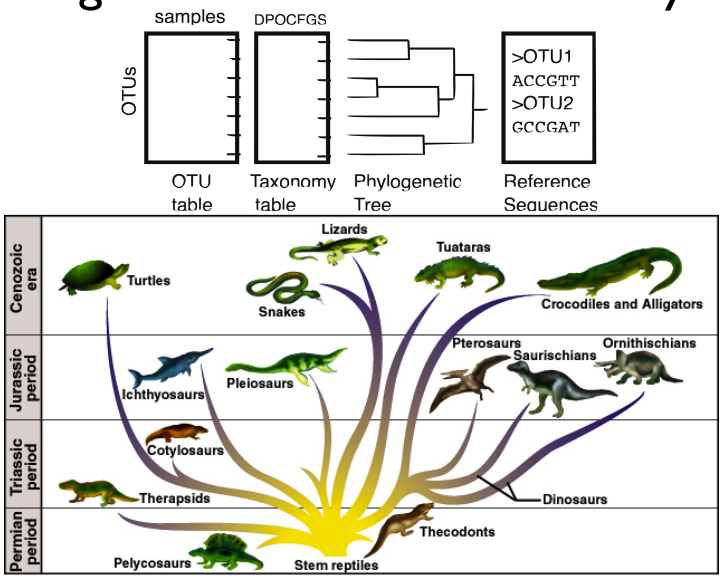
Using Trees in Microbiome Analysis

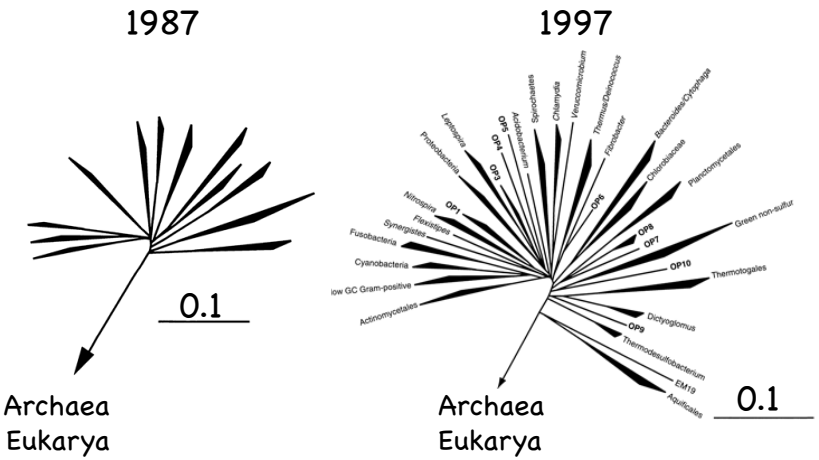


Using Trees in Microbiome Analysis

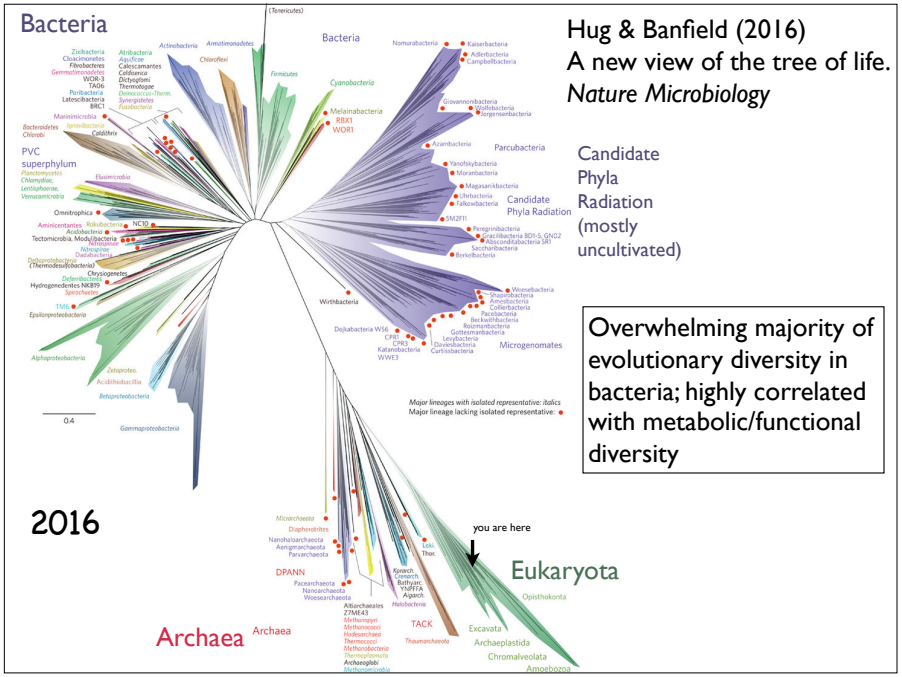
- Phylogenetic (Evolutionary) Trees
  - Tree-Building ("quick" overview)
  - Tree formats (Newick, Ape's "phylo")
  - Manipulating Trees in phyloseq/ape
  - Tree plots (Examples, how to interpret)
- Using Trees and contingency tables together
  - UniFrac and variants
  - DPCoA

Phylogenetic Trees

Evolutionary Tree, Known Bacteria



Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313), 734–740.



Hug & Banfield (2016)
A new view of the tree of life.
Nature Microbiology

Candidate Phyla Radiation (mostly uncultivated)

Overwhelming majority of evolutionary diversity in bacteria; highly correlated with metabolic/functional diversity

Phylogenetic Trees

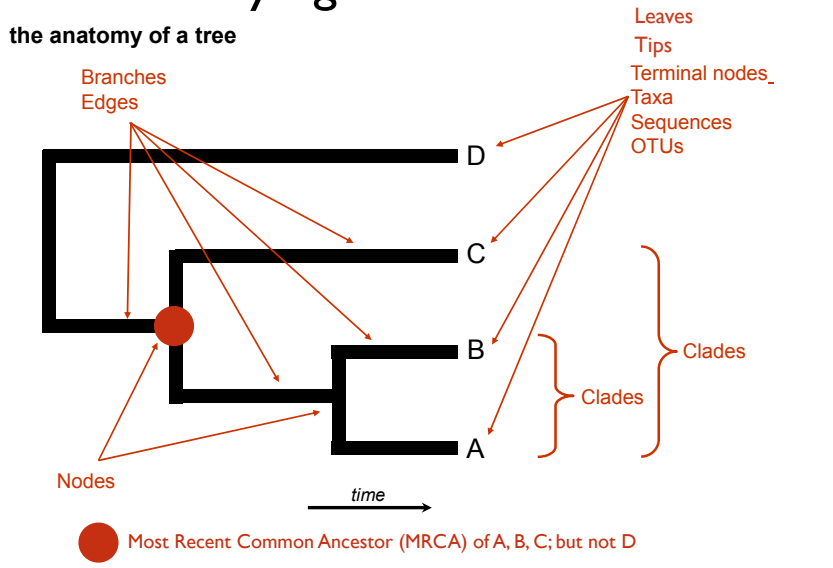
Motivations:

(1) Reconstructing evolutionary history from incomplete information

(2) Robust summary of the similarity of related biological sequences
(a lot like hclust)

The data - biological sequences
- often proteins, sometimes DNA/RNA (16S rRNA), etc.

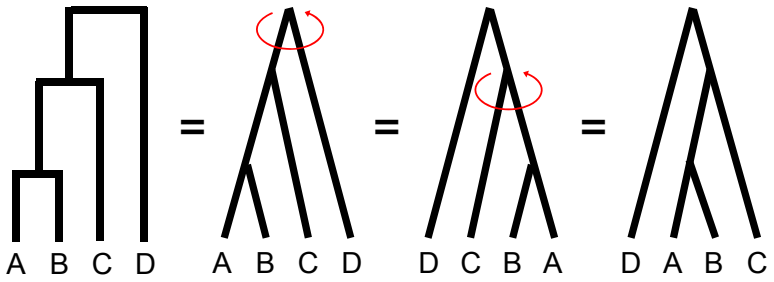
Phylogenetic Trees Nomenclature



Adapted from N. Provart & D. Guttman

Phylogenetic Trees

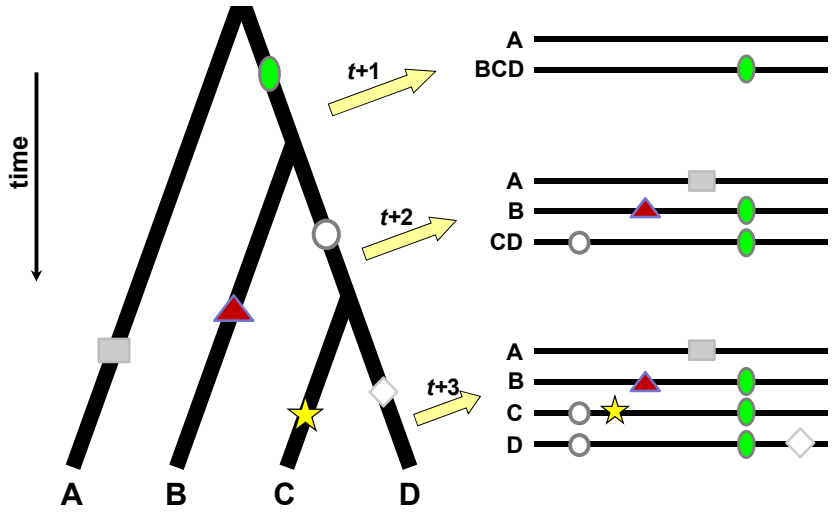
Rotating internal nodes is not meaningful:



2^{N-1} possible arrangements for a particular rooting

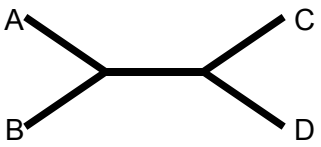
Adapted from N. Provart & D. Guttman

Phylogenetic Trees example

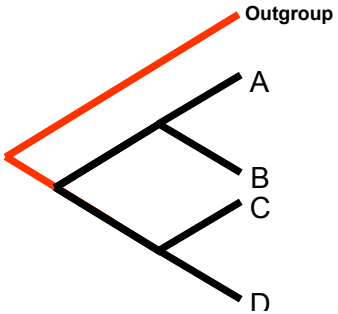


Adapted from N. Provart & D. Guttman

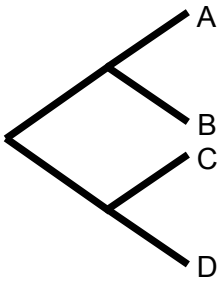
Using an "Outgroup"



Outgroup Rooting



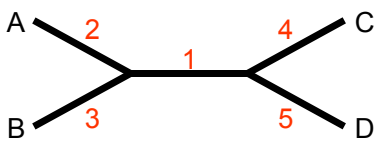
Midpoint Rooting



Adapted from N. Provart & D. Guttman

Rooting Trees

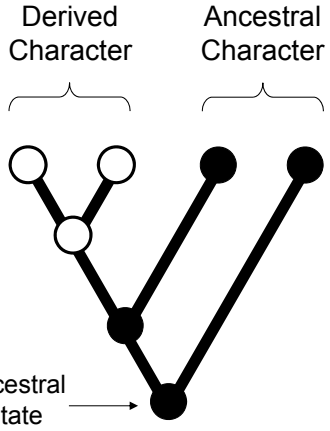
Unrooted Tree



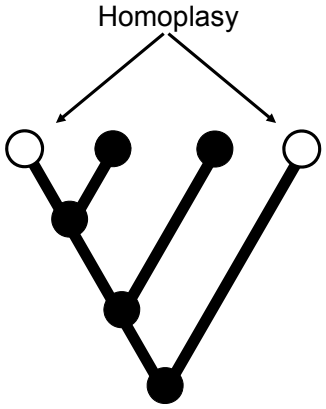
Rooted Trees – have one node from which all other nodes descend
– imply direction corresponding to evolutionary time

Adapted from N. Provart & D. Guttman

More Terminology



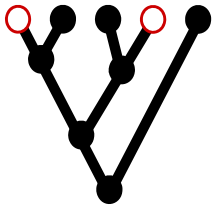
Homology - Similarity due to common ancestry



Homoplasy - Similarity due to parallel evolution, convergent evolution, or secondary loss

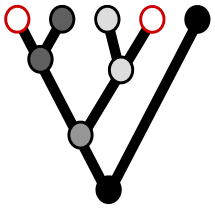
Adapted from N. Provart & D. Guttman

Forms of homoplasy...



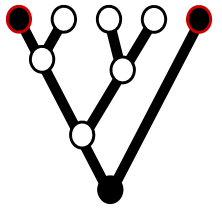
Parallel Evolution

Independent evolution of same character from same ancestral state



Convergent Evolution

Independent evolution of same character from different ancestral state

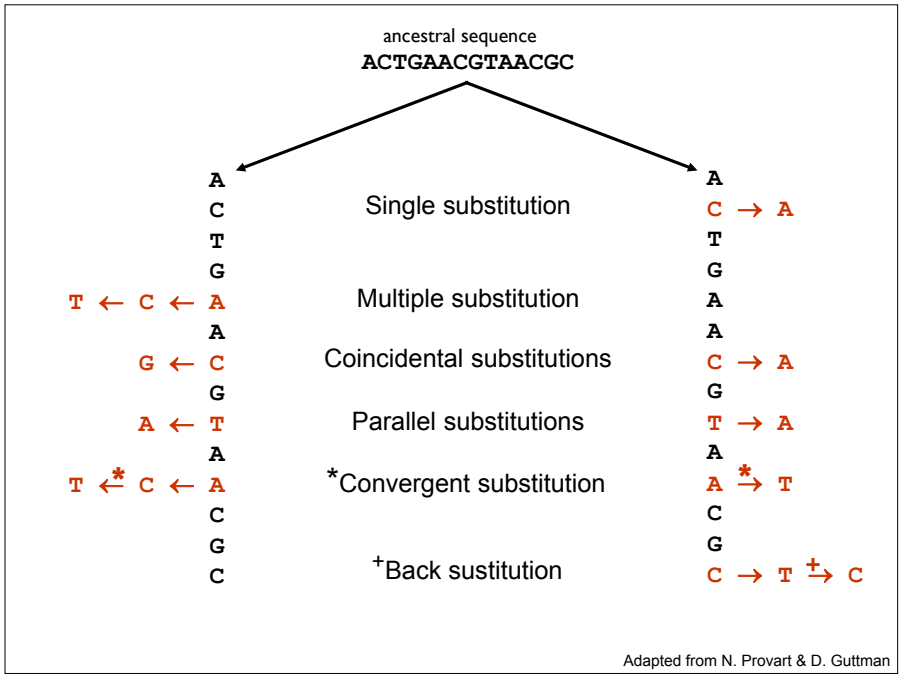


Secondary Loss

Reversion to ancestral state

E.g. Ni-Fe and Fe-only hydrogenases: highly-similar enzymatic activity, no detectable shared ancestry

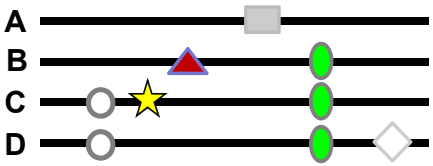
Adapted from N. Provart & D. Guttman



Phylogenetic Tree Construction Methods

Multiple Sequence Alignment:

All tree-building begins with multiple-alignment



- Naïve multiple sequence alignment is NP-complete.
- Students typically don't want to spend time multiple alignment details.
- Just read about / use one of the following multiple-alignment algorithms:

ClustalW	Thompson, J. D., Higgins, D. G., & Gibson, T. J. (1994). *CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment...* Nucleic Acids Research, 22(22), 4673–4680.
Muscle	Edgar, R. C. (2004). *MUSCLE: multiple sequence alignment with high accuracy and high throughput.* Nucleic Acids Research, 32(5), 1792–1797.
MAFFT	Katoh, Misawa, Kuma, Miyata 2002 (Nucleic Acids Res. **30**:3059-3066) *MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform.*
Mauve, Lagan, etc.	Whole genome alignment...

NOTE: You will not create a meaningful tree from a meaningless alignment. Spending time selecting the appropriate alignment tools and checking your alignment is usually a worthwhile thing to do.

Phylogenetic Tree Construction Methods

Distance-based tree methods

UPGMA	Bad, don't use. Implemented as guesses in better, more complex algorithms for m-alignment / tree construction
Neighbor-Joining	Also not very good, only use if other methods intractable, or use as initial guess for parsimony or ML tree.

Character-based (discrete) tree methods

Maximum Parsimony

Maximum Likelihood

Bayesian Methods

Phylogenetic Tree Construction Methods

Distance Methods

Relationships based upon sequence similarity.

Advantages

- Computationally fast.
- Single "best tree" found.

Disadvantages

- Assumptions
  - additive distances (always)
  - molecular clock (sometimes)
- Information loss occurs due to data transformation
- Uninterpretable branch lengths
- Single "best tree" found.

Phylogenetic Tree Construction Methods

UPGMA

Not much point in discussing. Not very good.
You know how to do it from clustering lecture(s).

Details:

* Assumes rates of evolution are same among different lineages (severely unrealistic)
* Very sensitive to unequal evolutionary rates
* Tends to be reliable only if data/phylogeny is essentially ultrametric (severely unrealistic)

Phylogenetic Tree Construction Methods

Neighbor Joining

Saitou, N., & Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, *4*(4), 406–425.

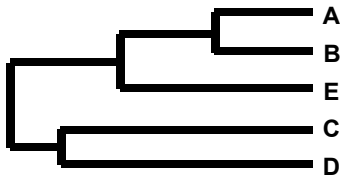
1. Calculate pairwise distances
2. Create distance matrix
3. Determine net divergence for each terminal node
4. Create rate-corrected distance matrix
5. Identify taxa with minimum rate-corrected distance
6. Connect taxa with minimum rate-corrected distance via a new node, and determine their distance from this new node
7. Determine the distance of new node from rest of taxa or nodes
8. Regenerate distance matrix
9. Return to step 2

Adapted from N. Provart & D. Guttman

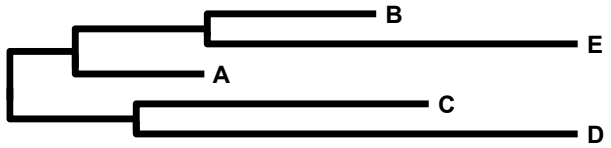
UPGMA vs. Neighbour-Joining

	A	B	C	D	E
A	-	17	21	31	23
B		-	30	34	21
C			-	28	39
D				-	43
E					-

UPGMA



NJ



Adapted from N. Provart & D. Guttman

Phylogenetic Trees

Character-based (discrete) Methods

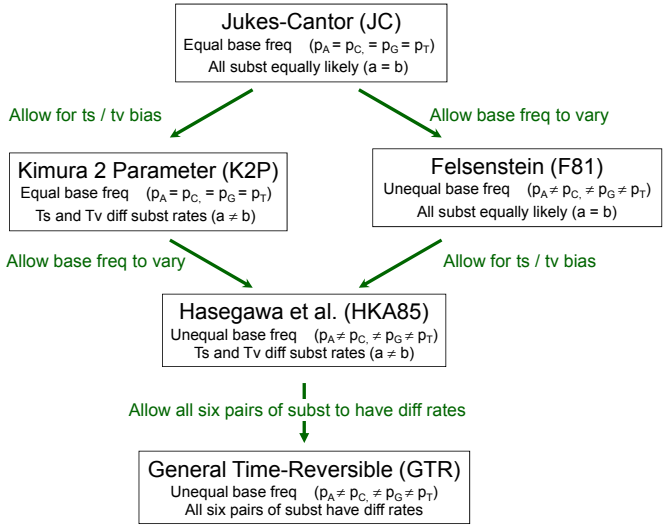
Maximum Parsimony

Maximum Likelihood

Bayesian Methods

These methods attempt to map the history of gene sequences onto a tree. (And decide what the tree looks like)

Models of Sequence Evolution



Adapted from N. Provart & D. Guttman

Phylogenetic Trees

Maximum Parsimony

Works under the principle of "Occam's razor"

Farris (1983), has a justification for parsimony:
"minimizes requirements of ad hoc hypotheses of homoplasy".

Analogy is made between homoplasies and residuals, (part of the data that the tree does not explain), minimizing homoplasies is akin to minimizing residuals in regression.

Based on the assumption that "evolution is parsimonious" which means that there should be no more evolutionary steps than necessary.

The best tree(s) minimize the number of changes between ancestors and descendants.

Under independence of each of the characters, this has a clear combinatorial translation.

Phylogenetic Trees

Maximum Parsimony

Implementation:

- In parsimony, the score is simply the minimum number of mutations that could possibly produce the data.
- Pro: There are fast algorithms that guarantee that any tree can be scored correctly
- Con: There are lots of possible trees to choose between...

Math people:
If you take it in terms of distance on a graph the inner points are what are known as Steiner points and the problem of finding the tree is equivalent to the Steiner tree problem...

Drawbacks:

- the score of a tree is completely determined by the minimum number of mutations among all of the reconstructions of ancestral sequences.
- fails to account for the fact that the number of changes is unlikely to be equal on all branches in the tree.
  - As a result, susceptible to "long-branch attraction", in which two long branches that are not adjacent on the true tree are inferred to be closest relatives
- in practice this is still pretty good... ML/Bayesian better

Phylogenetic Trees

Maximum Likelihood

Attempts to answer the question:

- What is the probability of observing the data, given a particular model of evolution and evolutionary history?
  - data = MSA
  - model = transition probabilities, base frequencies, rate heterogeneity...
  - evolutionary history = phylogenetic tree

Evaluates the likelihood of every substitution of every possible tree.

All possible trees are considered, and the number of substitutions that must have occurred are calculated.

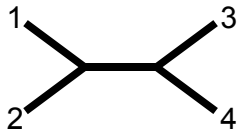
The tree with the highest likelihood is assumed to be the correct tree.

Phylogenetic Trees

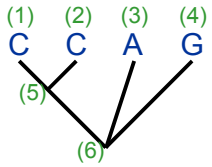
Maximum Likelihood

```
          1.........j........N
1         C...GGACACGTTTA...C
2         C...AGACACCTCTA...C
3         C...GGATAAGTTAA...C
4         C...GGATAGCCTAG...C
```

Unrooted tree for the 4 taxa

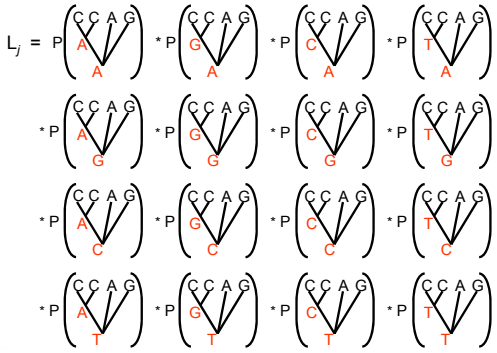


Arbitrarily rooted tree for site *j*



Phylogenetic Trees

Maximum Likelihood



Phylogenetic Trees

Maximum Likelihood

Likelihood of the tree = product of the likelihoods for each site.

$$L = L_1 \times L_2 \times \ldots \times L_N = \prod_{j=1}^{N} L_j$$

Usually evaluated as the sum of the log likelihoods.

$$\ln L = \ln L_1 + \ln L_2 + \ldots + \ln L_N = \sum_{j=1}^{N} \ln L_j$$

ML evaluates:

- all possible ancestral states
  - at all variable site
    - in all possible tree topologies

→The most likely (best) tree is the topology that has the highest overall likelihood.

Phylogenetic Trees

Maximum Likelihood

Advantages of ML methods

- Based on explicit evolutionary models.
- Permits statistical evaluation of the likelihood of specific tree topologies.
- Often returns many equally likely trees.
- Usually outperforms other methods.

Disadvantages

- Computationally very intensive.
- Often returns many equally likely trees.

Adapted from N. Provart & D. Guttman

Bayesian Approach to Phylogeny Estimation

Approach:

Uses the likelihood function

Typically implemented using same models of evolutionary change used in ML

Metropolis-Hastings - Metropolis-Coupled Markov Chain Monte Carlo (MC^3)

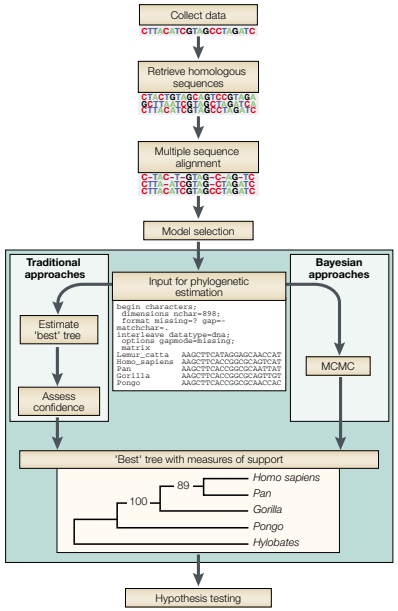
Assumptions:

Same set of parameter choices for evolutionary model as for ML

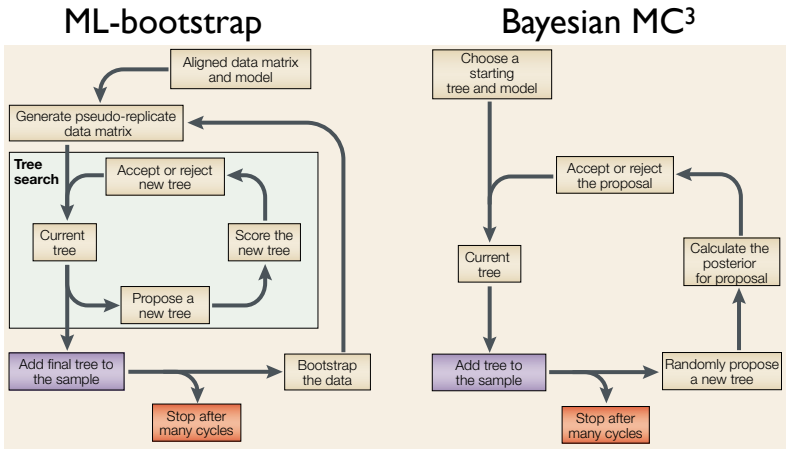
Must also choose initial set of prior probabilities.

Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews Genetics*, *4*(4), 275–284.

Ronquist, F. and J.P. Huelsenbeck. (2003) MrBayes3: Bayesian phylogenetic inference... *Bioinformatics*, **19**, 1572–1574.



Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews Genetics*, *4*(4), 275–284.



Holder, M., & Lewis, P. O. (2003). Phylogeny estimation: traditional and Bayesian approaches. *Nature reviews Genetics*, *4*(4), 275–284.

Phylogenetic Tree Construction Methods — Recommended Software

```
phangorn - MP, ML, and Bayesian tree estimation
ape      - tree-handling in R, tree-build, graphics
picante  -
phyloseq - integrated tree-abundance and graphics
ggtree - ggplot2-specific for trees
```

NJ, UPGMA, PAUP*, PhyML, RaxML, MrBayes (including "cloud" MrBayes)

RAxML — Stamatakis, A. (2006). RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics (Oxford, England)*, *22*(21), 2688–2690.

MrBayes — Huelsenbeck, J. P., & Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics (Oxford, England)*, *17*(8), 754–755.

BEAUti / BEAST 1.7 — Drummond, A. J., Suchard, M. A., Xie, D., & Rambaut, A. (2012). Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Molecular biology and evolution.* http://beast.bio.ed.ac.uk/

http://en.wikipedia.org/wiki/List_of_phylogenetics_software

Phylogenetic Tree Construction Methods

But we're not going to build trees in this workshop…

Why we won't:
- There are many manually-curated public trees
- Optimal tree is not really known, lots to argue over
- For our purposes small differences should not matter

Why you might want to calculate a new tree:
- You have counts from non-16S rRNA gene
- Have concatenated whole genome sequence data
- Basically any time you have new biological sequence data for which a public reference tree is not available

Tree file format, data representation: Newick

Green Genes Tree in Newick format:

```
(((((((836:0.06877,
((549322:0.00892,522457:0.01408)1.000:0. ,
314761:0.09977)0.161:0.01566)0.882:0.00924,
(((311539:0.0484 (((174835:0.01627,
(34207:0.00082,45996:0.00334)0.863:0.00433
1.000.3:0.09792)1.000.4:0.04652,(((((945:0.08077,
(178877:0.01342,
(29928:0.00726,35548:0.00187)0.748:0.01216)
1.000.5:0.05924)0.975:0.01729, ...;
```

A simple Newick tree with branch lengths is noted:
((1 : 1, 4 : 1) : 3, ((2 : 1, 3 : 1), 5 : 2) : 1);

http://evolution.genetics.washington.edu/phylip/newick_doc.html

Tree file format, data representation: phylo (ape)

Terminology and Notations:

branch: edge, vertex
node: internal node
degree: the number of edges that meet at a node
tip: terminal node, leaf, node of degree 1
n: number of tips
m: number of nodes

http://ape-package.ird.fr/misc/FormatTreeR_24Oct2012.pdf

Tree file format, data representation: phylo (ape)

Definition of the Class "phylo"
The class "phylo" is used to code "acyclical" phylogenetic trees. These trees have no reticulations, and all their internal nodes are of degree 3 or more, except the root (in the case of rooted trees) which is of degree 2 or more. An object of class "phylo" is a list with the following mandatory elements:

1. A numeric matrix named `edge` with two columns and as many rows as there are branches in the tree;
2. A character vector of length n named `tip.label` with the labels of the tips;
3. An integer value named `Nnode` giving the number of (internal) nodes;
4. An attribute class equal to "phylo".

In the matrix edge, each branch is coded by the nodes it connects: tips are coded 1, . . . , n, and internal nodes are coded n+ 1, . . . , n+m (n+ 1 is the root). Both series are numbered without gaps.

`edge.length`, `node.label`, `root.edge` are optional annotation slots in "phylo" list

http://ape-package.ird.fr/misc/FormatTreeR_24Oct2012.pdf

Tree file format, data representation: phylo (ape)

The "ape::phylo" edge-matrix has the following properties:

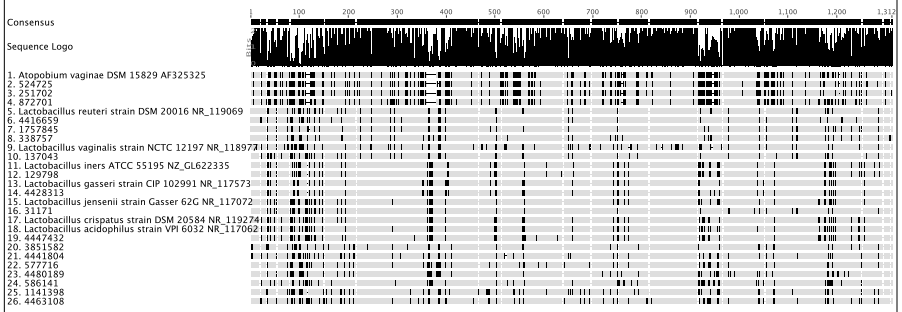
1. The first column has only values greater than n (thus, values less than or equal to n appear only in the second column).
2. All nodes appear in the first column at least twice.
3. The number of occurrences of a node in the first column is related to the nature of the node: twice if it is dichotomous (i.e., of degree 3), three times if it is trichotomous (degree 4), and so on.
4. All elements, except the root n + 1, appear once in the second column.

Example Tree Plots: "How to *Read* a Tree"

Exercise:
Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database

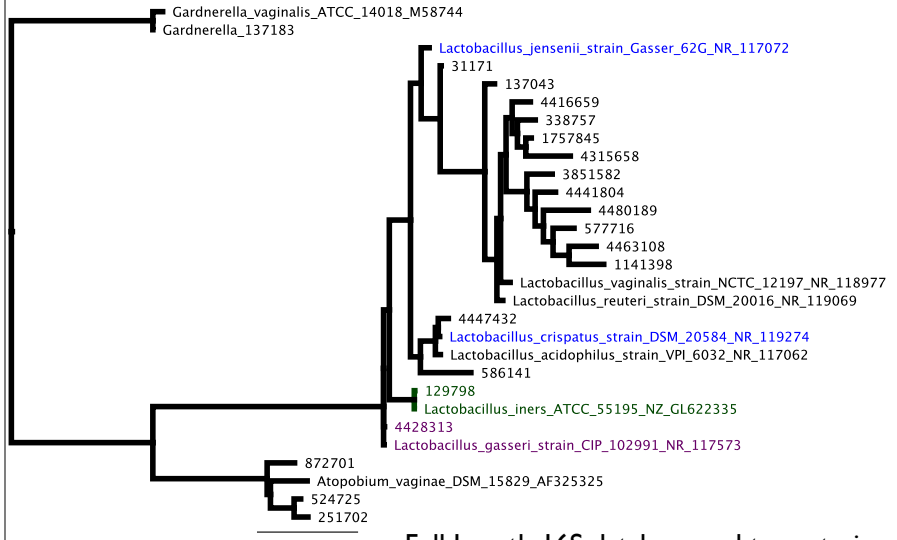
Research Motivation:
Does the region of 16S rRNA gene in my data actually discriminate *Lactobacillus* species?

Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



Full Length 16S database and type strains

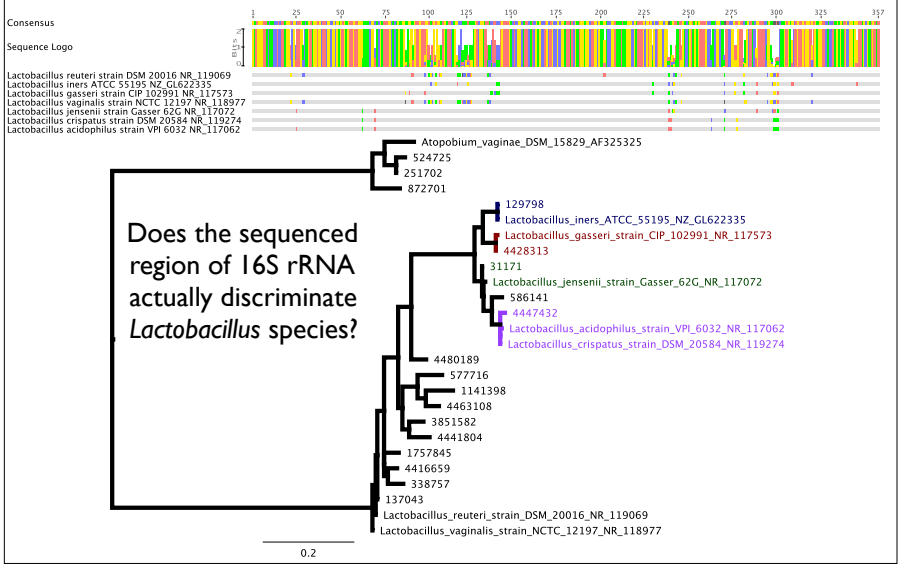
Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



Full Length 16S database and type strains

41

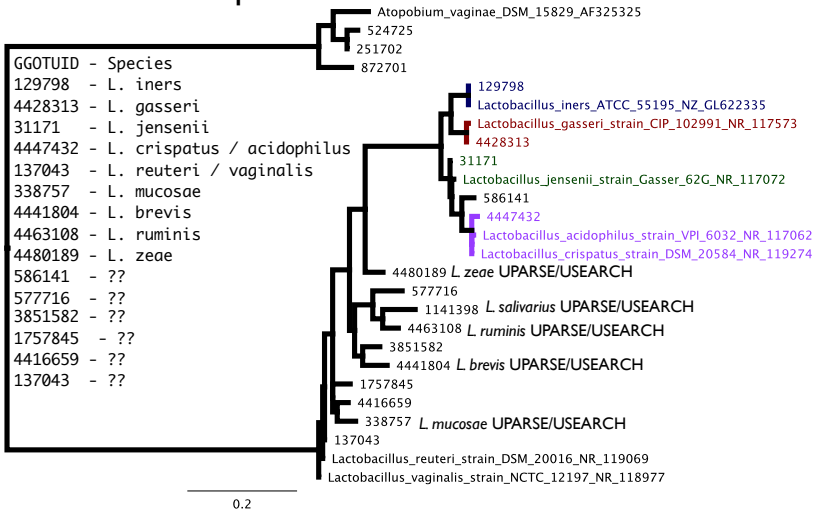
Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



Does the sequenced region of 16S rRNA actually discriminate *Lactobacillus* species?

42

Example 1: Determine species names of unlabeled *Lactobacillus* species in the GreenGenes database



43

Manipulating Trees in phyloseq

- Trees are automatically pruned to match data operations on other parts of phyloseq object
- Use standard taxa functions
  - prune_taxa(), filter_taxa(), subset_taxa()
- Agglomeration
  - tip_glom()
  - tax_glom()
- ape functions after accession:
  - plot.tree(phy_tree(physeq))
  - root(phy_tree(physeq), …)

44

(Tree-based) Distances between microbiomes

Community Distance

Communities are a vector of abundances:
$\mathbf{x} = \{x_1, x_2, x_3, \ldots\}$

E. coli: ●●●
P. fluorescens: ●
B. subtilis: ●
P. acnes:
D. radiodurans:
H. pylori: ●●●●●●●
L. crispatus:

$\mathbf{x} = \{3,1,1,0,0,7,0\}$

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Community Distance Properties

- Range from 0 to 1
- Distance to self is 0
- If no shared taxa, distance is 1
- Triangle inequality (metric)
- Joint absences do not affect distance (biology)
- Independent of absolute counts (metagenomics)

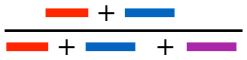
Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

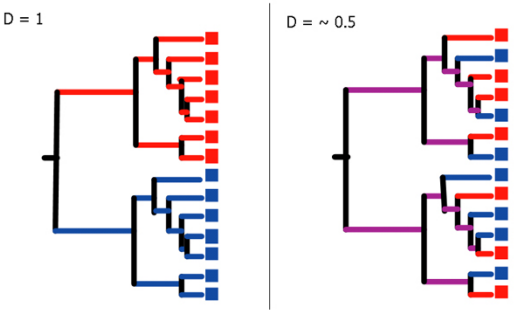
The Distance Spectrum

	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Unifrac

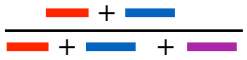
Dist(x, y) = 

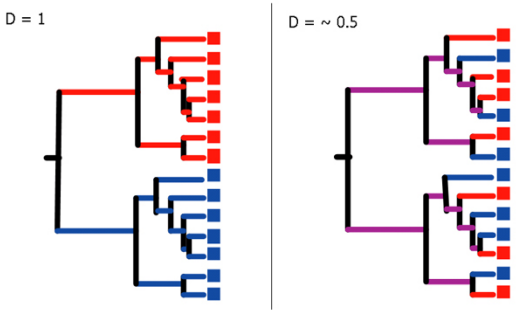


Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone and Knight (2008)

Unifrac

Dist(x, y) = 

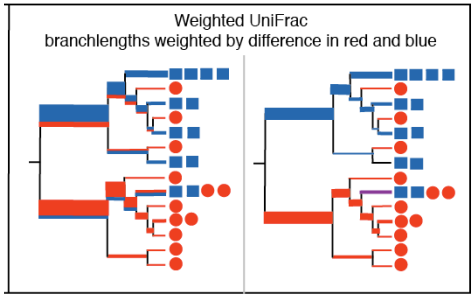


Intuition: Fraction of shared **tree** unique to one of the communities

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone and Knight (2008)

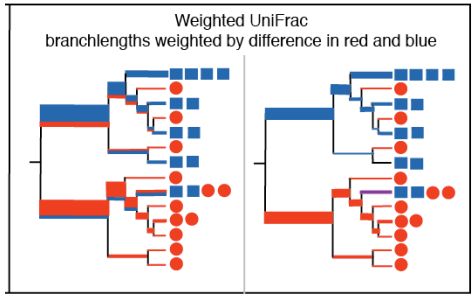
Weighted Unifrac



Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

Lozupone et al. (2007)

Weighted Unifrac



Intuition: The cost of turning one distribution into the other; where the cost is the amount of "dirt" moved times the distance by which it is moved.

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

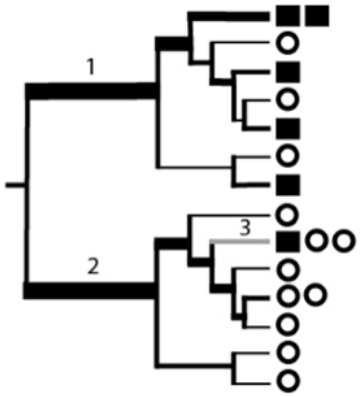
Lozupone et al. (2007)

Weighted UniFrac Distance

A modification of (unweighted) UniFrac

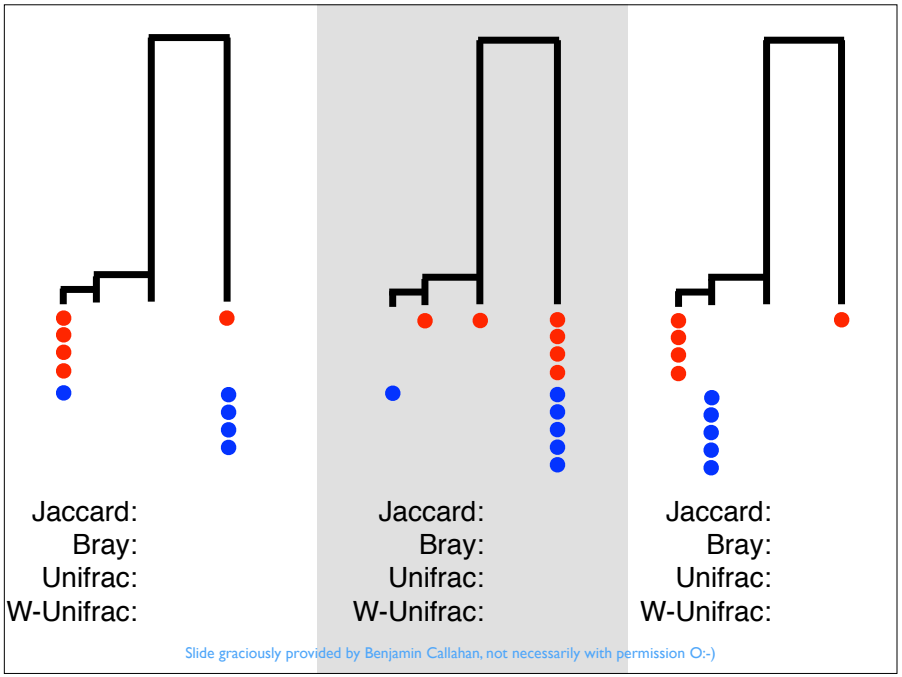
$$\sum_{i=1}^{n} b_i \times \left| \frac{A_i}{A_T} - \frac{B_i}{B_T} \right|$$

- n = number of branches in the
- b_i = length of the ith branch
- A_i = number of descendants of ith branch in group A
- A_T = total number of sequences in group A



Lozupone et al., 2007

53



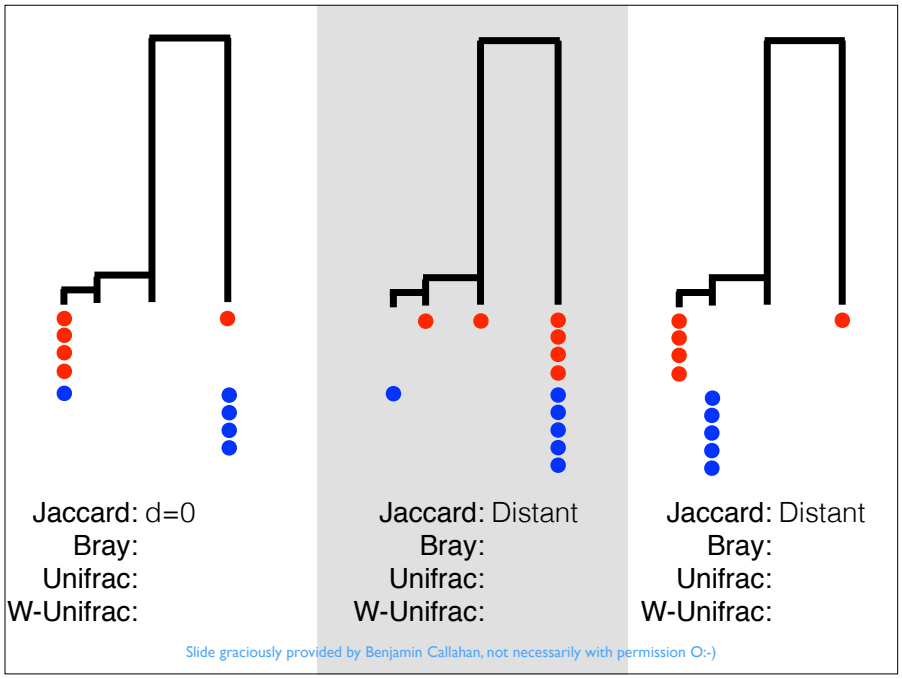
Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

Jaccard:
Bray:
Unifrac:
W-Unifrac:

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

54



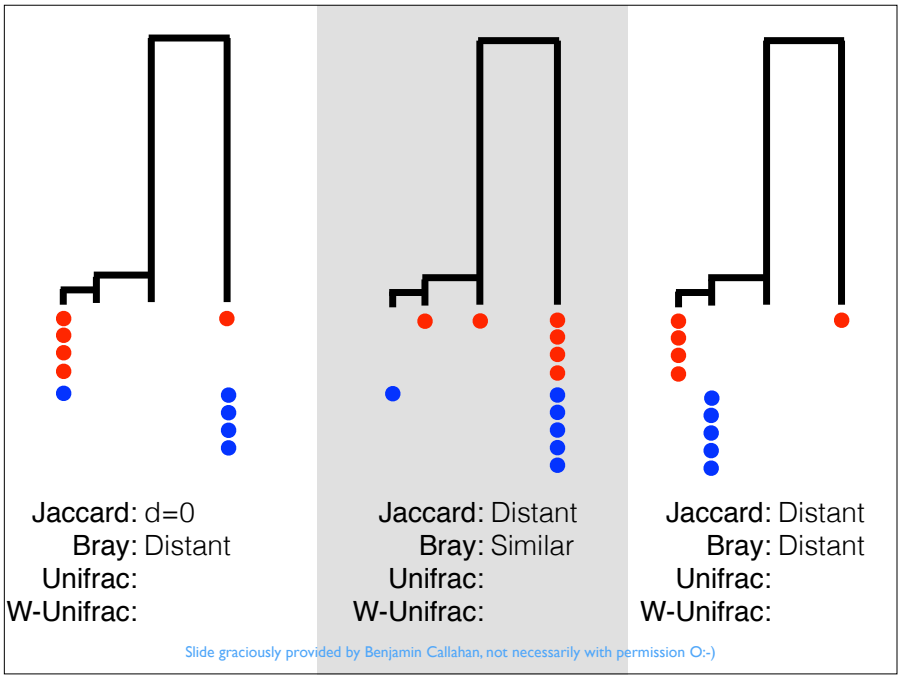
Jaccard: d=0
Bray:
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray:
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray:
Unifrac:
W-Unifrac:

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

55



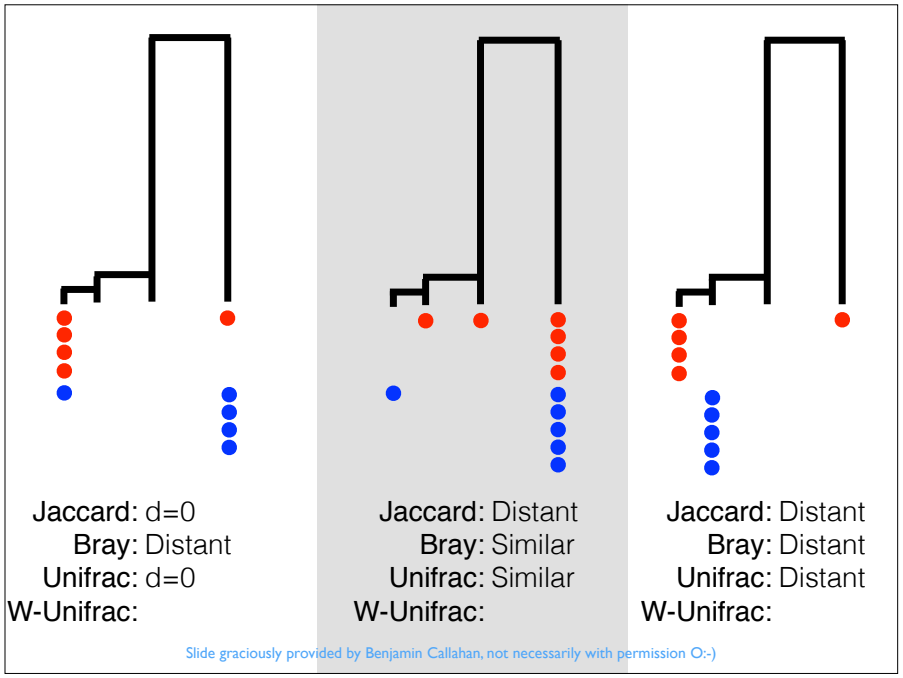
Jaccard: d=0
Bray: Distant
Unifrac:
W-Unifrac:

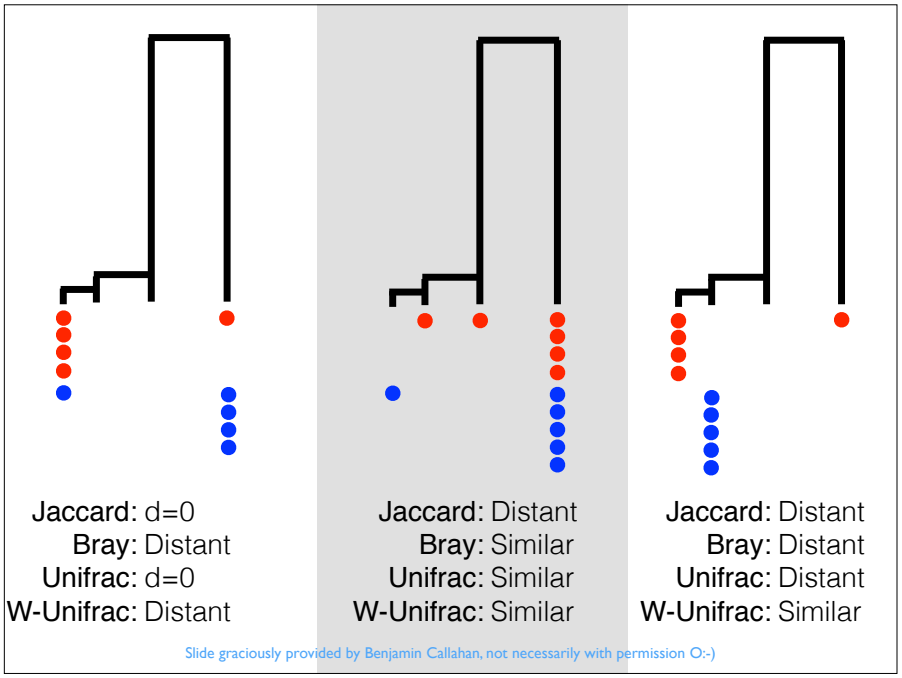
Jaccard: Distant
Bray: Similar
Unifrac:
W-Unifrac:

Jaccard: Distant
Bray: Distant
Unifrac:
W-Unifrac:

Slide graciously provided by Benjamin Callahan, not necessarily with permission O:-)

56





The Distance Spectrum

	Categorical	Phylogenetic
Presence/ Absence	Jaccard	Unifrac
Quantitative Abundance	Bray-Curtis	Weighted Unifrac

phyloseq distances

manhattan
euclidean
canberra
bray
kulczynski
jaccard
gower
altGower
morisita-horn
mountford
raup
binomial
chao
cao
jensen-shannon
unifrac
weighted-unifrac
...

That's great, Joey… What do we do with these distances???

Alex is going to go over ordination methods for interpreting the distance matrix derived from comparing all the samples in your data…

What we learned here was…

- A survey about how to think about trees
- How trees are represented and interact with phyloseq
- An introduction about different definitions for a distance between two microbiomes

End

Comparison of UniFrac and DPCoA

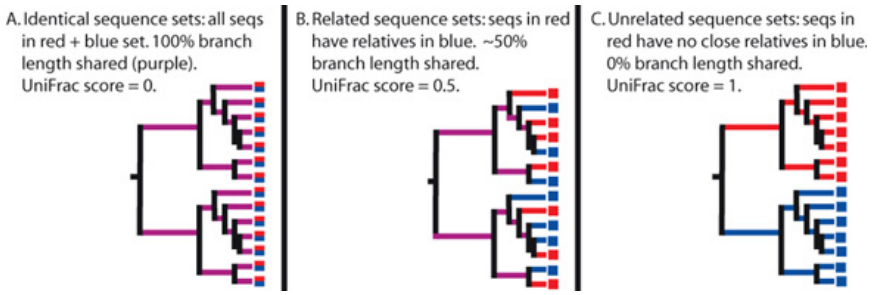
Original description	New formula	Properties
square root of Rao's distance based on the square root of the patristic distances	$[\sum_i b_i (A_i/A_T - B_i/B_T)^2]^{1/2}$	Most sensitive to outliers, least sensitive to noise, upweights deep differences, gives OTU locations
$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	$\sum_i b_i \lvert A_i/A_T - B_i/B_T \rvert$	Less sensitive to outliers/more sensitive to noise than DPCoA
fraction of branches leading to exactly one group	$\sum_i b_i 1\{\frac{A_i/A_T - B_i/B_T}{A_i/A_T + B_i/B_T} \geq 1\}$	Sensitive to noise, upweights shallow differences on the tree

Summary of the methods under consideration. "Outliers" refers to highly abundant OTUs, and noise refers to noise in detecting low-abundance OTUs (see Fukuyama and Holmes, 2012)

Pavoine, et al. (2004). From dissimilarities among species to dissimilarities among communities: a double principal coordinate analysis. *Journal of Theoretical Biology*, *228*(4), 523–537

(Unweighted) UniFrac Distance

A proposal for using the phylogenetic tree and OTU table

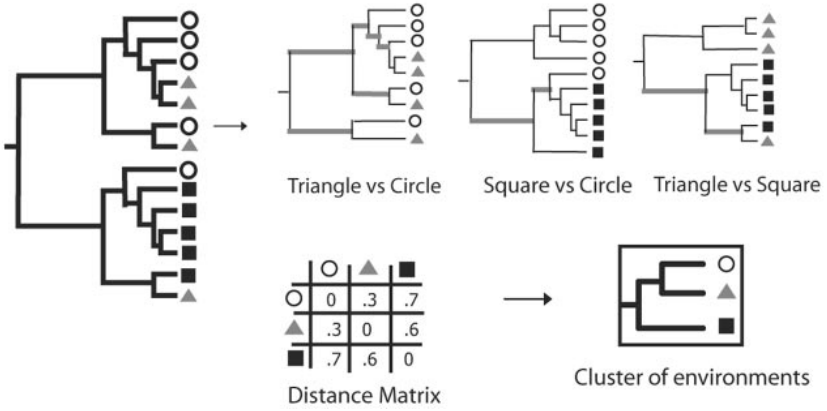


"Since we compared environments on a large scale, the ability of particular lineages of organisms to survive in each environment is more likely to represent the relevant aspects of similarity between environments than the relative abundance of each surviving lineage"

Lozupone & Knight (2005) *Applied and Environmental Microbiology*

(Unweighted) UniFrac Distance

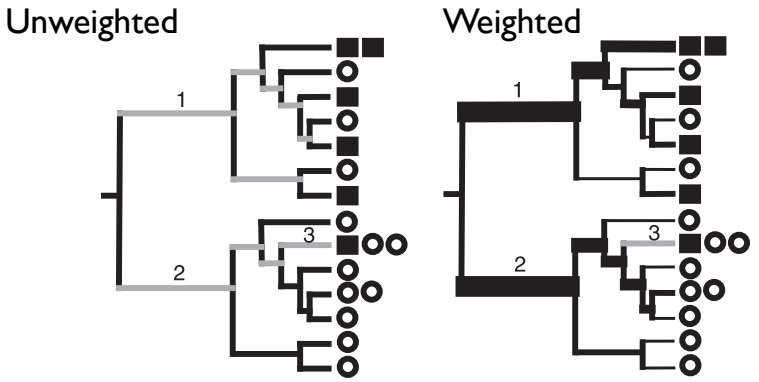
A proposal for using the phylogenetic tree and OTU table



Lozupone & Knight (2005) *Applied and Environmental Microbiology*

UniFrac Comparison

(Fraction of branch lengths not shared)



gray branches have no weight

Lozupone, et al (2007). Quantitative and qualitative… *Applied and Environmental Microbiology*