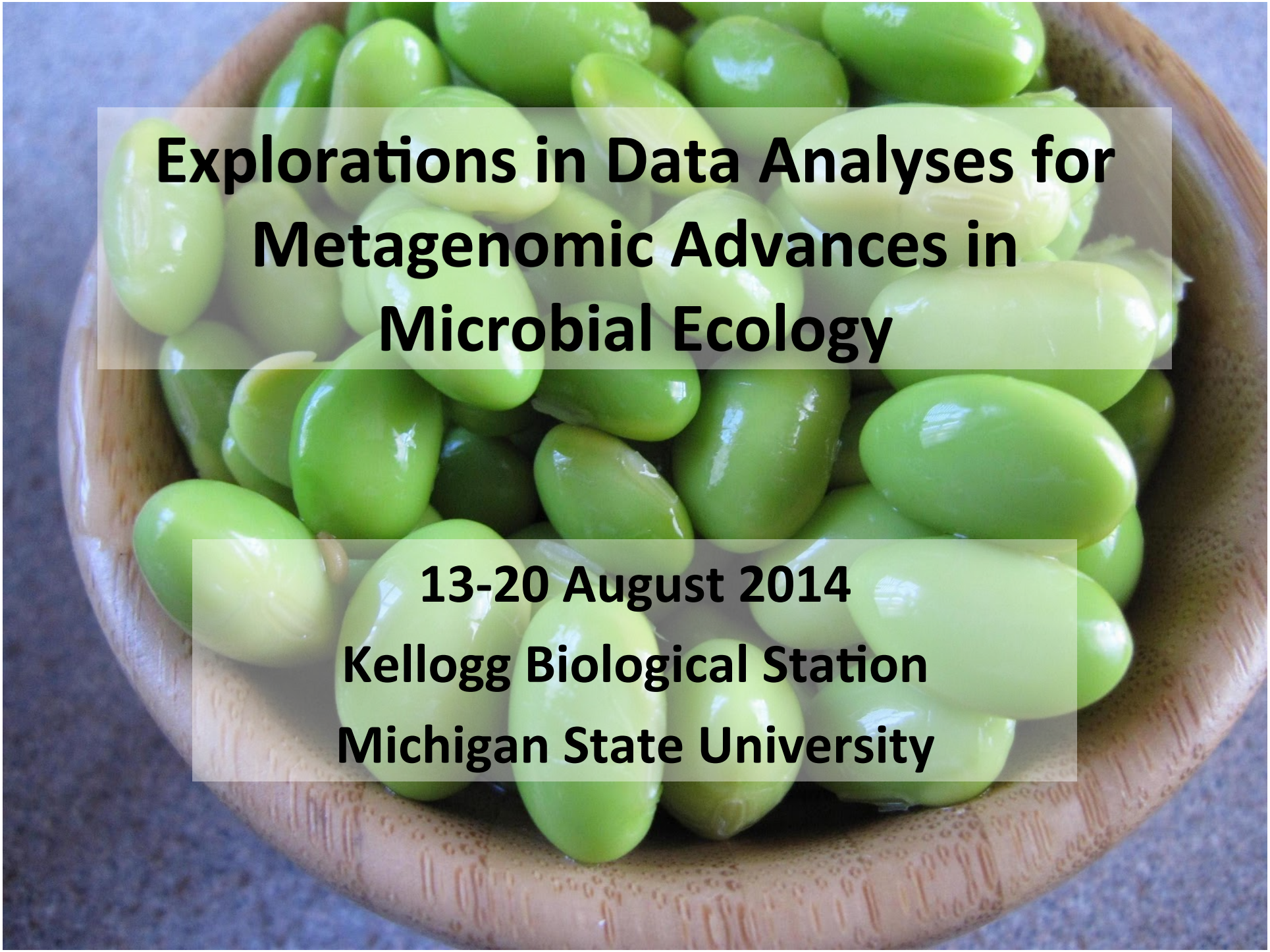


Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology

13-20 August 2014

Kellogg Biological Station

Michigan State University

Review

- **Beta diversity** describes *comparative diversity* between communities or changes in a community

- **Resemblance metrics** quantify pair-wise differences between communities, and can include information about OTU abundances & phylogenetic representation

- An **resemblance matrix** is a square, sample-by-sample table of all pairs of resemblance

Questions about beta diversity?

Overview Day 4

- Using categorical and quantitative variables to explore community patters
- Non-parametric tests
  - Hypothesis tests for clusters: dispersion versus centroids
- Visualization: ordinations, heatmaps, clusters
- Linking quantitative metadata data to community patterns

Analysis of beta-diversity is informed by:

- Associated environmental/quantitative variables*
  - Examples: 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"*

Clusters & Gradients

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

Also called: factors, qualitative variables

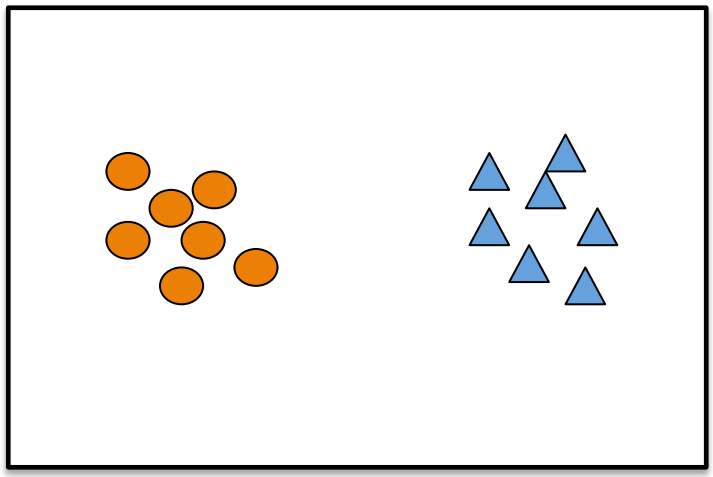
Gradients = Do communities change with known environmental changes? (*e.g.,* over time?)

Also called: continuous, quantitative, vector variables

Clusters: Testing for effects of treatment

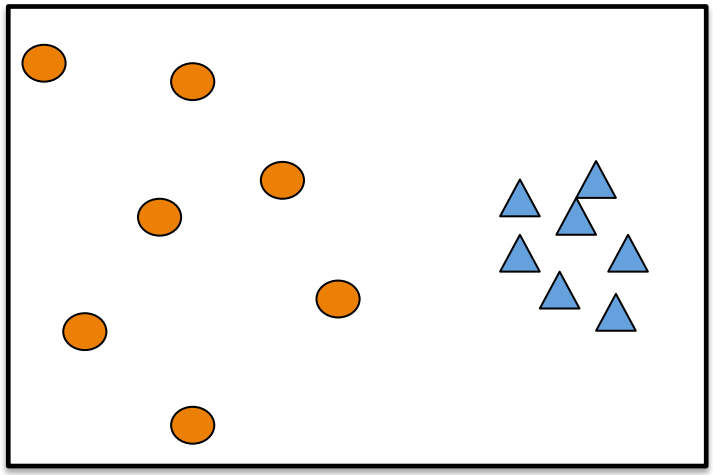
Permutation-based analyses to test hypotheses about group differences in
CENTROID (mean) or **DISPERSION (spread, variability)**

A. Different centroid, same spread

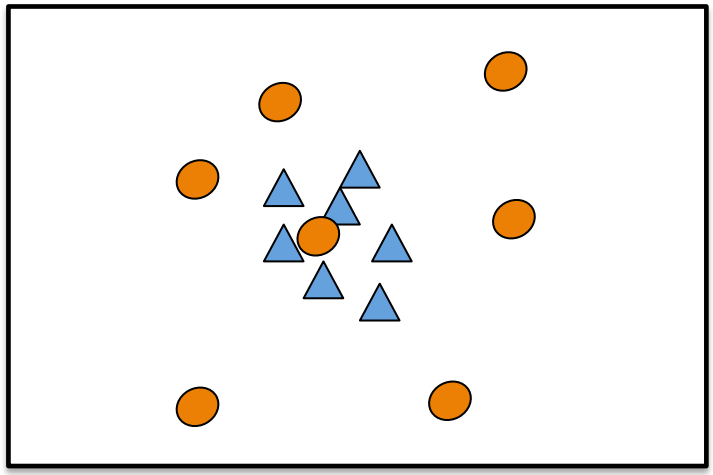


Test name	Centroid (mean)	Spread (variability)
PERMANOVA	X	X
MRPP	X	X
ANOSIM	X	X
PERMDISP		X

B. Different centroid, different spread



C. Same centroid, different spread

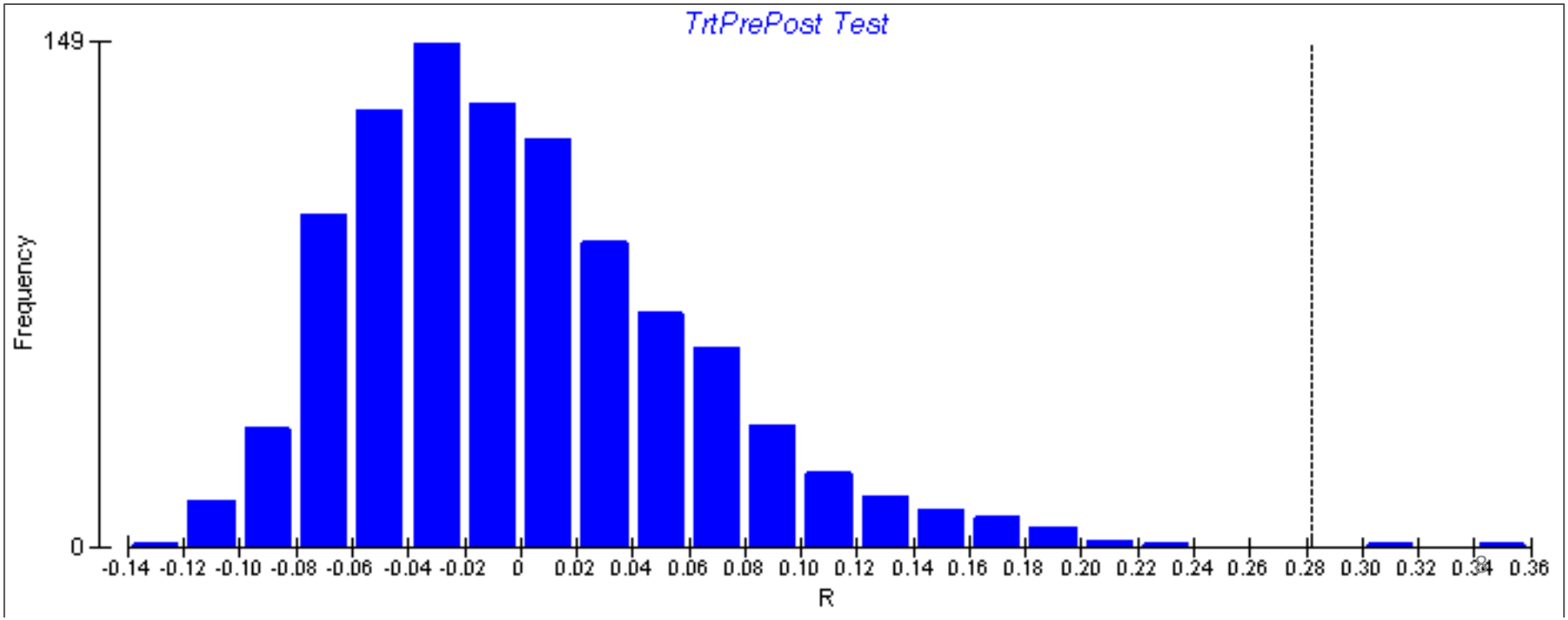


Non-parametric hypothesis tests

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.



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

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

TABLE 1 Four hypothesis tests for differences in community structure (mean or variation) among prespray and postspray untreated and treated soil microbial communities, assessed using each of four resemblance metrics[a]

		Differences in mean or variation in community structure			Differences in variation in community structure (PERMDISP)
Sample group	Metric	PERMANOVA	MRPP	ANOSIM	
Culture independent	Bray-Curtis	n.s. ($P = 0.070$)	n.s. ($P = 0.058$)	n.s. ($P = 0.166$)	n.s. ($P = 0.434$)
	Modified Gower log10	n.s. ($P = 0.082$)	n.s. ($P = 0.087$)	n.s. ($P = 0.176$)	n.s. ($P = 0.127$)
	Morisita-Horn	n.s. ($P = 0.233$)	n.s. ($P = 0.177$)	n.s. (0.438)	n.s. ($P = 0.388$)
	Sørenson	**$R^2 = 0.131, P = 0.054$**	n.s. ($P = 0.079$)	n.s. ($P = 0.136$)	n.s. ($P = 0.535$)
StrR cultured	Bray-Curtis	**$R^2 = 0.14, P = 0.004$**	**deltaA = 0.6371, $P = 0.008$**	**$R = 0.12, P = 0.011$**	n.s. ($P = 0.284$)
	Modified Gower log10	**$R^2 = 0.12, P = 0.001$**	**deltaA = 1.15, $P = 0.001$**	**$R = 0.10, P = 0.002$**	n.s. ($P = 0.144$)
	Morisita-Horn	n.s. ($P = 0.096$)	n.s. ($P = 0.105$)	n.s. ($P = 0.094$)	n.s. ($P = 0.057$)
	Sørenson	**$R^2 = 0.13, P = 0.001$**	**deltaA = 0.6625, $P = 0.001$**	**$R = 0.15, P = 0.002$**	n.s. ($P = 0.155$)
Cultured	Bray-Curtis	**$R^2 = 0.13, P = 0.024$**	**deltaA = 0.55, $P = 0.02$**	**$R = 0.102, P = 0.016$**	n.s. ($P = 0.276$)
	Modified Gower log10	**$R^2 = 0.12, P = 0.001$**	**deltaA = 1.162, $P = 0.001$**	**$R = 0.12, P = 0.001$**	n.s. ($P = 0.766$)
	Morisita-Horn	n.s. ($P = 0.257$)	n.s. ($P = 0.164$)	n.s. ($P = 0.236$)	n.s. ($P = 0.367$)
	Sørenson	**$R^2 = 0.12, P = 0.001$**	**deltaA = 0.670, $P = 0.001$**	**$R = 0.186, P = 0.001$**	n.s. ($P = 0.617$)

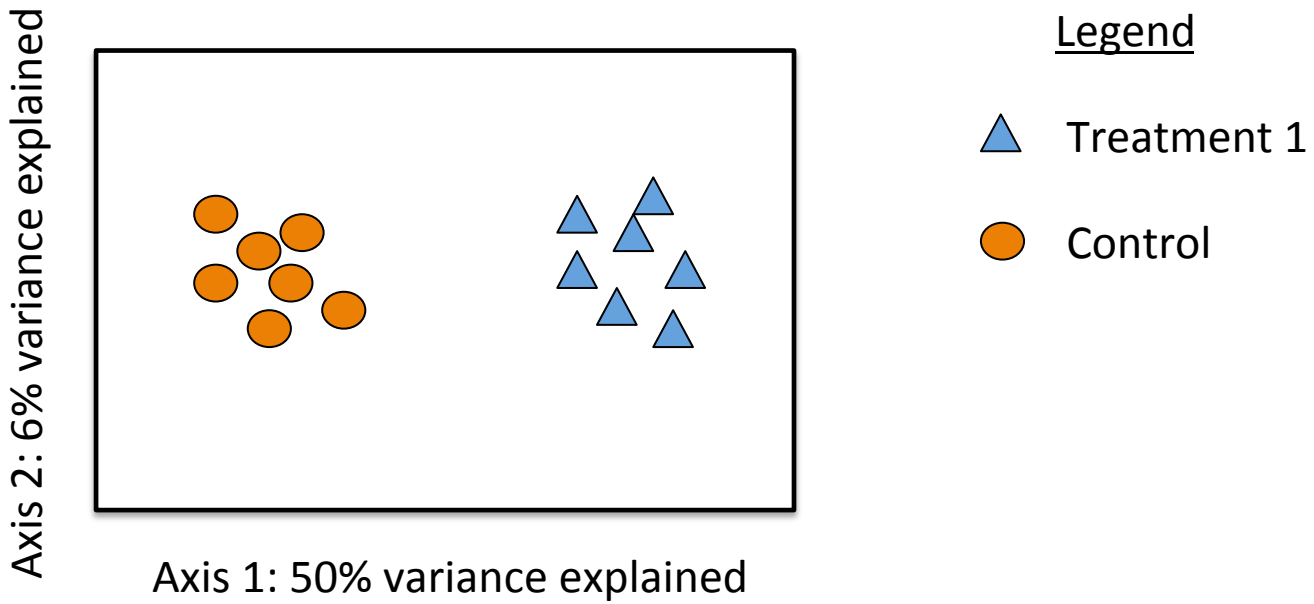
[a] Significant test results are shown in bold. n.s., not significant ($P > 0.05$); PERMANOVA, permuted analysis of variance; MRPP, multiple-response permutation procedure; ANOSIM, analysis of similarity; PERMDISP, permuted analysis of multivariate dispersion.

Shade *et al*. 2013 AEM

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.

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.

Cluster analysis

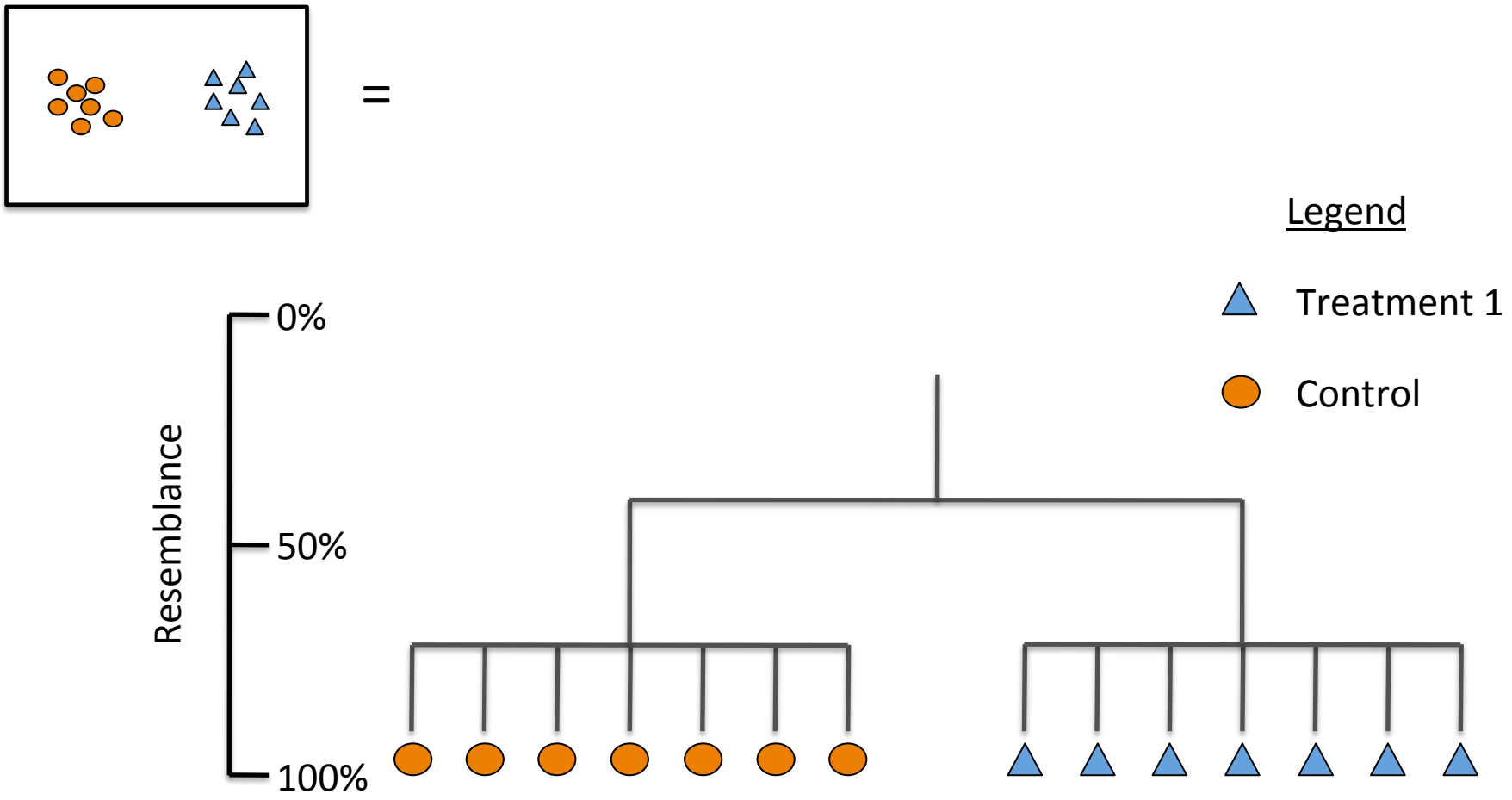
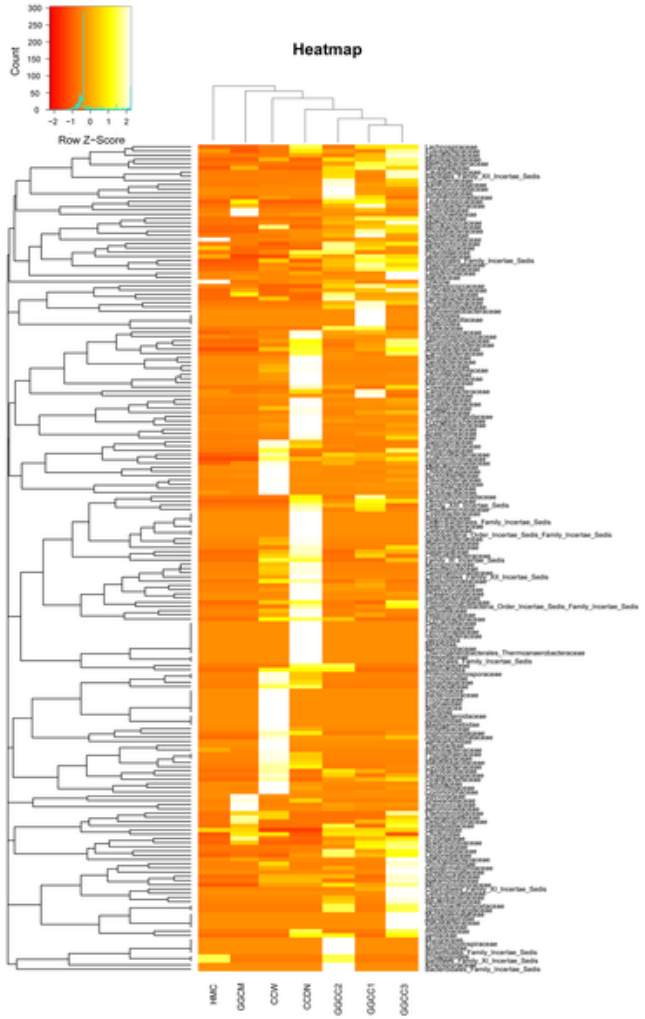


Figure 6. Bacterial distribution among the seven samples.



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

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