2}{2n_2}$$

- Estimator for θ, the actual number of phylotypes ("classes") – i.e. richness
  - d : total number of observed phylotypes
  - n_1 : number of phylotypes observed only once ('*singleton*')
  - n_2 : number of phylotypes observed only twice ('*doubleton*')
- Often written as:

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{{f_1}^2}{2f_2}$$

- Modified forms usually used, to allow for cases where f_2 is 0
  - Such as:

$$S_{\text{Chao1}} = S_{\text{obs}} + \frac{{f_1}^2}{2(f_2+1)} - \frac{f_1 f_2}{2(f_2+1)^2}$$

  - E.g. Kemp & Aller (2004)
  - Other forms exist in the literature
    - E.g. in Gotelli & Colwell (2011)

Chao estimators

- **Chao1**: Particularly appropriate for communities where **most phylotypes are relatively rare** (Chao, 1987; Kemp & Aller, 2004)
  - This probably describes the gut microbiome? (Discuss….)
- **ACE** (Chao & Lee, 1992) : considers all observed phylotypes as either '**rare**' or '**abundant**', and uses the numbers of each, as well as the number of **singletons**, explicitly
- Some assessments using earlier sequencing platforms for 16S rDNA
  - (thus, very small sample sizes – and much longer sequences - compared to today)
  - E.g. Kemp & Aller (2004)
  - also used hypothetical, model distributions of frequencies
  - concluded **Chao1 well-suited for estimating phylotype richness from prokaryotic 16S rDNA**
  - ACE did not perform as well

- **Jackknife estimators** for abundance data
  - First order: $S_{jackknife1} = S_{obs} + f_1$
  - Second-order: $S_{jackknife2} = S_{obs} + 2f_1 - f_2$
- Burnham & Overton (1979)
- See also
  - Gotelli & Colwell (2011)
  - Hortal *et al.* (2006)
  - and references therein
- Many other estimators

Estimating richness from *incidence*

i.e. from how many samples a phylotype is observed in

Estimating richness from incidence

- Requires multiple samples
  - In contrast to abundance-methods
- Abundance in each sample is relevant only in the consideration of whether:
  - The frequency is zero
  - The frequency is non-zero
- Sizes of non-zero frequencies are **irrelevant**

- Chao2 (Chao, 1987)

$$S_{\text{Chao2}} = S_{\text{obs}} + \frac{{q_1}^2}{2q_2}$$

  - Identical in form to Chao1
    - But q_1 is the number of phylotypes which occur in only 1 sample
    - q_2 is the number of phylotypes which occur in only 2 samples
- ICE (Lee & Chao, 1994)
- Jackknife estimators for incidence
  - E.g. Smith & van Belle (1984)

- Many richness estimators exist for both abundance- and incidence-based frequencies
- For a description of some of these, see:
  - Gotelli & Colwell (2011)
  - Hortal *et al.* (2006)

How reliable is all this?

- How concerned should we be with richness (numbers of types)?
- Consider this from two points of view
  1. What are we actually interested in, given what we are able to sample?
  2. Rigorous assessments of different richness estimators

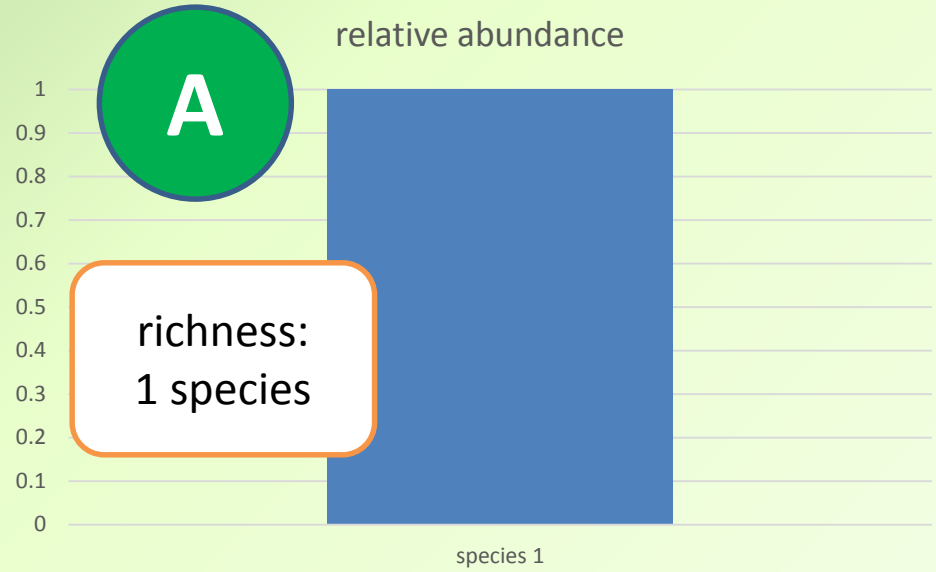
What are we interested in?

- Given the expected uncertainty in determining the exact number of (phylo)types, should we be more interested in :
  - determining the number of types which can be reliably observed?
  - determining the number of types which we actually care about?
- which is another way of asking:
  - How miniscule does an **actual abundance in the original community** need to be, in order for us to treat it the same as if it was zero?

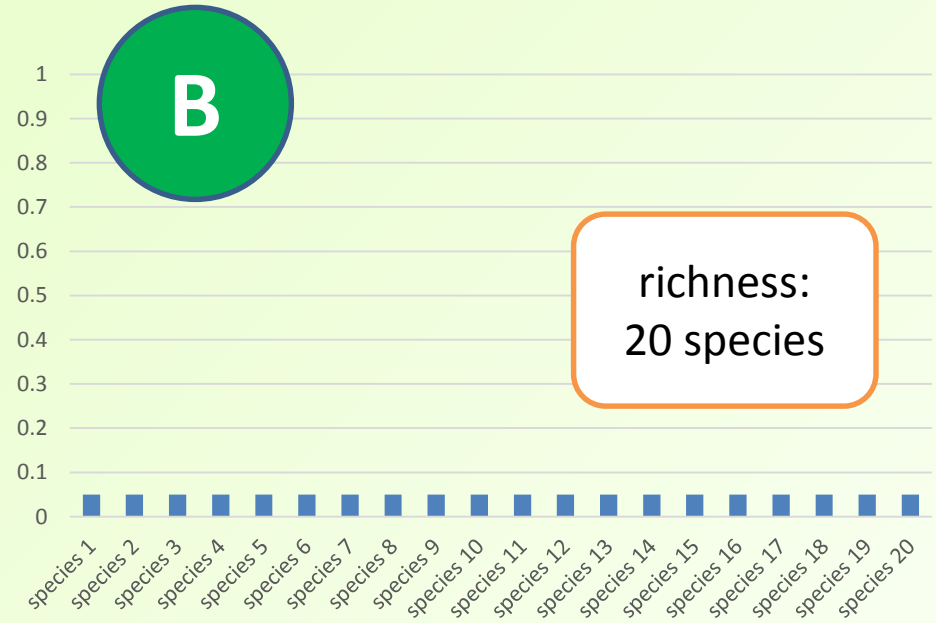
…which is another way of describing the limitation of richness

In this example, A, B and C represent true relative abundances in communities

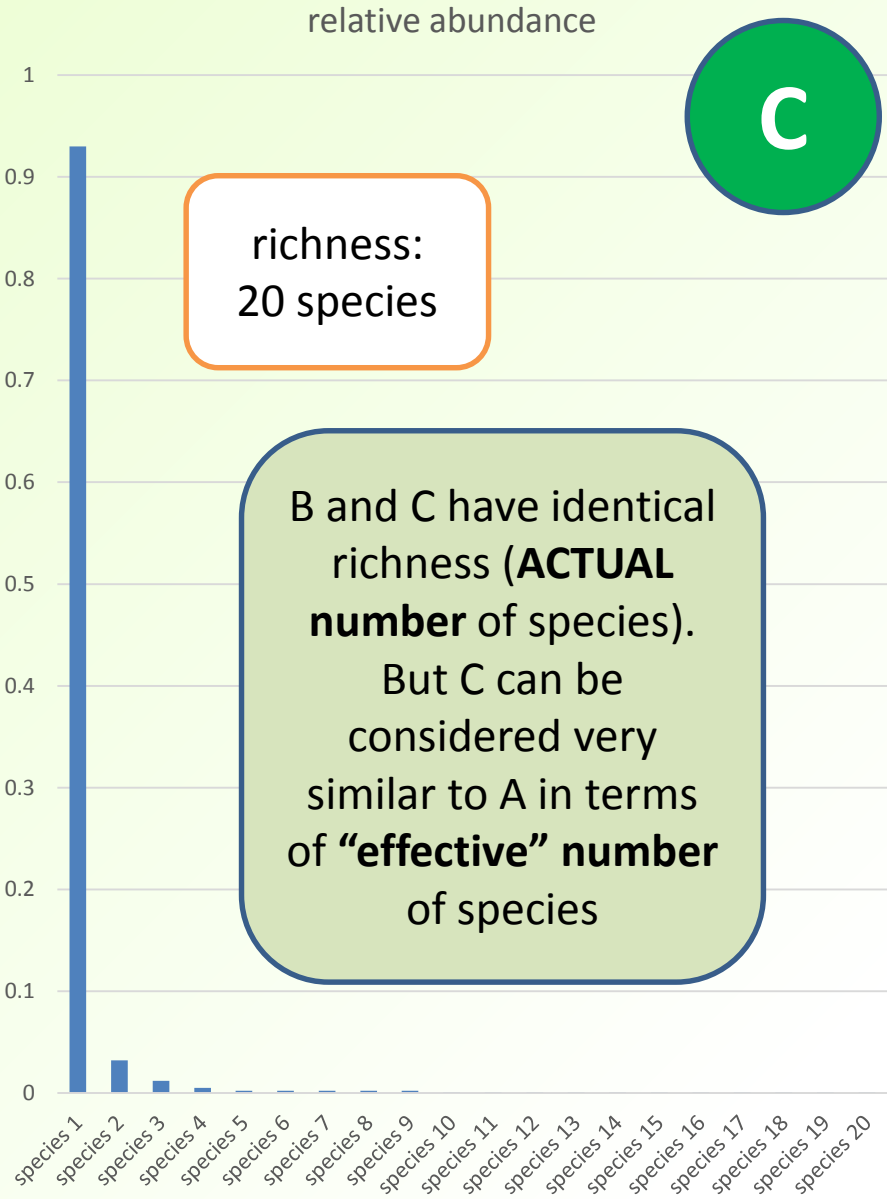
(rather than observations in samples)



richness:
1 species



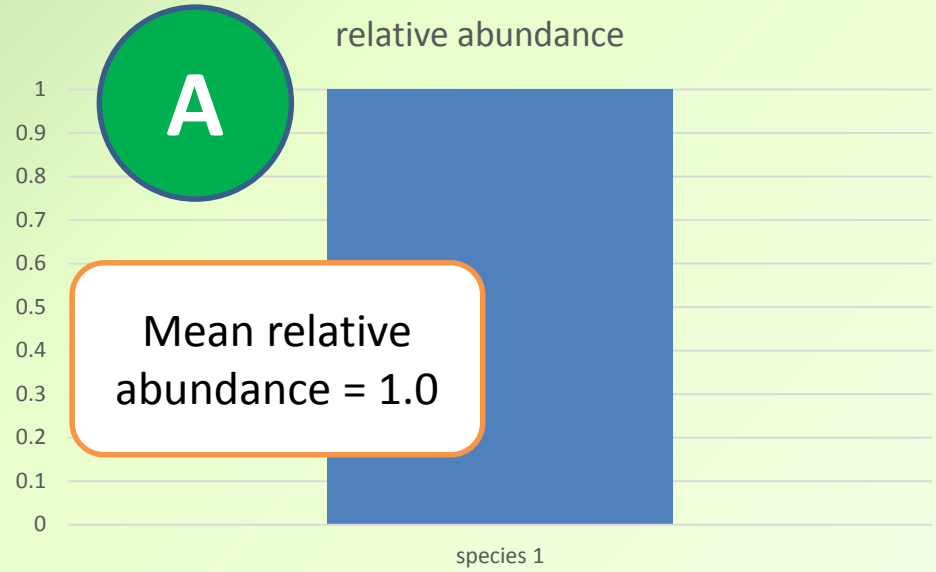
richness:
20 species

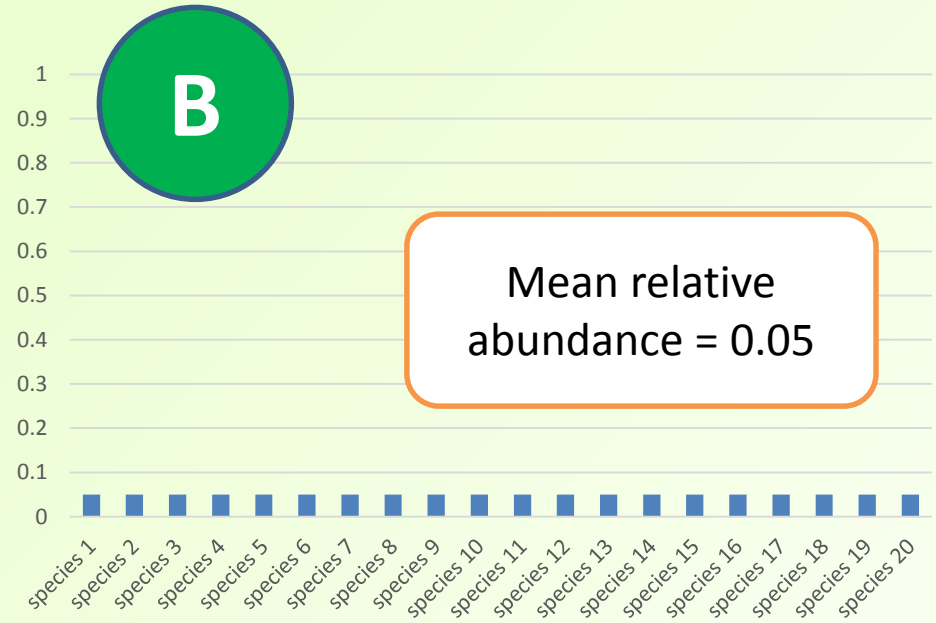


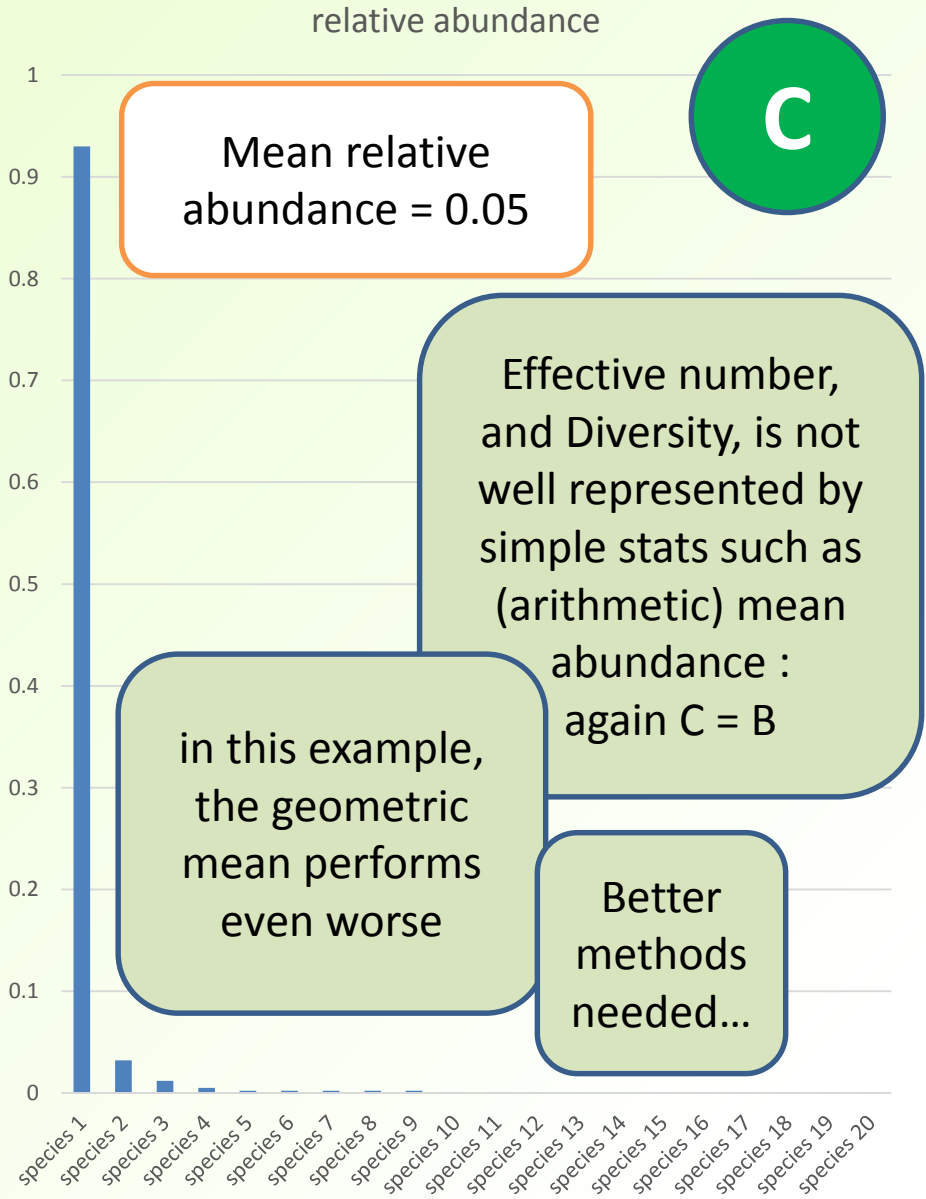
richness:
20 species

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

- The **effective number** of phylotypes results from a consideration of "**dominance**" versus "**evenness**", and can be quantified (by various methods).
- It is also simply related to measures of **diversity**
  - Which describe distributions of relative abundance
  - More in the next session…
- It also relates to our ability to reliably and reproducibly estimate the number of phylotypes by sampling
  - The effective number is more reproducible than the actual number


A
relative abundance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
species 1
Mean relative abundance = 1.0


B
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
species 1
species 2
species 3
species 4
species 5
species 6
species 7
species 8
species 9
species 10
species 11
species 12
species 13
species 14
species 15
species 16
species 17
species 18
species 19
species 20
Mean relative abundance = 0.05


relative abundance
C
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
species 1
species 2
species 3
species 4
species 5
species 6
species 7
species 8
species 9
species 10
species 11
species 12
species 13
species 14
species 15
species 16
species 17
species 18
species 19
species 20
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…

Assessment of estimators

- Numerous in the literature
- That Haegeman *et al.* (2013) paper again:
- "Species richness cannot be estimated from sample data alone"
- "We claim that sample data is always consistent with very different community structures"
- "computation shows that the rarefaction curves do not depend on the abundance distribution of the rare species"
- "We have shown that the number of species in a community cannot be reliably estimated from sample data"
- For anyone who has analysed many sets of 16S-sequenced samples from many experiments, it may be a relief to hear all this…

Recommendations

- Measurements of richness are easy to obtain from your data
- Don’t use measurements of richness
  - At least, quote them
  - but do not rely on them as a descriptor of your samples
- Bad news for Richness
- Better news for Diversity?

References (1)

- Burnham K.P. and Overton W.S. (1979) Robust estimation of population size when capture probabilities vary among animals, *Ecology* **60:** 927-236
- Chao A. (1984) Nonparametric Estimation of the Number of Classes in a Population, *Scand J. Stat.* **11** (4): 265-270
- Chao A. (1987) Estimating the Population Size for Capture-Recapture Data with Unequal Catchability, *Biometrics* **43** (4): 783-791
- Chao A. and Lee S.-M. (1992) Estimating the number of species in a stochastic abundance model, *Biometrics*, **43**: 783-791
- Good I.J. (1953) The Population Frequencies of Species and the Estimation of Population Parameters, *Biometrika* **40** (3,4): 237-264
- Gotelli, N.J. and Colwell R.K. (2011) Estimating species richness, in *Biological Diversity: Frontiers in Measurement and Assessment*, Chapter 4, pp 39-54, Eds Magurran AE and McGill BJ, Oxford University Press

References (2)

- 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
- Hortal J., Borges P.A.V. and Gaspar C. (2006) Evaluating the performance of species richness estimators: sensitivity to sample grain size, *J. Anim. Ecol.* **75:** 274-287
- Kemp P.F. and Aller J.Y. (2004) Estimating prokaryotic diversity: When are 16S rDNA libraries large enough? *Limnol. Oceanogr. Meth.* **2:** 114-125
- Lee S.-M. and Chao A. (1994) Estimating population size via sample coverage for closed capture-recapture models, *Biometrics* **50:** 88-97
- Smith E.P. and van Belle G. (1984) Nonparametric estimation of species richness, *Biometrics* **40:** 119-129