

Using Phylogenetic Structure to Assess the Evolutionary Ecology of Microbiota

TJS

iSEEM Call

Dec. 15, 2014

How are Microbes Distributed In Nature?

- A major question in microbial ecology
- Used to assess putative function of taxa:
 - Core taxa: those common to a set of communities.
May be critical or keystone taxa
 - Interacting taxa: those that correlate in abundance across samples
 - Environmental interactions: those taxa that correlate with environmental covariates across samples

Measuring OTU Distributions

1. Generate 16S sequences from a variety of communities
2. Classify/cluster sequences into OTUs
3. Calculate each OTU's abundance in each sample
4. Evaluate the OTU by sample matrix to assess OTU distributions

OTU Matrices are Frequently Sparse

	OTU 1	OTU 2	OTU 3	OTU 4
Sample 1	7	1	0	0
Sample 2	0	3	5	0
Sample 3	3	0	0	5
Sample 4	0	0	10	0

Create several challenges:

1. Inference: Lots of tests
2. Little overlap: Hard to correlate OTU distributions

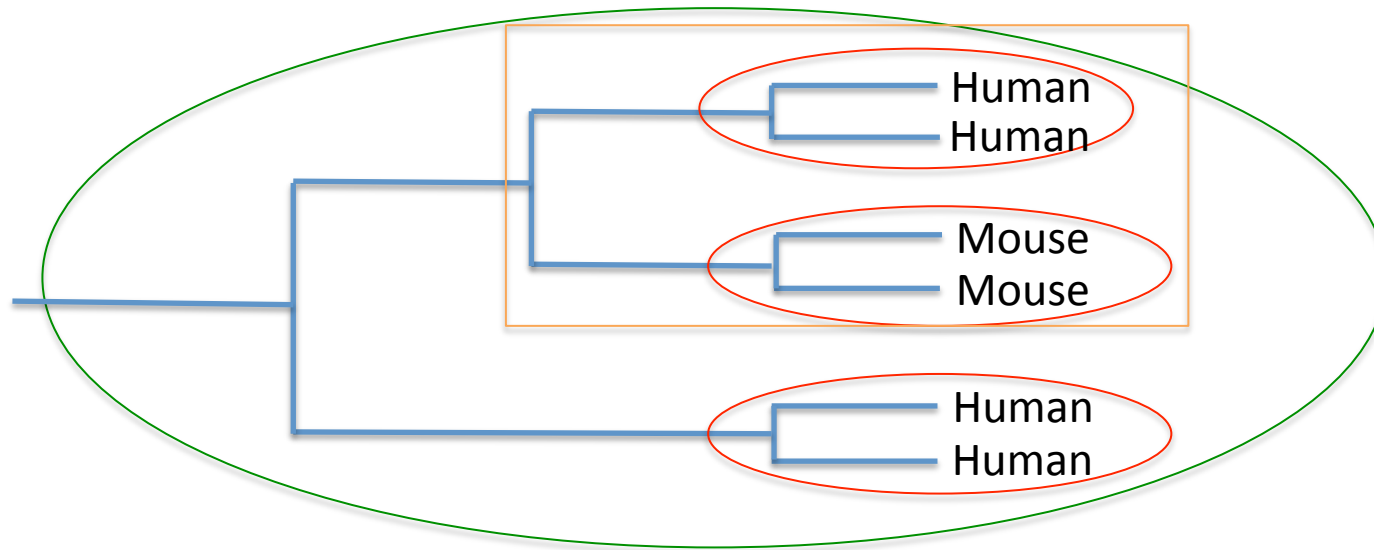
Phylotyping Often Increases Overlap

- Classifying sequences into taxonomic groups often decreases data sparsity and increases frequency of occurrence
 - e.g., Krych PLOS ONE 2013
- Suggests that non-overlapping OTUs are closely related in phylogeny
- Taxonomy is an imprecise method of grouping
 - e.g., not phylogenetically consistent

Considering Phylogenetic Structure May Improve Resolution of Interesting Taxa

1. Build a tree using 16S sequences from communities of interest
2. Annotate tree tips with community identifiers
3. Build a samples by clades matrix:
 1. Traverse tree and, for each node, measure
 1. The samples each monophyletic clade is found in
 2. The abundance of the clade in each sample

An Example



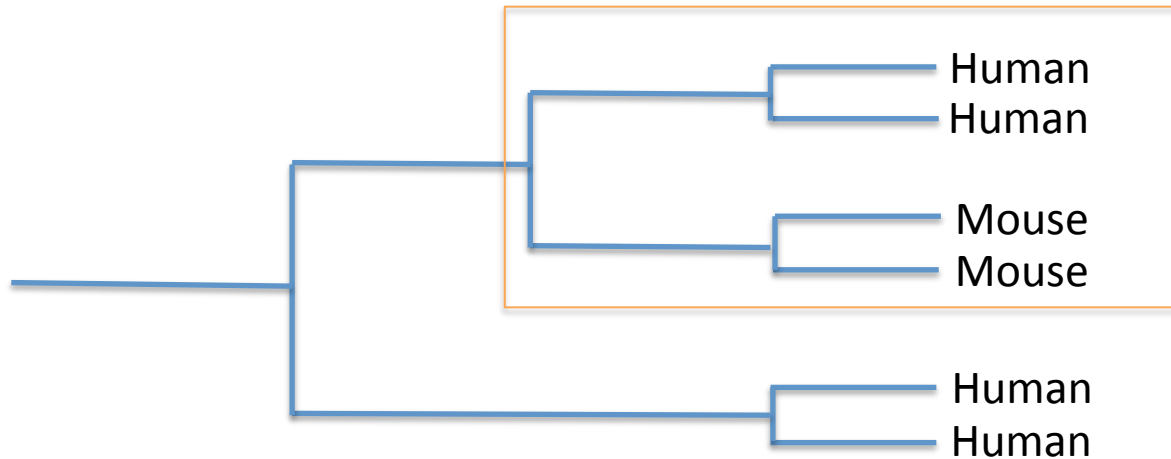
- Genus Designation
- OTU Designations
- Core Monophyletic Clade

1. At the OTU level, there are no core groups
2. While the genus is core, it contains additional lineages that may may complicate statistics (e.g., correlation)
3. By walking the tree, we can identify the specific monophyletic clade that is common to humans and mice

Benefits of Assessing Distributions of Clades

- Reduces sparsity of the data
- Improves identification resolution
- Incorporates evolutionary information into assessment of distribution

Benefits of Evolutionary Info: Core Taxa



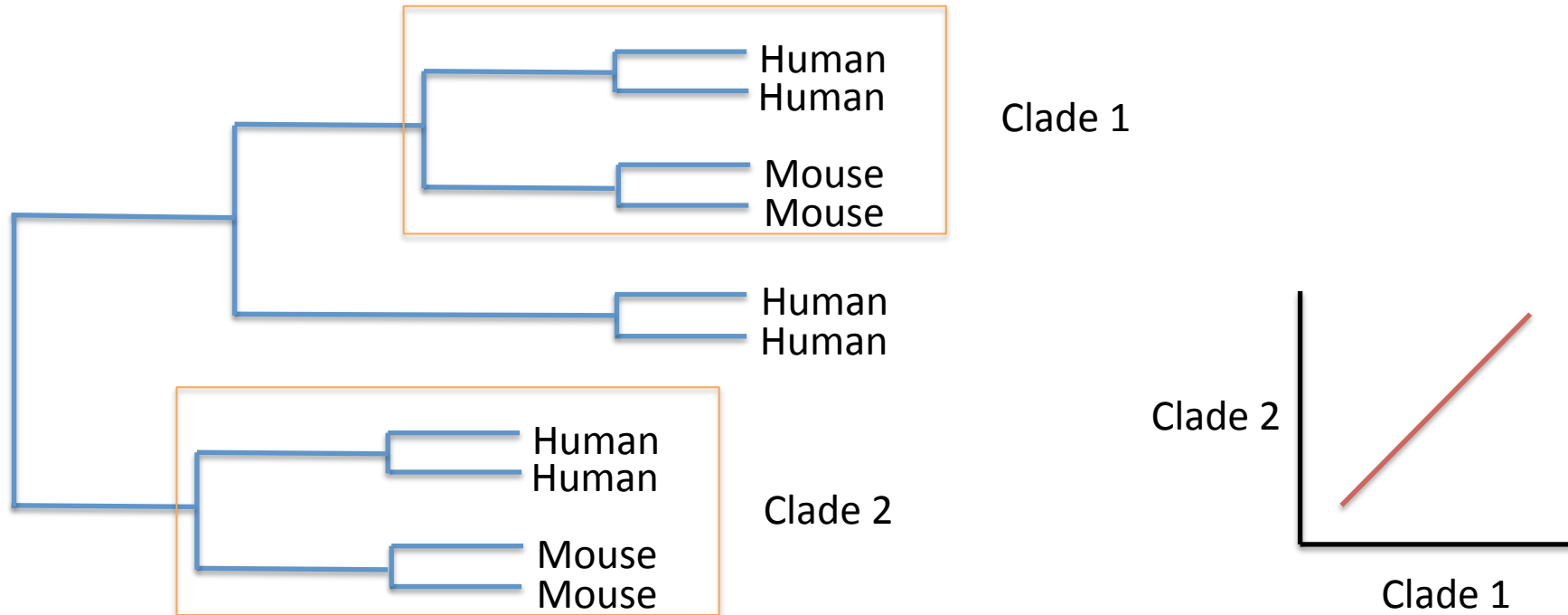
Provides hypotheses about the evolution of ecological functions:

- e.g., this ancestor may have evolved a function critical to the maintenance, operation, etc. of these communities

Provides opportunity to explore co-evolution (in the case of host-associated microbiota):

- Is structure of this clade concordant with structure of host-phylogeny?
Obviously not interesting in the case of two hosts.

Benefits of Evolutionary Info: Interacting Taxa



Provides hypotheses about robustness of interaction

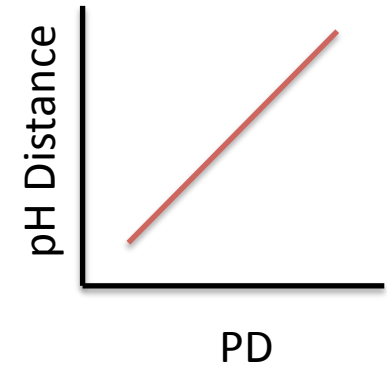
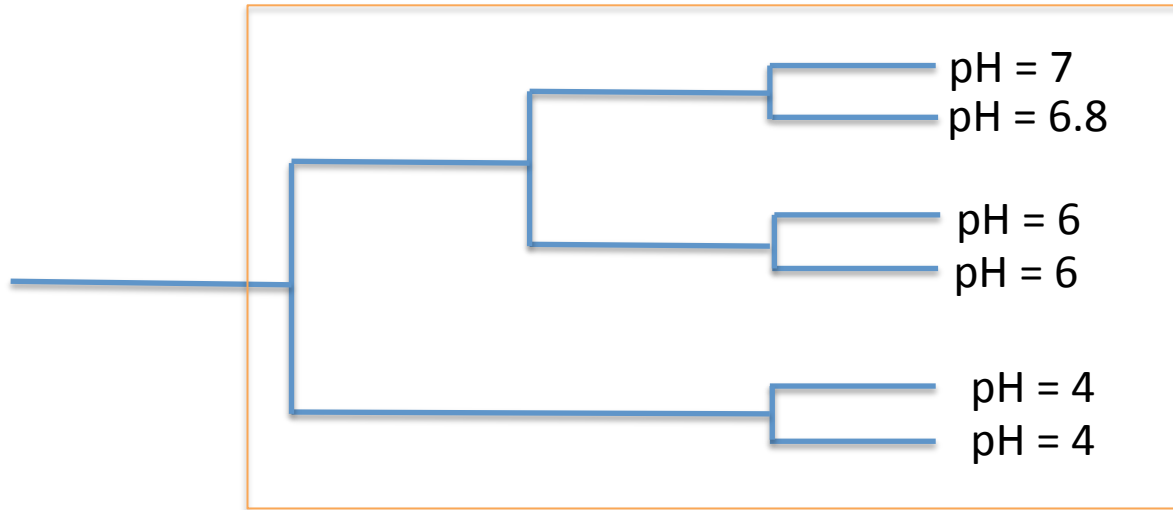
- e.g., Any random individual from clade 1 may produce a function needed for any random individual to survive

Provides hypotheses about the evolution of interaction:

- e.g., these ancestors may have directly interacted, interaction maintained

Potential to discover co-evolution between interacting clades if concordant subtrees

Benefits of Evolutionary Info: Ecological Interaction



Provides framework to quantify potential evolutionarily conserved environmental interactions

Maybe this is Neat, but It's Probably Slow

- Lots of 16S data being generated
- Tree assembly is error prone with large volumes of data – they may profoundly impact results
- Tree walking can be very slow

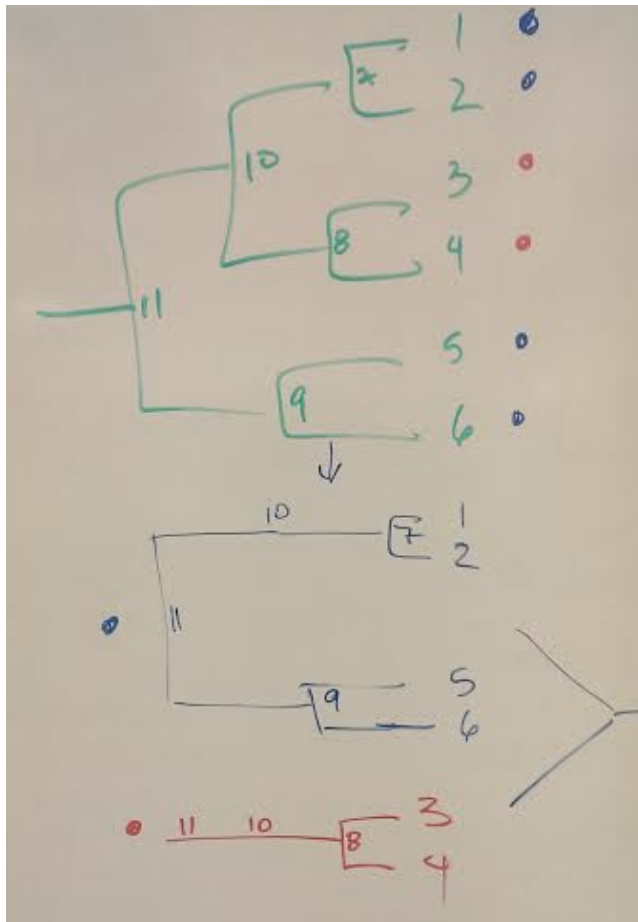
Solution: Borrow From FastUniFrac

Classify sequences into a reference tree
(e.g., GreenGenes)

Extract each sample's subtree

Use a ref sequence-to-all-ancestors map to
quantify abundance of each node for each
sample

Will miss novel lineages, but avoids walking
the tree.



	1	2	3	4	5	6	7	8	9	10	11
0	1	1	0	0	1	1	2	0	2	2	4
•	0	0	1	1	0	0	0	2	0	2	2