

HU CFAR QIIME Workshop

15 September 2014

www.qiime.org

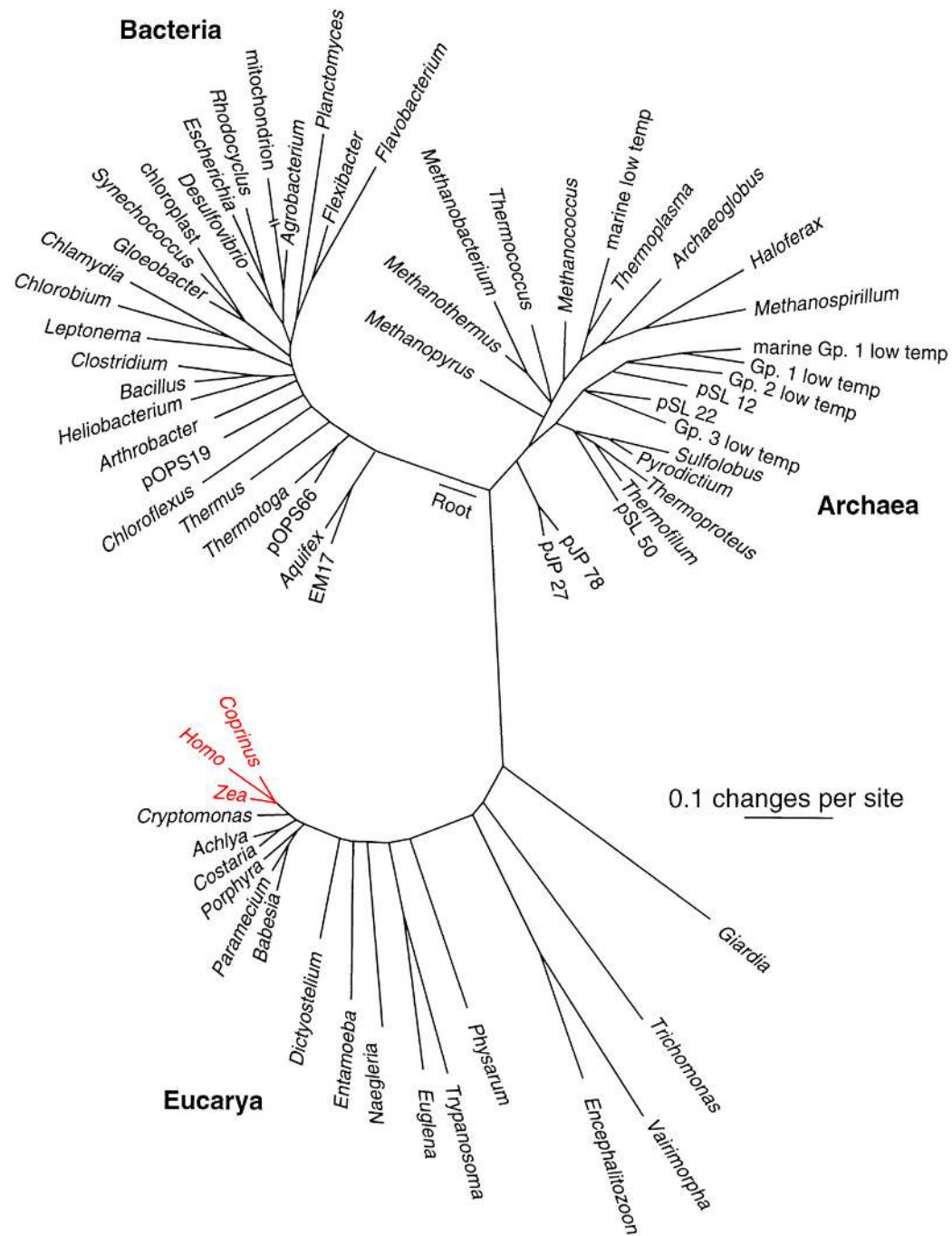
John Chase

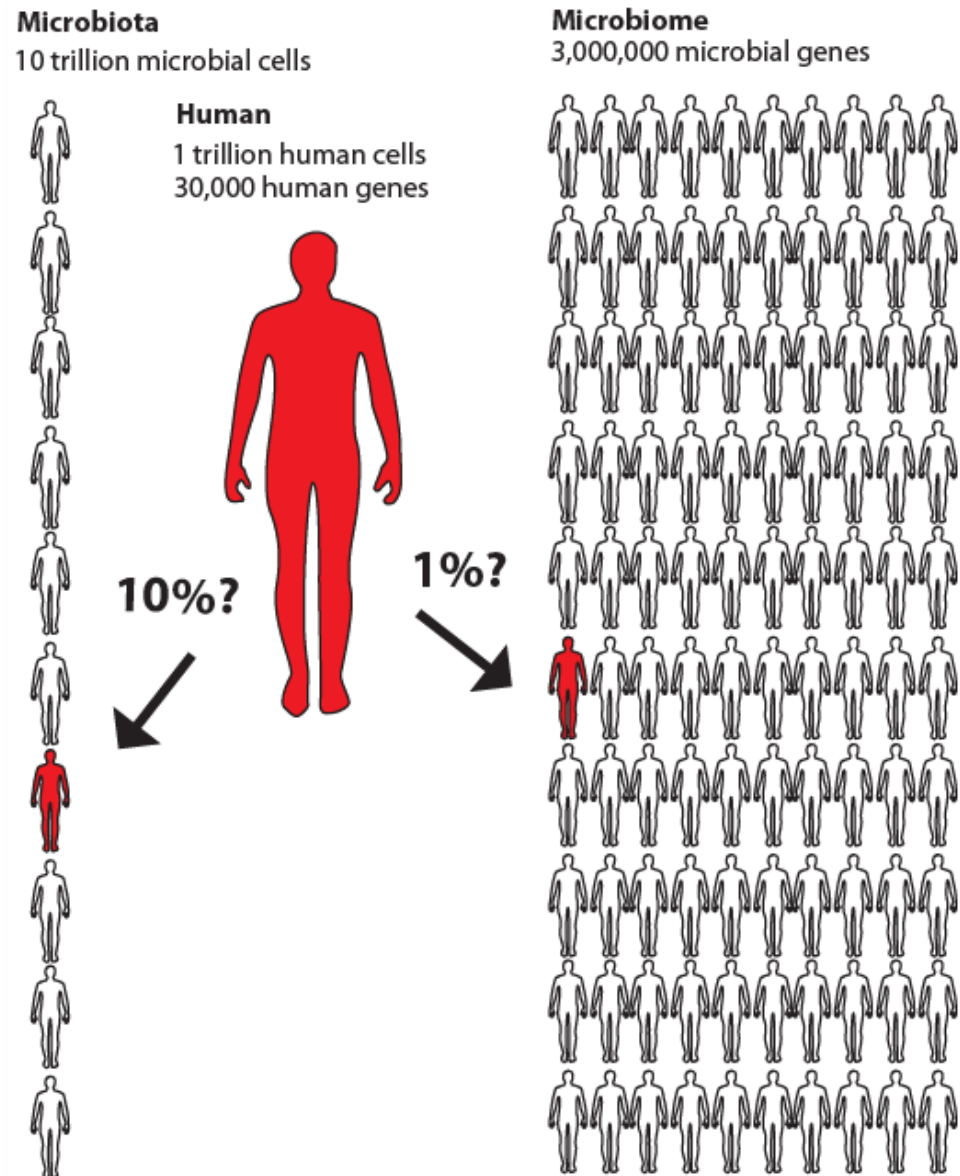
Yoshiki Vazquez-Baeza

www.caporasolab.us

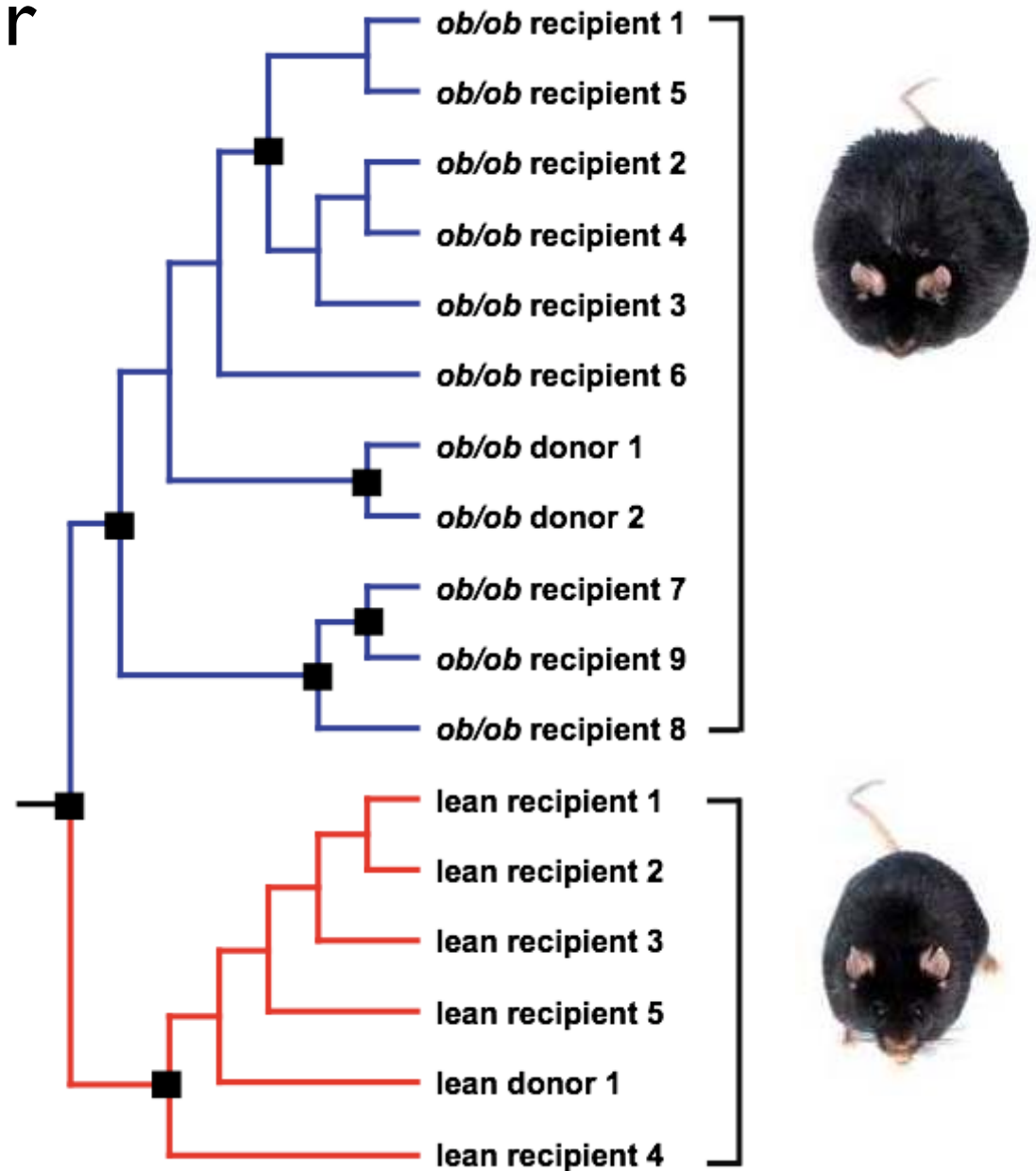
<https://knightlab.ucsd.edu>

www.applied-bioinformatics.org





Do differences in our microbiota matter?



Peter J. Turnbaugh et al., Nature 2006

An obesity-associated gut microbiome with increased capacity for energy harvest

Microbes rarely live or act alone.

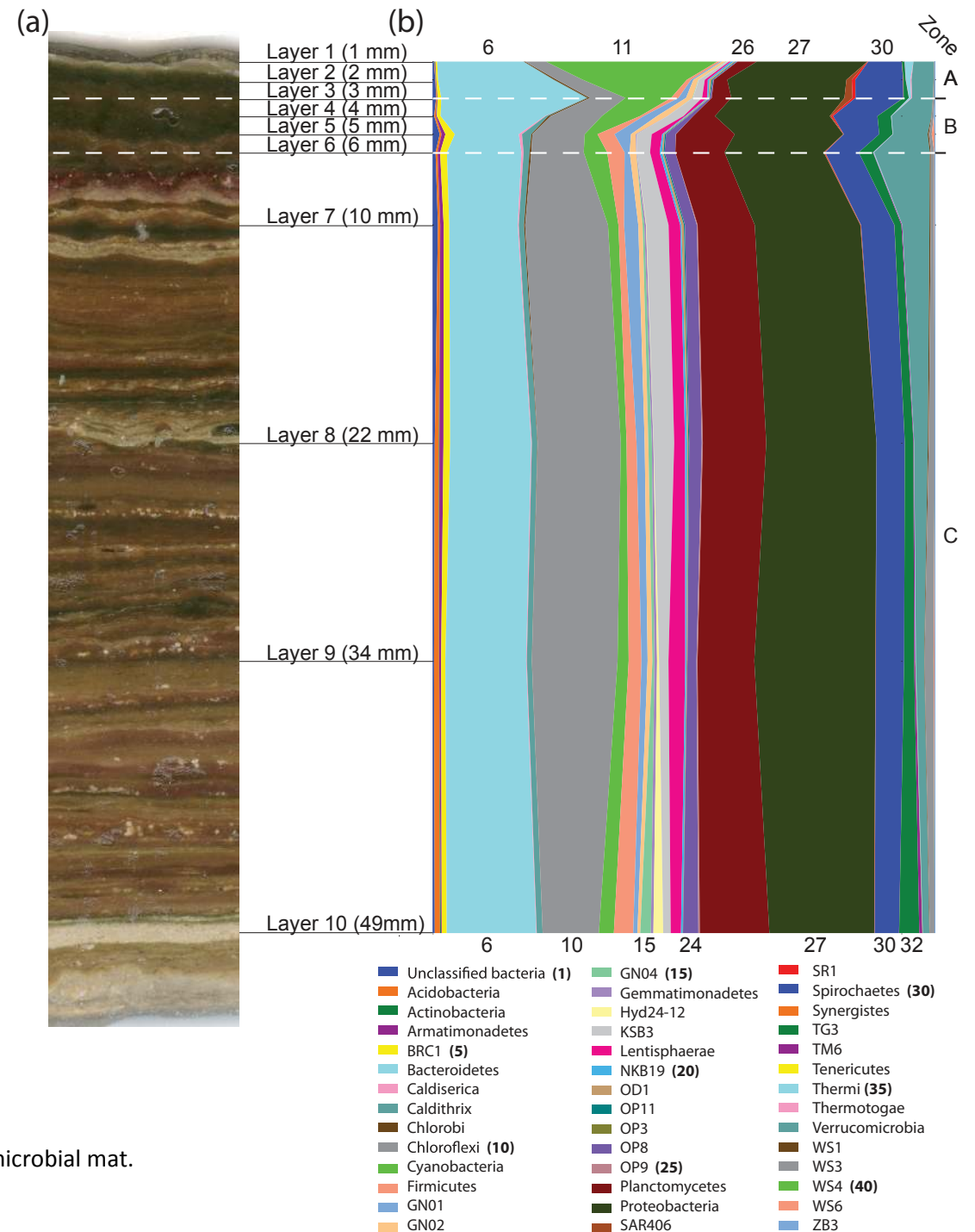


Image source:

Phylogenetic stratigraphy in the Guerrero Negro hypersaline microbial mat.

Harris, Caporaso *et al.* (2012)

International Society for Microbial Ecology Journal

Culturing microbes is hard

Back of the envelope calculation: less than 13% of bacterial species* have a representative that has been grown in culture.

Many recent advances are based on *culture-independent* approaches for studying microbial communities.



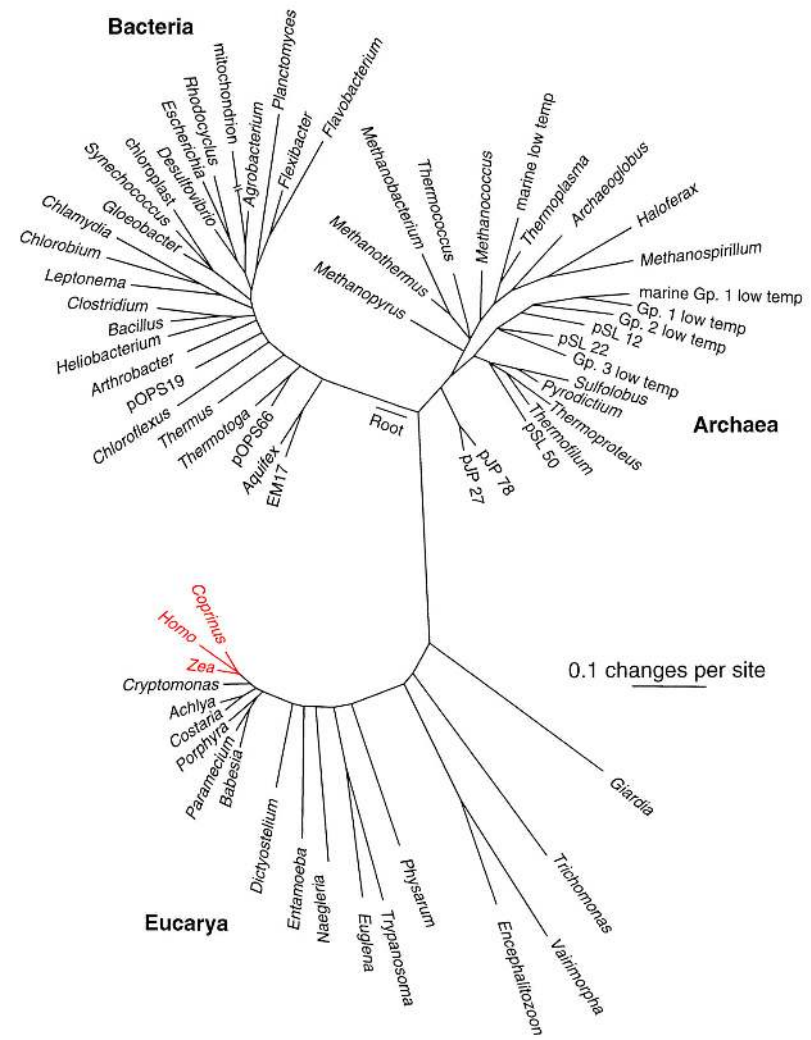
Bacillus anthracis in culture

* Defined as 97% OTUs in the Greengenes 13_5 reference database.

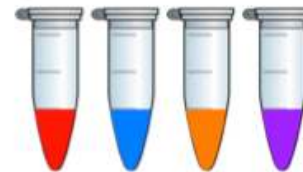
Culture-independent investigation of microbial communities

All cellular life has a shared evolutionary history, and some genes are shared by all organisms.

The sequence of those genes can be used as a *genetic fingerprint* for different organisms.



Collect samples



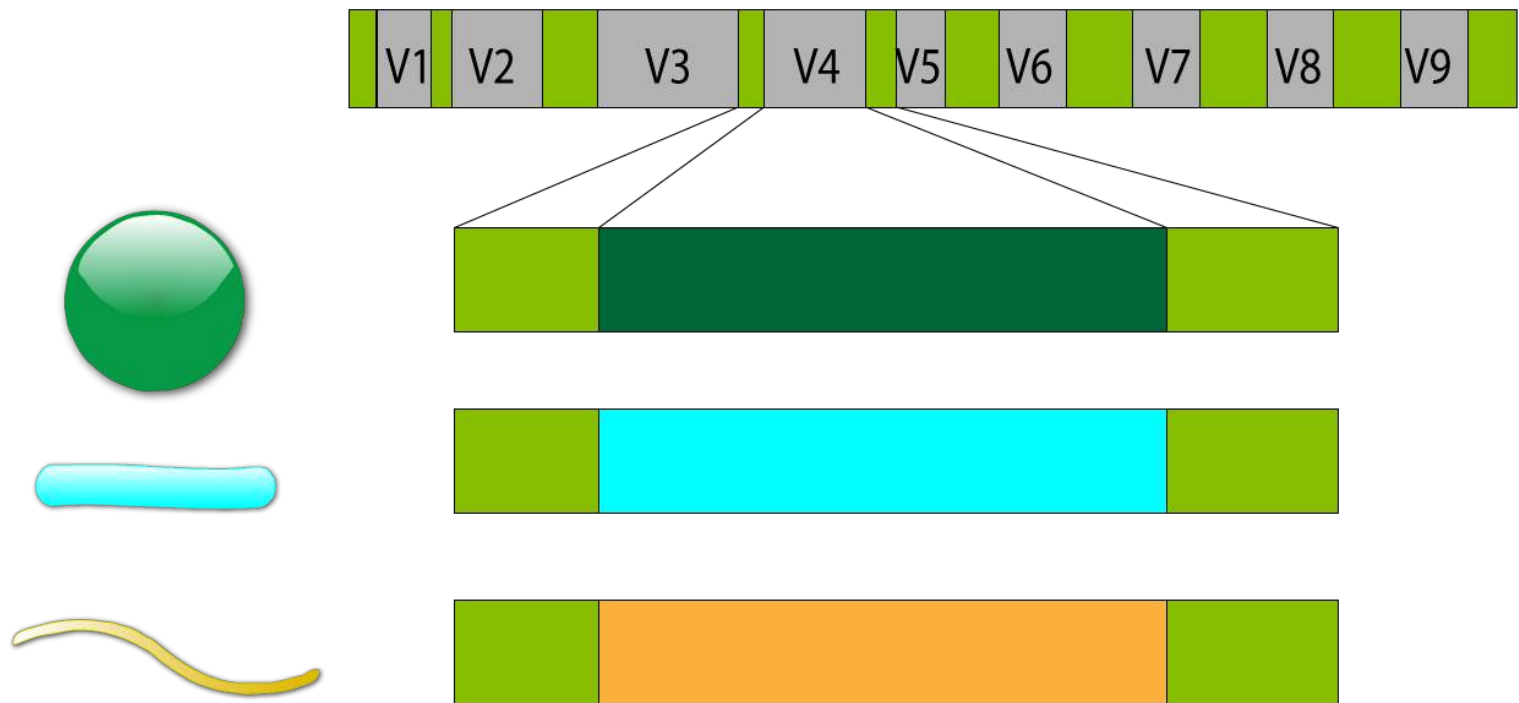
Extract DNA

(you can do this at home!)



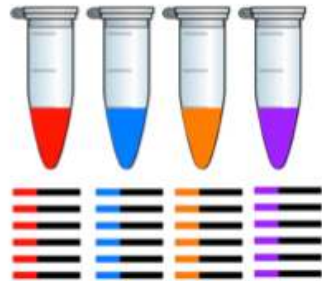
Isolate the *small subunit ribosomal RNA* gene to “fingerprint” different microbial organisms. (PCR)

Why 16S rRNA?



Sequence the rRNA from all samples on a “high-throughput” DNA sequencer

Per-sample rRNA

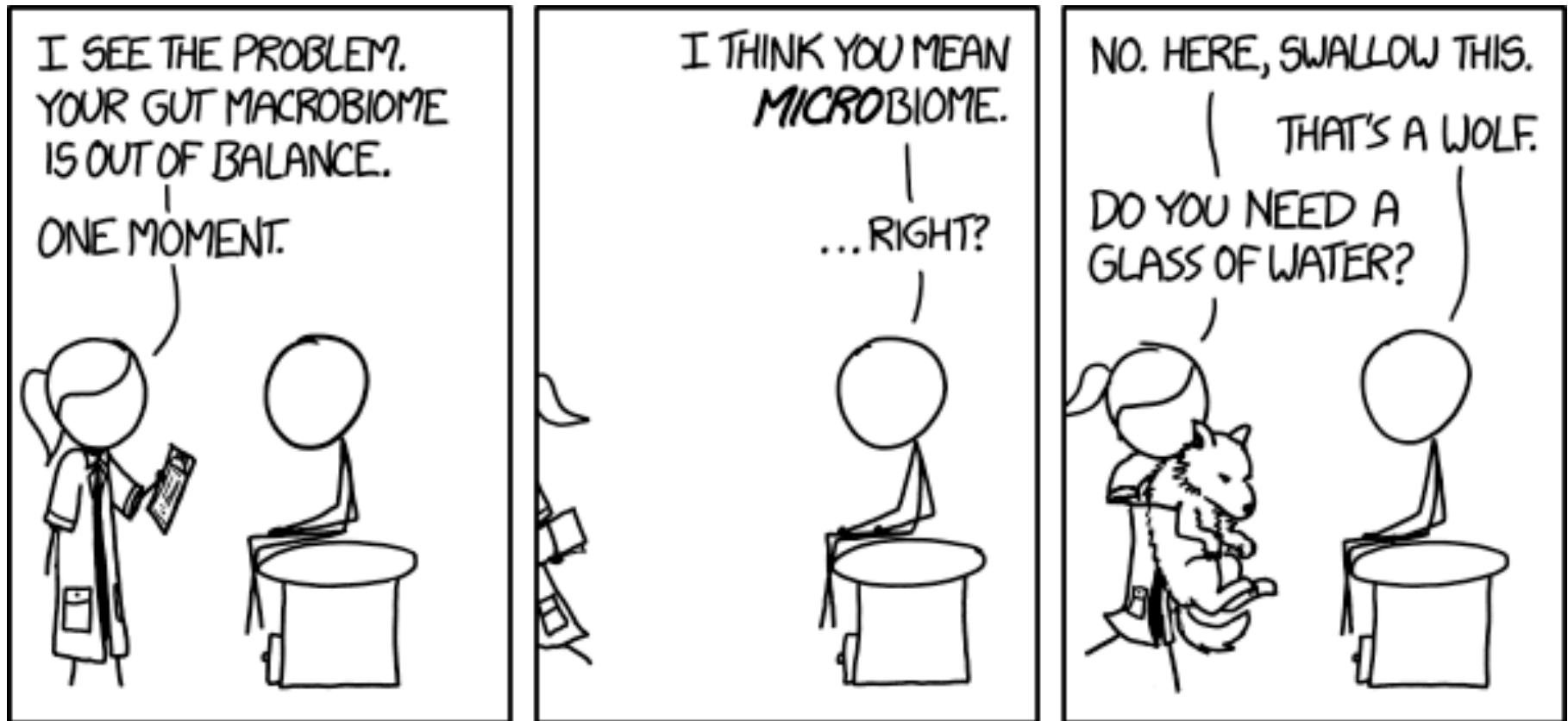


Pool samples
and sequence



>GCACCTGAGGACAGGCATGAGGAA...
>GCACCTGAGGACAGGGGAGGAGGA...
>TCACATGAACCTAGGCAGGACGAA...
>CTACCGGAGGACAGGCATGAGGAT...
>TCACATGAACCTAGGCAGGAGGAA...
>GCACCTGAGGACACGCAGGACGAC...
>CTACCGGAGGACAGGCAGGAGGAA...
>CTACCGGAGGACACACAGGAGGAA...
>GAACCTTCACATAGGCAGGAGGAT...
>TCACATGAACCTAGGGGCAAGGAA...
>GCACCTGAGGACAGGCAGGAGGAA...

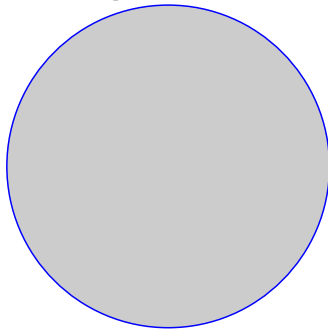
Analyzing Microbial Diversity



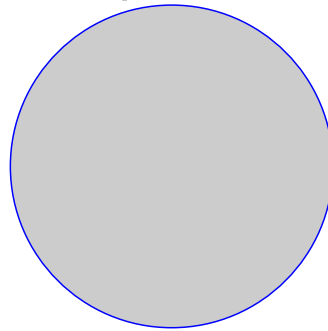
Sequences are clustered into Operational
Taxonomic Units (OTUs)

Similar sequences are grouped together

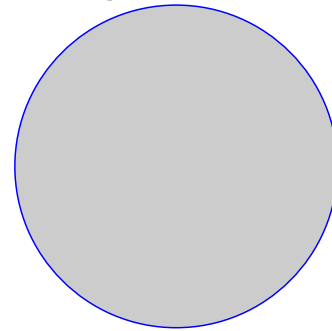
Reference
Sequence 1



Reference
Sequence 2



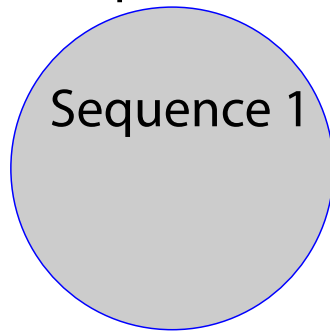
Reference
Sequence 3



Sequence 1
Sequence 2
Sequence 3
Sequence 4
Sequence 5
Sequence 6

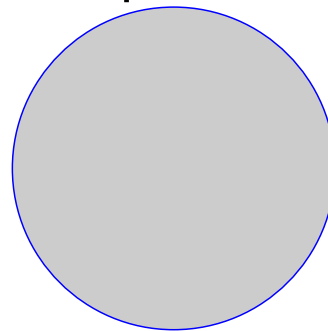
Similar sequences are grouped together

Reference
Sequence 1

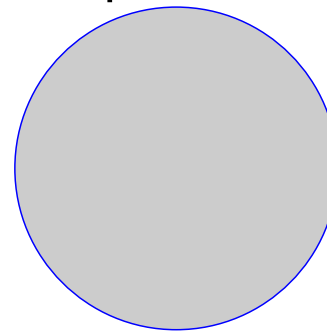


Sequence 1

Reference
Sequence 2



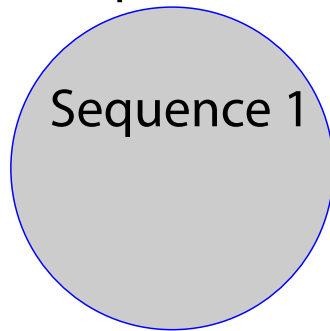
Reference
Sequence 3



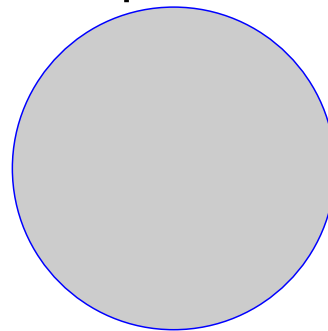
Sequence 2
Sequeunce 3
Sequence 4
Sequence 5
Sequence 6

Similar sequences are grouped together

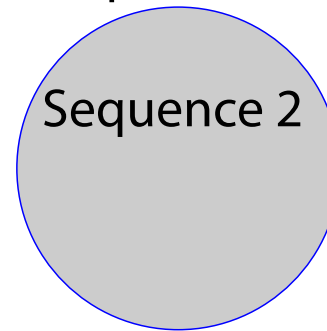
Reference
Sequence 1



Reference
Sequence 2



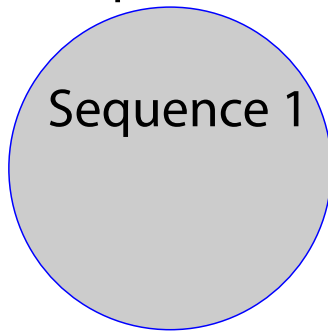
Reference
Sequence 3



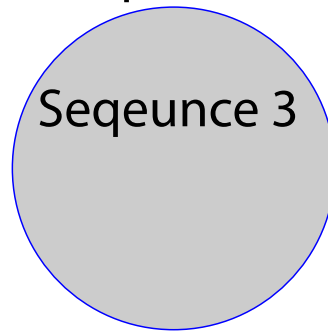
Sequence 3
Sequence 4
Sequence 5
Sequence 6

Similar sequences are grouped together

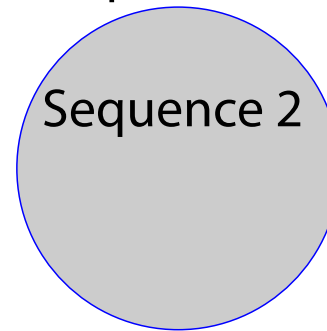
Reference
Sequence 1



Reference
Sequence 2



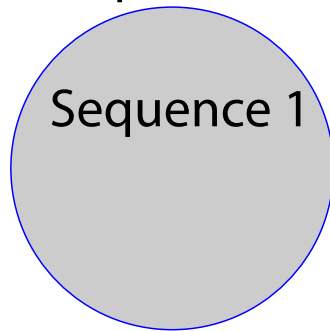
Reference
Sequence 3



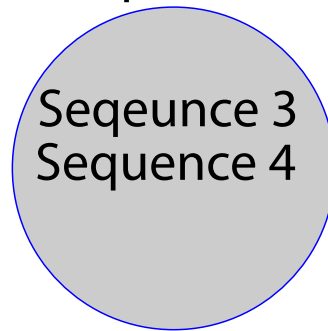
Sequence 4
Sequence 5
Sequence 6

Similar sequences are grouped together

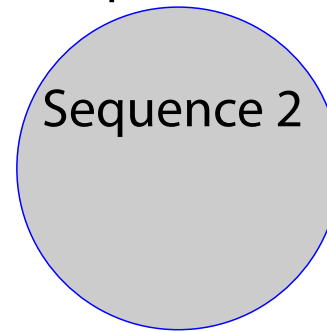
Reference
Sequence 1



Reference
Sequence 2



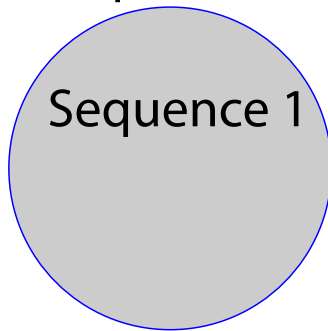
Reference
Sequence 3



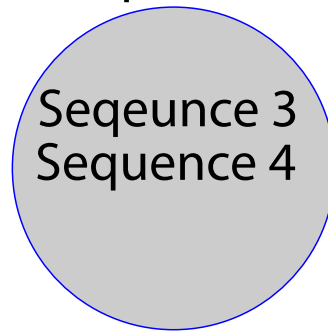
Sequence 5
Sequence 6

If a sequence does not match the database it is assigned to a new cluster

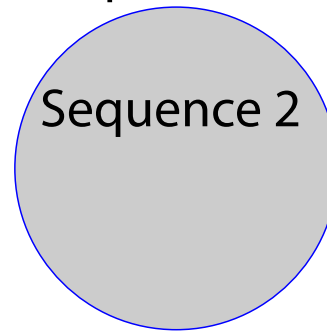
Reference
Sequence 1



Reference
Sequence 2



Reference
Sequence 3



New Cluster



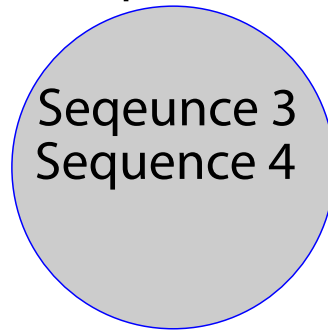
Sequence 6

This continues until all sequences
are assigned

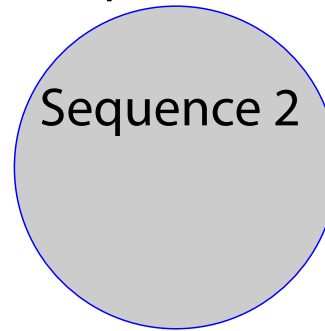
Reference
Sequence 1



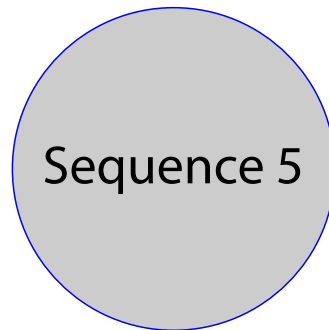
Reference
Sequence 2



Reference
Sequence 3



New Cluster



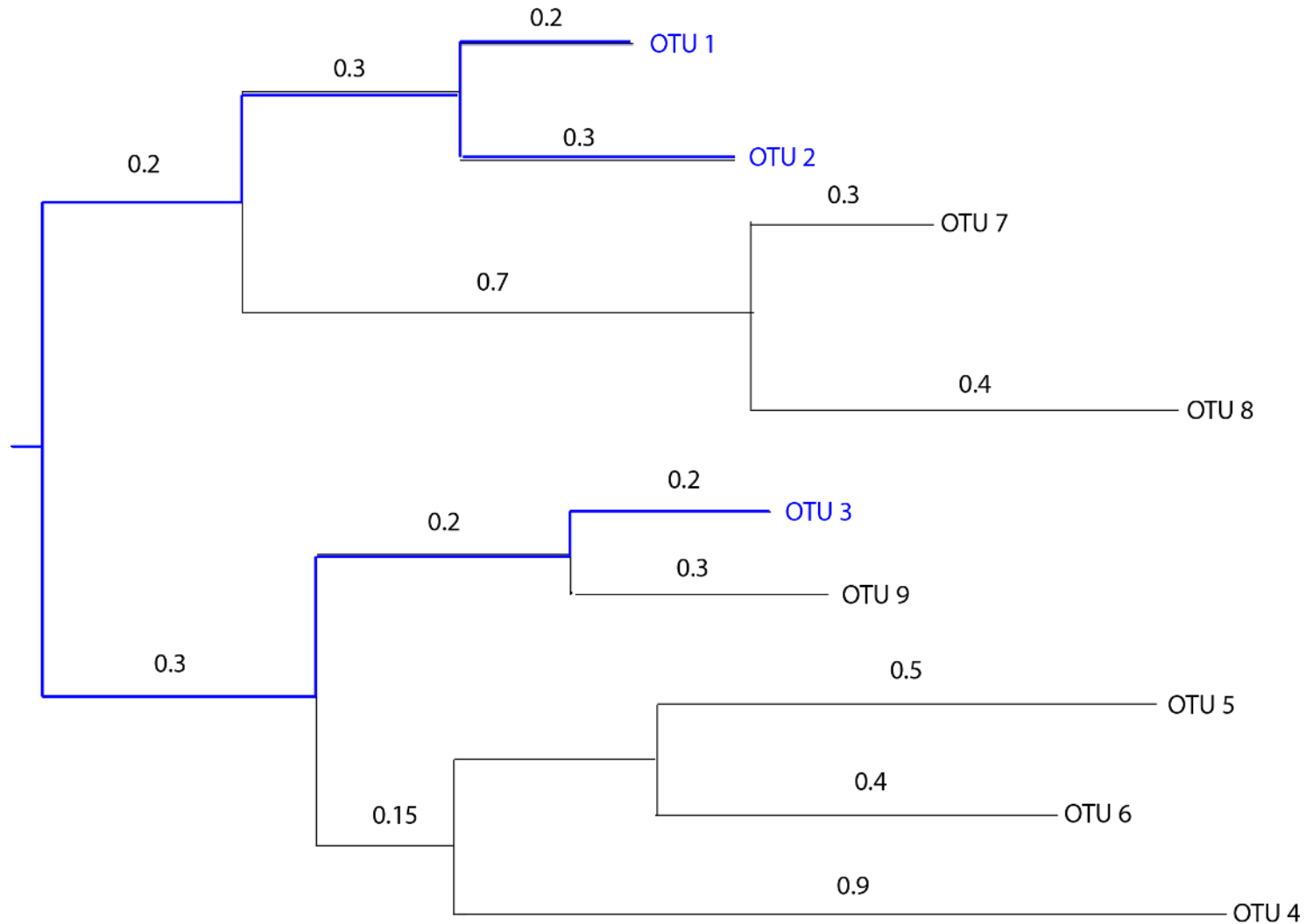
A representative sequence is
chosen from each cluster and an
OUT table is made

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

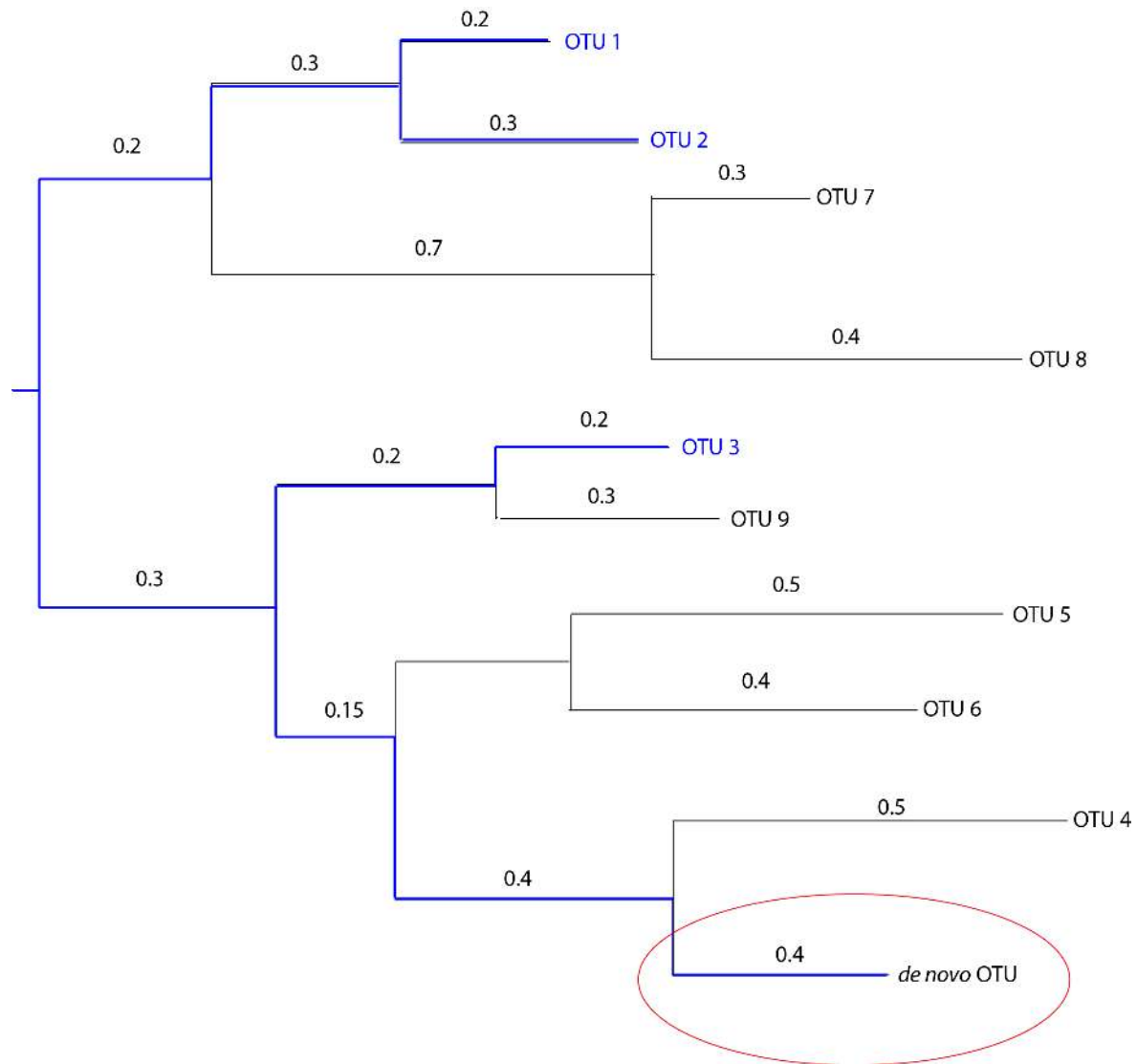
Metadata is recorded for each sample that is taken

	BodySite	Date	Sex
Sample1	gut	01/10/1970	male
Sample2	palm	01/10/1970	male
Sample3	gut	01/25/1970	female
Sample4	palm	01/25/1970	female

A table may be created from an existing phylogenetic tree



New “*de novo*” OTUs can be added to a tree, however phylogenetic analyses should not be performed with a tree that does not contain all OTUs



Comparing microbial communities

Who is there?

How many unique “species” are there?

How similar (or different) are pairs of samples?

How do we compare the communities (Samples)?

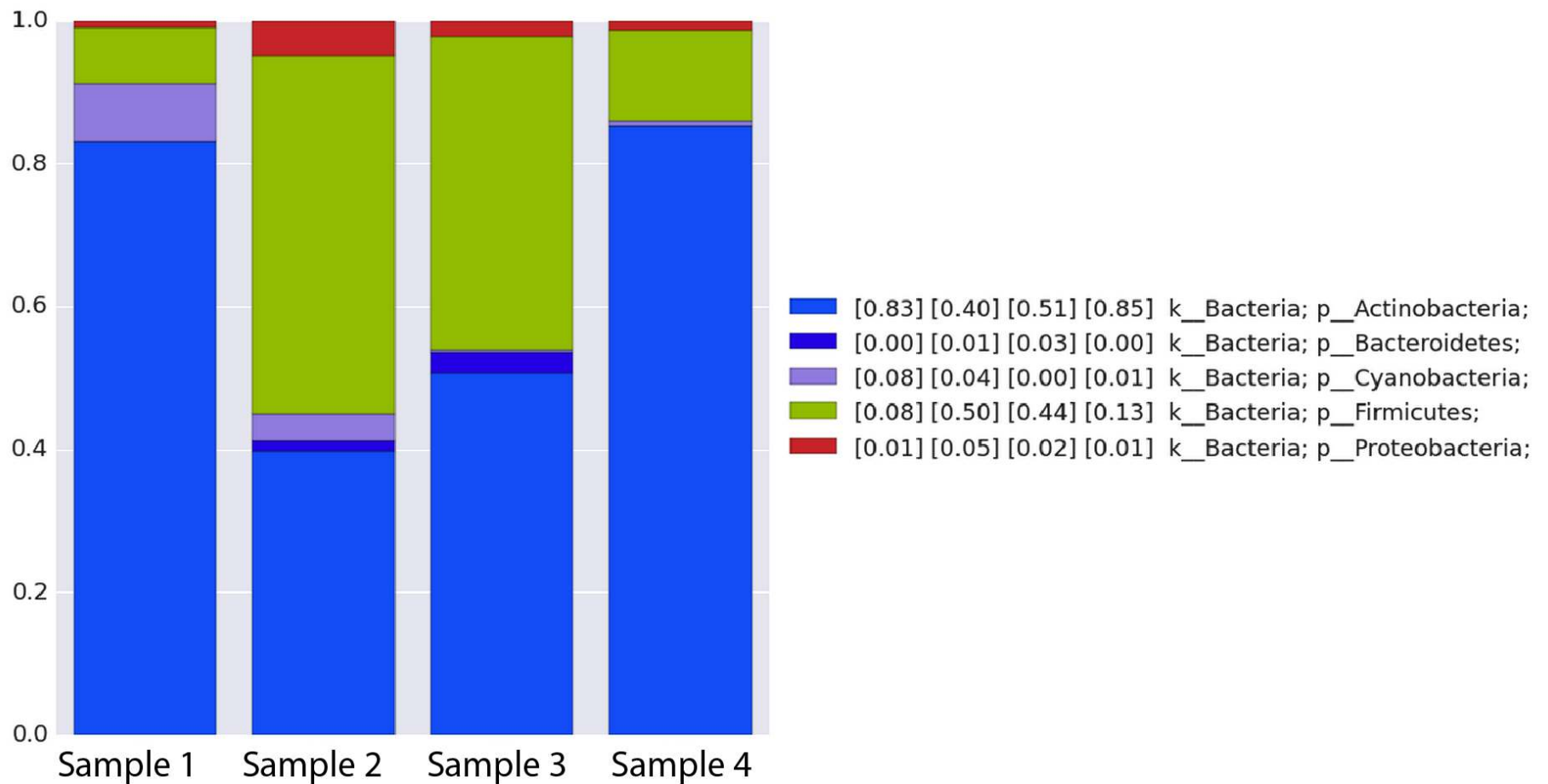
- Taxonomy Summary
- Alpha Diversity (within sample diversity)
- Beta Diversity (between sample diversity)

Some Terminology

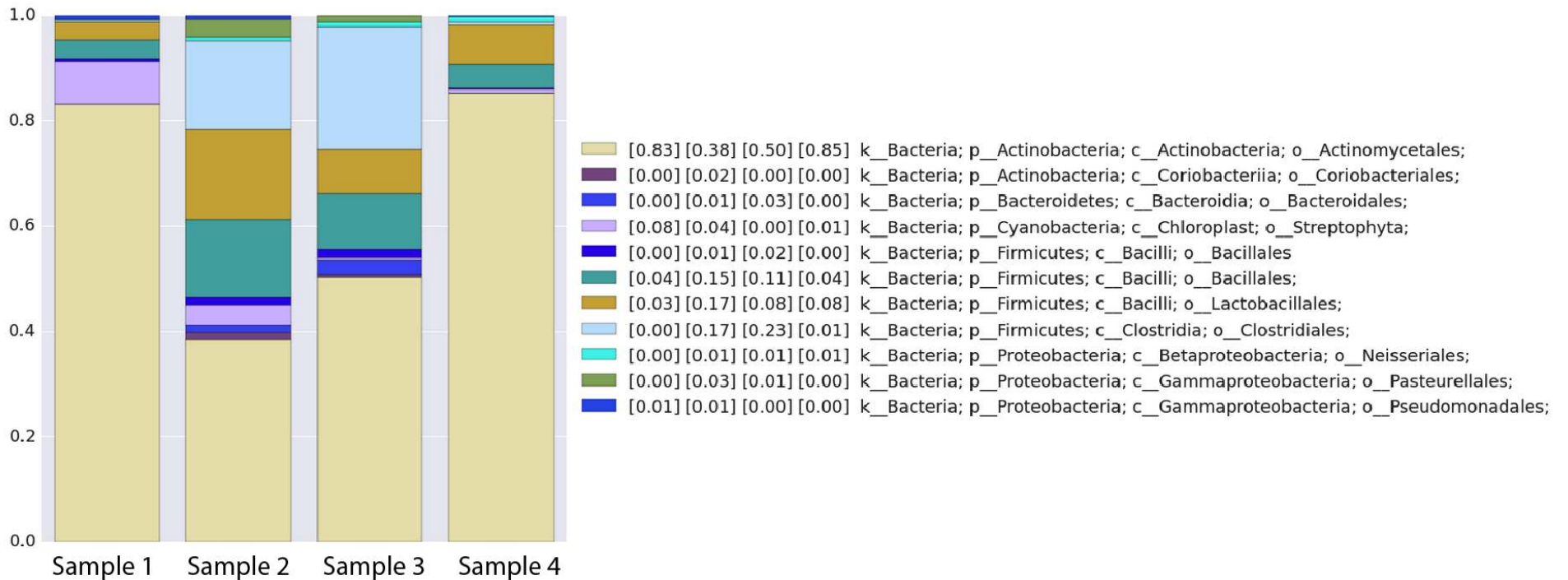
- Qualitative metric - Presence absence of an organism
- Quantitative metric - Takes into account the abundances of different organisms
- Non-phylogenetic metric - Treats all OTUs as being equally related
- Phylogenetic metric - Incorporates evolutionary relationships between the OTUs

Taxonomy Summary

Phylum level differences by sample

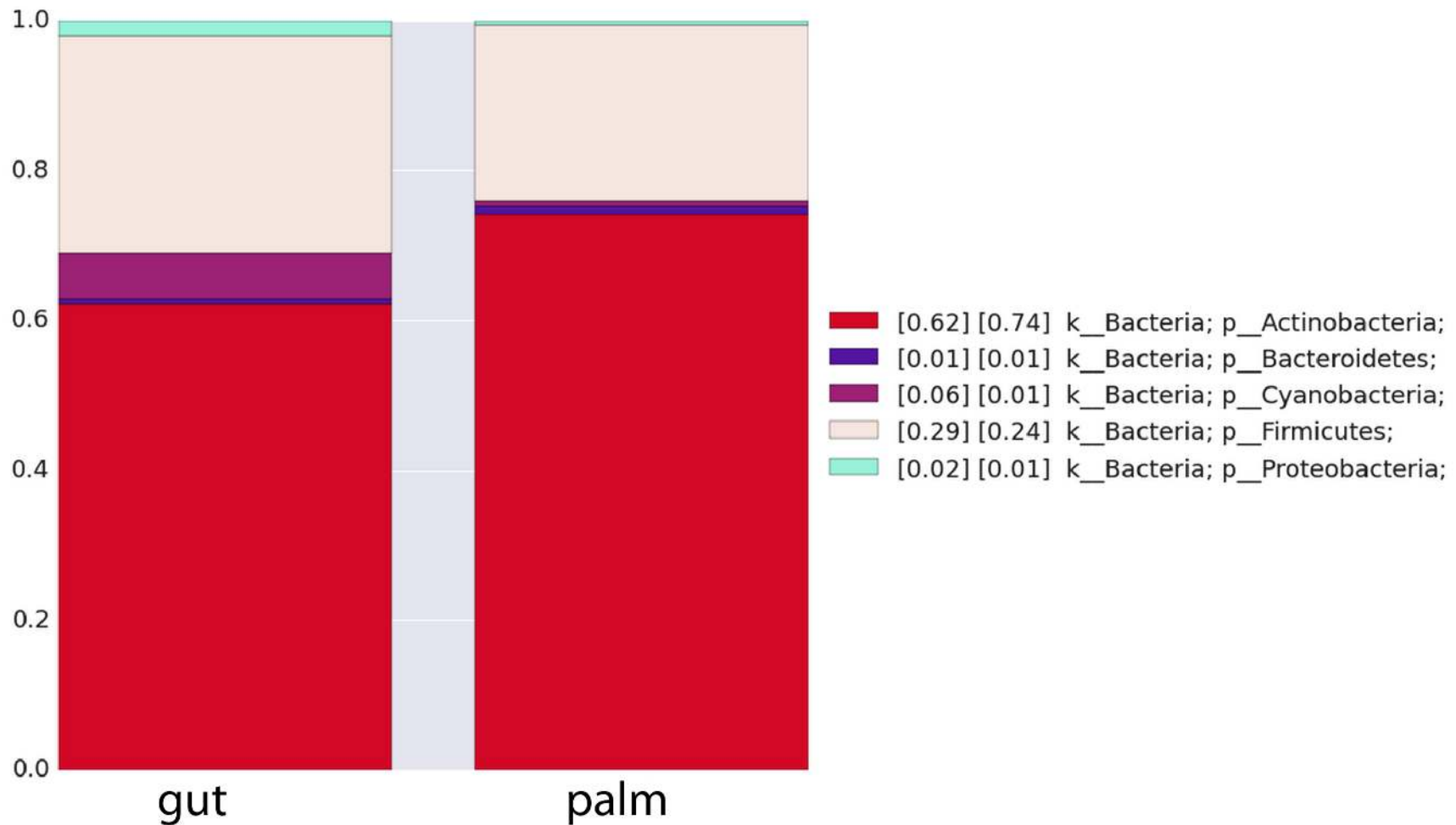


Order level differences by sample



Samples can be combined to look for category differences

(not real data!)



Are relative abundances of OTUs statistically different between groups?

- Kruskal-Wallis (non-parametric)
 - Groups must contain > 5 samples
 - Commonly used for marker gene surveys
- Anova (parametric)
 - Equal variance
 - Normality of residuals
 - Independence
 - Usually violated by marker gene surveys
- *Many others*

Kruskal-Wallis test for OTU differences

OTU	Test Stat	P-value	Bonferroni Corrected p-value	Gut mean	Palm mean
k__Bacteria; p__Firmicutes	10.86	0.001	0.01	1.12E-10	23.751
k__Bacteria; p__Bacteroidetes	5.07	0.005	0.05	0.003	7.767
k__Bacteria; p__Proteobacteria	2	0.078	0.078	1.151	1.616

Alpha Diversity (within sample diversity)

Observed Species: a non-phylogenetic qualitative metric

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

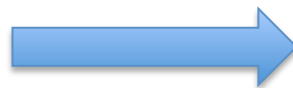
Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Observed species

Sample A 3
Sample B 3
Sample C 3



Conclusion

$A = B = C$

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Alpha diversity (Observed Species)

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

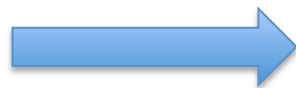
Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Observed species

Sample A 3
Sample B 3
Sample C 3



Conclusion

$A = B = C$

Phylogenetic Diversity: a qualitative phylogenetic metric

Sample A

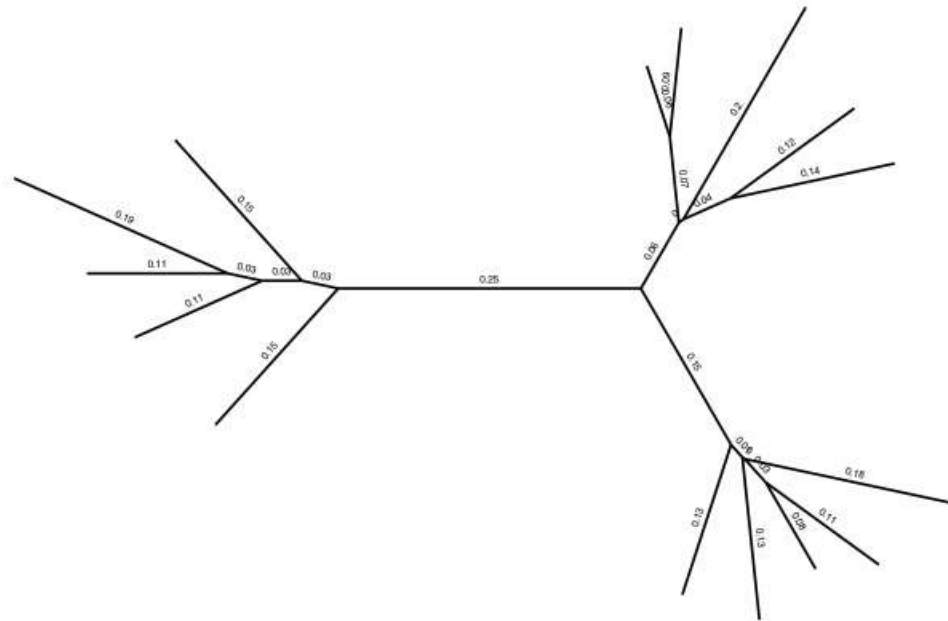
Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli

Sample C

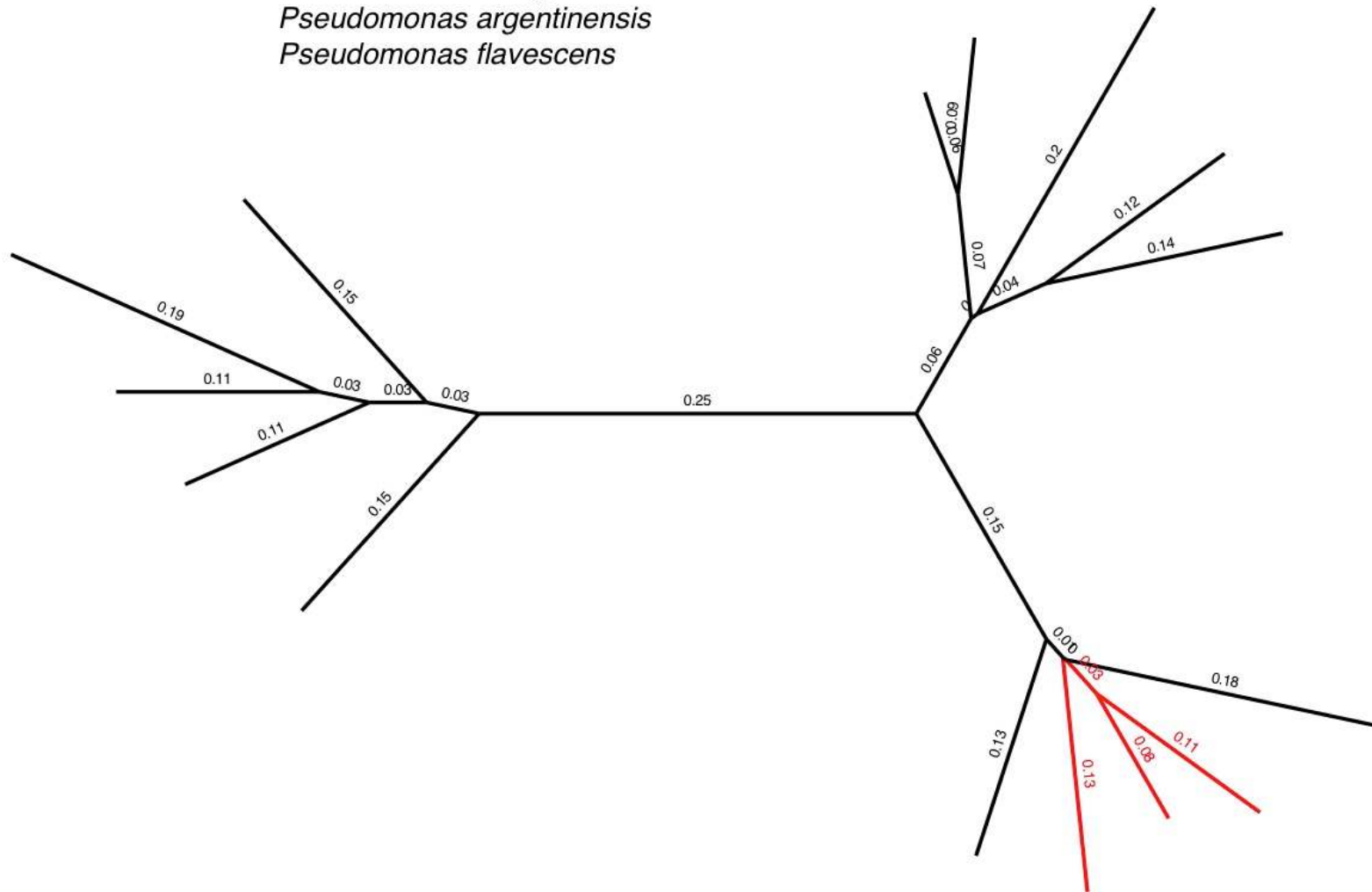
Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



Alpha diversity

Sample A

Pseudomonas aeruginosa
Pseudomonas argentinensis
Pseudomonas flavescens

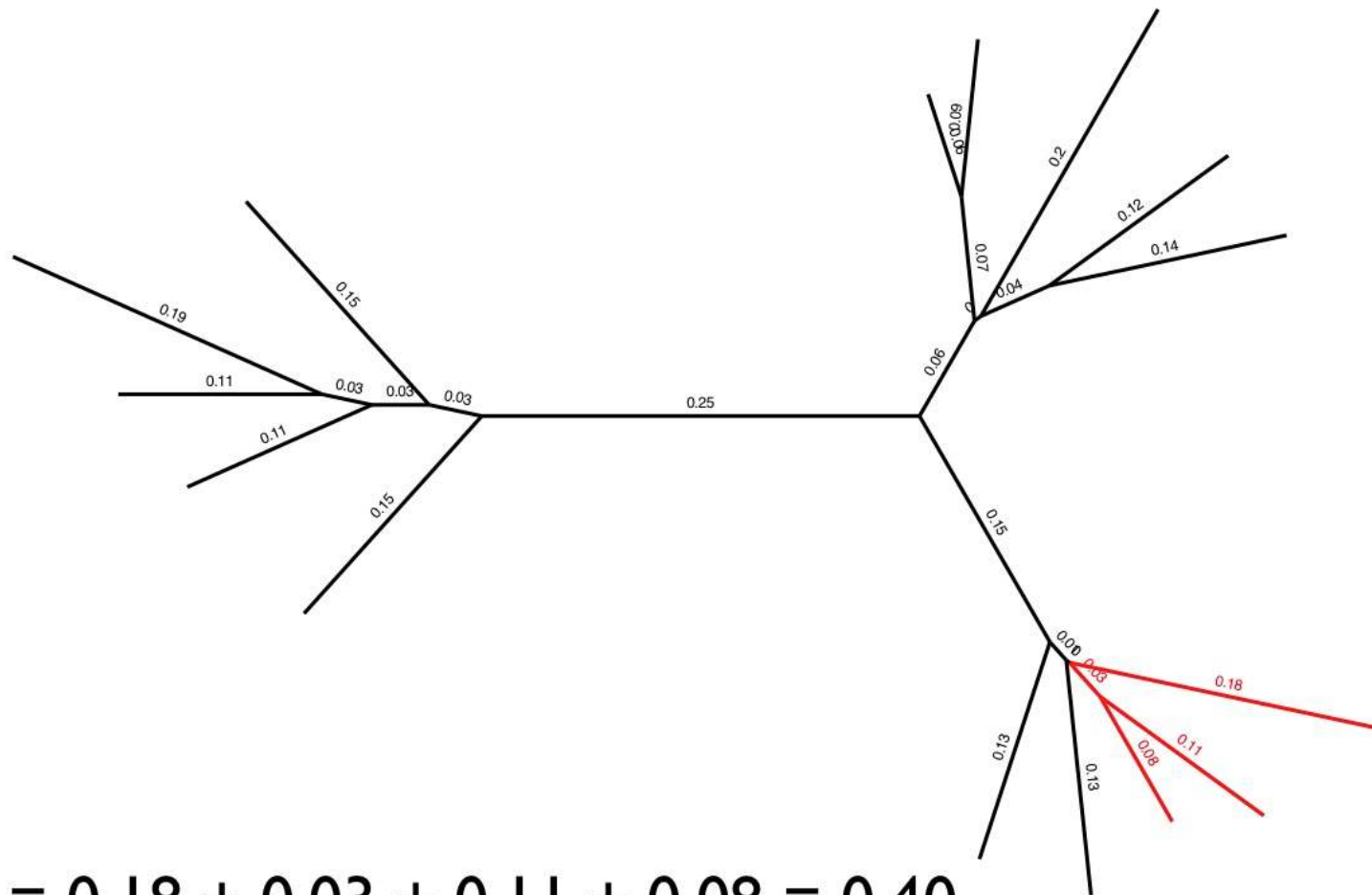


$$PD = 0.13 + 0.03 + 0.11 + 0.08 = 0.35$$

Alpha diversity

Sample B

Pseudomonas aeruginosa
Pseudomonas argentinensis
Escherichia coli



$$PD = 0.18 + 0.03 + 0.11 + 0.08 = 0.40$$

Sample C

Phylogenetic tree showing relationships between *Pseudomonas aeruginosa*, *Giardia lamblia*, and *Methanobrevibacter smithii*. The tree is rooted at the top and branches downwards. The branches are color-coded: black for *Pseudomonas aeruginosa*, red for *Giardia lamblia*, and black for *Methanobrevibacter smithii*. Bootstrap values are indicated on the branches.

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

Alpha 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

PD = 0.35 < PD = 0.40 < PD = 0.95

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

Alpha rarefaction

Sample A
alpha div=20

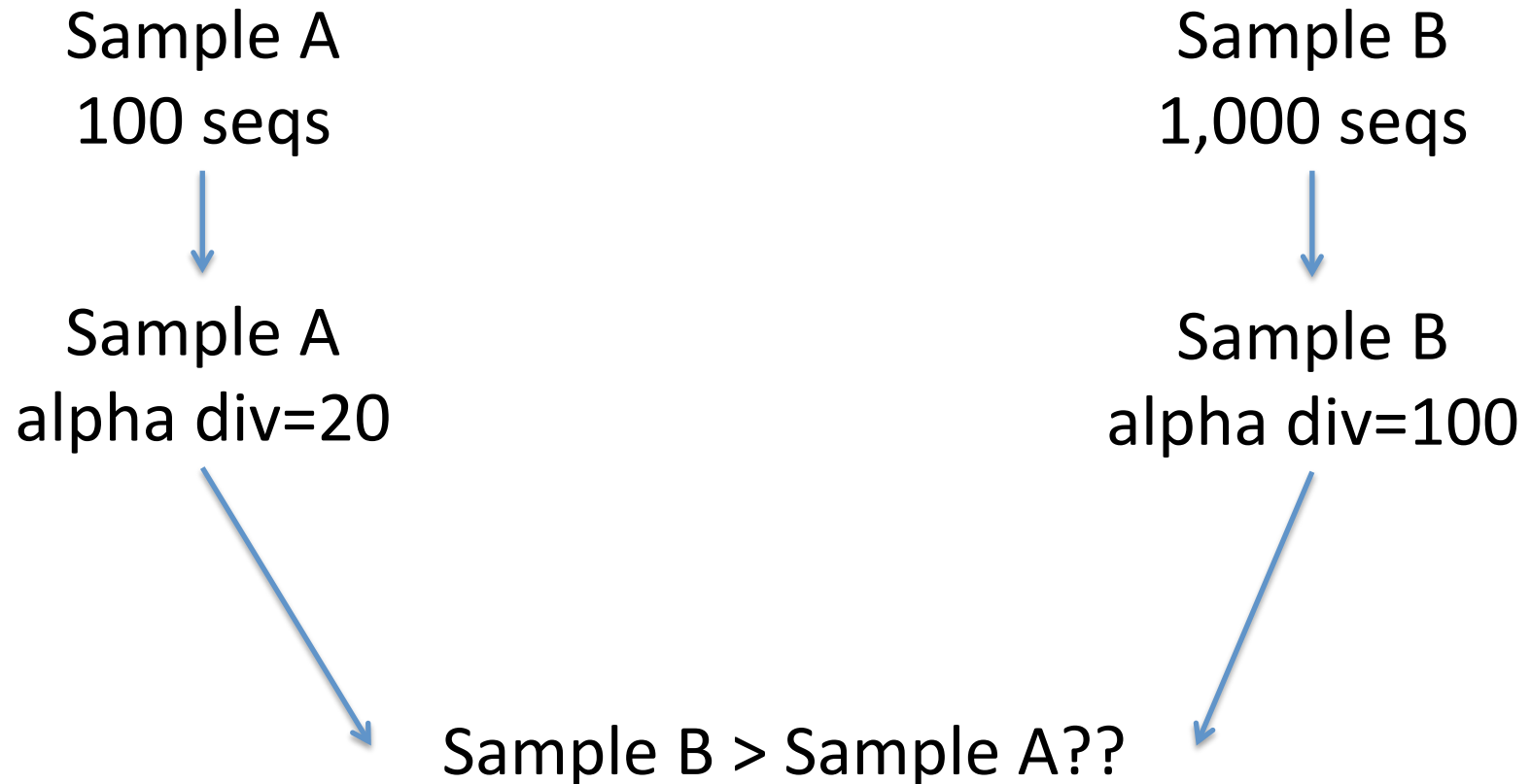
Sample B
alpha div=100



Sample B > Sample A

The diagram illustrates the concept of alpha rarefaction by comparing two samples. Sample A, with an alpha diversity of 20, and Sample B, with an alpha diversity of 100, are shown at the top. Two blue arrows point from each sample towards a central red text statement, 'Sample B > Sample A', which indicates that Sample B has a higher alpha diversity than Sample A.

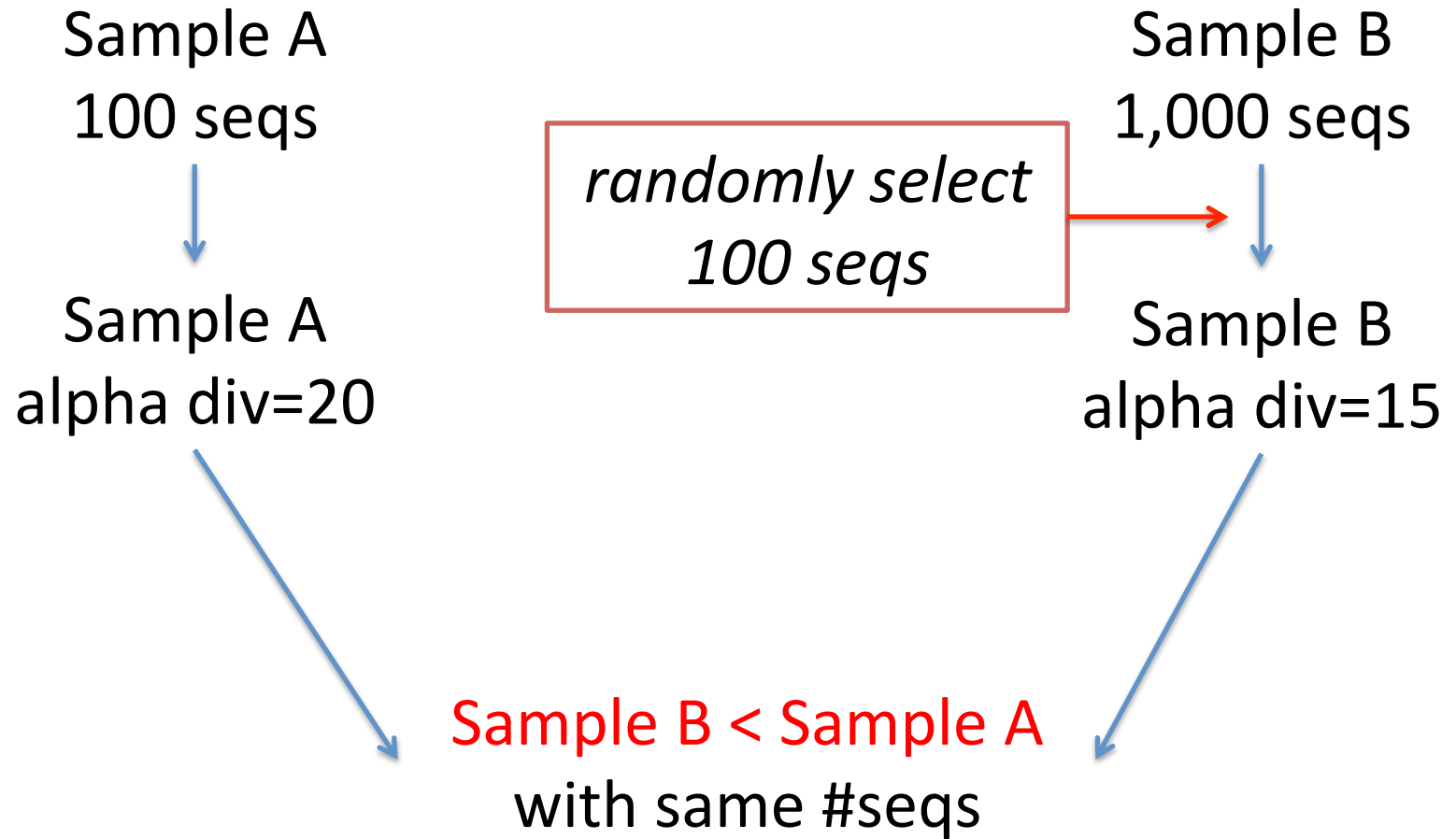
Alpha rarefaction



Alpha rarefaction

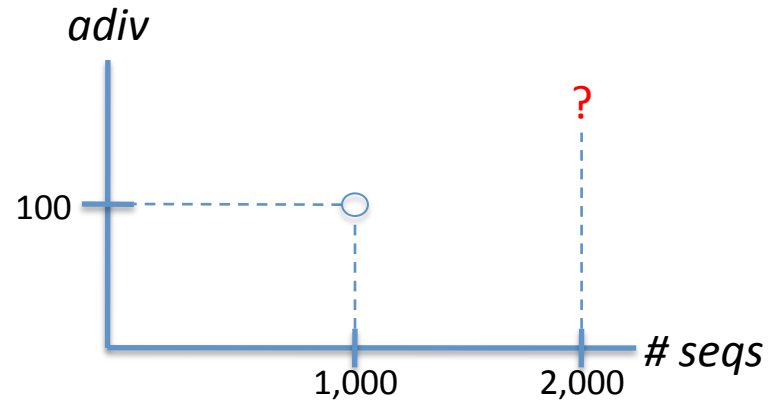


Alpha rarefaction



Multiple alpha rarefaction

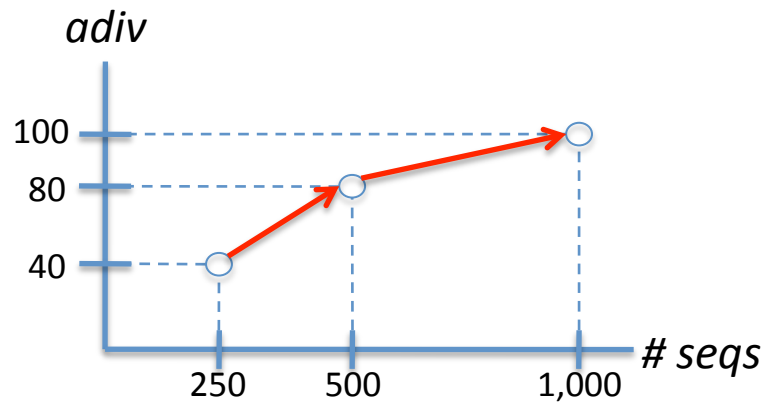
Sample A
Alpha div = 100
with 1,000 seqs



What if we had 2,000 seqs?

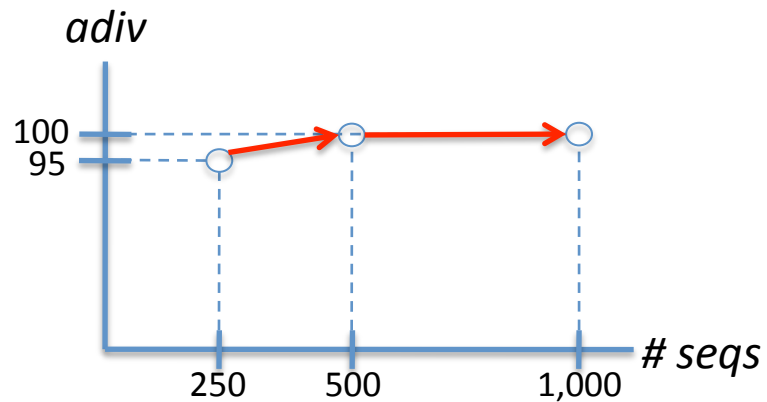
Repeatedly calculate alpha div
at **decreasing** number of seqs

Multiple alpha rarefaction



Higher sequencing effort might result
in higher observed diversity

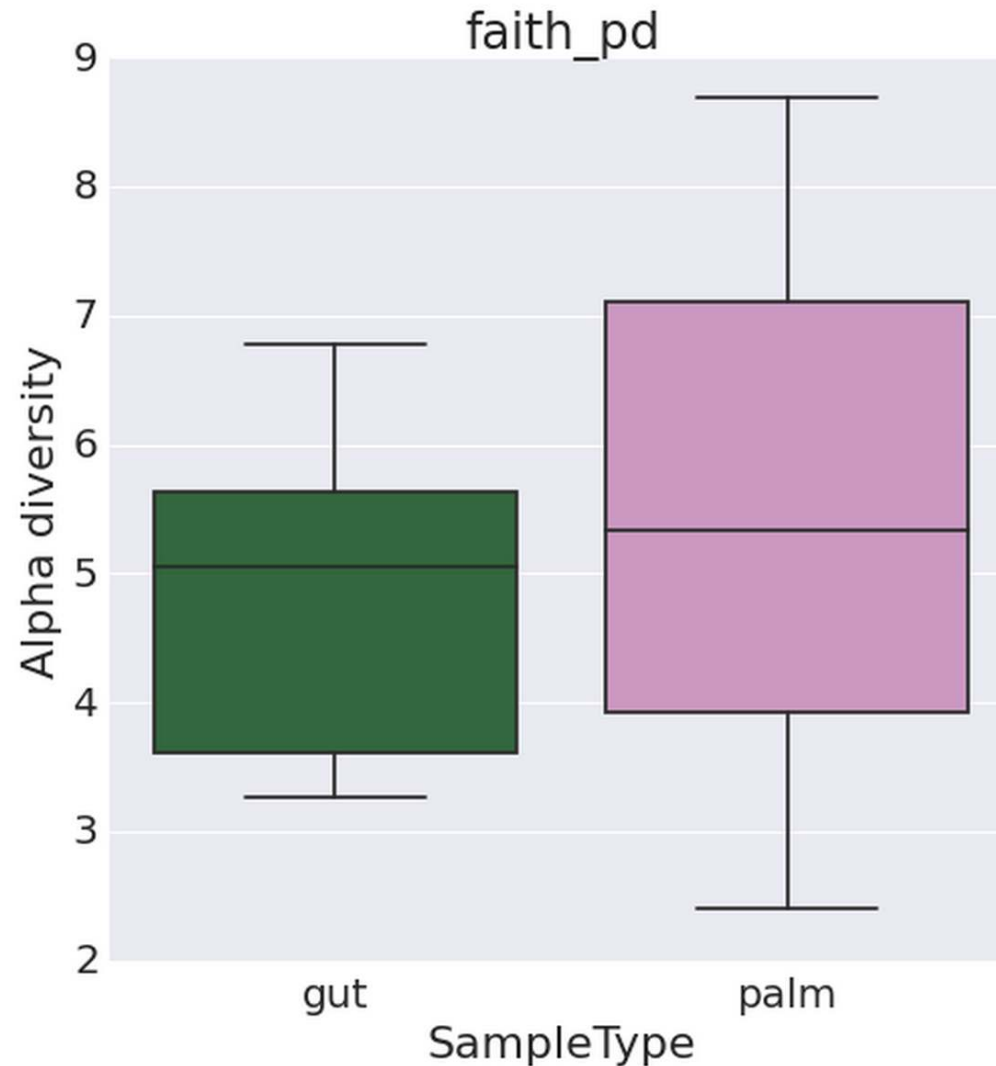
Multiple alpha rarefaction



Higher sequencing effort will probably
not add to observed diversity

- Rarefaction results in a loss of data
- Can be difficult to determine optimal rarefaction level
- Rarefaction increases error
- Not always the best option
 - Small libraries/sample size

Are skin samples more diverse than gut?



Comparing alpha diversity between categories

- T-test
 - Parametric
 - Not usually appropriate
 - Non-parametric
 - Mont carlo permutations
 - More commonly used

Beta Diversity
(between sample diversity)

Alpha Diversity

Sample A

Wolbachia melophagi
Lactobacillus delbruecki
Pseudomonas flavescens

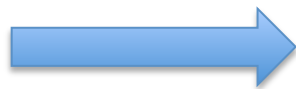
Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii

Sample B

Pseudomonas aeruginosa
Giardia lamblia
Escherichia coli

**Alpha Diversity
Conclusion**



A = B = C

Sample A

Wolbachia melophagi
Lactobacillus delbruecki
Pseudomonas flavescens



Sample C

Pseudomonas aeruginosa
Giardia lamblia
Methanobrevibacter smithii



Sample B

Pseudomonas aeruginosa
Giardia lamblia
Escherichia coli

Bray-Curtis dissimilarity

Non-phylogenetic, quantitative method

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

BC = dissimilarity

Sample

j = jth sample

k = kth sample

i = ith observation

X = Value

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator:
 $|0 + 0| = 0$

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator:

$$|0 - 0| = 0$$

$$|1 - 7| = 6$$

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator

$$|0 - 0| = 0$$

$$|1 - 7| = 6$$

$$|4 - 0| = 4$$

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator:

$$|0 - 0| = 0$$

$$|1 - 7| = 6$$

$$|4 - 0| = 4$$

Denominator

$$0 + 0 = 0$$

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator:

$$|0 - 0| = 0$$

$$|1 - 7| = 6$$

$$|4 - 0| = 4$$

Denominator:

$$0 + 0 = 0$$

$$1 + 7 = 8$$

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator:

$$|0 - 0| = 0$$

$$|1 - 7| = 6$$

$$|4 - 0| = 4$$

Denominator

$$0 + 0 = 0$$

$$1 + 7 = 8$$

$$4 + 0 = 4$$

Bray-Curtis dissimilarity

$$BC_{jk} = \frac{\sum_i |X_{ij} - X_{ik}|}{\sum_i (X_{ij} + X_{ik})}$$

To calculate distance
from Sample 1 to
Sample 2

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

Numerator:

$$|0 - 0| = 0$$

$$|1 - 7| = 6$$

$$|4 - 0| = 4$$

Denominator

$$0 + 0 = 0$$

$$1 + 7 = 8$$

$$4 + 0 = 4$$

$$BC = 10/12$$

Unweighted Unifrac

A phylogenetic qualitative metric

$$U_{AB} = \frac{\textit{unique}}{\textit{observed}}$$

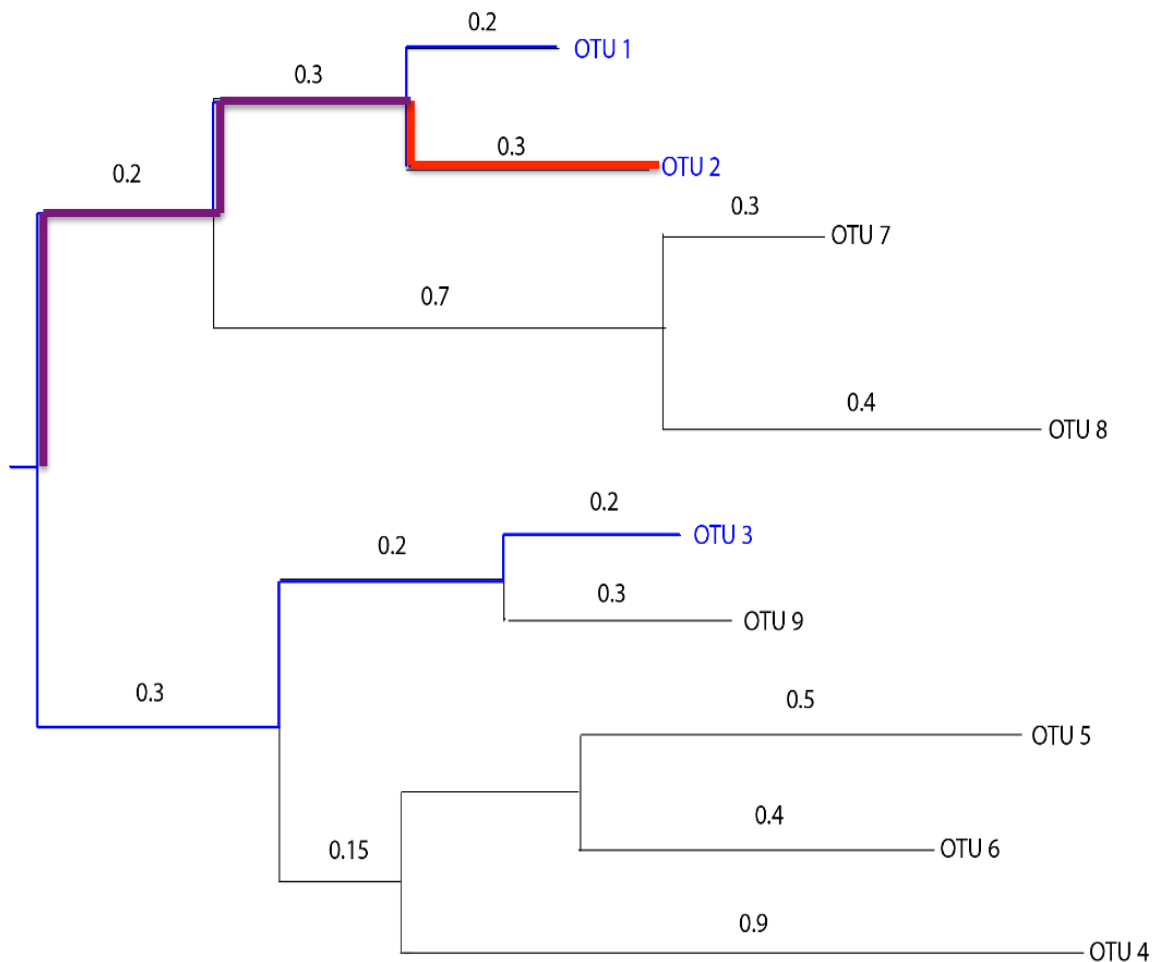
Unique: the sum of the branch lengths that only lead to an observations found in one of samples being compared

Observed: the sum of the branch lengths that lead to an observation in either sample

	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

$$U_{AB} = \frac{\text{unique}}{\text{observed}}$$

Sample 2 to Sample
3 Distance



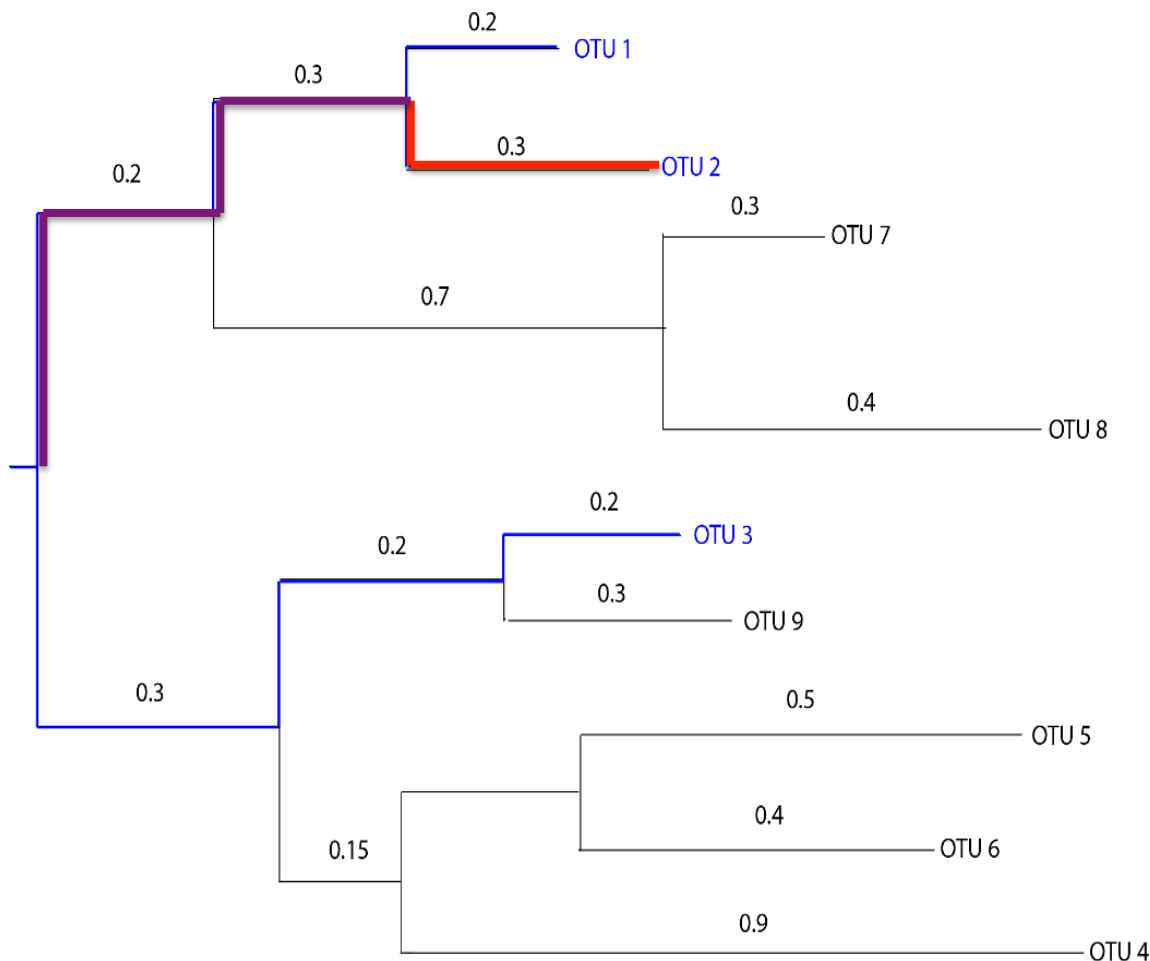
	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

$$U_{AB} = \frac{\text{unique}}{\text{observed}}$$

Sample 2 to Sample 3 Distance

$$\text{Unique} = 0.2 + 0.3 + 0.3 + 0.2 + 0.2 + 0.2$$

$$\text{Observed} = 0.2 + 0.3 + 0.2 + 0.3 + 0.3 + 0.2 + 0.2 + 0.2$$



	Sample 1	Sample 2	Sample 3	Sample 4
OTU 1	0	0	4	6
OTU 2	1	7	0	1
OTU 3	4	0	4	2

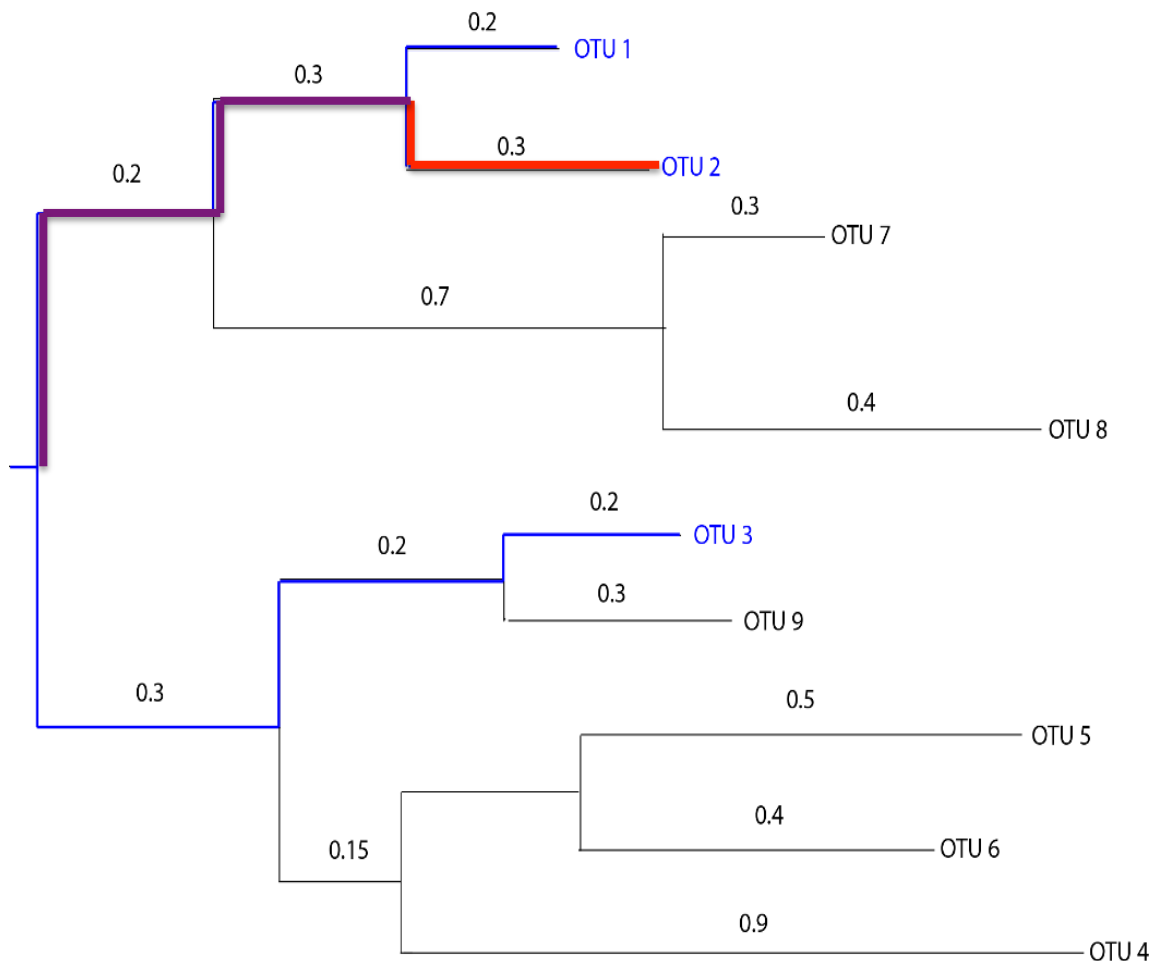
$$U_{AB} = \frac{\text{unique}}{\text{observed}}$$

Sample 2 to Sample 3 Distance

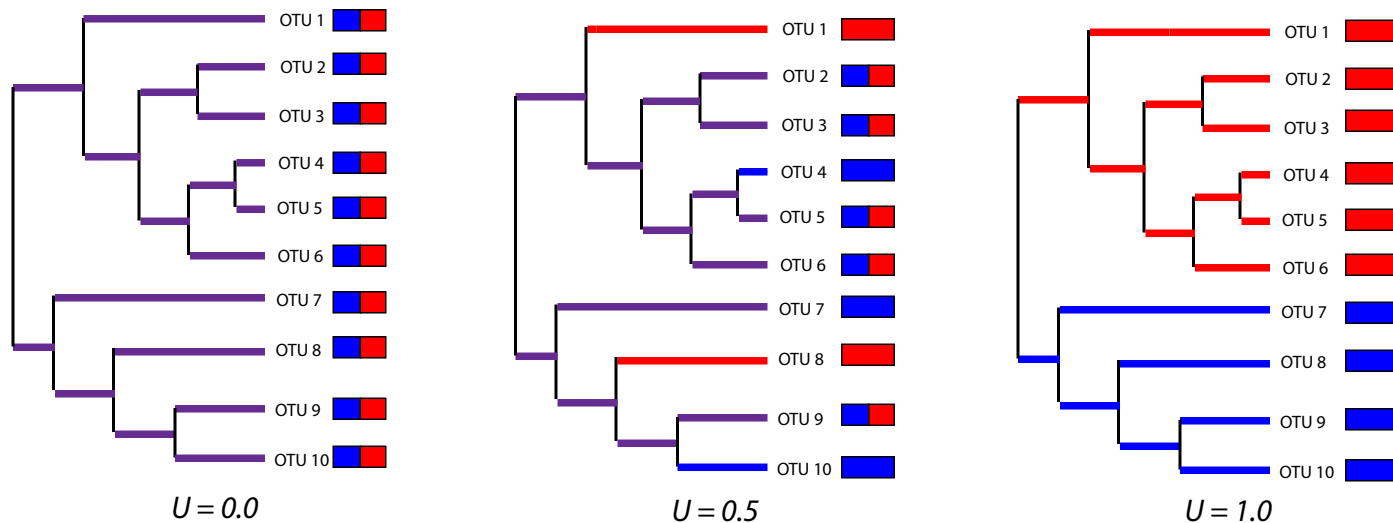
Unique = $0.2 + 0.3 + 0.3 + 0.2 + 0.2 + 0.2$

$$\begin{aligned} \text{Observed} &= 0.2 + 0.3 \\ &+ 0.2 + 0.3 + 0.3 + \\ &0.2 + 0.2 + 0.2 \end{aligned}$$

$$U_{AB} = 1.4/1.9$$



Unweighted Unifrac: a phylogenetic measure of the dissimilarity of microbial communities



$$U_{AB} = \frac{\text{unique}}{\text{observed}}$$

where:

unique : the unique branch length, or branch length that only leads to OTU(s) observed in sample *A* or sample *B*

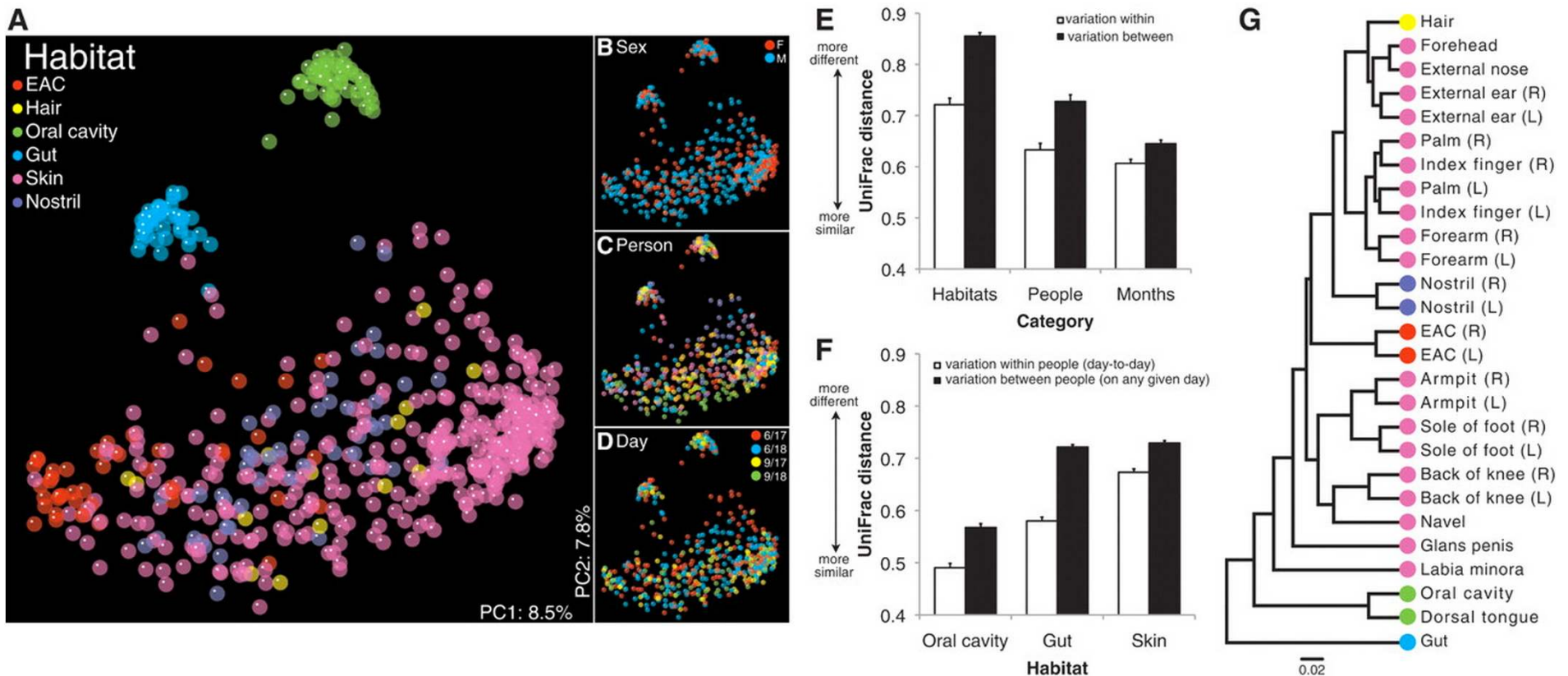
observed : the total branch length observed in either sample *A* or sample *B*

Result of beta diversity is a distance matrix

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

	unweighted_unifrac_dm.txt															
1	Stillton4R2	Stillton4R3	Stillton4R1	HCanyon3R3	HCanyon3R2	HCanyon3R1	Halls9R2	HCanyon2R2	HCanyon2R3	HCanyon2R1	Halls9R1	Stillton10R3	HCanyon0R1	HCanyon0R2	HCanyon0R3	HCa
2	Stillton4R2	0.0	0.382273294624	0.391416675288	0.560309484808	0.553938232028	0.566136031815	0.557134987546	0.531719852875	0.53655901824	0.567041909667	0.574831935502	0.474244845718	0.47910046195	0.47910046195	
3	Stillton4R3	0.382273294624	0.0	0.394399899497	0.560309484808	0.58366320358	0.58366320358	0.58366320358	0.58366320358	0.58366320358	0.58366320358	0.58366320358	0.58366320358	0.58366320358	0.58366320358	
4	Stillton4R1	0.391416675288	0.394399899497	0.0	0.580628929798	0.583767102601	0.584463513906	0.573828216898	0.550892933103	0.566056393448	0.584611276938	0.607144746402	0.469209424282	0.469209424282	0.469209424282	
5	HCanyon3R3	0.560309484808	0.586303332083	0.580628929798	0.0	0.354600316745	0.345731612479	0.458450145834	0.412488628393	0.385127601086	0.384525100123	0.46823652766	0.585402372425	0.585402372425	0.585402372425	
6	HCanyon3R2	0.553938232028	0.58366320358	0.583767102601	0.354600316745	0.0	0.36256949665	0.462707750398	0.414528760004	0.385453766442	0.380087766632	0.46109426257	0.574488221279	0.574488221279	0.574488221279	
7	HCanyon3R1	0.566136031815	0.589363298118	0.584463513906	0.345731612479	0.36256949665	0.0	0.452351806146	0.415483559719	0.398796424875	0.381981707704	0.485596583752	0.580926122772	0.580926122772	0.580926122772	
8	Halls9R2	0.557134987546	0.588066262733	0.573828216898	0.458450145834	0.462707750398	0.452351806146	0.0	0.447883445295	0.429943464459	0.409064513124	0.344264504725	0.561110815583	0.561110815583	0.561110815583	
9	HCanyon2R2	0.531719852875	0.560923487354	0.550892933103	0.412488628393	0.414528760004	0.415483559719	0.447883445295	0.0	0.404179520995	0.388727659604	0.468345488157	0.549328940024	0.549328940024	0.549328940024	
10	HCanyon2R3	0.53655901824	0.57767776062	0.566056393448	0.385127601086	0.385453766442	0.398796424875	0.429943464459	0.404179520995	0.0	0.361444528983	0.439367245599	0.566330894525	0.566330894525	0.566330894525	
11	HCanyon2R1	0.567041909667	0.601572335069	0.584611276938	0.384525100123	0.380087766632	0.381981707704	0.409064513124	0.388727659604	0.361444528983	0.0	0.454633280595	0.582635621853	0.582635621853	0.582635621853	
12	Halls9R1	0.574831935502	0.609758664376	0.607144746402	0.46823652766	0.46109426257	0.485596583752	0.344264504725	0.468345488157	0.439367245599	0.454633280595	0.0	0.578254150137	0.578254150137	0.578254150137	
13	Stillton10R3	0.474244845718	0.479910046195	0.469209424282	0.585402372425	0.574488221279	0.580926122772	0.561110815583	0.549328940024	0.566330894525	0.582635621853	0.578254150137	0.578254150137	0.578254150137	0.578254150137	
14	HCanyon0R1	0.776189708931	0.779315452075	0.781071344042	0.69071344042	0.69246412288	0.683086513009	0.727520276229	0.709578414217	0.711736423222	0.707912554104	0.745189206089	0.7	0.745189206089	0.745189206089	
15	HCanyon0R2	0.797241309582	0.795032626325	0.797809039119	0.713804444578	0.714717508733	0.697490417772	0.75335887058	0.723508997775	0.728837315629	0.722257776622	0.761081079305	0.7	0.761081079305	0.761081079305	
16	HCanyon0R3	0.78412436272	0.791711071574	0.795527048923	0.693476495058	0.699412891734	0.688360016409	0.73248914053	0.713925200799	0.708354096579	0.708462144382	0.747778043238	0.7	0.747778043238	0.747778043238	
17	HCanyon7R3	0.539194289149	0.563745794923	0.557216919336	0.457709184042	0.467073058252	0.473888308307	0.42997132225	0.455576623999	0.448029116277	0.441566044862	0.444788573194	0.5	0.444788573194	0.444788573194	
18	HCanyon7R2	0.665877547192	0.683833494738	0.674471865383	0.55308292175	0.563316168524	0.562611936812	0.546938312206	0.558741865535	0.562533096516	0.534453165883	0.558141564634	0.6	0.558141564634	0.558141564634	
19	HCanyon7R1	0.554857668962	0.578652540309	0.570402839551	0.475303182122	0.474309745536	0.462587318718	0.43076590681	0.463222699451	0.452021582021	0.42664191335	0.44552755863	0.5	0.44552755863	0.44552755863	
20	HCanyon1R1	0.746066596617	0.754645795641	0.759778994683	0.654674045733	0.651121732227	0.640930093945	0.703134531793	0.67682938985	0.675957137688	0.67695919301	0.714539249748	0.7	0.714539249748	0.714539249748	
21	HCanyon1R3	0.751522854919	0.764074726397	0.767226018039	0.641116735477	0.65043468158	0.629924888988	0.702785457767	0.675191009098	0.668258446703	0.66343689429	0.713256418398	0.7	0.713256418398	0.713256418398	
22	HCanyon1R2	0.748585453295	0.753698061339	0.759152313976	0.666322115311	0.666043304584	0.654075859985	0.711448710816	0.680563093126	0.679964026606	0.678341781539	0.728364004942	0.7	0.728364004942	0.728364004942	
23	HCanyon10R2	0.56835786215	0.584149104488	0.584183424759	0.460055758676	0.464168854592	0.470680372129	0.458866733827	0.4814190318	0.460896823727	0.442543616102	0.4619867362	0.5	0.4619867362	0.4619867362	
24	HCanyon10R3	0.50798042641	0.588843077275	0.591942476159	0.50320590177	0.498226026993	0.503094520399	0.450352212318	0.496962587514	0.490816424808	0.497919293873	0.463277414447	0.5	0.463277414447	0.463277414447	
25	HCanyon10R1	0.505926071682	0.528347698837	0.529350507225	0.472413250039	0.479553556468	0.478832889376	0.455918418773	0.458027517992	0.428733419477	0.456719821741	0.457712834343	0.5	0.457712834343	0.457712834343	
26	HCanyon11R3	0.582831143875	0.594413158276	0.60681715155	0.463621293703	0.45842280462	0.486656215192	0.430493752745	0.47439497384	0.445815433526	0.432013882772	0.43706115715	0.5	0.43706115715	0.43706115715	
27	HCanyon11R2	0.52855255543	0.55774407831	0.555177188694	0.432763917311	0.429877059678	0.440567014981	0.41332044653	0.426696806769	0.398447088329	0.396059530935	0.433778641456	0.5	0.433778641456	0.433778641456	
28	HCanyon11R1	0.555250897098	0.579991048797	0.575126874355	0.453486254954	0.44588780311	0.456712334902	0.419232406891	0.466869791255	0.430325765526	0.428966840449	0.431167398777	0.5	0.431167398777	0.431167398777	
29	HCanyon6R2	0.568527068157	0.58513246508	0.583868790939	0.430174126838	0.419369519662	0.432708499988	0.416828113391	0.445815626173	0.426928785146	0.410839210322	0.422526407071	0.5	0.422526407071	0.422526407071	
30	HCanyon6R3	0.566346578975	0.587780110737	0.586940661236	0.438634245182	0.425883025257	0.449519180577	0.424862337935	0.447100684715	0.425434475646	0.426218627557	0.424693270252	0.5	0.424693270252	0.424693270252	
31	HCanyon6R1	0.58392769946	0.586778207344	0.581534515565	0.434912987895	0.424307243199	0.441765941119	0.415236907762	0.428454635151	0.428113957425	0.406954252133	0.422536839493	0.5	0.422536839493	0.422536839493	
32	HCanyon5R1	0.558910457622	0.57470972671	0.562821643814	0.440681773256	0.451596106791	0.442162621284	0.435763316753	0.417737787339	0.416631599644	0.455877139948	0.455877139948	0.5	0.455877139948	0.455877139948	
33	HCanyon5R3	0.559987492698	0.5820196982	0.584250873735	0.450336948189	0.426574240559	0.448524323812	0.439745925494	0.440643510478	0.417613041188	0.413563495656	0.464820696561	0.5	0.464820696561	0.464820696561	
34	HCanyon5R2	0.577951660858	0.593568671618	0.596389572645	0.422545765978	0.409515246412	0.434403209926	0.429247944328	0.45387029905	0.431590973836	0.41366014112	0.436175614058	0.5	0.436175614058	0.436175614058	
35	HCanyon4R1	0.516189922282	0.555158871988	0.544022608263	0.416449154884	0.420465768689	0.437380878857	0.46015301609	0.459537309642	0.425192263132	0.439771880651	0.474121489459	0.5	0.474121489459	0.474121489459	
36	HCanyon4R2	0.601827528628	0.632734373589	0.628005145571	0.41666304996	0.417716489314	0.443390188238	0.497188855337	0.473766452031	0.447416361784	0.454050692681	0.502442274303	0.6	0.502442274303	0.502442274303	
37	HCanyon4R3	0.624113459554	0.653406187405	0.649892943176	0.438848993112	0.452700670776	0.45666604112	0.506388612937	0.493725468949	0.479954328128	0.462327913958	0.510888415471	0.6	0.510888415471	0.510888415471	
38	HCanyon12R1	0.513376891954	0.547194076841	0.519066521533	0.471737648119	0.464638931219	0.461841527684	0.430362902794	0.439277983407	0.419105789781	0.41992903626	0.466361101825	0.5	0.466361101825	0.466361101825	
39	HCanyon12R2	0.552892298156	0.560516044793	0.549257756132	0.504271040435	0.504749255647	0.517681811493	0.453014756694	0.480827841616	0.489487879426	0.485996535935	0.465436661211	0.5	0.465436661211	0.465436661211	
40	HCanyon12R3	0.553185666616	0.585465060164	0.585665024134	0.439712101711	0.431336995109	0.463612059381	0.409326408365	0.435088785406	0.414920465304	0.407840817122	0.422739541765	0.5	0.422739541765	0.422739541765	
41	HCanyon8R1	0.552014316498	0.559504417239	0.562836483618	0.483602818412	0.470838841272	0.485053828711	0.446654297414	0.465192835869	0.464892486774	0.441733317446	0.466837938413	0.5	0.466837938413	0.466837938413	
42	HCanyon8R2	0.547616166832	0.576958110793	0.578527645361	0.458907829212	0.452778768072	0.456800424019	0.420176245854	0.449481849055	0.430019204227	0.41302664081	0.432050808518	0.5	0.432050808518	0.432050808518	
43	HCanyon8R3	0.549015567252	0.573444281355	0.580500957664	0.446205572204	0.44699609932	0.453171155433	0.4327591563	0.439870648001	0.425623148959	0.408822227859	0.446705079363	0.5	0.446705079363	0.446705079363	
44	Halls8R2	0.535707795107	0.569114464296	0.574510728774	0.46528819939	0.452053418068	0.460087502347	0.379587194176	0.458489263366	0.421181481291	0.440290750744	0.366881097176	0.5	0.366881097176	0.3668	

Pairwise distances between samples are the basis of most microbiome surveys



Bacterial Community Variation in Human Body Habitats Across Space and Time.
Costello et al. *Science* (2009)

Polar Ordination

$$a = \frac{D^2 + D1^2 - D2^2}{2 \times D}$$

D = is distance between the endpoints

D1 is distance between the current sample and endpoint 1

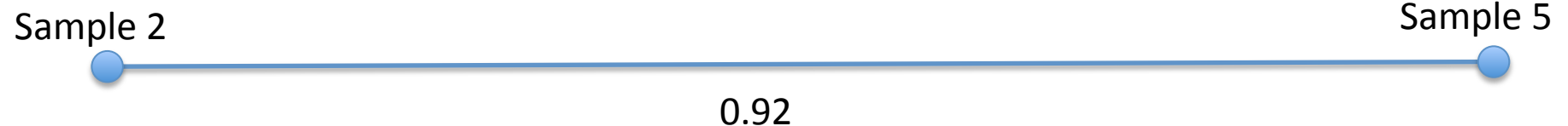
D2 is distance between sample and endpoint 2.

Step 1. Identify the largest distance in the distance matrix.

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

Step 2. Define a line with the two samples contributing to that distance as follows:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0



Step 3. Compute the location of each other sample on that axis as follows:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

$$a = \frac{D^2 + D1^2 - D2^2}{2 \times D}$$

Sample 1:

$$\frac{0.92^2 + 0.35^2 - 0.9^2}{2 \times 0.92} = 0.86$$

Sample 2



Sample 5



0.92

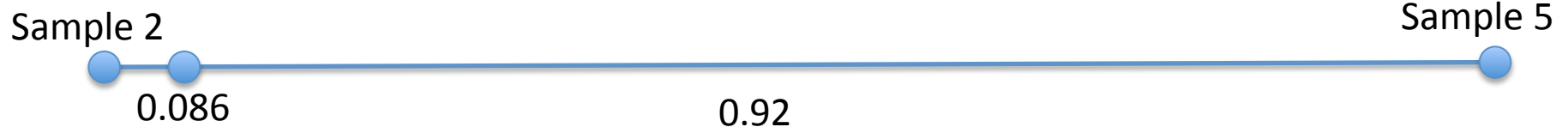
Step 3. Compute the location of each other sample on that axis as follows:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

$$a = \frac{D^2 + D1^2 - D2^2}{2 \times D}$$

Sample 1:

$$\frac{0.92^2 + 0.35^2 - 0.9^2}{2 \times 0.92} = 0.086$$



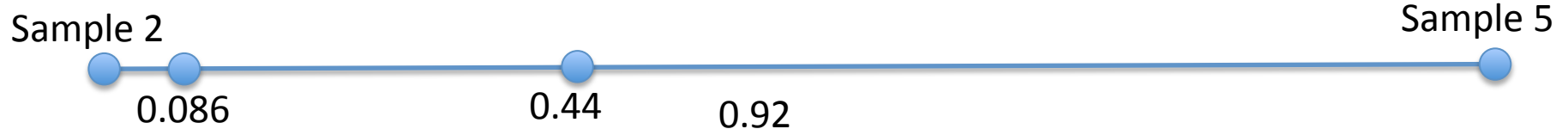
Step 3. Compute the location of each other sample on that axis as follows:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

$$a = \frac{D^2 + D1^2 - D2^2}{2 \times D}$$

Sample 2:

$$\frac{0.92^2 + 0.86^2 - 0.88^2}{2 \times 0.92} = 0.44$$

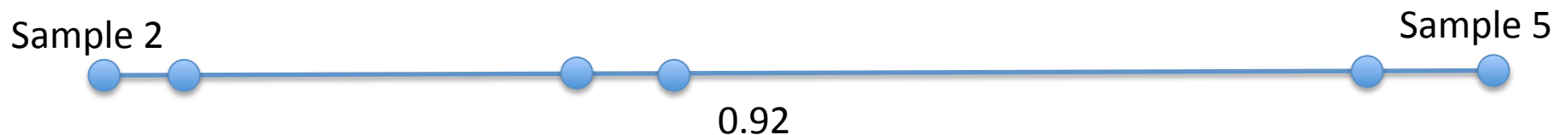


Step 3. Compute the location of each other sample on that axis as follows:

	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

$$a = \frac{D^2 + D1^2 - D2^2}{2 \times D}$$

Repeat for all samples



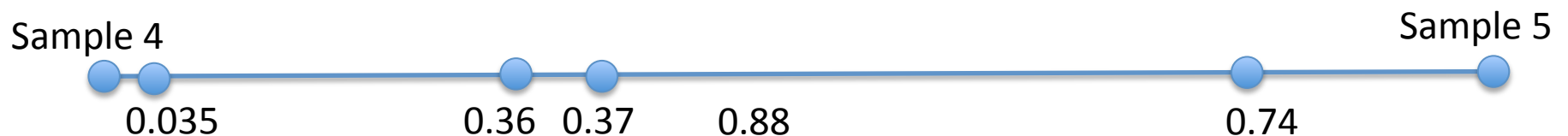
Step 4. Find the next largest distance that could be used to define an *uncorrelated axis*. This step can be labor-intensive to do by hand - usually you would compute all of the axes, along with correlation scores.

Step 5. Compute the location of each other sample on that axis as follows:

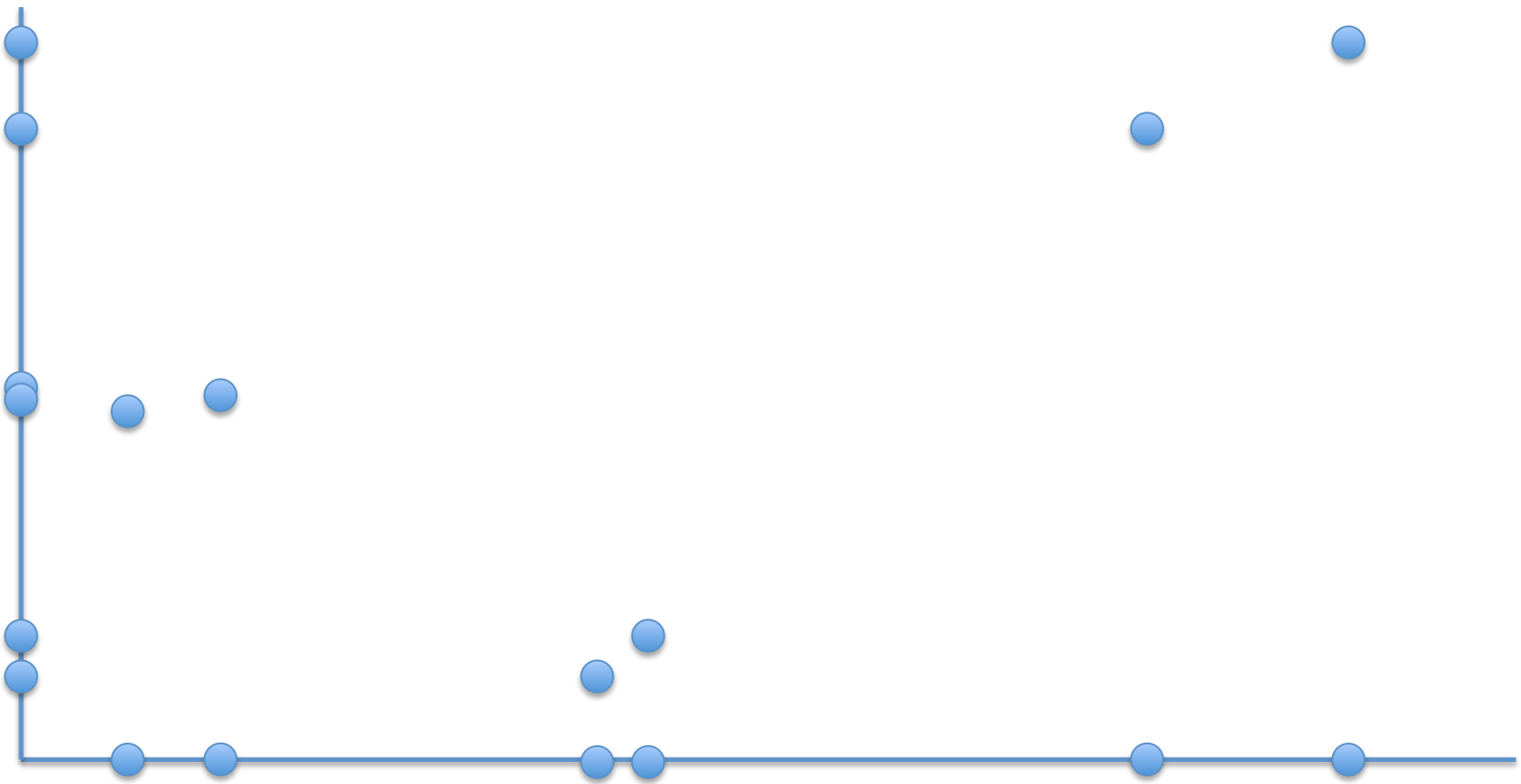
	Sample 1	Sample 2	Sample 3	Sample 4	Sample 5	Sample 6
Sample 1	0	0.35	0.83	0.83	0.9	0.9
Sample 2	0.35	0	0.86	0.85	0.92	0.91
Sample 3	0.83	0.7	0	0.25	0.88	0.87
Sample 4	0.83	0.1	0.4	0	0.88	0.88
Sample 5	0.9	0.92	0.88	0.88	0	0.5
Sample 6	0.9	0.91	0.87	0.88	0.5	0

$$a = \frac{D^2 + D1^2 - D2^2}{2 \times D}$$

Repeat for all samples



Step 6. Plot pairs of points:

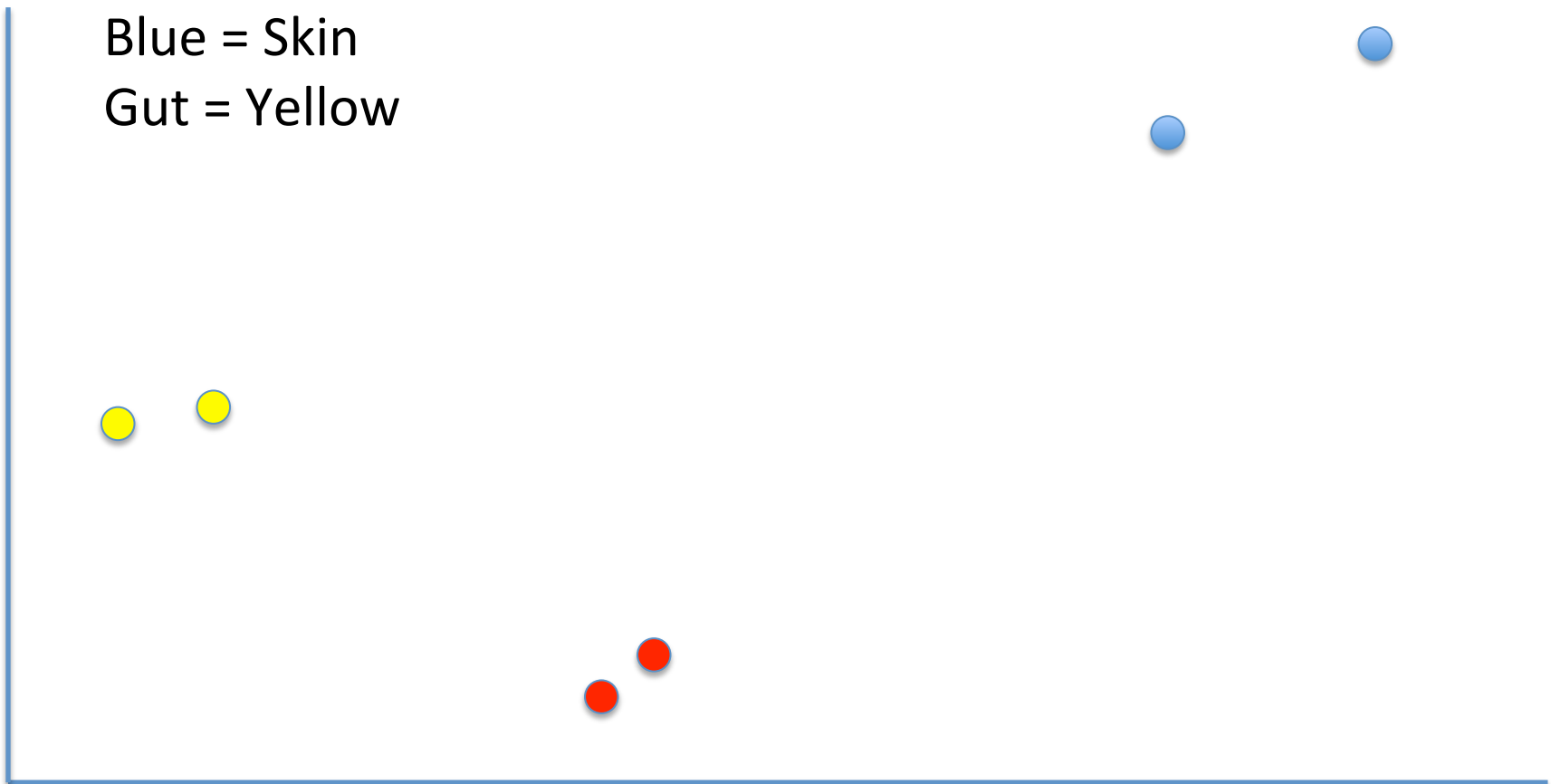


Step 7. Color by metadata category :

Red = Tongue

Blue = Skin

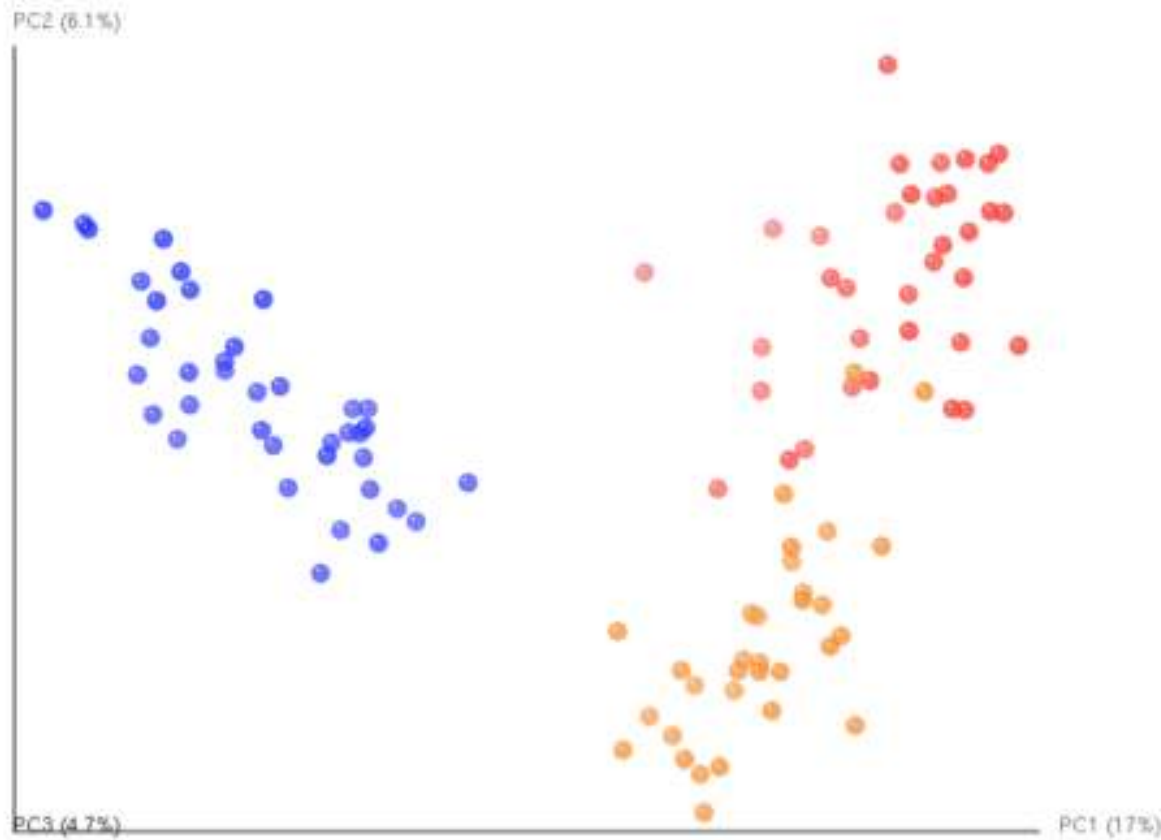
Gut = Yellow



- Anosim
 - Permutation
 - Compares within and between distances
 - R statistic -1 to 1. 0 indicates random grouping
 - P-values can be unreliable with a large sample size
 - Rank based
- Permanova
 - Works with any distances (as opposed to manova)
 - Permutations
- T-test
 - Not likely appropriate for a

Variation in sampling depth also needs
to be controlled for beta diversity!

Variation in sampling depth is an important consideration



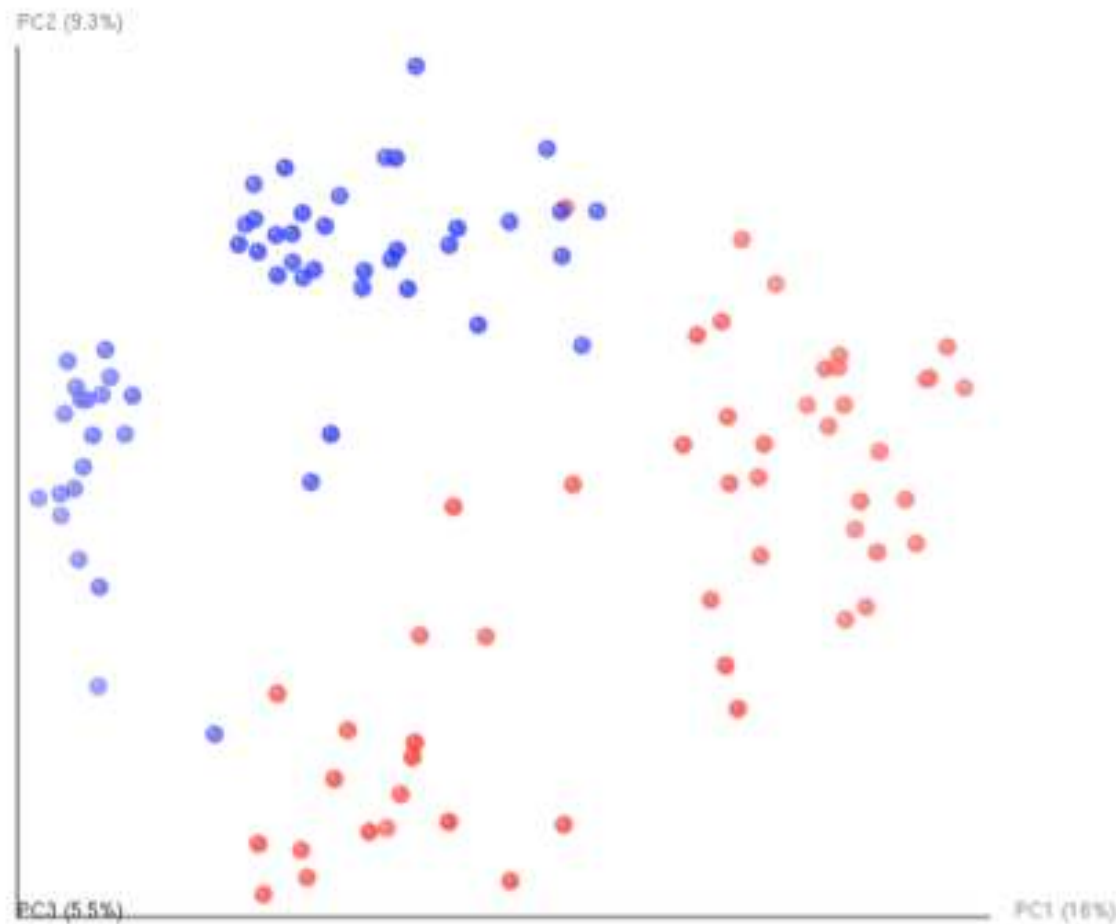
Human skin, colored
by individual, at 500
sequence/sample

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 sampling depth, at either 50 (blue) or 500 (red) sequences/sample

Image/analysis credit: Justin Kuczynski

Data reference:

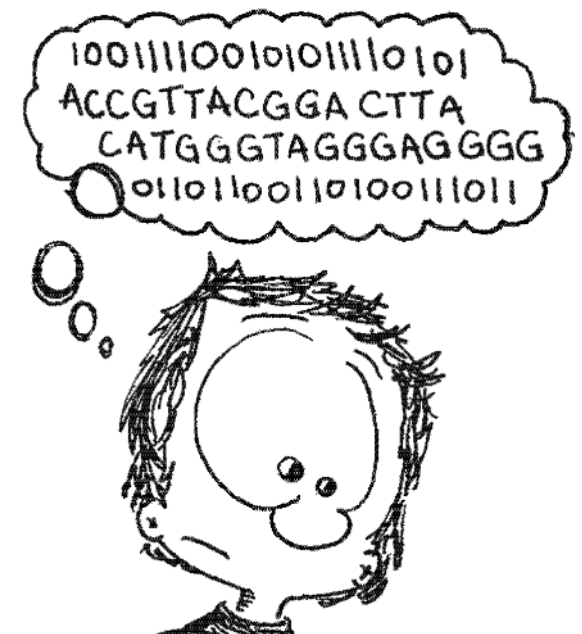
Forensic identification using skin bacterial communities. Fierer N, Lauber CL, Zhou N, McDonald D, Costello EK, Knight R. Proc Natl Acad Sci U S A. 2010 Apr 6;107(14):6477-81.

An Introduction To Applied Bioinformatics

An Introduction to **A**pplied **B**ioinformatics (or IAB) is a **free, open source interactive text** that introduces readers to core concepts of bioinformatics in the context of their implementation and application.

- To learn more about who should read IAB or how to read IAB, see [Reading IAB](#).
- For an overview of the content, see the [IAB Table of Contents](#).
- If you're interested in writing or editing content or code for IAB, see [CONTRIBUTING.md](#).

<http://readiab.org>

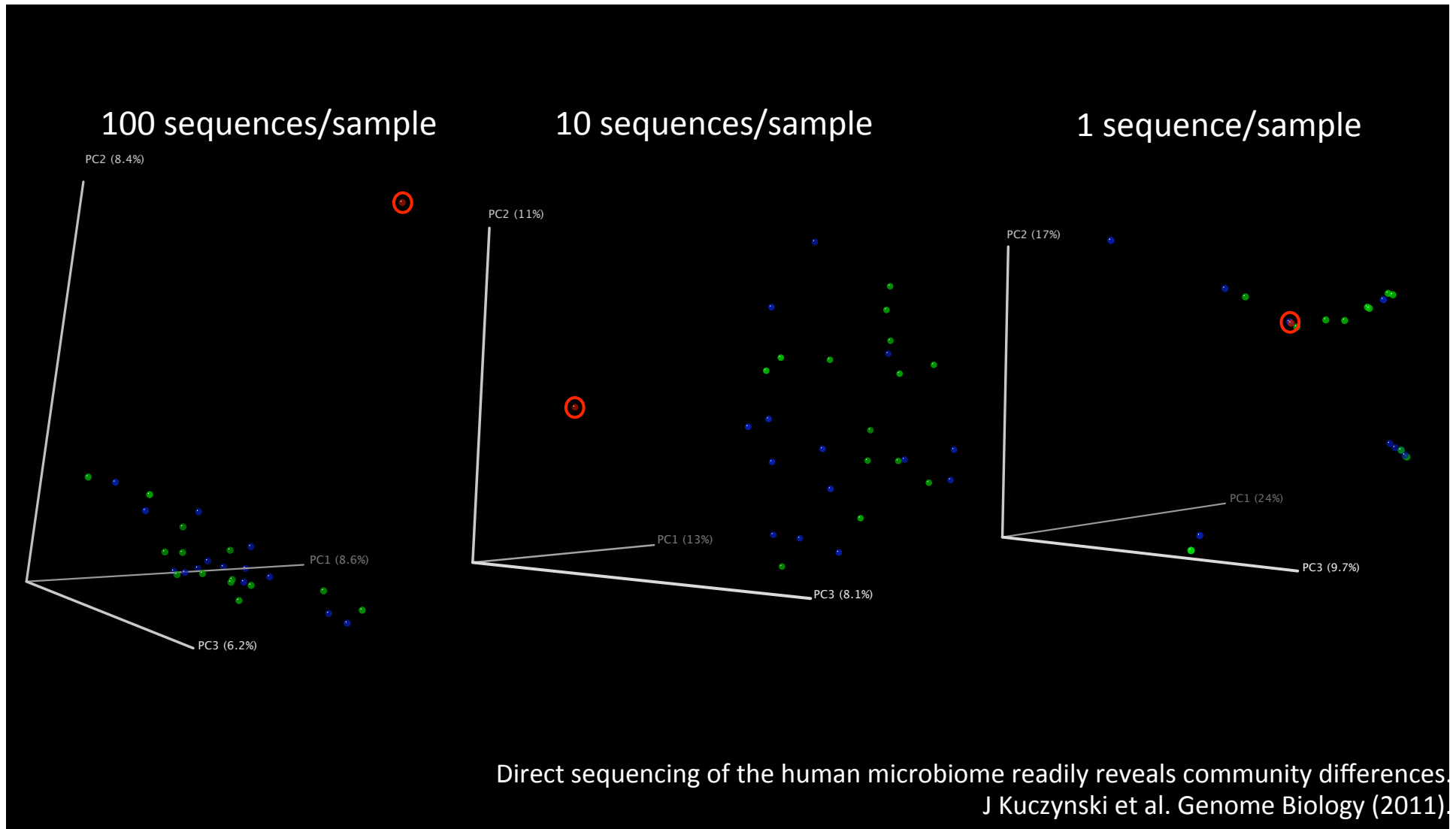


How deep is deep enough?

It depends on the question...

- Differences between community types: not many sequences.
- Rare biosphere: more (but be careful about sequencing noise!)

How deep is deep enough?



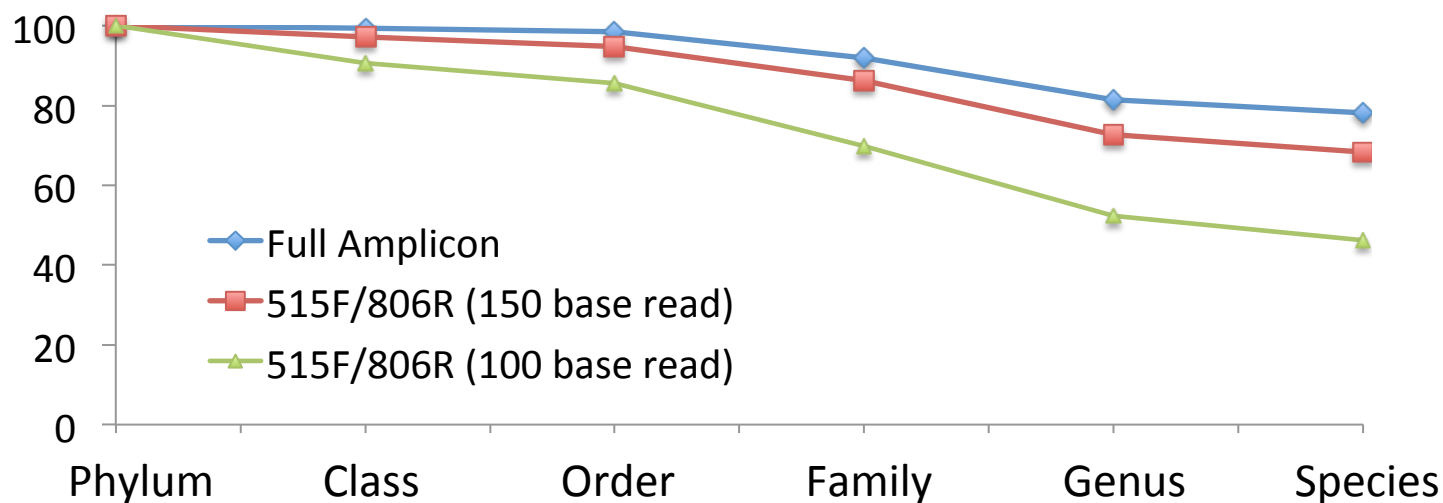
Genome Biol. 2010;11(5):210. Epub 2010 May 5.

Direct sequencing of the human microbiome readily reveals community differences.

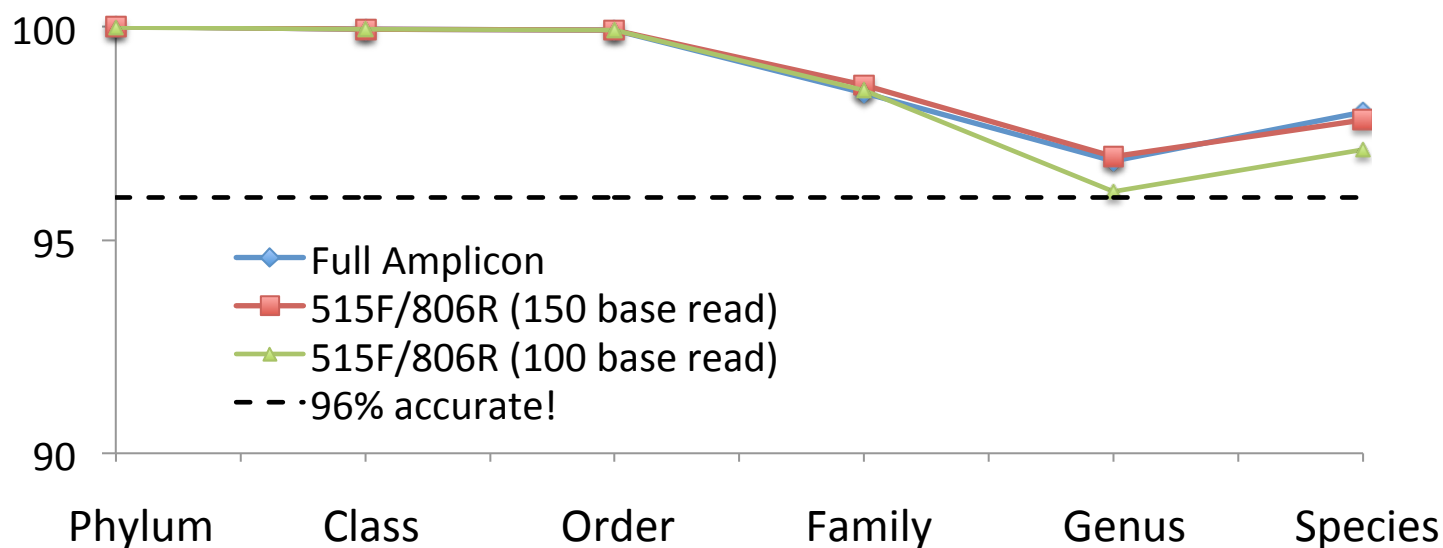
Kuczynski J, Costello EK, Nemergut DR, Zaneveld J, Lauber CL, Knights D, Koren O, Fierer N, Kelley ST, Ley RE, Gordon JI, Knight R.

Figure 1

Fraction of Greengenes *simulated reads* classified by taxonomic level using the RDP Classifier (80% confidence)

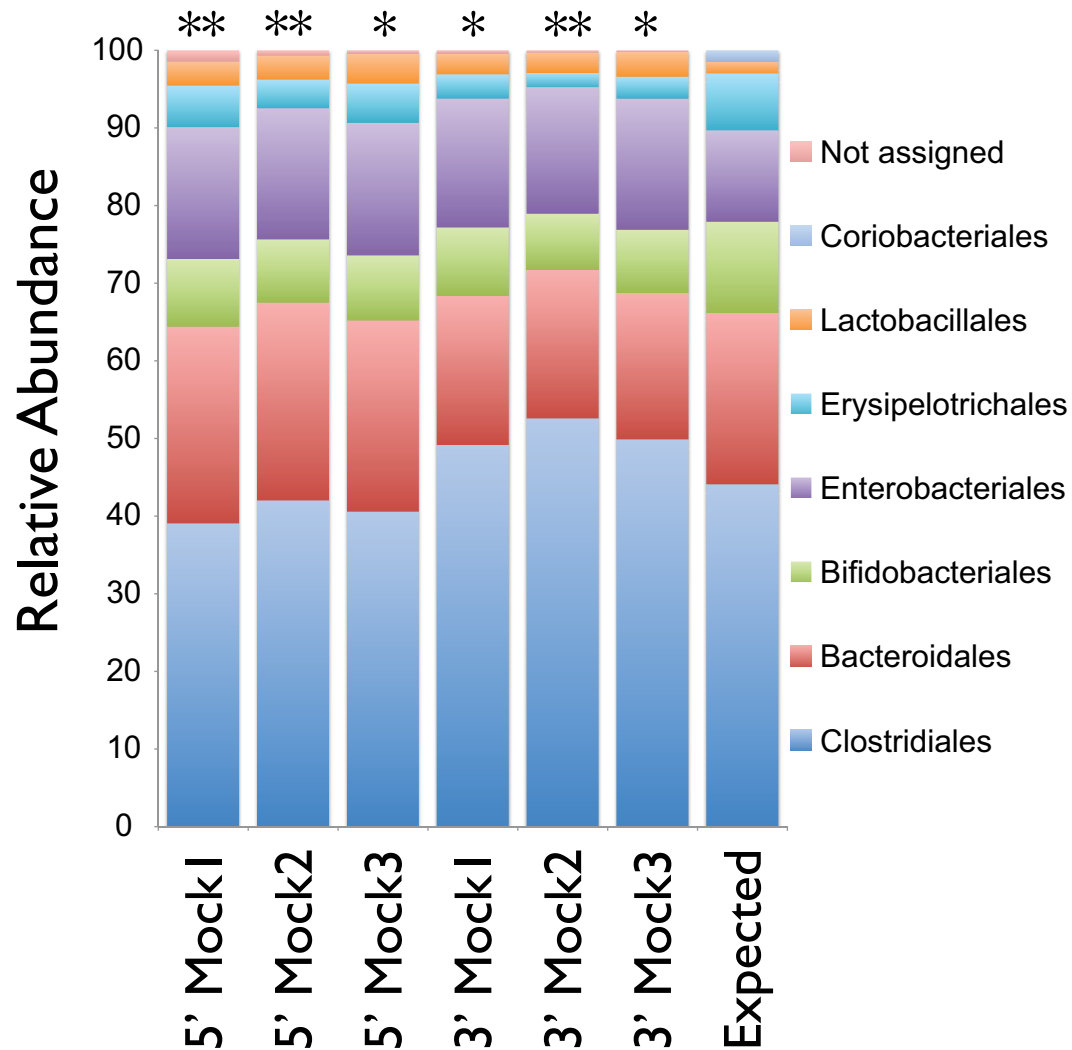


Accuracy of classified reads



Acknowledgement: Tony Walters

Can accurate taxonomy assignments be achieved?



Order-level taxonomy assignments

G-test (goodness of fit)

** $p < 0.01$

* $p < 0.05$

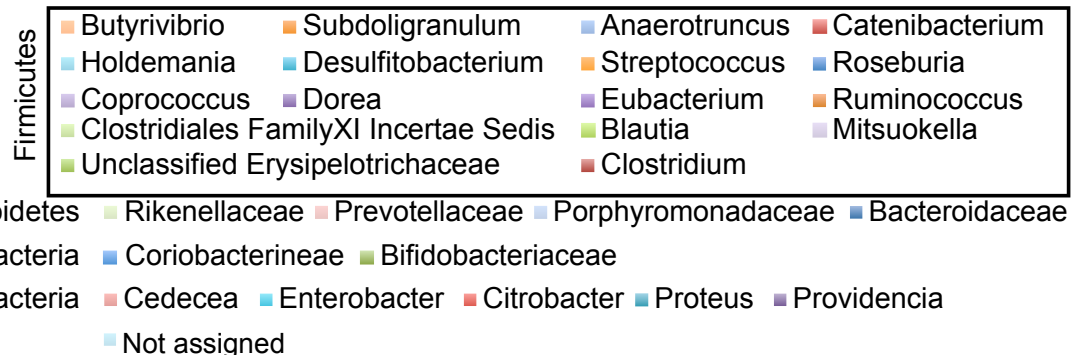
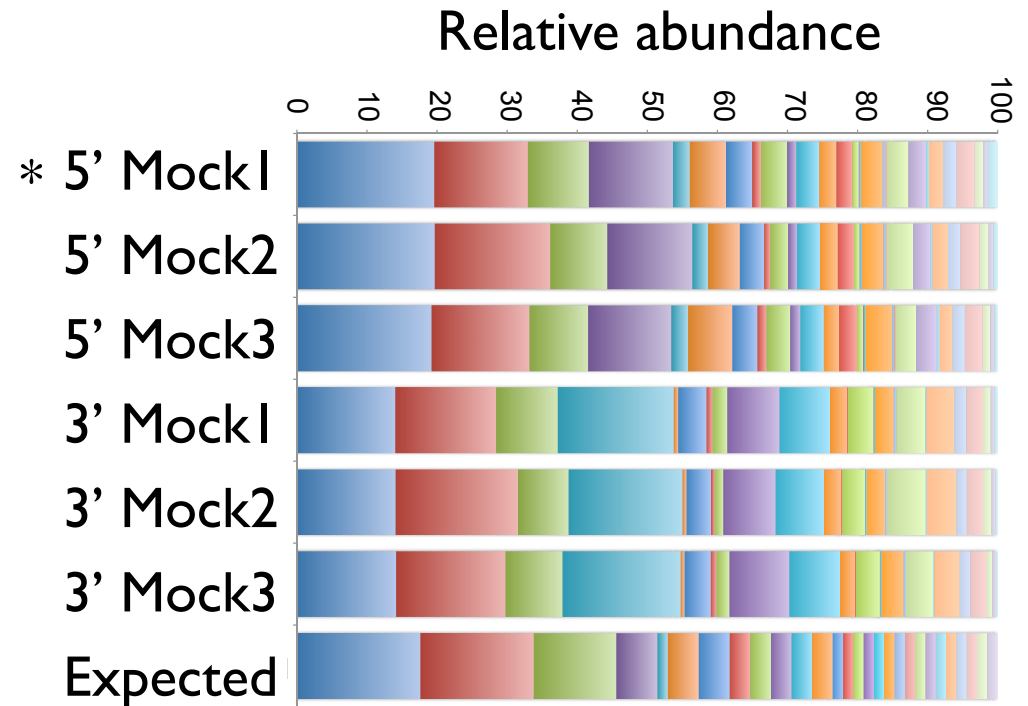
Can accurate taxonomy assignments be achieved?

Genus-level taxonomy assignments

G-test (goodness of fit)

** $p < 0.01$

* $p < 0.05$



Slides compiled by:

Greg Caporaso

John Chase

Jose Clemente

Antonio Gonzalez Peña

Rob Knight

Cathy Lozupone

Daniel McDonald

Jai Rideout

Yoshiki Vázquez Baeza



This work is licensed under the Creative Commons Attribution 3.0 United States License. To view a copy of this license, visit

<http://creativecommons.org/licenses/by/3.0/us/> or send a letter to Creative Commons, 171 Second Street, Suite 300, San Francisco, California, 94105, USA.

Feel free to use or modify these slides, but please credit us by placing the following attribution information where you feel that it makes sense:

Slides derived from QIIME educational materials www.qiime.org.