Using Phylogenetic Structure to Assess the Evolutionary Ecology of Microbiota

TJS

iSEEM Call

April 2015

How are Microbes Distributed In Nature?

- A major question in microbial ecology
- Used to assess properties of taxa:
  - *Core taxa*: those common to a set of communities. May be critical or keystone organisms
  - *Intertaxon interactions*: 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 (or phylotypes)
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

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:

   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

Example: Identification of Clades Common to Myriad Samples (Core)

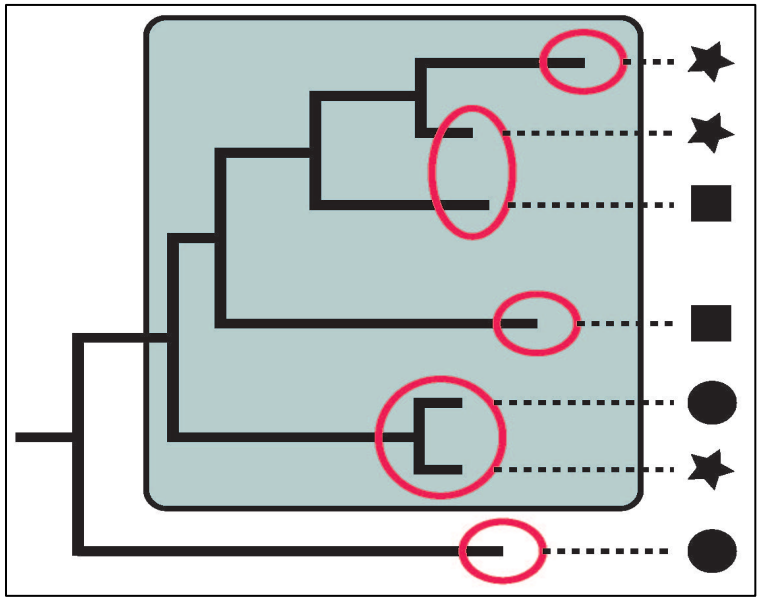
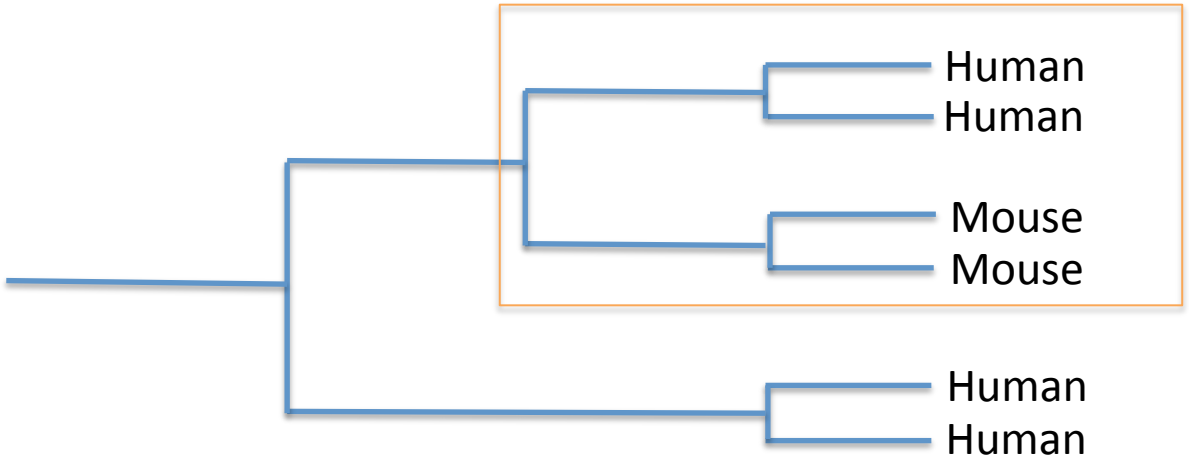


Fig. 1: A sub-tree of the total 16S phylogeny that contains a core clade. 16S sequences derived (dashed lines) from various communities (solid shapes) can be related via a phylogeny (solid lines) or clustered into OTUs (red circles). In this example, no OTU is common to all communities, but a monophyletic clade is (shaded area), indicating that the common ancestor may have evolved and subsequently maintained a function critical to these communities. Note that interesting clades (*e.g.*, core clades) may also be discovered at the sub-OTU level.

Benefits of Assessing Distributions of Clades

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

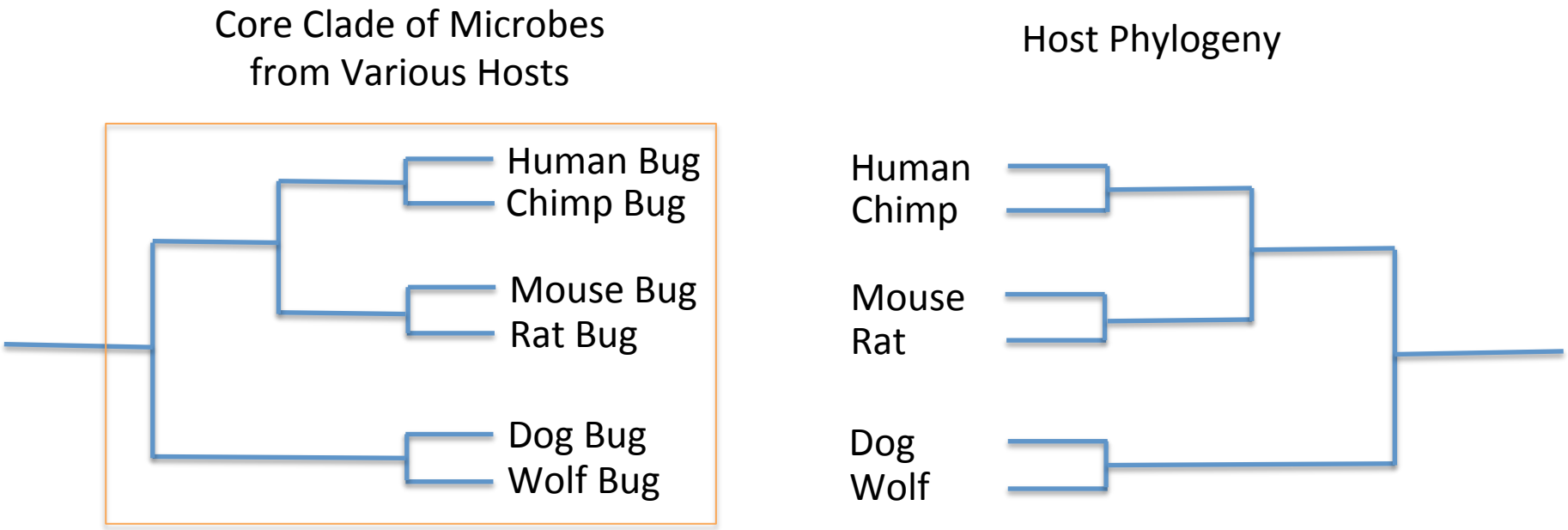
Benefits of Evolutionary Info: Core Taxa



Provides hypotheses about the evolution of ecological functions:

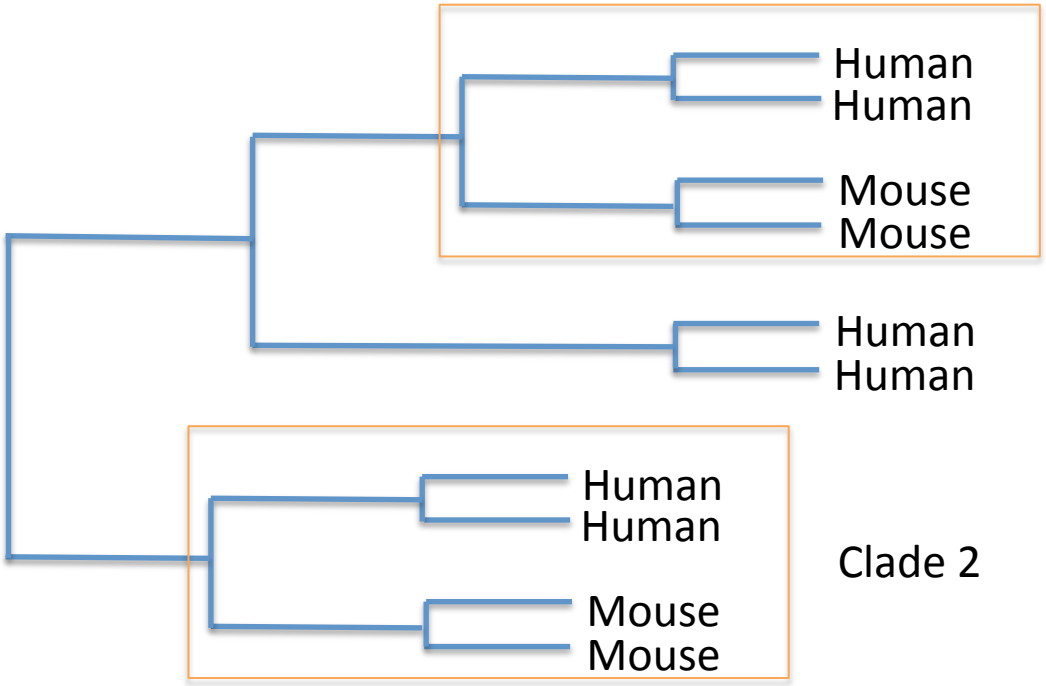
- e.g., this ancestor may have evolved a function critical to the maintenance, operation, etc. of these communities

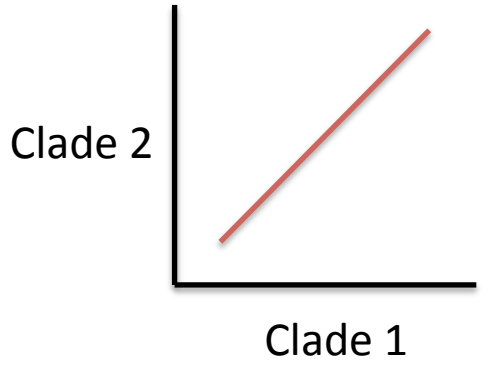
Benefits of Evolutionary Info: Co-diversification with Host



Can identify clades of host-associated microbiota that have co-diversified with their hosts

Benefits of Evolutionary Info: Interacting Taxa





Provides hypotheses about robustness of interaction

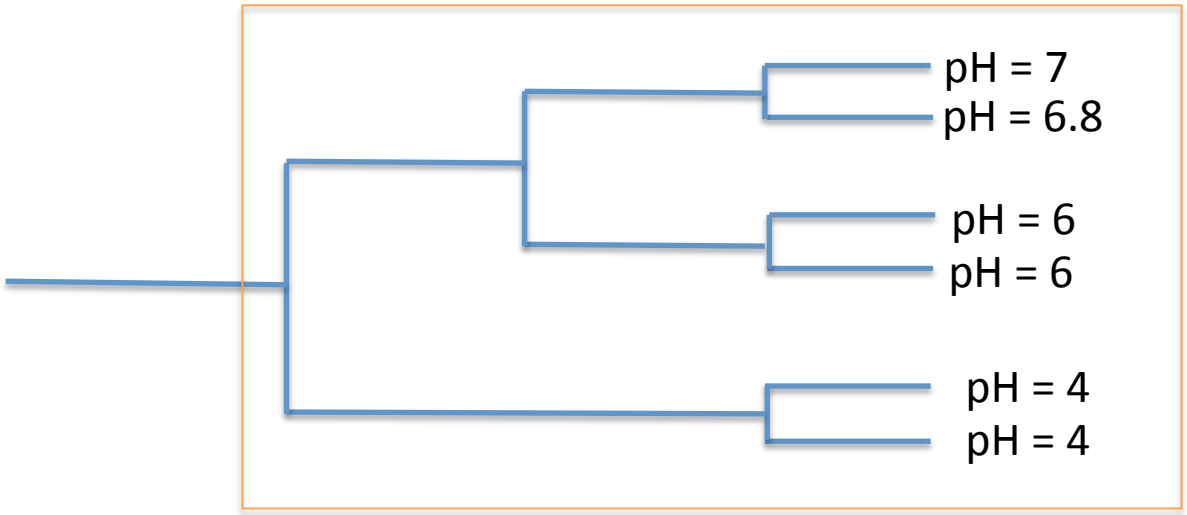
- e.g., Any random individual from clade 1 may produce a function needed for any random individual to survive

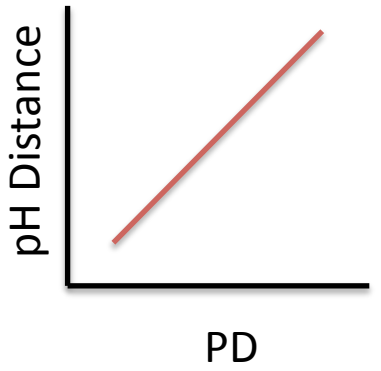
Provides hypotheses about the evolution of interaction:

- 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

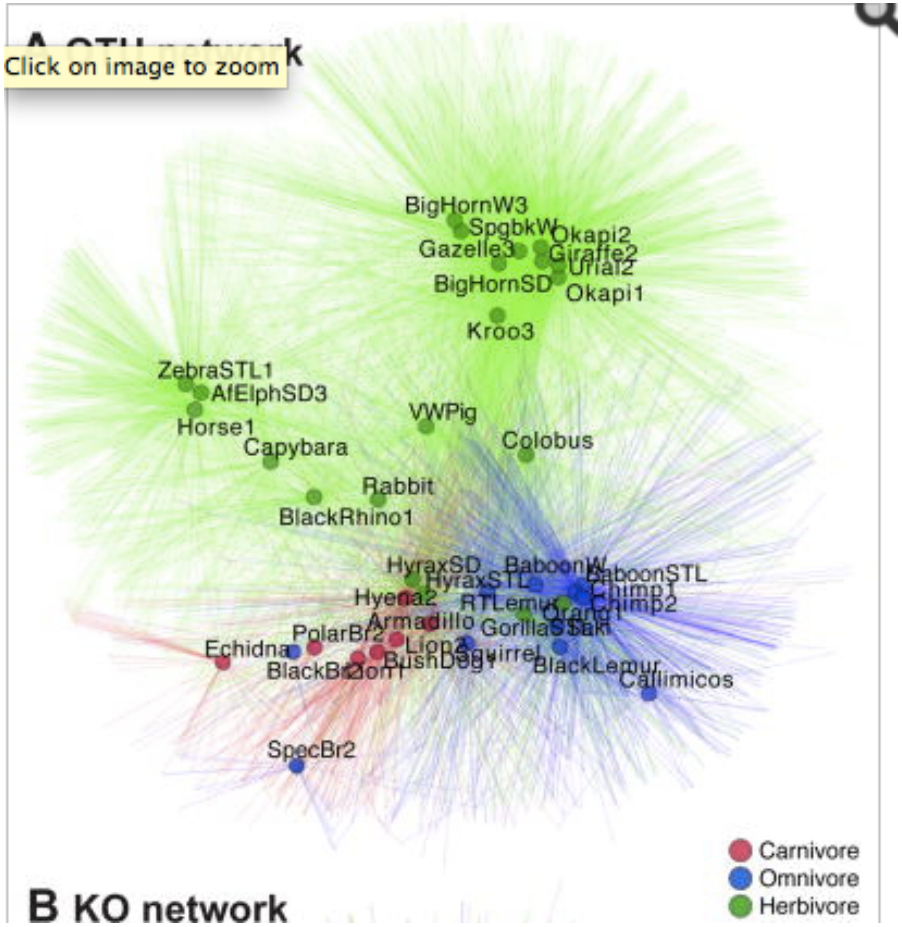




Provides framework to quantify potential evolutionarily conserved environmental interactions

Proof-of-Principal

- Muegge et al Science 2011
- Found that microbiomes grouped by dietary preference over phylogeny, no core OTUs
- Do specific clades co-diversify?
- Are specific clades common to all hosts?



Proof-of-Principal

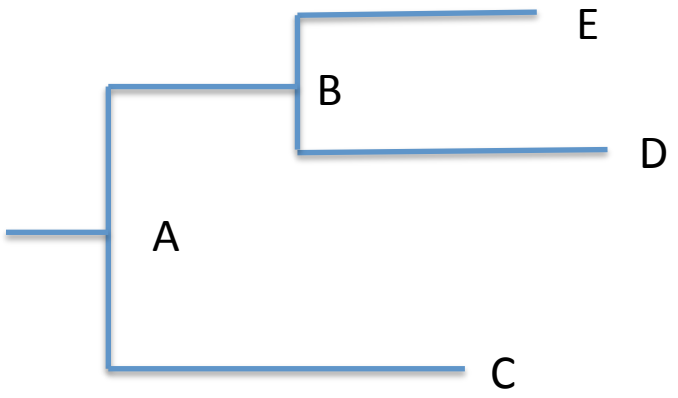
- Used their data to build *de novo* 16S tree
- Focused on the 6 non-human primates
  - 11 clades common to all samples
    - one within *Prevotella*
  - Identified clades that co-vary across samples
    - Clade w/in *Barnesiella* co-varies with another in the Peptostreptococcaceae

Maybe this is Neat, but It is Slow

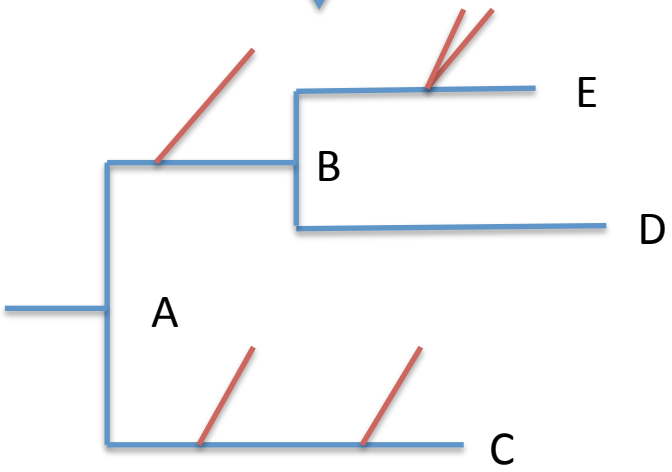
- Lots of 16S data being generated, and tree walking is rarely efficient
- Tree assembly is error prone with large volumes of data and errors may profoundly impact results

Solution: Place Reads on a Reference Tree

Greengenes Reference Tree
Tips are Reference OTUs



Place Reads



Build Edge to Clade Matrix Once

```
AB -> A
AC -> A
BE -> B, A
BD -> D, A
```

Quantify Abundance of Clades
Using only Placed Reads

A	B
5	2

Challenges with this Method

1. For large trees, pplacer is very slow

   Potential solutions:

   - Cut out subtrees (e.g., phyla), place reads into each one and classify into best hit.
   - Classify sequences into reference OTUs used to build the tree

2. Accuracy of pplacer on 16S data is not well described

   Potential solutions:

   - Statistical simulations
   - Compare to de novo tree

Next Steps

1. Explore an implement these proposed solutions
2. Identify null models of clade diversification (O'Dwyer)
3. Apply to real data
    1. Muegge
    2. Kembel
    3. Ochman
    4. Non-host associated