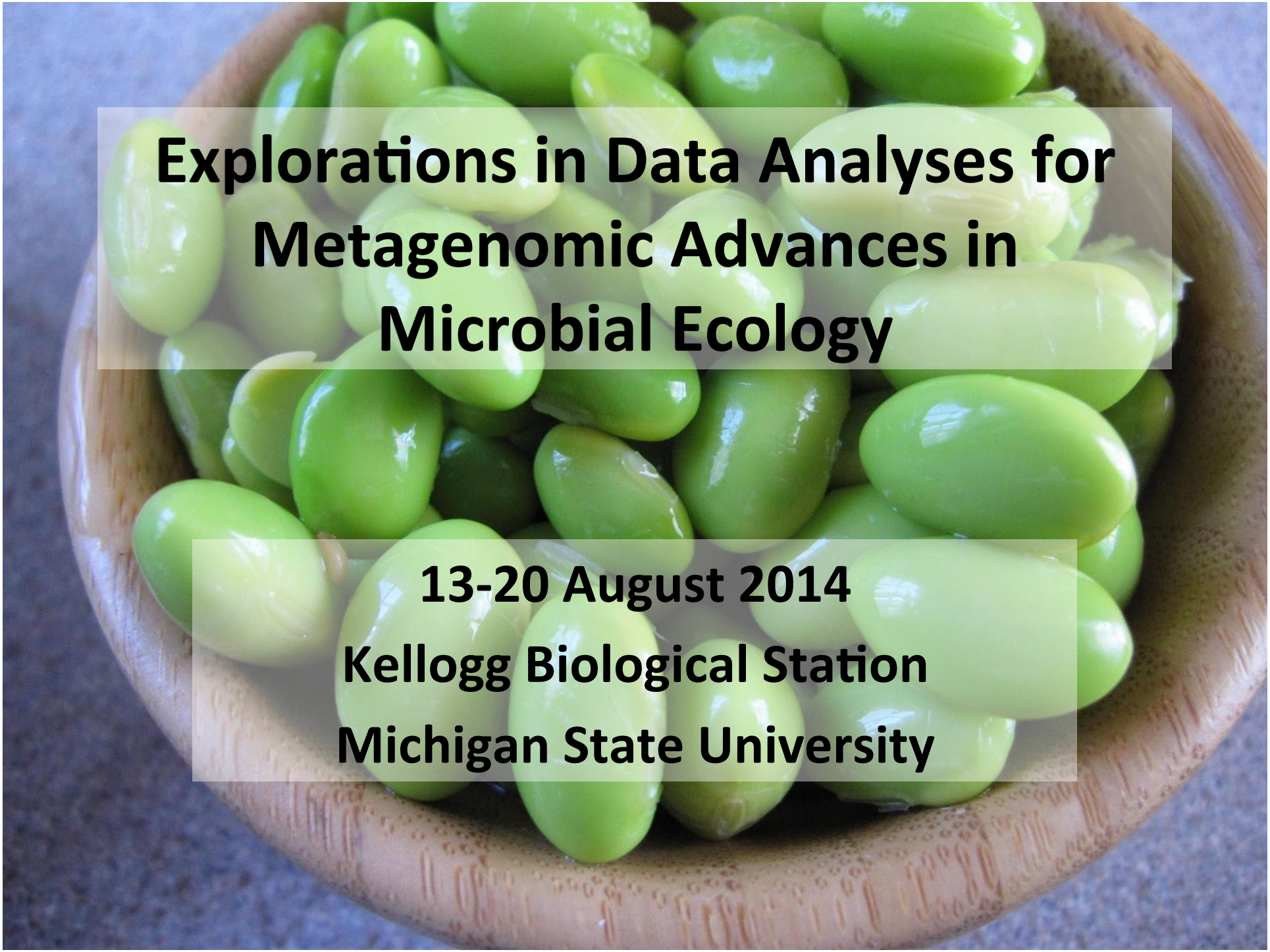

Explorations in Data Analyses for Metagenomic Advances in Microbial Ecology
13-20 August 2014
Kellogg Biological Station
Michigan State University

Review from QIIME Intro

- 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)
- Next gen. sequencing offers a cost-effective way to assay metagenomic DNA of microbial communities from any environment
- There are lots of choices for analyzing 16S 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**

Tutorial review: What happened yesterday?

- Merged paired-end reads: **pandaseq**
- Picked OTUs: based on 97% sequence identity: **pick_otus.py**
- Picked one sequence to represent each OTU: **pick_rep_set.py**
- Make an alignment of those representative sequences for future tree building and other analyses: **align_seqs.py**

Questions from yesterday?

Lecture 2: Alpha Diversity

- Alpha diversity in all of its glory
- The advantage of phylogenetic information
- Rarefaction
- What does a community look like, data-style?
- A note on the .biom v. OTU table format

Alpha 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>.

Alpha diversity

- Alpha 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.*

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.

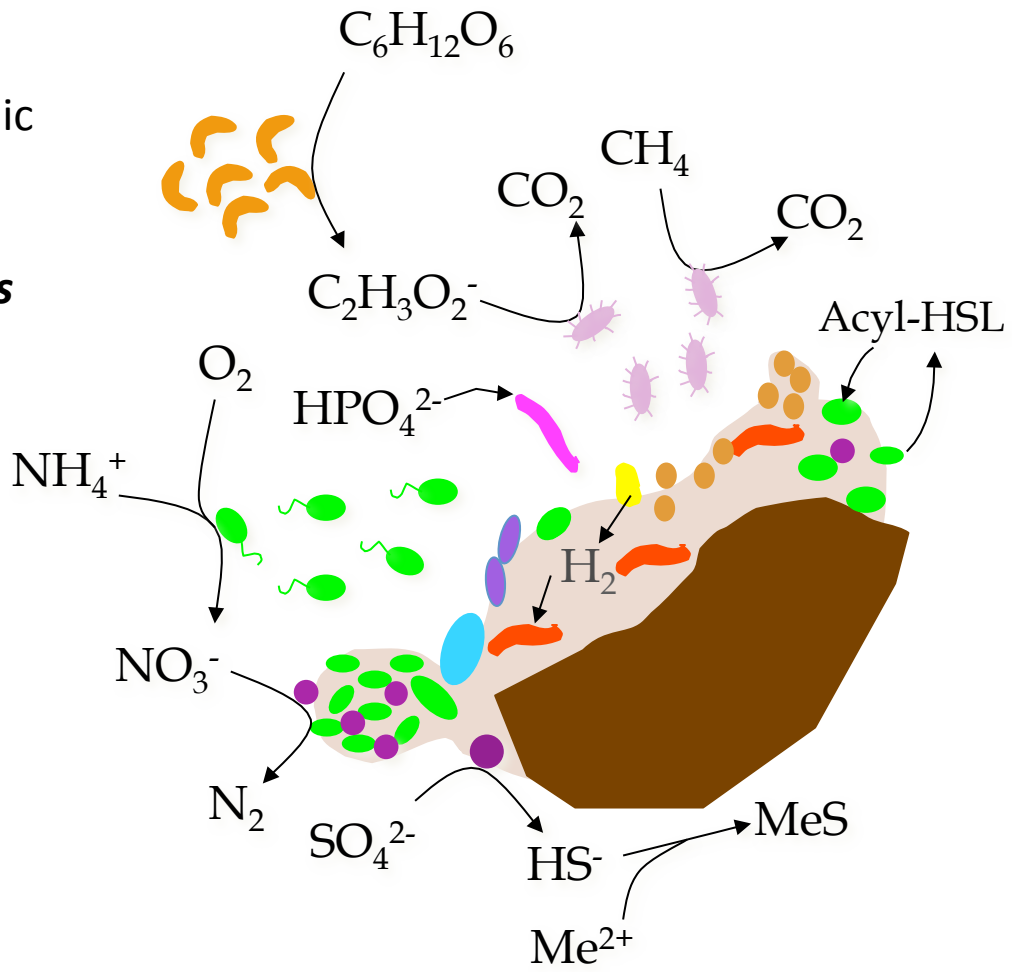
Alpha (within-sample) diversity

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***

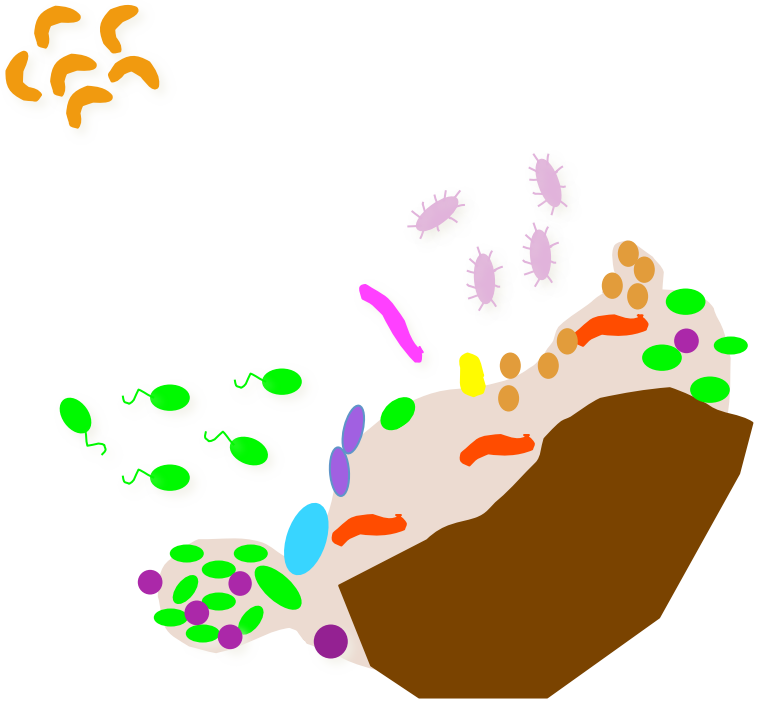


Alpha diversity

OTU

Richness: How many OTUs?

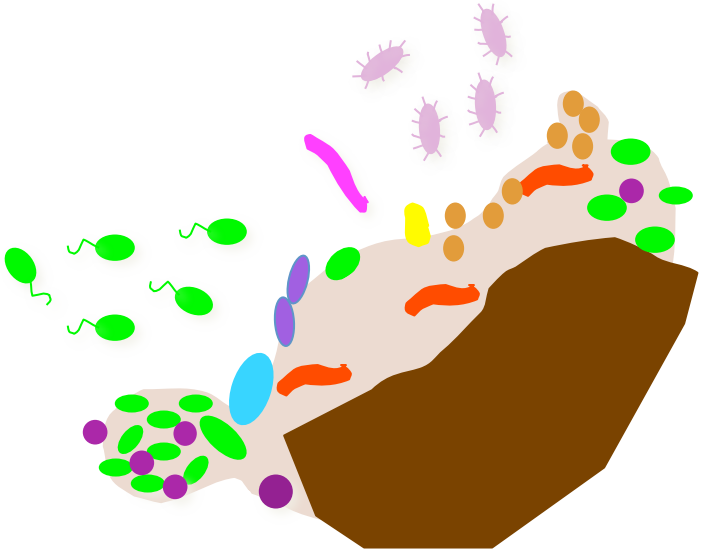
Richness = 11 OTUs



Alpha diversity

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

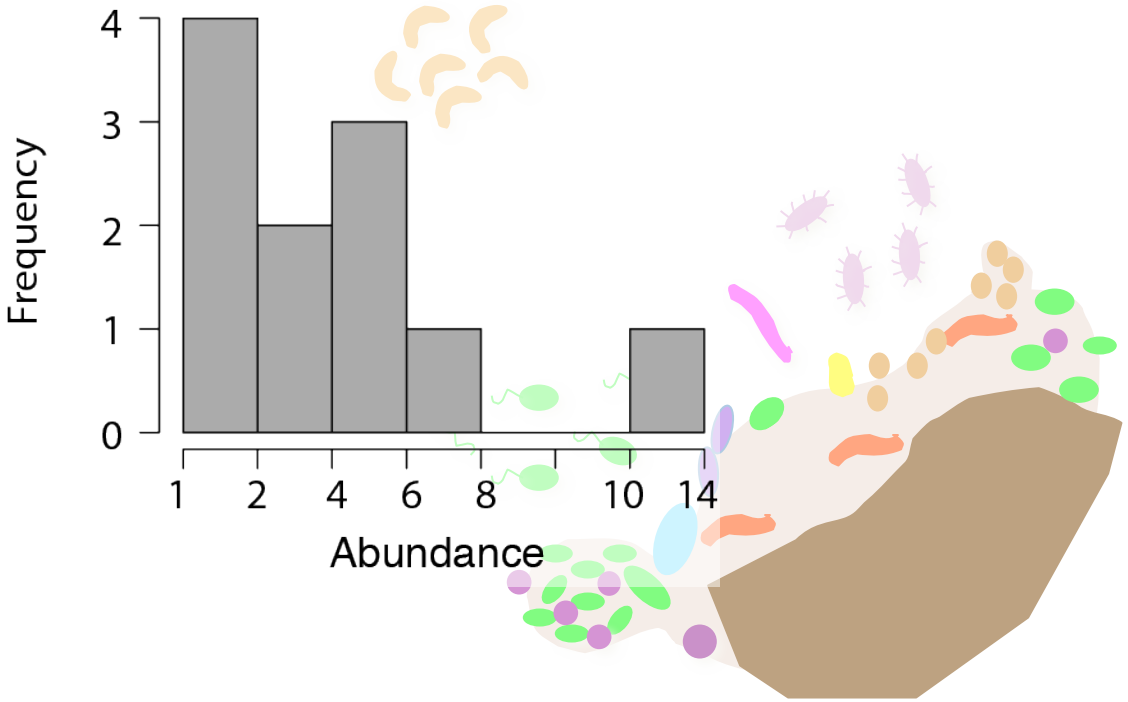
OTU	**Count**: No. seq, no. individuals (e.g., FISH), biomass, etc.
	6
	1
	4
	8
	1
	5
	3
	13
	5
	1
	2



Alpha diversity

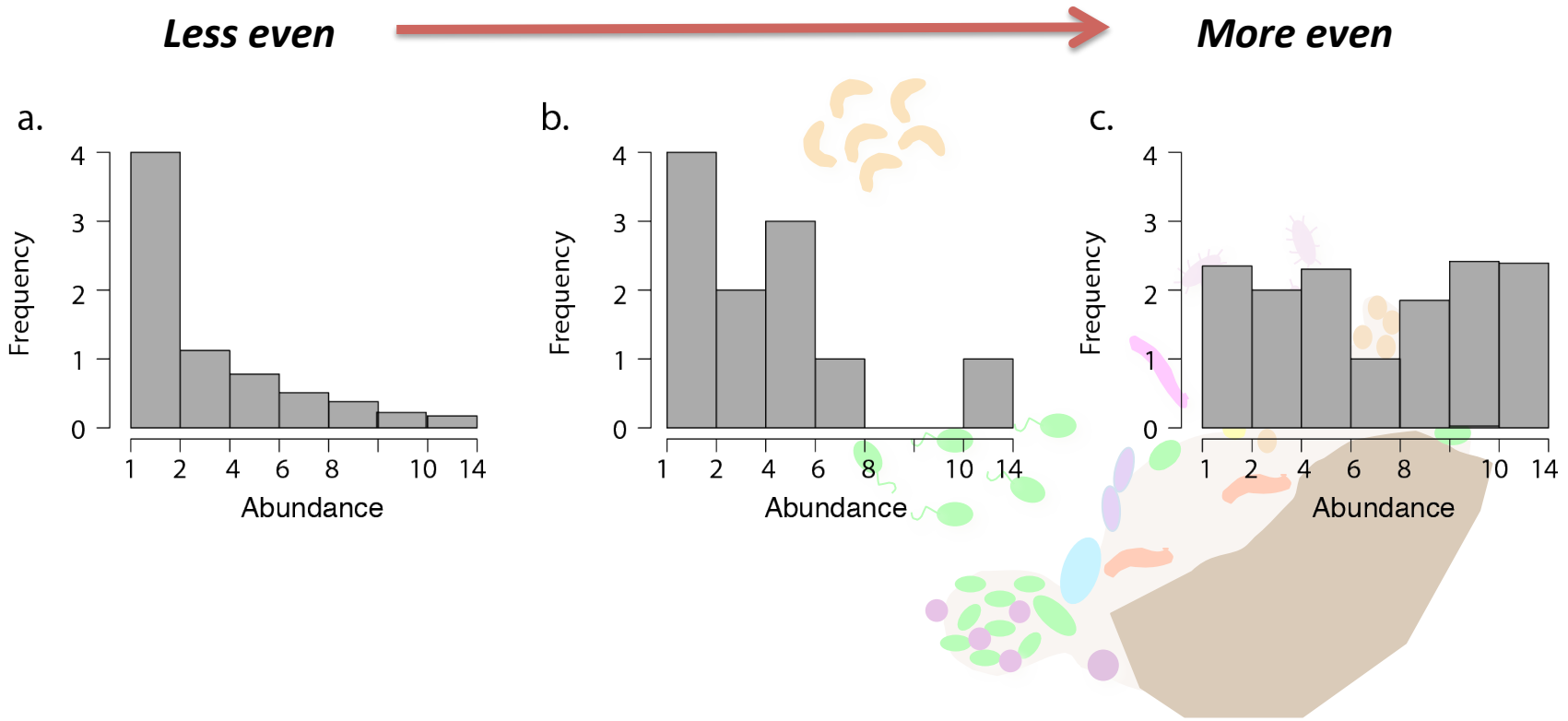
Evenness: What is the distribution of abundances in the community?

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



Alpha diversity

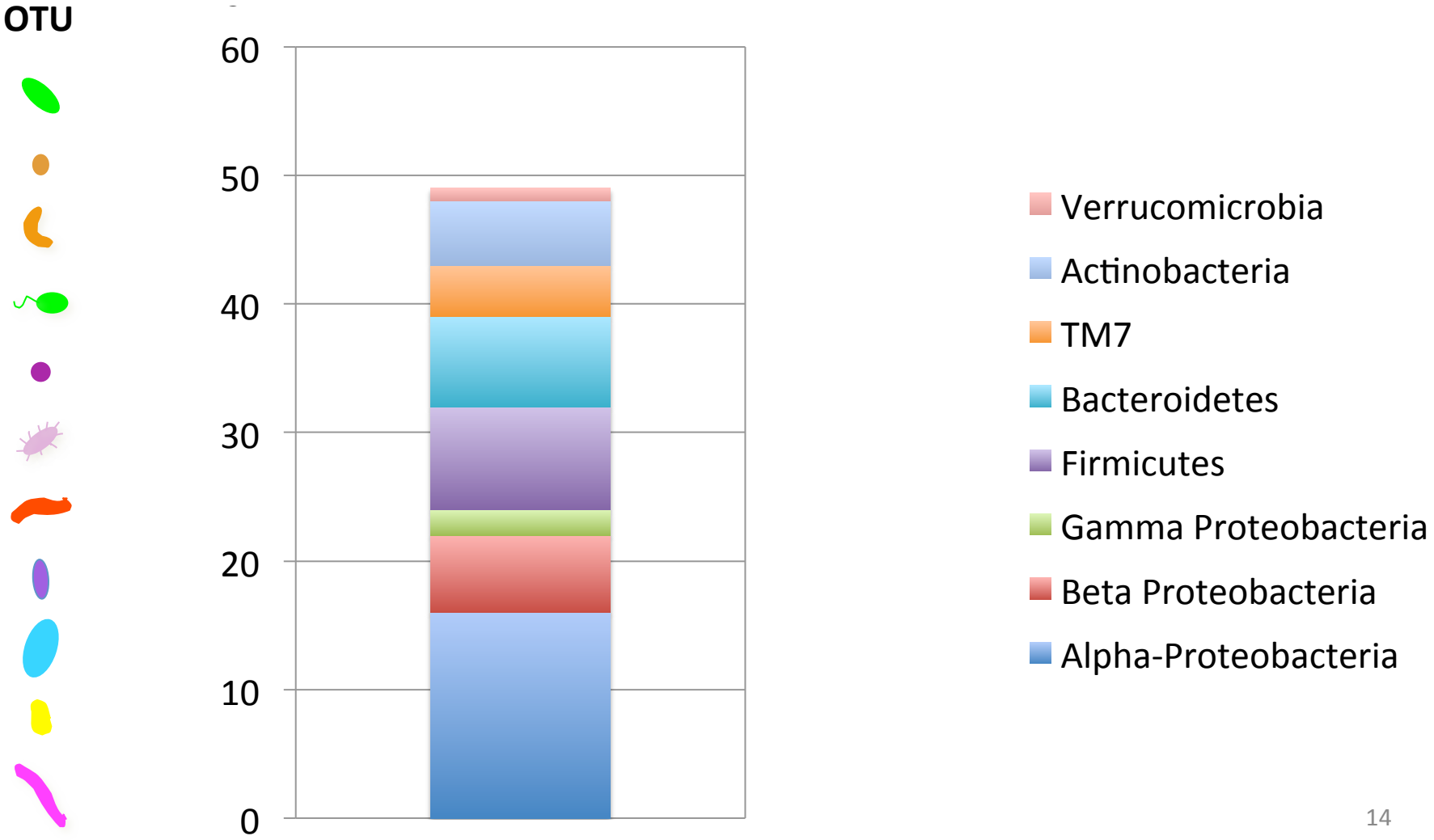
Evenness: What is the distribution of abundances in the community?



Common metric = Pielou's evenness

Alpha diversity

Composition: What is the distribution of taxonomic affiliations?

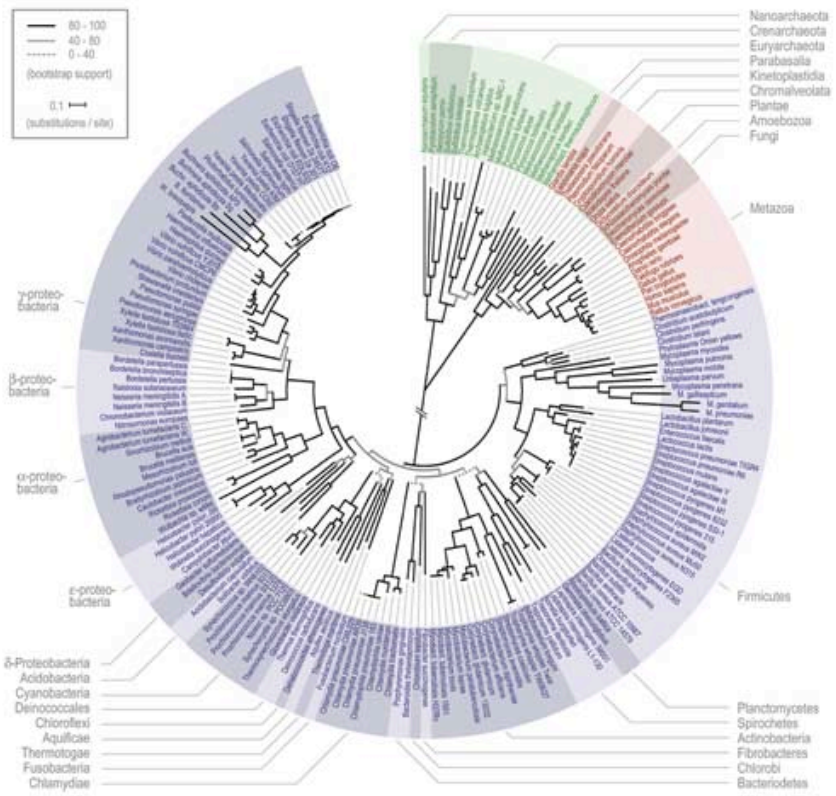


Alpha diversity: The advantages of phylogenetic information

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



Rarefaction exercise

Rarefaction

- Because of sequencing artifacts and experimental design, there can be quite a range of quality sequences returned for each sample.
- Rarefaction : Sub-sampling of sequences to achieve an even number across all samples within a dataset.
- **This is very important for being able to compare diversity across samples.**
- Choose a rarefaction depth that maximizes the number of samples that can be included at the most informative sequencing depth.

What does a community look like, data-style?

- “OTU table” – the original
- .Biom table – more concise & faster computing for extra large datasets
- New formats : biom2, coming online RIGHT NOW

Kinds of Community matrix/ OTU Table

Raw
Straight up counts*

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

**note: sequencing data must be rarefied (re-sampled) to an equal sequencing effort!*

Relative
Percent or proportion
(Standardizes for sampling effort?

	Caterpillar 1	Caterpillar 2	Caterpillar 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)

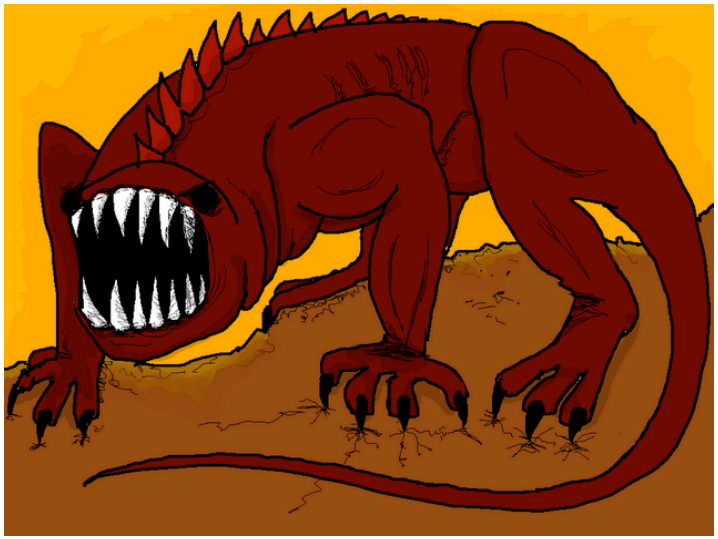
	Caterpillar 1	Caterpillar 2	Caterpillar 3
OTU 1	0	1	1
OTU 2	1	1	1
OTU 3	1	1	1

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

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! Biom formats are 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
```

Old-style OTU table - 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