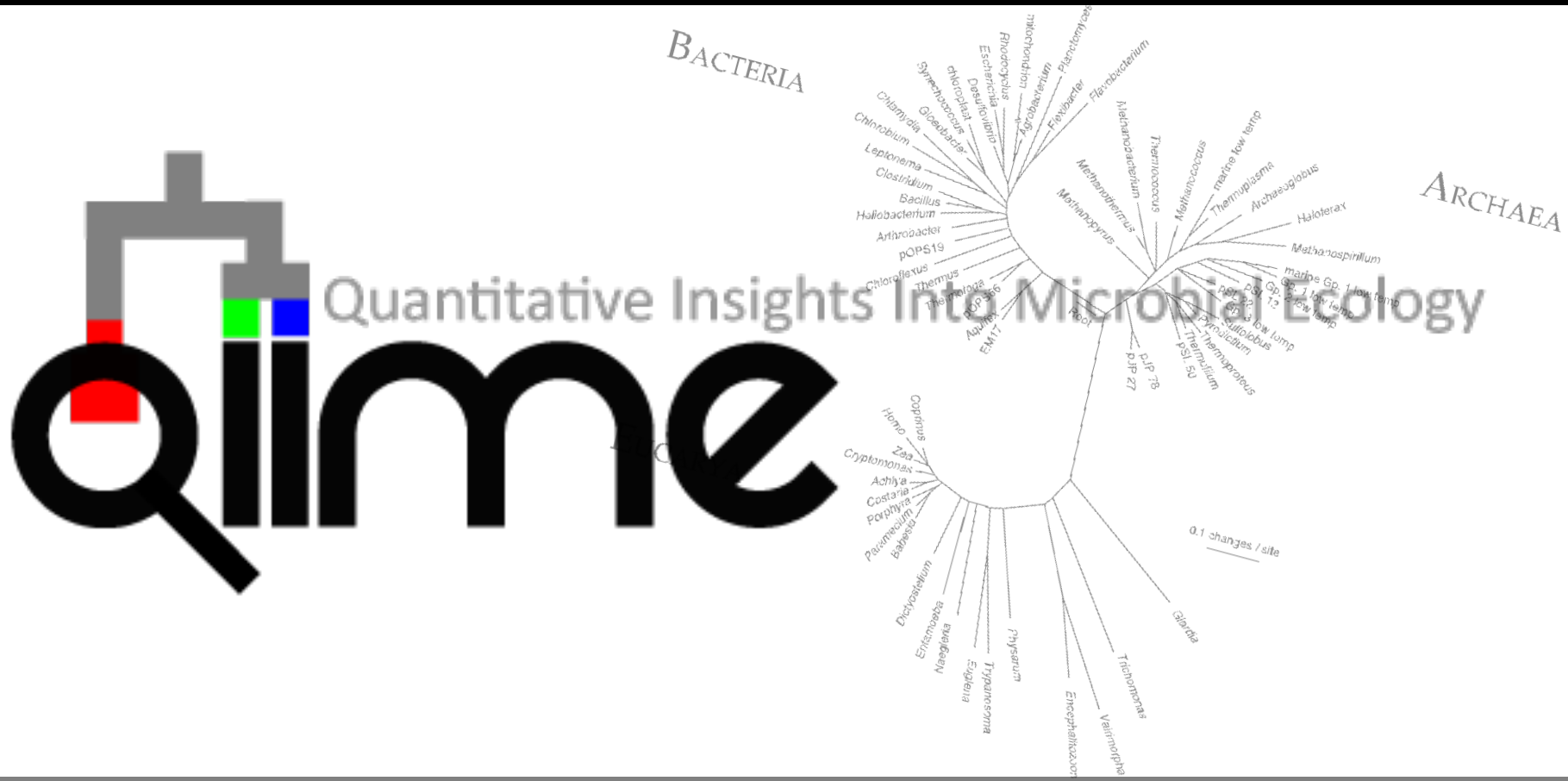
Microbial Communities Profiling via QIIME



KNIGHTLAB UCSD

Today's schedule

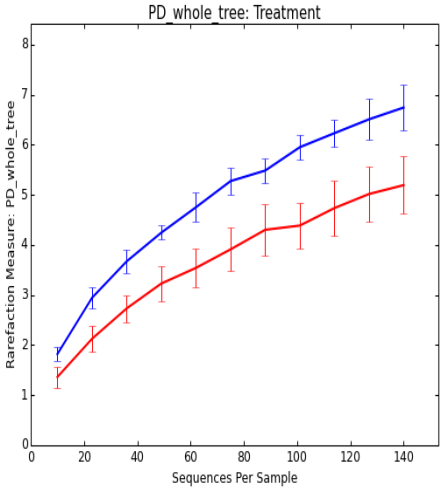
- Day 3.

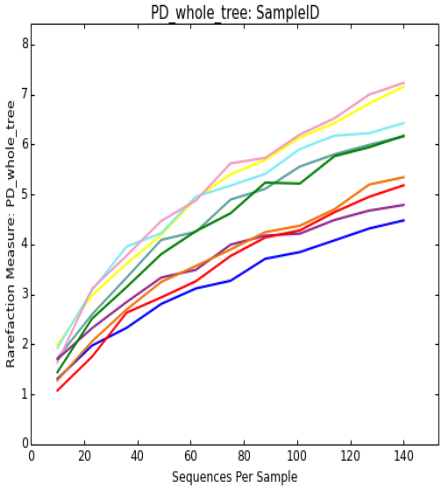
  9:00am – 3:00pm: Diversity Analyses and Running Commands

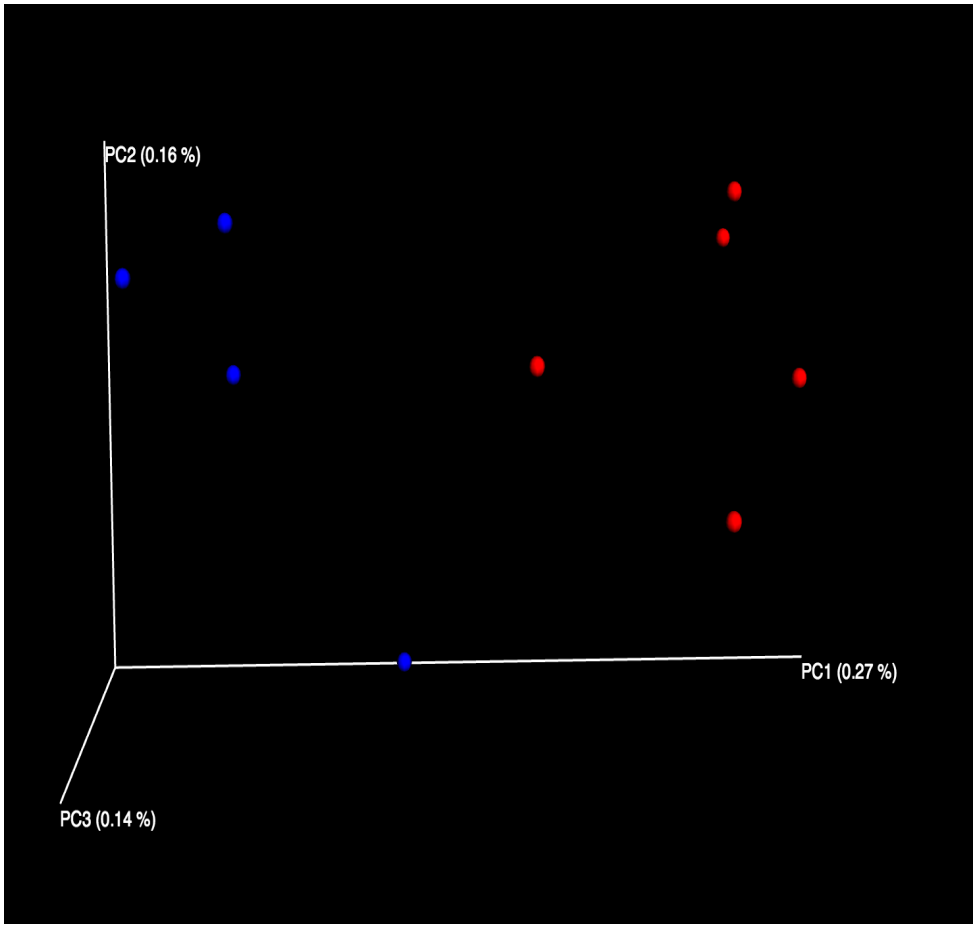
  3:30 – 5:00pm: Discuss Group Project

Diversity Analyses

Diversity analysis

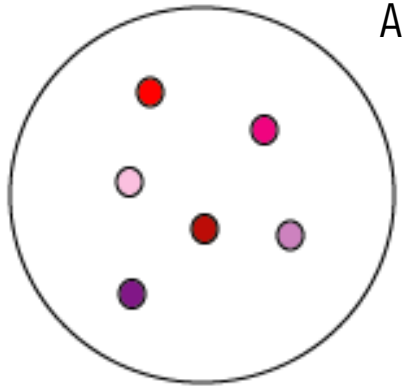


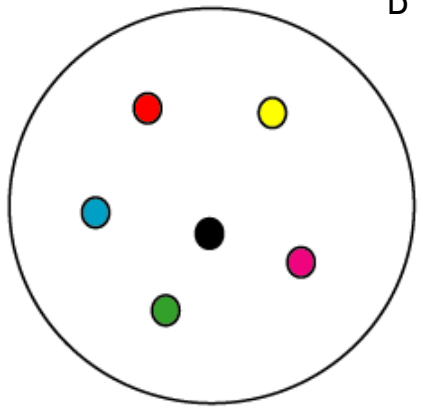




How do we describe and compare diversity?

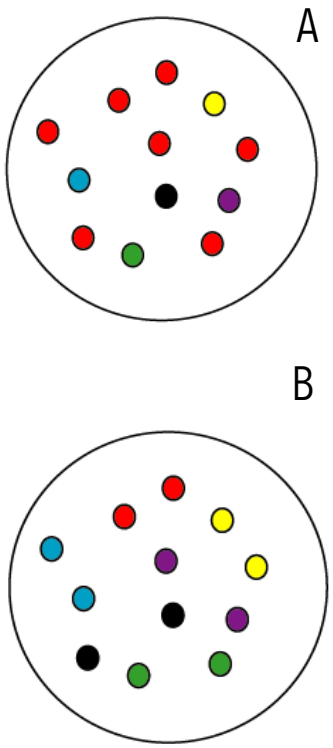
- α Diversity:
  - "How many species are in a sample?"
    - (e.g. 6 colors in A and 6 in B)
  - e.g.: Are polluted environments less diverse than pristine?
- β Diversity:
  - "How many species are shared between samples?"
    - (e.g. 2 shared colors between A and B)
  - e.g.: Does the microbiota differ with different disease states?





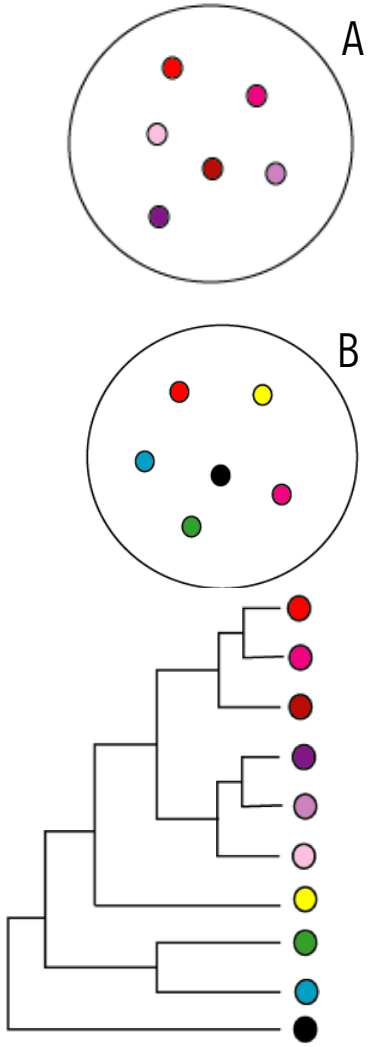
Quantitative versus Qualitative measures

- Qualitative: Considers presence absence only
  - α: How many species are in a sample?
    - e.g.: 6 colors in both A and B.
  - β: How many species are shared between samples?
    - e.g.: A and B are identical because the same colors are present in both.
- Quantitative: Also considers relative abundance.
  - α: Accounts for "evenness":
    - e.g. B, where the population is evenly distributed across the 6 species, is more diverse than A, where all species are present but red dominates.
  - β: Samples will be considered more similar if the same species are numerically dominant versus rare.
    - e.g. B and A no longer look identical because of differences in abundance.



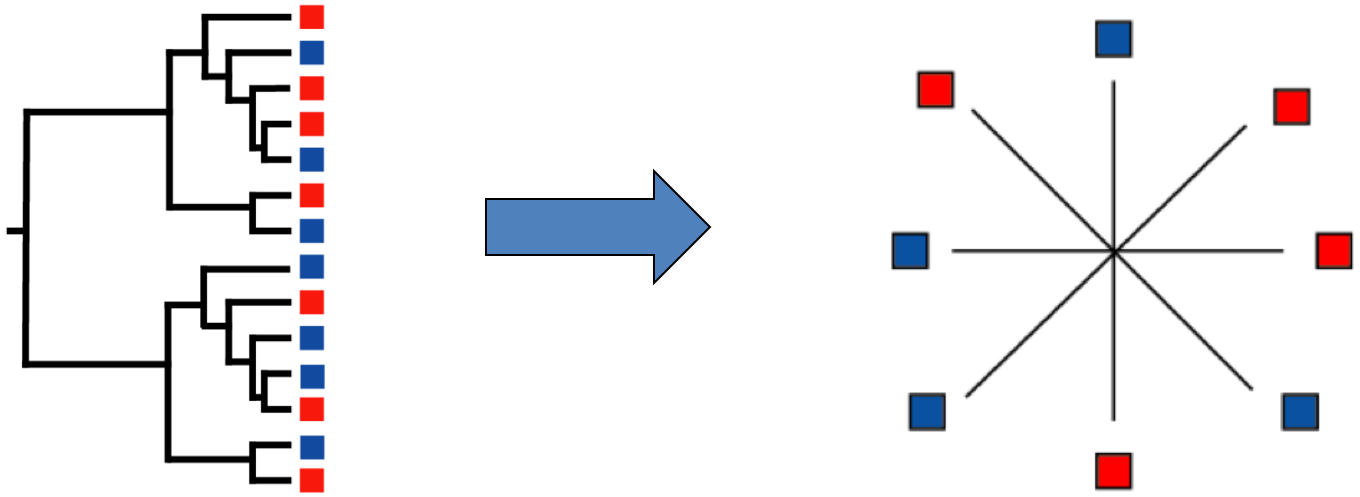
What is a phylogenetic diversity measure?

- α Diversity:
  - Taxon: "How many species are in a sample?"
  - Phylogenetic: "How much phylogenetic divergence is in a sample?"
    - (e.g. B more individually diverse than A - more divergent colors)
- β Diversity:
  - Taxon: "How many species are shared between samples?"
  - Phylogenetic: "How much phylogenetic distance is shared between samples?"
    - (only related colors from B are in A)



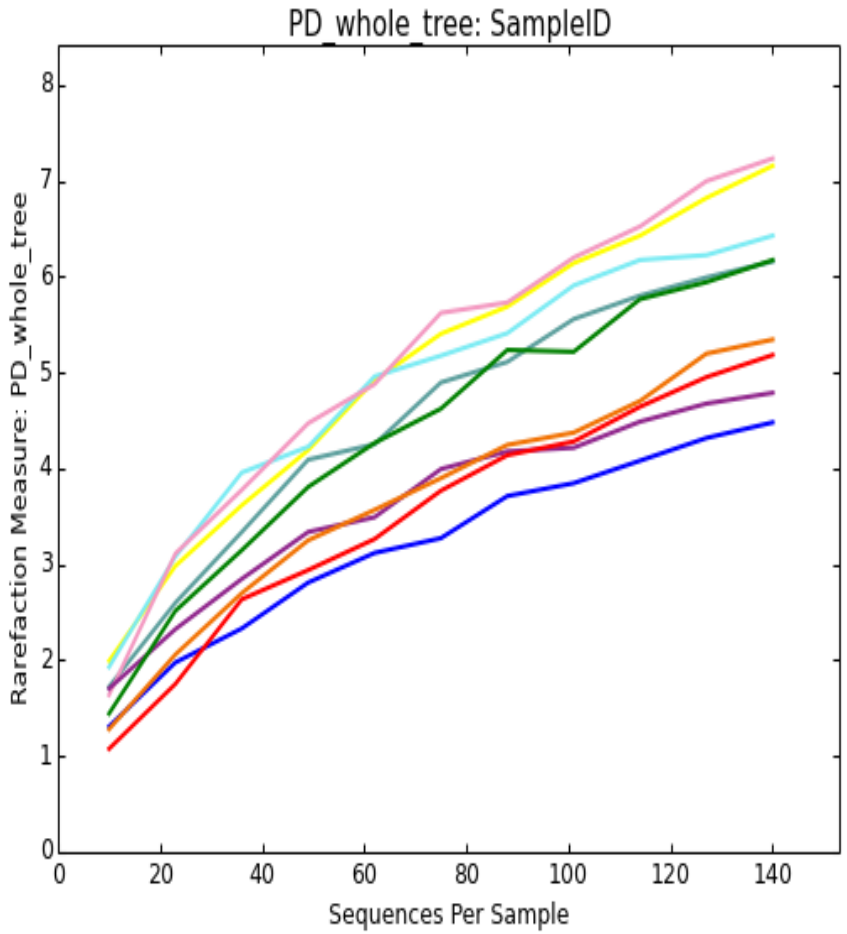
Advantages of phylogenetic techniques.

- Phylogenetically related organisms are more likely to have similar roles in a community.
- Taxon-based methods assume a "star phylogeny"
- Phylogeny and Taxon-based methods can be complementary.



Alpha diversity

Alpha diversity



Basic alpha diversity measure: count number of OTUs.

other measures can be:

- phylogenetic (PD)
- estimators (chao1)
- other statistics (evenness)

PD Rarefaction

- Plot the amount of branch length against the # of observations.
- Shape of curve allows for estimating how far we are from sampling all of the phylogenetic diversity.
- Allows for comparison of phylogenetic diversity between samples.

Comparing diversity

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Observed species

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Observed species:

Sample A 3
Sample B 3
Sample C 3

Conclusion:

Samples A, B, and C are equally diverse.

Phylogenetic Diversity (PD)

Sample A

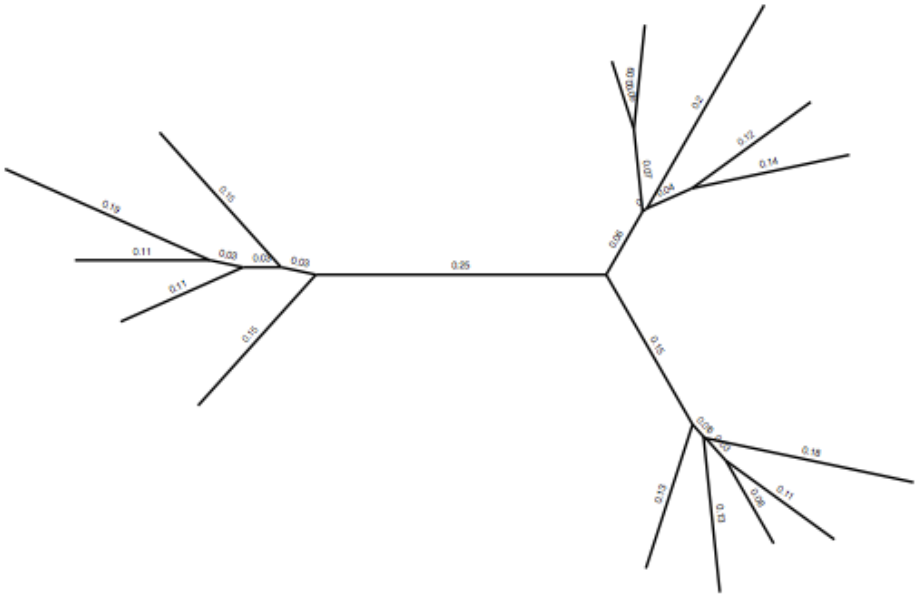
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

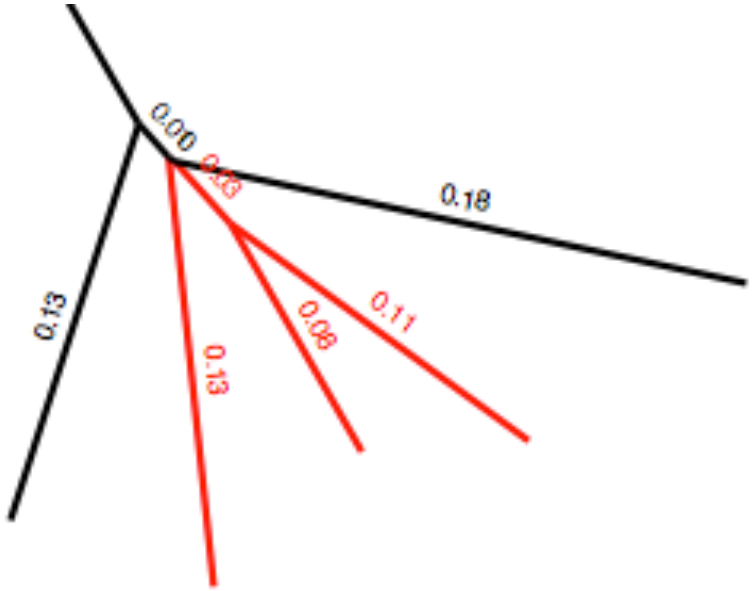
Sample C

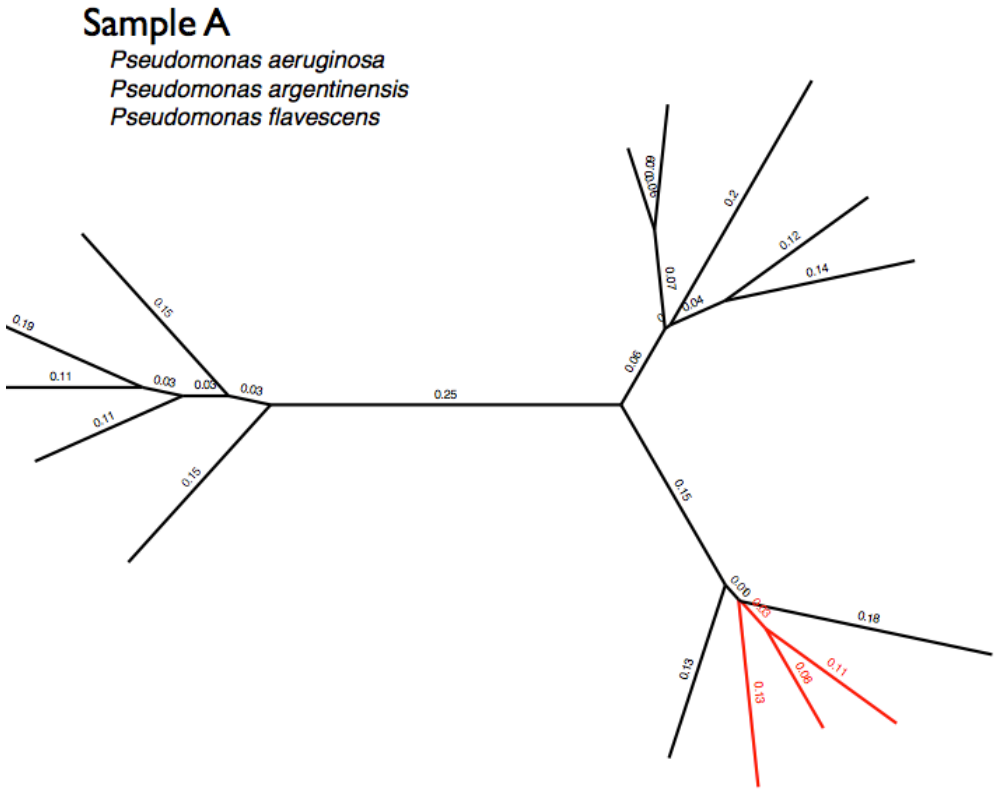
Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

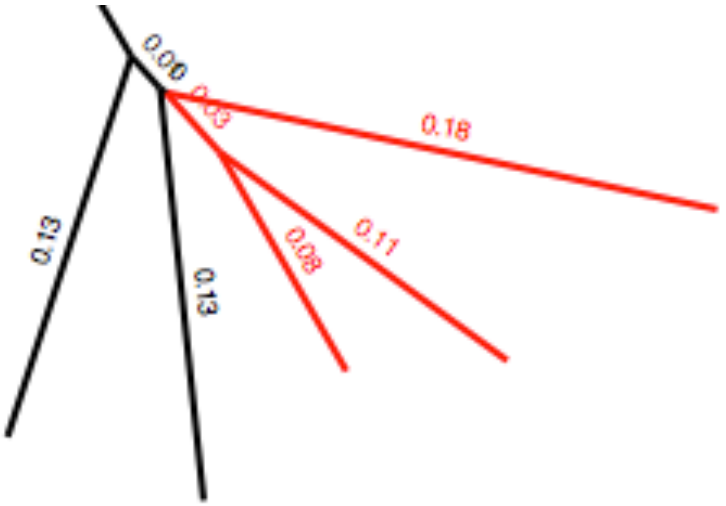
Phylogenetic Diversity (PD)

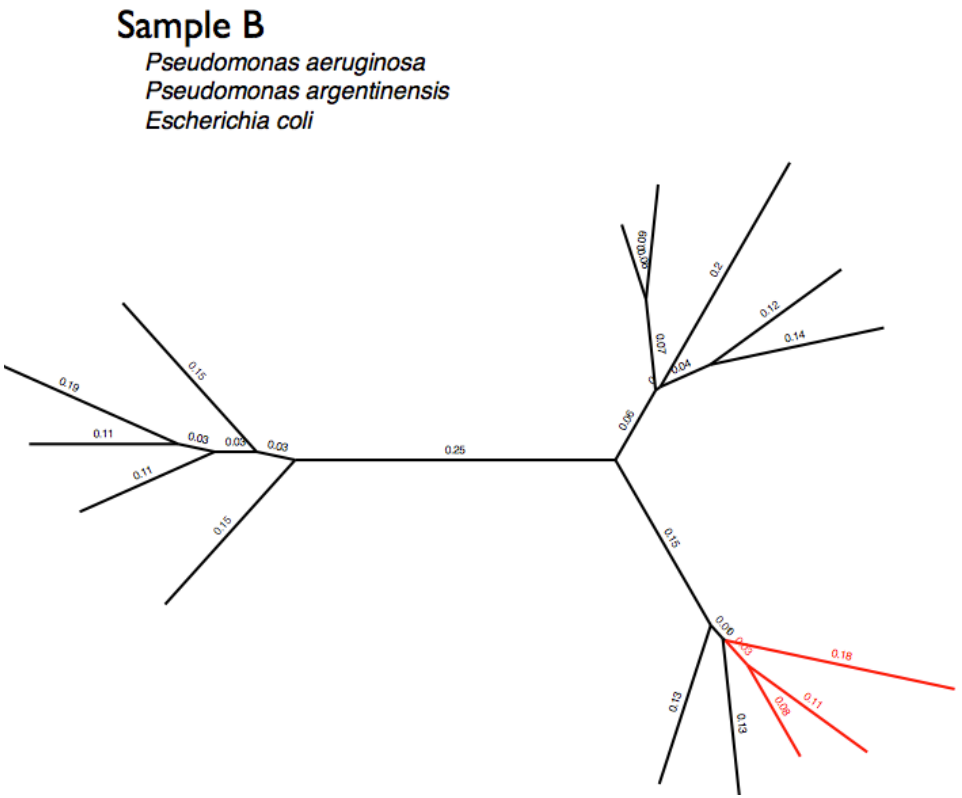




PD = 0.13 + 0.03 + 0.11 + 0.08 = 0.35

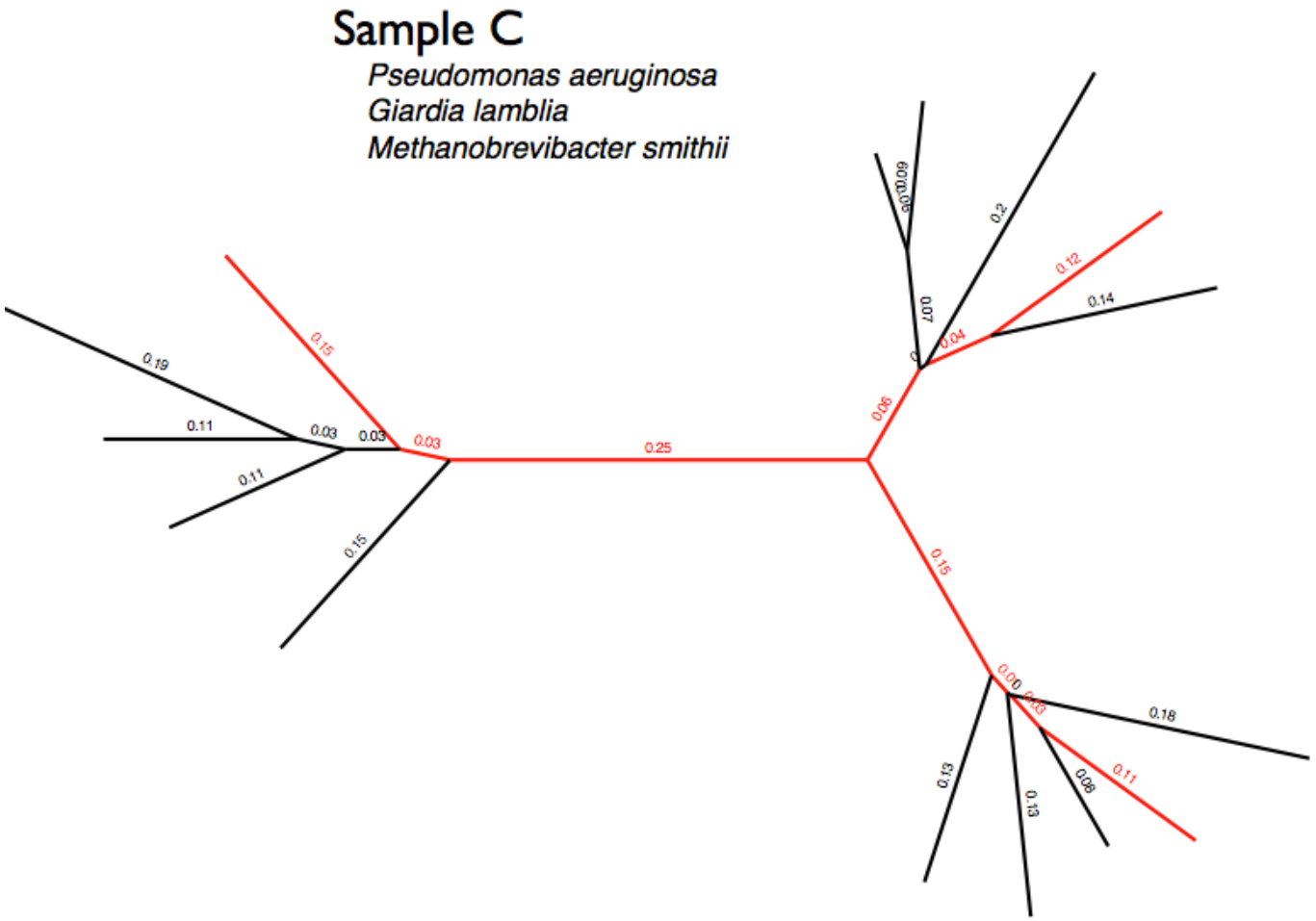
Phylogenetic Diversity (PD)





PD = 0.18 + 0.03 + 0.11 + 0.08 = 0.40

Phylogenetic Diversity (PD)



PD = 0.15+0.03+0.25+0.06+0.04+0.12+0.15+0.01+0.03+0.11= 0.95

Phylogenetic Diversity (PD)

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

PD = 0.35

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

PD = 0.40

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

PD = 0.95

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Phylogenetic Diversity (PD)

Sample A
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B
Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C
Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

PD = 0.35 < PD = 0.40 < PD = 0.95

Conclusion:

Sample C is more diverse than sample B, which is more diverse than sample A.

Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Beta diversity

Should you rarefy?

PLOS | COMPUTATIONAL BIOLOGY

Browse | For Authors | About Us

Search

advanced search

OPEN ACCESS PEER-REVIEWED

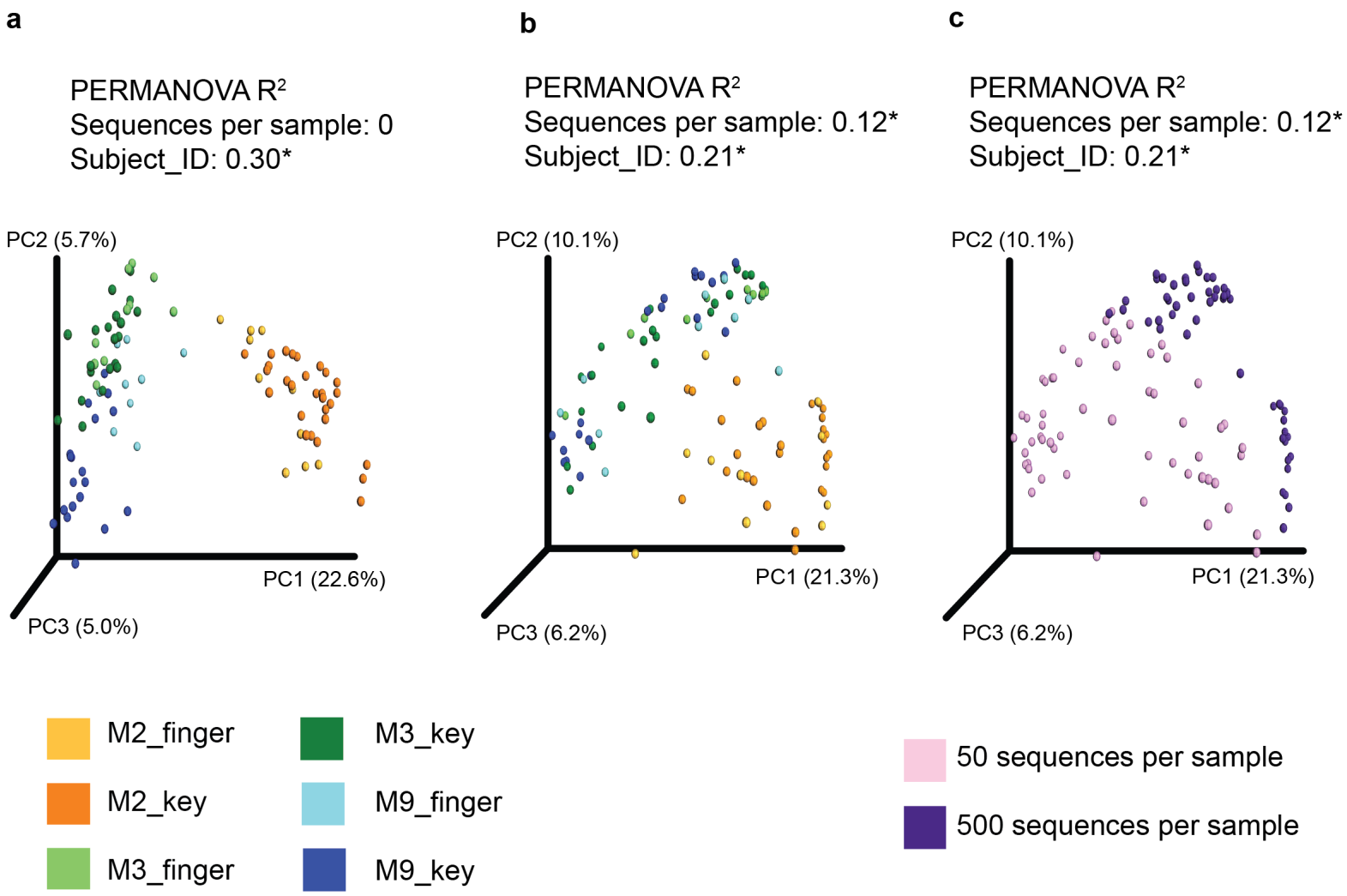
RESEARCH ARTICLE

Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible

Paul J. McMurdie, Susan Holmes

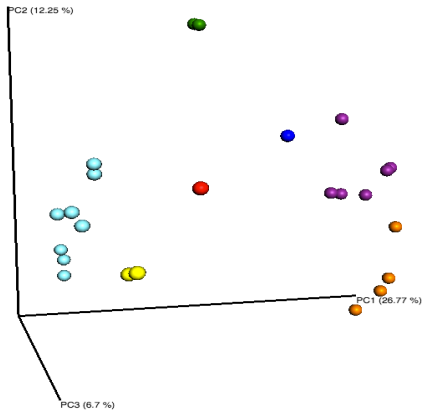
Published: April 3, 2014 • DOI: 10.1371/journal.pcbi.1003531

Data normalization

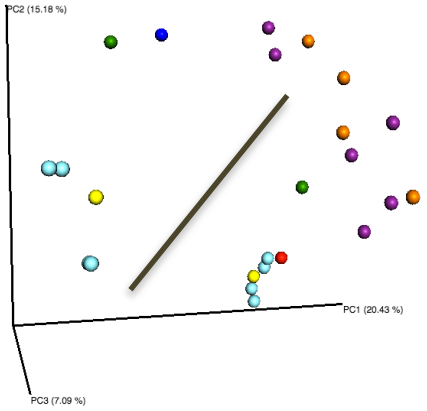


Normalization Methods on Actual Data, Unweighted UniFrac

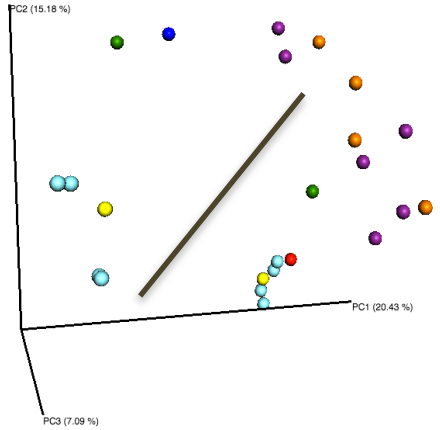
rarefy: PERMANOVA R^2
Library_Size: 0.045*
Host_Bodysite: 0.56***



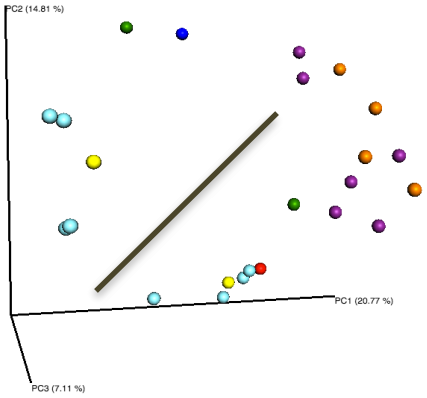
None: PERMANOVA R^2
Library_Size: 0.14***
Host_Bodysite: 0.40***



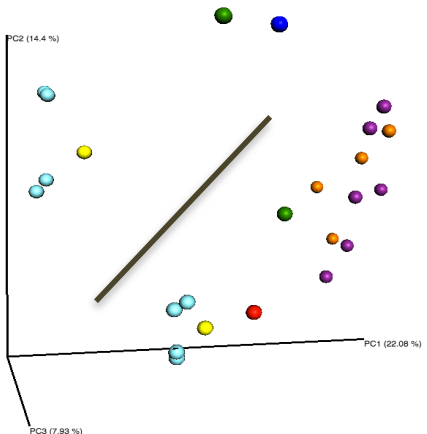
Proportion: PERMANOVA R^2
Library_Size: 0.14***
Host_Bodysite: 0.40***

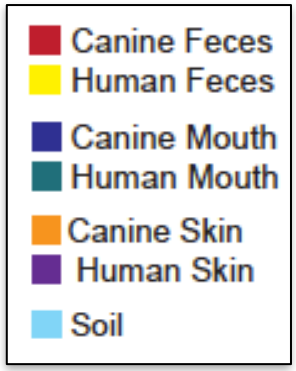


CSS: PERMANOVA R^2
Library_Size: 0.13***
Host_Bodysite: 0.42***



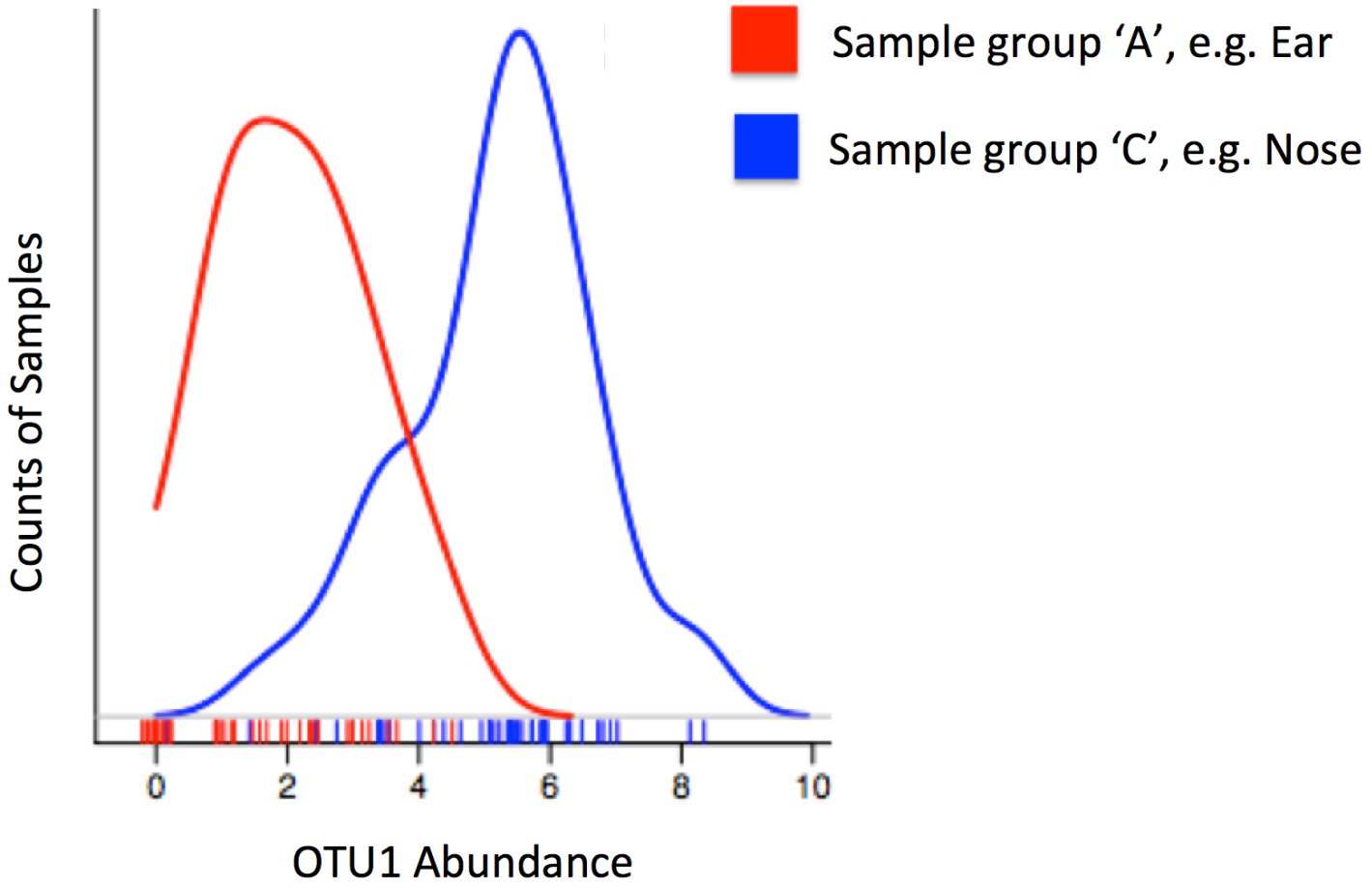
DESeq: PERMANOVA R^2
Library_Size: 0.12***
Host_Bodysite: 0.46***



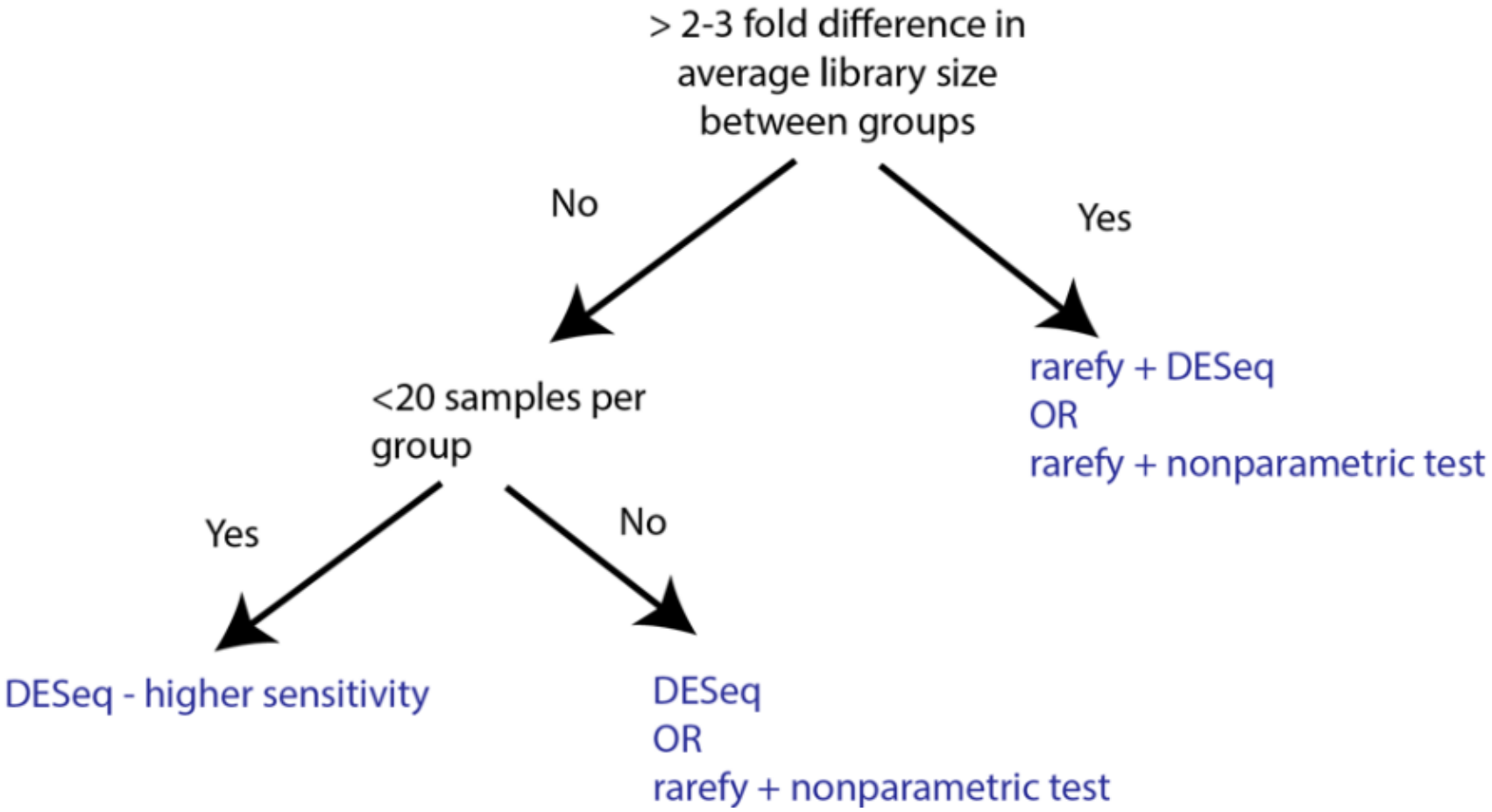


Weiss et al., *in review*. (Pre-print available)

Differential Abundance Testing: Finding Significant Changes in Abundance



Proposed Differential Abundance Testing Strategy



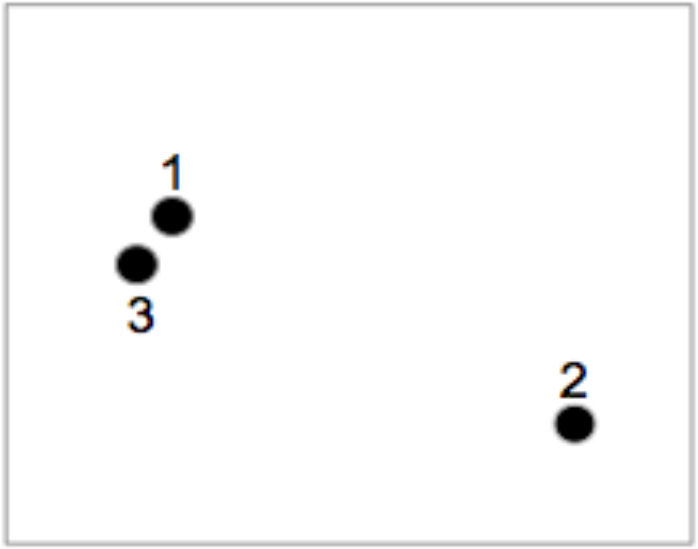
Weiss et al., *in review*. (Pre-print available)

Just to be clear

Normalization ≠ Differential abundance testing

Beta diversity

	1	2	3
OTU1	4	0	4
OTU2	4	0	4
OTU3	0	7	1
OTU4	0	7	0
…	…	…	…



Principal Coordinates Analysis (PCoA)

- distance based multivariate approach
- each sample becomes one point, and the closer together points are, the more similar the samples they represent are
- **dependent on positions of all other points**

name	formula	ref
bray-curtis	$D_{ab} = \frac{\sum_j \lvert A_{aj} - A_{bj} \rvert}{\sum_j (A_{aj} \quad A_{bj})}$	[1]
canberra	$D_{ab} = \frac{1}{Num_{a \cap b}} \sum_j \frac{A_{aj} \quad A_{bj}\rvert}{A_{aj} + A_{bj}}$	[1]
χ^2 (chi-squared)	$D_{ab} = \sqrt{A_{++} \sum_j \left(\frac{1}{A_{-j}} \left(\frac{A_{aj}}{A_{a-}} - \frac{A_{bj}}{A_{b+}} \right)^2 \right)}$	[1]
chord	$D_{ab} = \sqrt{\sum_j \left(\frac{A_{aj}}{\sqrt{\sum_j A_{aj}^2}} \quad \frac{A_{bj}}{\sqrt{\sum_j A_{bj}^2}} \right)^2}$	[2]
euclidean	$D_{ab} = \sqrt{\sum_j (A_{aj} \quad A_{bj})^2}$	[1]
gower	$D_{ab} = \sum_j \frac{\lvert A_{aj} - A_{bj} \rvert}{(max_{\forall i} A_{ij} \quad min_{\forall i} A_{ij})}$	[1]
he⎵nger	$D_{ab} = \sqrt{\sum_j \left(\sqrt{\frac{A_{aj}}{A_{a\cdot}}} - \sqrt{\frac{A_{bj}}{A_{b\cdot}}} \right)^2}$	[3]
kulczynski (quantitative)	$D_{ab} \quad 1 - \frac{1}{2} \left(\frac{\sum_j min(A_{aj}, A_{bj})}{\sum_j A_{aj}} - \frac{\sum_j min(A_{aj}, A_{bj})}{\sum_j A_{bj}} \right)$	[1]
manhattan	$D_{ab} \quad \sum \lvert A_{aj} - A_{bj} \rvert$	[1]
morisita-horn	$\frac{2 \sum_j^J A_{aj} A_{bj}}{\left(\frac{\sum_j A_{aj}^2}{\sum_j A_{aj}} \quad \frac{\sum_j A_{bj}^2}{\sum_j A_{bj}} \right) \sum_j A_{aj} * \sum_j A_{bj}}$	[4]
pearson	$D_{ab} - 1 - \frac{\sum_j (A_{aj} \quad (\frac{A_a}{M})) \sum_j (A_{bj} \quad (\frac{A_b}{M}))}{\sqrt{\sum_i (A_{aj} - (\frac{A_a}{M}))^2} \sqrt{\sum_i (A_{aj} - (\frac{A_b}{M}))^2}}$	[5]
soergel	$D_{ab} = \frac{\sum_j \lvert A_{aj} - A_{bj} \rvert}{\sum_j max(A_{aj}, A_{bj})}$	[6]
spearman	$D_{ab} - 1 - \frac{6 \sum_j (rank(A_{aj}) \quad rank(A_{bj}))^2}{M(M-1)}$	[7]
species profile	$D_{ab} - \sqrt{\sum_j \left(\frac{A_{aj}}{A_{a-}} - \frac{A_{bj}}{A_{b-}} \right)^2}$	[2]
	qualitative methods	
χ^2 (qualitative)	see quantitative χ^2	see above
chord	see quantitative chord	see above
euclidean	see quantitative euclidean	see above
hamming	$D_{ab} - Num_a + Num_b - 2Num_{a \cap b}$, see also manhattan	[8]
jaccard	$D_{ab} = 1 \quad (\frac{Num_{a \cap b}}{Num_a - Num_b - Num_{a \cap b}})$, , see also soergel	[9]
lennon	$D_{ab} \quad 1 - \frac{Num_{a \cap b}}{Num_{a \cap b} - min(Num_a - Num_{a \cap b}, Num_b - Num_{a \cap b})}$	[10]
ochiai	$D_{ab} - 1 - \frac{Num_{a \cap b}}{\sqrt{Num_a Num_b}}$	[11]
pearson	see quantitative pearson	see above
dice	$D_{ab} - 1 - \frac{2Num_{a \cap b}}{Num_a - Num_b}$	[12]

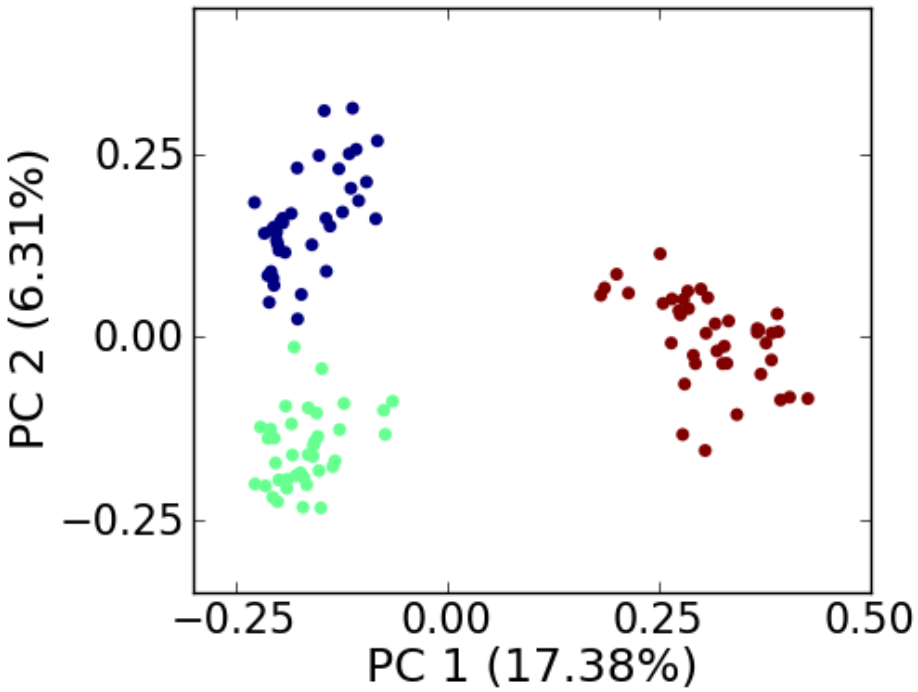
A variety of community dissimilarity measures exist

name	formula	ref
bray-curtis	$D_{ab} = \frac{\sum_j \lvert A_{aj} - A_{bj} \rvert}{\sum_j (A_{aj} \quad A_{bj})}$	[1]
canberra	$D_{ab} = \frac{1}{Num_{a \cap b}} \sum_j \frac{A_{aj} \quad A_{bj}\rvert}{A_{aj} + A_{bj}}$	[1]
χ^2 (chi-squared)	$D_{ab} = \sqrt{A_{++} \sum_j \left(\frac{1}{A_{-j}} \left(\frac{A_{aj}}{A_{a-}} - \frac{A_{bj}}{A_{b+}} \right)^2 \right)}$	[1]
chord	$D_{ab} = \sqrt{\sum_j \left(\frac{A_{aj}}{\sqrt{\sum_j A_{aj}^2}} \quad \frac{A_{bj}}{\sqrt{\sum_j A_{bj}^2}} \right)^2}$	[2]
euclidean	$D_{ab} = \sqrt{\sum_j (A_{aj} \quad A_{bj})^2}$	[1]
gower	$D_{ab} = \sum_j \frac{\lvert A_{aj} - A_{bj} \rvert}{(max_{ij} A_{ij} \quad min_{ij} A_{ij})}$	[1]
he□nger	$D_{ab} = \sqrt{\sum_j \left(\sqrt{\frac{A_{aj}}{A_{a\cdot}}} - \sqrt{\frac{A_{bj}}{A_{b\cdot}}} \right)^2}$	[3]
kulczynski (quantitative)	$D_{ab} \quad 1 - \frac{1}{2} \left(\frac{\sum_j min(A_{aj}, A_{bj})}{\sum_j A_{aj}} - \frac{\sum_j min(A_{aj}, A_{bj})}{\sum_j A_{bj}} \right)$	[1]
manhattan	$D_{ab} \quad \sum \lvert A_{aj} - A_{bj} \rvert$	[1]
morisita-horn	$\frac{2 \sum_j^J A_{aj} A_{bj}}{\left(\frac{\sum_j A_{aj}^2}{\sum_j A_{aj}} \quad \frac{\sum_j A_{bj}^2}{\sum_j A_{bj}} \right) \sum_j A_{aj} * \sum_j A_{bj}}$	[4]
pearson	$D_{ab} - 1 - \frac{\sum_j (A_{aj} \quad (\frac{A_a}{M})) \sum_j (A_{bj} \quad (\frac{A_b}{M}))}{\sqrt{\sum_i (A_{ai} - (\frac{A_a}{M}))^2} \sqrt{\sum_i (A_{bi} - (\frac{A_b}{M}))^2}}$	[5]
soergel	$D_{ab} = \frac{\sum_j \lvert A_{aj} - A_{bj} \rvert}{\sum_j max(A_{aj}, A_{bj})}$	[6]
spearman	$D_{ab} - 1 - \frac{6 \sum_j (rank(A_{aj}) \quad rank(A_{bj}))^2}{M(M-1)}$	[7]
species profile	$D_{ab} - \sqrt{\sum_j \left(\frac{A_{aj}}{A_{a-}} - \frac{A_{bj}}{A_{b-}} \right)^2}$	[2]
	qualitative methods	
χ^2 (qualitative)	see quantitative χ^2	see above
chord	see quantitative chord	see above
euclidean	see quantitative euclidean	see above
hamming	$D_{ab} - Num_a + Num_b - 2Num_{a \cap b}$, see also manhattan	[8]
jaccard	$D_{ab} = 1 \quad (\frac{Num_{a \cap b}}{Num_a - Num_b - Num_{a \cap b}})$, , see also soergel	[9]
lennon	$D_{ab} \quad 1 - \frac{Num_{a \cap b}}{Num_{a \cap b} - min(Num_a - Num_{a \cap b}, Num_b - Num_{a \cap b})}$	[10]
ochiai	$D_{ab} - 1 - \frac{Num_{a \cap b}}{\sqrt{Num_a Num_b}}$	[11]
pearson	see quantitative pearson	see above
dice	$D_{ab} - 1 - \frac{2Num_{a \cap b}}{Num_a - Num_b}$	[12]

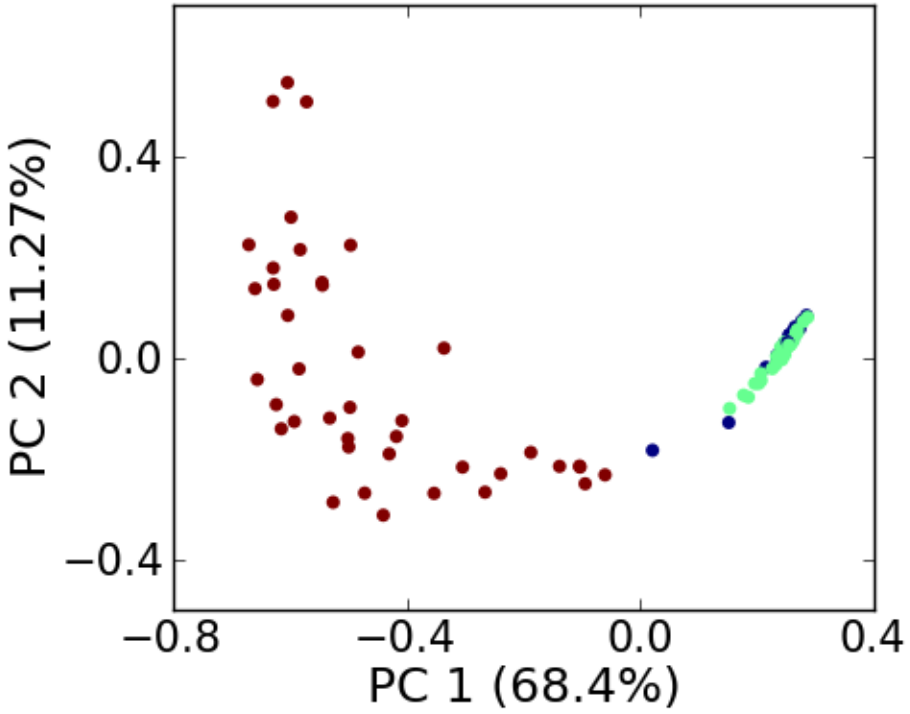
Is metric choice important?

Absolutely

metric 1 (Jaccard)



metric 2 (Morisita-Horn)



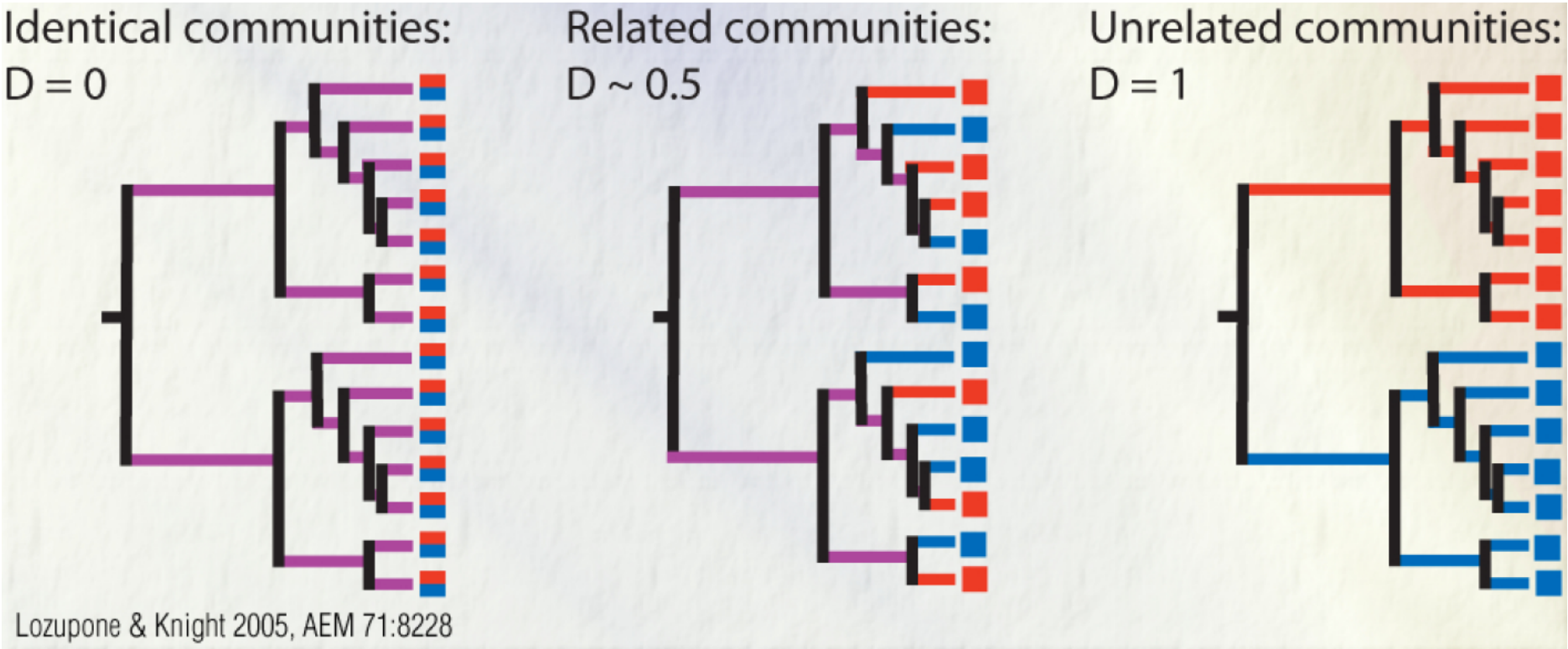
Fierer et al. *PNAS* 2010

NATURE METHODS | ANALYSIS

Microbial community resemblance methods differ in their ability to detect biologically relevant patterns

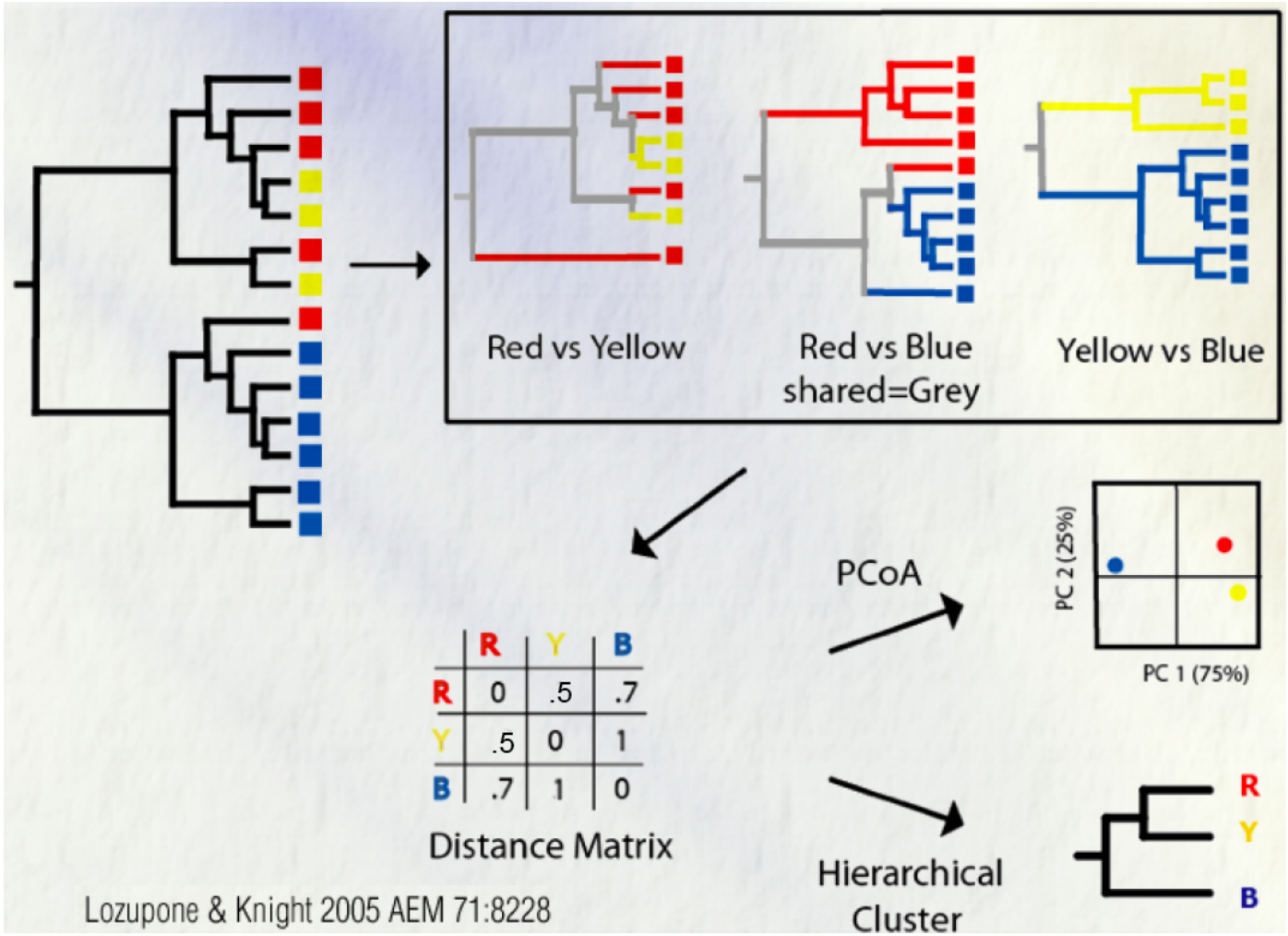
Justin Kuczynski, Zongzhi Liu, Catherine Lozupone, Daniel McDonald, Noah Fierer & Rob Knight

UniFrac metric can be used in PCoA



(cited over 1500 times)

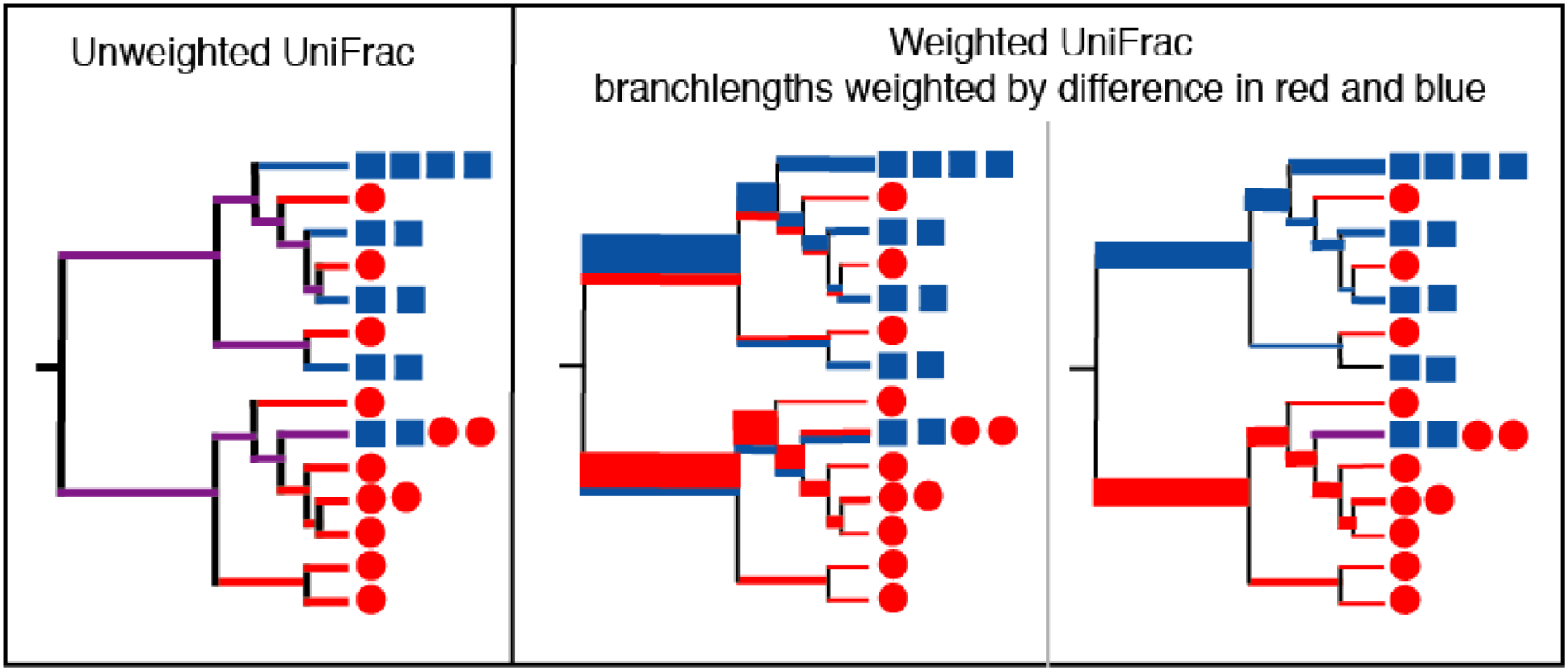
UniFrac distance matrix



UniFrac

"Unweighted"
Qualitative

"Weighted"
Quantitative



Lozupone et al., 2007. Appl Environ Microbiol 73:1576

Core diversity analysis

core_diversity_analyses.py

-i <biom format otu table>

-m <mapping file>

-t <tree file>

-o <output directory>

-e <rarefaction level>

-c <mapping category/categories>

-aO <number of cores if parallel>

What's actually happening:

- alpha_rarefaction.py
  - multiple_rarefactions.py
  - alpha_diversity.py
  - collate_alpha.py
  - make_rarefaction_plots.py
- beta_diversity_through_plots.py
  - beta_diversity.py
  - principal_coordinates.py
  - make_emperor.py
- summarize_taxa_through_plots.py
  - summarize_otu_by_cat.py
  - summarize_taxa.py
  - plot_taxa_summary.py
- make_distance_boxplots.py
- compare_alpha_diversity.py
- group_significance.py

Running Illumina Tutorial