DTU

DTU Health Technology
Bioinformatics

Quantitative metagenomics

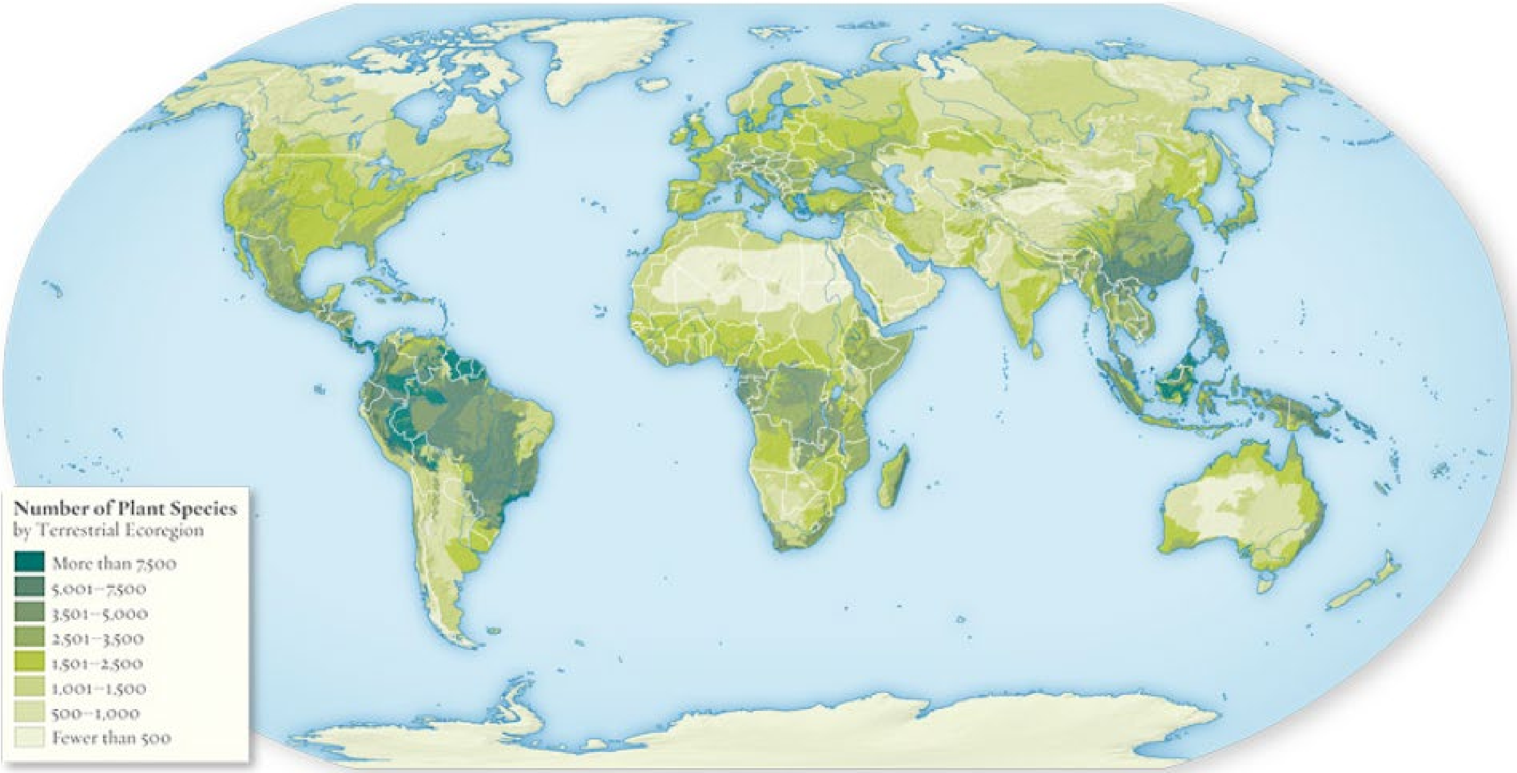
Gisle Vestergaard
Associate Professor
Section of Bioinformatics
Technical University of Denmark
gisves@dtu.dk

Menu

- Diversity measurements
  - Alpha & beta diversity
- Normalization
- Sequencing data is compositional data

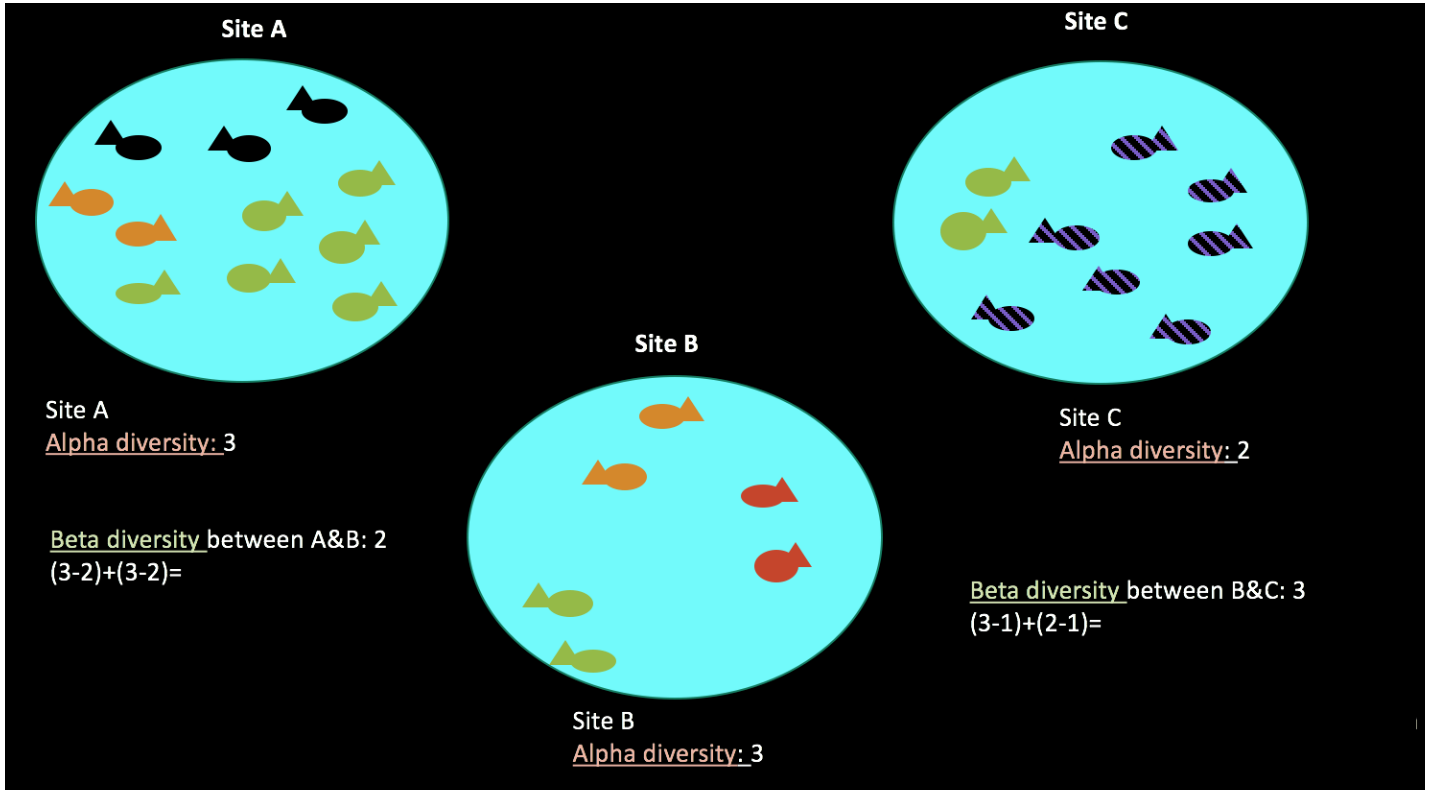
Classical measures

- Abundance
- Richness
- Rarefaction
- Diversity
  - Alpha
  - Beta



Describing the spatial component of biodiversity

- Alpha diversity (within sample)
- Beta diversity (between samples)
- We can compare both alpha and beta diversity



Site A
Alpha diversity: 3

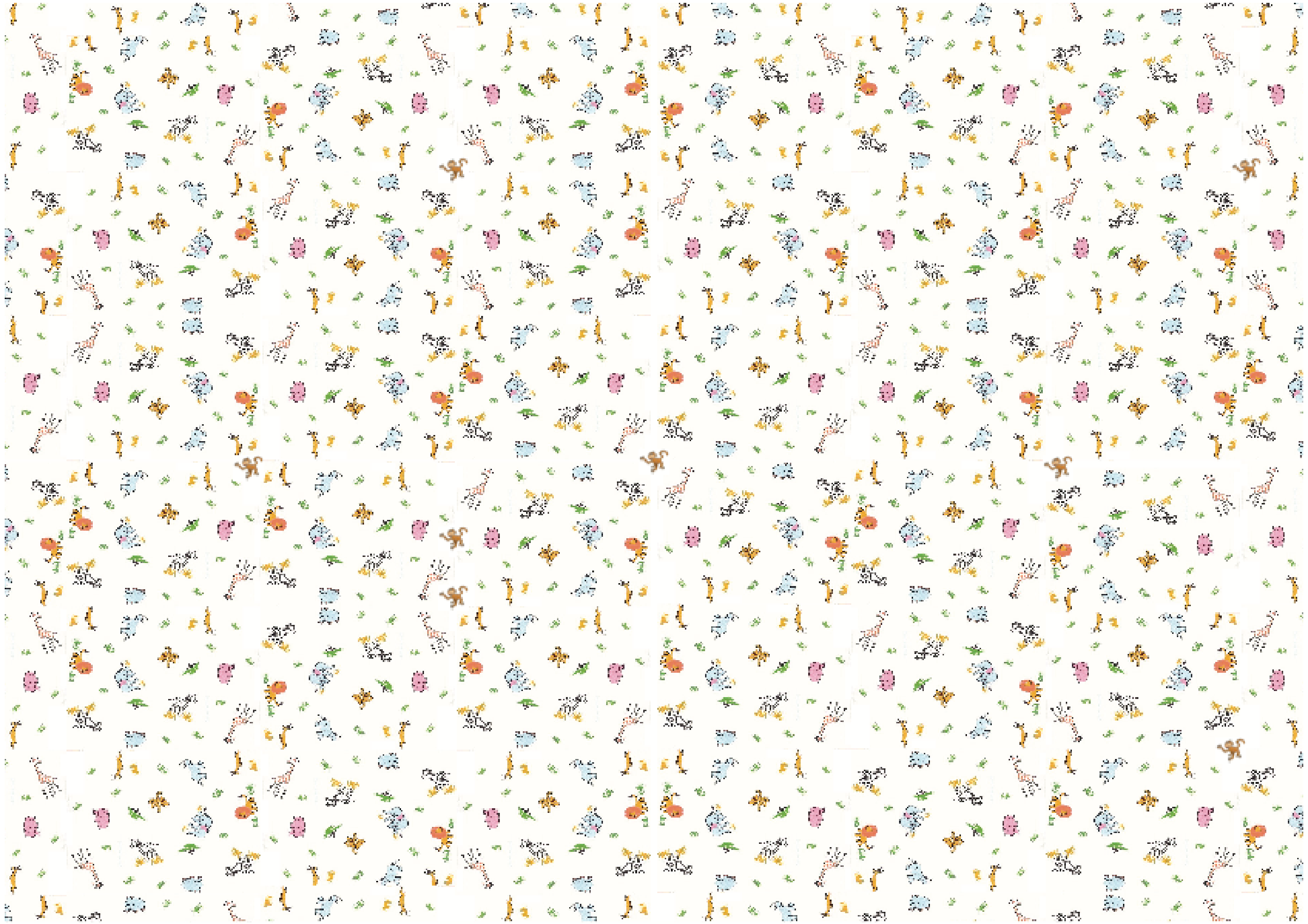
Beta diversity between A&B: 2
(3-2)+(3-2)=

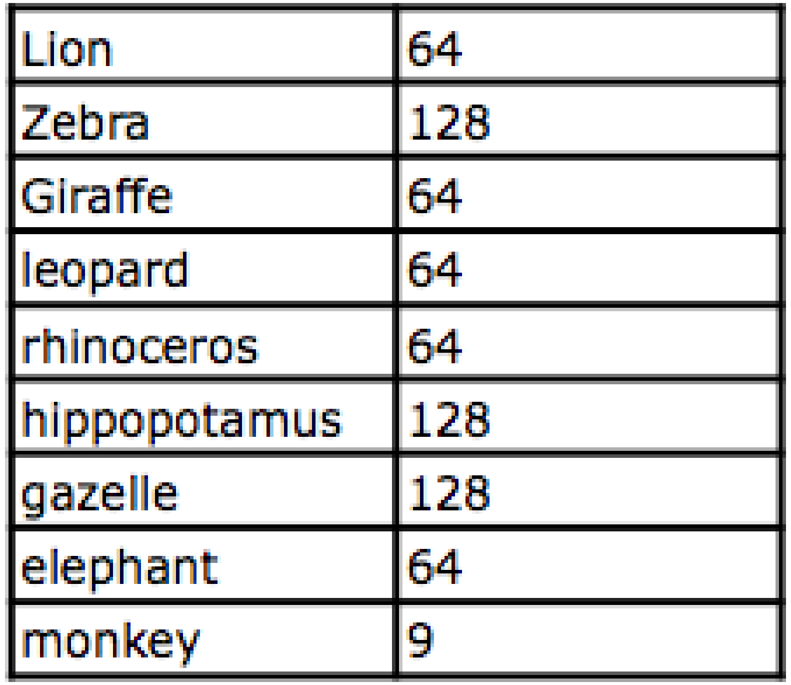
Site C
Alpha diversity: 2

Beta diversity between B&C: 3
(3-1)+(2-1)=

Site B
Alpha diversity: 3

Abundance (counts)

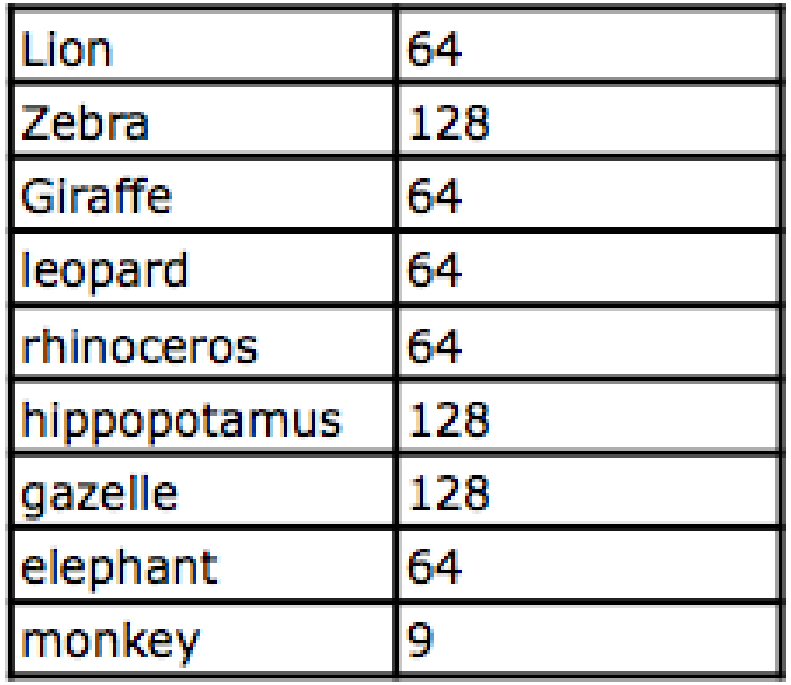




Lion	64
Zebra	128
Giraffe	64
leopard	64
rhinoceros	64
hippopotamus	128
gazelle	128
elephant	64
monkey	9

Species richness

- The number of different species in a system

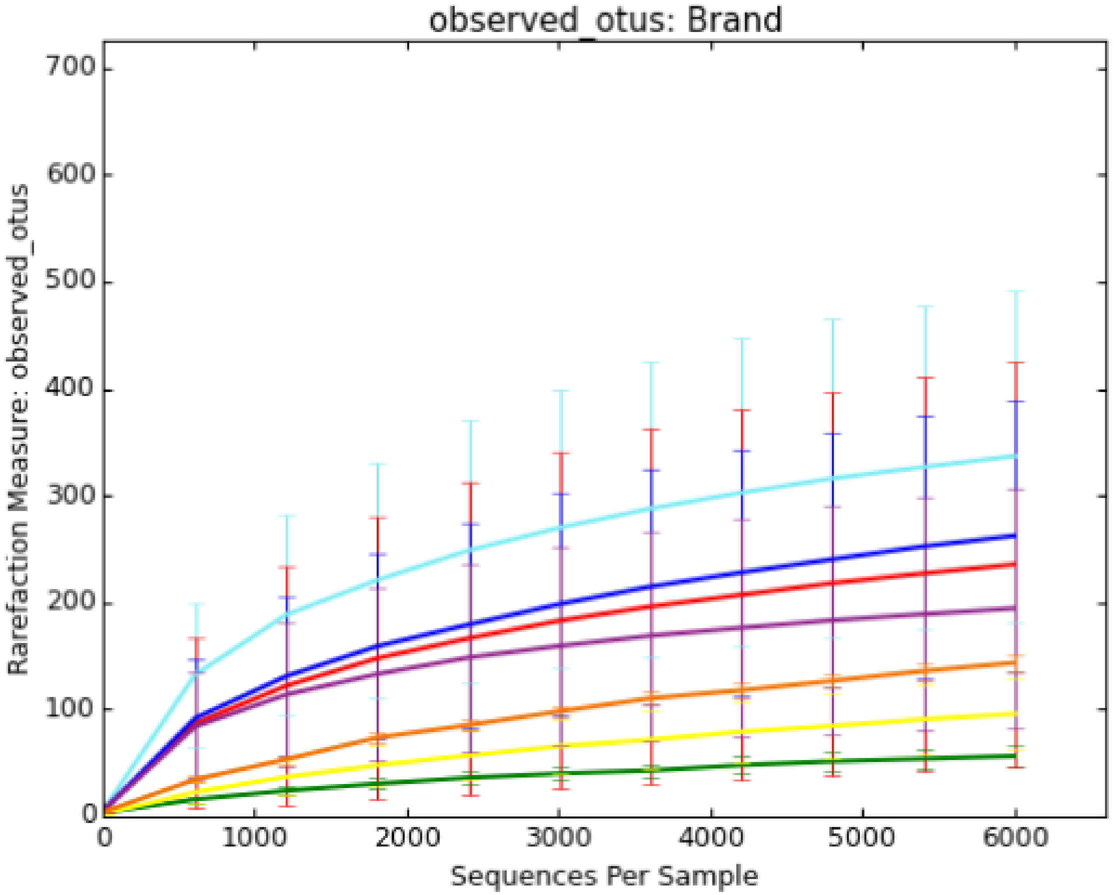


Lion	64
Zebra	128
Giraffe	64
leopard	64
rhinoceros	64
hippopotamus	128
gazelle	128
elephant	64
monkey	9

9 observed species

Rarefaction

- Species richness is a function of our no. observations
- When have we sampled enough?
- Mostly used for 16s rRNA amplicons…why?



Shannon index

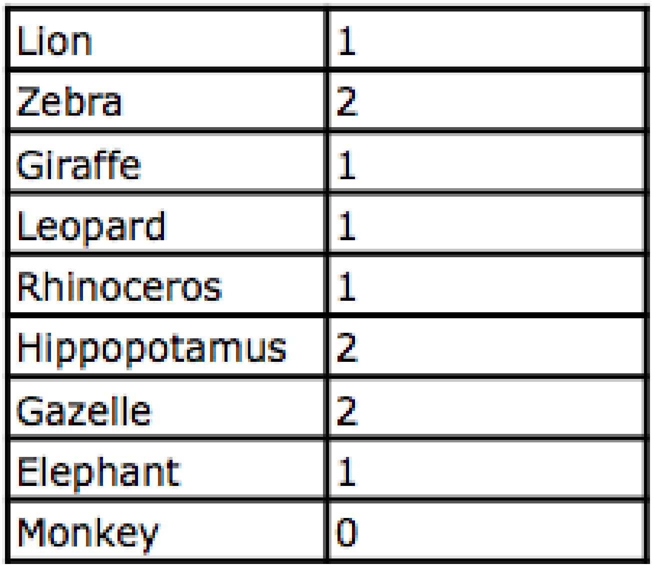
- Incorporates species richness & eveness
- Quantify the entropy (information content)
- Quantifies the uncertainty (degree of surprise) associated with a prediction
- The Shannon index increases as both the richness and the evenness of the community increase
- Typical values are generally between 1.5 and 3.5 in most ecological studies, and the index is rarely greater than 4

$$H' = -\sum_{i=1}^{R} p_i \ln p_i \qquad H' = -(\ln p_1^{p1} + \ln p_2^{p2} + \ln p_3^{p3} + \cdots + \ln p_R^{pR})$$

P_i = species proportion

R = observed species

Alpha diversity



Lion	1
Zebra	2
Giraffe	1
Leopard	1
Rhinoceros	1
Hippopotamus	2
Gazelle	2
Elephant	1
Monkey	0

$$H' = -(\ln p_1^{p1} + \ln p_2^{p2} + \ln p_3^{p3} + \cdots + \ln p_R^{pR})$$

11 animals (NOT species) meaning each animal is 0.09 of the total abundance

H' = -(ln($0.09^{0.09}$) + ln($0.18^{0.18}$) + … = 2.0

Bray-curtis dissimilarity

$0 \leq B \leq 1$

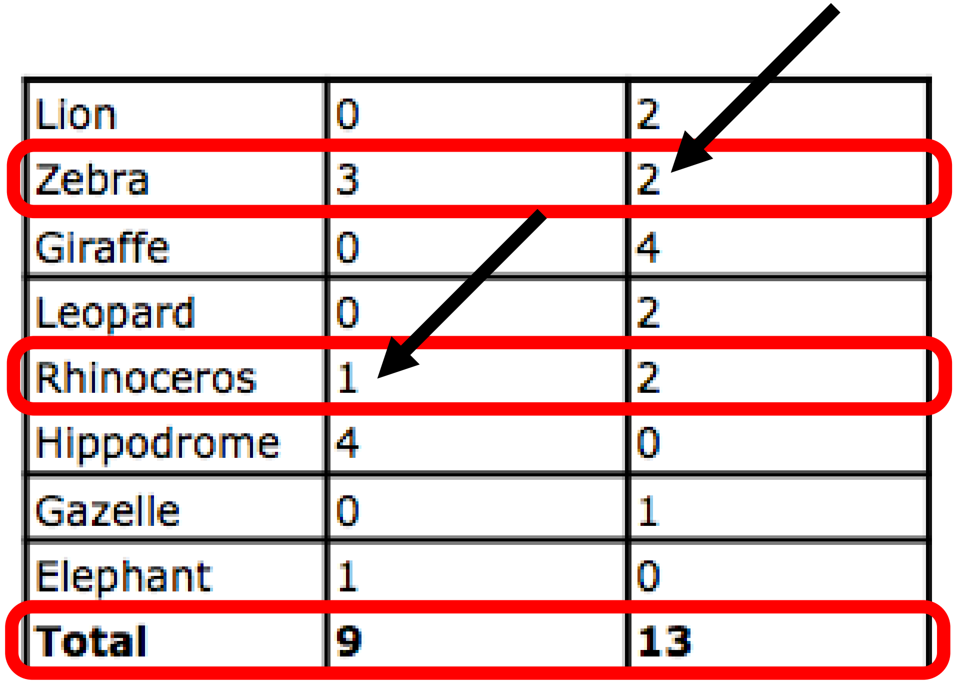
$B_{ij} = 1 - 2C_{ij} / (S_i + S_j)$

C = sum of the lowest count of all common species

S = total count of the sample

1 means that they do not share anything

$B_{s1s2} = 1 - 2*(2+1) / (9 + 13) = 0.73$



Lion	0	2
Zebra	3	2
Giraffe	0	4
Leopard	0	2
Rhinoceros	1	2
Hippodrome	4	0
Gazelle	0	1
Elephant	1	0
Total	**9**	**13**

Sampling effect

- To be fair we should sample equally in the systems we investigate

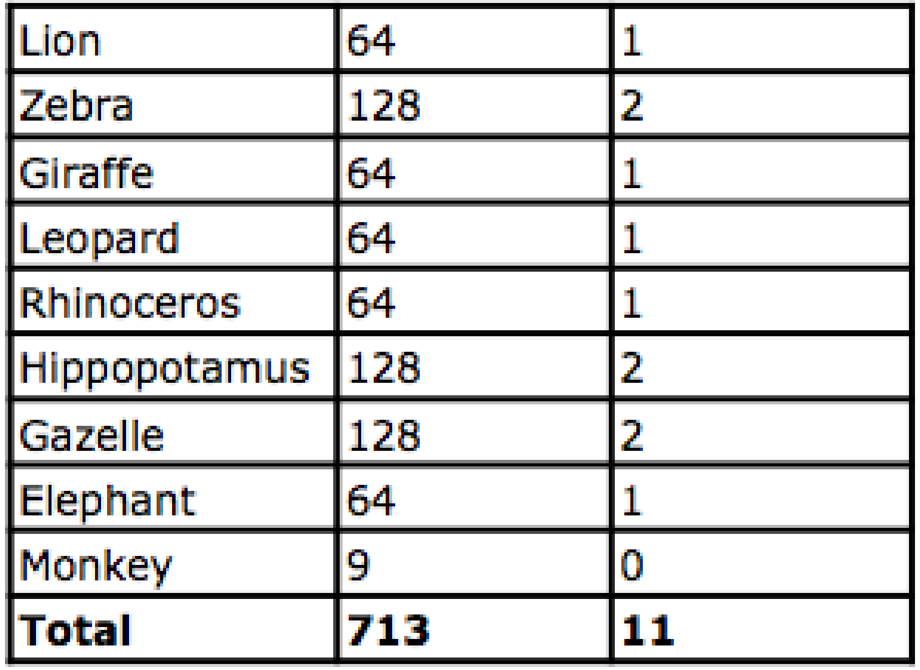
Sample sizes

Sample sizes

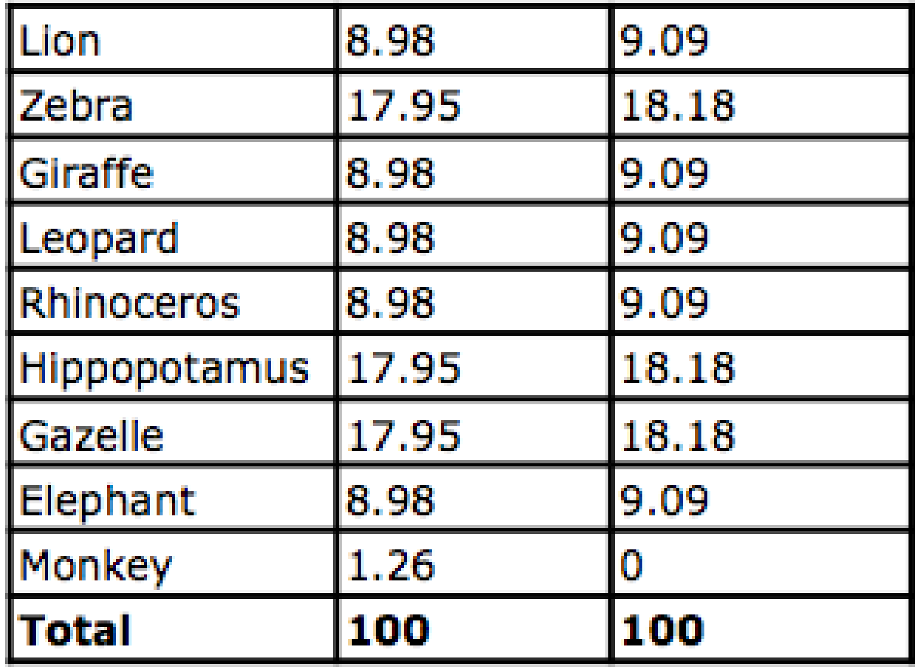
- Accounting for different sample sizes:
  - Normalise to sample size
  - Rarefy (downsize) samples
  - Statistically model the variance

Normalizing

$$N = n_i/n_{tot}$$



Lion	64	1
Zebra	128	2
Giraffe	64	1
Leopard	64	1
Rhinoceros	64	1
Hippopotamus	128	2
Gazelle	128	2
Elephant	64	1
Monkey	9	0
Total	**713**	**11**

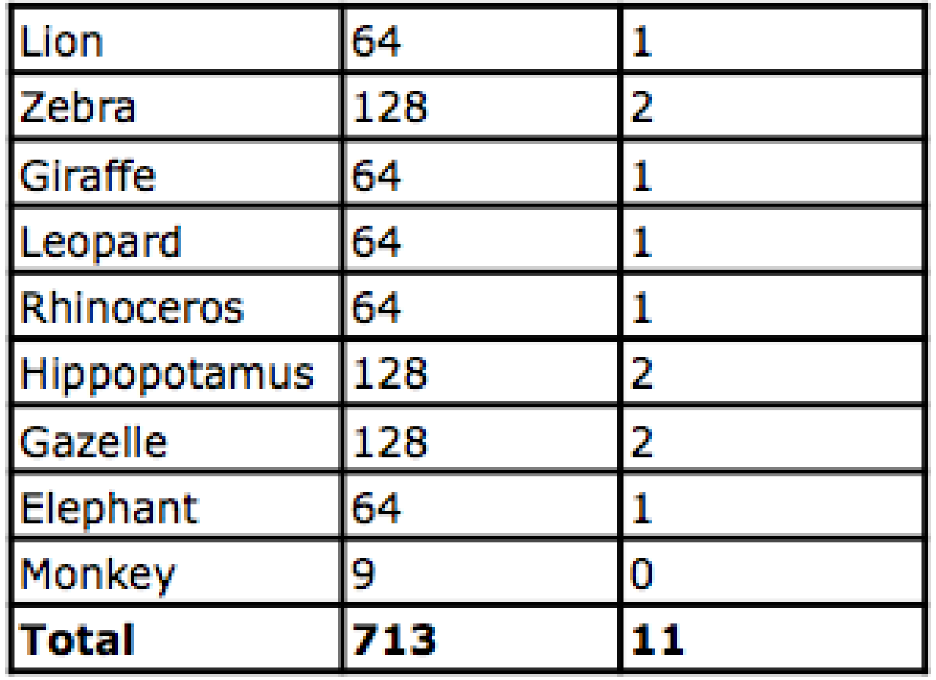


Lion	8.98	9.09
Zebra	17.95	18.18
Giraffe	8.98	9.09
Leopard	8.98	9.09
Rhinoceros	8.98	9.09
Hippopotamus	17.95	18.18
Gazelle	17.95	18.18
Elephant	8.98	9.09
Monkey	1.26	0
Total	**100**	**100**

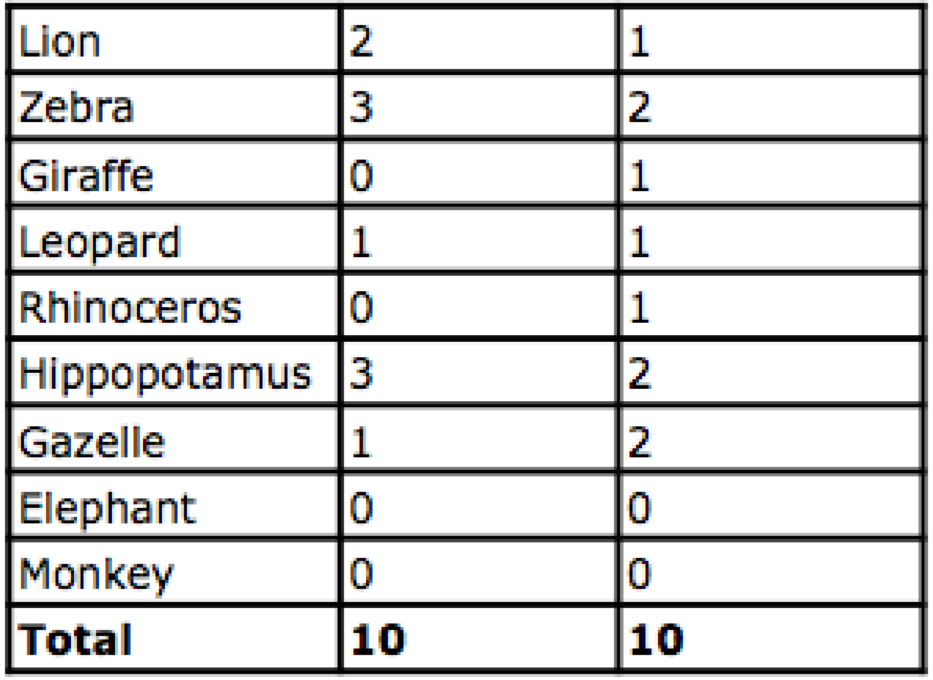
Issue with different sampling power (higher chance of observing rare species) and does not take compositional nature into account

Downsize / rarefy

Resample x amount of observations



Lion	64	1
Zebra	128	2
Giraffe	64	1
Leopard	64	1
Rhinoceros	64	1
Hippopotamus	128	2
Gazelle	128	2
Elephant	64	1
Monkey	9	0
Total	**713**	**11**



Lion	2	1
Zebra	3	2
Giraffe	0	1
Leopard	1	1
Rhinoceros	0	1
Hippopotamus	3	2
Gazelle	1	2
Elephant	0	0
Monkey	0	0
Total	**10**	**10**

Downsize / rarefy

- Select the target depth carefully
- The more reads we keep the more sensitive
- We may have to remove samples with few counts
- We might throw away a lot of data
- Still does not take compositional nature of data into account

Compositional data

- Arbitrary total
  - Sequencing depth never 100%
- Species can co-exist without abundance inter-influences
  - Independence between abundance is affected by the capacity of the sequencing instrument
  - Sequencing instrument has fixed number of slots

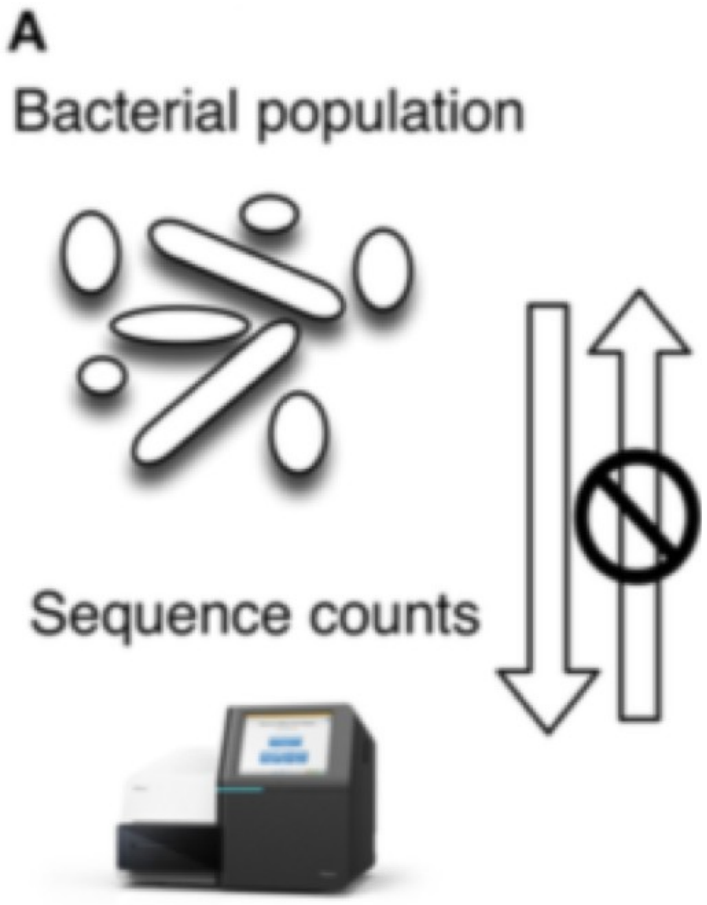
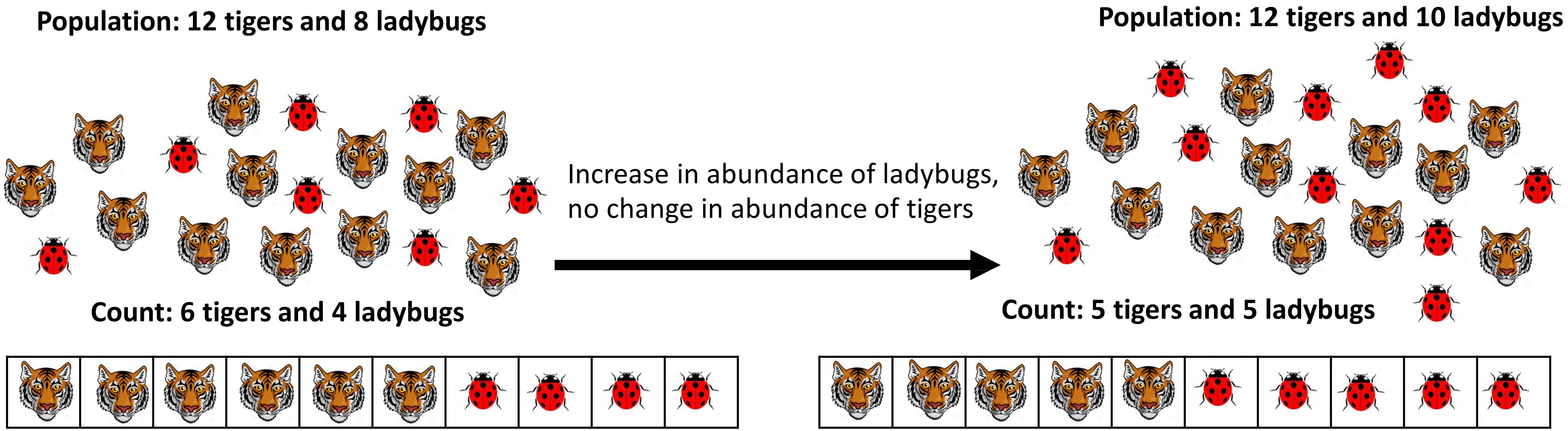


Figure from: Gloor, Gregory B. *et al.*, Microbiome Datasets Are Compositional: And This Is Not Optional. *Frontiers in Microbiology* **8** (2017)

Compositional data problems

- **Example**: an environment containing both tigers and ladybugs
  - The abundances of the two are not affected by each other
  - If the abundance of the ladybugs increases some of the slots with tigers must instead be filled by ladybugs
  - i.e. the two environmentally independent species are affecting the read count of each other



Population: 12 tigers and 8 ladybugs

Count: 6 tigers and 4 ladybugs

Increase in abundance of ladybugs, no change in abundance of tigers

Population: 12 tigers and 10 ladybugs

Count: 5 tigers and 5 ladybugs

Dealing with compositional data

- Statistically model the variance & heteroscedasticity

- Use packages developed for RNA-seq such as DESeq2 and edgeR

- DESeq2 takes raw counts divided by sample-specific size factors determined by median ratio of gene counts relative to geometric mean per gene
(See this link for a brilliant explanation)