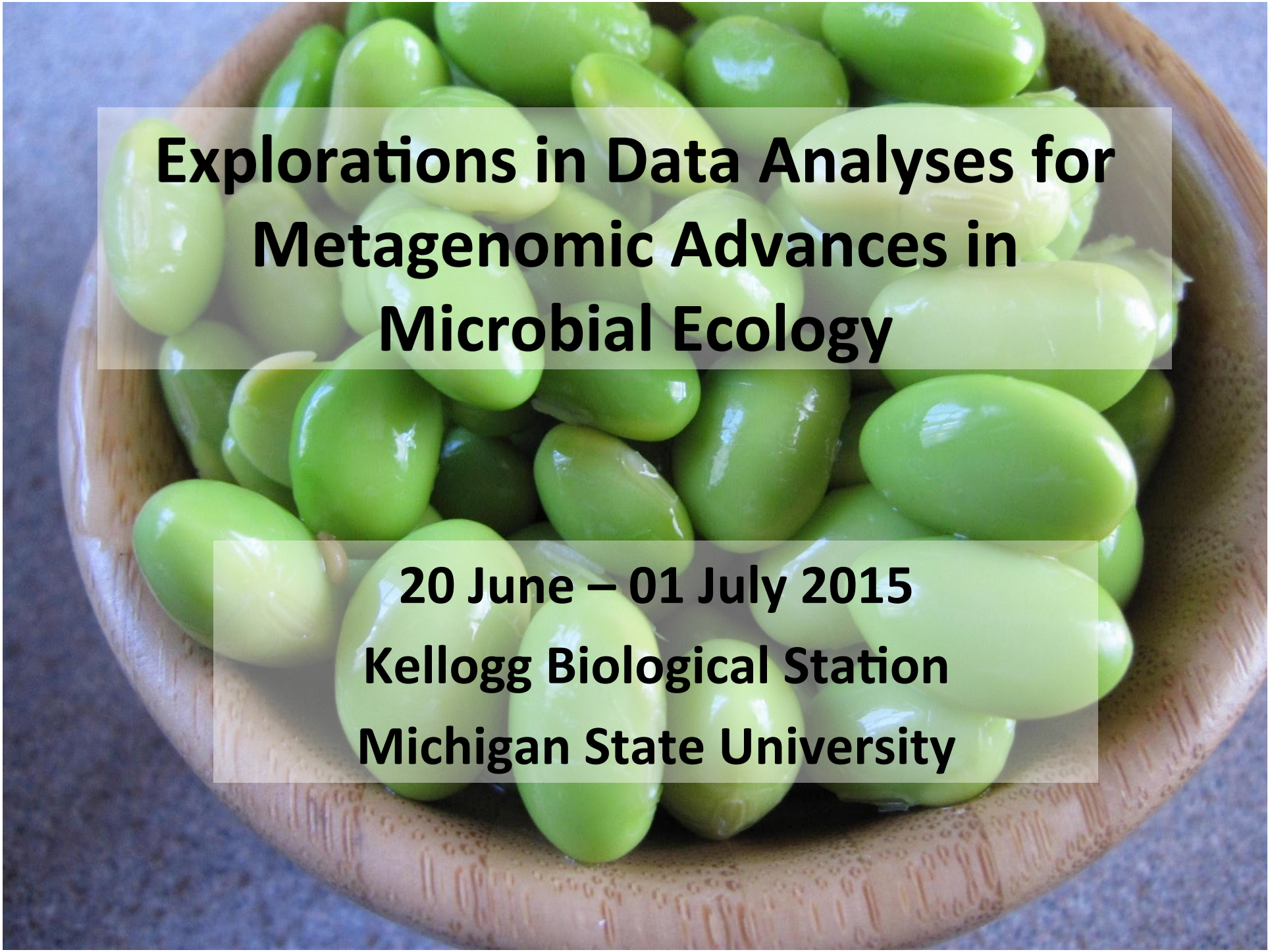
WELCOME BACK!

- Good morning! We missed you!

- Make a name tent (back table)
- Grab a red and a blue sticky note
- **Re-start your stopped EC2 instance, copy new public DNS**
- **Connect to the instance (hint - use 'ssh')**

- Think about writing a blog post, and check out Mark's microbial analysis haikus on twitter.

- Get ready to QIIME it up!



Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology

20 June – 01 July 2015

Kellogg Biological Station

Michigan State University

Conceptual Review

- Microbial communities are local assemblages of microorganisms that interact with each other or with their environment

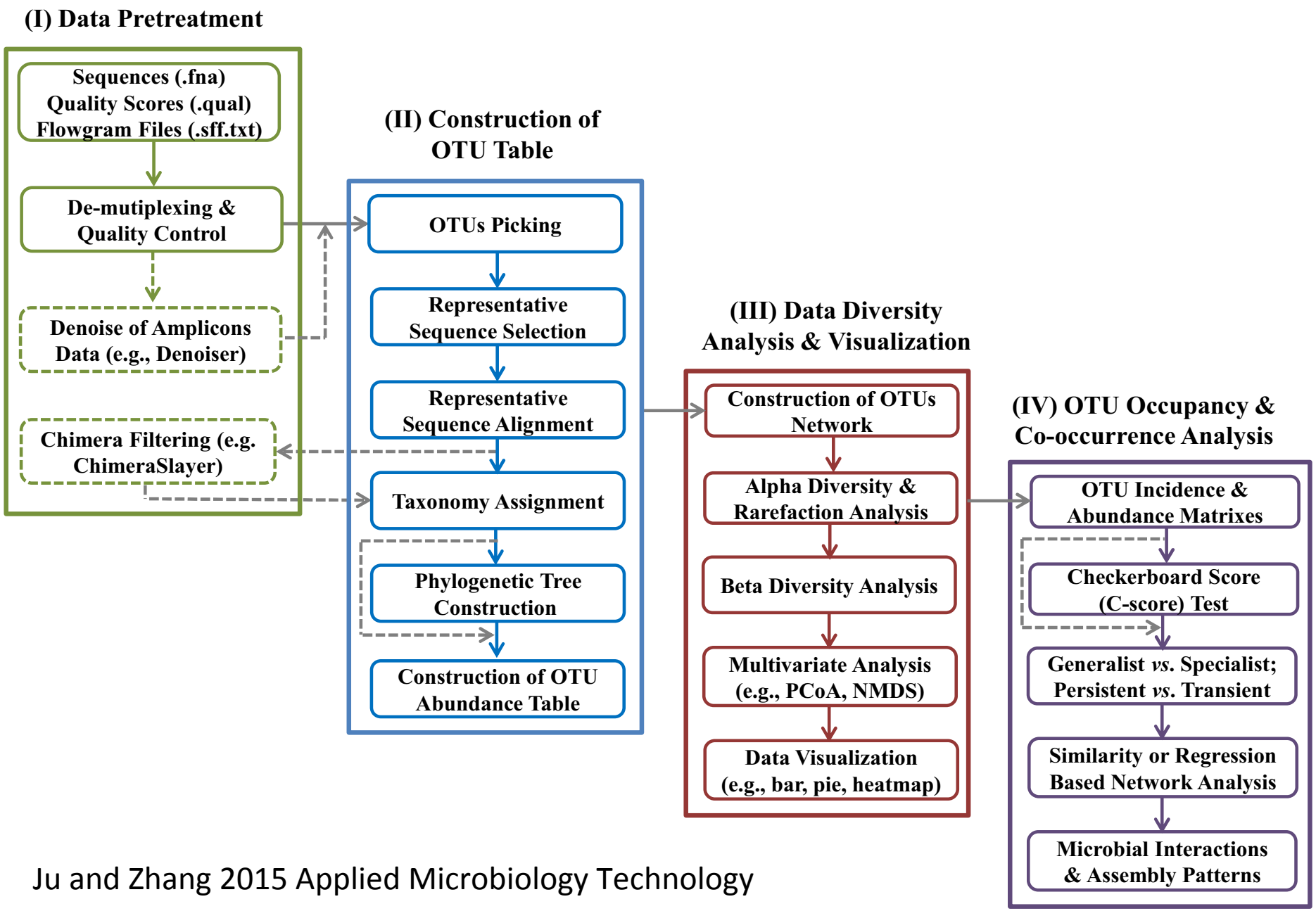
- "OTUs" – operational taxonomic units – as a microbial species definition (97% identity 16S rRNA)

- There are lots of choices for analyzing high-throughput sequencing data – choice of seq. platform, primers, variable region; depth of sampling; analysis methods – **most of these can be reasonably informed by the community of interest, the scientific question, and the experiment design**

Practical Review - yesterday

- Navigating in the Shell
- Starting an EC2 instance
- Connecting to the instance and transferring files
- Installing software on an instance
- Using FastQC to assess raw Illumina quality

Questions from yesterday?



Ju and Zhang 2015 Applied Microbiology Technology

Intro to amplicon sequence analysis

- Merging paired-end Illumina Reads – pandaseq

- QIIME:
  - Clustering sequences by 97% identity
  - Picking representative sequences
  - assigning taxonomy to sequences
  - Building and alignment and phylogenetic tree
  - Building an “even” OTU table : equal No. sequences per sample so that comparisons can be made
  - Calculating within-sample (alpha) diversity

What does a community look like, data-style?

- "OTU table" – the original
- .Biom table – more concise & faster computing for extra large datasets
  - Newer formats : "biom2"
  - See McDonald et al. 2012. "The Biological Observation Matrix…" *GigaScience*.

Kinds of Community matrix/ OTU Table

Raw
Straight up counts*

	Soil 1	Soil 2	Soil 3
OTU 1	0	3000	23
OTU 2	1	5	5
OTU 3	20	100	100

**note: data must be subsampled to an even sequencing effort!*

Relative
Percent or proportion

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

Binary
Presence/absence

	Soil 1	Soil 2	Soil 3
OTU 1	0	1	1
OTU 2	1	1	1
OTU 3	1	1	1

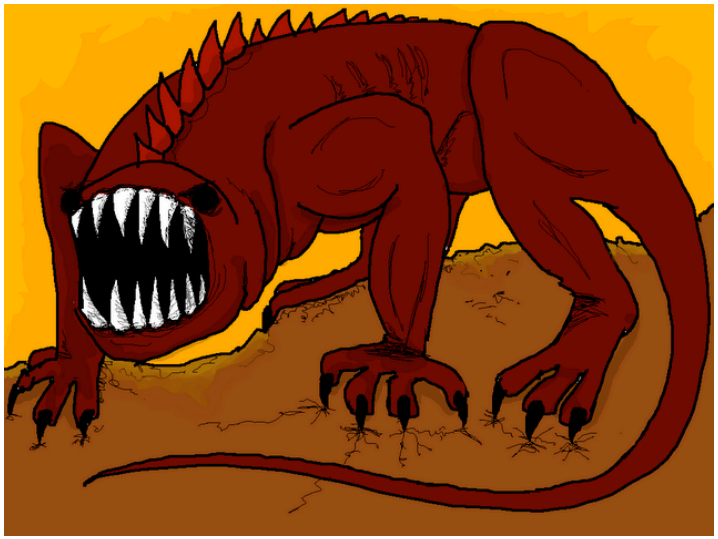
Information in an OTU table

- Number of occurrences (per sample and for the whole dataset)
- Total no. OTUs observed in the dataset
- Average abundance of OTUs
- Richness (no. OTUs per sample, mean, max, min, range)
- Number of singletons (OTUs detected only once in a dataset)
- Calculate: Diversity, Evenness (equitability of OTU abundances, including rarity and dominance)
- Number of samples (communities) in your dataset

- Dimensions of an OTU table: rows (taxa) x columns (samples/ communities)

Common features of microbial OTU tables

- Redundant: more than one taxa has the exact same pattern
- Unknown underlying distribution
- Contain many “zeros”
- Many samples and OTUs; computationally large



(A beast, hyperboleandahalf.blogspot.com)

Biom formatted OTU tables

- .biom format

Link:

http://biom-format.org

This is all changing very often!! Biom formats are constantly improved, keep up with when changes are anticipated

A dense representation of an OTU table:

OTU ID	PC.354	PC.355	PC.356
OTU0	0	0	4
OTU1	6	0	0
OTU2	1	0	7
OTU3	0	0	3

Traditional OTU table - microbial communities have lots of 0's

A sparse representation of an OTU table:

```
PC.354 OTU1 6
PC.354 OTU2 1
PC.356 OTU0 4
PC.356 OTU2 7
PC.356 OTU3 3
```

.biom formatted – only list present taxa

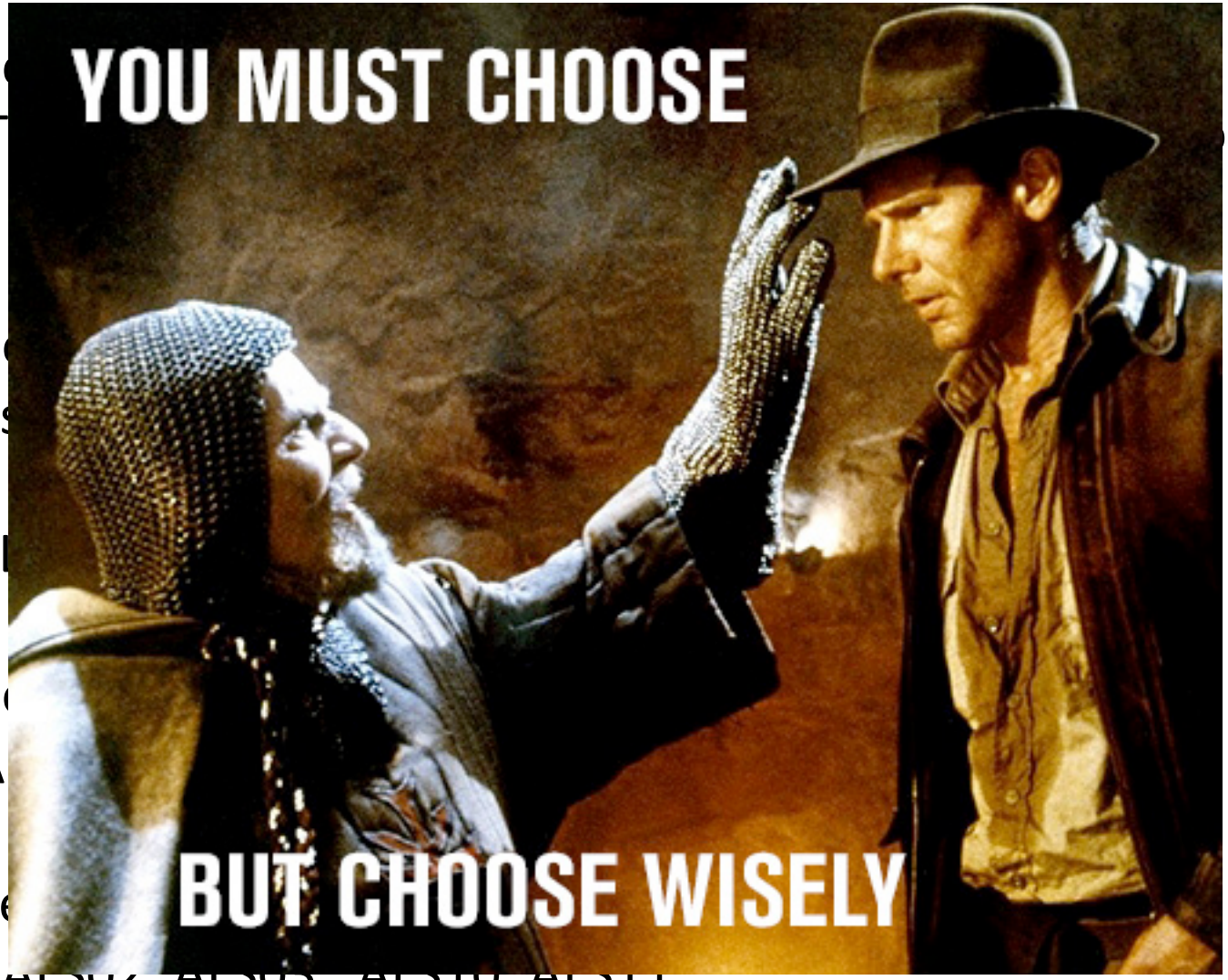
Naming Conventions

Exampl
20_A_T
rep1)

Exampl
Ashley'
A
Ashley

Exampl
ALS1, A

Improve
ALS01, ALS02, ALS03…ALS10, ALS11



Our samples, e.g.
C01_05102014_R1_D01

C01 – Centralia core site 1
Date 05102014 – 05 Oct 2014
R1 – core 1 (there were sometimes multiple cores from the same site)
D01 – DNA extraction replicate 1 D01- DNA extraction rep 1

…
F – forward read; R = Reverse read

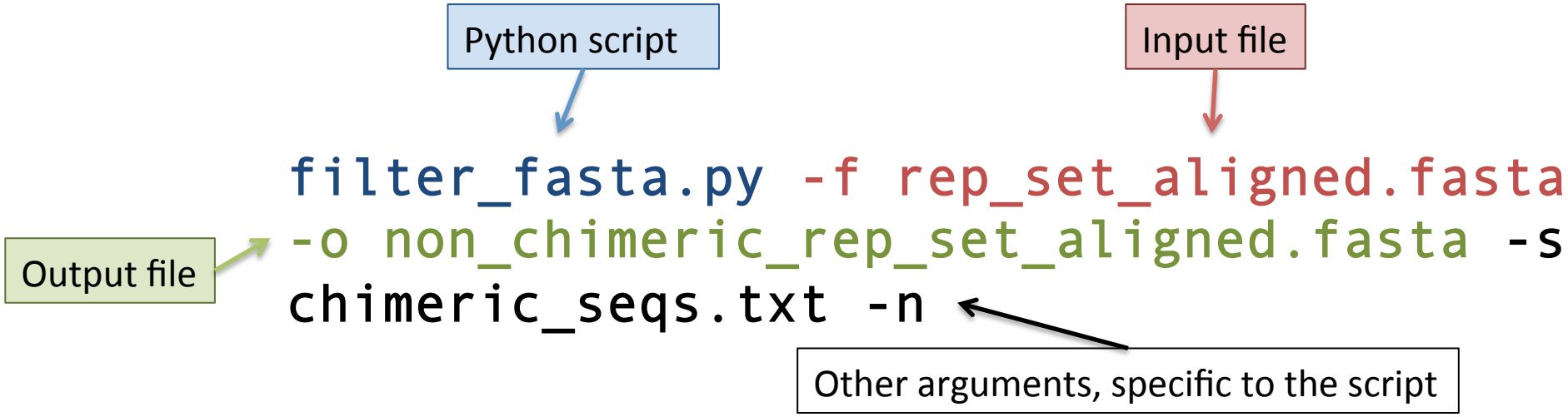
Approaches to Picking OTUs

- Reference based : percent identity to defined taxa populating in a reference database
  - Pros: you know the taxa are "real"!
  - Cons: Weird environments don't have many representatives in databases, only as good as your database

- De novo : percent identity to other sequences in the dataset; taxonomic assignment to the OTUs happens afterwards
  - Pros: Good for weird environments with low representation in databases
  - Cons: Computationally expensive, "greedy" algorithms can artificially inflate diversity

- Open reference : cluster against a reference db first, and anything that doesn't hit gets clustered de novo
  - Best of both worlds? Now can optimized so that new de novo OTUs are added to the original database and used subsequently in "reference" clustering
  - See Rideout et al. 2014 PeerJ

A look at **python** syntax & common arguments in QIIME



Other common QIIME arguments

-m analysis method, metric (sometimes map file)
-t tree file
-a alignment template file
-v verbose = good for troubleshooting
-h help
-f force overwrite of an existing directory

Tutorial: What we're about to do

- Practice subsampling a dataset to make it managable for workflow development
- Merge paired end reads with PANDASeq; move the sequences into QIIME
- Pick OTUs open reference - includes:
  - Quality control/ chimera check
  - Cluster at 97% identity
  - Pick representative sequence for the whole OTU
  - Assign taxonomy to the rep. sequence
  - Make an alignment of the rep. sequence
  - Build a tree from the alignment
  - Made OTU tables (biom + classic): **make_otu_table.py**
- Rarefy to an equal sequencing depth
- Calculated & visualized alpha diversity

Let’s analysis!

Analysis is hard, and it is completely normal to struggle.

- Workflow diagram
- Pace – just fine
- Like details of all the flags/ options

- Directory angst – where am I?
- Why aren't tutorials perfect? – too many changes
- Where do we modify the pipeline for our own datasets?
- Pace – too fast

Diversity part 1

- *What is diversity anyway?*
- The advantage of phylogenetic information
- Rarefaction
- What does a community look like, data-style?
- A note on the .biom v. OTU table format
- A note on naming conventions
- Intro to python scripting

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 <u>one sample</u>.

Whittaker introduces alpha, beta, gamma diversity (1972)

TAXON 21 (2/3): 213-251. MAY 1972

EVOLUTION AND MEASUREMENT OF SPECIES DIVERSITY*

*R. H. Whittaker***

Summary

Given a resource gradient (e.g. light intensity, prey size) in a community, species evolve to use different parts of this gradient; competition between them is thereby reduced. Species relationships in the community may be conceived in terms of a multidimensional coordinate system, the axes of which are the various resource gradients (and other aspects of species relationships to space, time, and one another in the community). This coordinate system defines a hyperspace, and the range of the space that a given species occupies is its niche hypervolume, as an abstract characterization of its intra-community position, or niche. Species evolve toward difference in niche, and consequently toward difference in location of their hypervolumes in the niche hyperspace. Through evolutionary time additional species can fit into the community in niche hypervolumes different from those of other species, and the niche hyperspace can become increasingly complex. Its complexity relates to the community's richness in species, its alpha diversity.

The confusion continues… for decades

Ecology Letters 2011

Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist

Abstract
A recent increase in studies of β diversity has yielded a confusing array of concepts, measures and methods. Here, we provide a roadmap of the most widely used and ecologically relevant approaches for analysis through a series of mission statements. We distinguish two types of β diversity: directional turnover along a gradient vs. non-directional variation. Different measures emphasize different properties of ecological data. Such properties include the degree of emphasis on presence/absence vs. relative abundance information and the inclusion vs. exclusion of joint absences. Judicious use of multiple measures in concert can uncover the underlying nature of patterns in β diversity for a given dataset. A case study of Indonesian coral assemblages shows the utility of a multi-faceted approach. We advocate careful consideration of relevant questions, matched by appropriate analyses. The rigorous application of null models will also help to reveal potential processes driving observed patterns in β diversity.

Marti J. Anderson,[1]* Thomas O. Crist,[2] Jonathan M. Chase,[3] Mark Vellend,[4] Brian D. Inouye,[5] Amy L. Freestone,[6] Nathan J. Sanders, Howard V. Cornell,[8] Liza S. Comita,[9] Kendi F. Davies,[10] Susan P. Harrison,[8] Nathan J. B. Kraft,[11] James C. Stegen[12] and Nathan G. Swenson[13]

Abstract There is a genuine need for consensus on a clear terminology in the study of species diversity given that the nature of the components of diversity is the subject of an ongoing debate and may be the key to understanding changes in ecosystem processes. A recent and thought-provoking paper (Jurasinski et al. Oecologia 159:15–26, 2009) draws attention to the lack of precision with which the terms alpha, beta, and gamma diversity are used and proposes three new terms in their place. While this valuable effort may improve our understanding of the different facets of species diversity, it still leaves us far from achieving a consistent terminology. As such, the conceptual contribution of these authors is limited and does little to elucidate the facets of species diversity. It is, however, a good starting point for an in-depth review of the available concepts and methods.

Keywords Alpha diversity · Beta diversity · Gamma diversity · Species richness · Turnover

A consistent terminology for quantifying species diversity?

Claudia E. Moreno · Pilar Rodríguez

Oecologia 2010

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

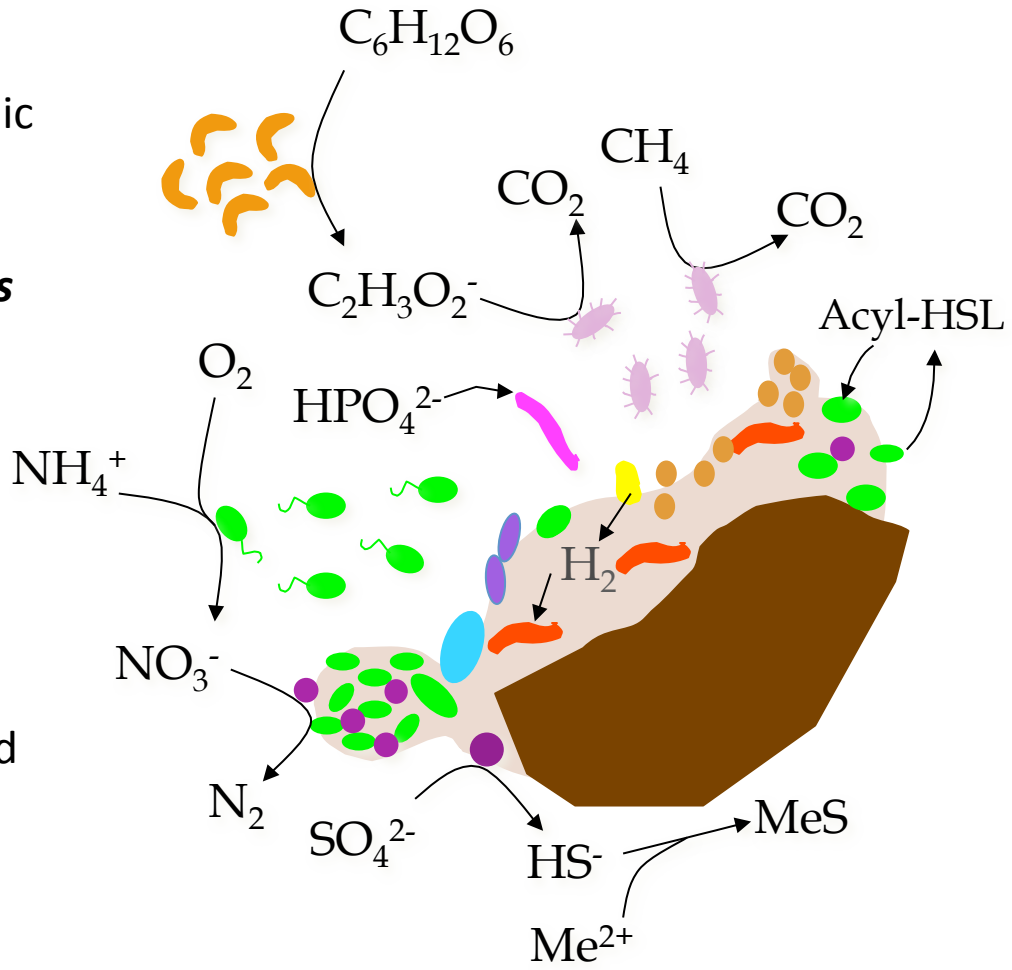
Information about the community that we can glean from metagenomic sequencing

•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?

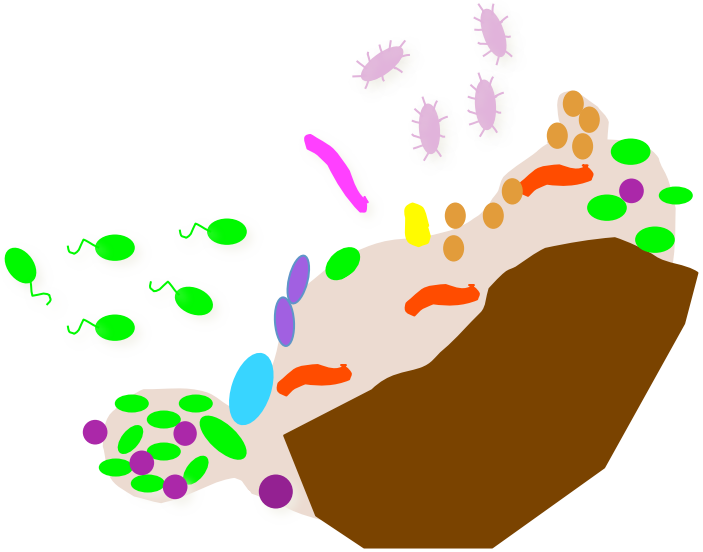


Richness

OTU

Richness: How many OTUs?

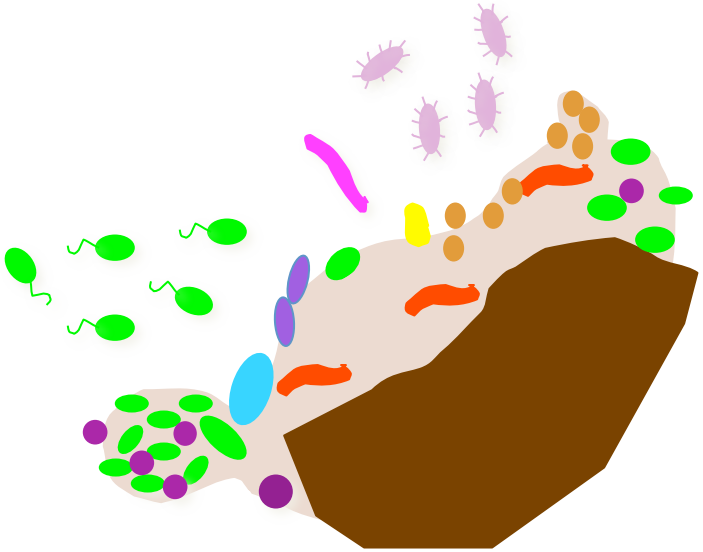
Richness = 11 OTUs



Evenness

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

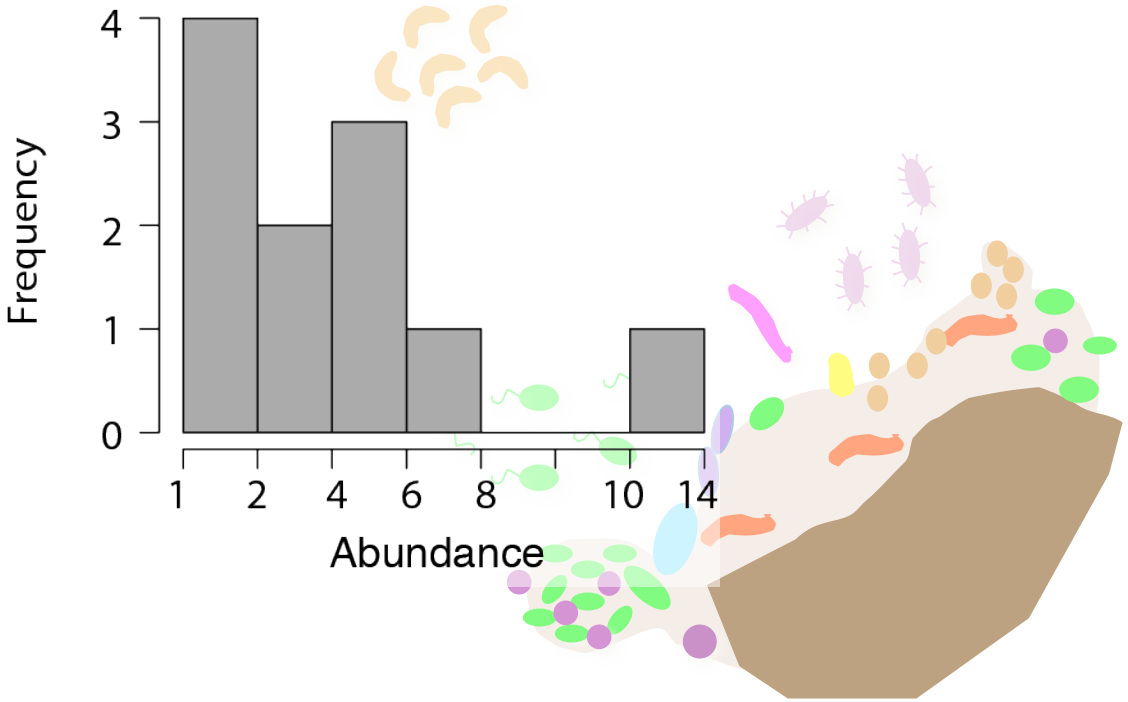
OTU	**Count**: No. seq, no. individuals (e.g., FISH), biomass, etc.
	6
	1
	4
	8
	1
	5
	3
	13
	5
	1
	2



Evenness

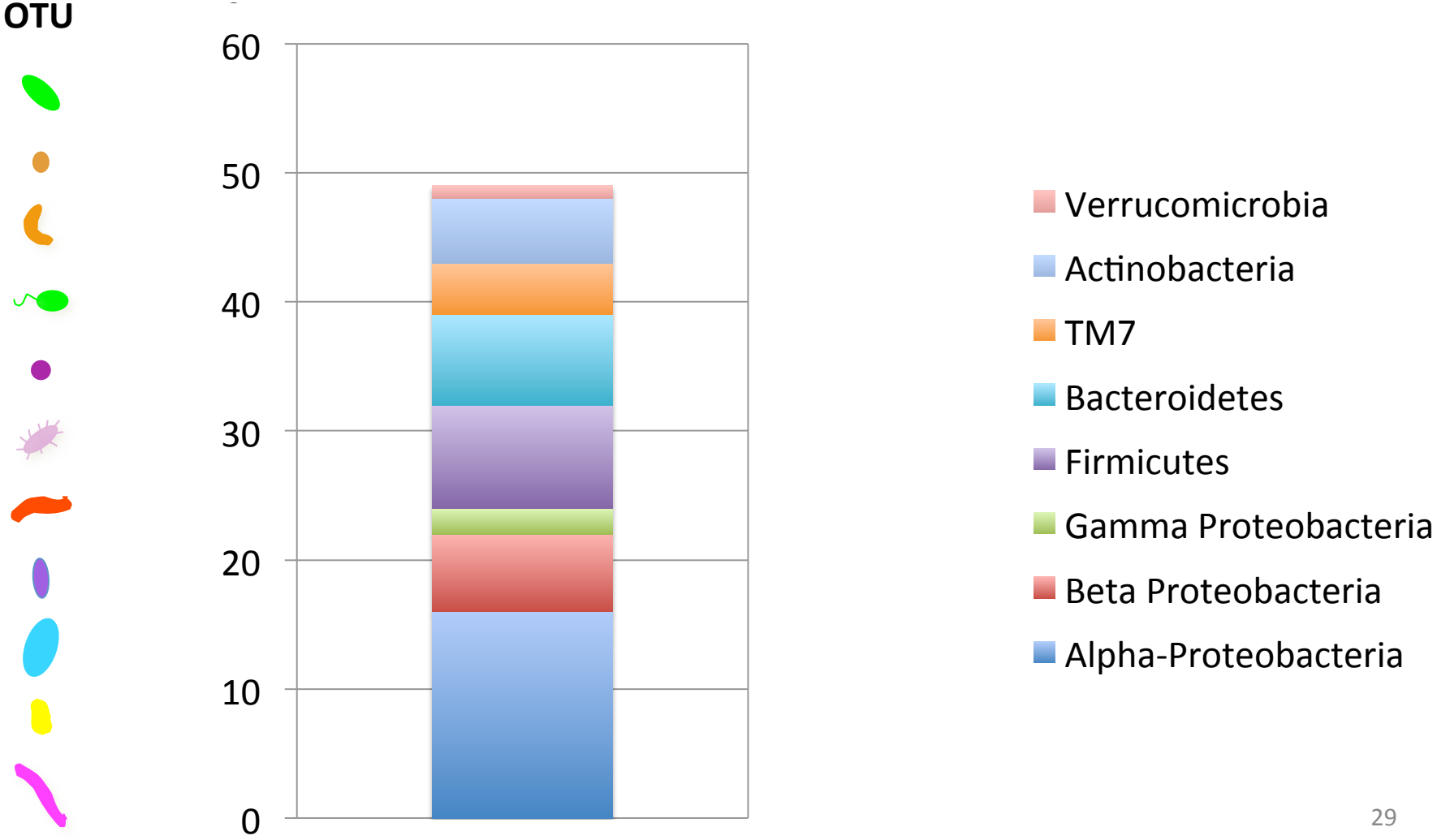
Evenness: What is the distribution of abundances in the community?

OTU	**Count**:
	13
	8
	6
	5
	5
	4
	3
	2
	1
	1
	1



Membership and Composition

Composition: Who is there?

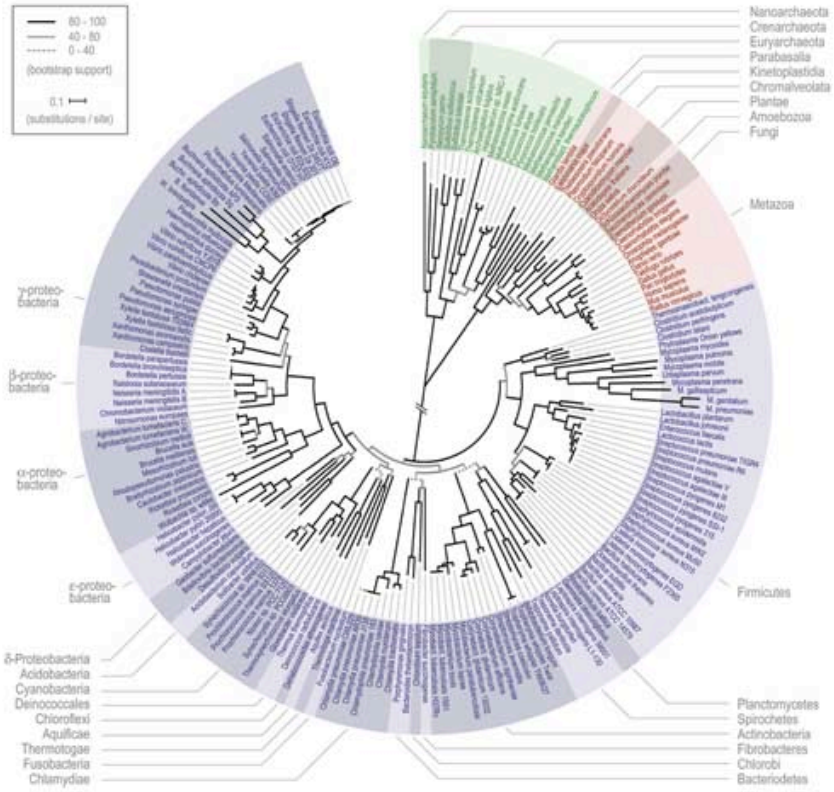


The advantages of phylogeny

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



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.
- 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.

Let’s analysis!

Diversity Part 2

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.

- 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 we're about to do

- Pick OTUs open reference includes:
  - Quality control/ chimera check
  - Cluster at 97% identity
  - Pick representative sequence for the whole OTU
  - Assign taxonomy to the rep. sequence
  - Make an alignment of the rep. sequence
  - Build a tree from the alignment
  - Made OTU tables (biom + classic): **make_otu_table.py**
- Rarefied to an equal sequencing depth: **alpha_rarefaction.py**
- Calculated & visualized alpha diversity: **alpha_diversity.py, summarize_taxa_through_plots.py**

Questions?

Amplicon sequence analysis continued

- In QIIME:
  - Calculating comparative (beta) diversity
  - Visualizing patterns of comparative diversity
  - Hypothesis testing: clusters and gradients
  - Dealing with the QIIME biom table: conversion and nuances

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:** 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

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

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

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

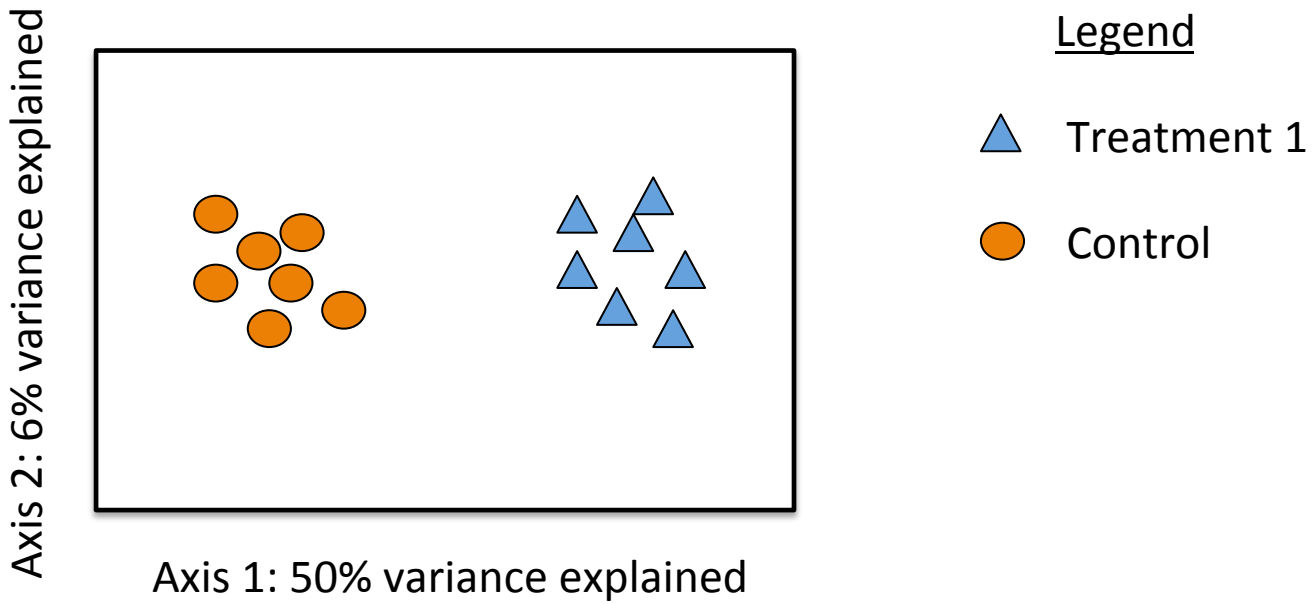
Metric name / *Accounts for*	Sørensen	Bray-Curtis	Weighted UniFrac	Unweighted UniFrac
Composition	X	X	X	X
OTU abundances?		X	X	
Phylogenetic diversity?			X	X

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

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

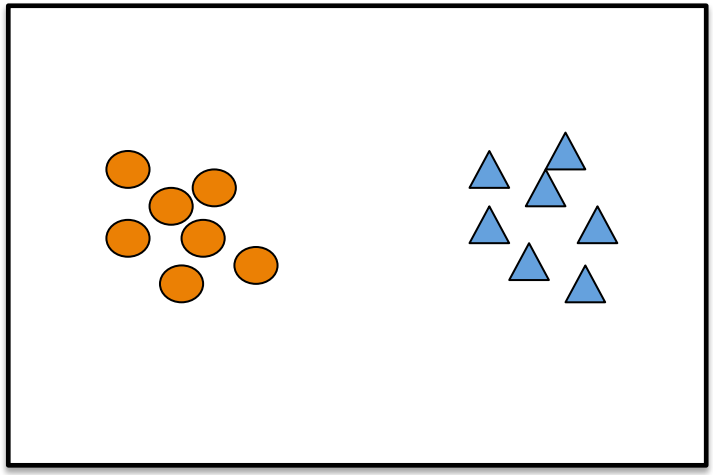
- Non-metric multidimensional scaling (NMDS)
- Principle coordinates analysis (PCoA)
- Correspondence analysis (CA)

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

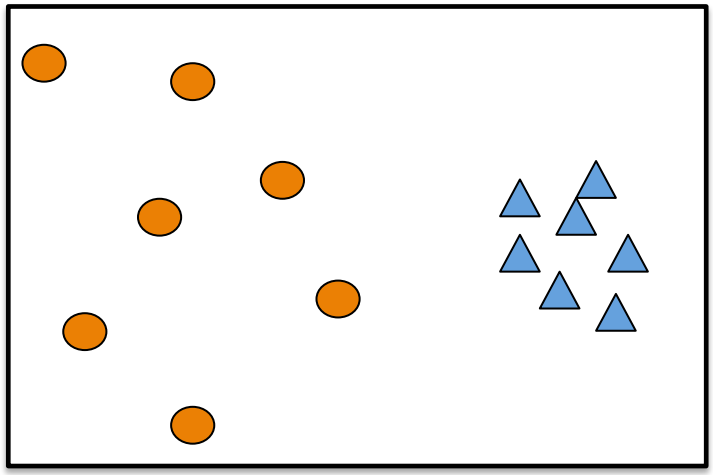
How do we look at ordinations?

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

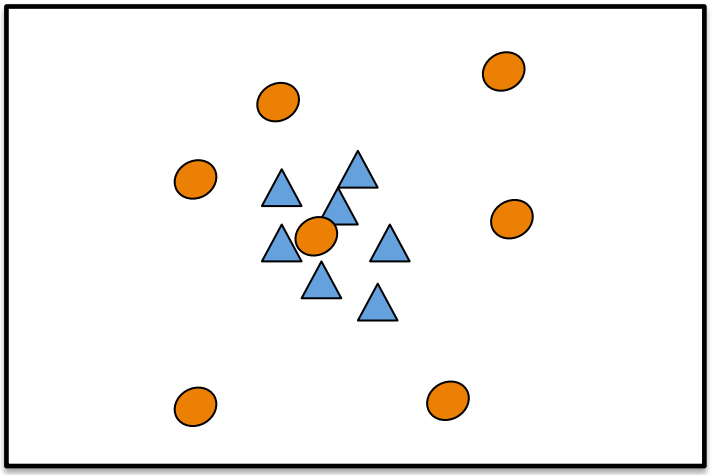
A. Different centroid, same spread



B. Different centroid, different spread

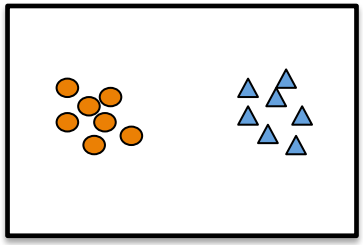


C. Same centroid, different spread

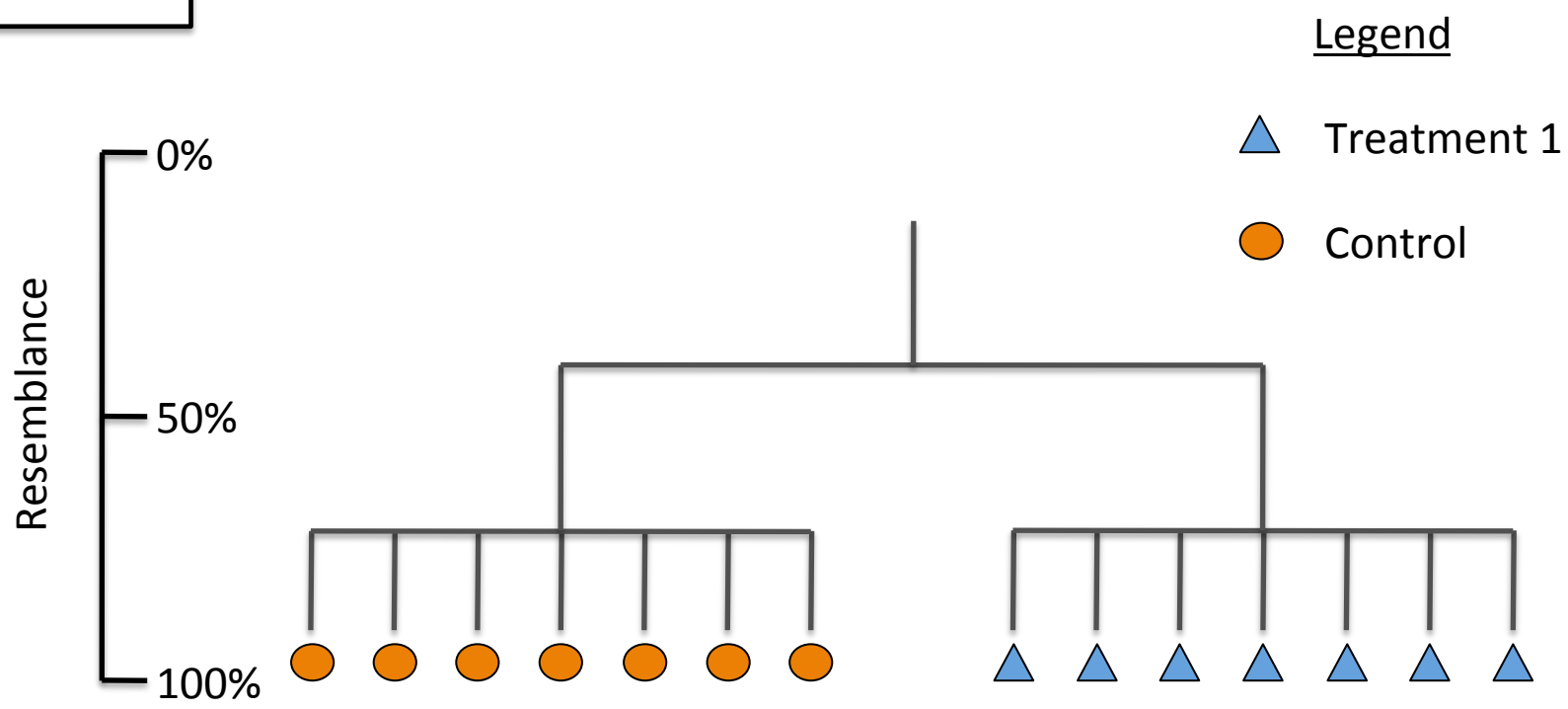


Visualizing communities: dendrograms

A different way of visualizing the same data

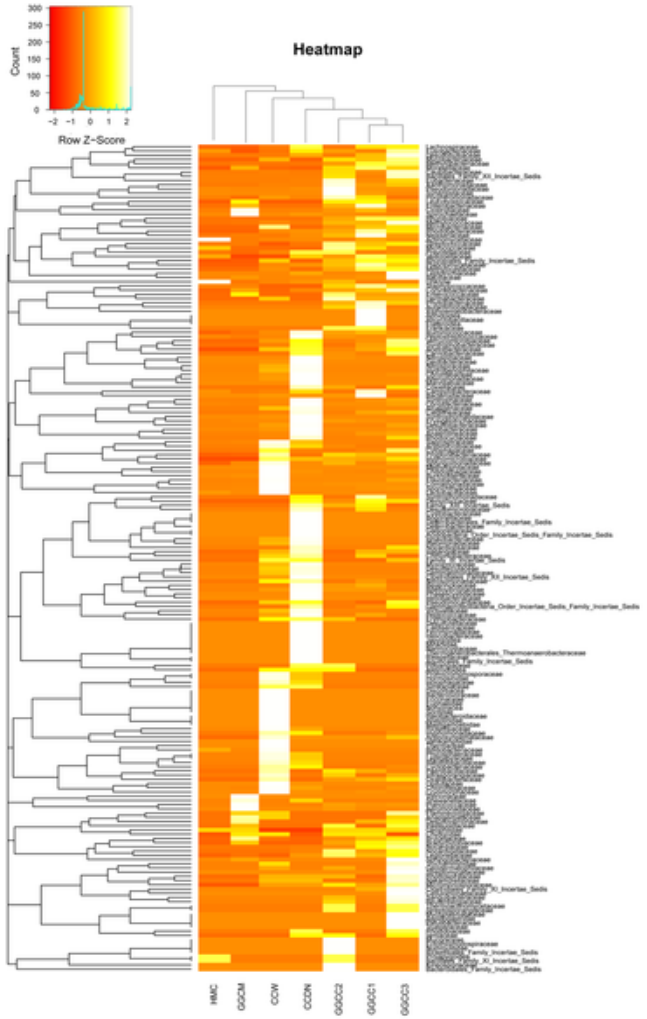


=



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

PLOS | ONE

Discovering patterns: Clusters & Gradients

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

Also called: factors, qualitative variables

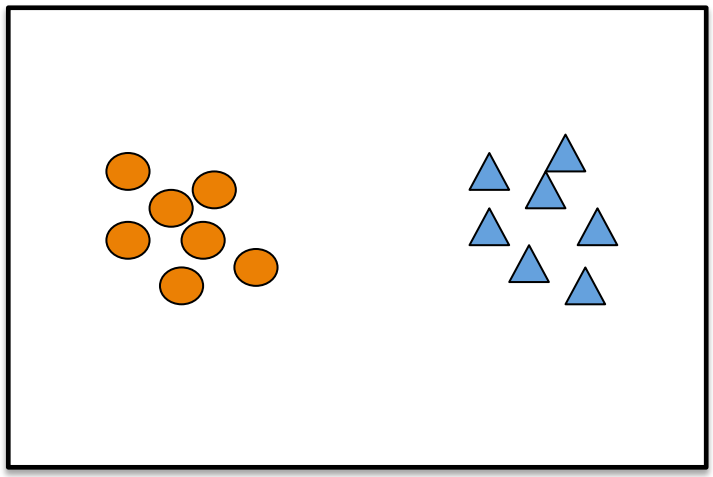
Gradients = Do communities change with known environmental changes? (*e.g.,* over time?)

Also called: continuous, quantitative, vector variables

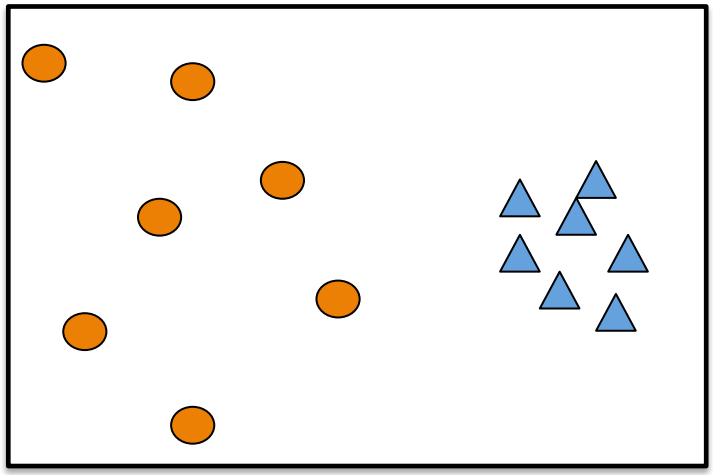
How do we interpret ordinations?

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

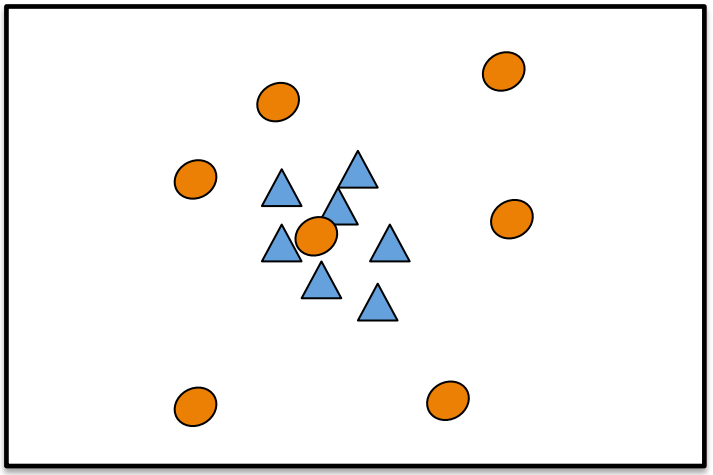
A. Different centroid, same spread



B. Different centroid, different spread



C. Same centroid, different spread

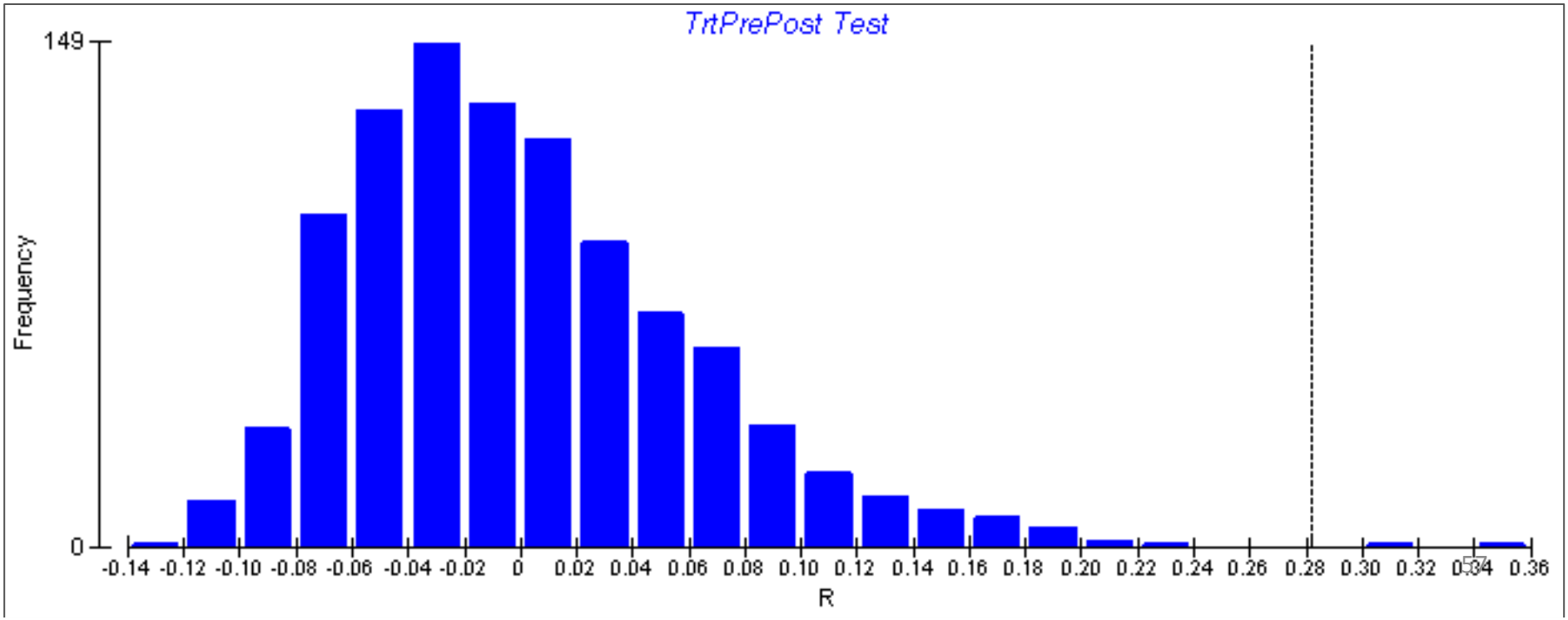


Non-parametric hypothesis tests

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.

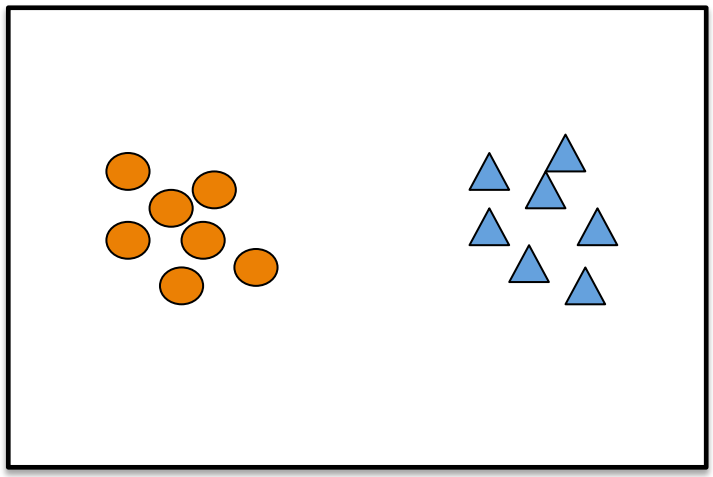
It is important to use these tests for microbial beta diversity, as the assumptions of underlying normal distributions of most parametric tests (e.g., ANOVA) are violated.



Clusters: Testing for differences in *a priori* groups

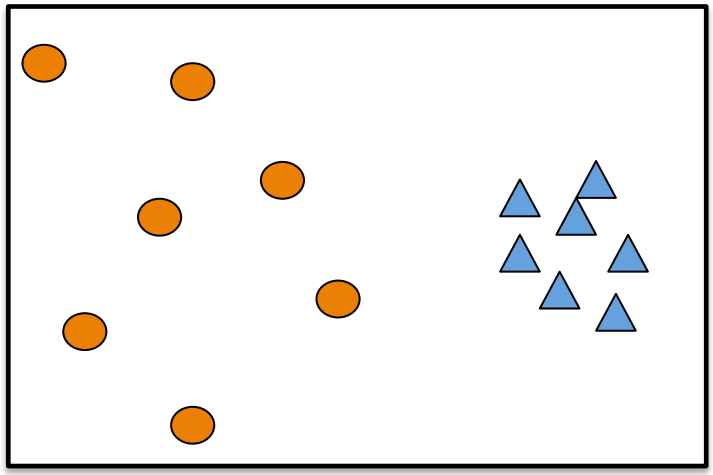
Permutation-based analyses to test hypotheses about group differences in
CENTROID (mean) or **DISPERSION (spread, variability)**

A. Different centroid, same spread

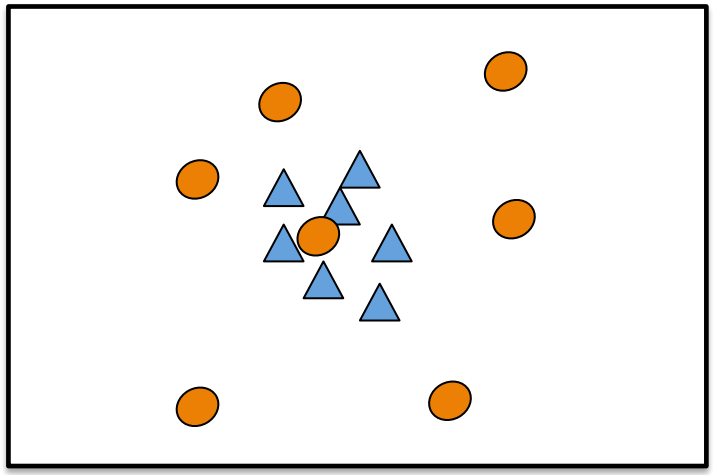


Test name	Centroid (mean)	Spread (variability)
PERMANOVA	X	X
MRPP	X	X
ANOSIM	X	X
PERMDISP		X

B. Different centroid, different spread



C. Same centroid, different spread



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

Gradients: Linking environmental and community data

1. Mantel Test

Community Resemblance

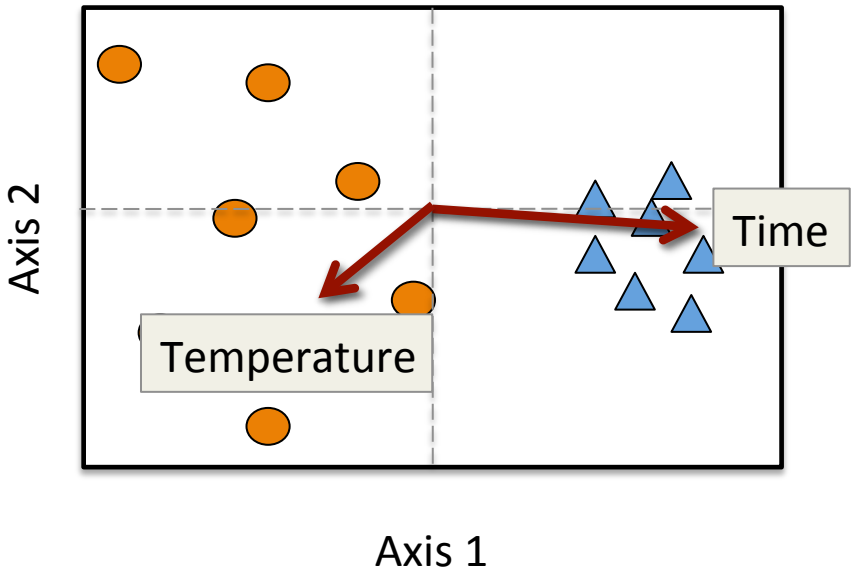
	Caterpillar 1	Caterpillar 2	Caterpillar 3
Caterpillar 1	0		
Caterpillar 2	0.966	0	
Caterpillar 3	0.179	0.787	0

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



Practicalities: Getting data into R from QIIME

Biom to classic OTU table

-See end of QIIME 3 tutorial : biom convert

For importing resemblance tables

1. read.table(header=TRUE, row.names =1; sep =“\t”)
2. as.dist()

Let’s analysis!