

Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology

10 August 2017 – Ashley Shade
shadeash@msu.edu
Kellogg Biological Station
Michigan State University

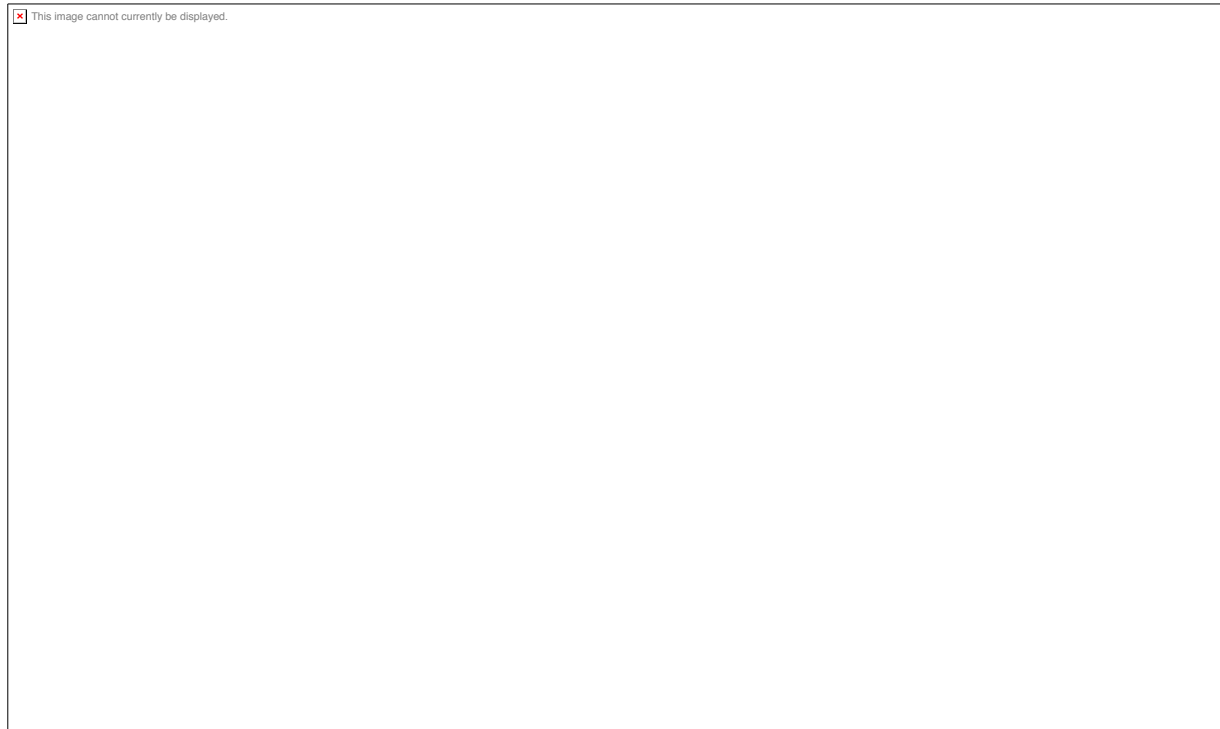
Diversity Outline

- The confusion about diversity
- Within-Sample (alpha) diversity
- Subsampling
- Comparative (beta) diversity
- Directing your data exploration to your question
- Categories of metadata for comparative diversity analysis: categories/clusters v. gradients
- Resemblances
- Ordination
- Hypothesis testing

Diversity in all of its glory


- **“Diversity” is a vague word.** In ecology, it has there are many types of diversity (*e.g.*, alpha, beta, gamma), and there are many components to that contribute to those types.
- Alpha diversity refers to the diversity inherently descriptive of one sample.

Whittaker introduces alpha, beta, gamma diversity (1972)




The confusion continues... for decades

Ecology Letters 2011

 This image cannot currently be displayed.

 This image cannot currently be displayed.

 This image cannot currently be displayed.

 This image cannot currently be displayed.

Oecologia 2010

We don't measure diversity well

- Diversity does not have an absolute value upon which multiple scales or methods of measuring will agree (think about temperatures: it doesn't matter if you measure in Celsius or Fahrenheit, you still measure the "same" value of temperature)
- Diversity depends on OTU definitions, method of measurements, method of data processing, and choice of diversity index/metric.
 - It is very difficult to compare alpha diversity across studies or investigators
 - Thus, great care is needed for interpreting such comparisons.

Within-sample (aka *alpha*) diversity

- Within-sample diversity includes:
 - Richness (number of taxa)
 - Evenness (distribution of the abundances of taxa)
 - Phylogenetic diversity (breadth of phylogenetic representation)
 - *Composition (who's there – identity of the taxa)
- Combinations of the above components are used to calculate other diversities: Shannon diversity, Simpson, *etc.*

Within-sample diversity

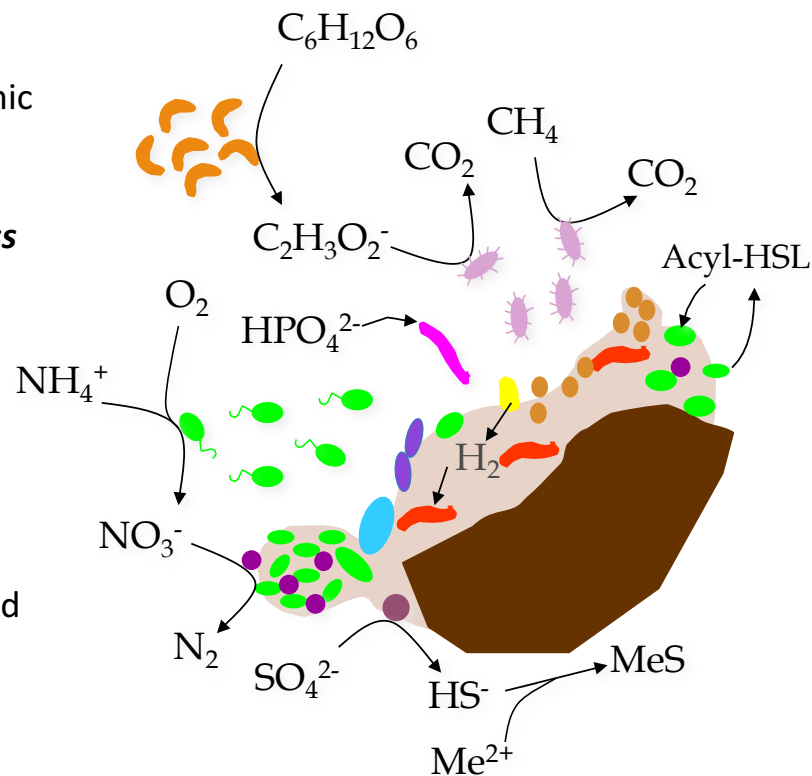
Information about the community that we can glean from metagenomic sequencing

- A certain number of OTUs- **richness**

- Each OTU is present in a certain abundance- collectively, **evenness**

- Each OTU has a taxonomic assignment- **composition**

- **Phylogenetic breadth** - how related are the lineages represented in the community?



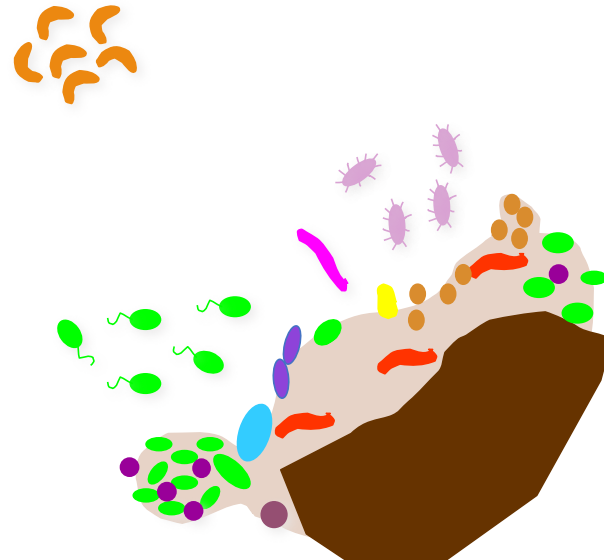
Richness

OTU

Richness: How many OTUs?



Richness = 11 OTUs



Evenness

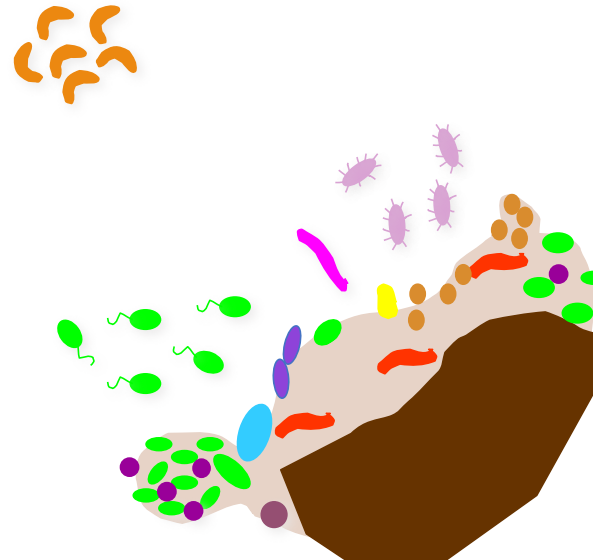
Evenness: What is the distribution of abundances in the community?

OTU

Count:

No. seq, no. individuals (e.g., FISH), biomass, etc.

	6
	1
	4
	8
	1
	5
	3
	13
	5
	1
	2



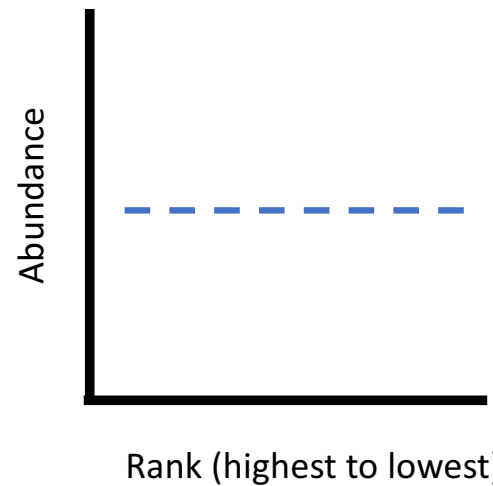
Evenness

Evenness: What is the distribution of abundances in the community?

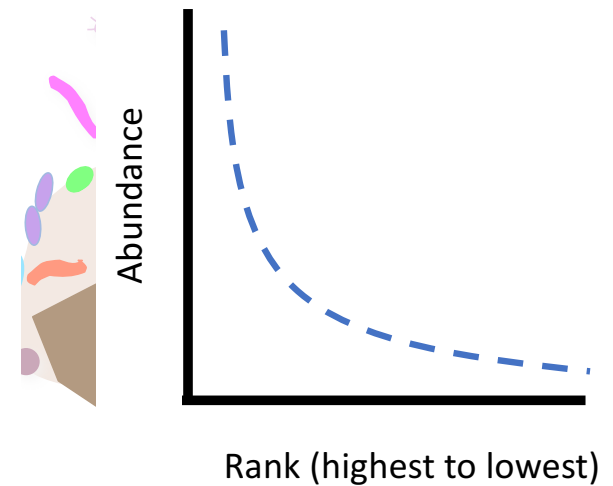
OTU Count:

	13
	8
	6
	5
	5
	4
	3
	2
	1
	1
	1
	1

Even

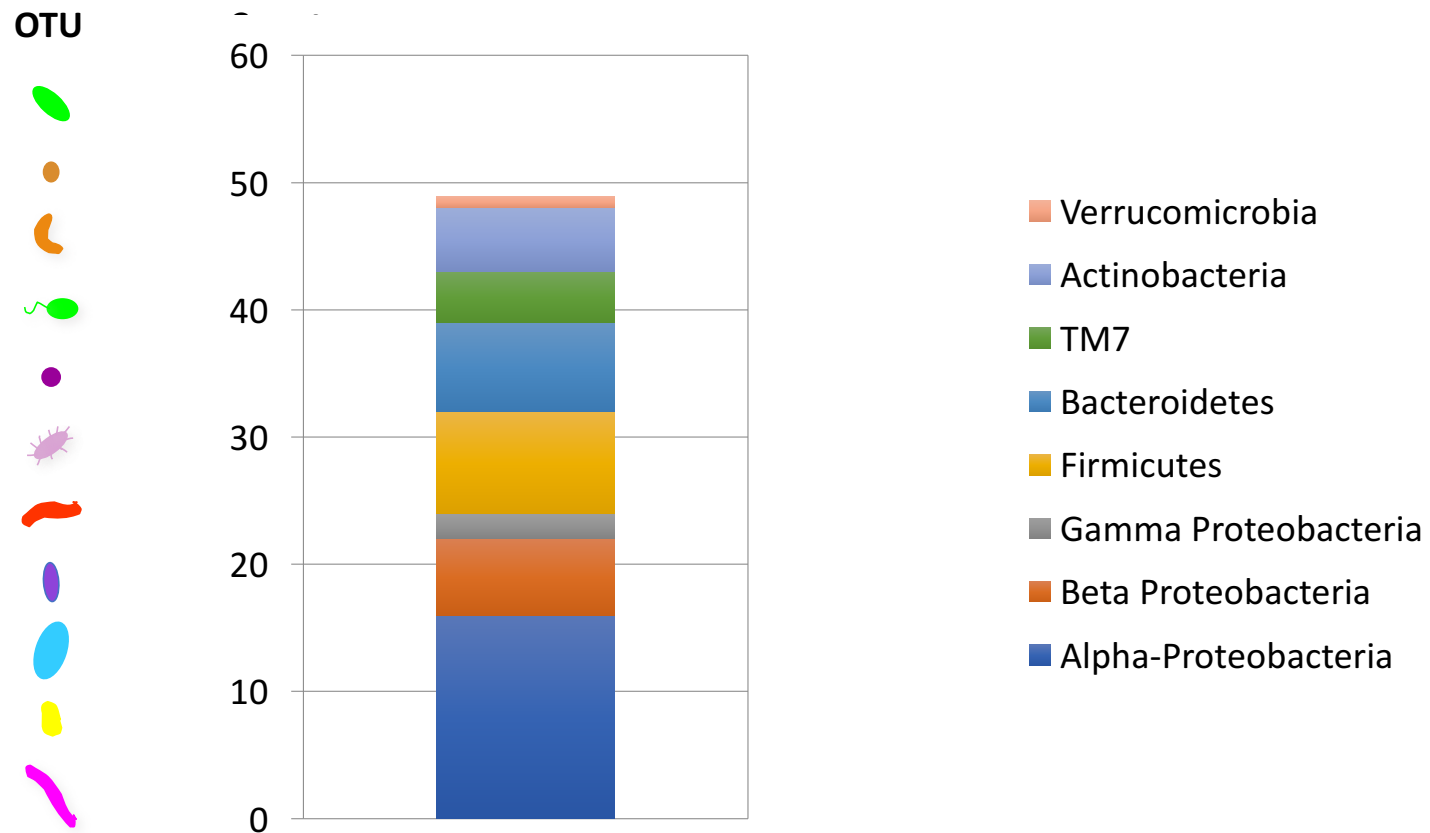


Uneven














Membership and Composition

Composition: Who is there?

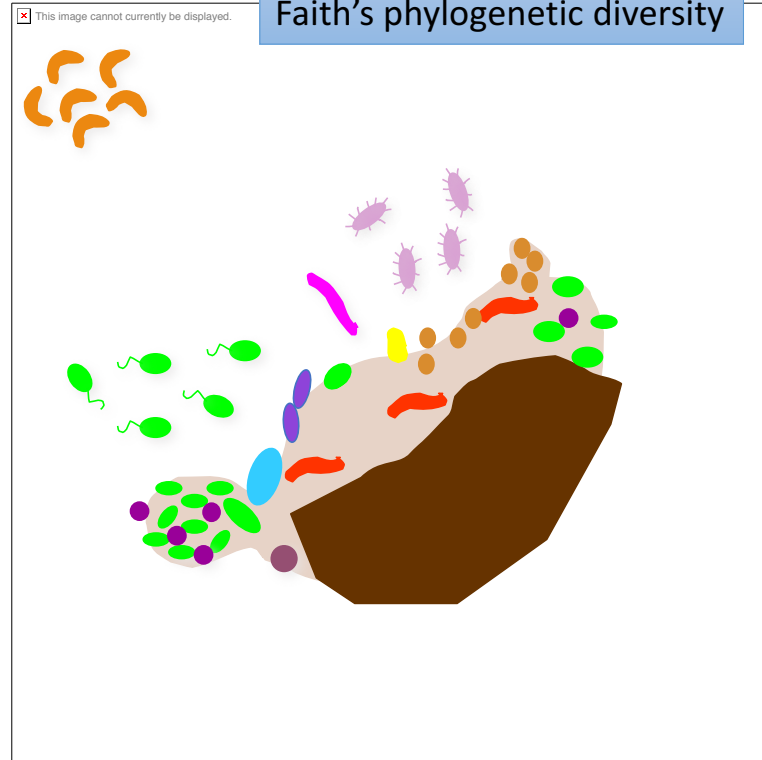


The advantages of phylogeny

Phylogenetic diversity: What is the breadth of phylogenetic representation?

OTU	Count:
	13 Alpha-proteobacteria
	8 Firmicutes
	6 Beta-proteobacteria
	5 Bacteroidetes
	5 Actinobacteria
	4 TM7
	3 Alpha-proteobacteria
	2 Gamma-protobacteria
	1 Bacteroidetes
	1 Bacteroidetes
	1 Verrucomicrobia

Common metric =
Faith's phylogenetic diversity



A final word about within-sample diversity

- Higher diversity isn't necessarily better or healthier. Diversity is the outcome of ecological processes, not an ecological process in itself.

Subsampling: Get “even”

- Because of sequencing artifacts and experimental design, there can be quite a range of quality sequences returned for each sample.
- Sub-sampling of sequences to achieve an even number across all samples within a dataset allows for comparing diversity across samples.
- **This is very important for being able to compare diversity across samples.**
 - Analogy: You are a tree ecologist. Would it be reasonable to directly compare the forest diversity (assessed by counting different types of trees) in a 1x1 m plot to a 1000 x 1000 km plot? The second plot has 1000x the coverage as the first and thus the comparison is unsound.
 - This is a question of the observational effort. We can only compare communities observed at an equal observational effort.
- Choose a sequencing depth that maximizes the number of samples that can be included at the most informative sequencing depth.
- If you have obvious outliers in sequencing depth (e.g., the median depth is 70K and you have one sample with 1000 sequences), get rid of it and save yourself heartache.

Rarefaction for estimating diversity given coverage

- I call this “exploratory” rarefaction
- It lets us understand how well we’ve observed our community
- Calculate a diversity of choice (e.g., richness)
- The question asked: Given an increase in sequencing (observation/coverage) effort, how likely are we to observe a new taxon?
- Exploratory rarefaction happens to understand your coverage BEFORE subsampling to an even sequencing effort

 This image cannot currently be displayed.

Diversity Part 1 Review

- **Within-sample diversity** describes a single community/ sample, and includes metrics of **richness**, **evenness**, **phylogenetic diversity**, and other summative metrics of diversity.
- Because sequencing success can be highly variable, **rarefaction** is used to ensure an even-depth of sequences across communities that will be compared.
- Alternative normalization methods cannot substitute for low quality data or a poor sequencing run!

Outline: Comparative (beta) diversity

- What questions can you ask about your microbial communities?
- Comparative diversity
 - Calculating community resemblance
 - Visualizations: Ordinations, heatmaps, dendrograms
- Gradients versus categories (clusters)
 - Statistical analysis of clusters: Hypothesis testing for differences in categorical groups
 - Statistical analysis of gradients: Linking environmental variables to changes in community structure

Questions about microbial communities

- Summary information for each community: *Within-sample (alpha) diversity*
- Differences between communities: *Comparative (beta) diversity*

Ask yourself: What is the purpose of the analysis?

1. **Exploration/Discovery:** hypothesis generating, perfect for observational studies, includes visualizations like ordinations and clustering
2. **Hypothesis testing:** address a specific question (*e.g.*, are there differences among treatment groups?), and usually permutation-based (non-parametric) p-value

What questions do you want to ask about your microbial communities?

Comparative diversity

- Space / Time
- Categories (e.g., fire-affected, recovered)
- Gradients/empirical measurements (e.g., pH, temperature, chemistry)
- Look forward to Stuart's R lecture on category/gradient analyses!

Analysis of comparative diversity is informed by:

- Associated environmental/quantitative variables*
 - Examples: temperature, red blood cell counts, glucose levels, dissolved oxygen, temperature, acidity, time, % mortality, etc.
- Associated categorical/descriptive/qualitative variables*
 - Examples: treatment groups, male/female, control/treatment, age groups, before/after

** Environmental and categorical variables often are linked to samples in a single “mapping file”*

Comparative diversity requires a measure of pair-wise community **resemblance**

- Resemblance = distance, similarity, dissimilarity
- Important decisions in choosing a resemblance metric:
 - Weighted v. Unweighted
 - Phylogenetic v. Taxonomic
- All pairs of resemblances are included in a sample by sample **resemblance (distance/similarity) matrix**
 - Simplifies the data and the analysis
- Choice of resemblance metric will influence the outcome of community analysis

Calculating resemblance: Bray-Curtis Example

$$d_{jk} = (\text{sum } \text{abs}(x_{ij} - x_{ik}) / (\text{sum } (x_{ij} + x_{ik})))$$

Where d_{jk} is the Bray-Curtis index between samples j and k
and x is the (relative) abundance of taxa l

See Legendre and Legendre book: *Numerical Ecology*.
Chapter 7: “Ecological resemblance” for a comprehensive
discussion of All the Resemblances Ever.

Making a Resemblance Matrix

1. OTU table (usually relativized)

	Soil 1	Soil 2	Soil 3
OTU 1	0	0.966	0.179
OTU 3	0.047	0.002	0.039
OTU 3	0.953	0.032	0.782



2. Chose appropriate resemblance (*e.g.*, Bray Curtis, UniFrac)

3. Create a square (observation x observation) resemblance matrix from pair-wise comparisons.

	Soil 1	Soil 2	Soil 3
Soil 1	0		
Soil 2	0.966	0	
Soil 3	0.179	0.787	0

Examples of Resemblance metrics

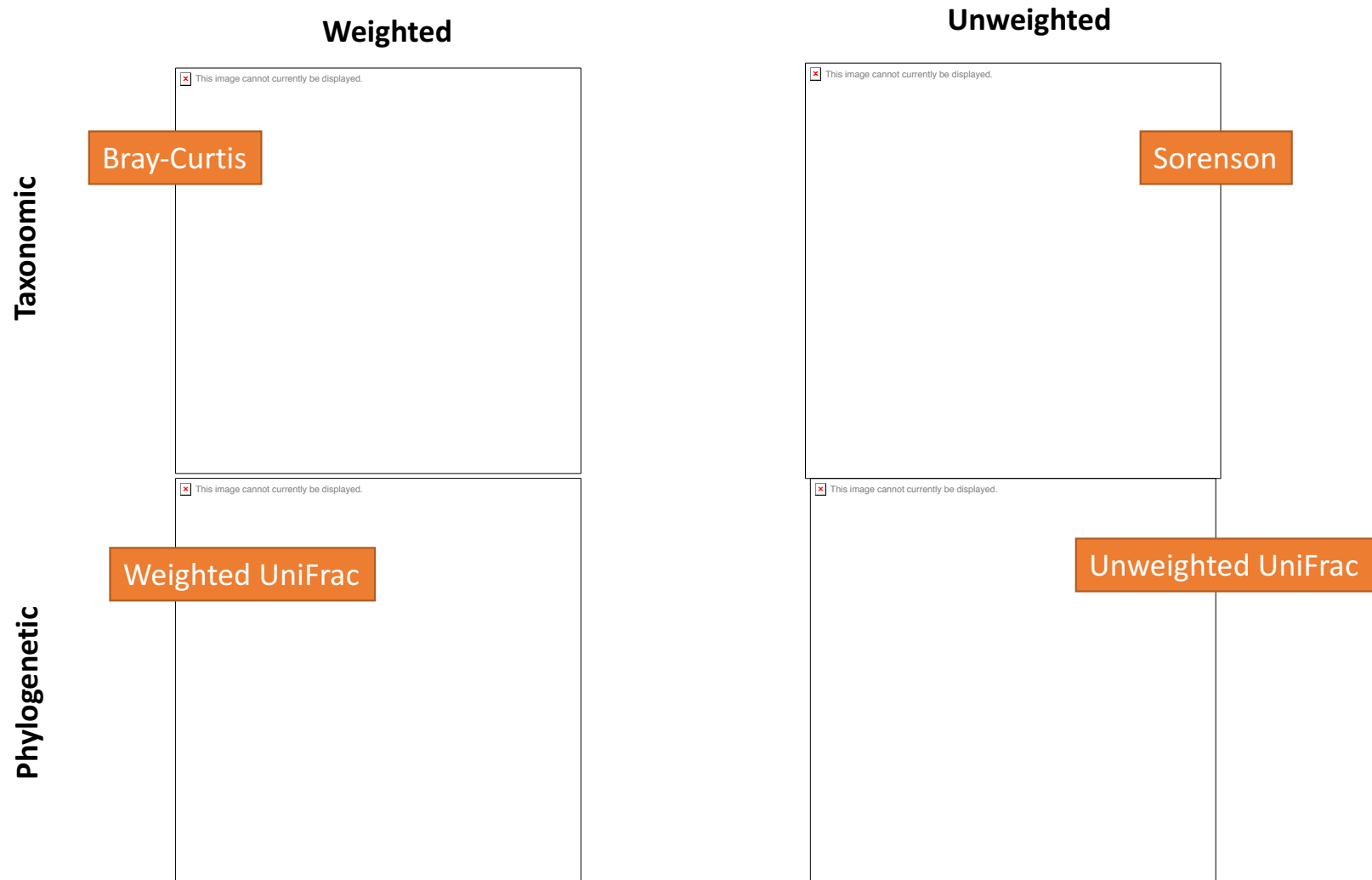


<i>Metric name</i>	Sørensen	Bray-Curtis	Weighted UniFrac	Unweighted UniFrac
<i>Accounts for</i>				
Composition	X	X	X	X
OTU abundances?		X	X	
Phylogenetic diversity?			X	X

We can compare different distance/similarity measures to deduce the most important components of community structure for the overarching patterns observed

(This does not mean we “pick” the best ordination for our hypothesis - *confirmation bias*)

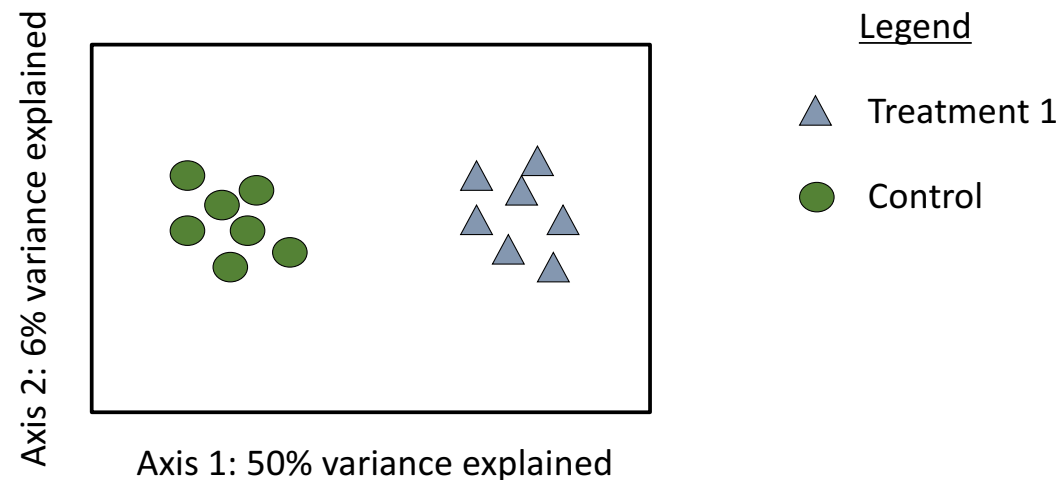
Example: Centralia coal mine fire-affected soils



Useful community visualization tools

- **Ordination** : Calculated from community resemblance; relationships are represented by distances between symbols
- **Heatmap** : Calculated from count/abundance data; The abundance of each taxon relative to the others depicted by color
- **Dendrogram**: Calculated from community resemblance; similar communities fall into same cluster.

Visualizing communities: ordination



2 or 3 dimensional representation of the data

Each symbol is one community (compared by the chosen resemblance metric)

The distance between symbols represents the extent of differences between communities

First axis often explains most variance in the data, variation explained should be provided.

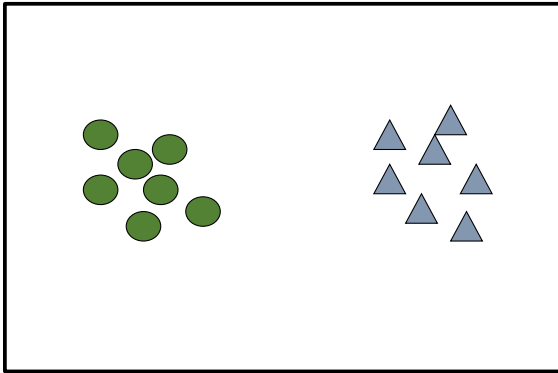
Types of ordinations

- **Non-metric multidimensional scaling (NMDS)**
 - Makes fewest assumptions, rank based, non-metric so cannot subsequently correlate to environmental measures
- **Principal coordinates analysis (PCoA)**
 - Assumes linear response of taxa, can use any resemblance
 - Special case Principal Components Analysis (PCA): based on Euclidean distance and not appropriate for communities (for environmental variables, PCA is okay)
 - Constrained version: Redundancy Analysis (RDA): variation in communities are *constrained* to be explained only by the measured variables. This is only okay if the ordination of the PCoA matches the RDA, which means that the important/most explanatory environmental variables were measured. Constrained analyses can be used for variance partitioning.
- **Correspondence analysis (CA)**
 - Assumes unimodal response of taxa, uses chi-squared distance
 - Constrained version: Constrained (Canonical) Correspondence Analysis (CCA)
 - Detrended correspondence analysis- if there is a large “horseshoe” or “arch” affect because too much variability is being squished onto too few dimensions; not often advisable.
- Avoid: Principle components analysis (PCA) for communities, Redundancy analysis (RDA) in situations that do not have a strong gradient (e.g., time) and detrended analyses *unless you really know what you are doing*.

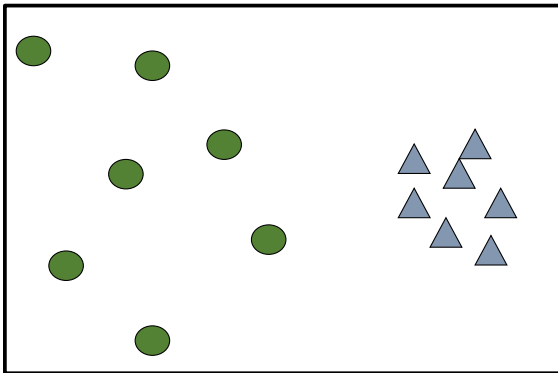
How do we look at ordinations?

Think about: **CENTROID (mean)** or **DISPERSION (spread, variability)**

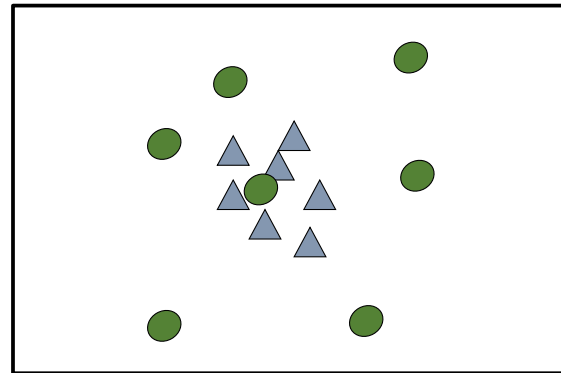
A. Different centroid, same spread



B. Different centroid, different spread

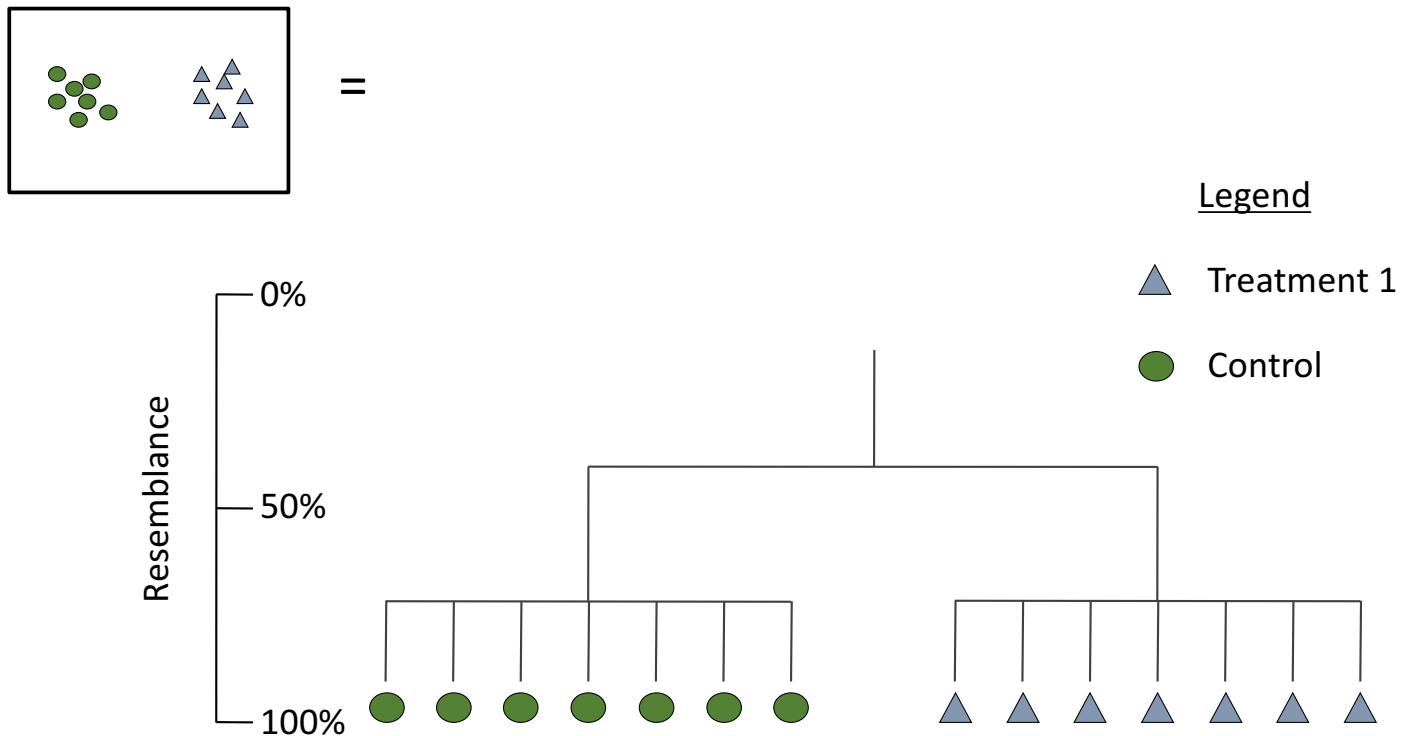


C. Same centroid, different spread



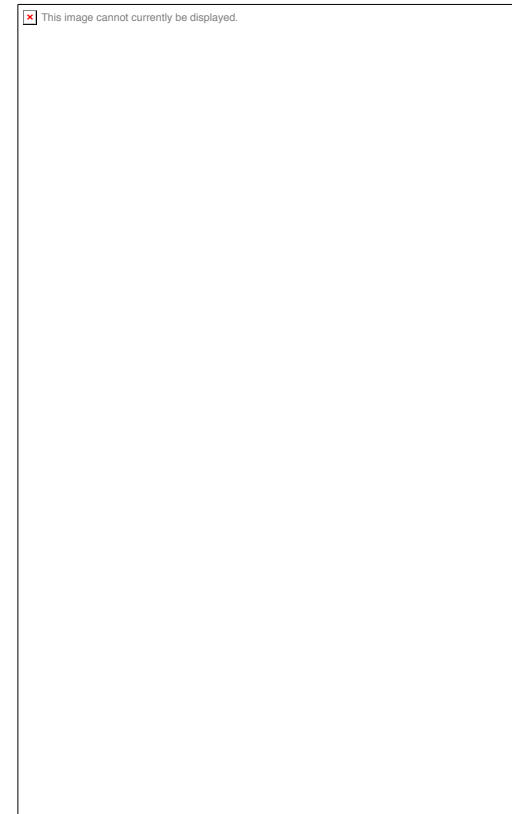
Visualizing communities: dendrograms

A different way of visualizing the same data

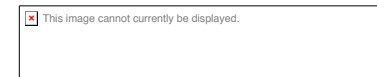


Visualizing communities: heatmaps

Figure 6. Bacterial distribution among the seven samples.



Wu S, Wang G, Angert ER, Wang W, et al. (2012) Composition, Diversity, and Origin of the Bacterial Community in Grass Carp Intestine. PLoS ONE 7(2): e30440. doi:10.1371/journal.pone.0030440
<http://www.plosone.org/article/info:doi/10.1371/journal.pone.0030440>



Discovering patterns: Clusters & Gradients

Clusters = Are groups different? (*e.g.*, Treatment v. Control)

Also called: factors, qualitative variables

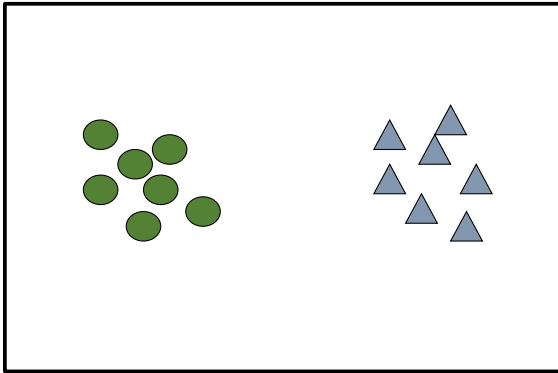
Gradients = Do communities change with known environmental changes? (*e.g.*, over time?)

Also called: continuous, quantitative, vector variables

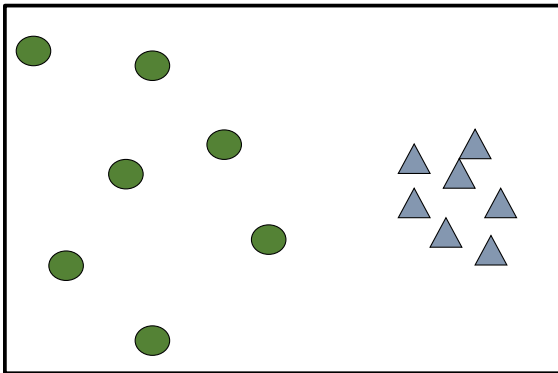
How do we interpret ordinations?

Think about: **CENTROID (mean)** or **DISPERSION (spread, variability)**

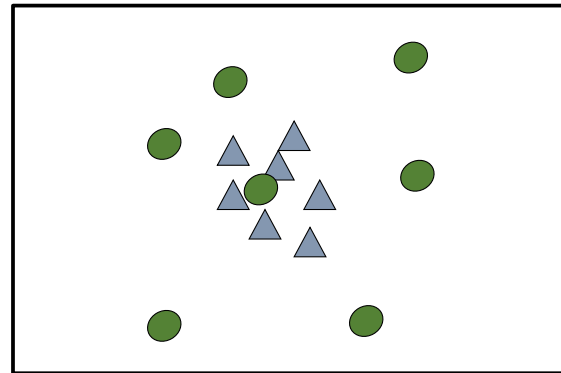
A. Different centroid, same spread



B. Different centroid, different spread



C. Same centroid, different spread




Non-parametric hypothesis tests

Non-parametric tests are used to test hypotheses of multivariate data when the underlying distribution of the data is unknown.

Non-parametric tests randomly re-sample the dataset to create a re-shuffled distribution, calculate a test statistic for each random distribution, and then ask the probably of finding the *actual* statistic given the random re-sampling distribution of the data.

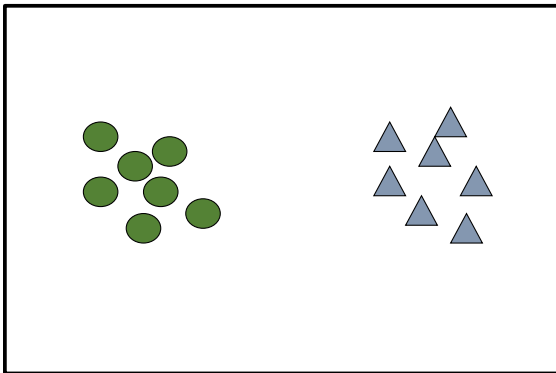
It is important to use these tests for microbial beta diversity, as the assumptions of underlying normal distributions of most parametric tests (e.g., ANOVA) are violated.

 This image cannot currently be displayed.

Clusters: Testing for differences in *a priori* groups

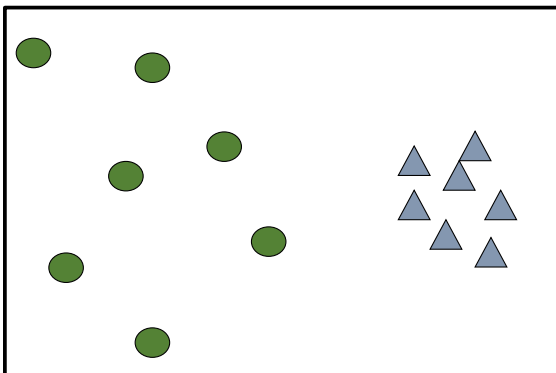
Permutation-based analyses to test hypotheses about group differences in
CENTROID (mean) or **DISPERSION (spread, variability)**

A. Different centroid, same spread

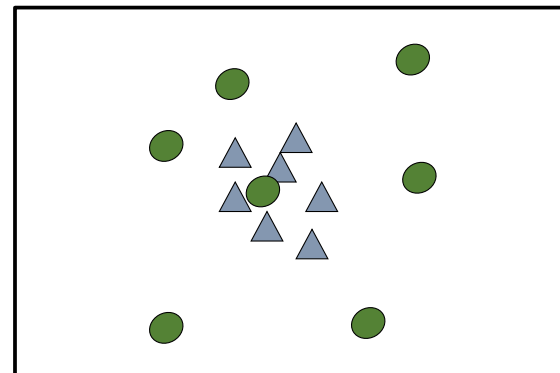


<i>Test name</i>	Centroid (mean)	Spread (variability)
PERMANOVA	X	X
MRPP	X	X
ANOSIM	X	X
PERMDISP		X

B. Different centroid, different spread



C. Same centroid, different spread



Hypothesis tests

- **Analysis of similarity (ANOSIM)**: rank based, least sensitive, makes fewest assumptions, permuted p-value
- **Multi-Response Permutation Procedure (MRPP)**: metric, permuted p-value
- **Permuted analysis of variance (PERMANOVA)**: assumes pseudo-F distribution, can accommodate a range of ANOVA-type experimental designs
- **Permutated analysis of dispersion (PERMDISP)**: p-values from permutation of residuals – tests specifically for differences in dispersion around centroid

A paper where every hypothesis test is used with every resemblance. Ever.

- (just kidding)
- (kind of)
- The methods are useful.

 This image cannot currently be displayed.

Shade *et al.* 2013 AEM

Gradients: Linking environmental and community data

1. Mantel Test

Community Resemblance

	Caterpillar 1	Caterpillar 2	Caterpillar 3
Caterpillar 1	0		
Caterpillar 2	0.966	0	
Caterpillar 3	0.179	0.787	0

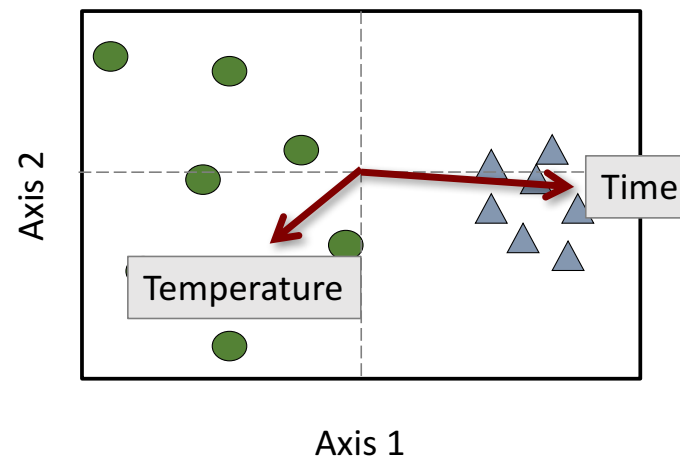


Pearson's correlation
Permuted p value

Time / environ. distance

	Caterpillar 1	Caterpillar 2	Caterpillar 3
Caterpillar 1	0		
Caterpillar 2	1	0	
Caterpillar 3	10	3	0

2. Vector fitting to ordination axis score



Exercise: Make your own ordination

	A	B	C	D	E
OTU1	0	2	3	4	8
OTU2	5	6	7	0	0
OTU3	20	10	7	0	0
OTU4	5	12	5	12	0
OTU5	0	0	4	7	11
OTU6	0	0	4	7	11