Lies my instructor told me


silva
high quality ribosomal RNA database
elixir
Core Data Resource
de.NBI
GERMAN NETWORK FOR BIOINFORMATICS INFRASTRUCTURE
Home SILVAngs Browser Search ACT Download Documentation Projects FISH & Probes Jobs Contact
Database SSU r138.1
Taxonomy SILVA
Show Cart
Cart: 0
Show
Clear
Download
SILVA ▸ Bacteria ▸ Proteobacteria
SILVA
(3)
Archaea
Bacteria
Eukaryota
Bacteria
(89) (252)
Margulisbacteria
Marinimicrobia (SAR406 clade)
MAT-CR-M4-B07
MBNT15
Methylomirabilota
Myxococcota
NB1-j
Nitrospinota
Nitrospirota
NKB15
Patescibacteria
Planctomycetota
Proteobacteria
RCP2-54
Rs-K70 termite group
SAR324 clade(Marine group B)
Spirochaetota
Sva0485
Proteobacteria
(4) (285)
Alphaproteobacteria
Gammaproteobacteria
Magnetococcia
Zetaproteobacteria
New taxonomy
Old taxonomy

ICG INSTITUTE FOR COMPARATIVE GENOMICS

DALHOUSIE UNIVERSITY

Diversity, statistics and data visualization

Rob Beiko and Diana Haider

(with thanks to Jacob Nearing and others)

Dalhousie University

Learning Objectives

- By the end of this lecture, you will be able to:
  - **Distinguish** key types and classes of diversity
  - **Describe** compositionality and why it's important
  - **Recognize** different types of statistical analysis

The Plan

(1) Diversity analysis (~20 minutes)
(2) Statistics (~20 minutes)
(3) Machine learning etc., if time remains

Diversity analysis

Communities

- **Community** – the group of things (species, etc.) that occupy the same location at the same time

- Note that community interactions are hypotheses – don't assume that everyone in a given location is talking to everyone else

Richness and Diversity

- **Richness** – The count of “things”

$$R$$

- **Diversity** – The count of “things” with some consideration of evenness

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

Proportion of “thing” *i*

The richness and diversity of *what* exactly?

- Species!
- Genera!
- Phyla!
- OTUs!
- ASVs!
- Unique sequences!
- Functional genes!

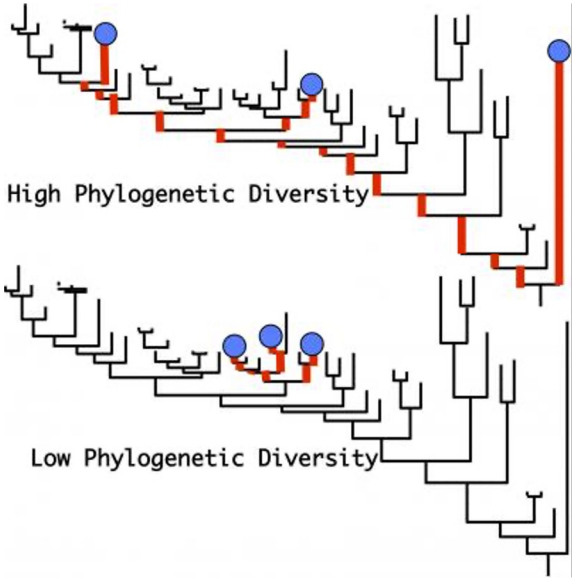
Not all approaches make sense for all types of “things”

Criteria that differentiate diversity measures

- **Observed** vs **estimated**
- **Non-phylogenetic** vs **phylogenetic**
- **Unweighted** vs **weighted by abundance**

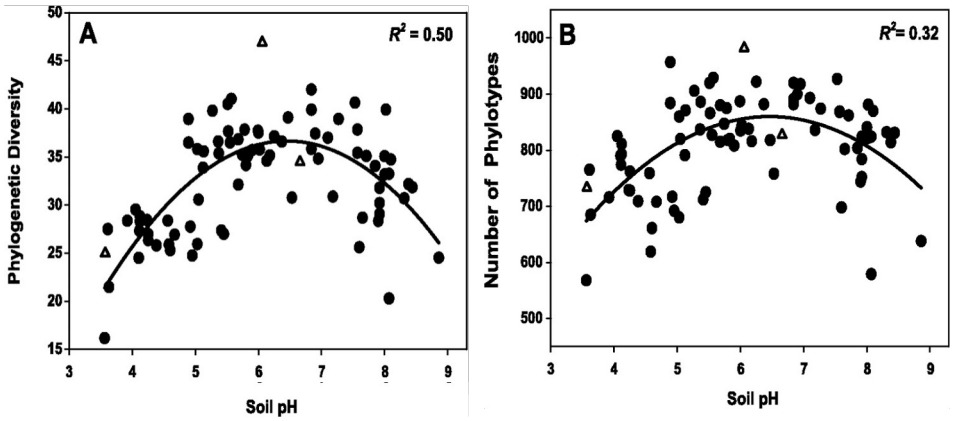
“Alpha” diversity: diversity at a single site

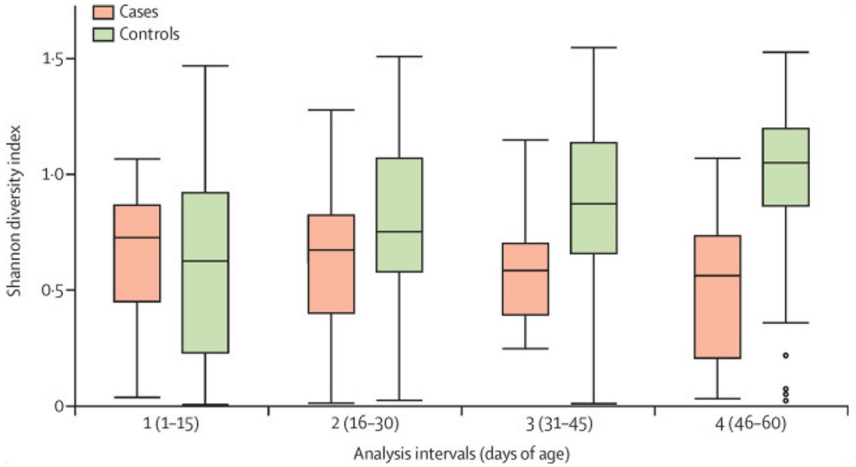
	Richness	Diversity
Non-phylogenetic	Species count	Shannon index
Phylogenetic	Unweighted phylogenetic “diversity”	Abundance-weighted phylogenetic diversity



https://www.technology.org/2014/07/18/scientists-enlist-big-data-guide-conservation-efforts/

What can we do with alpha diversity?





Very low birth weight (< 1500 g) vs control babies

Lauber et al. (2009) *Applied and Environmental Microbiology*

Warner et al. (2016) *The Lancet*

“Beta” diversity: diversity between sites

- Compare the communities at two or more sites (typically pairwise)
- **More-dissimilar communities = greater beta-diversity**
  - identical communities should have beta diversity of 0
  - similarity is often = (1 – diversity)
- Similar criteria as alpha-diversity measures: phylogenetic vs non-phylogenetic, weighted vs non-weighted

Jaccard similarity: unweighted, non-phylogenetic

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Where

A = community #1
B = community #2
|X| = number of taxa in a given set
∩ = Intersection
U = Union

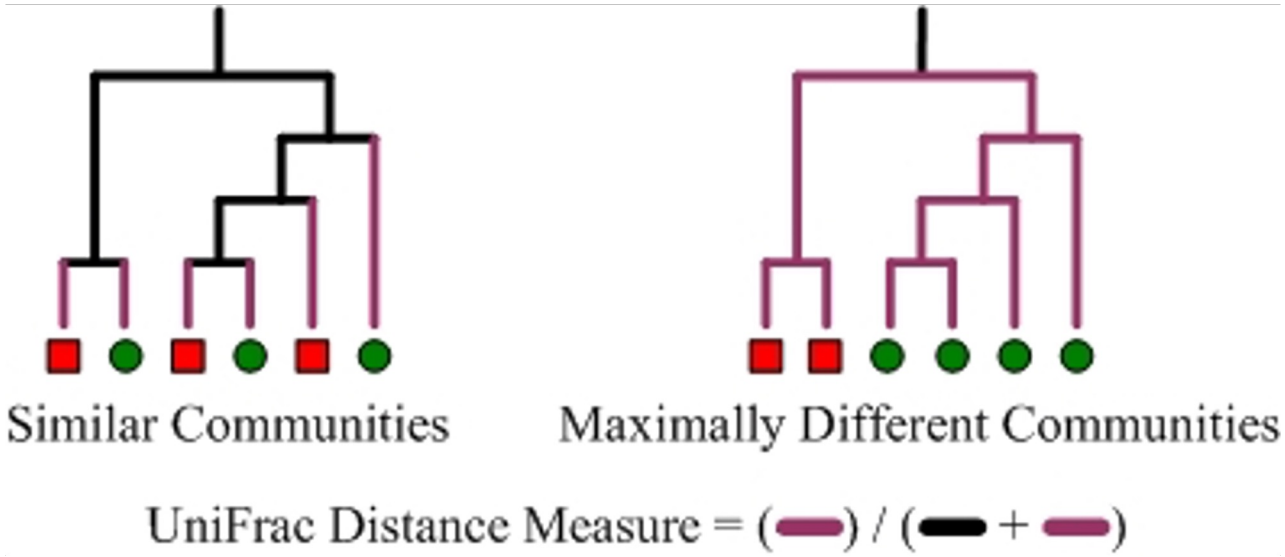
Bray-Curtis Dissimilarity: weighted, non-phylogenetic

$$BC_{ij} = 1 - \frac{2C_{ij}}{S_i + S_j}$$

Where

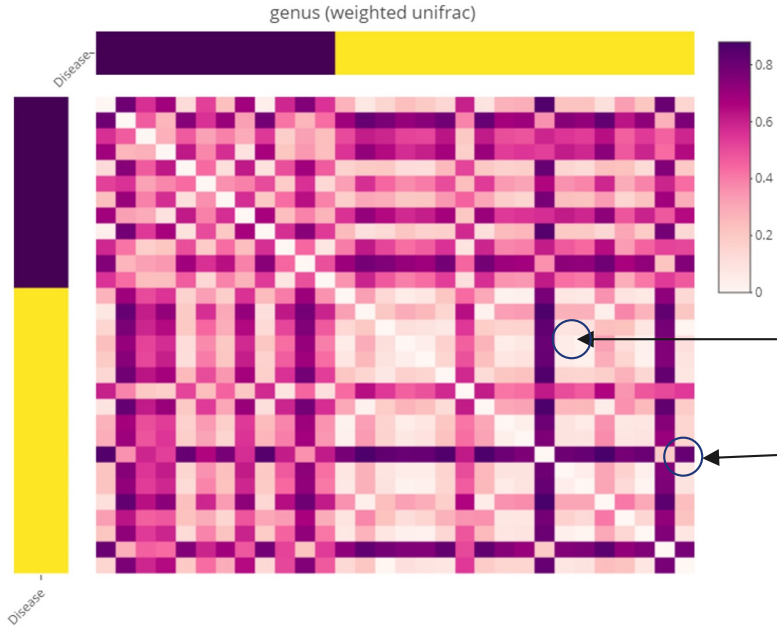
$BC_{i,j}$ = Bray-Curtis
S_i / S_j = # of “specimens” from sites *i* and *j*
$C_{i,j}$ = Smaller count of each species from either *i* or *j*

UniFrac: phylogenetic, weighted or unweighted



https://mothur.org/wiki/unweighted_unifrac_algorithm/

Beta-diversity matrices



- Show similarity or distance between each pair of communities

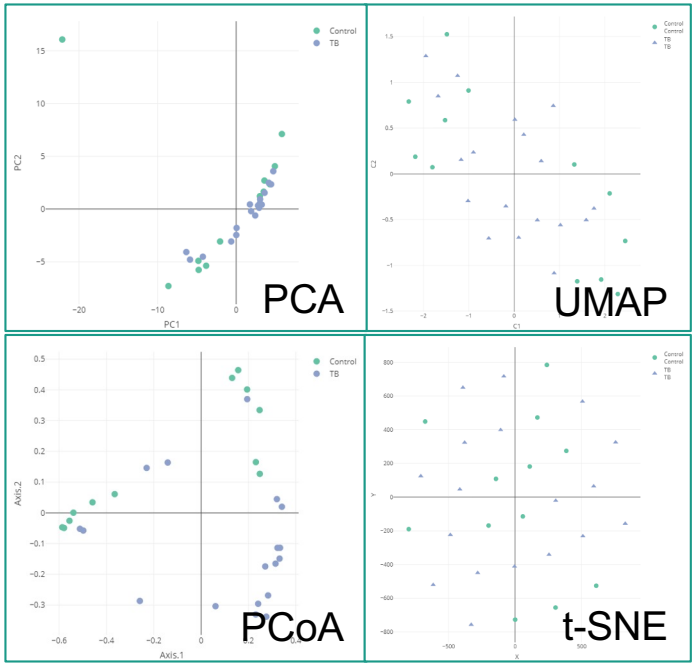
These two communities are SIMILAR to one another

These two communities are highly DISSIMILAR

Visualizations made with Animalcules (https://github.com/compbiomed/animalcules)

What can we do with beta diversity?

- Dimensionality reduction – ordination
- Many different ways to do this:
  - Euclidean (PCA)
  - Custom distances (PCoA, NMDS)
  - Global / Local Structure preservation (UMAP, t-SNE)



Visualizations made with Animalcules (https://github.com/compbiomed/animalcules)

What can we do with beta diversity?

- Hierarchical clustering

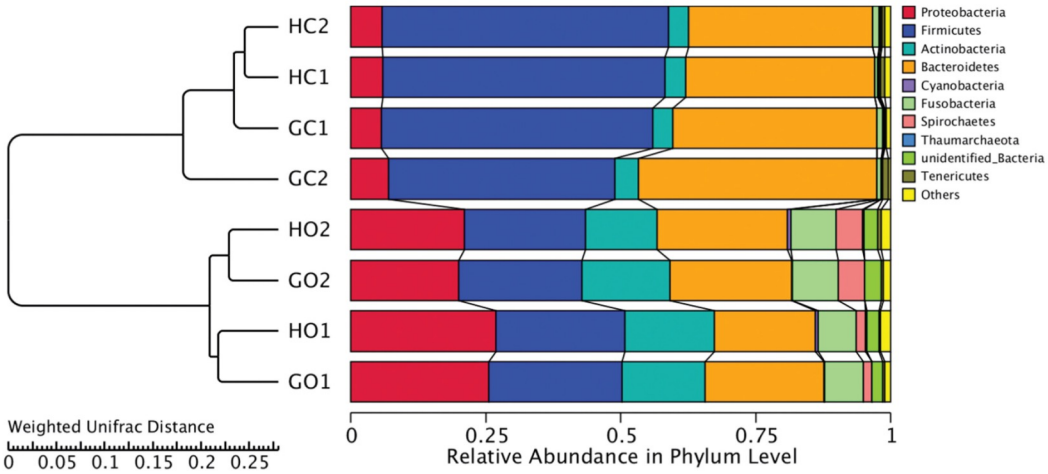
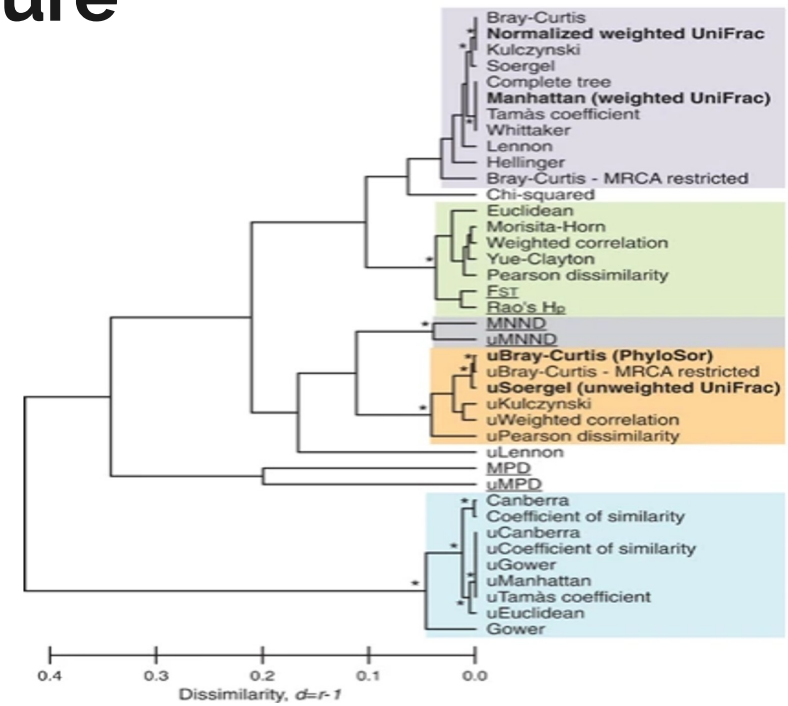


Figure 3. UPGMA analysis based on weighted UniFrac distances with the result of the clustering tree shown on the left and the distribution diagram of the top 10 phylum abundances shown on the right. GC1: GDM. intestinal, GC2: periodontitis+GDM. intestinal, HC1: healthy control. intestinal, HC2: periodontitis. intestinal, GO1: GDM. oral, GO2: periodontitis+GDM. oral, HO1: healthy control. oral, HO2: periodontitis. oral. Comparisons of community structures among groups were performed using AMOVA analysis.

Zhang et al. (2021) *Journal of Oral Microbiology*

Choosing a diversity measure

- Some seemingly different measures tend to give similar results (e.g., Bray-Curtis, weighted UniFrac)
- No single measure is best in all circumstances
- So choose a couple that are *really* different from each other



Parks and Beiko (2013) *ISME Journal*

Statistics and Machine Learning

disclaimer

we will almost certainly

not

make it through

all of these slides

but that is ok

Do you have a hypothesis?

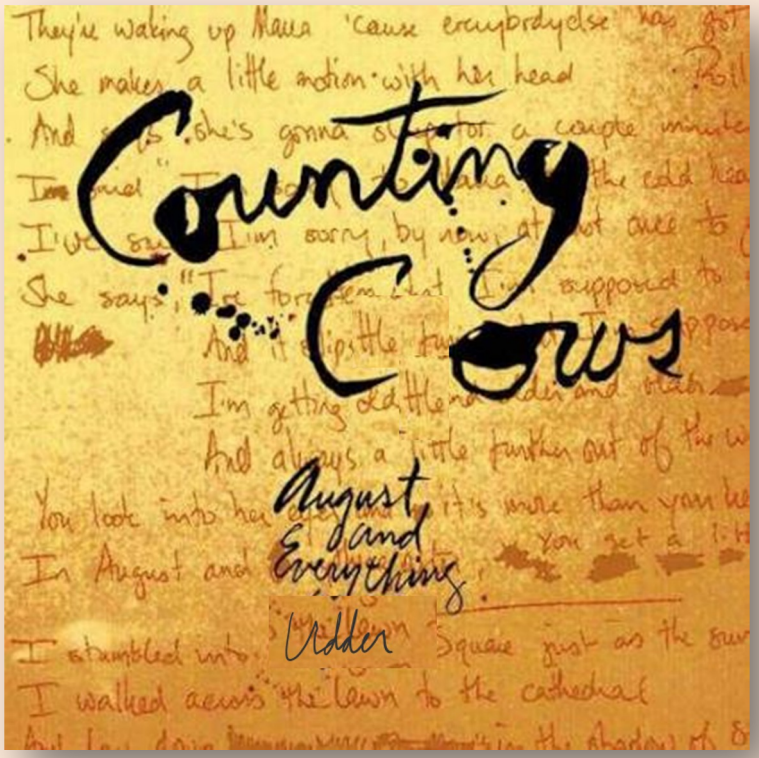
- Yes!
  - What are the predictions?
  - How do we test them?

- No!
  - FINE.
  - Well, we can still explore the data

Fun things that break statistics

- Weird distribution of observations
  - lots of 0s
  - *So many 0s*
  - Different *kinds* of 0s!
- Proportions rather than counts: non-independence
- Hierarchies: functional, phylogenetic, taxonomic

Compositionality

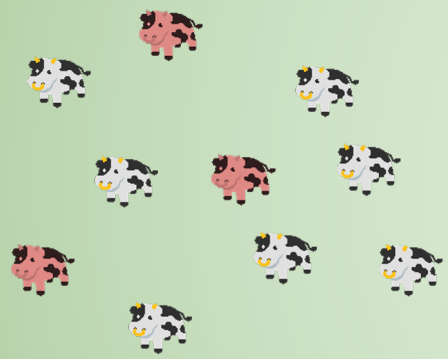


N = 10

N = 30

Same **count**, different **proportion**

N = 10

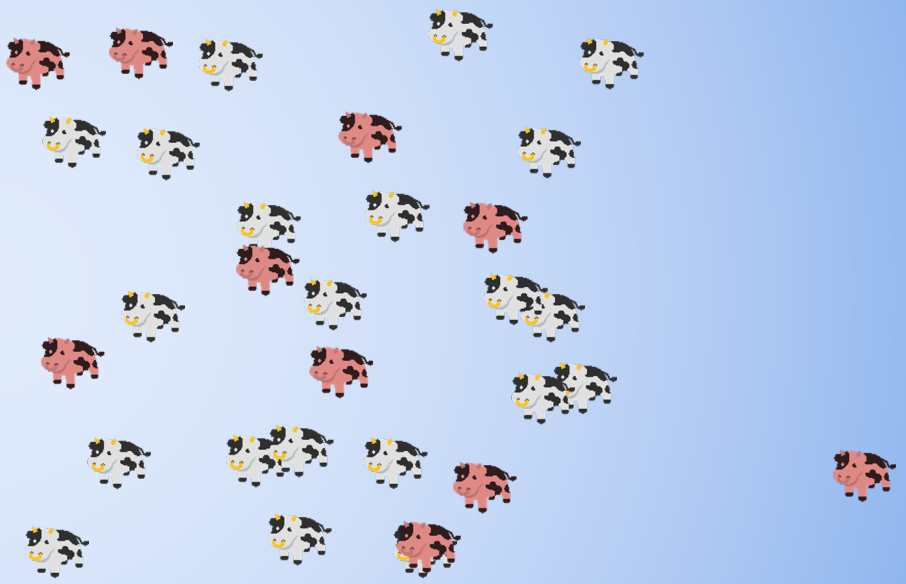


N = 30

Different **count**, same **proportion**

N = 10

N = ~~30~~ 10

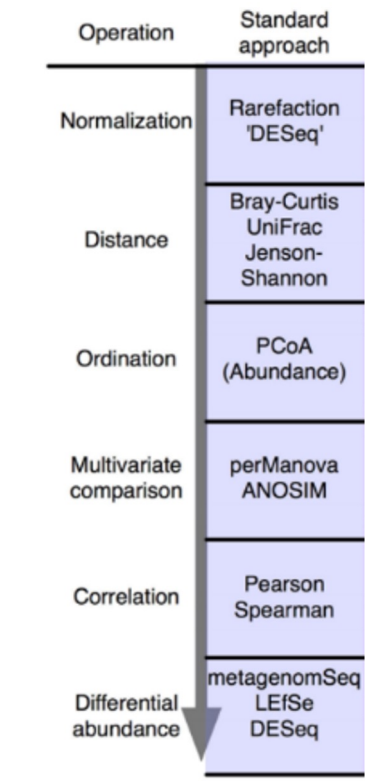


If we can only count ten instances, then we don't know!

This is what we get when we sequence DNA or RNA!!

- Our “counts” are (almost) always limited by the capacity of the sequencer, except in cases where there is vanishingly little diversity (e.g. most placental samples)
- Variation in read count does not reflect variation in cells / 16S genes
- There are ways to assess absolute counts (as mentioned by Corinne); these can give context to the count data.

So what?



Discarding of information (rare taxa → 0)

Outsized influence of most abundant (but possibly boring) taxa

Spurious, unstable correlations

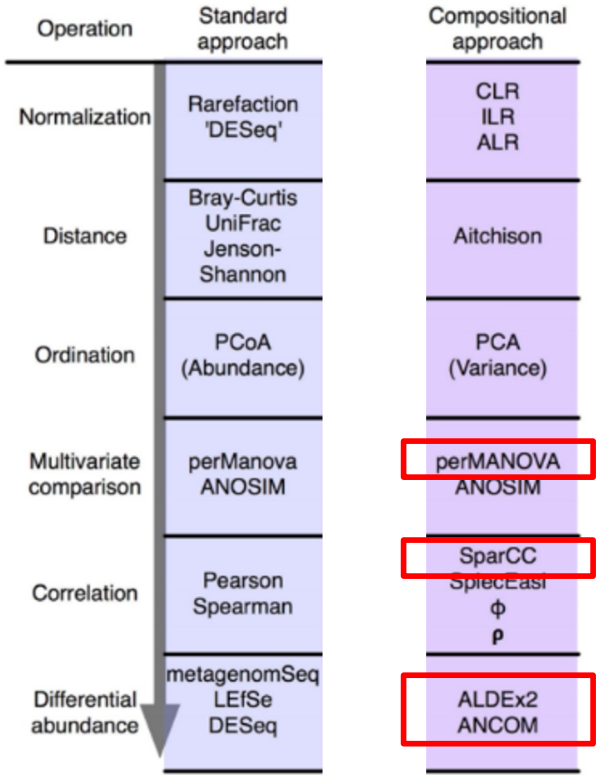
Incorrect estimates of variance

Gloor et al. *Front Microbiol*, 2017 (PMID 29187837)

That being said…

- The field is adopting techniques that account for compositionality
  - Rarefaction is still in widespread use but alternatives exist
  - Distances are still mostly B-C, UniFrac, …
  - Correlation analysis is a bit more “with it”
- The impact of biases in “Standard” approaches heavily depends on properties such as diversity and correlation structure (Friedman and Alm, 2012)
- Best practices are evolving – the key is to be aware of the key assumptions and pitfalls

Let's talk about...



Gloor et al. *Front Microbiol*, 2017 (PMID 29187837)

Testing for significant differences between and among groups

	2 categories	>2 categories
Parametric	T-test	ANOVA
Non-parametric (rank)	Mann-Whitney U	Kruskal-Wallace
Permutation	Permutation t-test	perMANOVA

Basic principle: are differences **between groups** significantly greater than differences **within groups?**

ANOVA

- Parametric!
  - A good thing
  - Also a bad thing
- Is the sum of squared differences *between* groups significantly larger than the sum of squared differences *within* groups?
- ANOVA can tell you *if* a difference exists, but not *where* - post-hoc tests required!!
- MANOVA for multivariate responses

Kruskal-Wallace

- The nonparametric answer to ANOVA
- Turn your observations into **ranked data**
- Do the medians of different groups differ significantly?
  - In other works, can my ranks beat up your ranks?
- If the result is significant, you again need to run post hoc tests to find out *where* the significant differences lie

perMANOVA

- Use permutations to simulate a null distribution
- perMANOVA looks for the difference between *centroids* of some dissimilarity measure; this can be anything from means to Bray-Curtis or what have you
- Can accommodate fancy experimental designs

Differential abundance

- What features (ASVs, OTUs, species, pathways, etc) are different between two or more samples?
- Useful for identifying “good guys” or “bad guys”, key functional genes, biomarkers
- Maybe we have a specific hypothesis (test only one thing!) and maybe we don’t (test a bunch of things!)

Log-ratio transformations: Breaking down the compositional wall

- Divide the values in each sample vector by some quantity, and take the logarithm

- Divide by what?
  - Some magic invariant feature (?): the *additive* log ratio
  - The geometric mean: *centred* log ratio

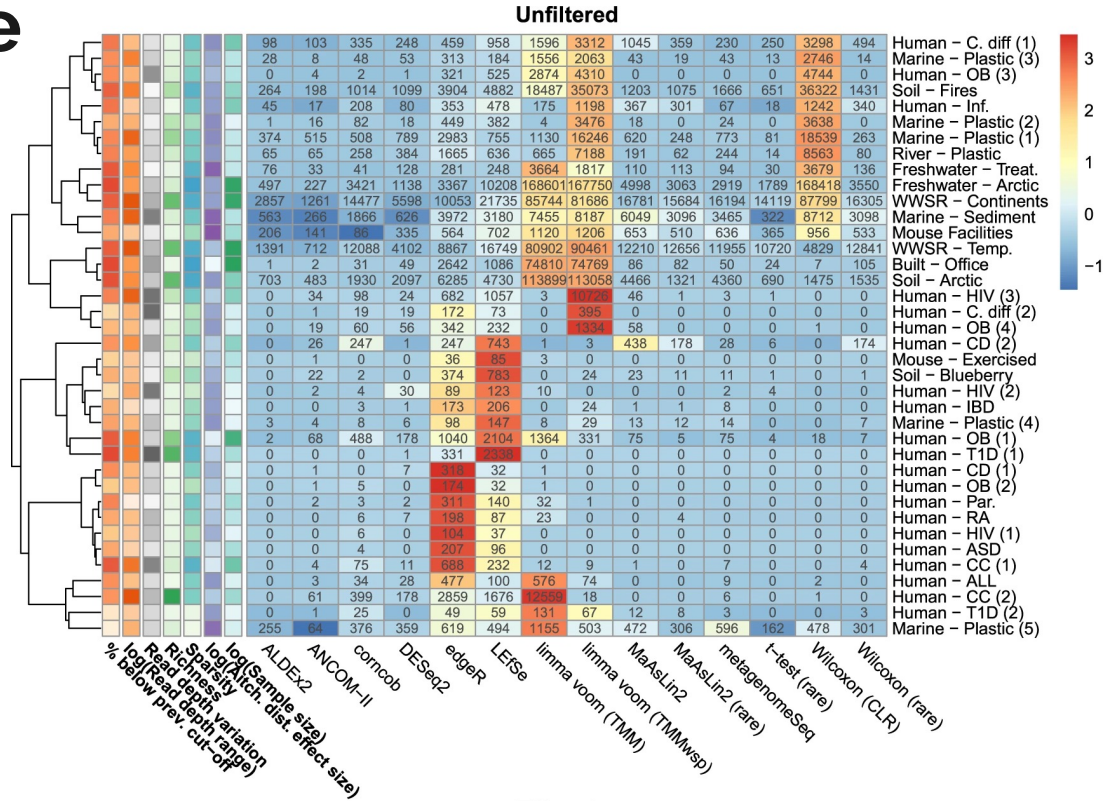
ALDEx2: accounting for compositionality

1. Take counts; add a “uniform” prior (in this case, 0.5) which avoids the awkwardness of log(0)
2. Sample counts many times to generate probabilities: samples with few counts will have higher variances
3. CLR transform!!!
4. Significance tests: Welch’s *t*, Wilcoxon rank
5. Correct for multiple tests!!

Fernandes et al. *Microbiome*, 2014 (PMID 24910773)

The Unfortunate Truth

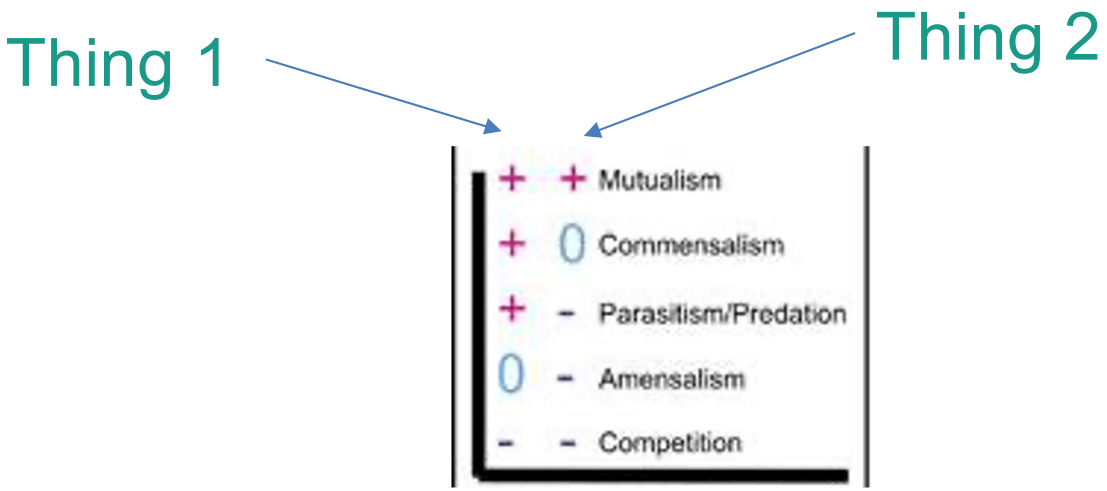
- Many of these methods will identify a different number of significant taxa within your sample



Nearing et al. (2022) *Nat Comms*

Correlations

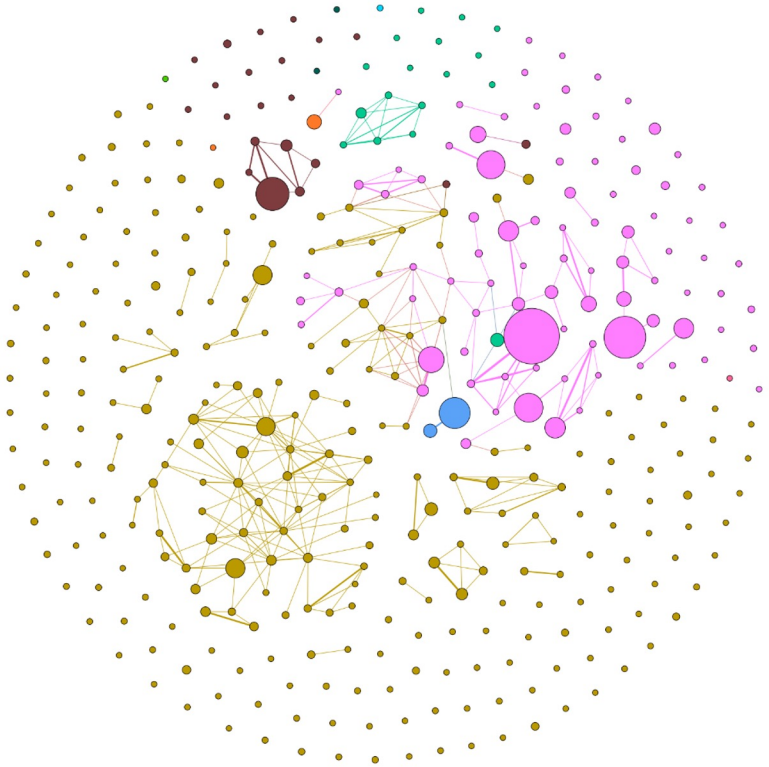
Goal: Infer different types of ecological interaction by examining shared abundance patterns among taxa in all samples in a study



Weiss et al. (2016) *ISME J*

Correlation networks

Compute correlations between all pairs of entities (e.g., OTUs or ASVs), then **threshold** by test statistic or p-value to build a network



Edges in the network

To find the Pearson correlation between taxa or functions X and Y, line up the corresponding samples and apply this formula:

$$r = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \pi r^2$$

Covariance

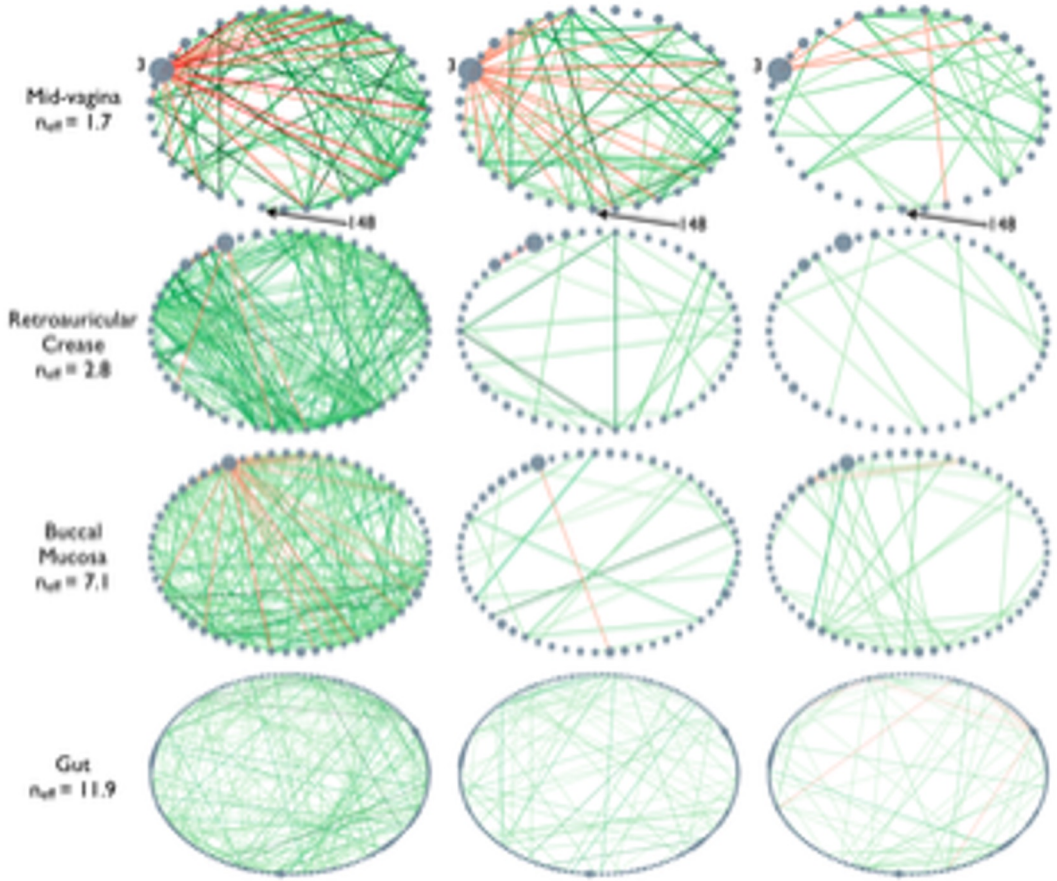
Standard deviation of X and Y

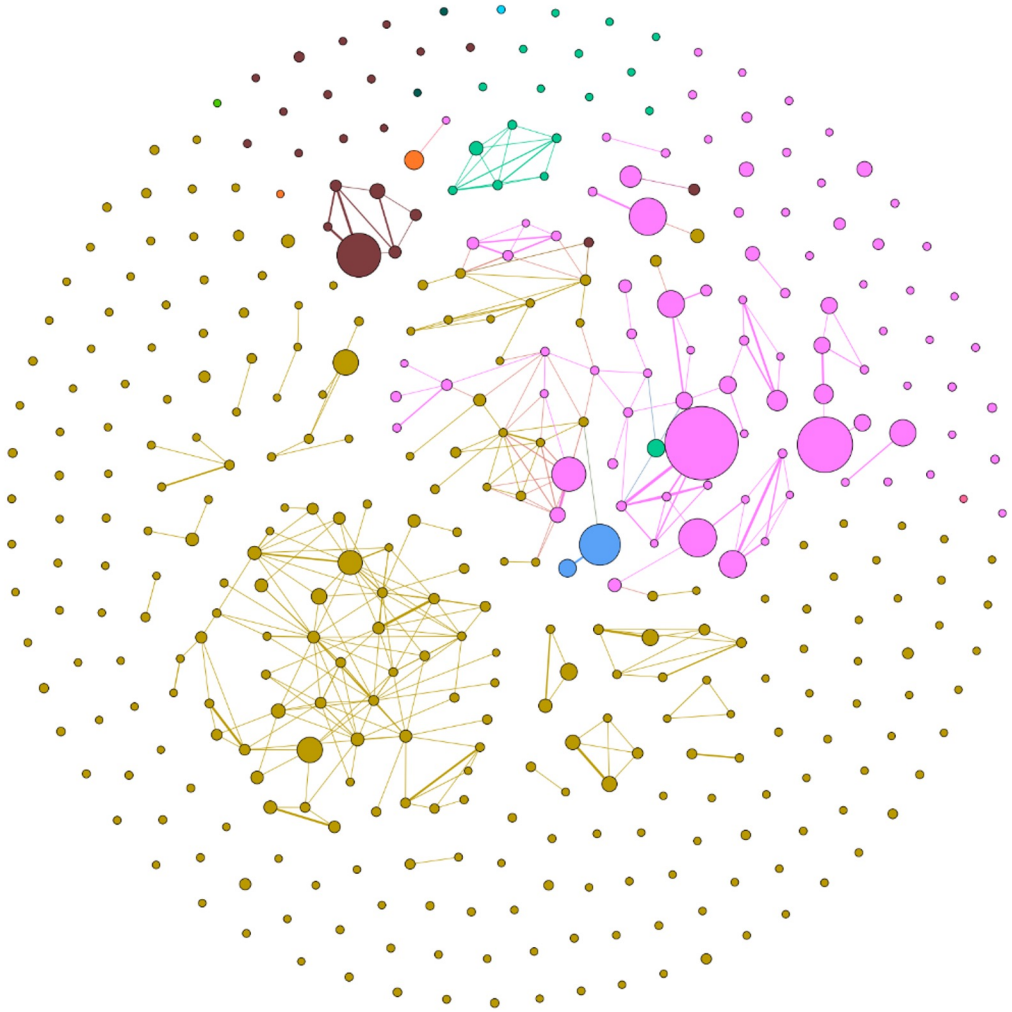
Assumes bivariate normality, sensitive to outliers
Spearman: similar test, applied to **ranks**

Sparse Correlations for Compositional Data (SparCC)

- Key principles: there are lots of features, but relatively few correlations
- Aitchinson's test: are there any dependencies?
- Statistical significance is based on simulation of many variables with **no correlation**.

Friedman and Alm (2012) *PLoS Comp Biol* (PMID 23028285)


Mid-vagina
n_{ott} = 1.7
Retroauricular Crease
n_{ott} = 2.8
Buccal Mucosa
n_{ott} = 7.1
Gut
n_{ott} = 11.9
Pearson (real data)
Pearson (fake data)
SparCC



p__Firmicutes	(61.96%)
p__Bacteroidetes	(15.15%)
p__	(13.7%)
p__Proteobacteria	(3.91%)
p__Actinobacteria	(3.28%)
p__Fusobacteria	(0.68%)
p__Verrucomicrobia	(0.34%)
p__Lentisphaerae	(0.3%)
p__Cyanobacteria	(0.26%)
p__Tenericutes	(0.26%)
p__Synergistetes	(0.13%)
p__Prot	(0.04%)

Correlation network of OTUs, size proportional to abundance

Machine Learning

Machine Learning – the leap (?)

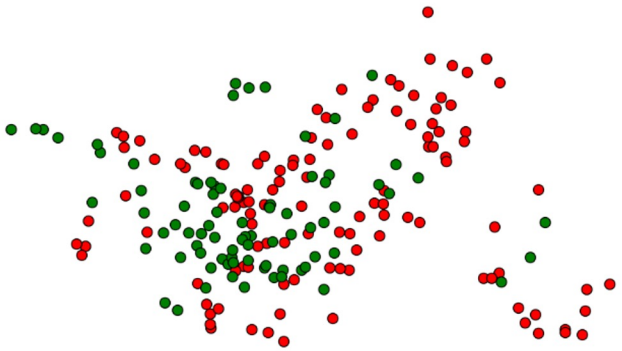
- Is there a difference between statistics and machine learning? apart from terminology
- Does statistics have a monopoly on probability density functions? (no)
- Is iterative training exclusive to machine learning? (no)
- Is machine learning alone concerned with predictive accuracy? (no)

Why use machine learning then?

- Its models generally have more free parameters to tweak – can “tailor” the predictor to different attributes of the data set
  - But watch for overfitting!
  - And models you can’t understand!

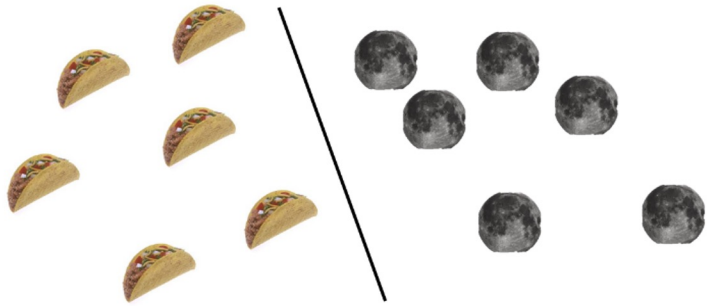
- Different methods perform well on different types of data
  - TAANSTAFL (no method wins in all cases)

Unsupervised - Clustering and Correlation



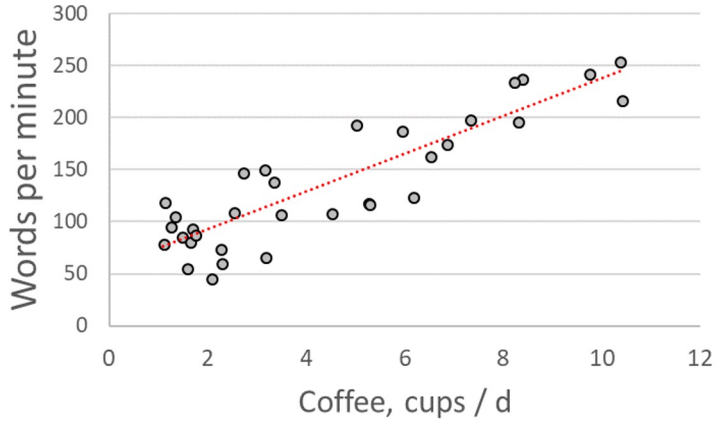
Supervised – Classification

Distinguish 2 or more discrete classes based on underlying features



Supervised – Regression

Predict quantitative values based on 1 or more features



Key things to keep in mind

- Holy frig there are a lot of classifiers

- Key attributes to think about:
  - Bias (too simple) vs variance (absurd number of parameters)
  - Do you care about interpretability?
  - Do you want training to finish this decade?
  - Does anything about the problem suggest a particular choice of classifier?

Generalization

A sufficiently complex classifier can learn pretty much anything in your training set

the ability of the machine to learn any training set without error. A machine with too much capacity is like a botanist with a photographic memory who, when presented with a new tree, concludes that it is not a tree because it has a different number of leaves from anything she has seen before; a machine with too little capacity is like the botanist's lazy brother, who declares that if it's green, it's a tree. Neither can generalize well. The exploration and

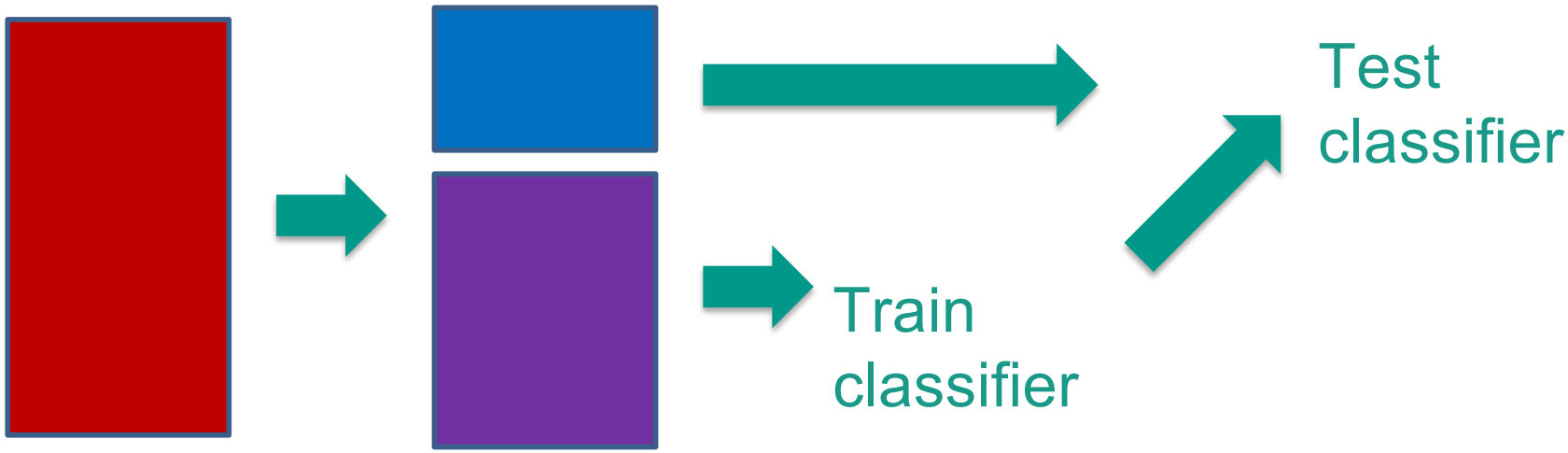
A tree

Not a tree

So you need to **test** your classifier on new data

Data set splitting (*holdout* method)

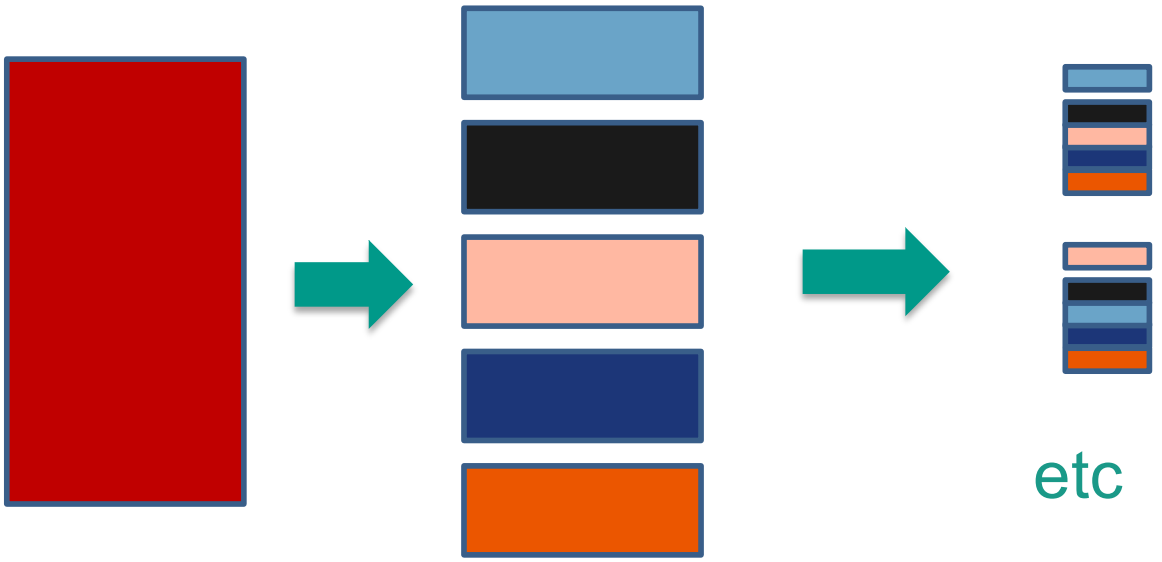
Use a fraction of available cases as the *training set*, reserve the remainder for a *test* set



Cross-validation

5-fold cross-validation

Repeated training with different subsets



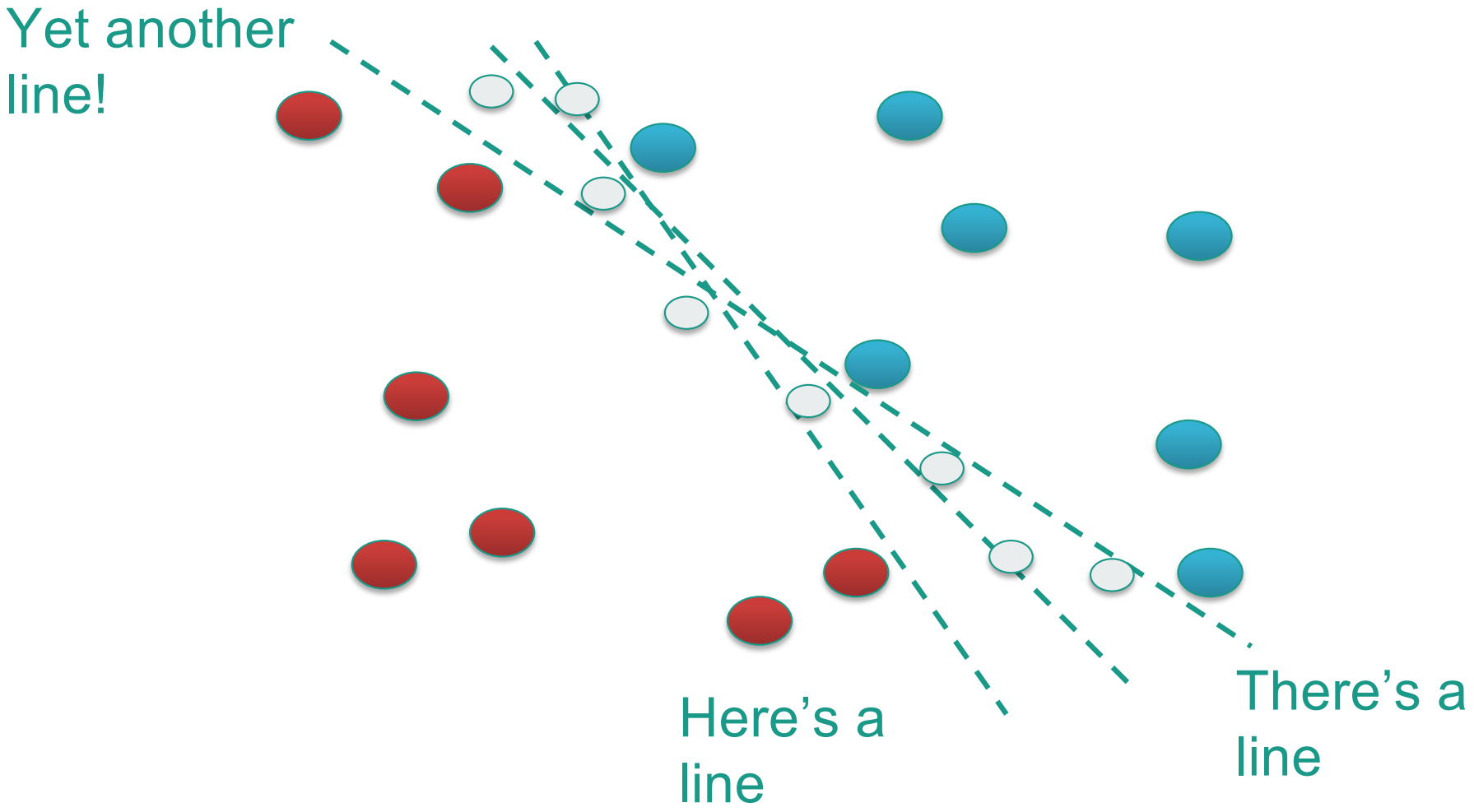
The *cross-validation score* is the average performance on all test sets

So let's pick a classifier.

- **Support vector machines** are based on a simple principle: try to fit a model that gives the best chance of generalizing well

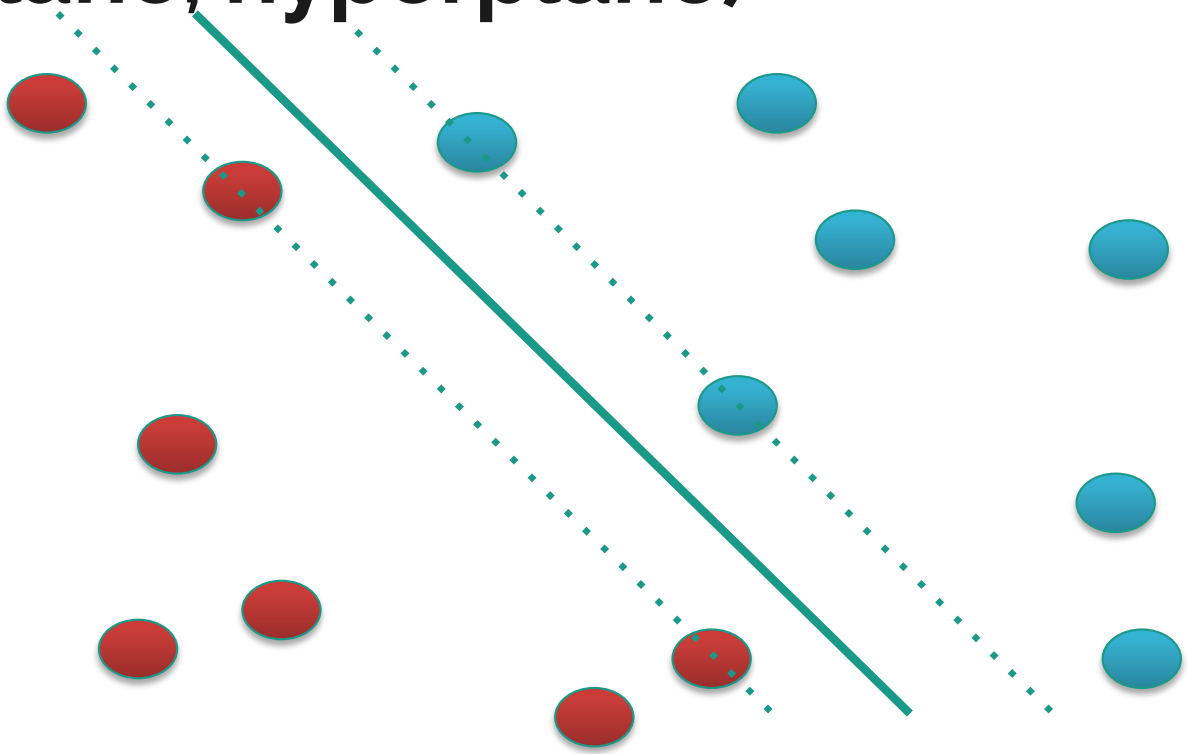
- Let's start with a simple example: a **linearly separable** data set

A two-dimensional, *linearly separable* problem



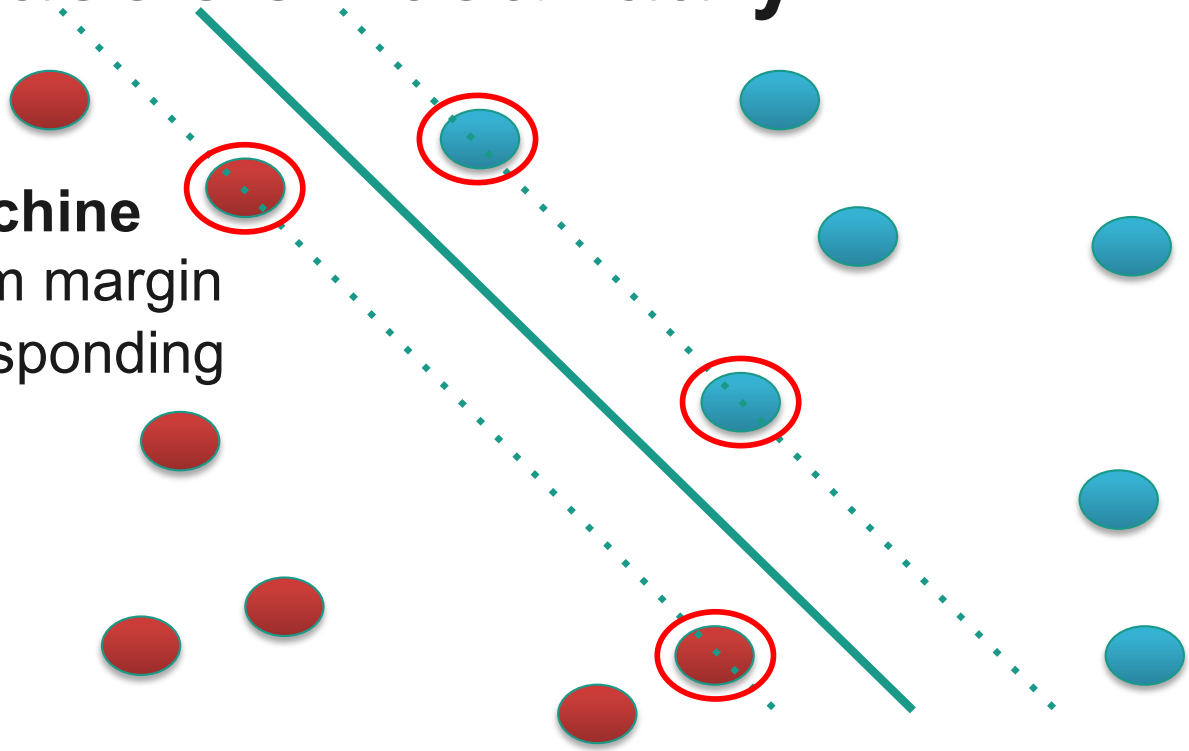
Define a maximum margin line (plane, hyperplane)

The **maximum margin hyperplane** separating two groups provides the optimal tradeoff between training set accuracy and function complexity



Only the SUPPORT VECTORS define the decision boundary

The **support vector machine** aims to find the maximum margin hyperplane and its corresponding support vectors

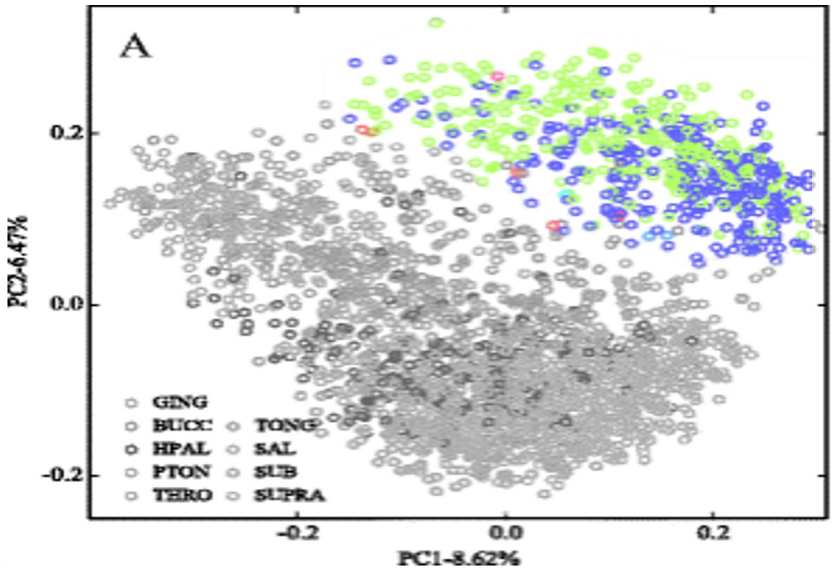


Things to know about SVMs

- They can actually handle non-separable problems
- Data not linearly separable? The **kernel trick** can be used to transform your data into other spaces (e.g., polynomial)
  - Kernels can be biologically inspired, which is cool
  - We tried UniFrac, which unfortunately sucks
- Training is iterative – can take a while
- No easy way to interpret the model (black box)

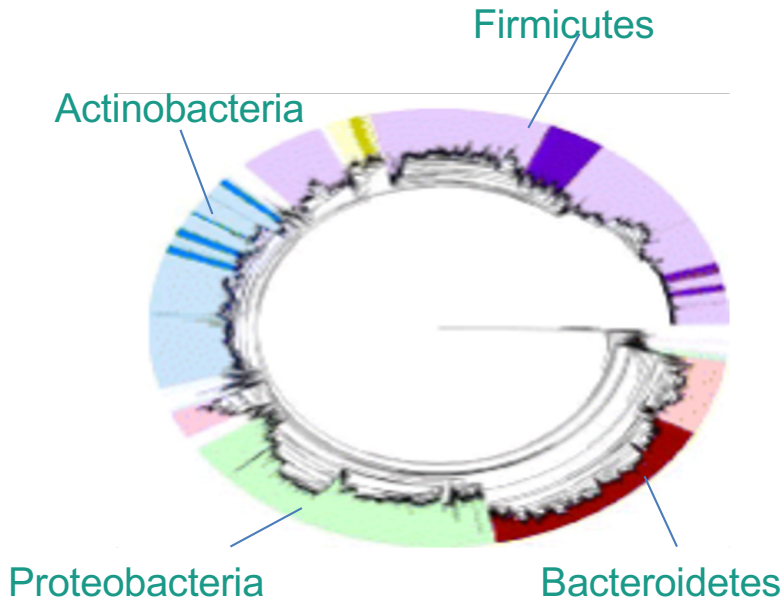
A fun example: classifying HMP plaque samples

About 300 samples each of supragingival and subgingival plaque



Data encoding

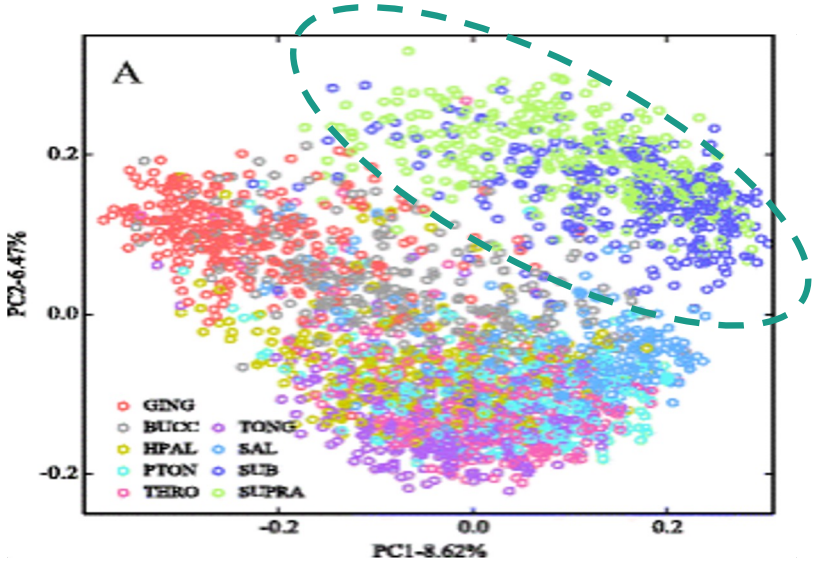
1. OTUs ☹
2. Phylogenetic trees
3. Predicted functions using PICRUSt



Too many features to begin with – use **feature selection** to narrow things down

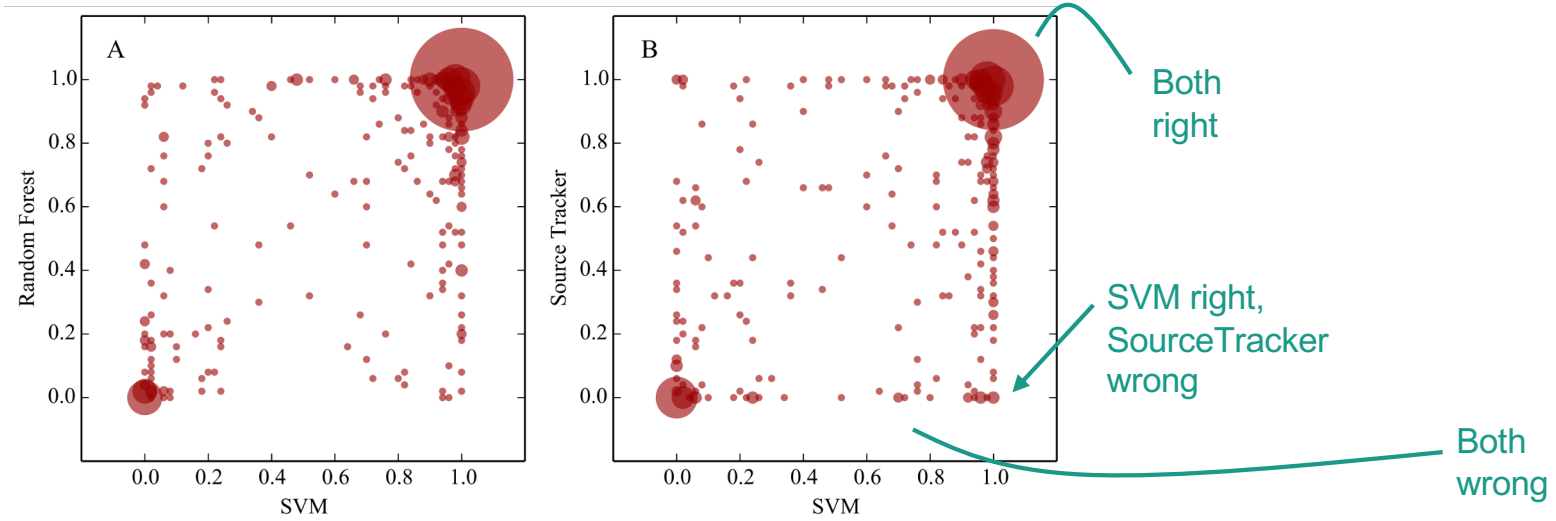
Accuracy

- OTU: 77-80%
- Clade: 80-81% (73.8% without feature selection!!)
- PICRUSt functions: 75-76%



What is the limit of classification?

- Probably not 100%
- Other classifiers get about the same accuracy, but on **different** subsets of samples
- 10% of samples appear to be hopeless



Summary

- Microbial community data make life difficult
  - Compositional
  - Lots of 0s
  - Biased data recovery
  - Hierarchical

- **All** methods have limitations you should be aware of
  - It's easy to get false positives!!


End of
Part III