Introducing Microbiome Bioinformatics

Part 10.

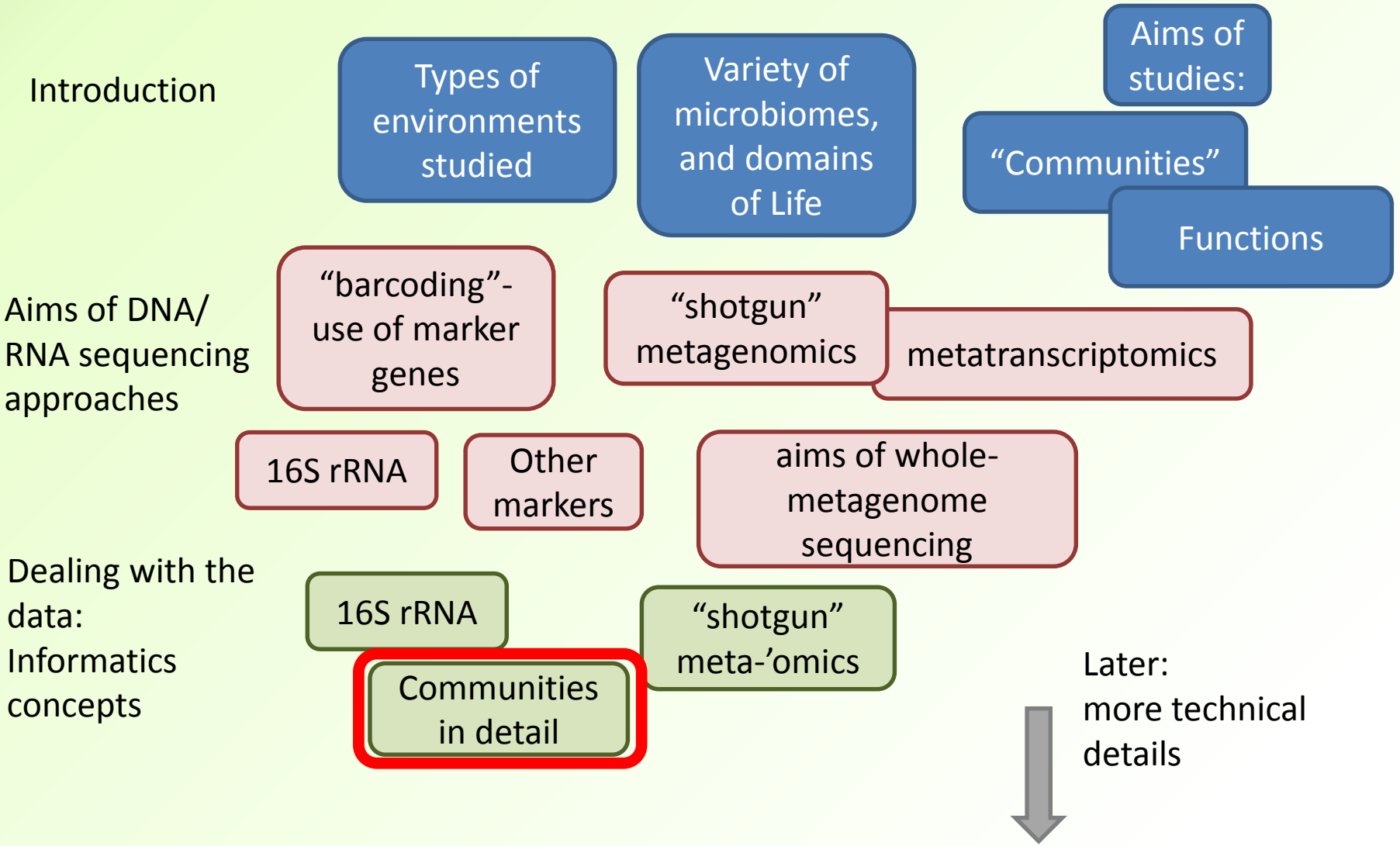
Microbial ecology –

Diversity (part 3)

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

- 9 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***
    - *Contact* ***Mark Fernandes,*** *or me*
- Informal and flexible
  - Please interrupt and ask questions
  - **Suggestions for topics for further focus**

Series of talks

Slideshows - http://ghfs1.quadram.ac.uk/ghfs/

- 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: 16/6/2017 – continuing microbial ecology: community diversity : true diversity
- Part 10: today – concluding diversity (for now);

Future talk(s)

- None planned for August 
- September
  - 8th
  - 22nd

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
- (16th June) :
  - True diversity
  - α-diversity, β-diversity, γ-diversity
- Today
  - Phylogenetic Diversity
  - intersample distances with UniFrac

Recap

α-diversity,
β-diversity,
γ-diversity

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?

γ-diversity is the Diversity of **all the samples collectively**

"constituent compositional unit"

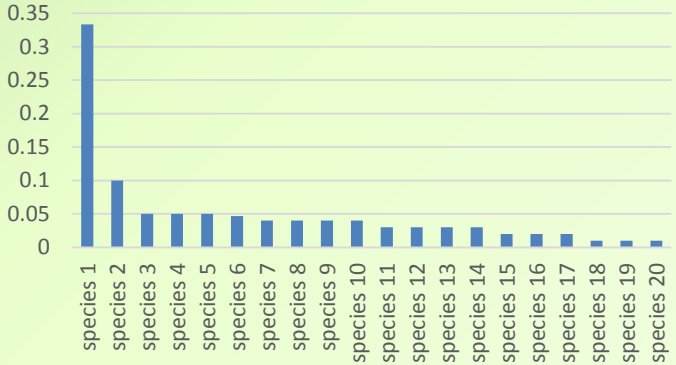
(such as a localised ecosystem in a larger region, or a sample from it)

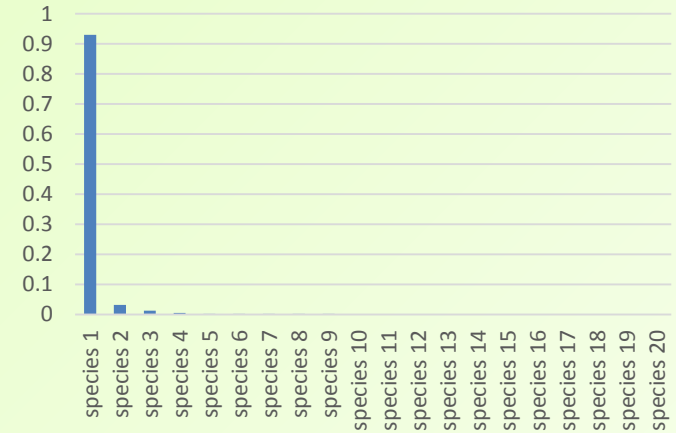
..is also equivalent to a constituent sampling unit in general

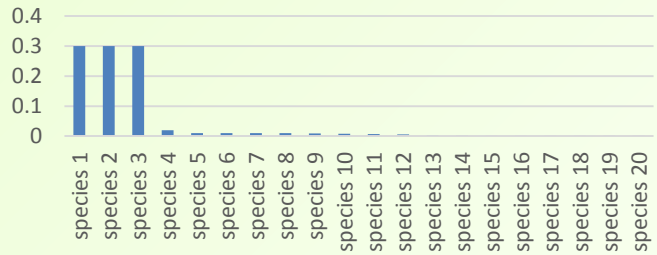
Such as multiple faecal samples from the same host

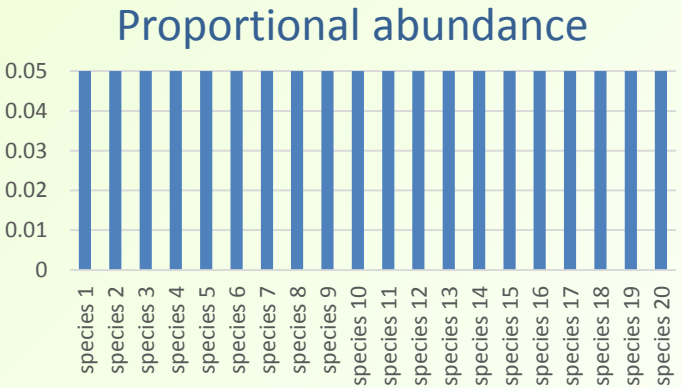
Note also that **we** are defining what each samping unit is, and how many units there are (this may seem obvious, but it's not the only way)

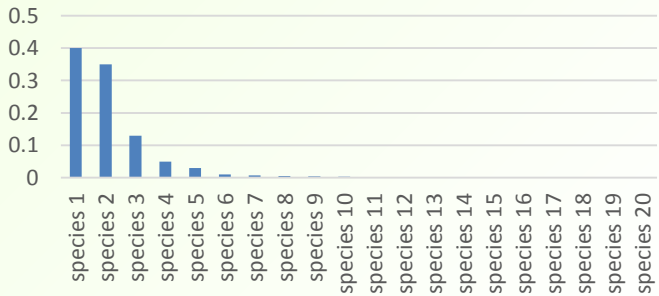
How to measure diversity?

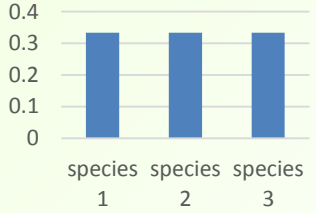


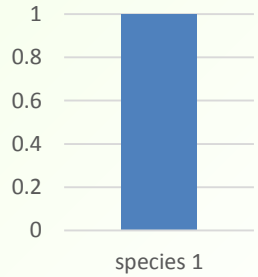












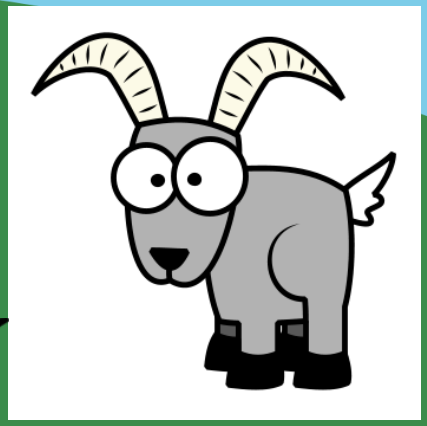
How to measure diversity?

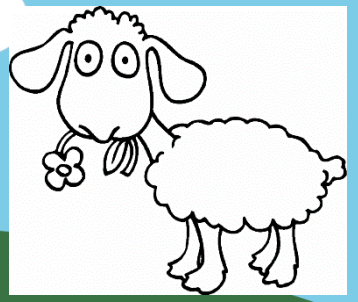
- Richness (number of species, OTUs etc):
  - not recommended
- A much better measure:
  - the “effective number of species”
  - This is the inverse of a **weighted generalised mean** of the proportional abundances
- Hill diversity of order q : $${}^{q}D = \left(\sum x_i^q \right)^{1/(1-q)}$$
- ${}^{0}D$ = richness
- The more useful ${}^{1}D$ and ${}^{2}D$ are closely related to the Shannon index and the Simpson index, respectively

What's wrong with this?

Ecosystem 1

Two constituent compositional units (samples)

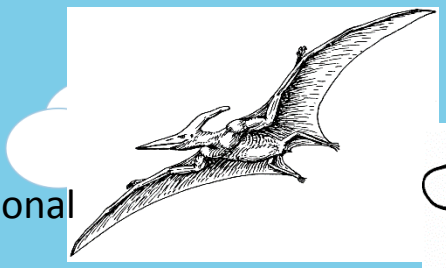


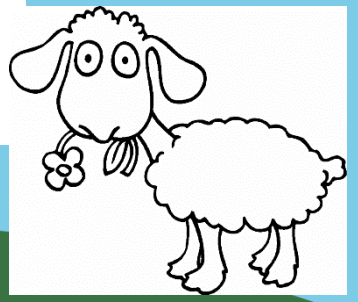


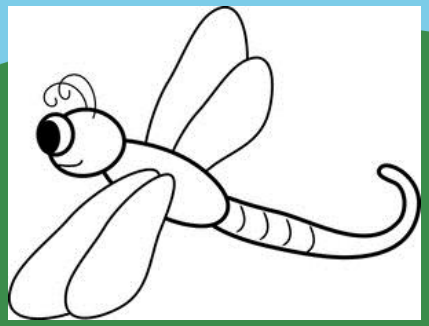
What's wrong with this?

Ecosystem 2

Two constituent compositional units (samples)







Assuming proportional abundances are the same as in Ecosystem 1
then both samples will have the same α diversities as before, and β and γ diversities will also be the same as in Ecosystem 1

Pros and cons

- True diversity measures we have seen so far (effective numbers of species, or OTUs etc) are fine:
- if you want a description of the distribution of abundances of different things
- But they treat all different species as equally different
- What if we want to take account of how similar or different the species are, within a sample?
  - Or the full complement of samples/environment?

Phylogenetic Diversity (Faith, 1992)

- The method is basically:
- Construct a phylogenetic tree
- Using whatever method
- Of **all the species** (or OTUs, etc…) **found in all of the samples in the study**
- The Phylogenetic Diversity (**PD**) of all samples collectively (γ-diversity) is the sum of all branch lengths
- The PD of a **single sample** (α-diversity) is the **sum of the lengths of all branches required to connect the species present in that sample**….
  - ***AND the root of the tree***

[See Fig 1(a) :

https://doi.org/10.1016/0006-3207(92)91201-3]

Fig 1(a) from Faith (1992) *Biological Conservation* **61,** 1-10

Group of 4 taxa (e.g. found in one sample or location) Bold lines constitute the ***minimum spanning path***

[See Fig 1(b) :

https://doi.org/10.1016/0006-3207(92)91201-3]

Fig 1(b) *Ibid*

[
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674678/figure/f3-ebo-02-121/
]

Fig 3 from Faith & Baker (2006), *Evol. Bioinform. Online* **2**, 121-128
https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2674678/figure/f3-ebo-02-121/

Comparing samples using phylogenetic trees

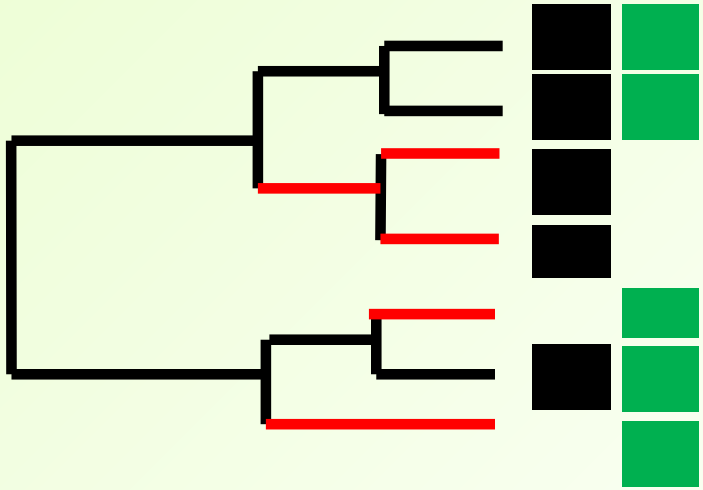
- PD is straightforward (once you have a tree) to calculate
- For individual samples
- For all samples of the experiment
- But, when comparing two samples, we really want to know:
- how much of the tree is shared by the two samples, and how much is not shared
- UniFrac (Lozupone & Knight, 2005)

UniFrac

- Provides a measure of **distance** between **two** samples
- It is the proportion of the total lengths of all the branches of the tree which are **not** shared
- Like PD, the original UniFrac ("unweighted") uses **incidence**
- i.e. a taxon (leaf of the tree) is either present or absent in each sample
- UniFrac = 0 means the two samples have exactly the same list of taxa
- UniFrac = 1 means that no branches are shared – the two samples have mutually exclusive lists of taxa

Taxon is present in Sample 1

Taxon is present in Sample 2



$$\frac{\text{Total of } \textbf{not shared}}{\text{Total of } \textbf{not shared} + \textbf{shared}}$$

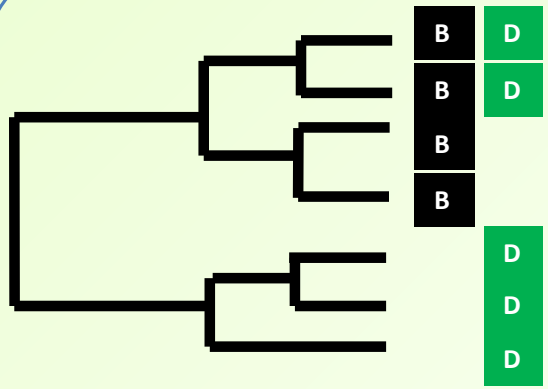
Samples →

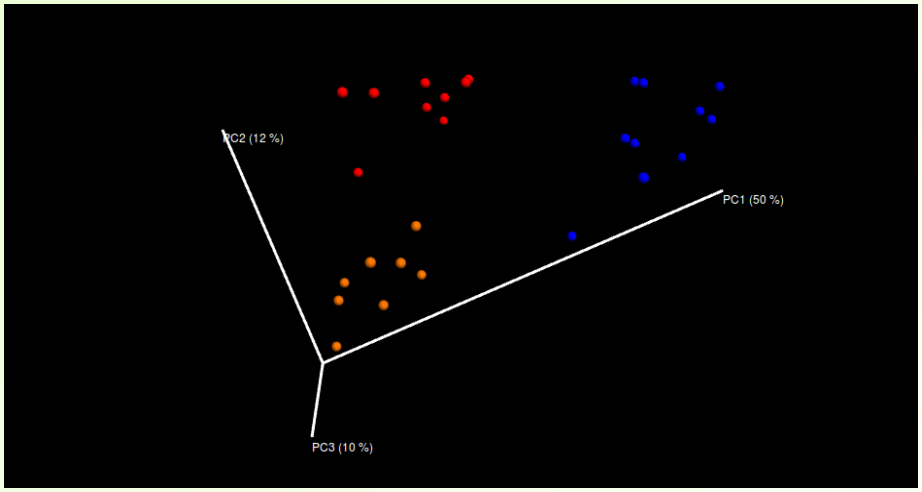
	A	B	C	D	E	F

OTUs/ taxa ↓

OTU frequencies in each sample…

phylogenetic relationships between taxa





PCoA

Samples →

	A	B	C	D	E
A					
B				*x*	
C					
D					
E					

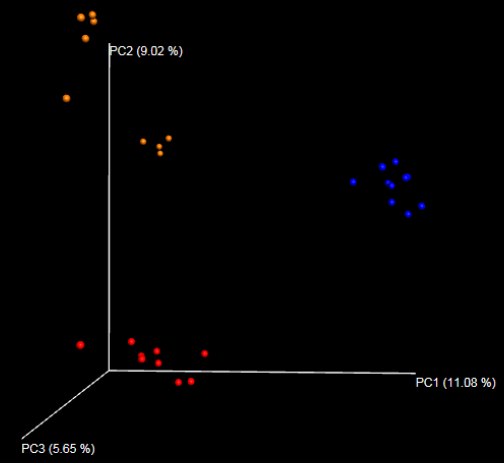
Samples ↓

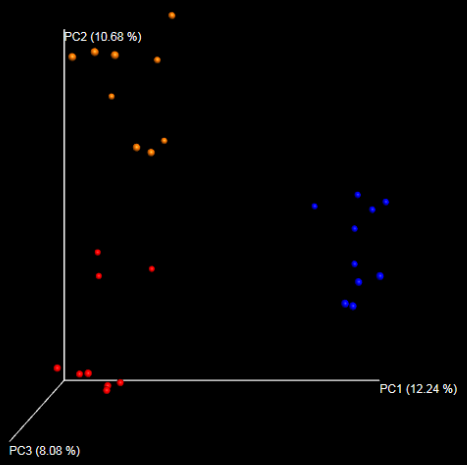
Inter-sample distances

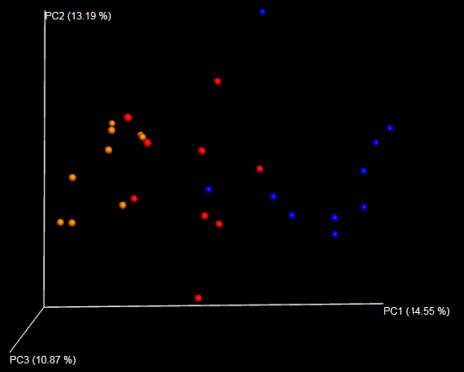
co-segregation over tree nodes
→ **distance metric**
(*e.g. weighted UniFrac*)

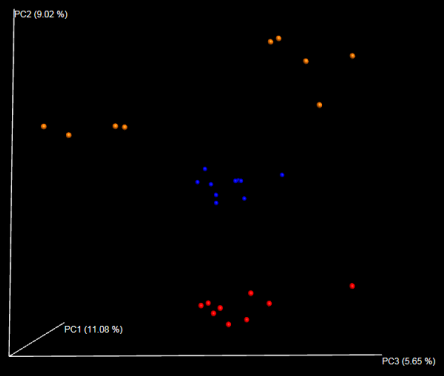
Weighted UniFrac

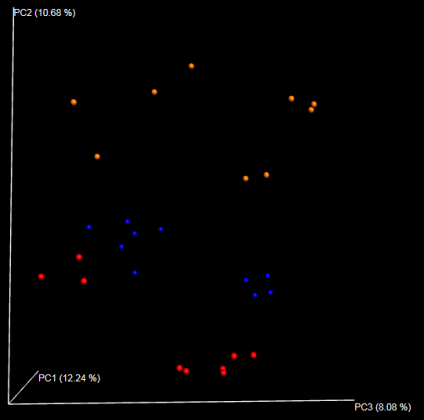
- Lozupone *et al.* (2007)
- Uses proportional abundance (instead of incidence)
- All the weighted branch lengths are summed
- Each branch is weighted by :
  - the **difference in proportional abundance** of the leaf (taxon) between the two samples
- Equal abundance → the branch is completely shared, as in unweighted UniFrac
- Unlike unweighted UniFrac, the maximum possible value is not 1.0, but determined by branch lengths
- If some taxa have evolved faster than others (longer branch lengths) – then these can lead to large values
- Normalisation: divide the result by the average distance of each **observation** (i.e. **sequence read**) from the root

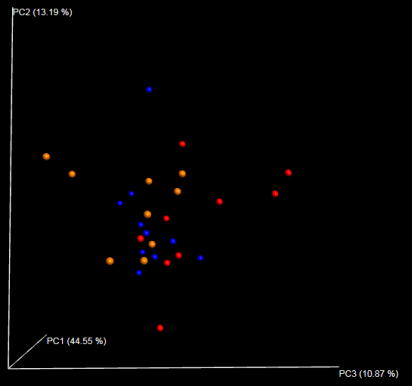

PC2 (9.02 %)
PC1 (11.08 %)
PC3 (5.65 %)


PC2 (10.68 %)
PC1 (12.24 %)
PC3 (8.08 %)


PC2 (13.19 %)
PC1 (14.55 %)
PC3 (10.87 %)


PC2 (9.02 %)
PC1 (11.08 %)
PC3 (5.65 %)


PC2 (10.68 %)
PC1 (12.24 %)
PC3 (8.08 %)


PC2 (13.19 %)
PC1 (44.55 %)
PC3 (10.87 %)

Why use PCoA on distance measures (such as UniFrac)?

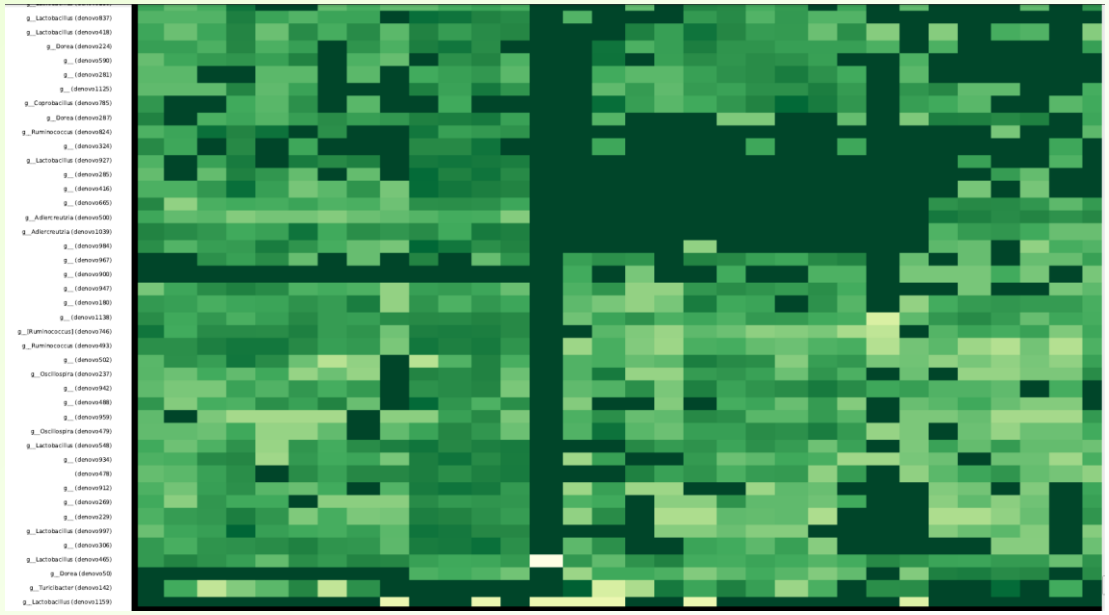
Compare and contrast with PCA:

Samples →

OTUs/ taxa ↓

	A	B	C	D	E	F

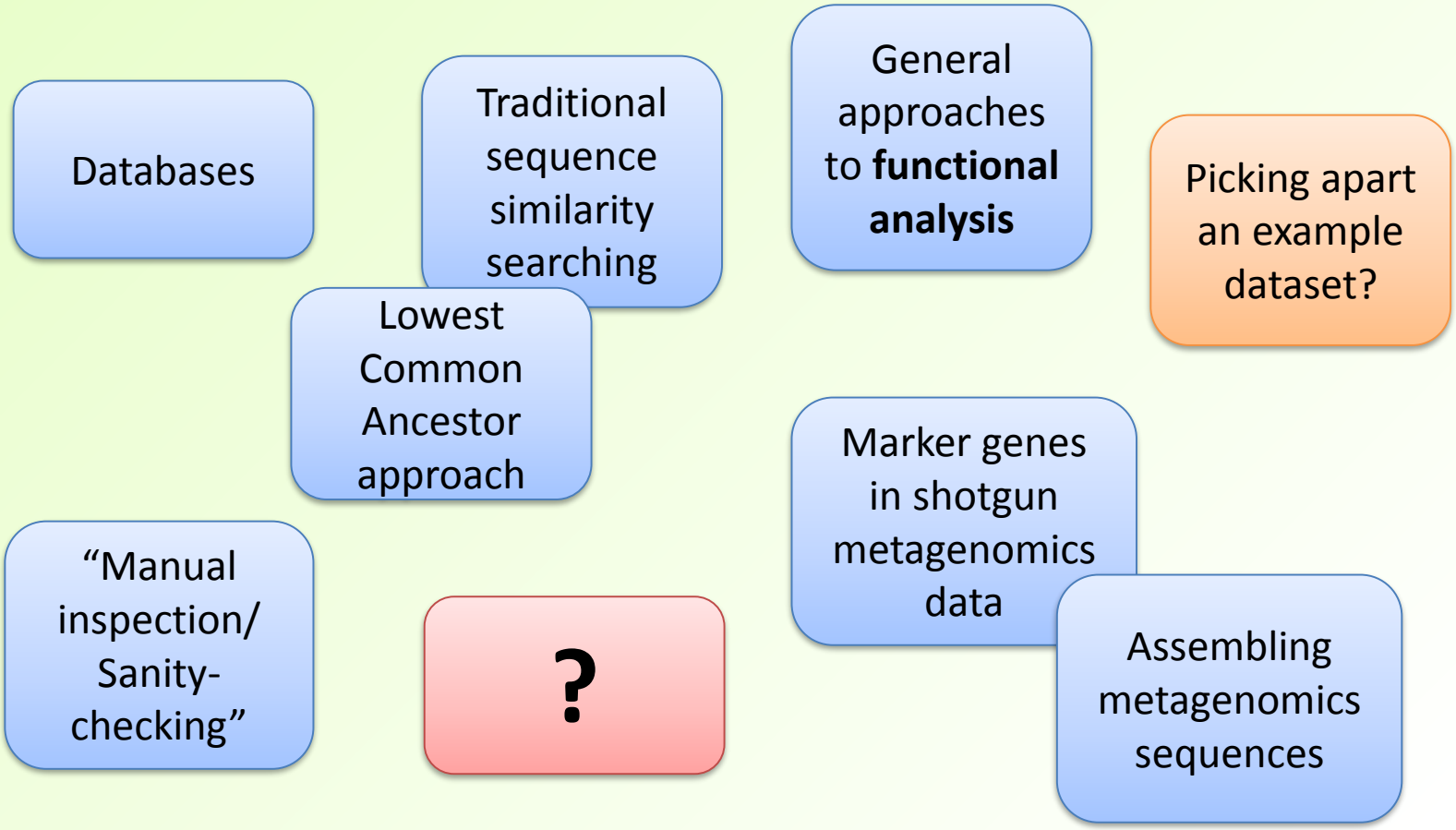
OTU frequencies in each sample…



References

- Faith (1992) Conservation evaluation and phylogenetic diversity, *Biological Conservation* **61,** 1-10
- Faith & Baker (2006) Phylogenetic diversity (PD) and biodiversity conservation: some bioinformatics challenges, *Evol. Bioinform. Online* **2**, 121-128
- Hill M.O. (1973) Diversity and evenness: A unifying notation and its consequences, *Ecology* **54:** 427-432
- Lozupone C.A. and Knight R. (2005) UniFrac: a New Phylogenetic Method for Comparing Microbial Communities, *Appl. Environ. Microbiol.* **71** (12): 8228-8235
- Lozupone C.A., Hamady M., Kelly S.T. and Knight R. (2007) Quantitative and Qualitative β Diversity Measures Lead to Different Insights into Factors That Structure Microbial Communities

What next?


Databases
Traditional sequence similarity searching
General approaches to **functional analysis**
Picking apart an example dataset?
Lowest Common Ancestor approach
Marker genes in shotgun metagenomics data
"Manual inspection/ Sanity-checking"
?
Assembling metagenomics sequences