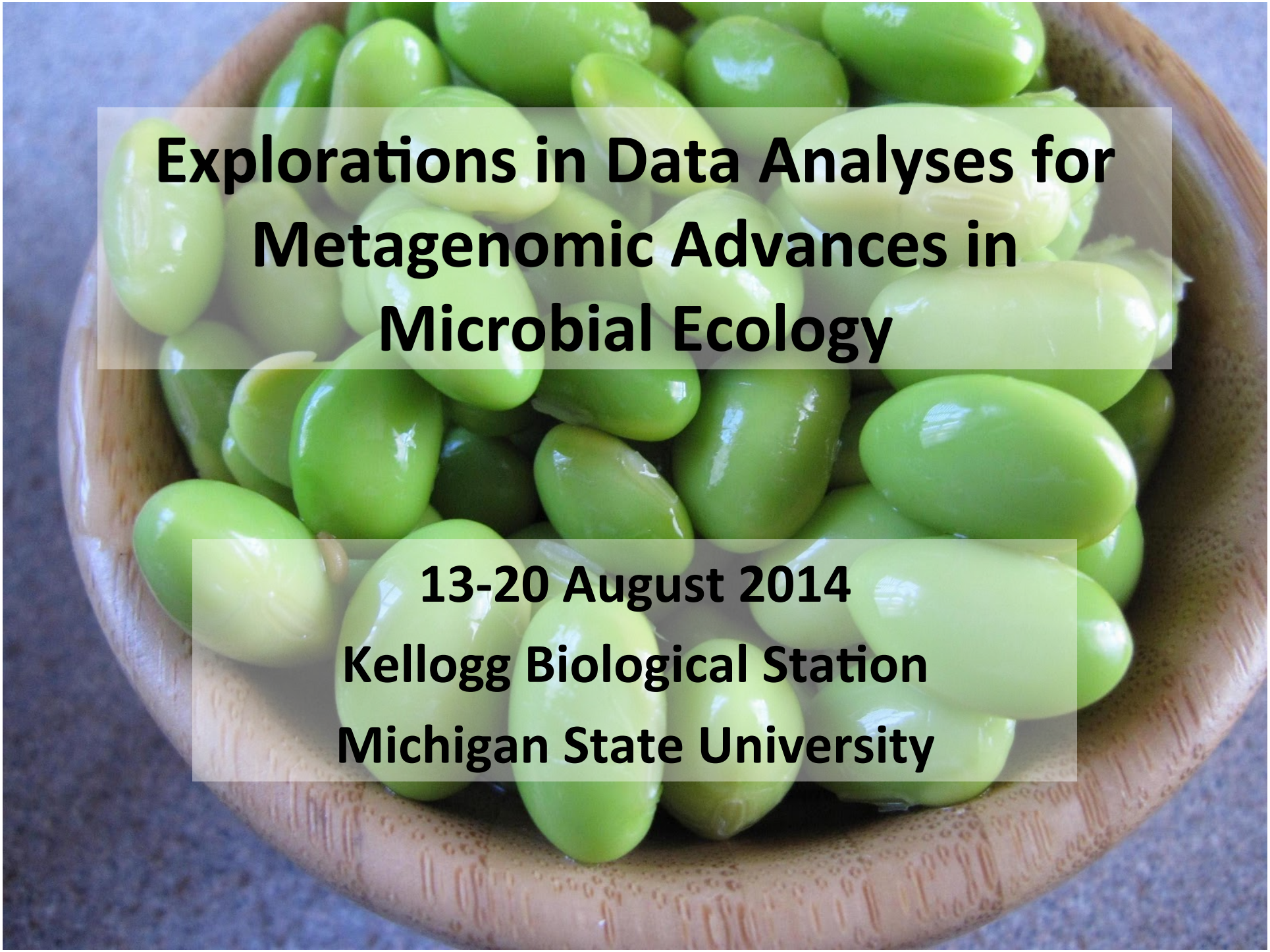

Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology
13-20 August 2014
Kellogg Biological Station
Michigan State University

Review from Lecture 2

- **Alpha 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.

- An **OTU table** is the input file for community analyses. It contains information about the abundance of each OTU within every sample. OTU tables can be (**classic, .txt**) or (.**biom**) format.

Tutorial: what did we do?

- Made taxonomic assignments of our sequences: **assign_taxonomy.py**
- Made OTU tables (biom + classic): **make_otu_table.py**
- Made a phylogenetic tree of our representative sequences: **make_phylogeny.py**
- Rarefied to an equal sequencing depth: **alpha_rarefaction.py**
- Calculated & visualized alpha diversity: **alpha_diversity.py, summarize_taxa_through_plots.py**

Questions from this morning?

Lecture 3: Beta diversity

- What questions can you ask about your microbial communities?
- Beta-diversity
- Gradients versus categories (clusters)
- Introduction to community resemblance
- Visualizing microbial communities

Questions about microbial communities

- Summary information for each community:
  Alpha diversity

- Differences between communities:
  Beta diversity

Comparative diversity

- Space / Time
- Categories (e.g., treatment v control)
- Gradients/empirical measurements (e.g., pH, blood sugar levels, temperature)

- Look forward to the R lecture on category/ gradient analyses!

Beta 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

Making a Resemblance Matrix

1. OTU table (usually relativized)

	Caterpillar 1	Caterpillar 2	Caterpillar 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.

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

Examples of Resemblance metrics

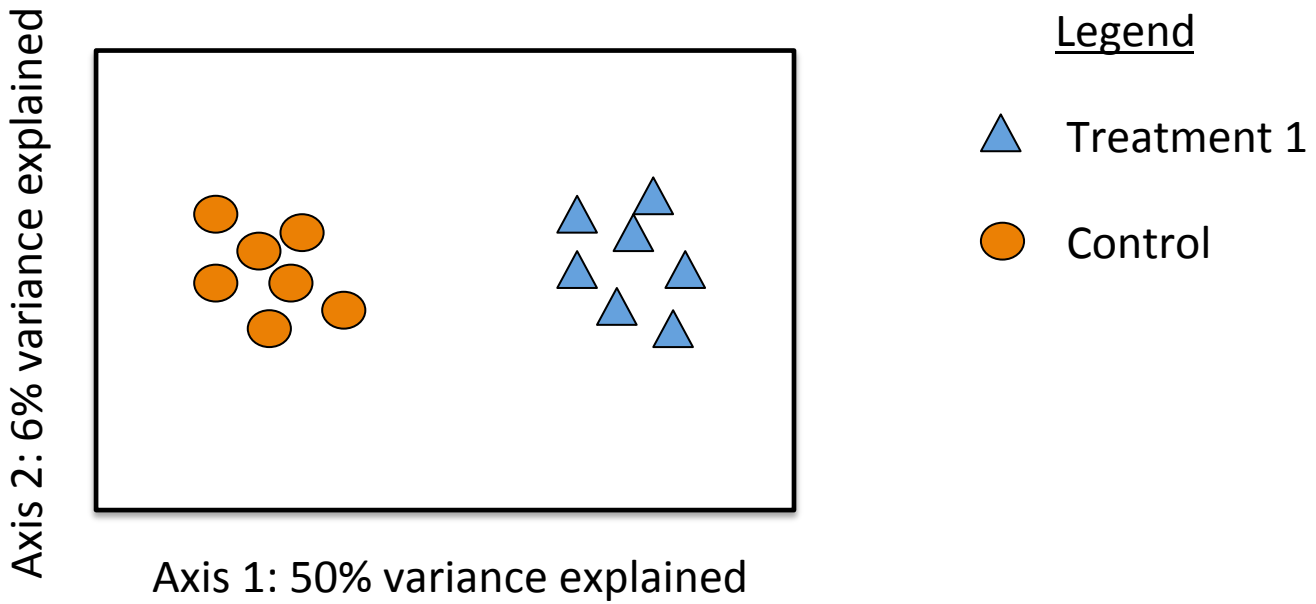
Metric name / *Accounts for*	Sørenson	Bray-Curtis	Weighted Unifrac	Unweighted Unifrac
Composition	X	X	X	X
OTU abundances?		X	X	
Phylogenetic diversity?			X	X

What is the purpose of the analysis?

1. **Exploration:** 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 p-value

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, should be labeled.

Types of ordinations

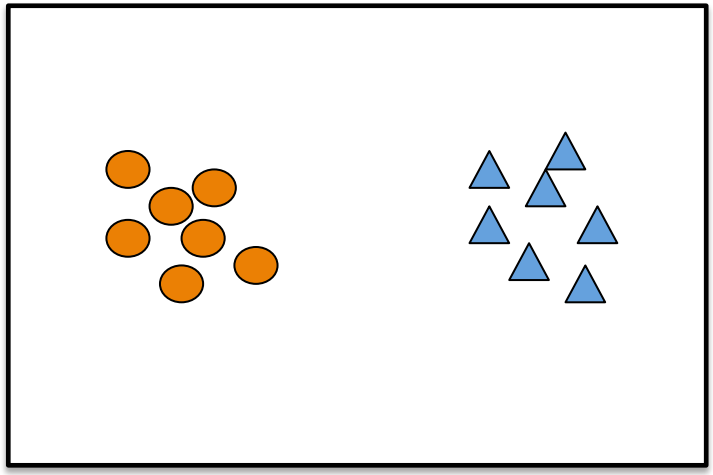
- Non-metric multidimensional scaling (NMDS)
- Principle coordinates analysis (PCoA)
- Correspondence analysis (CA)

- Avoid: Principle components analysis (PCA), Redundancy analysis (RDA) in some situations, and constrained analyses *unless you really know what you are doing*

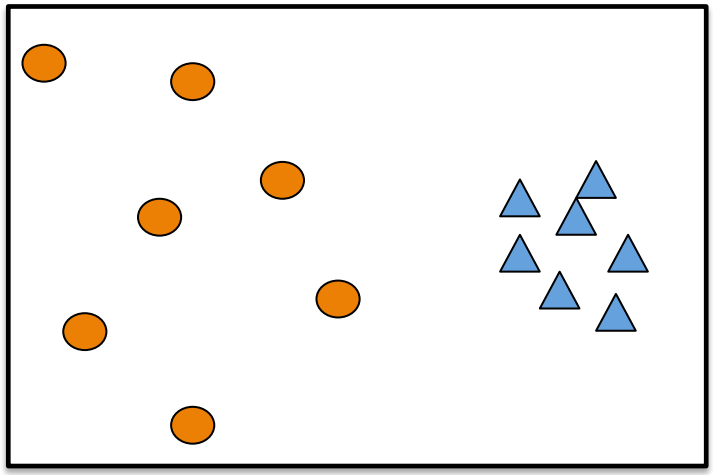
How do we look at ordinations?

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

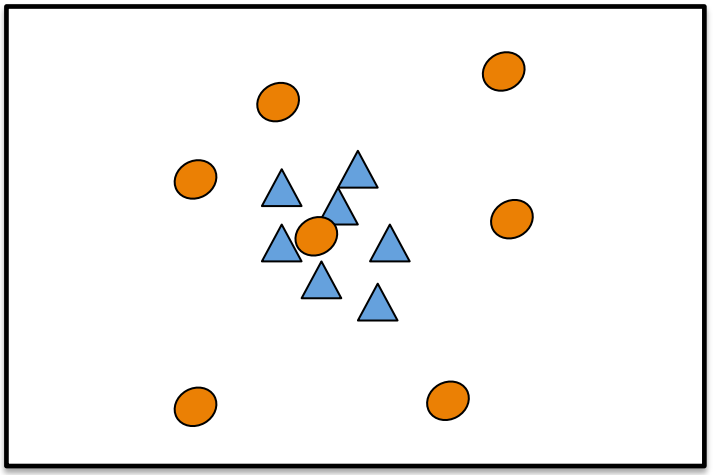
A. Different centroid, same spread



B. Different centroid, different spread



C. Same centroid, different spread



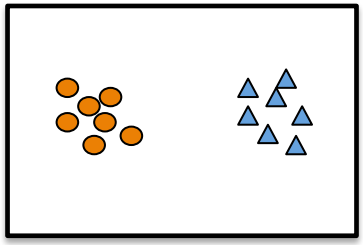
Types of ordinations

- Non-metric multidimensional scaling (NMDS)
- Principle coordinates analysis (PCoA)
- Correspondence analysis (CA)

- Avoid: Principle components analysis (PCA), Redundancy analysis (RDA) in some situations, and constrained analyses *unless you really know what you are doing*

Visualizing communities: clustering

A different way of visualizing the same data



=