Computing diversity metrics based on even sampling using QIIME

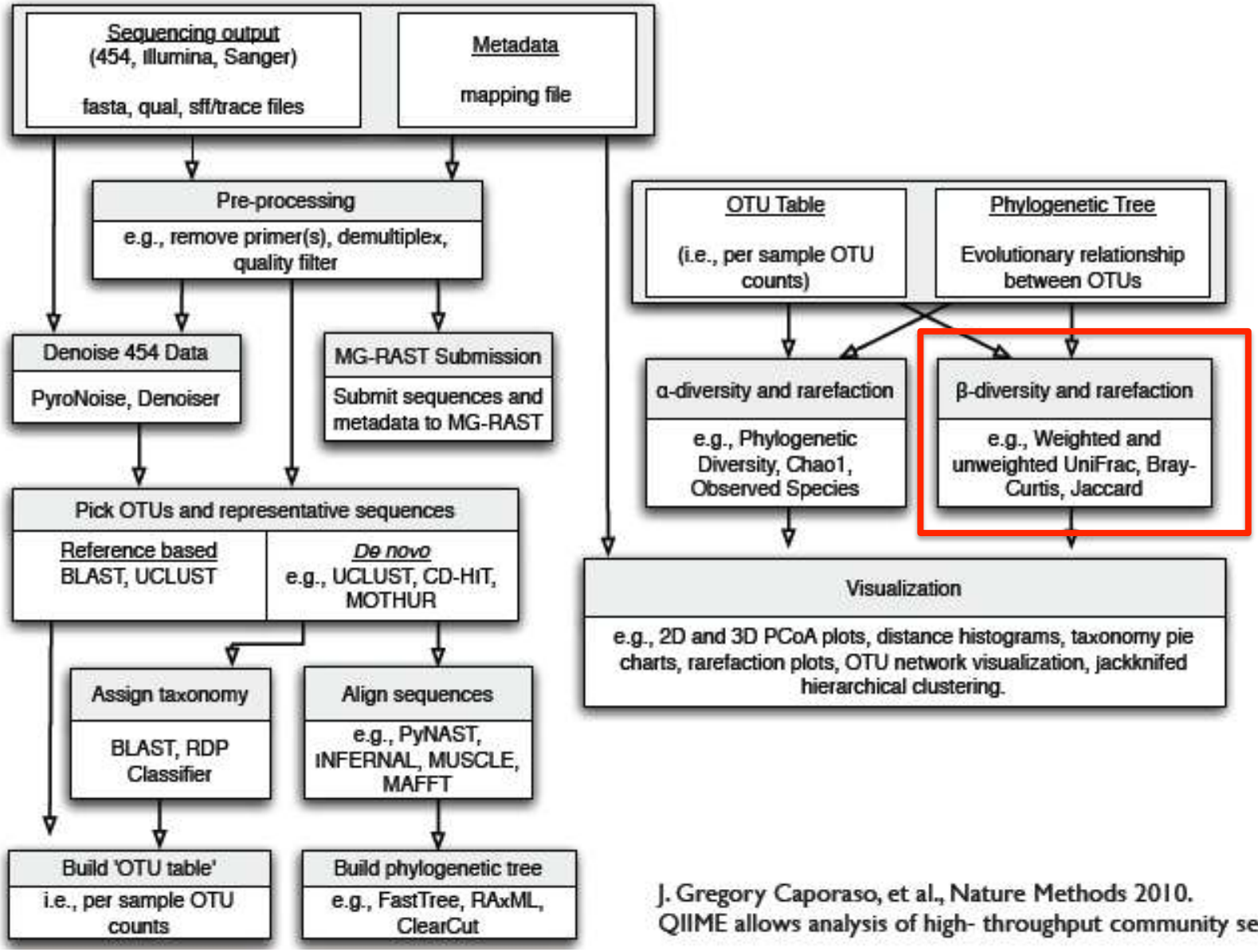
Bharath Prithiviraj,
Knight lab, Univ. of Colorado, Boulder.

Workshop on Comparative Genomics

Colorado State University, Fort Collins

July 20th 2011.

Quantitative Insights Into Microbial Ecology



J. Gregory Caporaso, et al., Nature Methods 2010.
QIIME allows analysis of high- throughput community sequencing data

Overview:

1. Use of Qualitative and Quantitative beta diversity metrics in community ecology.

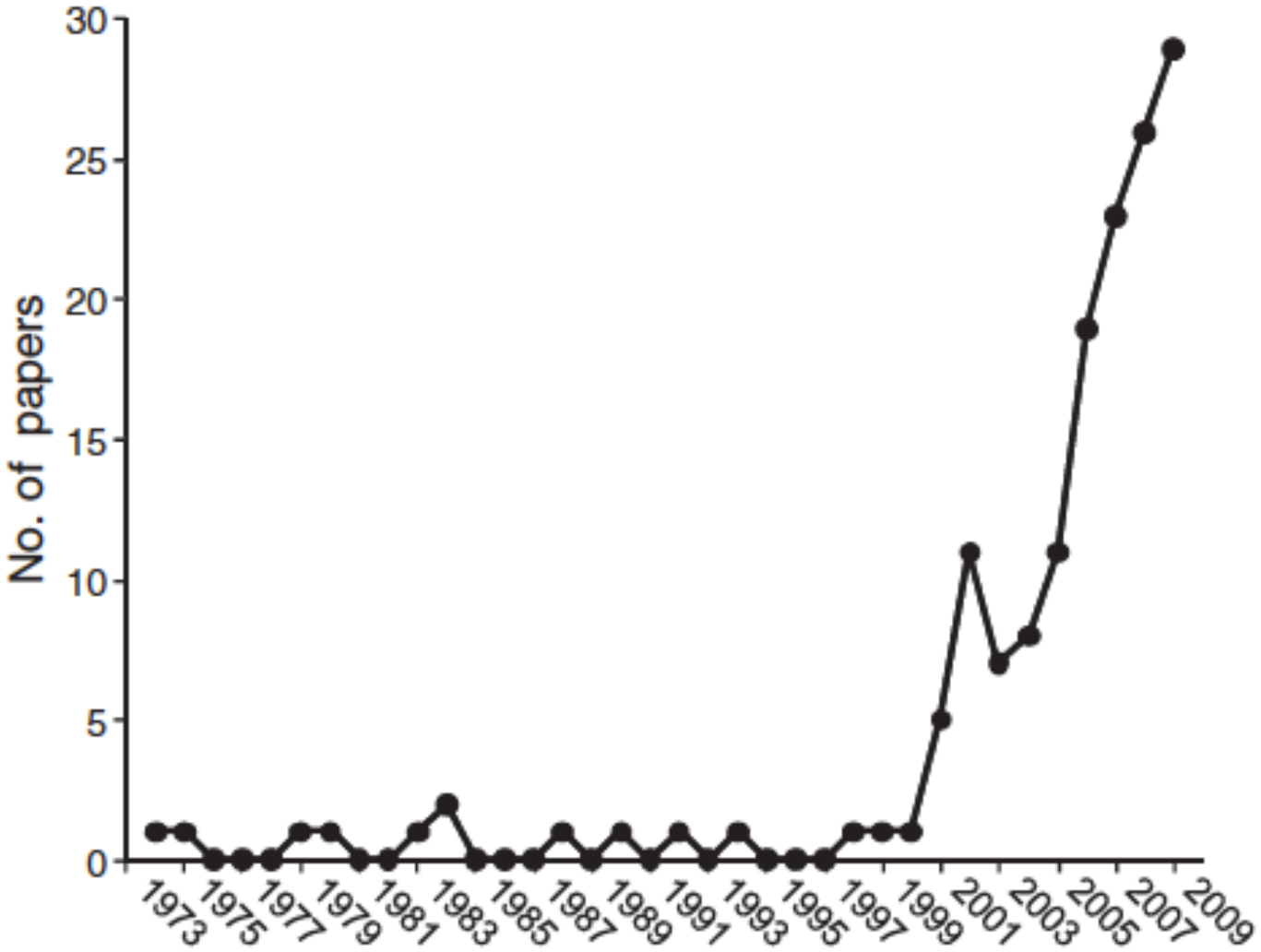
2. Phylogenetic diversity (PD) as a metric.

3. Example 1 : Germ Free mice transplant study using UniFrac metric and its comparison with Euclidean metric. (Caporaso *et al*.2010)

4. Example 2 : Keyboard study – Subset study on data for Variation with sampling depth.

5. Example 3 : Mammalian gut and human skin comparison. Kuczynski *et al*. 2010.

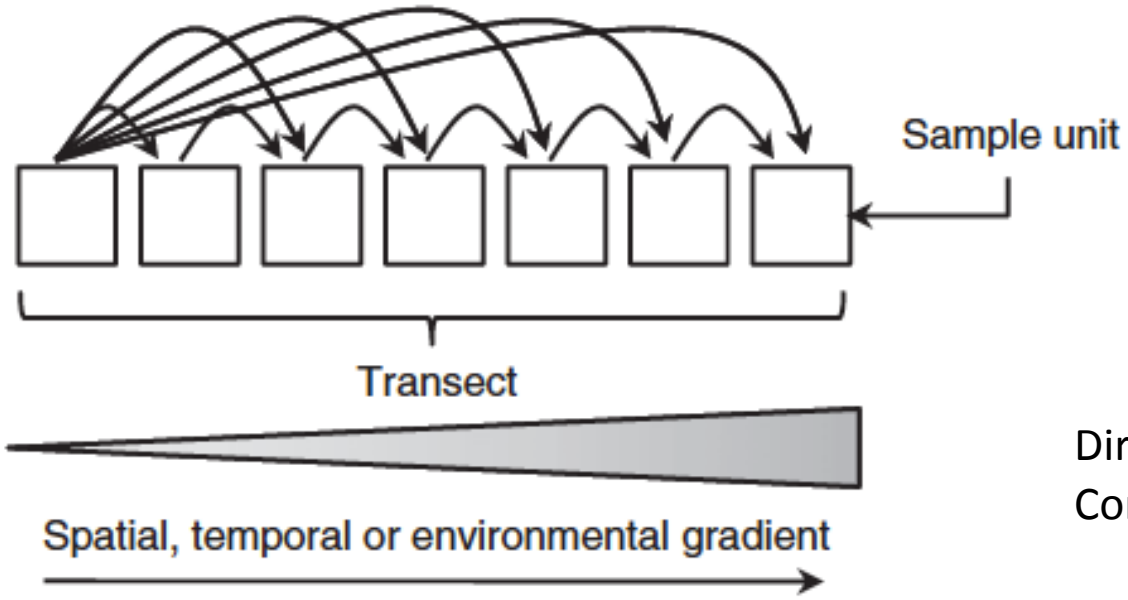
6. Even sampling using the tutorial dataset – QIIME command line task.



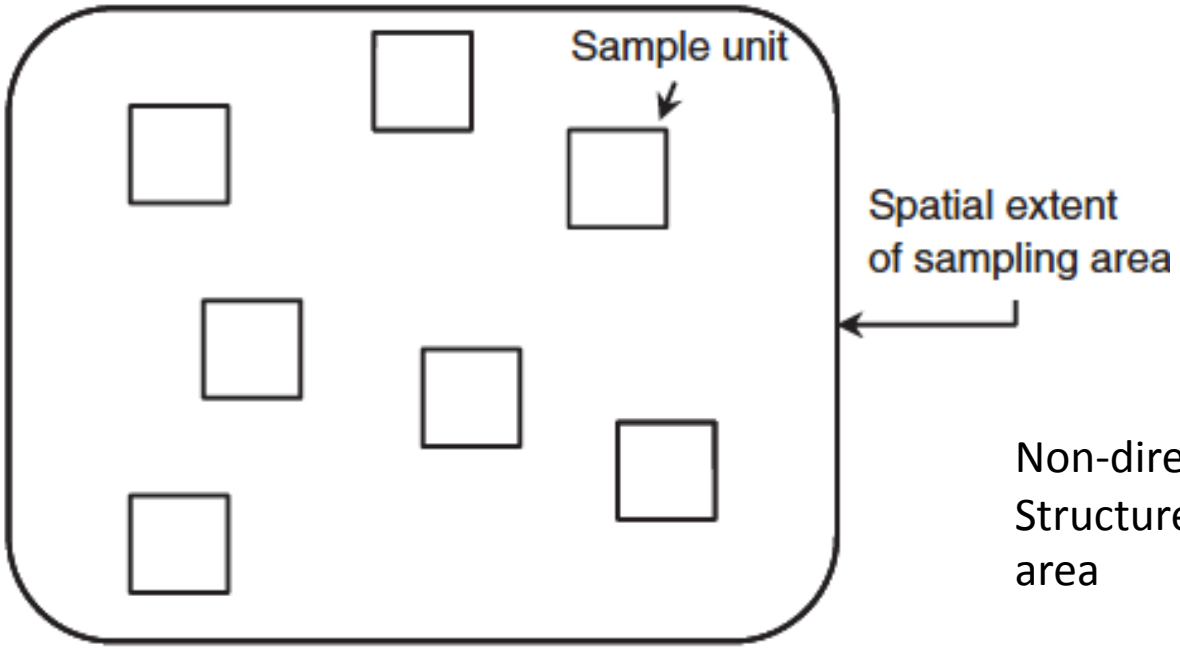
Plot showing the no. of papers from ISI web of science that cite Beta diversity in their title including Greek alphabet "β" diversity btw. 1973 and 2009.

Lead up from Whittaker's Taxon paper on the evolution and measurement of species diversity in 1972.

Anderson et al.2010

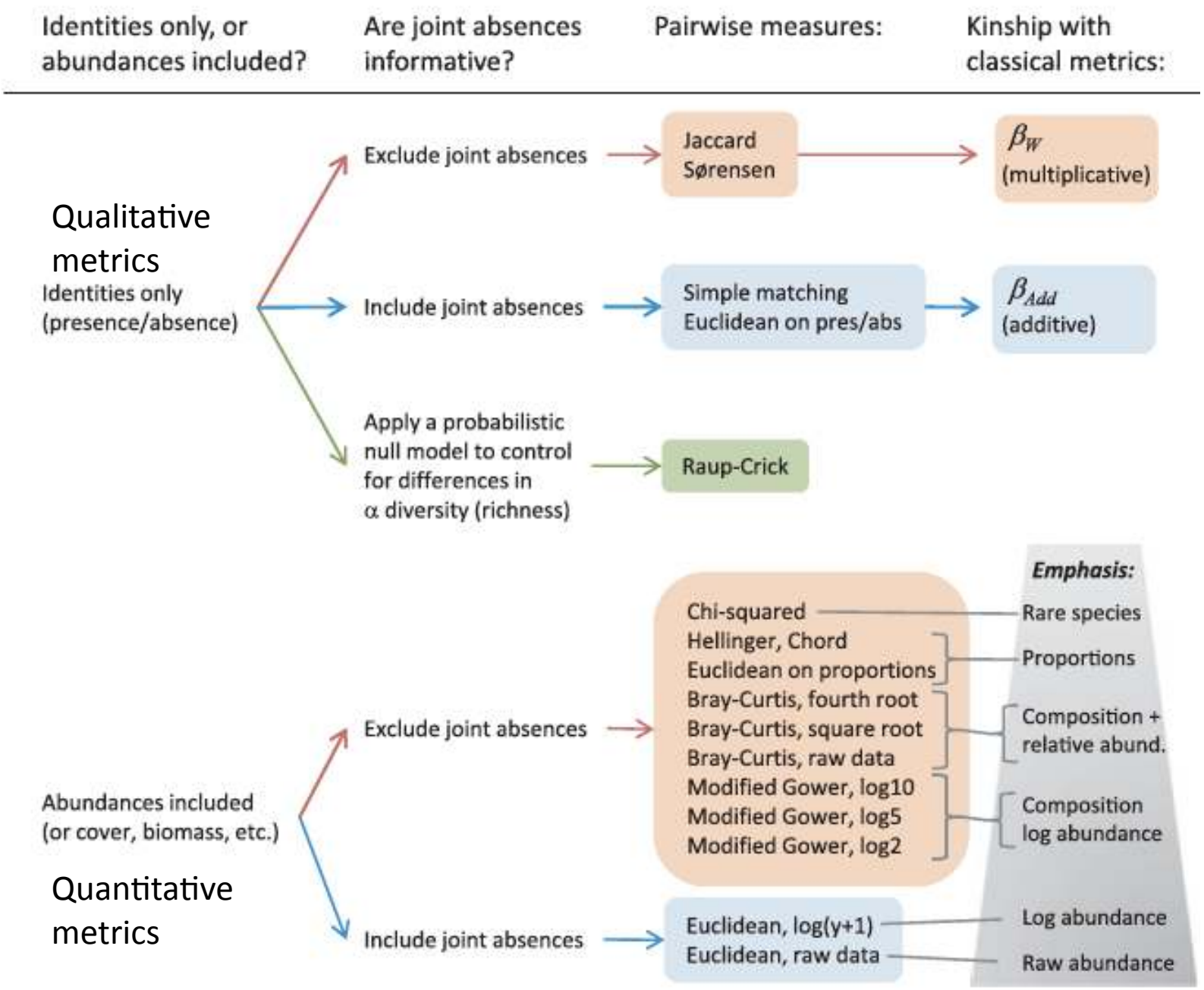


Directional Species turnover in
Community structure along a gradient



Non-directional Variation in community
Structure among samples within a given
area

Anderson et al.2010



Anderson et al.2010

Comparing microbial communities

What is there? How much is there?

α (i.e., within sample) diversity

How similar are samples?

β (i.e., between sample) diversity

Phylogenetic diversity measures account for evolutionary relatedness of observed OTUs

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Non-phylogenetic method: '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*

Species counts:

Sample A	3
Sample B	3
Sample C	3

Non-phylogenetic method: '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 methods: Phylogenetic Diversity (PD)

Sample A

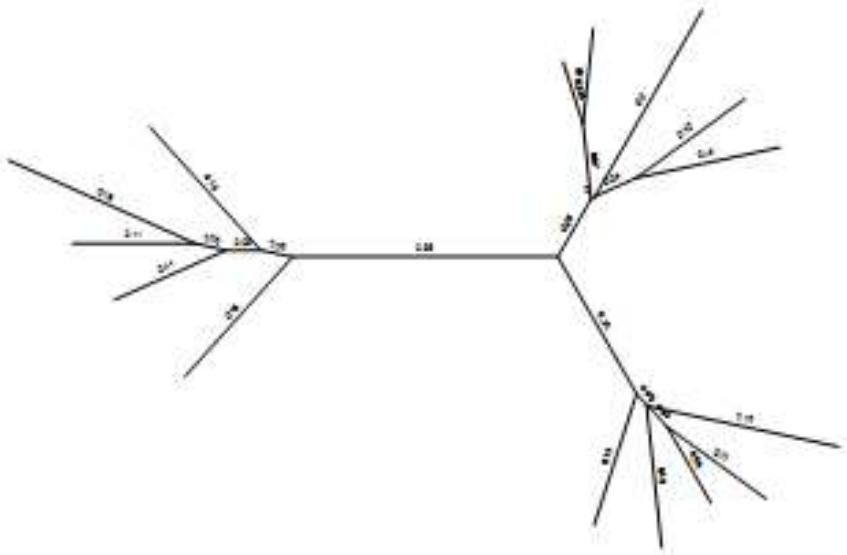
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

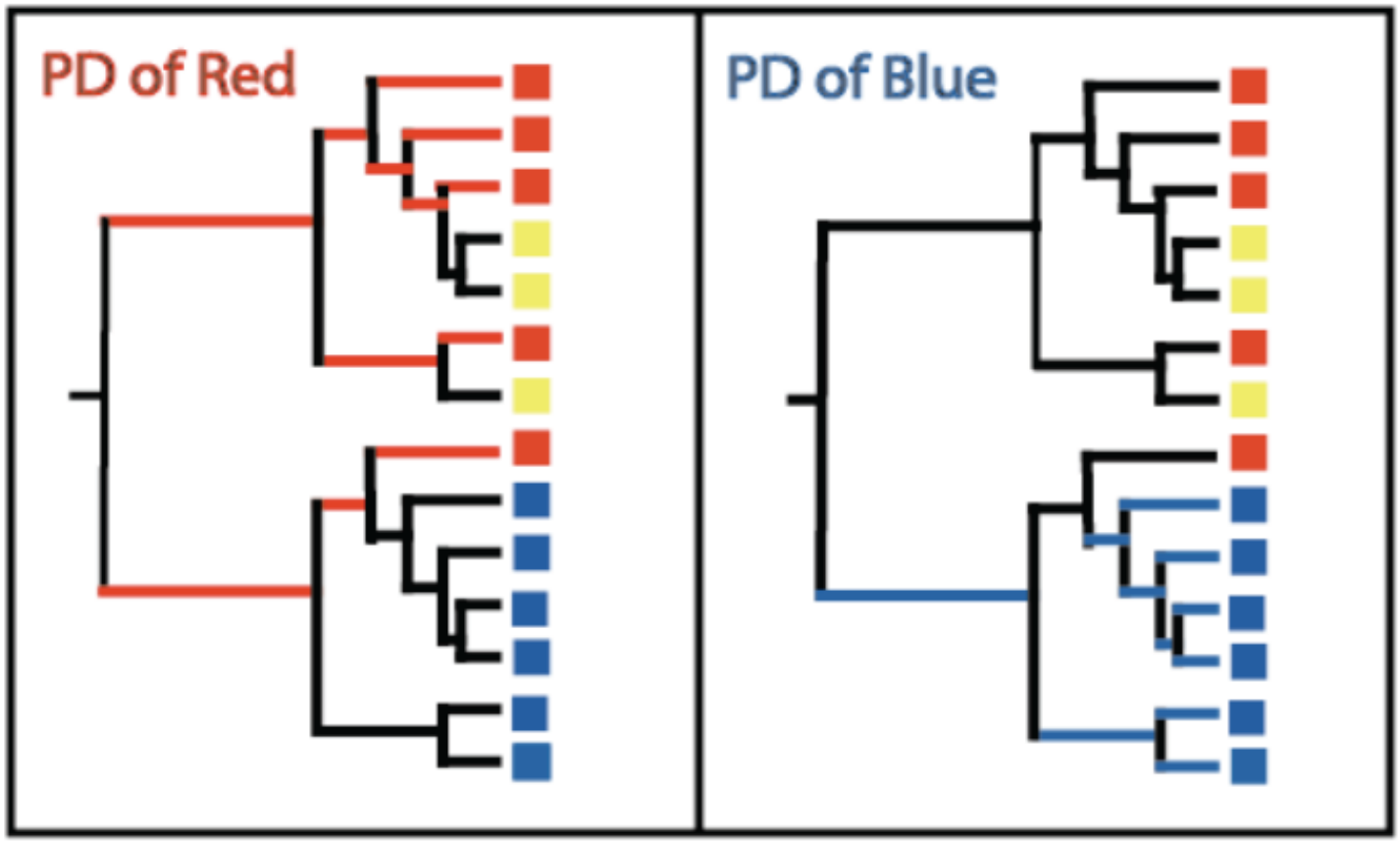
Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



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

Phylogenetic Diversity (PD): a qualitative, phylogenetic α-diversity metric

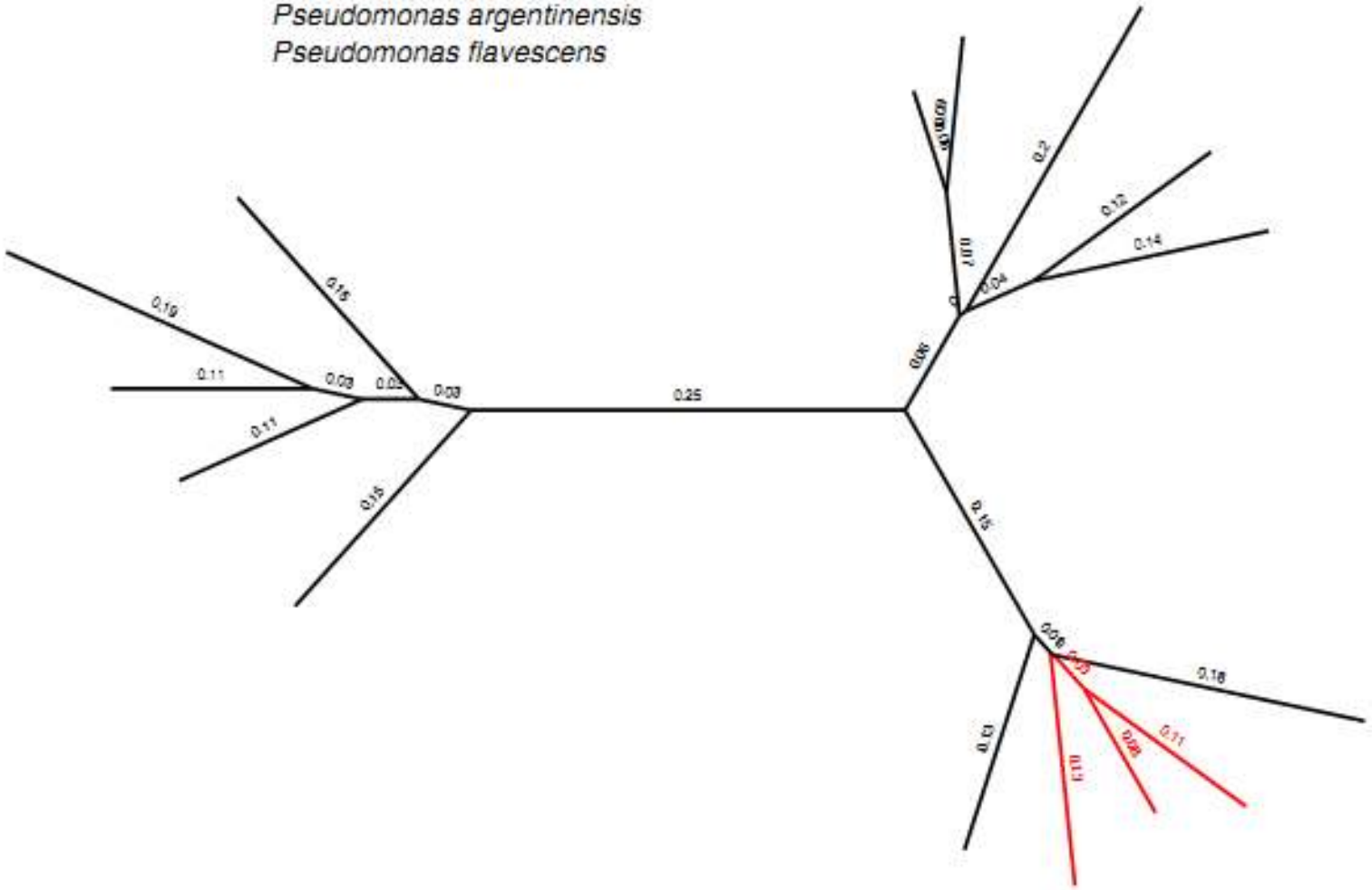


Sum of branch length covered by a sample

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

Sample A

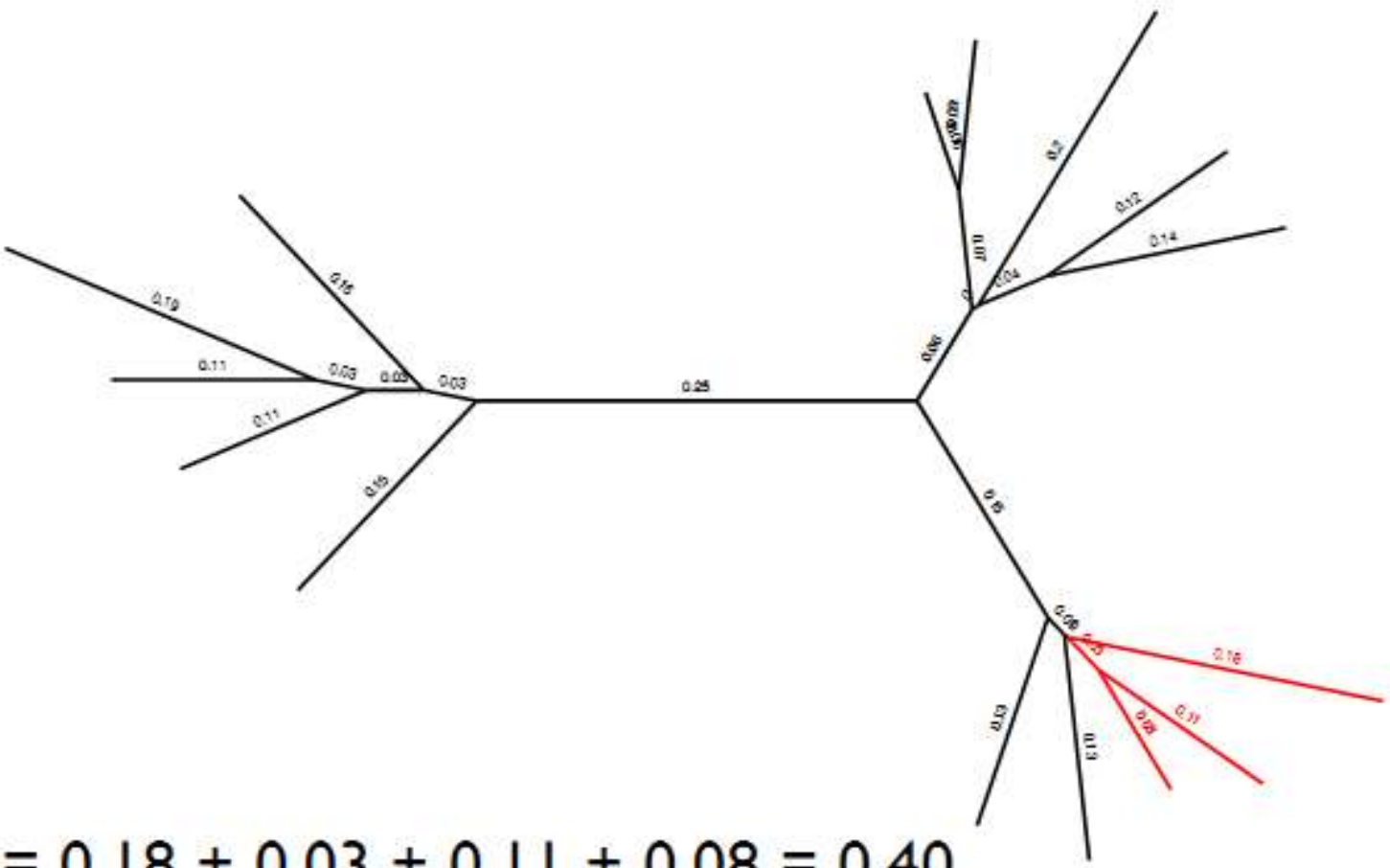
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens



PD = 0.13 + 0.03 + 0.11 + 0.08 = 0.35

Sample B

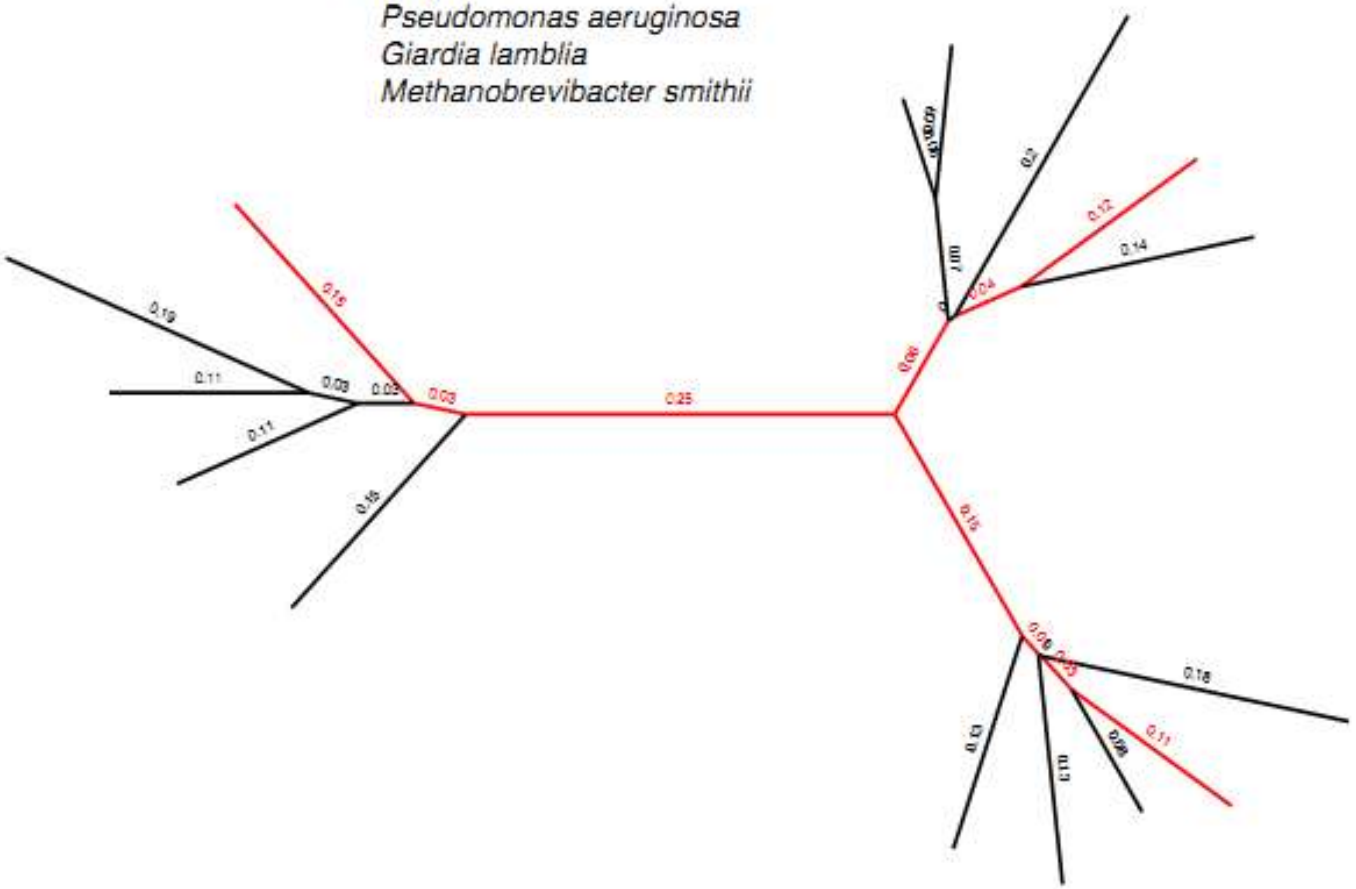
Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli



PD = 0.18 + 0.03 + 0.11 + 0.08 = 0.40

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



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

Phylogenetic methods: 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 methods: 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.

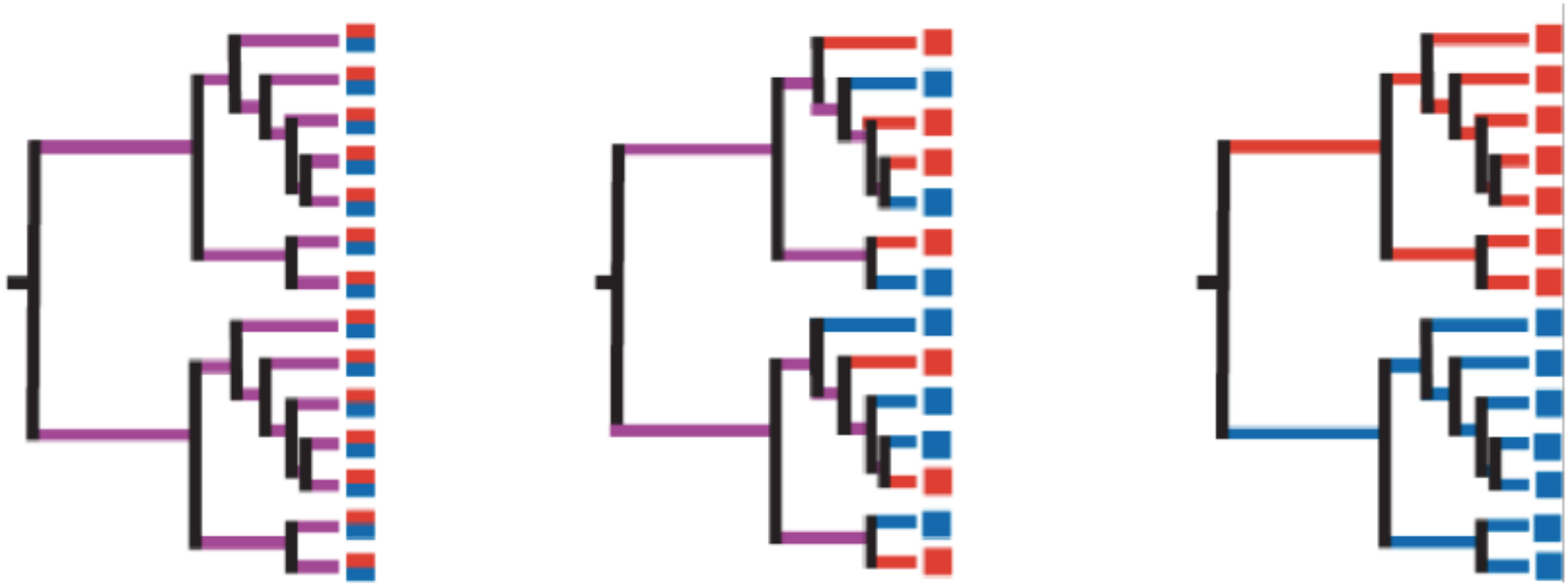
Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

UniFrac: a qualitative, phylogenetic β-diversity metric

Identical communities
D = 0.0

Related communities
D ~ 0.5

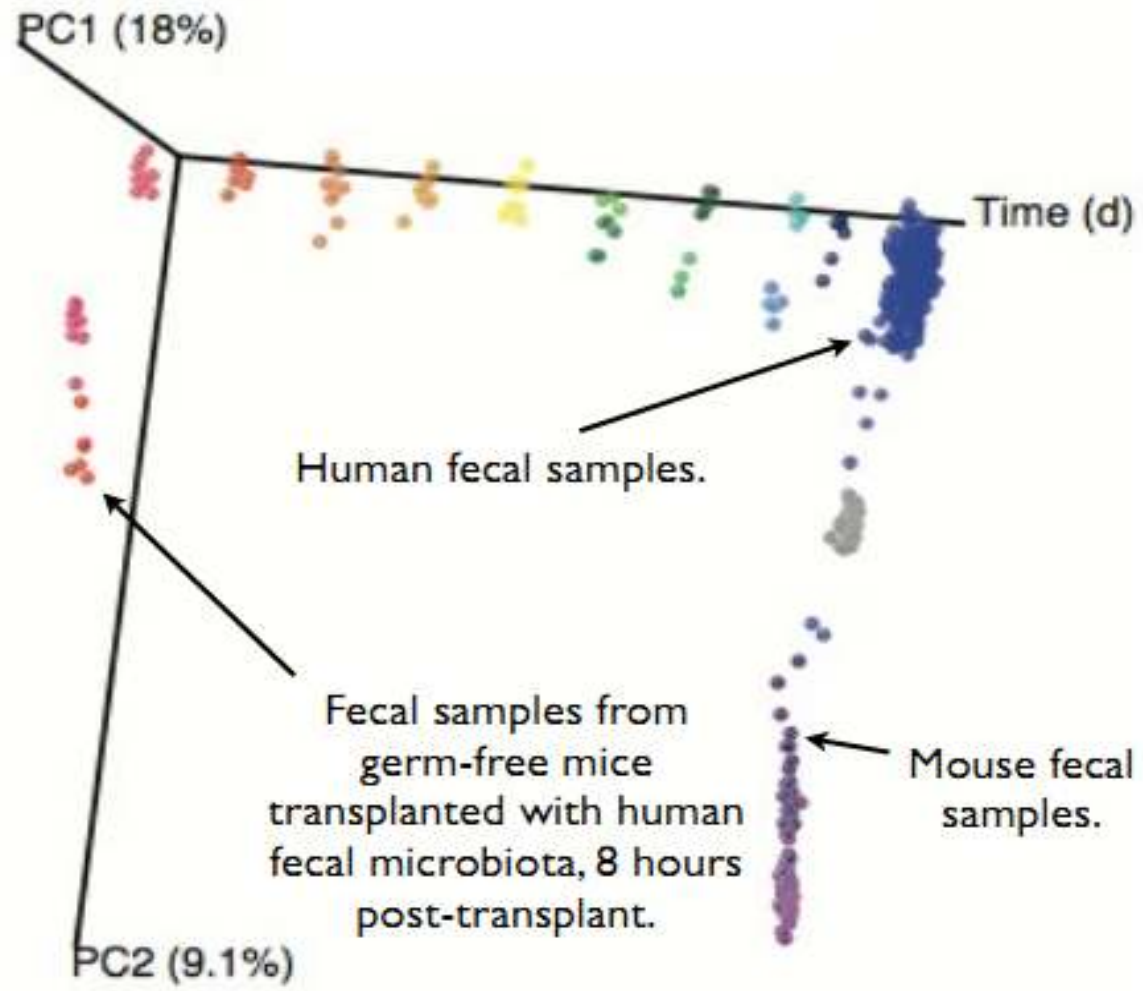
Unrelated communities
D = 1.0



Percent of observed branch length that is unique to either sample

Lozupone and Knight, 2005, Appl Environ Microbiol 71:8228

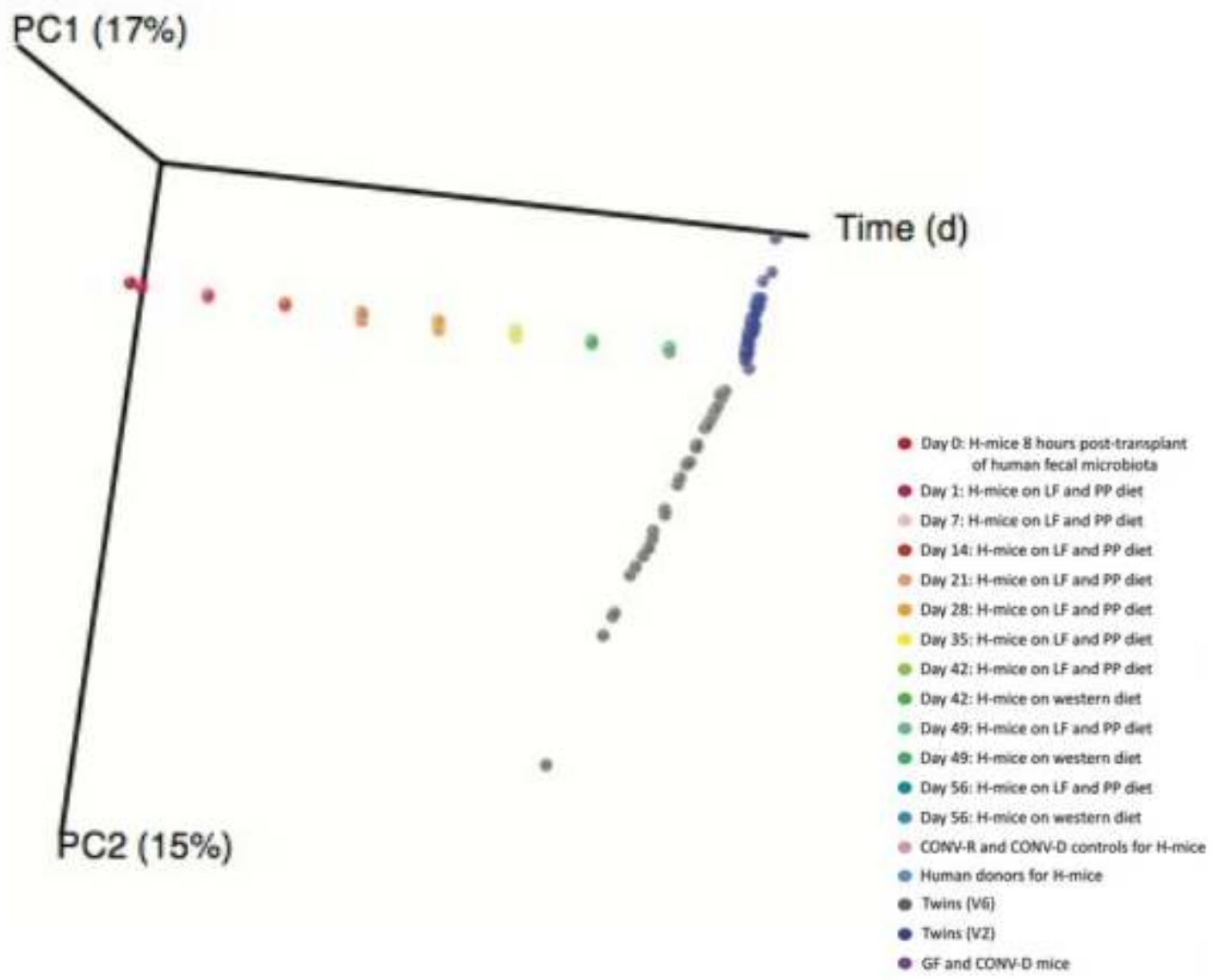
UniFrac distances between samples



- Day 0: H-mice 8 hours post-transplant of human fecal microbiota
- Day 1: H-mice on LF and PP diet
- Day 7: H-mice on LF and PP diet
- Day 14: H-mice on LF and PP diet
- Day 21: H-mice on LF and PP diet
- Day 28: H-mice on LF and PP diet
- Day 35: H-mice on LF and PP diet
- Day 42: H-mice on LF and PP diet
- Day 42: H-mice on western diet
- Day 49: H-mice on LF and PP diet
- Day 49: H-mice on western diet
- Day 56: H-mice on LF and PP diet
- Day 56: H-mice on western diet
- CONV-R and CONV-D controls for H-mice
- Human donors for H-mice
- Twins (V6)
- Twins (V2)
- GF and CONV-D mice

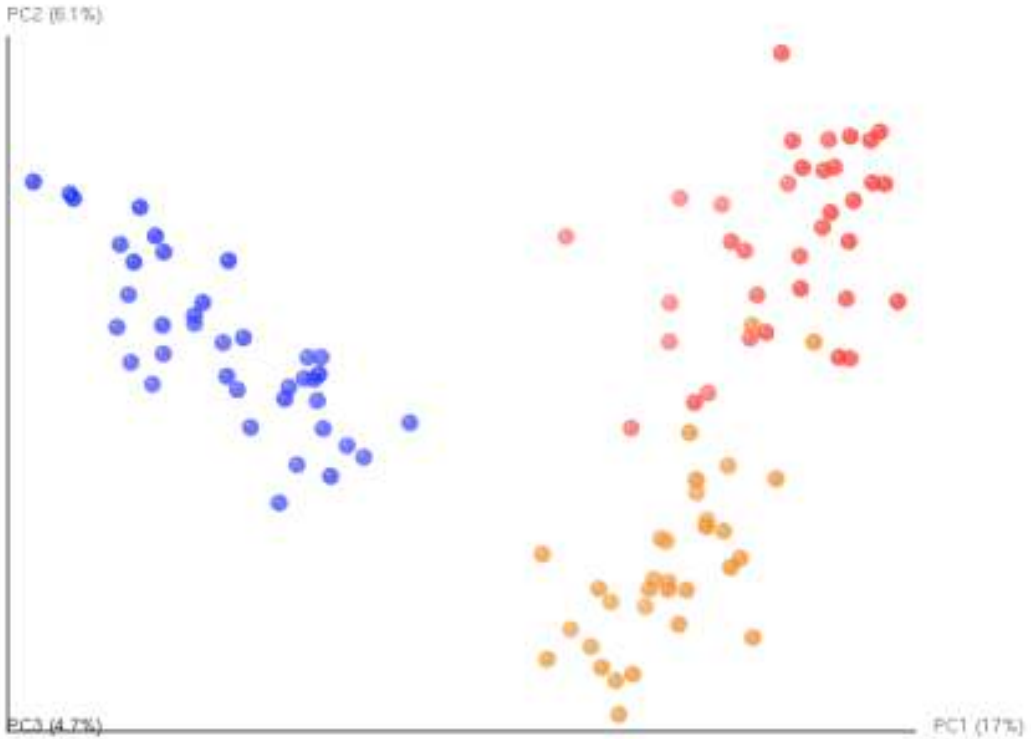
Caporaso et al., 2010

Euclidean distances between samples



Caporaso et al., 2010

Variation in sampling depth is an important consideration

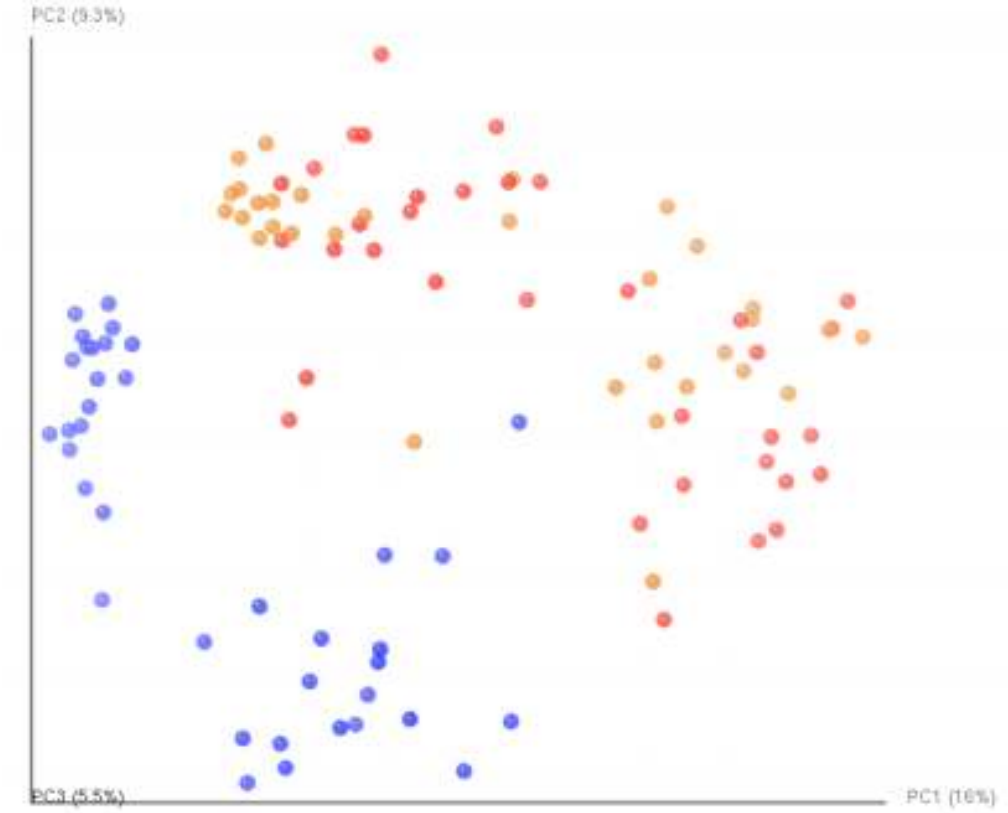


Human skin, colored by individual, at 500 sequence/sample

Image/analysis credit: Justin Kuczynski

Data reference:
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

Variation in sampling depth is an important consideration

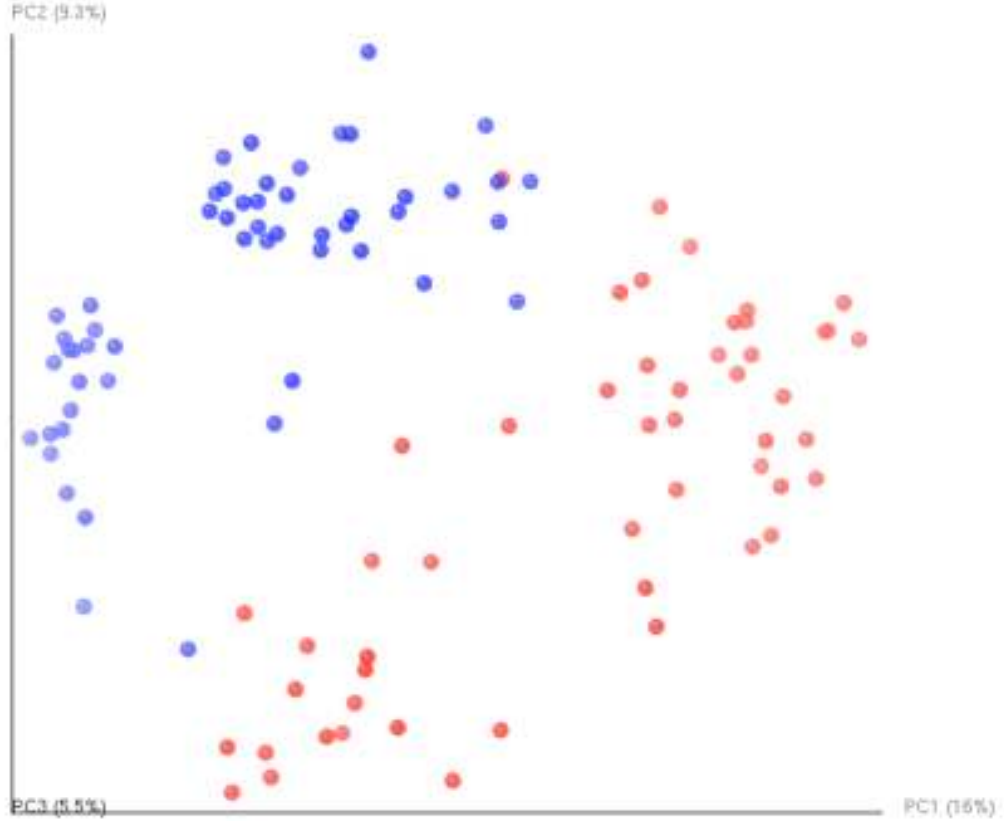


Human skin, colored by individual, at either 50 or 500 sequence/sample

Image/analysis credit: Justin Kuczynski

Data reference:
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

Variation in sampling depth is an important consideration



Human skin, colored by sampling depth, at either 50 or 500 sequence/sample

Image/analysis credit: Justin Kuczynski

Data reference:
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

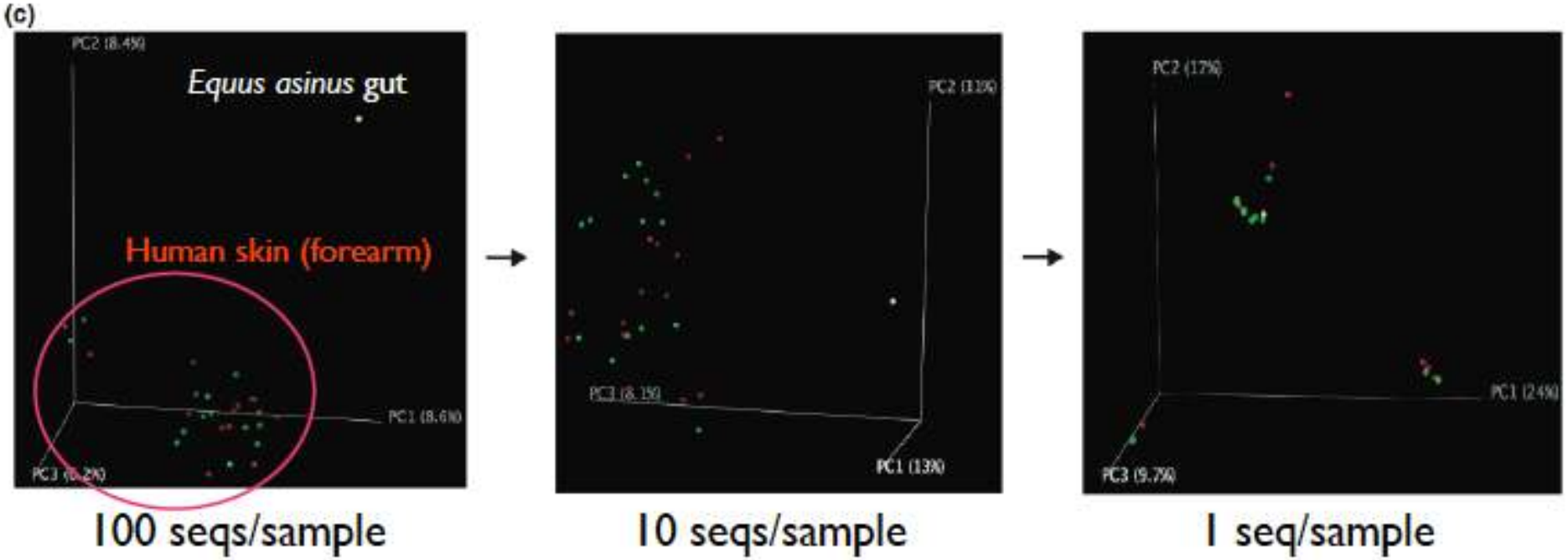
Kuczynski et al. Genome Biology 2010, 11:210
http://genomebiology.com/2010/11/5/210

Genome Biology

REVIEW

Direct sequencing of the human microbiome readily reveals community differences

Justin Kuczynski[1], Elizabeth K Costello[2], Diana R Nemergut[3], Jesse Zaneveld[1], Christian L Lauber[4], Dan Knights[5], Omry Koren[6], Noah Fierer[4], Scott T Kelley[7], Ruth E Ley[6], Jeffrey I Gordon[8] and Rob Knight[9,10]*



per_library_stats.py
QIIME script for summarizing the number of sequences in each sample, as well as some statistics about the distribution of sequences/sample

single_rarefaction.py
QIIME script for evenly sampling an OTU table to a specified number of sequences per sample