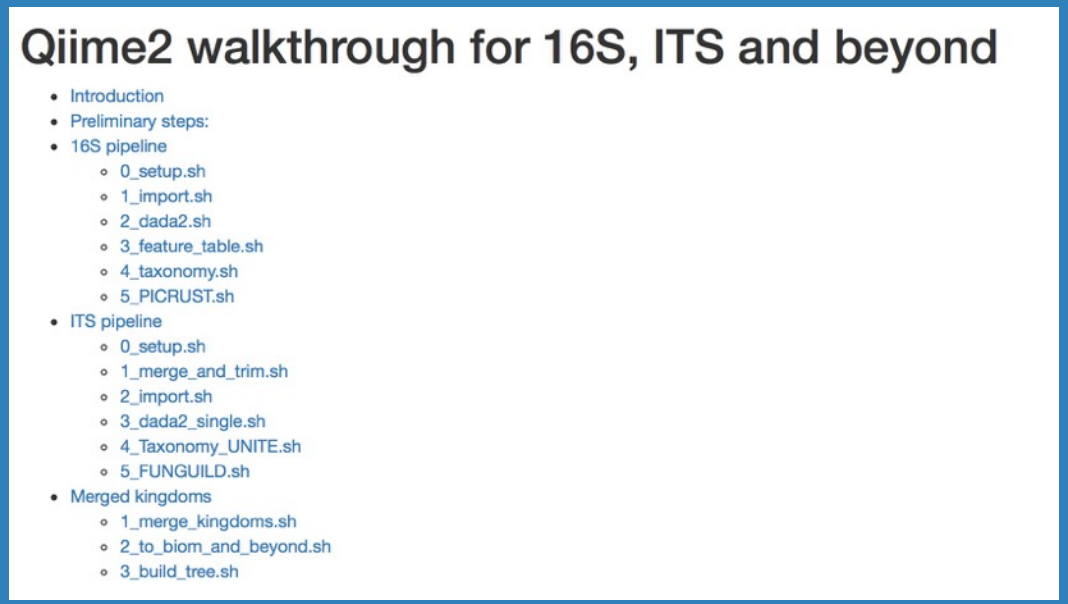
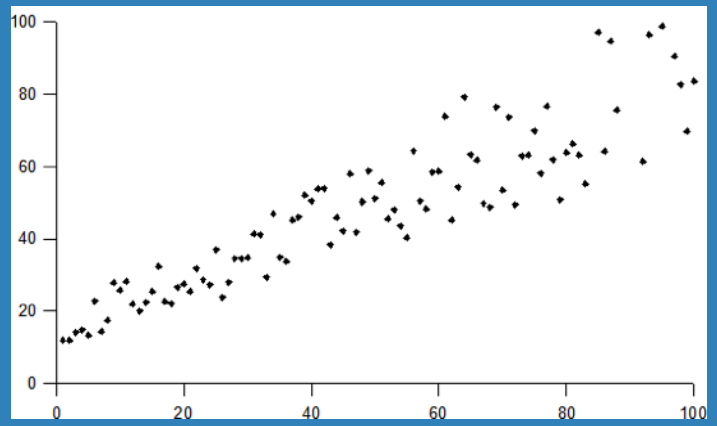
WHERE WE LEFT OFF

- 16S and ITS rRNA gene sequences (V4 region and ITS1, region respectively)
  - Joined paired ends
  - Filtered and trimmed low-quality base calls (hopefully)
  - Clustered into Amplicon Sequence Variants
  - Affiliated those sequences to their respective taxonomies
  - Predicted metabolic pathways/guilds
  - Merged the two kingdoms and produced the following:
    - A table file (without Chloroplasts/Mitochondria)
    - A normalized table file
    - A taxonomy file
    - A tree file

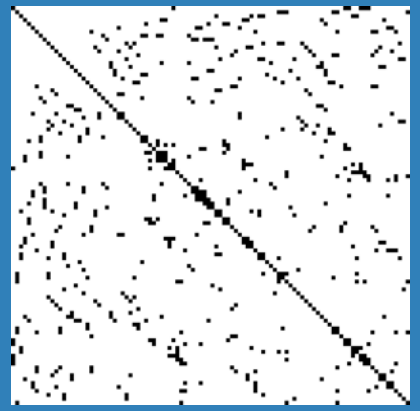


NEW CHALLENGES

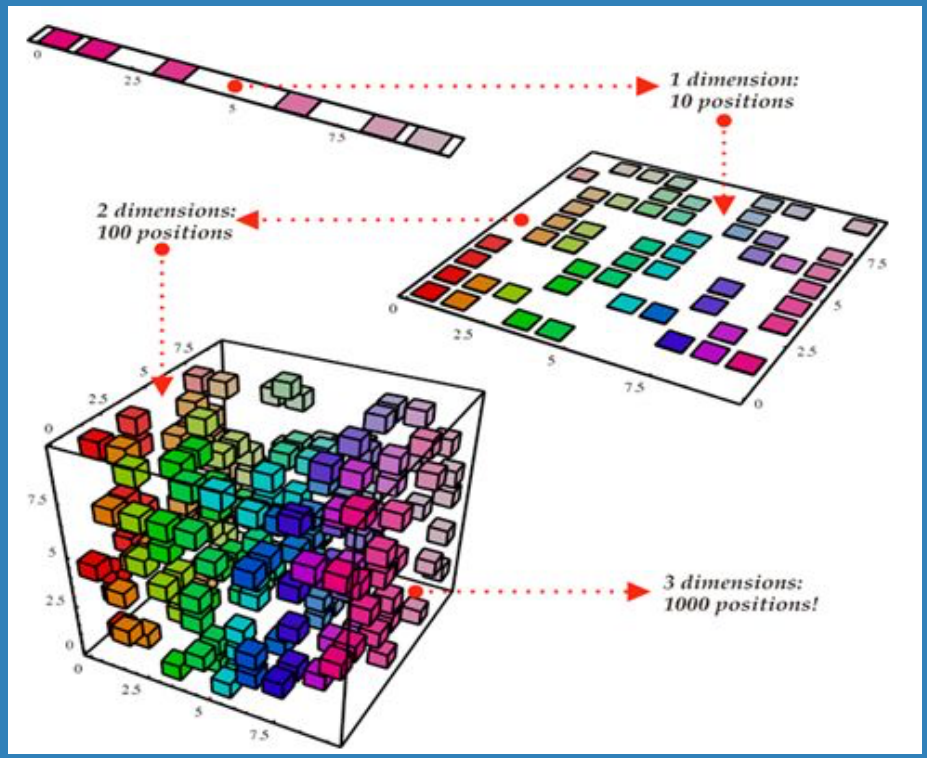
Heteroscedasticity



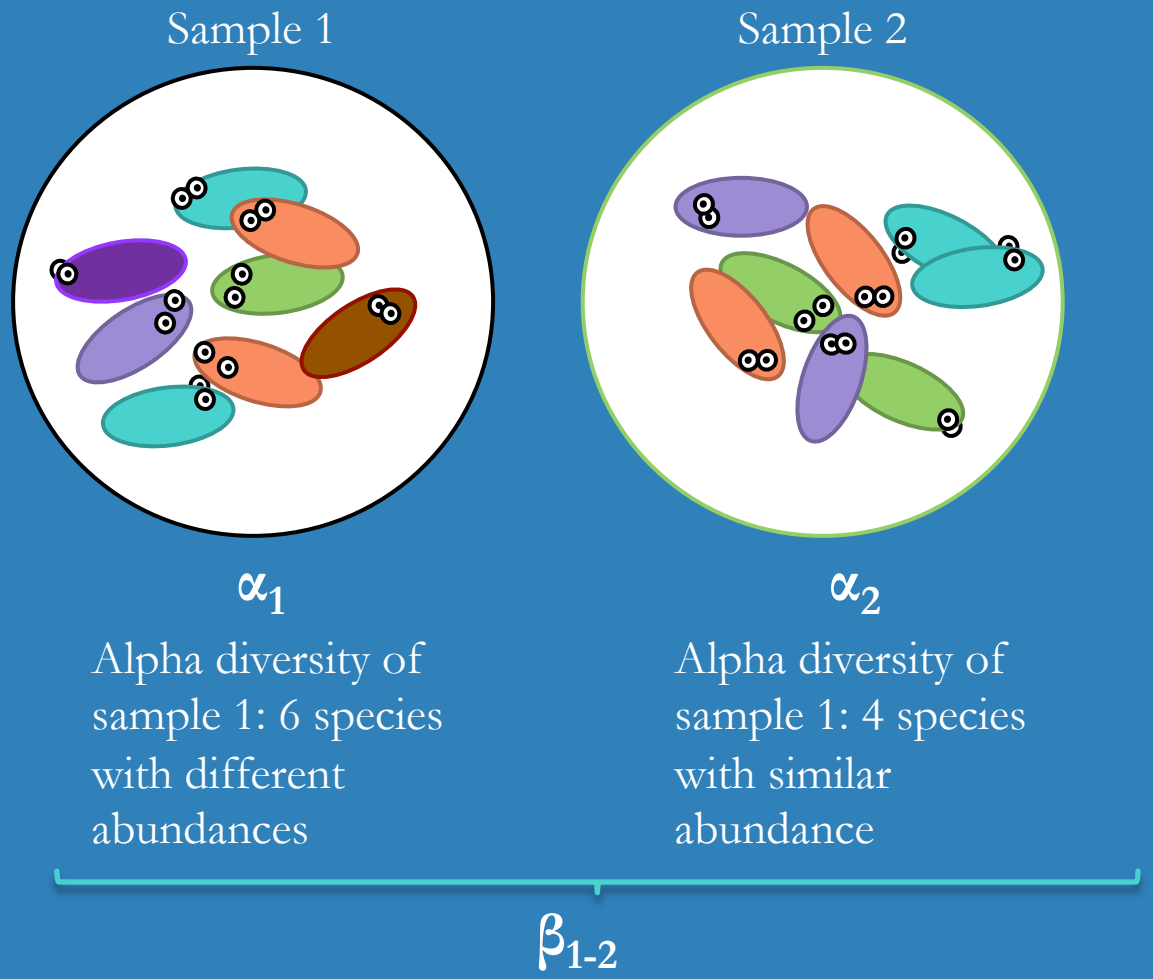
Sparsity



Dimensionality



DIVERSITY ANALYSIS



ALPHA DIVERSITY

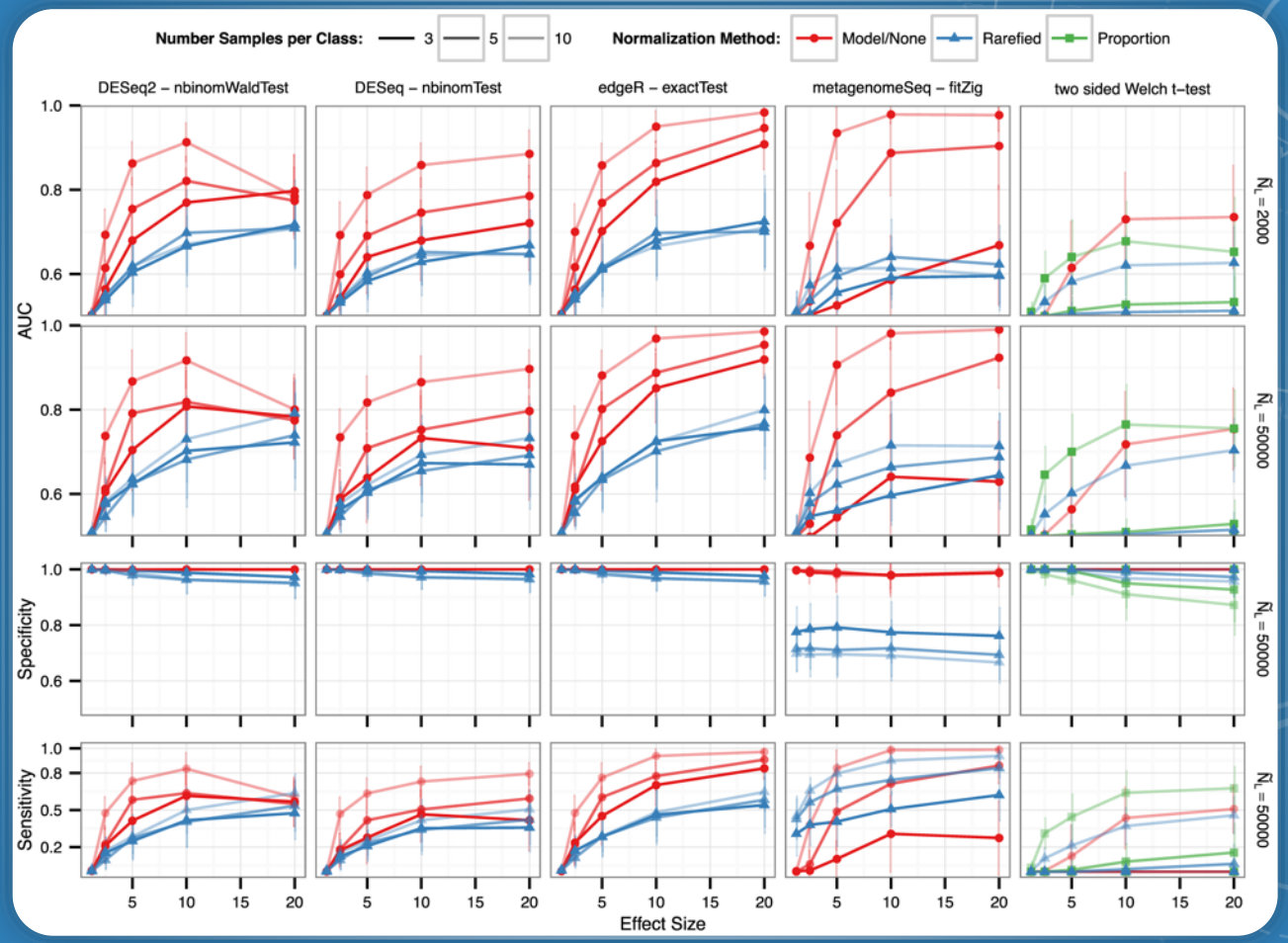
- DOES NOT NEED distribution assumptions:
  - DO NOT normalize your data
  - DO NOT filter your data
- Several indexes are possible (and often different names mean the same index), but they do different things, with different power, and measure different aspects:
  - Shannon, chao1, observed: measures *Richness*
  - Simpson: measures *Dominance*
  - Inverse Simpson: measures *Evenness*
- Rarefaction curves are a good proxy for data quality check!
  - Split the dataset into bins and plot the cumulative sums of the selected α diversity index
  - Look for plateau

BETA DIVERSITY

- Due to the nature of the data, making comparisons between samples is pretty difficult
- To compare samples diversities, you NEED to make them comparable, which means:
  - Normalize (i.e. fitting counts in a Gaussian curve within same range for all samples)
  - Reduce number of outliers
  - Etc.
- Introducing DISTANCES:
  - Indexes representing “dissimilarity” between samples
  - Make lots of assumptions
- Several indexes existing, including Bray-Curtis (works better on log- or root-transformed data), Jaccard (intersection over union), Manhattan (Euclidean distances), UniFrac (includes phylogeny)
- Overall, when plotting beta diversity, samples close together are more similar with each other

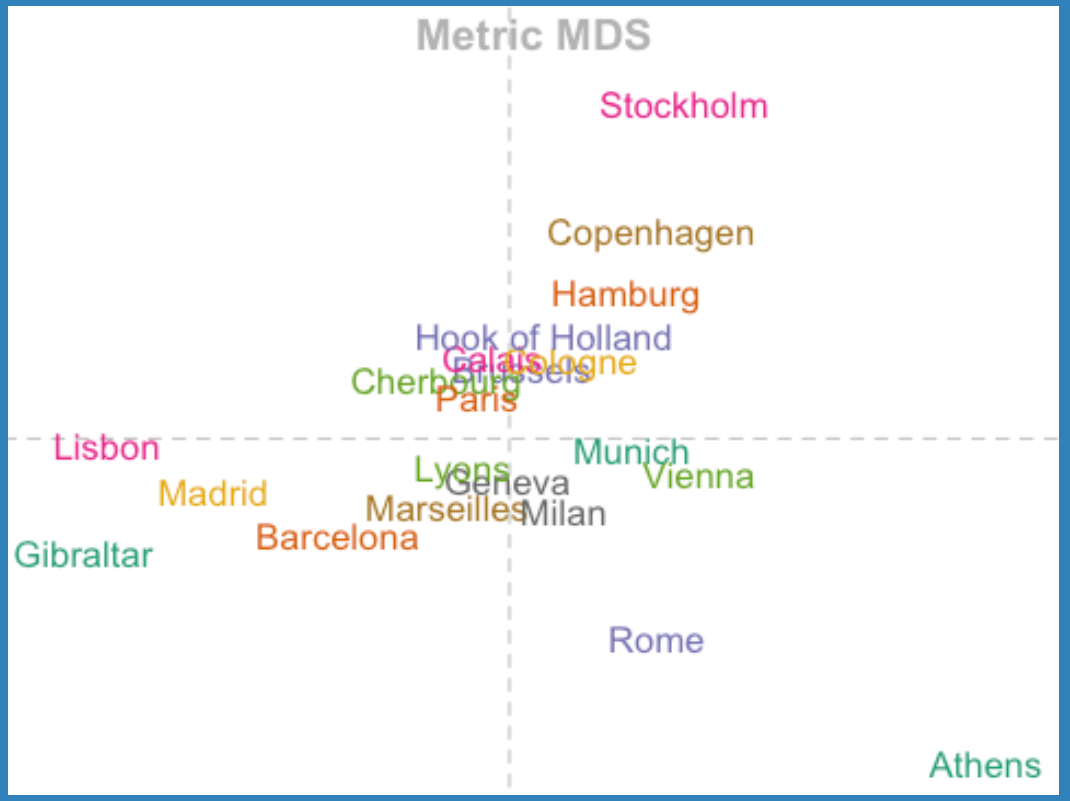
WASTE NOT WANT NOT: DO NOT RAREFY!

- Rarefying = resampling
  - Equalize number of counts/sample
  - With or without replacement
- Microbes are NEVER normally distributed:
  - Negative binomial distribution (mostly)
- Rarefying kills!
  - Does not solve heteroscedasticity
  - Loss of Accuracy, Sensitivity
  - Tendency to false positive
  - Increase of Type II errors (FN)
  - Arbitrary
- Flaws especially true for differential analyses on small number of samples



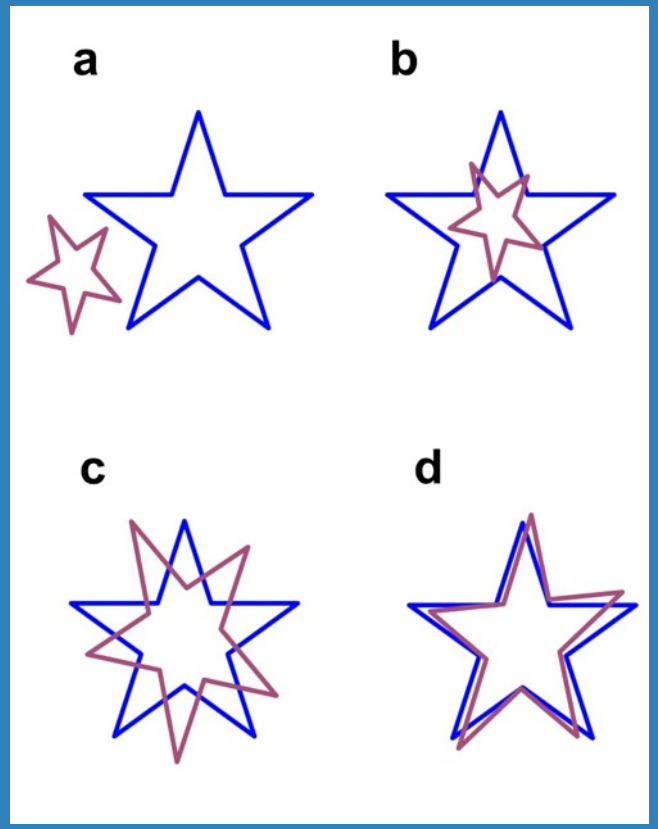
PCoA: PRINCIPAL COORDINATE ANALYSIS (or MDS)

- NOT PCA
- If I know distance between cities, but not their coordinates, how can I draw a map?
- Your count table is converted into a matrix of dissimilarity (using the diversity index chosen)
- May be impacted by high variance (so, you need to normalize)
- If some results in the dissimilarity matrices are negative, it leads to imaginary numbers in the eigenvectors
- The eigenvectors are not your variables, but are correlated with different percentages



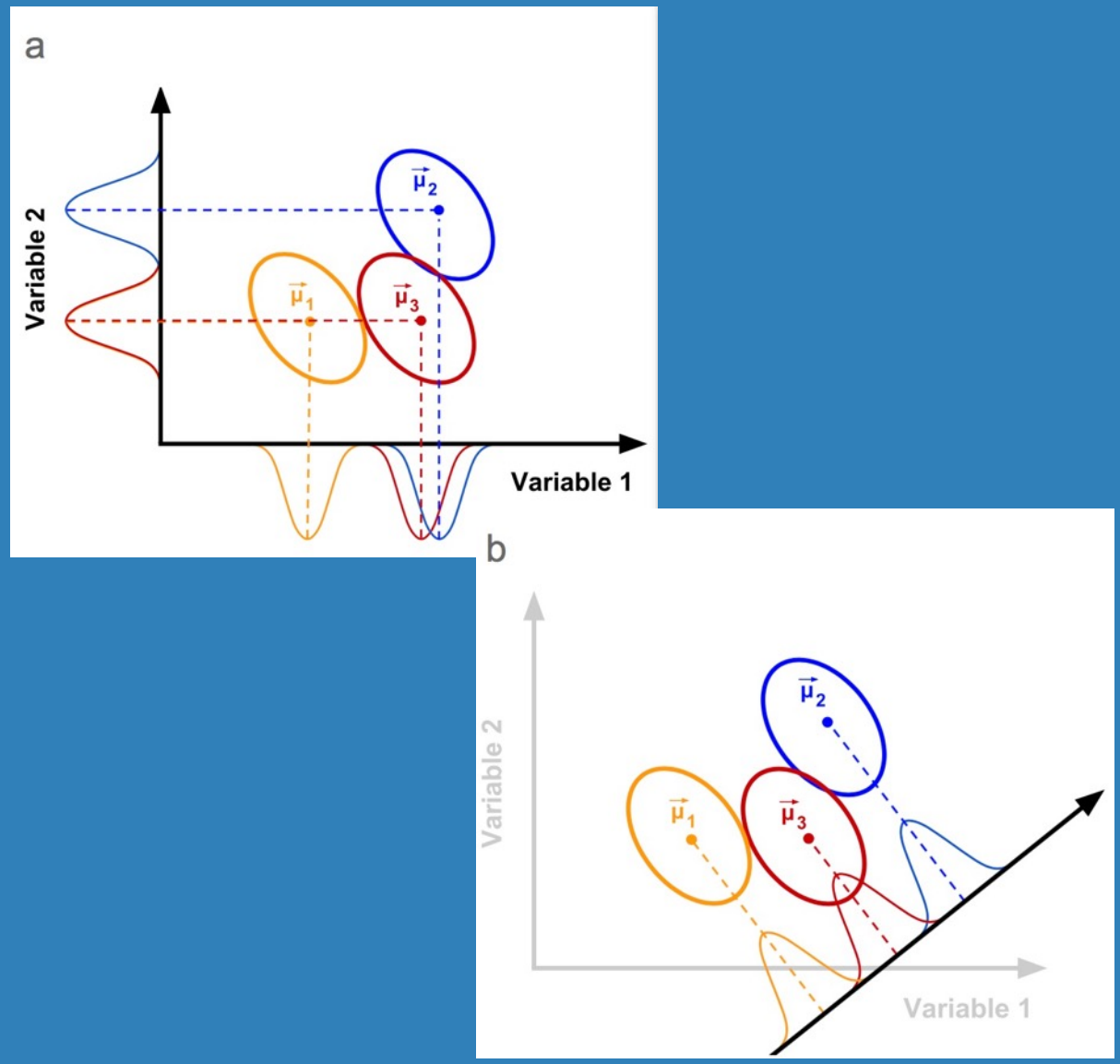
NMDS: NON-METRIC MULTIDIMENSIONAL SCALING

- Starts from a dissimilarity matrix
  - Ranking
- More robust than PCoA
- Applies Procrustes analysis to modify the matrix so that the eigenvectors are close to the original dimensions
  - Generate a ”stress value”
    - <0.1 good model
    - 0.1 < stress < 0.2 meh model
    - 0.2 < stress < 0.3 bad model
    - >0.3 random
- Eingenvectors are still NOT the original dimensions



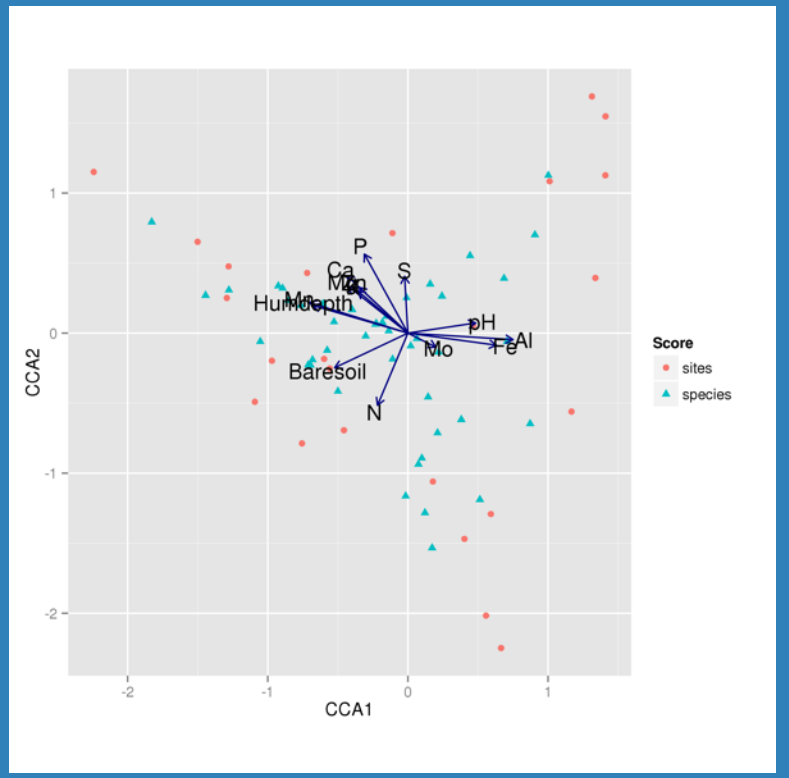
ANOSIM & PERMANOVA

- Both similar to ANOVA
- ANOSIM uses a dissimilarity matrix instead of the raw data
  - Finds differences between groups
  - Highly sensitive to dispersity
- PERMANOVA is a
  - Multivariate ANOVA (i.e. multiple factors influence multiple responses)
  - With PERmutations (solves the problem of limited number of samples)
  - Sensitive to dispersity



CCA: CANONICAL CORRESPONDENCE ANALYSIS

- Analyzes correspondences between a matrix of frequencies and a matrix of variables
  - How much the frequencies deviate from random per each variable?
- Correlate counts to variables (finally!)
- Does not try to maximize coverage of variance (unlike PCA)
- You can use ANOVA on CCA
- BEWARE of using only significant variables



MULTICOLLINEARITY

Why is it a problem?

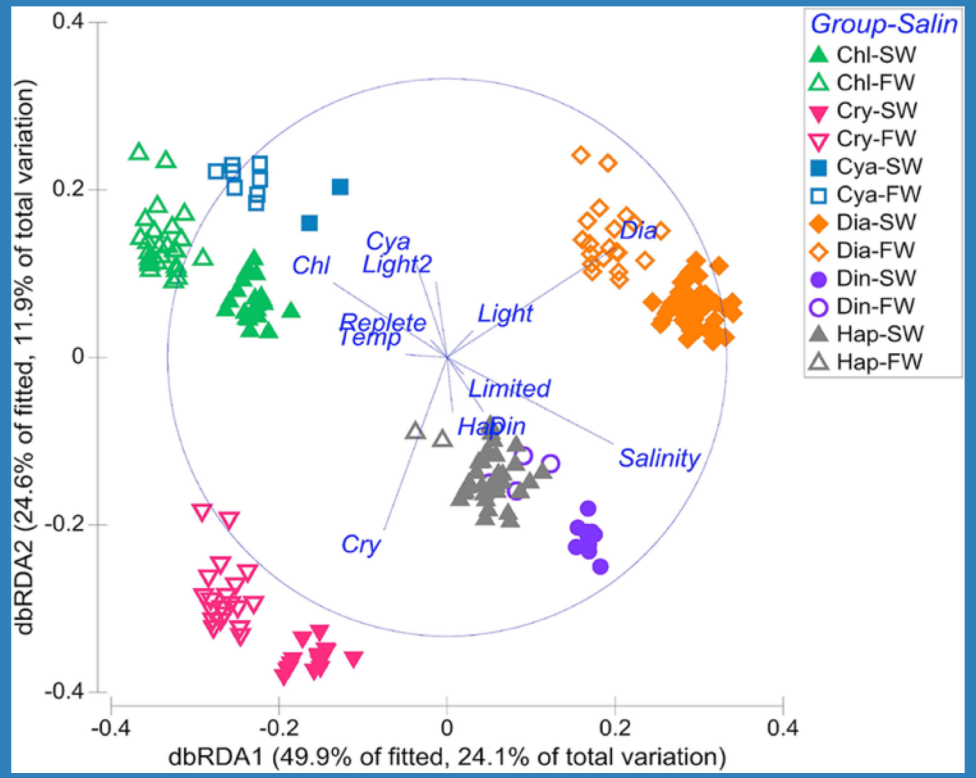
- Variables are highly correlated
- Thus, data are redundant
- Impacts of variables are wrongly estimated
- Small change of inputs -> high change of outputs
- Overfitting and huge errors

How can I solve it?

- Detect collinearity with VIF (variance inflation factor)
- Check dummy variable trap (encoding)
- Increase data
- Decrease number of variables

CAP (or db-RDA): CONSTRAINED ANALYSIS OF PRINCIPAL COORDINATES

- Basically a PCA where you constrain your components to your variables
  - Maximizes explanation of variance
- Similar (but different) to CCA
  - Still sensitive to collinearity
- BEWARE of using only significant variables



HOW ARE WE DOING?

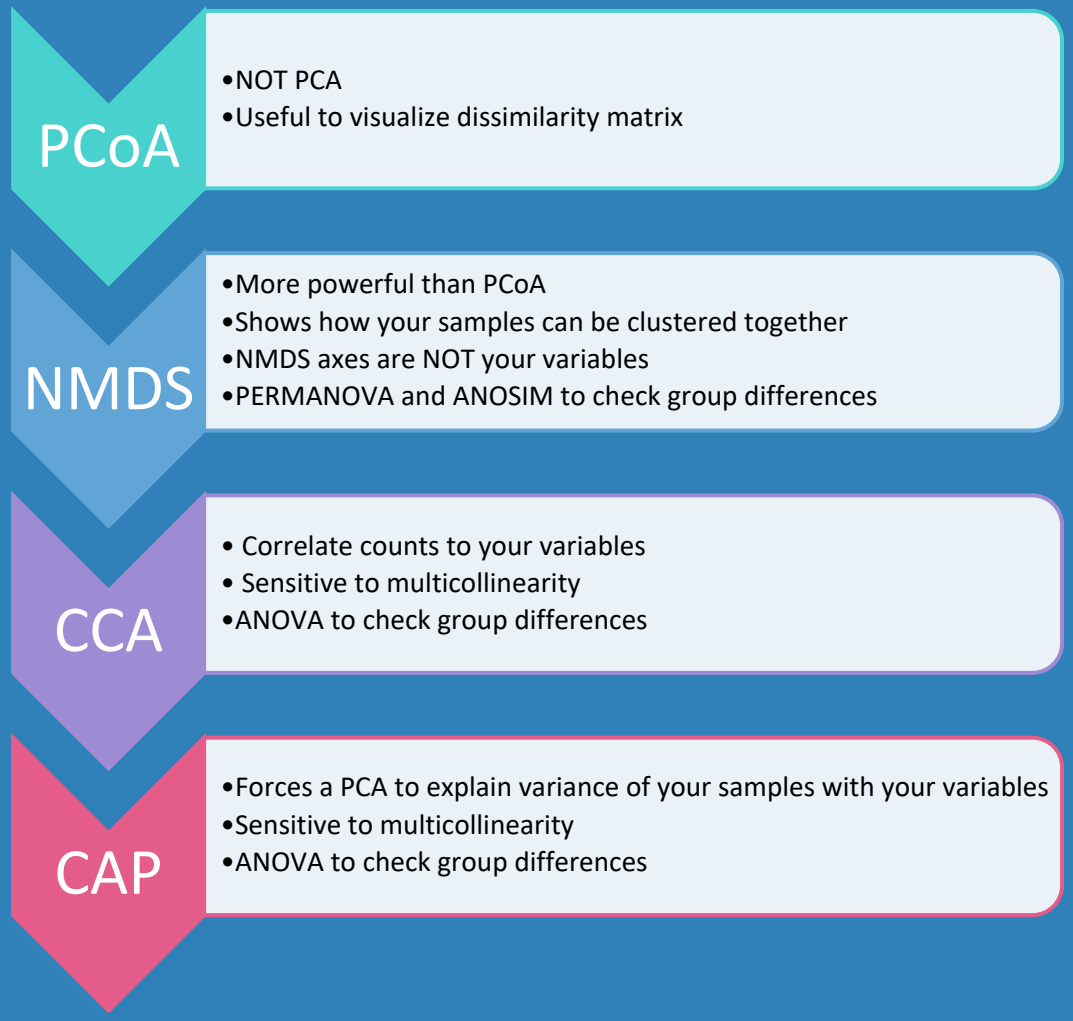
A

B

C

RECAP OF MULTIVARIATE ANALYSES

- You want to find out how the WHOLE microbiome is related to your Conditions or plant data
  - Possibly, which part of the microbiome is the most important
- α diversity: single index for each sample
- β diversity: distance index between samples
- DO NOT RAREFY



CODE