Microbiome Science

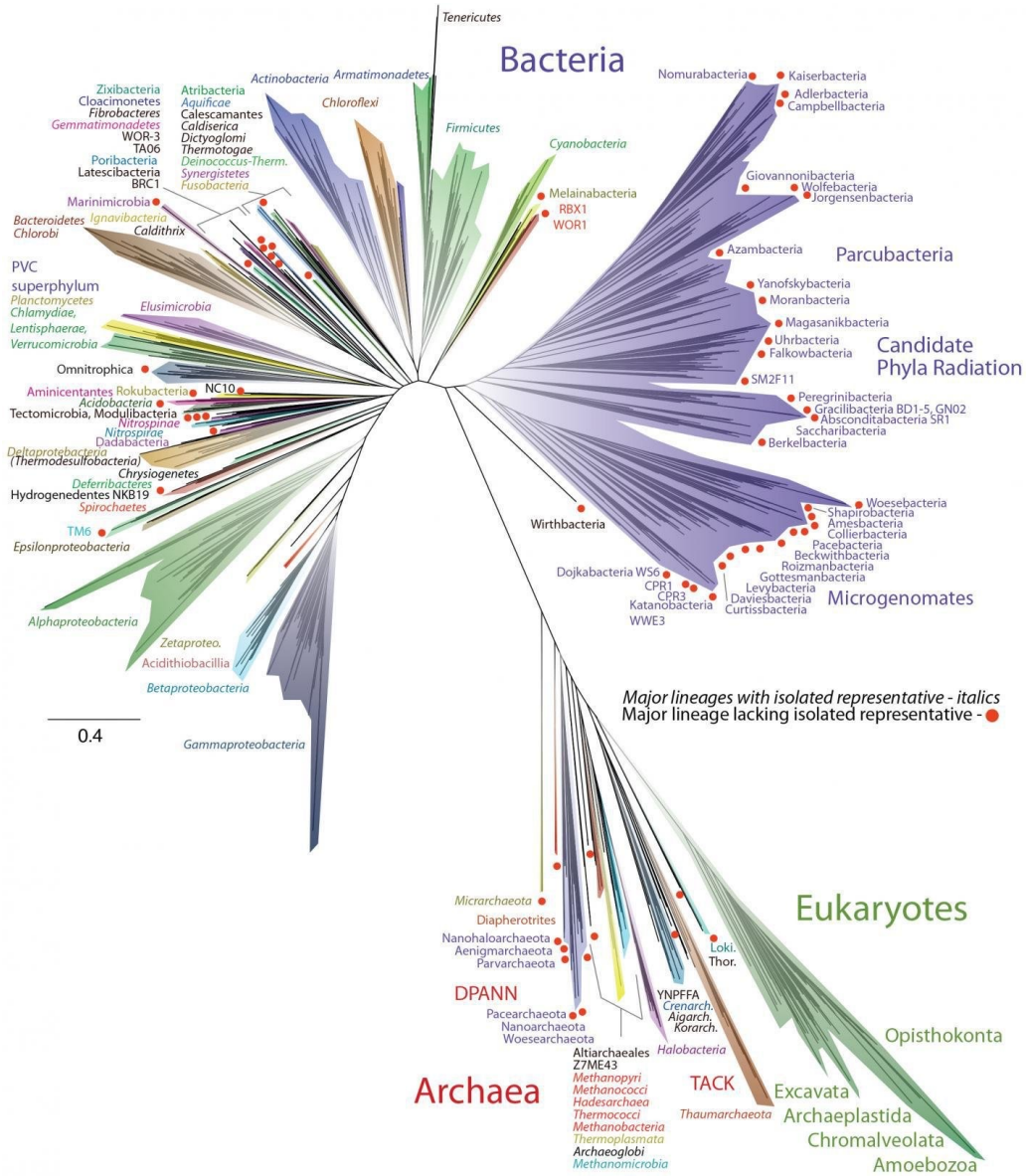
qCMB Introduction

Monday, March 20, 2023

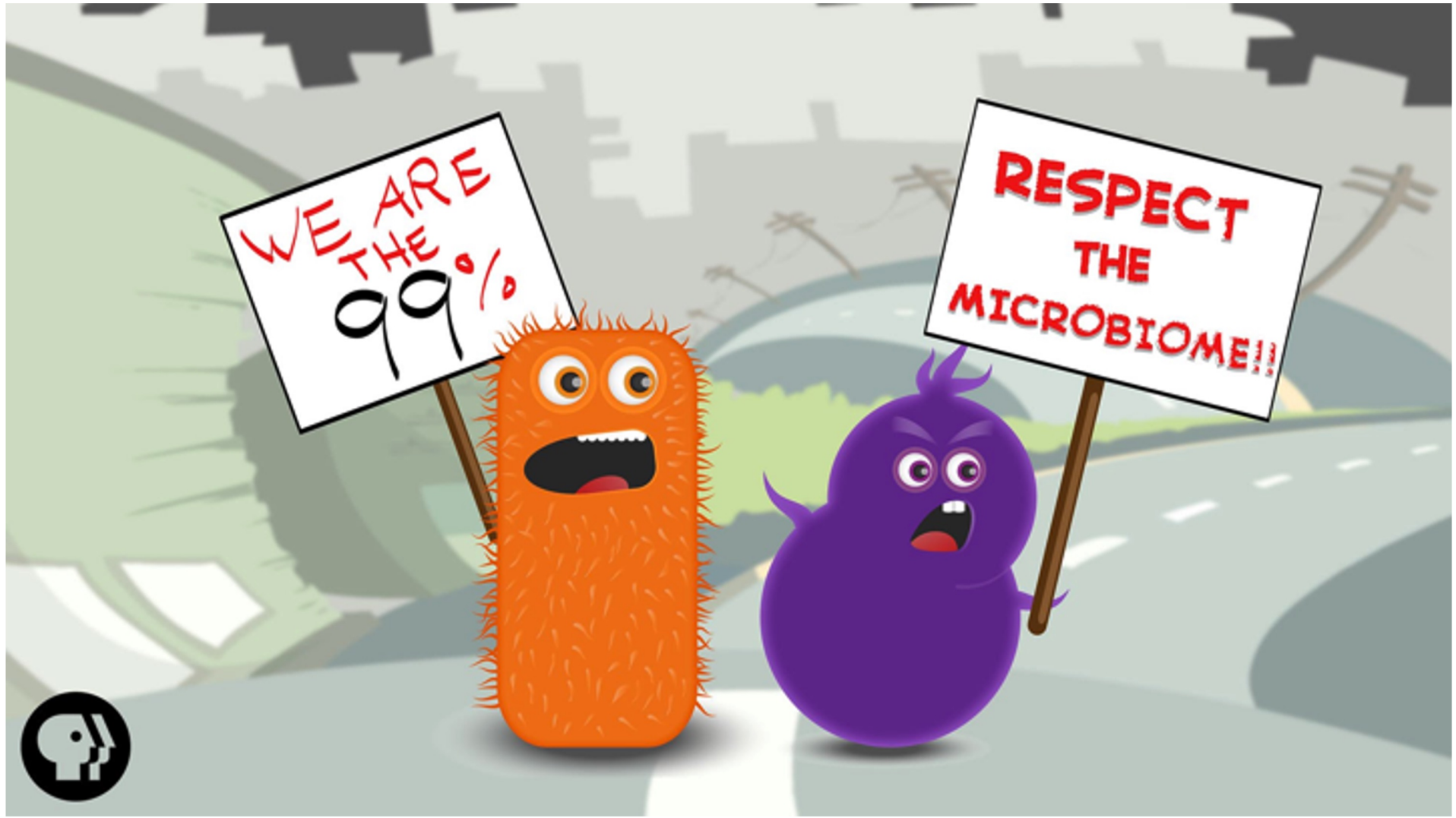
Microbiome Research

Microbiome – The genes and small molecules of microbes that are interacting in an environment.

We live in a microbial world



VIRUSES


WE ARE THE 99%
RESPECT THE MICROBIOME!!

MICROBIOME

Microbiology

Ecology

Genomics

Chemistry

We study …

structure,
diversity,
function,
communication

We sample ….

DNA,
RNA,
proteins,
metabolites

Metagenomics is the study of a collection of genetic material (genomes) from a mixed community of organisms.

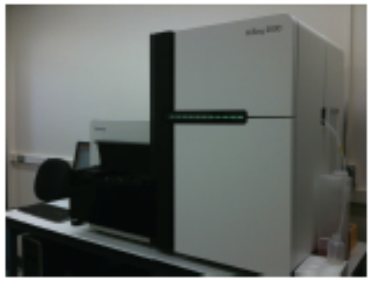
Also sometimes referred to as “shotgun metagenomics”

https://www.genome.gov/

What if you want to study a particular community like bacteria, fungi, or even macro plants, insects, etc?

Amplicon sequencing

Multiplex Thousands of Samples
with Error-Correcting Barcodes



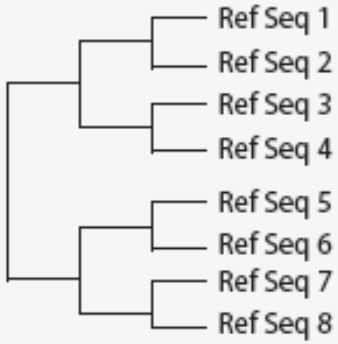
Pool Samples and Sequence

qiime
www.qiime.org

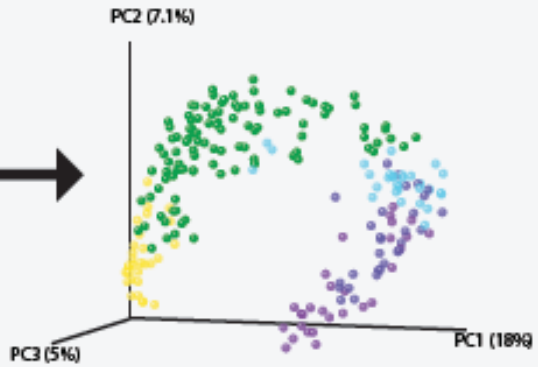
Process and Analyze Samples

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

Assign Sequences to Samples



Assign Millions of Sequences to
Clusters of Closely Related Organisms



Compare Samples Visually and Statistically

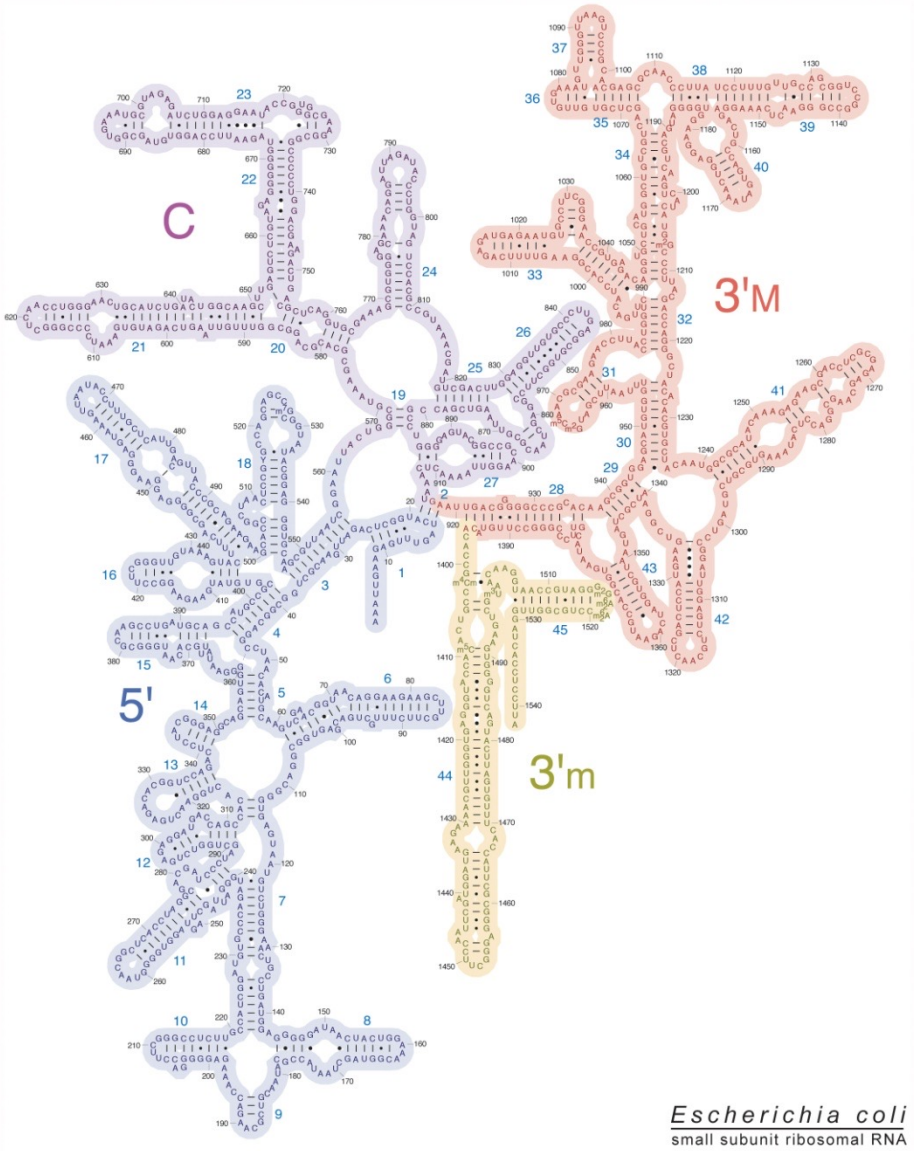
So what characteristics of a gene make it a good marker?

- Genes that are ubiquitous (e.g. important to the function of all living organisms)

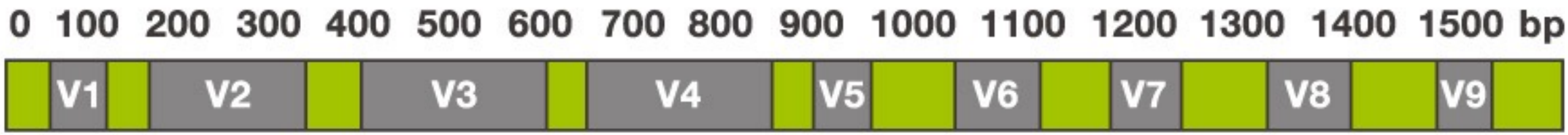
- Genes that contains both:
  - **Conserved region** – common between all microbes of interest e.g. a gene region present in all bacteria and archaea (so universal primers can find it)
  - **Variable region** – different between taxa contained within your microbial group of interest e.g. a region within a bacterial marker gene that differentiates *E. coli* or *P. aeruginosa*

16S ribosomal RNA

- 16S rRNA – part of the small subunit in prokaryotic ribosomes
  - Has multiple functions:
    - Binds to the Shine-Dalgarno sequence, a ribosomal binding site in bacterial and archaeal mRNA that is involved in recruiting the ribosome to initiate translation
    - Acts as a scaffold for ribosomal proteins
    - Helps to stabilize correct protein synthesis



16S rRNA amplicon sequencing



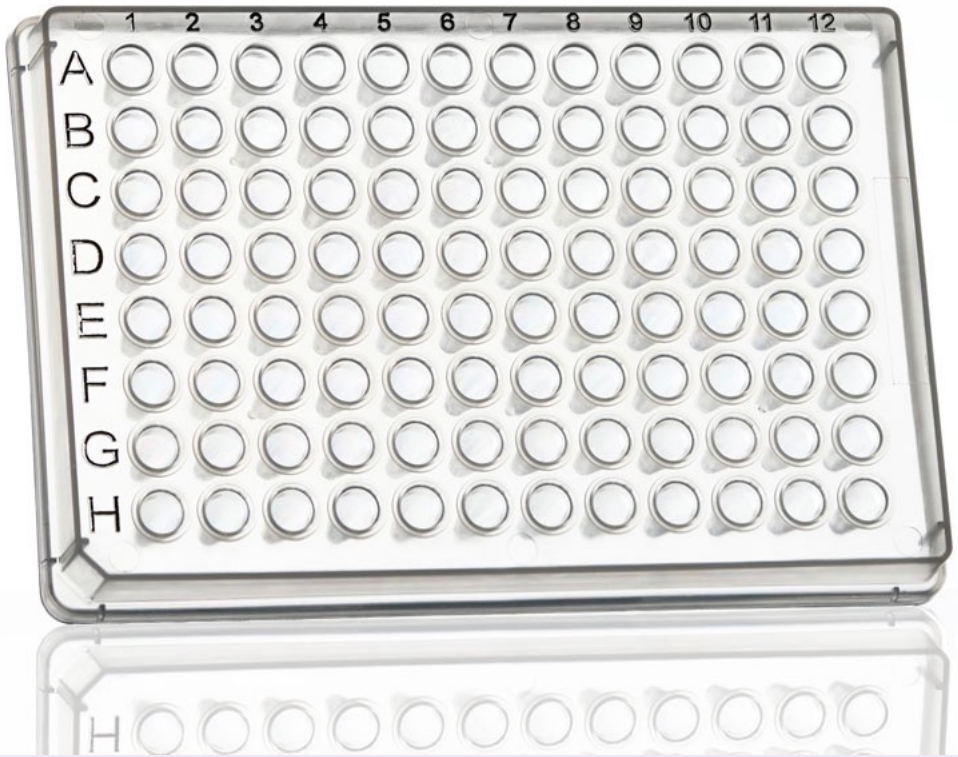
Sequencing primer	Barcode	Primer pad	Linker	515f
AATGATACGGCGACCACCGAGATCTACACGCT	XXXXXXXXXXXX	TATGGTAATT	GT	GTGYCAGCMGCCGCGGTAA

Plate Name(s)	Plate Number	Well Position		Sequence	Barcode
IL_515fBC_Jed_Arch_1	Plate 1	A1		AATGATACGGCGACCACCGAGATCTACACGCTAGCCTTCGTCGCTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	AGCCTTCGTCGC
IL_515fBC_Jed_Arch_1	Plate 1	A2		AATGATACGGCGACCACCGAGATCTACACGCTTCCATACCGGAATATGGTAATTGTGTGYCAGCMGCCGCGGTAA	TCCATACCGGAA
IL_515fBC_Jed_Arch_1	Plate 1	A3		AATGATACGGCGACCACCGAGATCTACACGCTAGCCCTGCTACATATGGTAATTGTGTGYCAGCMGCCGCGGTAA	AGCCCTGCTACA
IL_515fBC_Jed_Arch_1	Plate 1	A4		AATGATACGGCGACCACCGAGATCTACACGCTCCTAACGGTCCATATGGTAATTGTGTGYCAGCMGCCGCGGTAA	CCTAACGGTCCA
IL_515fBC_Jed_Arch_1	Plate 1	A5		AATGATACGGCGACCACCGAGATCTACACGCTCGCGCCTTAAACTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	CGCGCCTTAAAC
IL_515fBC_Jed_Arch_1	Plate 1	A6		AATGATACGGCGACCACCGAGATCTACACGCTTATGGTACCCAGTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	TATGGTACCCAG
IL_515fBC_Jed_Arch_1	Plate 1	A7		AATGATACGGCGACCACCGAGATCTACACGCTTACAATATCTGTTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	TACAATATCTGT
IL_515fBC_Jed_Arch_1	Plate 1	A8		AATGATACGGCGACCACCGAGATCTACACGCTAATTTAGGTAGGTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	AATTTAGGTAGG
IL_515fBC_Jed_Arch_1	Plate 1	A9		AATGATACGGCGACCACCGAGATCTACACGCTGACTCAACCAGTTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	GACTCAACCAGT
IL_515fBC_Jed_Arch_1	Plate 1	A10		AATGATACGGCGACCACCGAGATCTACACGCTGCCTCTACGTCGTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	GCCTCTACGTCG

Sequencing primer	Barcode	Primer pad	Linker	515f
AATGATACGGCGACCACCGAGATCTACACGCT	XXXXXXXXXXXX	TATGGTAATT	GT	GTGYCAGCMGCCGCGGTAA

Plate Name(s)	Plate Number	Well Position		Sequence	Barcode
IL_515fBC_Jed_Arch_1	Plate 1	A1		AATGATA	GCCGCGGTAA	AGCCTTCGTCGC
IL_515fBC_Jed_Arch_1	Plate 1	A2		AATGATA	GCCGCGGTAA	TCCATACCGGAA
IL_515fBC_Jed_Arch_1	Plate 1	A3		AATGATA	GCCGCGGTAA	AGCCCTGCTACA
IL_515fBC_Jed_Arch_1	Plate 1	A4		AATGATA	GCCGCGGTAA	CCTAACGGTCCA
IL_515fBC_Jed_Arch_1	Plate 1	A5		AATGATA	GCCGCGGTAA	CGCGCCTTAAAC
IL_515fBC_Jed_Arch_1	Plate 1	A6		AATGATA	GCCGCGGTAA	TATGGTACCCAG
IL_515fBC_Jed_Arch_1	Plate 1	A7		AATGATA	CCGCGGTAA	TACAATATCTGT
IL_515fBC_Jed_Arch_1	Plate 1	A8		AATGATA	GCCGCGGTAA	AATTTAGGTAGG
IL_515fBC_Jed_Arch_1	Plate 1	A9		AATGATACGGCGACCACCGAGATCTACACGCTGACTCAACCAGTTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	GACTCAACCAGT
IL_515fBC_Jed_Arch_1	Plate 1	A10		AATGATACGGCGACCACCGAGATCTACACGCTGCCTCTACGTCGTATGGTAATTGTGTGYCAGCMGCCGCGGTAA	GCCTCTACGTCG

96-well plate



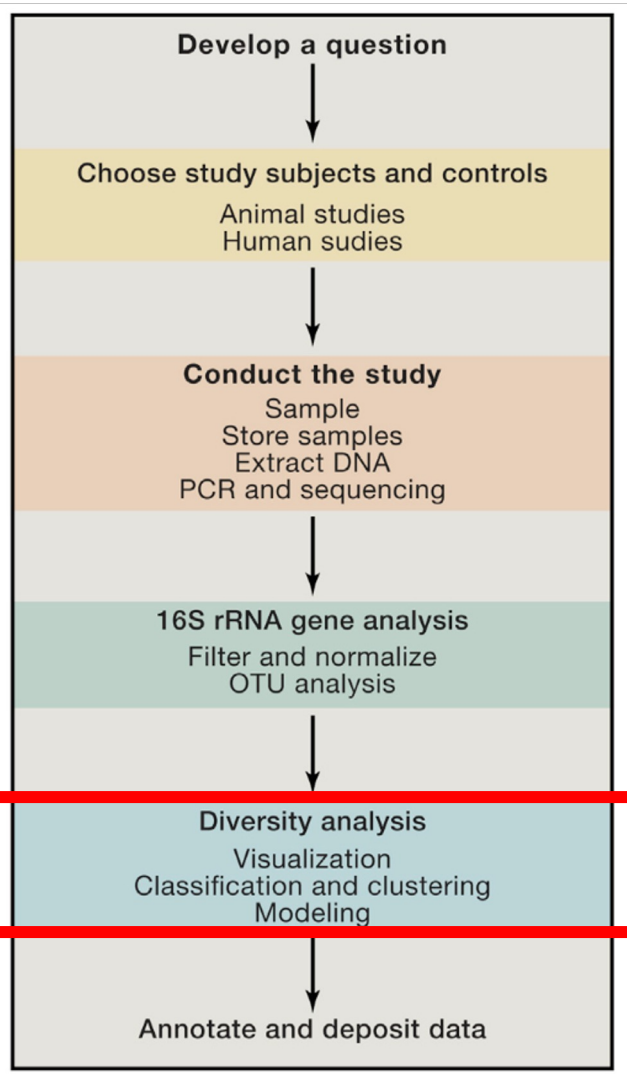
Performing a microbiome study

Study design
(power, randomization)
(metadata--standard + specific to your study)

Sample collection
(method, shipping/storage)
lab protocols
(include neg/pos controls)

Data processing

Data analysis



Sample Information

Sequence Data

Feature Table

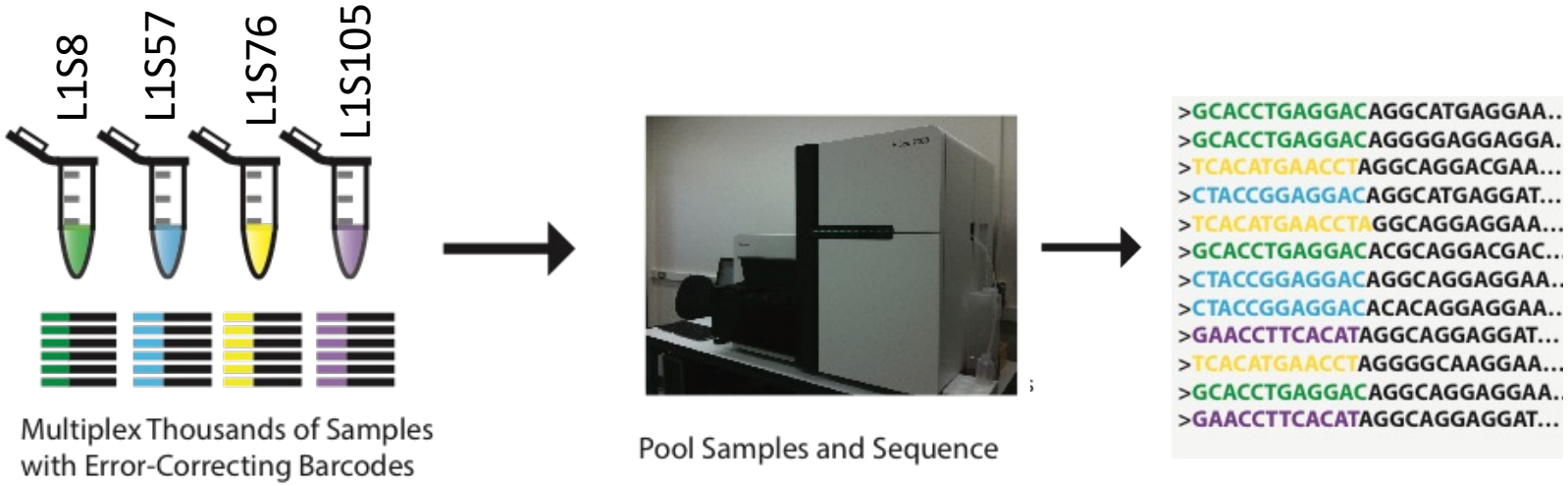
Summary Statistics

For more on these topics, see:
Conducting a Microbiome Study, by Goodrich et al. 2014
Reagent Contamination, Salter et al. 2014
Storage effects, by Song et al. 2016
Microbiome Quality Control (MBQC), by Sinha et al. 2017
MIMARKS, by Yilmaz et al. 2011
KatharoSeq low biomass workflow, by Minich et al. 2017

Other resources:
Earth Microbiome Project website
Human Microbiome Project website
American Gut Project website

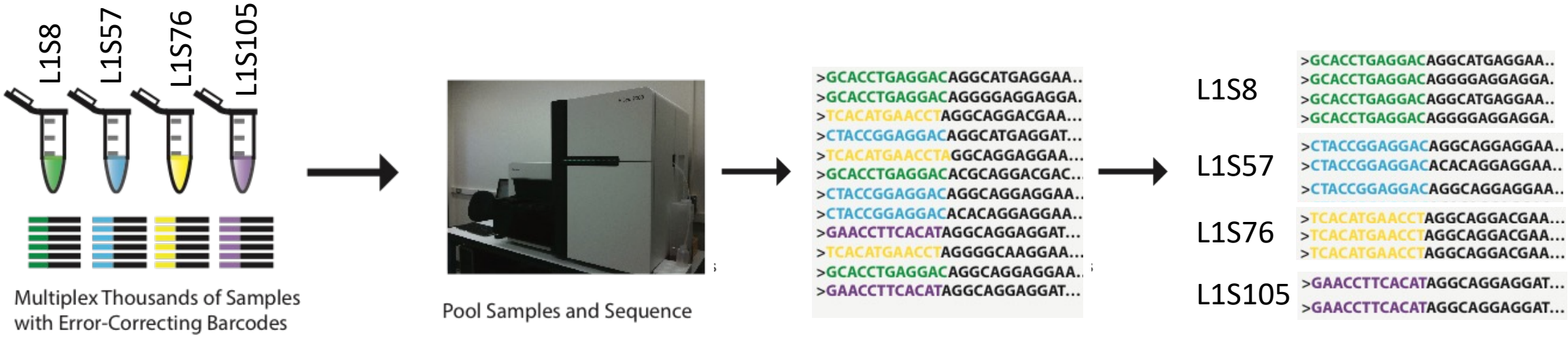
Demultiplexing

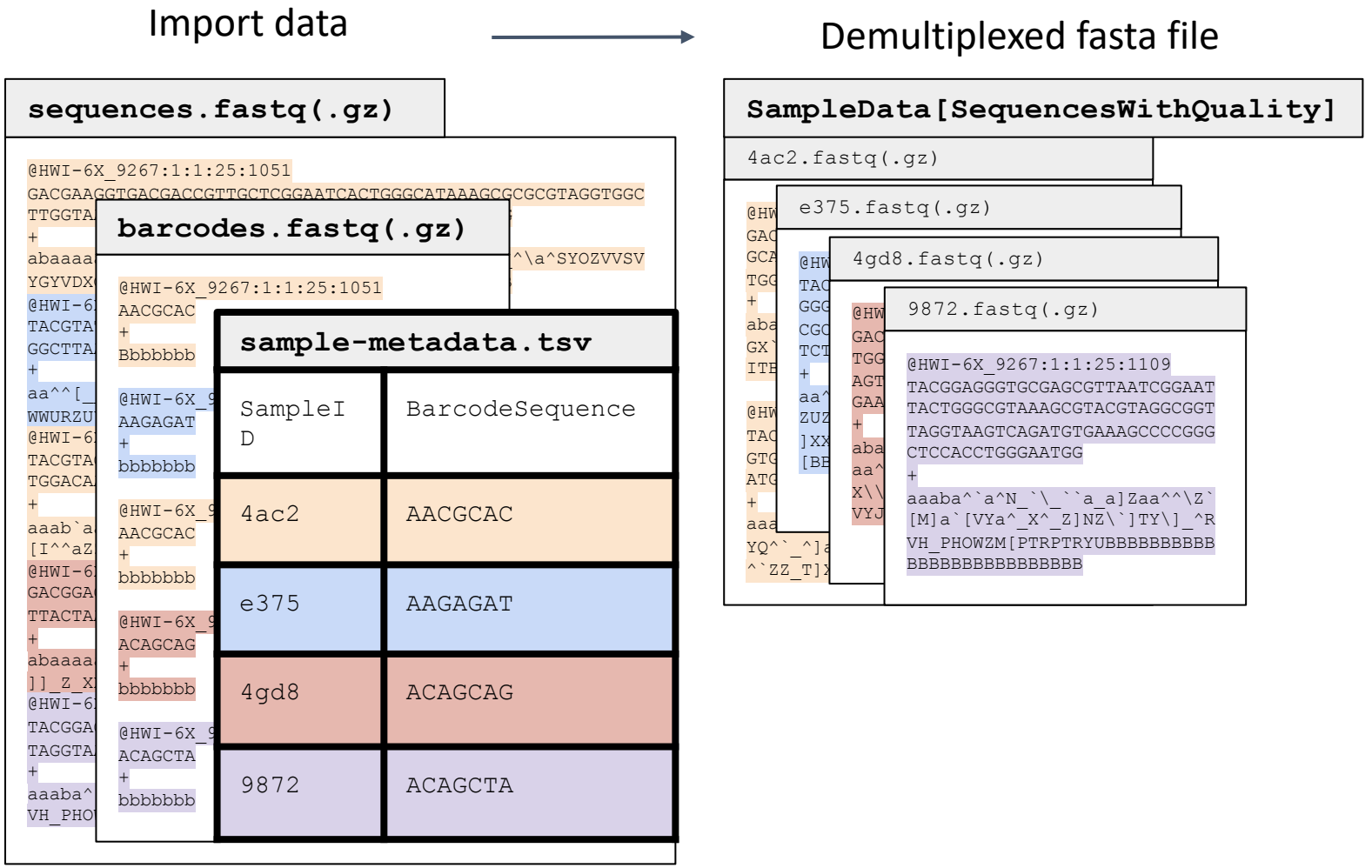
Bioinformatically undoing the multiplexing



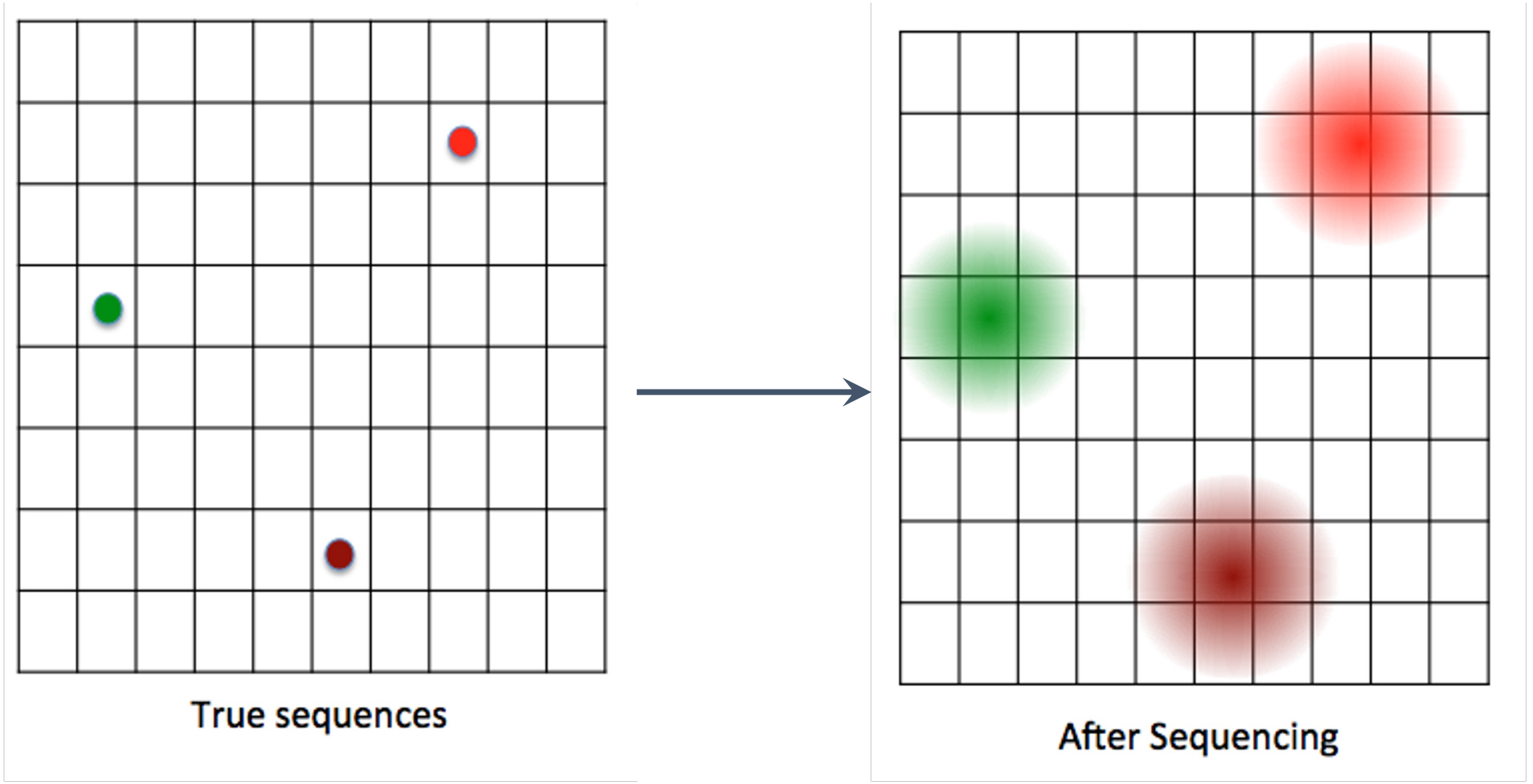
Demultiplexing

Bioinformatically undoing the multiplexing




Import data
Demultiplexed fasta file
sequences.fastq(.gz)
barcodes.fastq(.gz)
sample-metadata.tsv
SampleID
BarcodeSequence
4ac2
AACGCAC
e375
AAGAGAT
4gd8
ACAGCAG
9872
ACAGCTA
SampleData[SequencesWithQuality]
4ac2.fastq(.gz)
e375.fastq(.gz)
4gd8.fastq(.gz)
9872.fastq(.gz)
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGTACGTAGGCGGT
TAGGTAAGTCAGATGTGAAAGCCCCGGG
CTCCACCTGGGAATGG
+
aaaba^`a^N_`_``a_a]Zaa^^\Z`
[M]a`[VYa^_X^_Z]NZ\`]TY\]_^R
VH_PHOWZM[PTRPTRYUBBBBBBBBBB
BBBBBBBBBBBBBBBB

What normally happens during sequencing?



Feature table

SampleID	Feature #1	Feature #2	Feature #3	Feature #4	Feature #5
person1-time1	0	3,476	103	1,903	2,871
person1-time2	1,289	2,234	105	33	109
person2-time1	239	1,586	145	56	2,250
person2-time2	1,913	1,704	136	1,078	876

Import data

sequences.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGC
TTGGTA
+
abaaaa
YGYVDX
@HWI-6
TACGTA
GGCTTA
+
aa^^[_
WWURZU
@HWI-6
TACGTA
TGGACA
+
aaab`a
[I^^aZ
@HWI-6
GACGGA
TTACTA
+
abaaaa
]]_Z_X
@HWI-6
TACGGA
TAGGTA
+
aaaba^
VH_PHO
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
Bbbbbbb
@HWI-6X_9
AAGAGAT
+
bbbbbbb
@HWI-6X_9
AACGCAC
+
bbbbbbb
@HWI-6X_9
ACAGCAG
+
bbbbbbb
@HWI-6X_9
ACAGCTA
+
bbbbbbb
```

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

Demultiplexed fasta file

SampleData[SequencesWithQuality]

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGTACGTAGGCGGT
TAGGTAAGTCAGATGTGAAAGCCCCGGG
CTCCACCTGGGAATGG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`
[M]a`[VYa^_X^_Z]NZ\`]TY\]_^R
VH_PHOWZM[PTRPTRYUBBBBBBBBBB
BBBBBBBBBBBBBBBB
```

Feature table

FeatureTable[Frequency]

	feature 1	feature 2	feature 3	Feature 4	feature 5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

Representative sequences file

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGT
CCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG

>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGT
CTGATGTGAAAGGCTGGGGCTCAACCCCGGGACGG

>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGC
GCAGGGTTTAAGGCAATGGCTTAACTATTGTTCTC

>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGT
CAACTGTTAAATCTTGAGGCTCAACCTCGAAATCG

>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGT
CAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG
```

Taxonomic assignment of observed sequences

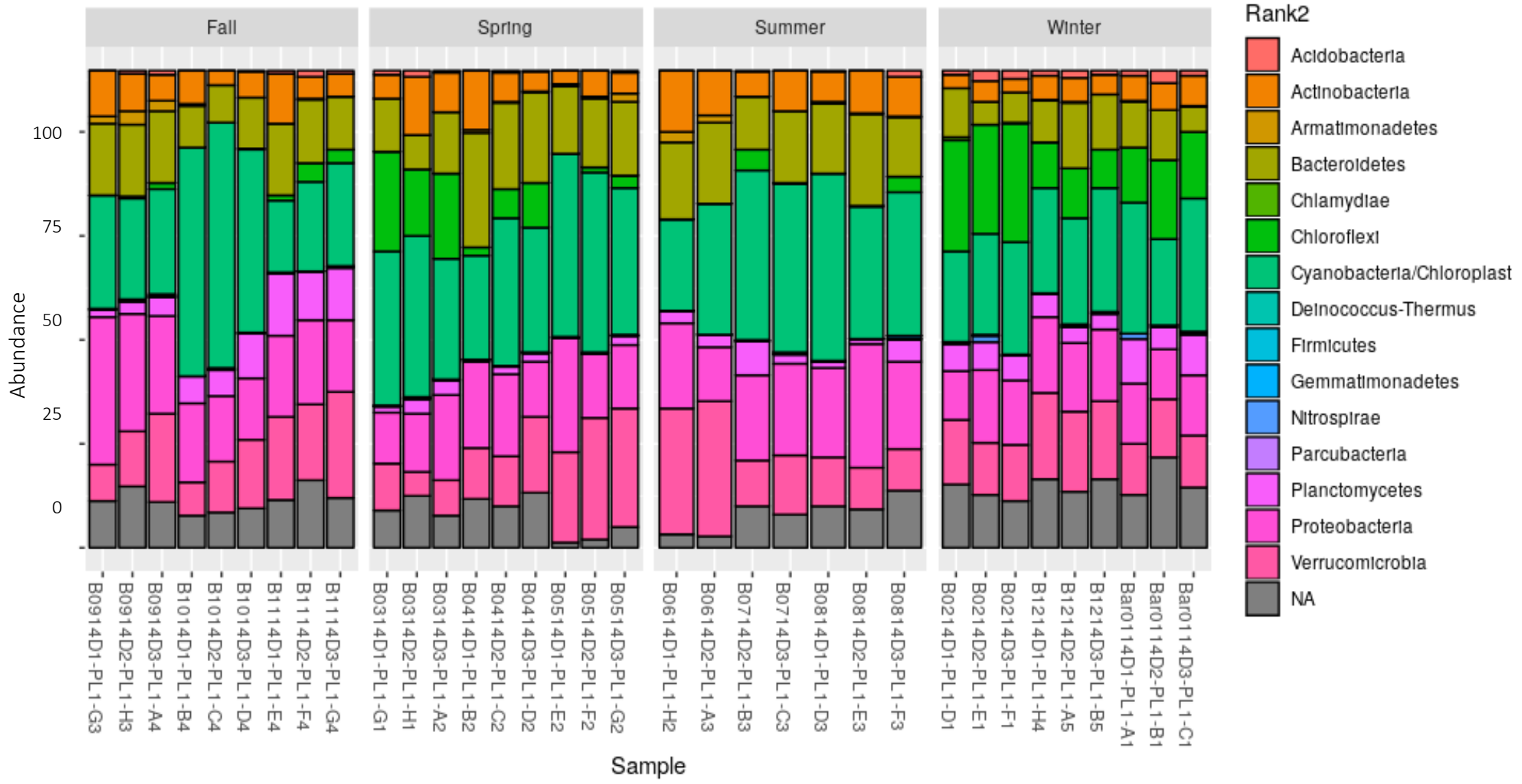
Assignment methods

- Compare directly to a reference database
  - VSEARCH, BLAST
- Predict using a machine learning classifier
  - RDP, QIIME2

Reference databases

- Greengenes (16S rRNA)
- SILVA (16S rRNA, 18S rRNA)
- UNITE (ITS- fungal)

Once you predict your taxonomy, you can visualize relative abundance of taxa in each sample with a bar plot



Using taxonomy for quality control of your data

Animal host-associated microbiomes

Mitochondrial 16S sequences:

k__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__mitochondria

Plant-associated microbiomes

Chloroplast 16S sequences:

k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;f__;g__

k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Chlorophyta;f__Trebouxiophyceae;g__

k__Bacteria;p__Cyanobacteria;c__Chloroplast;o__Streptophyta;f__;g__

Comparing microbial communities

How many different OTUs/ASVs are there? *Alpha diversity*

How similar are pairs of samples? *Beta diversity*

Who is there? *Taxonomic profiling, differential abundance testing*

What features of microbiomes differ?

Community richness (often referred to as alpha diversity)

Image source: http://miriadna.com/desctopwalls/images/max/Field-of-yellow-tulips.jpg

Image source: https://imgflip.com/memetemplate/62338435/Flower-garden

What features of microbiomes differ?

Community composition (often referred to as beta diversity)

Image source:
https://shawnacoronado.com/front-lawn-vegetable-garden-design/

Image source:
https://imgflip.com/memetemplate/62338435/Flower-garden

Non-phylogenetic diversity metrics assume that all taxa are equally related, so doesn't make assumptions about evolutionary relationships.

Phylogenetic diversity metrics incorporate evolutionary relationships between taxa, but assume that we know what those relationships are.

Import data

sequences.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGC
TTGGTA
+
abaaaa
YGYVDX
@HWI-6
TACGTA
GGCTTA
+
aa^^[_
WWURZU
@HWI-6
TACGTA
TGGACA
+
aaab`a
[I^^aZ
@HWI-6
GACGGA
TTACTA
+
abaaaa
]]_Z_X
@HWI-6
TACGGA
TAGGTA
+
aaaba^
VH_PHO
```

barcodes.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1051
AACGCAC
+
Bbbbbbb
@HWI-6X_9
AAGAGAT
+
bbbbbbb
@HWI-6X_9
AACGCAC
+
bbbbbbb
@HWI-6X_9
ACAGCAG
+
bbbbbbb
@HWI-6X_9
ACAGCTA
+
bbbbbbb
```

sample-metadata.tsv

SampleID	BarcodeSequence
4ac2	AACGCAC
e375	AAGAGAT
4gd8	ACAGCAG
9872	ACAGCTA

Demultiplexed fasta file

SampleData[SequencesWithQuality]

4ac2.fastq(.gz)

e375.fastq(.gz)

4gd8.fastq(.gz)

9872.fastq(.gz)

```
@HWI-6X_9267:1:1:25:1109
TACGGAGGGTGCGAGCGTTAATCGGAAT
TACTGGGCGTAAAGCGTACGTAGGCGGT
TAGGTAAGTCAGATGTGAAAGCCCCGGG
CTCCACCTGGGAATGG
+
aaaba^`a^N_`\_``a_a]Zaa^^\Z`
[M]a`[VYa^_X^_Z]NZ\`]TY\]_^R
VH_PHOWZM[PTRPTRYUBBBBBBBBBB
BBBBBBBBBBBBBBBB
```

Feature table

FeatureTable[Frequency]

	feature 1	feature 2	feature 3	Feature 4	feature 5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

rooted-tree.qza

Phylogeny[Rooted]

Representative sequences file

FeatureData[Sequence]

```
>feature5
GACGAAGGTGACGACCGTTGCTCGGAATCACTGGGCATAAAGCGCGCGTAGGTGGCTTGGTAAGT
CCATGGTGAAATCCCTCGGCTCAACCGAGGAACTG

>feature4
TACGTAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGATGGACAAGT
CTGATGTGAAAGGCTGGGGCTCAACCCCGGGACGG

>feature2
TACGTATGGGGCAAGCGTTATCCGGAATTATTGGGCGTAAAGAGTGCGTAGGTGGTGGCTTAAGC
GCAGGGTTTAAGGCAATGGCTTAACTATTGTTCTC

>feature1
GACGGAGGATGCAAGTGTTATCCGGAATCACTGGGCGTAAAGCGTCTGTAGGTGGTTTACTAAGT
CAACTGTTAAATCTTGAGGCTCAACCTCGAAATCG

>feature3
TACGGAGGGTGCGAGCGTTAATCGGAATTACTGGGCGTAAAGCGTACGTAGGCGGTTAGGTAAGT
CAGATGTGAAAGCCCCGGGCTCCACCTGGGAATGG
```

There are a lot of alpha diversity metrics

Observed ASVs – richness (# of ASVs)

Shannon Diversity Index – richness and evenness

Faith’s Phylogenetic diversity – sum of branch lengths

Observed OTUs (or Observed Species): non-phylogenetic, alpha diversity metric measuring richness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Observed OTUs
4ac2	
e375	

Observed OTUs (or Observed Species): non-phylogenetic, alpha diversity metric measuring richness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Observed OTUs
4ac2	3
e375	3

Shannon Diversity Index:
non-phylogenetic, alpha diversity metric measuring richness and evenness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	1.061
e375	1.064

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

Shannon Diversity Index:
non-phylogenetic, alpha diversity metric measuring richness and evenness

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	1.061
e375	1.064

$$H' = -\sum_{i=1}^{R} p_i \ln p_i$$

4ac2

Feature 1: 25/70=0.357

Feature 2: 30/70= 0.429

Feature 3: 15/70=0.214

= - [0.36(-1.03) + 0.43(-0.85)+ 0.21(-1.54)]

= -(-0.371 + -0.367 + -0.323) = 1.061

Why incorporate phylogeny in a diversity metric?

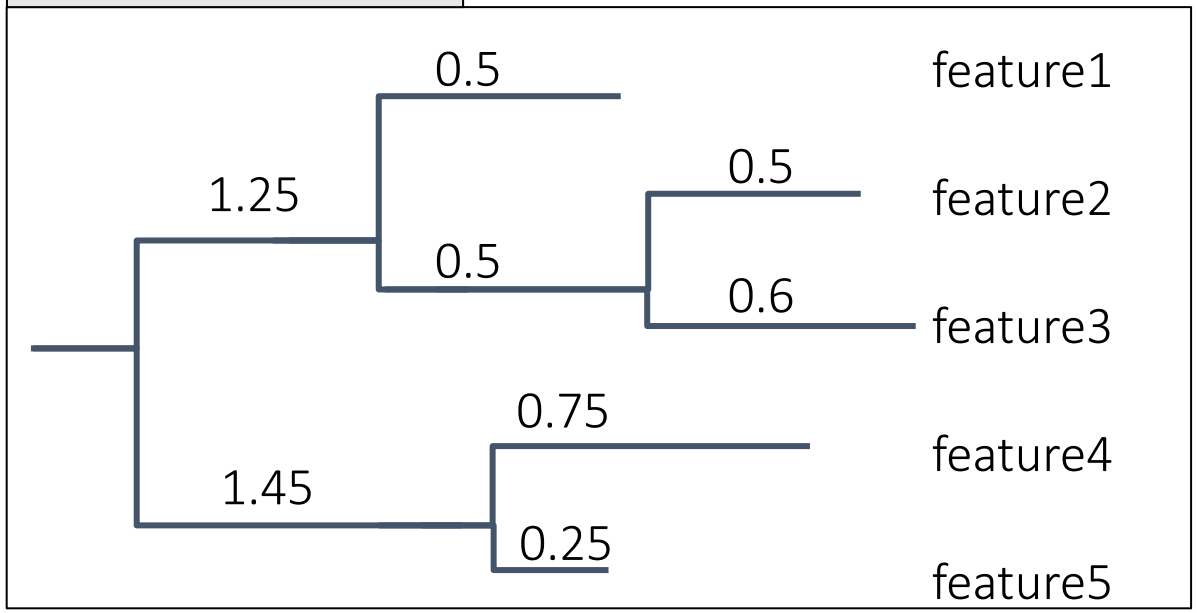
FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Shannon
4ac2	1.061
e375	1.064

Phylogeny[Rooted]



FeatureData[Taxonomy]

	Domain
feature1	Bacteria
feature2	Bacteria
feature3	Bacteria
feature4	Archaea
feature5	Archaea

Faith's Phylogenetic Diversity (PD): phylogenetic, alpha diversity metric measuring richness

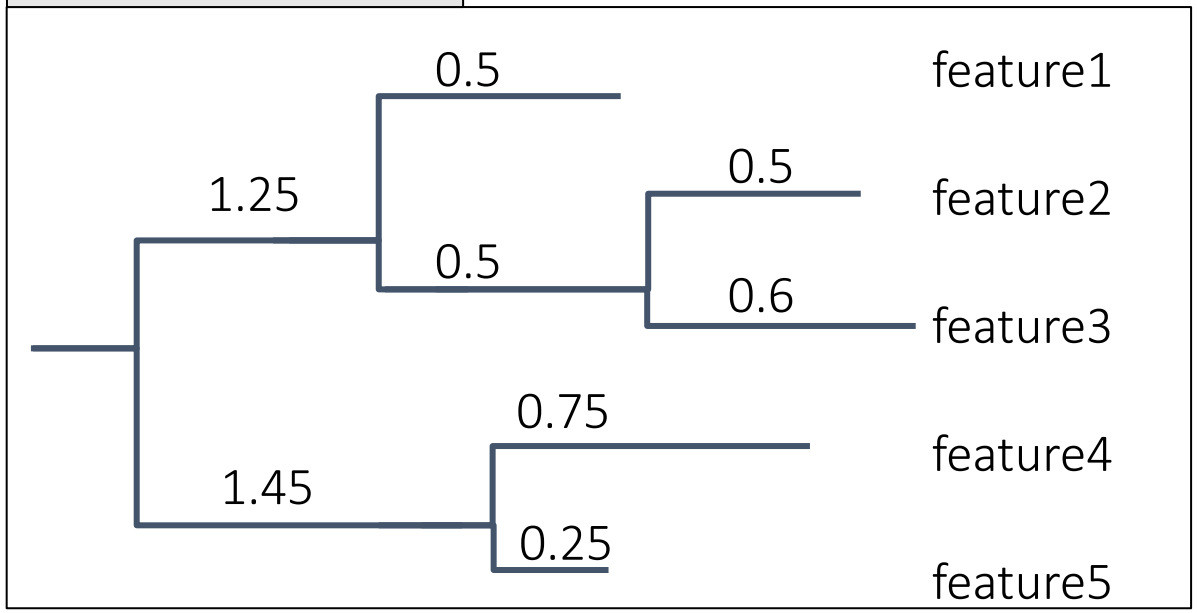
FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Faith's PD
4ac2	
e375	

Phylogeny[Rooted]



Sum of branch length covered by a sample.

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

Faith's Phylogenetic Diversity (PD): phylogenetic, alpha diversity metric measuring richness

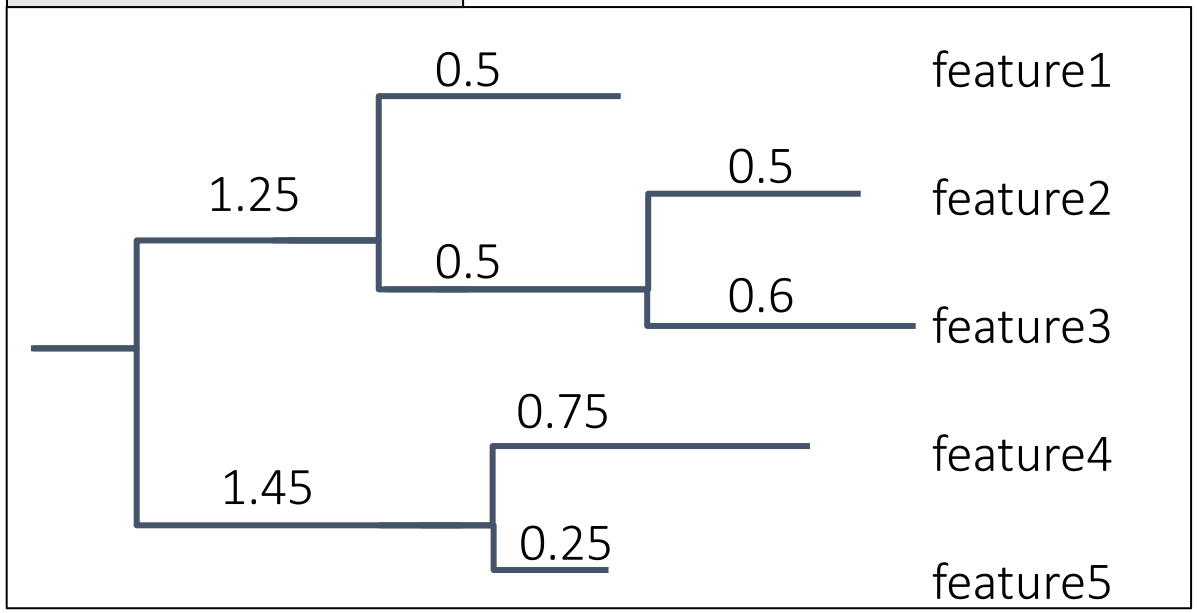
FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	25	30	15	0	0
e375	0	17	33	25	0

SampleData[AlphaDiversity]

	Faith's PD
4ac2	3.35
e375	5.05

Phylogeny[Rooted]



Sum of branch length covered by a sample.

4ac2 =1.25+0.5+0.5+0.5+0.6 = 3.35

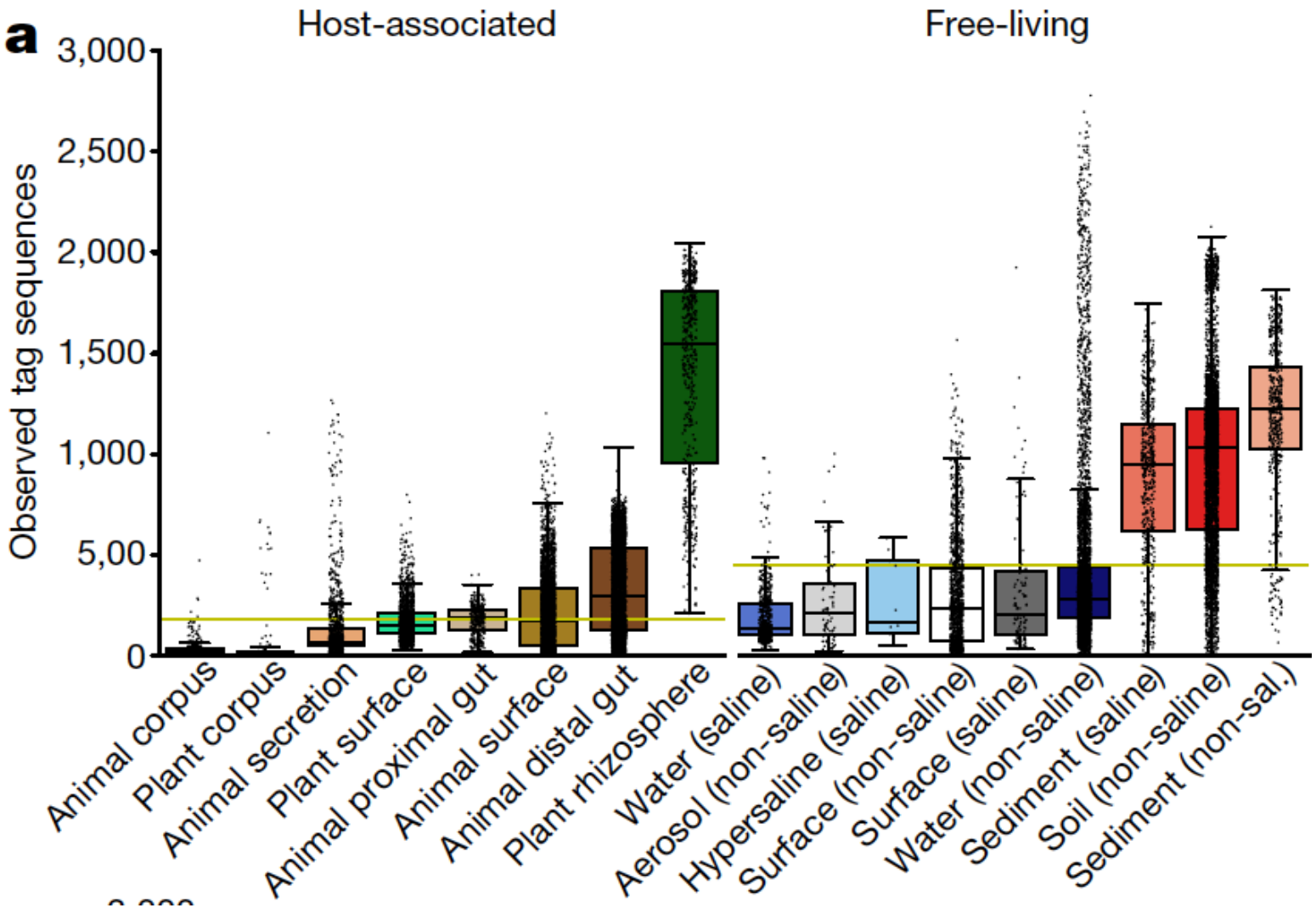
Faith DP (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation. 61:1-10.

Alpha diversity comparison

- visually with distribution comparison plots (discrete data) or scatter plots (continuous data)

- statistically with Kruskal-Wallis (discrete data) or Spearman correlation (continuous data)

What are the most diverse environments?



Thompson *et al.* 2017, *Nature*

End Day 1

Microbiome Science

amplicon sequencing analysis

qCMB Introduction

Wednesday, April 27, 2022

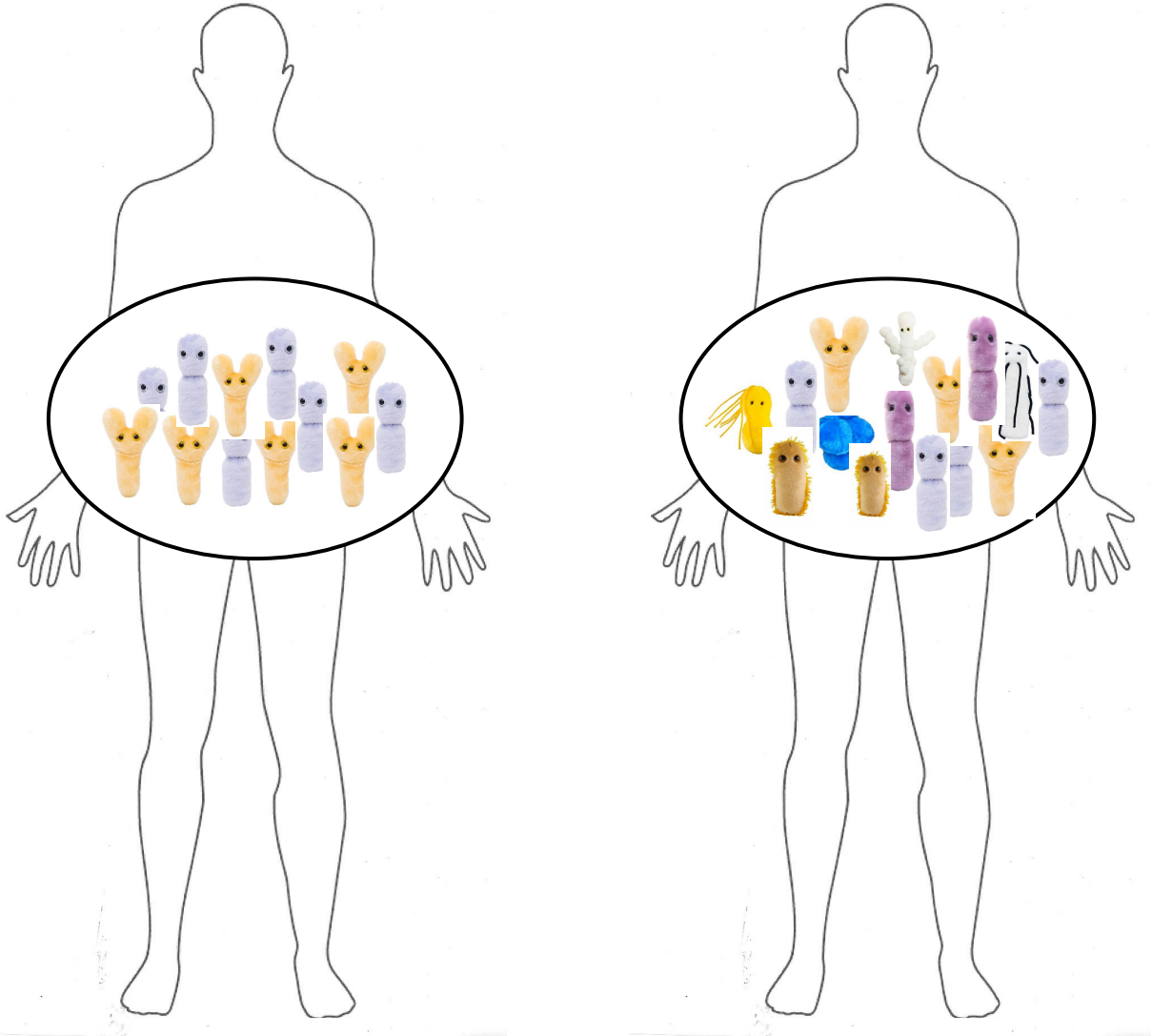
Comparing microbial communities

How many different OTUs/ASVs are there? *Alpha diversity*

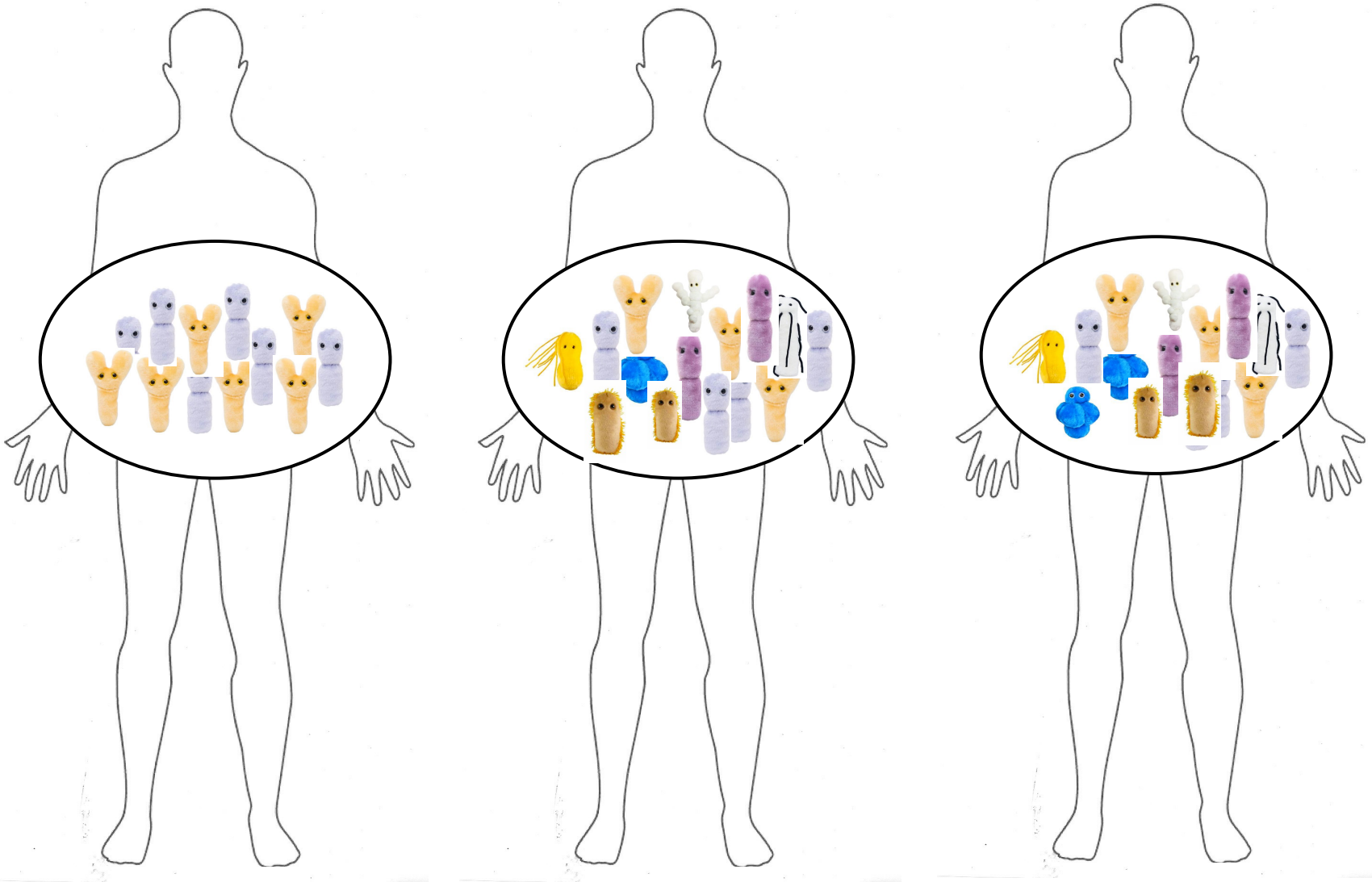
How similar are pairs of samples? *Beta diversity*

Who is there? *Taxonomic profiling, differential abundance testing*

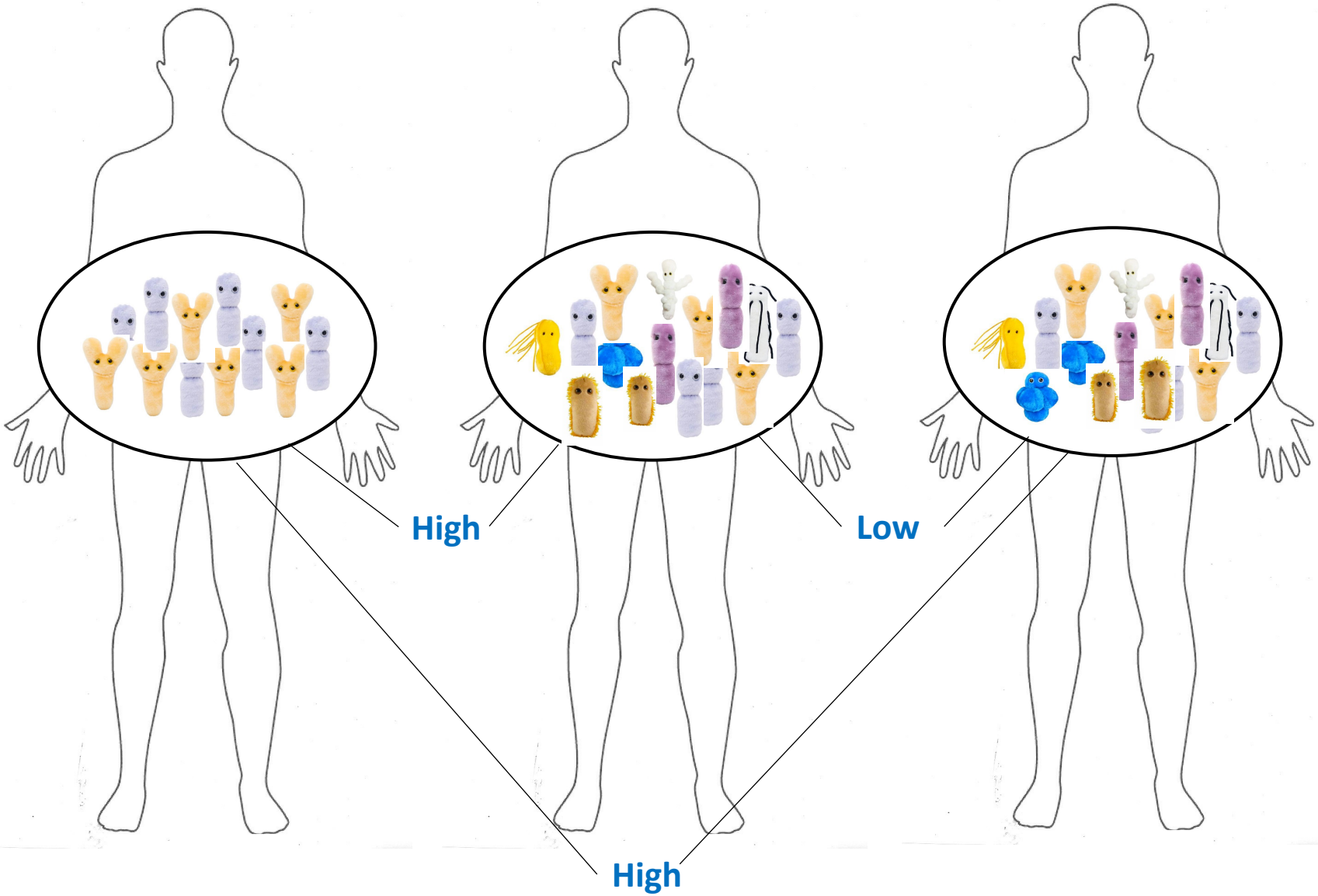
Alpha diversity: number of types of microbes



Beta diversity: difference between communities



Beta diversity: difference between communities



Bray-Curtis distance: non-phylogenetic beta diversity metric

FeatureTable[Frequency]

	feature1	feature 2	feature 3	feature 4	feature 5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$BC(A, B) = \frac{\sum_i |X_{iA} - X_{iB}|}{\sum_i (X_{iA} + X_{iB})}$$

X_{iA} : frequency of feature *i* in sample *A*

30+1+15+11+1 = 58
54+1+59+187+1 = 302
=58/302 = 0.19

DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2	0.0			
e375	0.19	0.0		
4gd8	0.15	0.07	0.0	
9872	0.65	0.69	0.70	0.0

Bray-Curtis distance: non-phylogenetic beta diversity metric

FeatureTable[Frequency]

	feature1	feature 2	feature 3	feature 4	feature 5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$BC(A, B) = \frac{\sum_i |X_{iA} - X_{iB}|}{\sum_i (X_{iA} + X_{iB})}$$

X_{iA} : frequency of feature *i* in sample *A*

DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2	0.0	0.19	0.15	0.65
e375	0.19	0.0	0.07	0.69
4gd8	0.15	0.07	0.0	0.70
9872	0.65	0.69	0.70	0.0

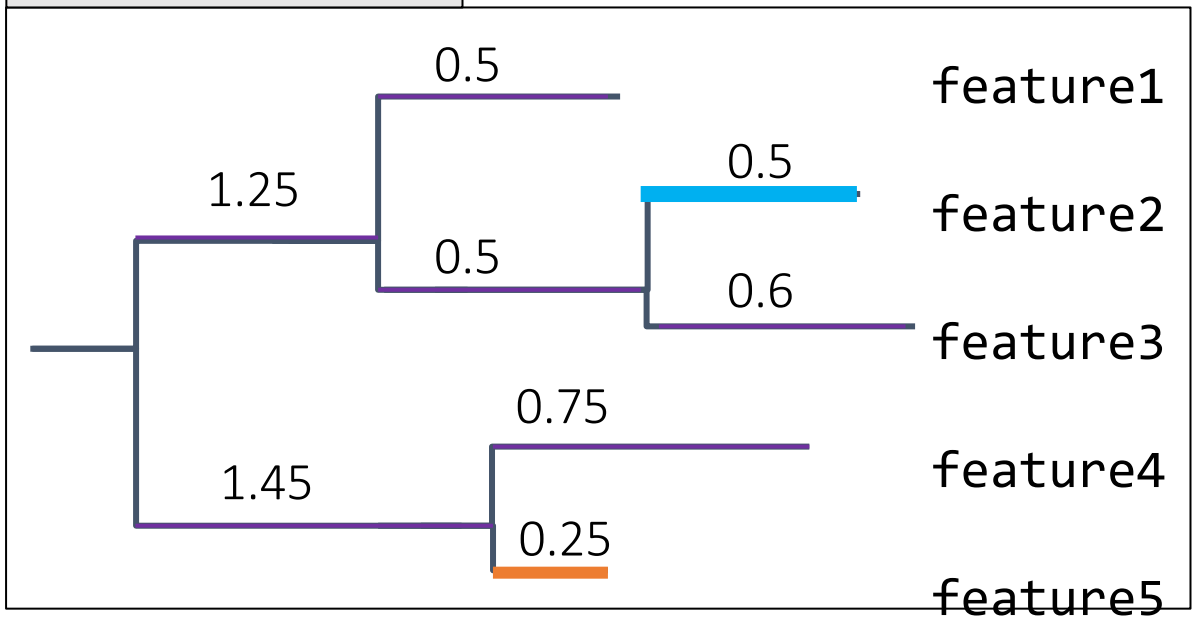
Unweighted UniFrac distance: phylogenetic beta diversity metric

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$UU(A, B) = \frac{\textit{sum of unique branch length}}{\textit{sum of observed branch length}}$$

Phylogeny[Rooted]



DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2	0.0			
e375		0.0		
4gd8			0.0	
9872				0.0

Unweighted UniFrac distance: phylogenetic beta diversity metric

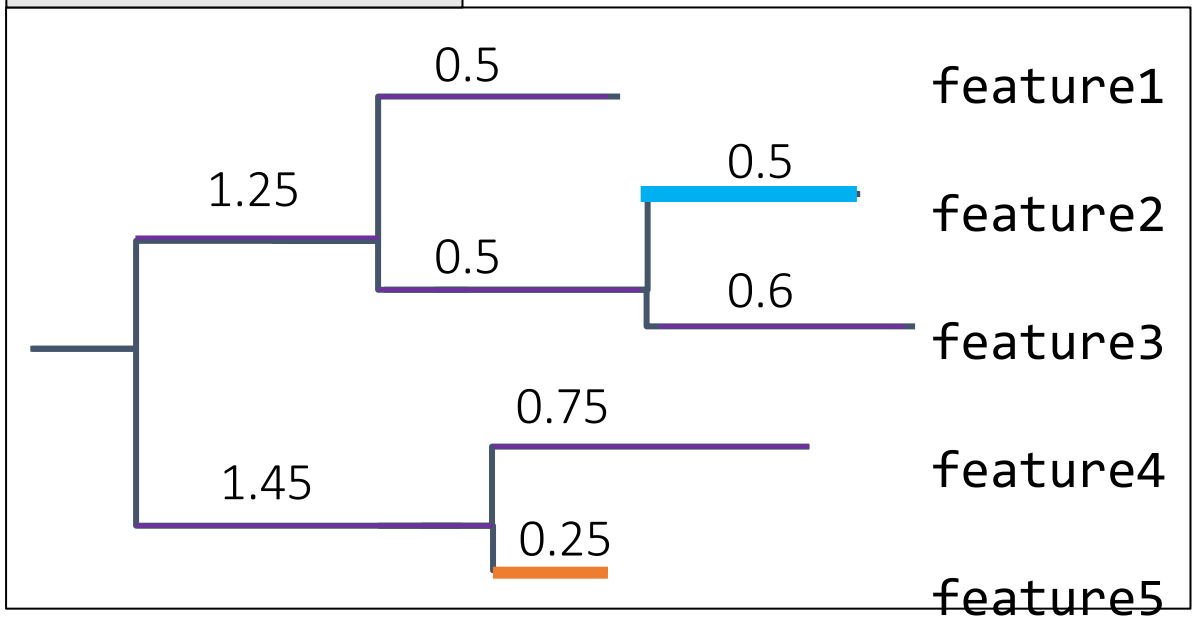
FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$UU(A, B) = \frac{\textit{sum of unique branch length}}{\textit{sum of observed branch length}}$$

Sum of unique 0.75
=1.25+.5+.5+.5+.6+1.45+.75 + 0.25= 5.8
0.75/5.8 =0.13

Phylogeny[Rooted]



DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2	0.0			
e375		0.0		
4gd8			0.0	
9872				0.0

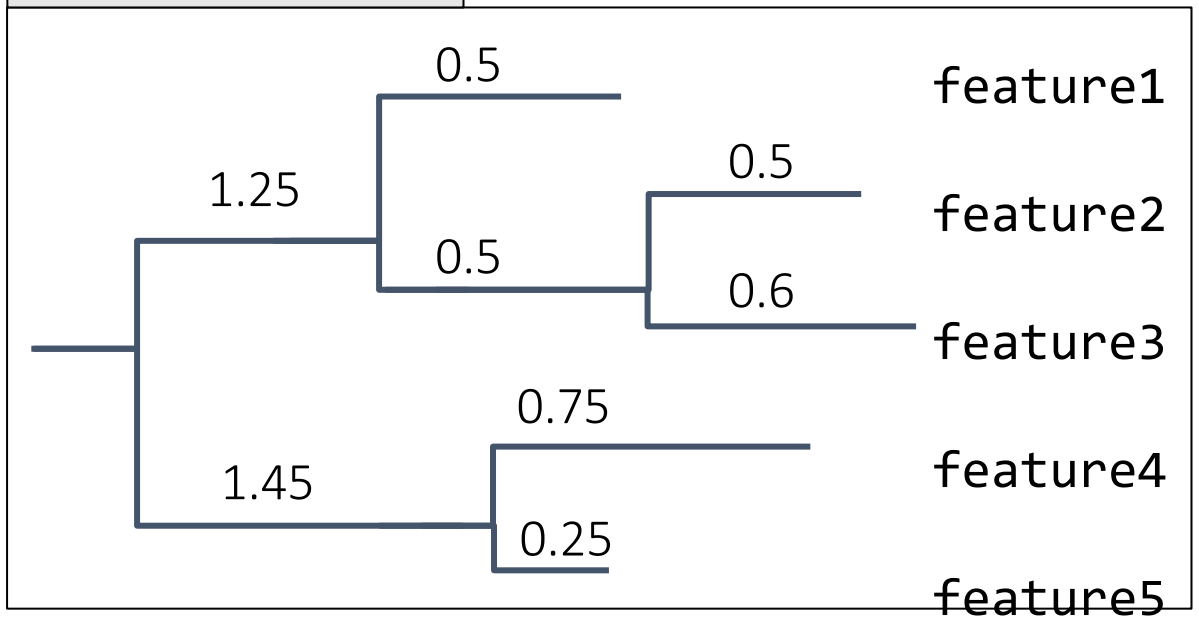
Unweighted UniFrac distance: phylogenetic beta diversity metric

FeatureTable[Frequency]

	feature1	feature2	feature3	feature4	feature5
4ac2	42	0	37	99	1
e375	12	1	22	88	0
4gd8	25	3	23	86	0
9872	0	0	87	12	0

$$UU(A, B) = \frac{\textit{sum of unique branch length}}{\textit{sum of observed branch length}}$$

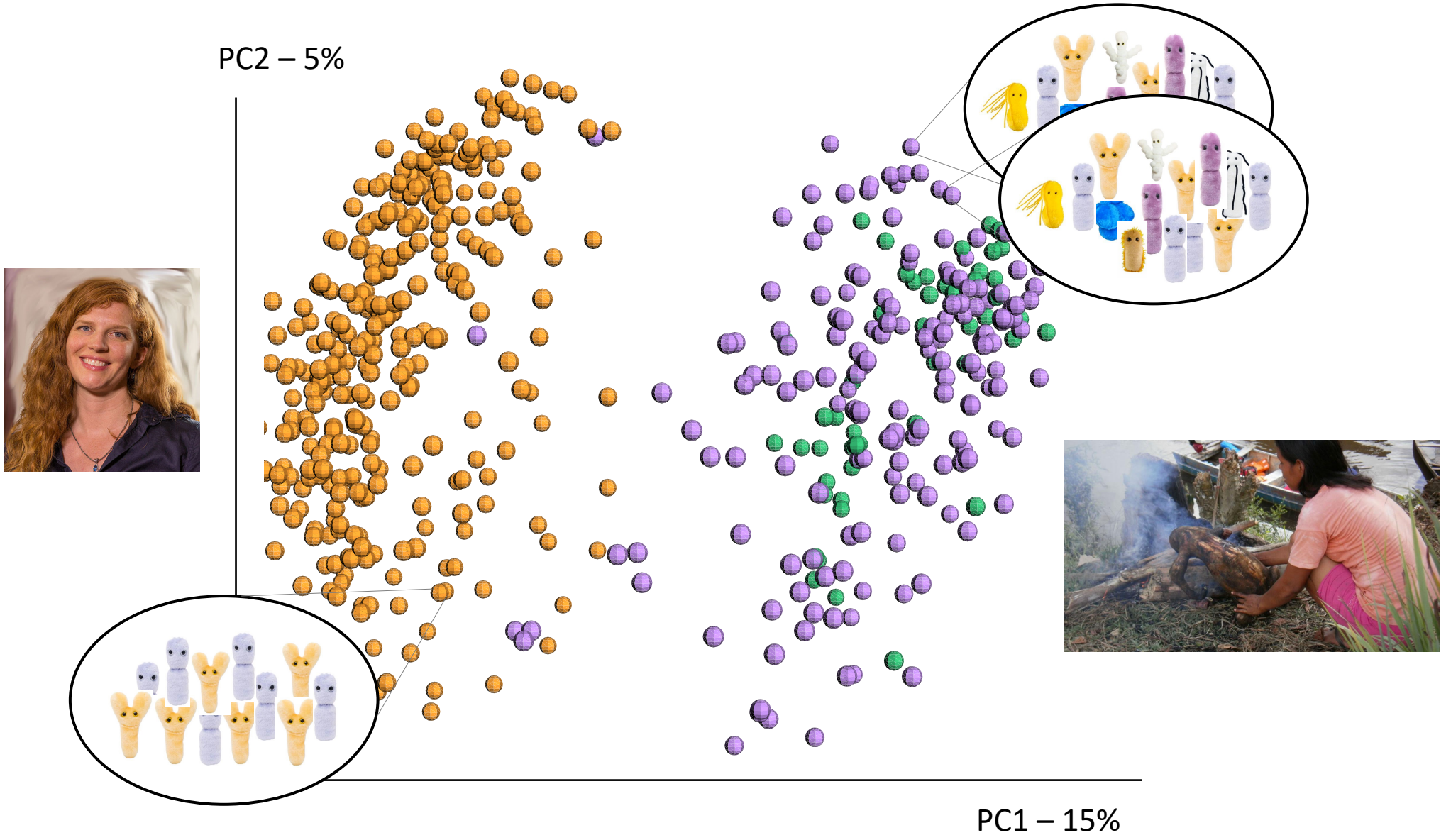
Phylogeny[Rooted]



DistanceMatrix

	4ac2	e375	4gd8	9872
4ac2	0.0	0.13	0.13	0.14
e375	0.13	0.0	0.0	0.18
4gd8	0.13	0.0	0.0	0.18
9872	0.14	0.18	0.18	0.0

PCoA: visualize beta diversity, unweighted Unifrac



Normalizing for sequencing depth

Diversity metrics are often impacted by the total number of sequences observed in samples

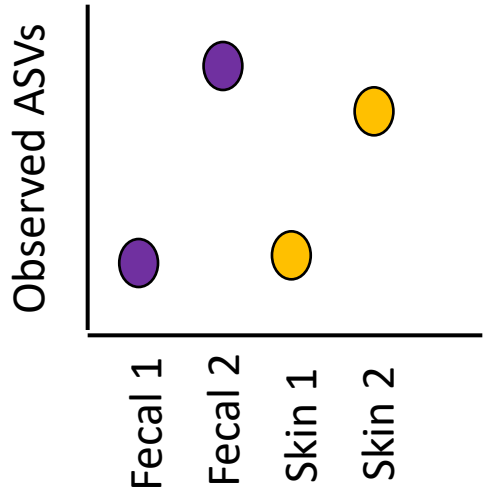
e.g. two samples can appear more similar just because they have similar sequencing depth..

fecal 1 – 100 reads

fecal 2 – 20,0000 reads

skin 1 – 105 reads

skin 2 – 17,000 reads



So we need to normalize the reads somehow

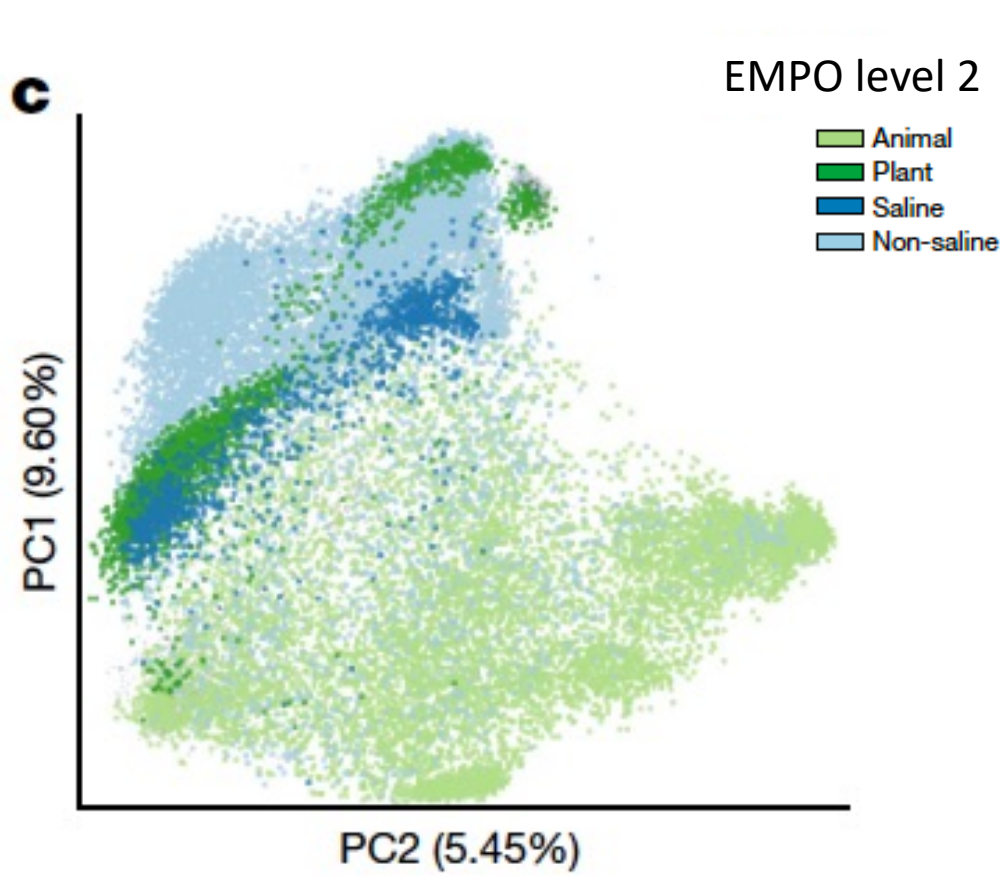
beta-group-significance

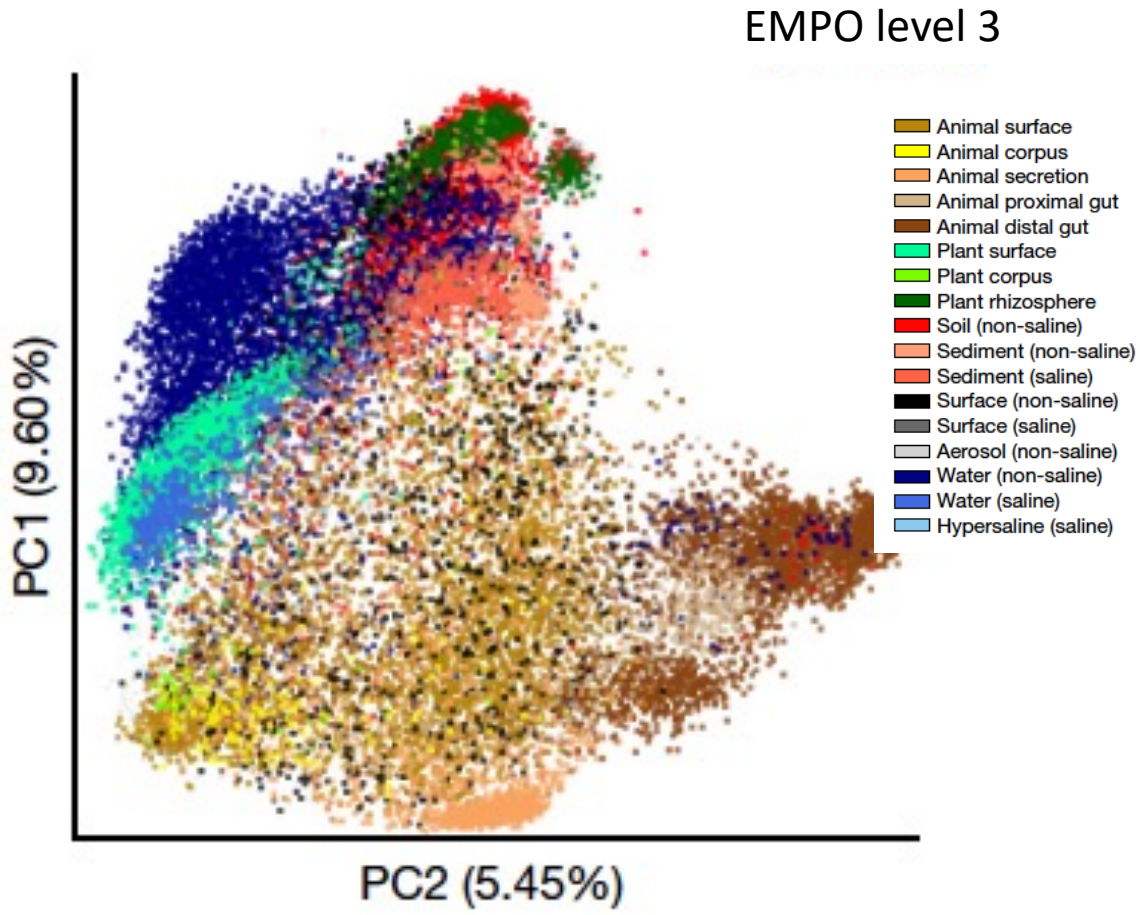
PERMANOVA - multivariate analysis of variance to determine significant difference between groups. Null = the centroids and dispersion of the groups as defined by measure space are equivalent for all groups.

Adonis allows for more than a one-way (multifactor) permanova

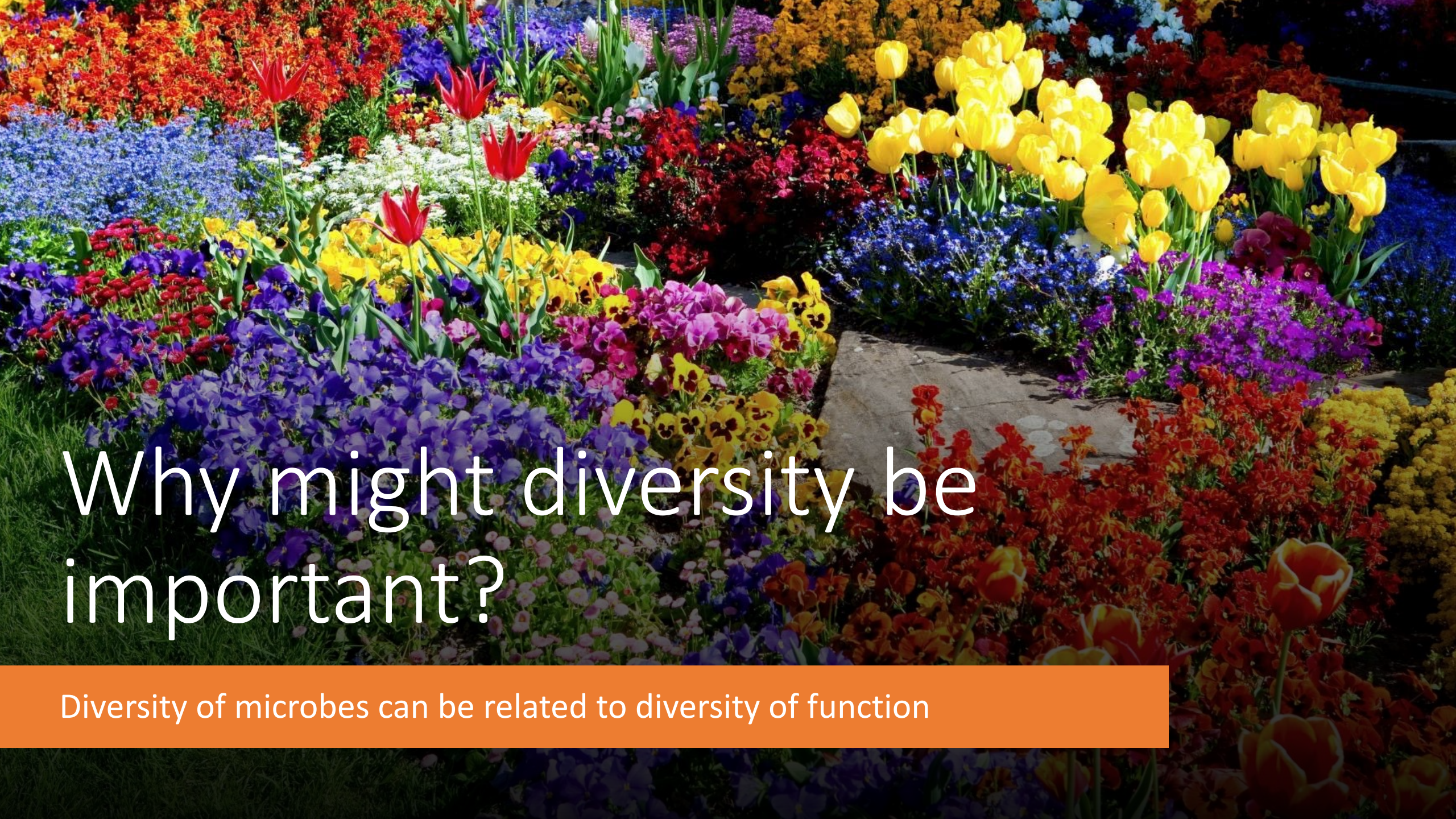
PERMDISP – used in conjunction with PERMANOVA. Null hypothesis: dispersion is equivalent

Earth Microbiome Project





Thompson *et al.* 2017, *Nature*



Why might diversity be important?

Diversity of microbes can be related to diversity of function

Leff *et al.* 2015

ITS 16S rRNA 16S rRNA shotgun

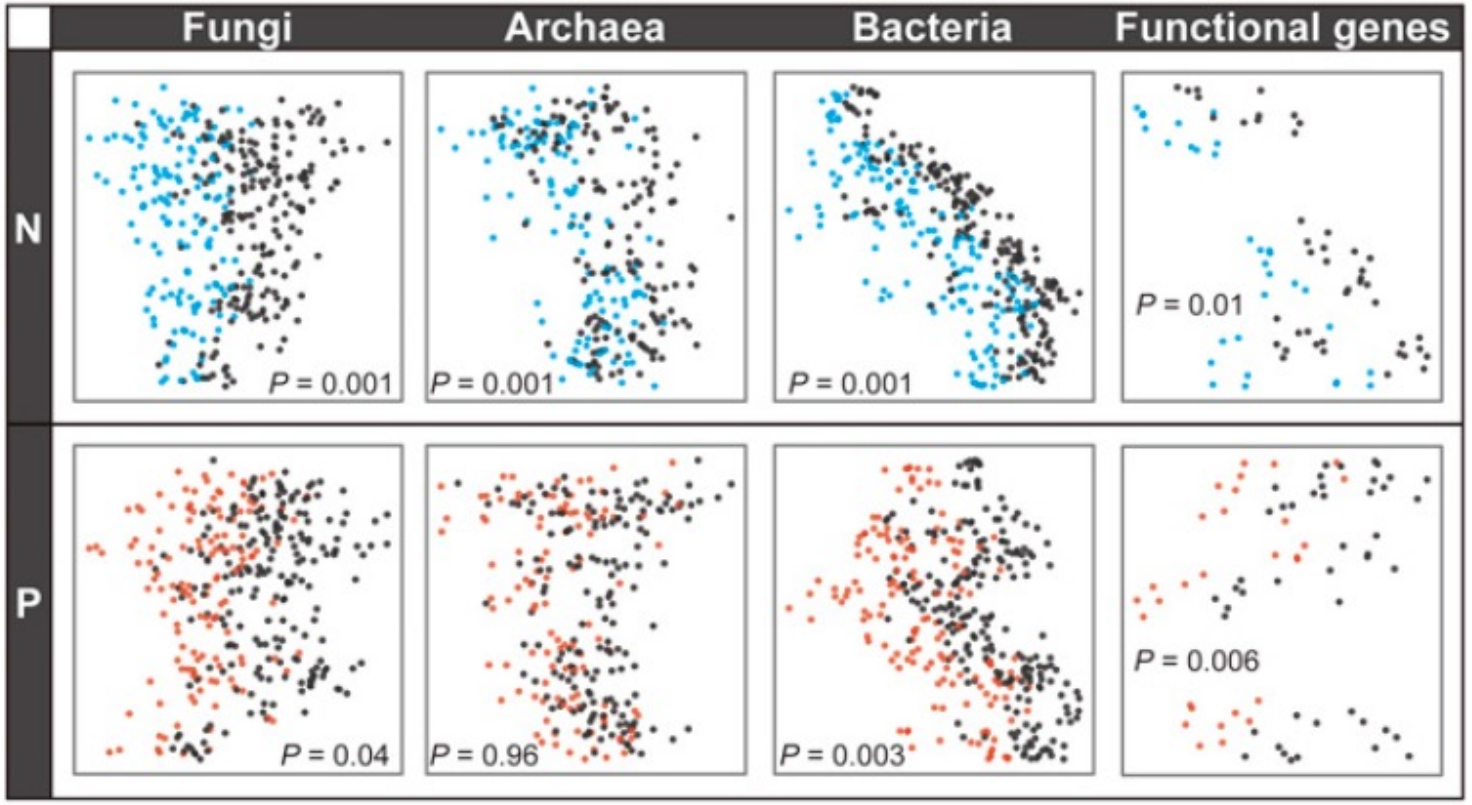
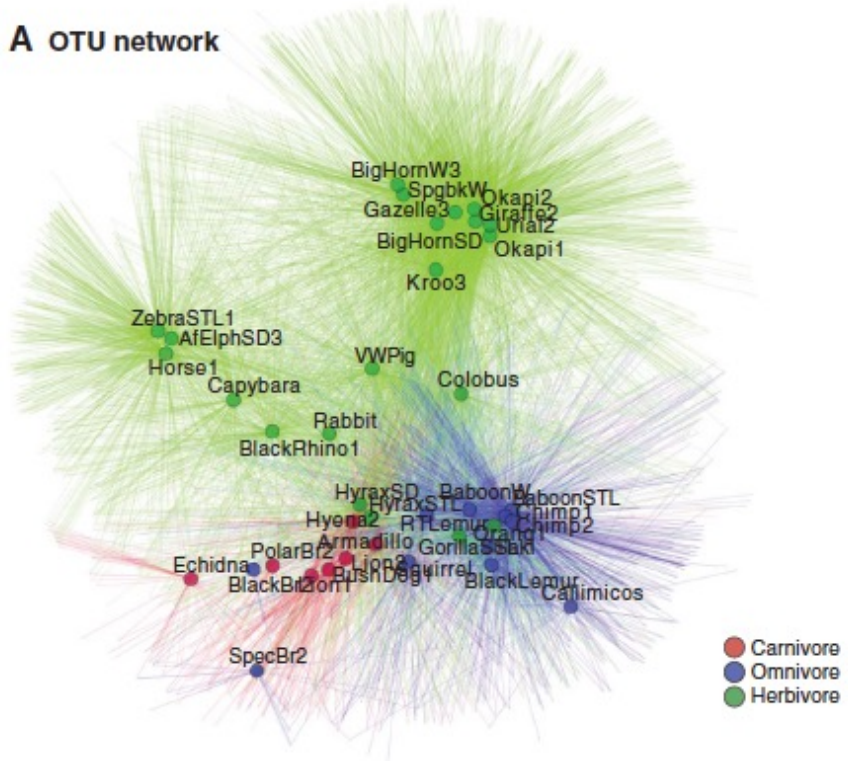


Fig. 1. Constrained ordinations showing differences between microbial communities from plots that did not receive the indicated nutrient (gray points) and from plots receiving N (blue) or P (red) additions (colored points). Colored points include samples receiving both nutrients. *P* values refer to permutational multivariate ANOVA results.

Diet Drives Convergence in Gut Microbiome Functions Across Mammalian Phylogeny and Within Humans

Brian D. Muegge,[1] Justin Kuczynski,[2] Dan Knights,[3] Jose C. Clemente,[3] Antonio González,[3] Luigi Fontana,[4,5] Bernard Henrissat,[6] Rob Knight,[2,7] Jeffrey I. Gordon[1]*



Muegge et al. 2012

ARTICLE

Received 26 Jul 2016 | Accepted 8 Dec 2016 | Published 23 Feb 2017

DOI: 10.1038/ncomms14319 OPEN

Unraveling the processes shaping mammalian gut microbiomes over evolutionary time

Mathieu Groussin[1,2,*], Florent Mazel[3,*], Jon G. Sanders[4], Chris S. Smillie[1,2,5], Sébastien Lavergne[3], Wilfried Thuiller[3] & Eric J. Alm[1,2,5]

Evolutionary trends in host physiology outweigh dietary niche in structuring primate gut microbiomes

Katherine R. Amato[1] · Jon G. Sanders[2,5] · Se Jin Song[2,5] · Michael Nute[3] · Jessica L. Metca
Luke R. Thompson[2,5] · James T. Morton[2,5,21] · Amnon Amir[2,5] · Valerie J. McKenzie[6] · Gregory Hu
Grant Gogul[7] · James Gaffney[7] · Andrea L. Baden[8] · Gillian A.O. Britton[9] · Frank P. Cuozzo[10] · An
Nathaniel J. Dominy[9] · Tony L. Goldberg[12] · Andres Gomez[13] · Martin M. Kowalewski[14] · Rebe
Andres Link[15] · Michelle L. Sauther[16] · Stacey Tecot[17] · Bryan A. White[18] · Karen E. Nelson[19] · Rebe
Rob Knight[2,5,21] · Steven R. Leigh[16]

Diet Versus Phylogeny: a Comparison of Gut Microbiota in Captive Colobine Monkey Species

Vanessa L. Hale[1] · Chia L. Tan[2] · Kefeng Niu[3,4] · Yeqin Yang[3] · Rob Knight[5,6] ·
Qikun Zhang[7] · Duoying Cui[8] · Katherine R. Amato[9]

Phyllostomid bat microbiome composition is associated to host phylogeny and feeding strategies

Mario Carrillo-Araujo[1†], Neslihan Taş[2†], Rocio J. Alcántara-Hernández[1], Osiris Gaona[1], Jorge E. Schondube[3], Rodrigo A. Medellín[4], Janet K. Jansson[5] and Luisa I. Falcón[1]*

The Bamboo-Eating Giant Panda Harbors a Carnivore-Like Gut Microbiota, with Excessive Seasonal Variations

Zhengsheng Xue,[a] Wenping Zhang,[b] Linghua Wang,[a] Rong Hou,[b] Menghui Zhang,[a] Lisong Fei,[b] Xiaojun Zhang,[a] He Huang,[b] Laura C. Bridgewater,[a] Yi Jiang,[c] Chenglin Jiang,[c] Liping Zhao,[a,d] Xiaoyan Pang,[a] Zhihe Zhang[b]

SPECIAL ISSUE: NATURE'S MICROBIOME

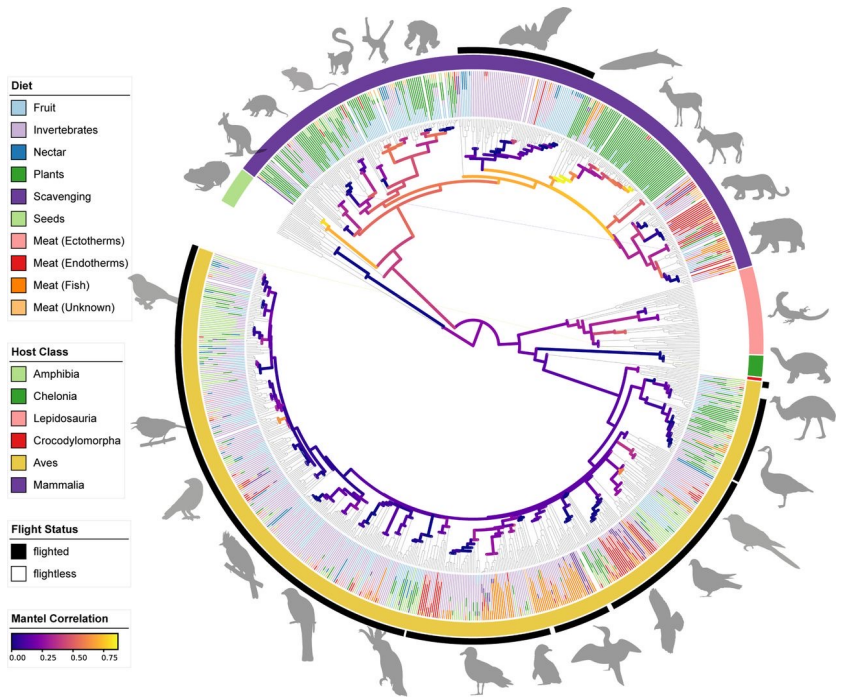
Convergence of gut microbiomes in myrmecophagous mammals

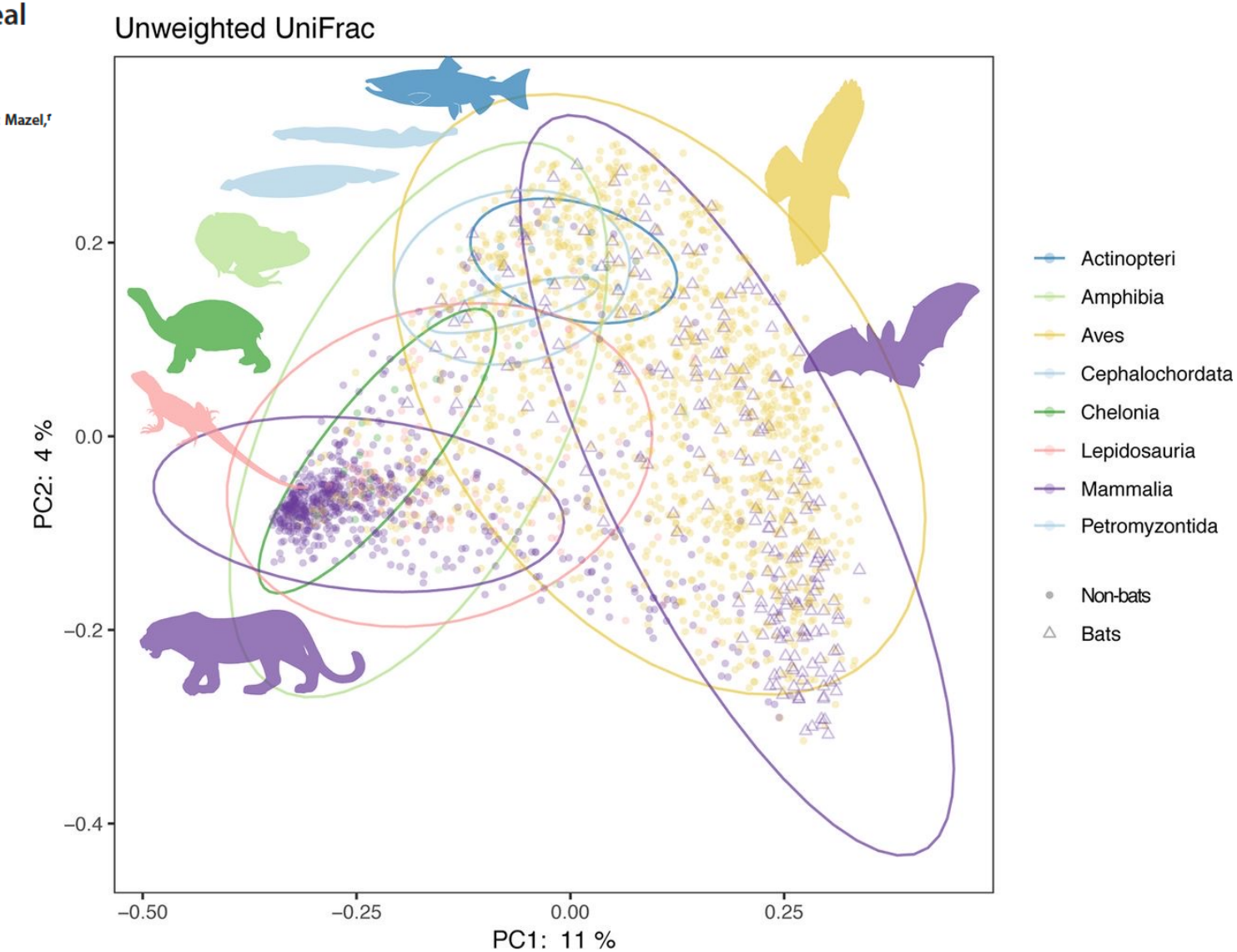
FRÉDÉRIC DELSUC,*†‡ JESSICA L. METCALF,‡ LAURA WEGENER PARFREY,‡ SE JIN SONG,‡§ ANTONIO GONZÁLEZ‡ and ROB KNIGHT†‡¶**

What does it mean when we see a lack of beta diversity pattern?

Comparative Analyses of Vertebrate Gut Microbiomes Reveal Convergence between Birds and Bats

Se Jin Song,[a] Jon G. Sanders,[a] Frédéric Delsuc,[b] Jessica Metcalf,[c] Katherine Amato,[d] Michael W. Taylor,[e] Florent Mazel,[f] Holly L. Lutz,[a,g] Kevin Winker,[h] Gary R. Graves,[i,j] Gregory Humphrey,[a] Jack A. Gilbert,[a] Shannon J. Hackett,[g] Kevin P. White,[k] Heather R. Skeen,[g,l] Sarah M. Kurtis,[m] Jack Withrow,[h] Thomas Braile,[h] Matthew Miller,[h,n] Kevin G. McCracken,[h,o,p,q,r] James M. Maley,[s] Vanessa O. Ezenwa,[t,u] Allison Williams,[t] Jessica M. Blanton,[v] Valerie J. McKenzie,[w] Rob Knight[a,x,y]





Caterpillars lack a resident gut microbiome

Tobin J. Hammer[a,b,1], Daniel H. Janzen[c], Winnie Hallwachs[c], Samuel P. Jaffe[d], and Noah Fierer[a,b]

Comparing microbial communities

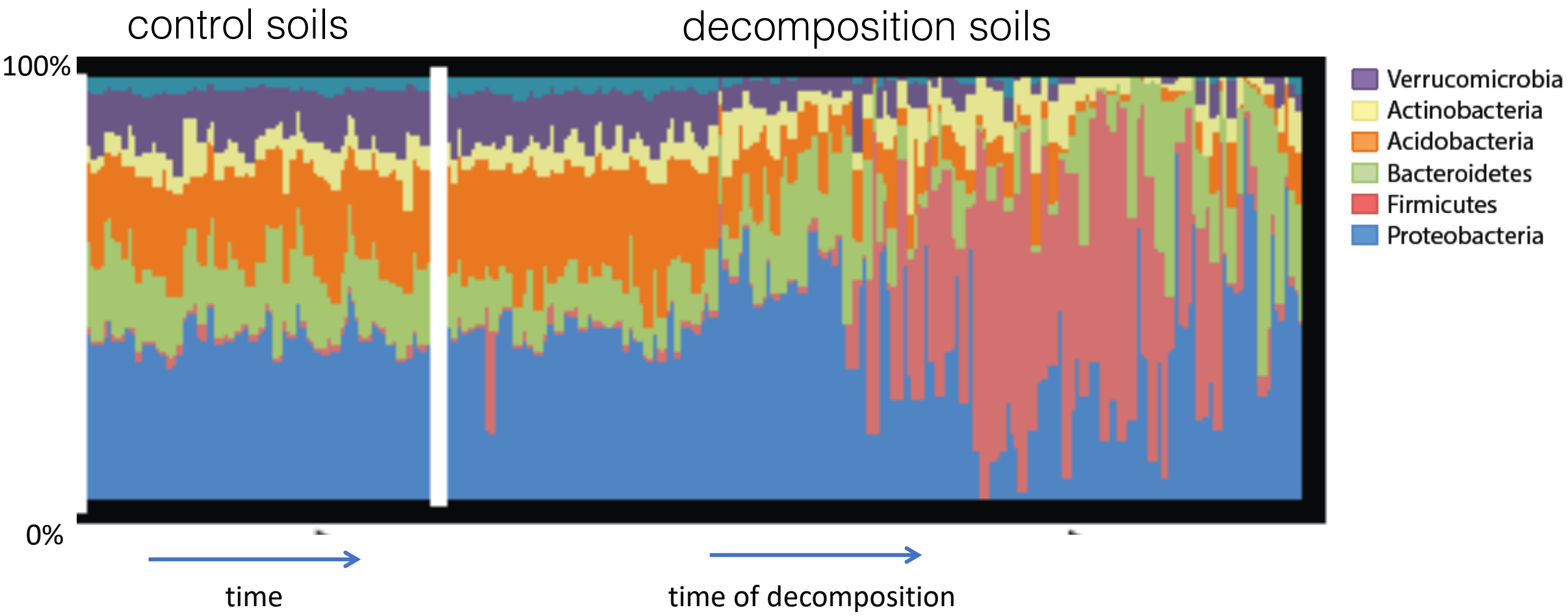
How many different OTUs/ASVs are there? *Alpha diversity*

How similar are pairs of samples? *Beta diversity*

Who is there? *Taxonomic profiling, differential abundance testing*

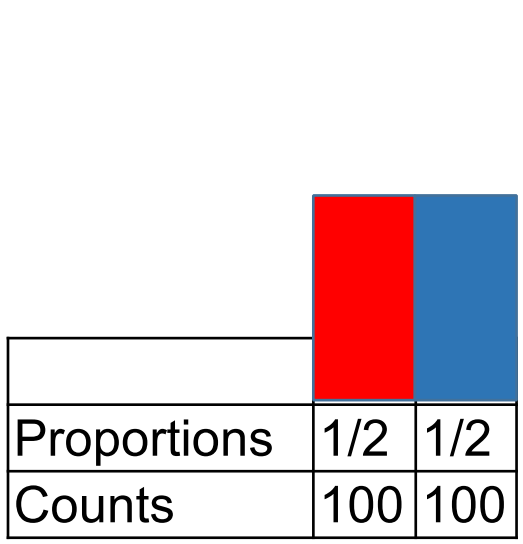
Data that are naturally described as proportions or probabilities, or with a constant or irrelevant sum, are referred to as compositional data.

What do these taxa plots really tell us?



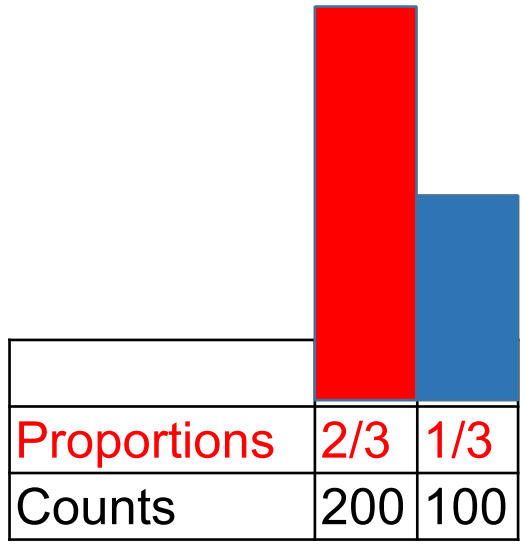
Metcalf *et al*. 2016

Compositionality

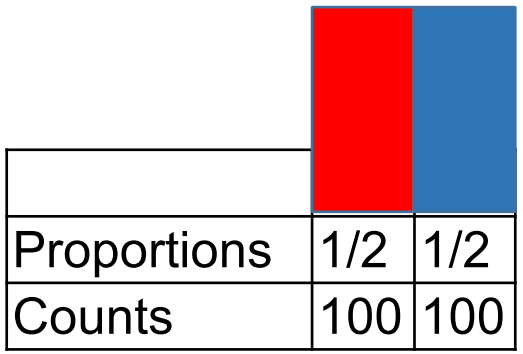


Proportions	1/2	1/2
Counts	100	100

Red doubled

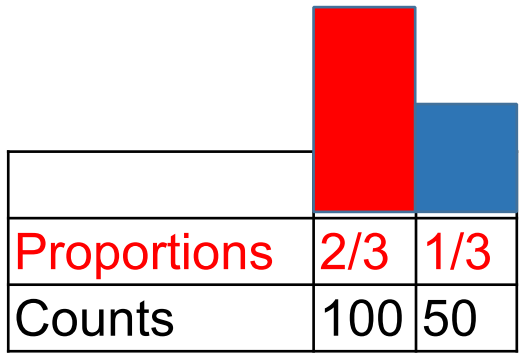


Proportions	2/3	1/3
Counts	200	100



Proportions	1/2	1/2
Counts	100	100

Blue halved

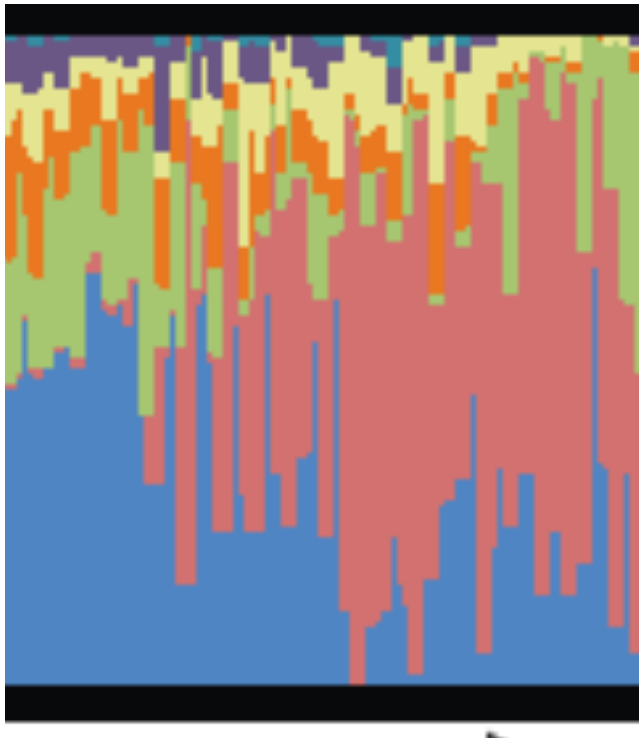


Proportions	2/3	1/3
Counts	100	50

Time point 1

Time point 2

- Cannot determine which is actually changing



Differential abundance testing is imperfect

Issues of compositionality and a lot of zeros

Non-parametric tests are problematic

e.g. Mann-Whitney/Wilcoxon rank-sum test for tests of two groups; the Kruskal-Wallis test for tests of multiple group

These don’t account for compositionality of amplicon sequencing data

ANCOM

ORIGINAL ARTICLE

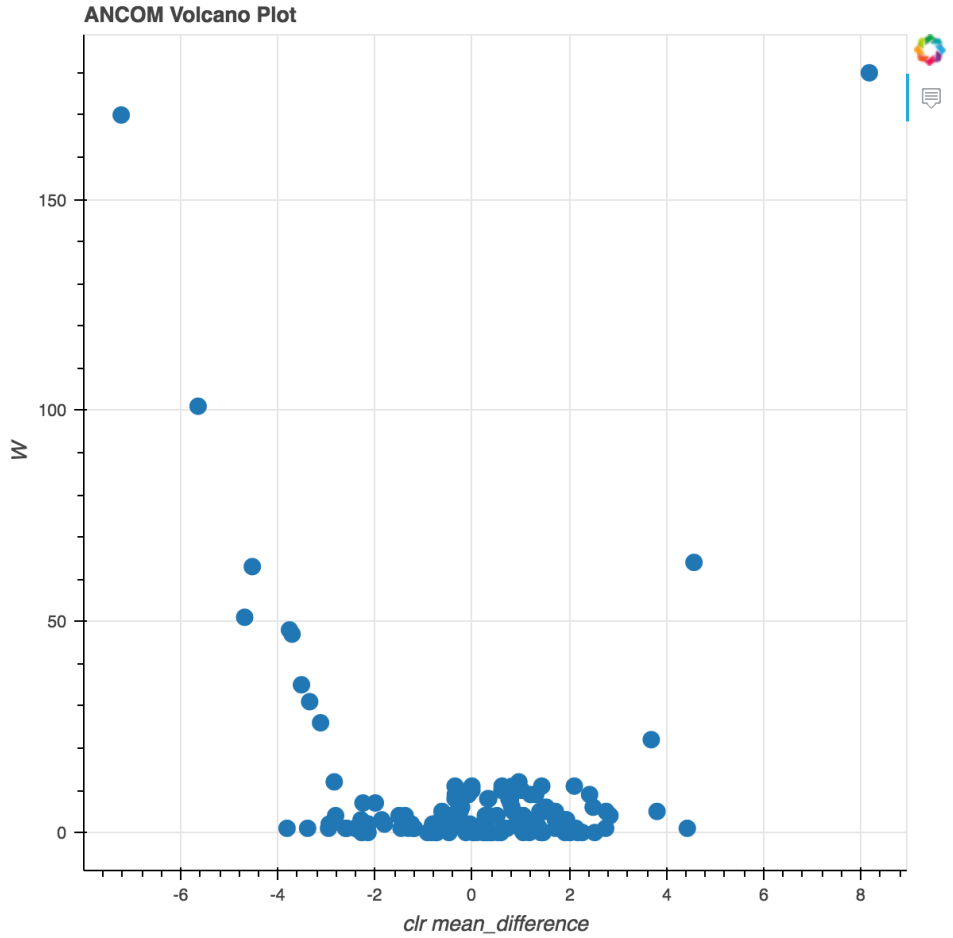
Analysis of composition of microbiomes: a novel method for studying microbial composition

Siddhartha Mandal[1], Will Van Treuren[2], Richard A. White[3], Merete Eggesbø[1], Rob Knight[4,5] and Shyamal D. Peddada[6*]

ANCOM – Analysis of Compositions of Microbiomes

Download complete table as CSV

Percentile	0.0	25.0	50.0	75.0	100.0	0.0	25.0	50.0	75.0	100.0
Group	subject-1	subject-1	subject-1	subject-1	subject-1	subject-2	subject-2	subject-2	subject-2	subject-2
4b5eeb300368260019c1fbc7a3c718fc	2223.0	2516.25	2722.0	2954.5	3328.0	1.0	1.00	1.0	1.0	1.0
868528ca947bc57b69ffdf83e6b73bae	1.0	1.00	1.0	1.0	1.0	1206.0	1631.25	1965.0	2187.0	2277.0



W= # of hypotheses rejected
The higher the number, the more likely it's a real difference

clr mean_difference (F-statistic) = magnitude of difference
If near zero, not that impressive

Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2

Michael I Love[1,2,3], Wolfgang Huber[2] and Simon Anders[2*]

Balances: a New Perspective for Microbiome Analysis

J. Rivera-Pinto,[a,b] J. J. Egozcue,[c] V. Pawlowsky-Glahn,[d] R. Paredes,[a,b,e,f] M. Noguera-Julian,[a,b,e]

Balance Trees Reveal Microbial Niche Differentiation

James T. Morton,[a,b] Jon Sanders,[a] Robert A. Quinn,[c] Daniel McDonald,[b] Antonio Gonzalez,[b] Yoshiki Vázquez-Baeza,[a,b] Jose A. Navas-Molina,[a,b] Se Jin Song,[a] Jessica L. Metcalf,[c] Embriette R. Hyde,[b] Manuel Lladser,[d] Pieter C. Dorrestein,[e] Rob Knight[a,b]

Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss[1], Zhenjiang Zech Xu[2], Shyamal Peddada[3], Amnon Amir[2], Kyle Bittinger[4], Antonio Gonzalez[2], Catherine Lozupone[5], Jesse R. Zaneveld[6], Yoshiki Vázquez-Baeza[7], Amanda Birmingham[8], Embriette R. Hyde[2] and Rob Knight[2,7,9*]

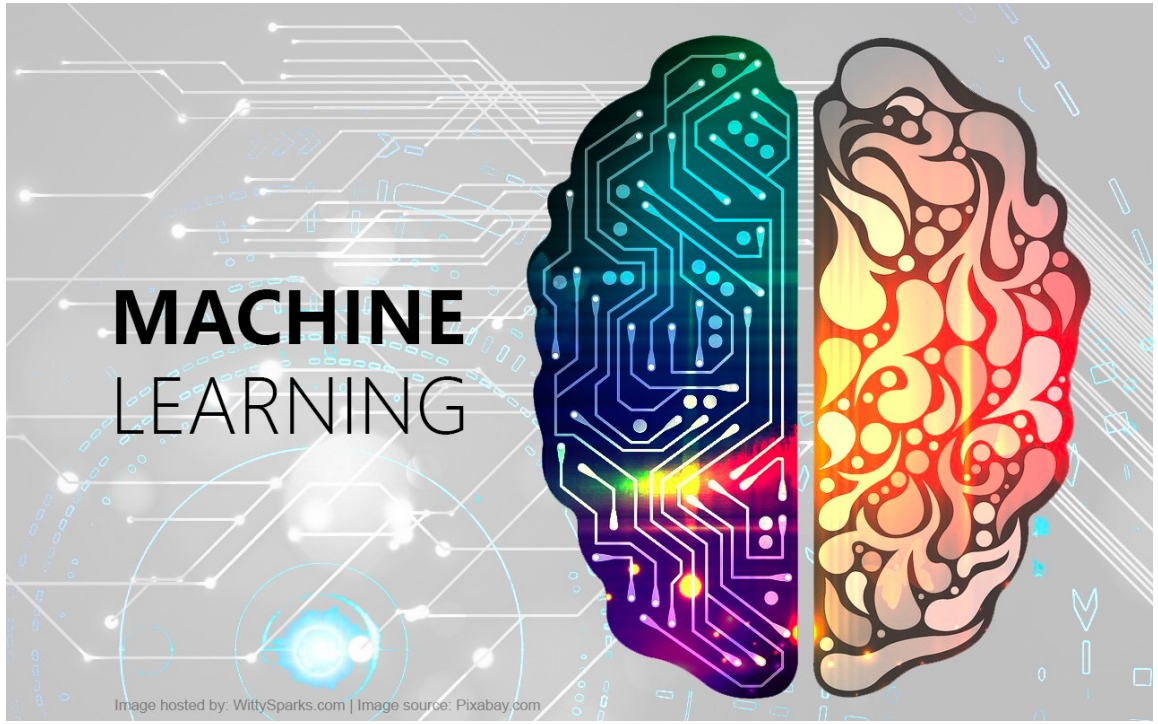
Normalization and microbial differential abundance strategies depend upon data characteristics

Sophie Weiss[1], Zhenjiang Zech Xu[2], Shyamal Peddada[3], Amnon Amir[2], Kyle Bittinger[4], Antonio Gonzalez[2], Catherine Lozupone[5], Jesse R. Zaneveld[6], Yoshiki Vázquez-Baeza[7], Amanda Birmingham[8], Embriette R. Hyde[2] and Rob Knight[2,7,9*]

- DESeq2 provides increased sensitivity on smaller datasets (<20 samples per group); however, it tends towards a higher false discovery rate with larger and/or very uneven library sizes

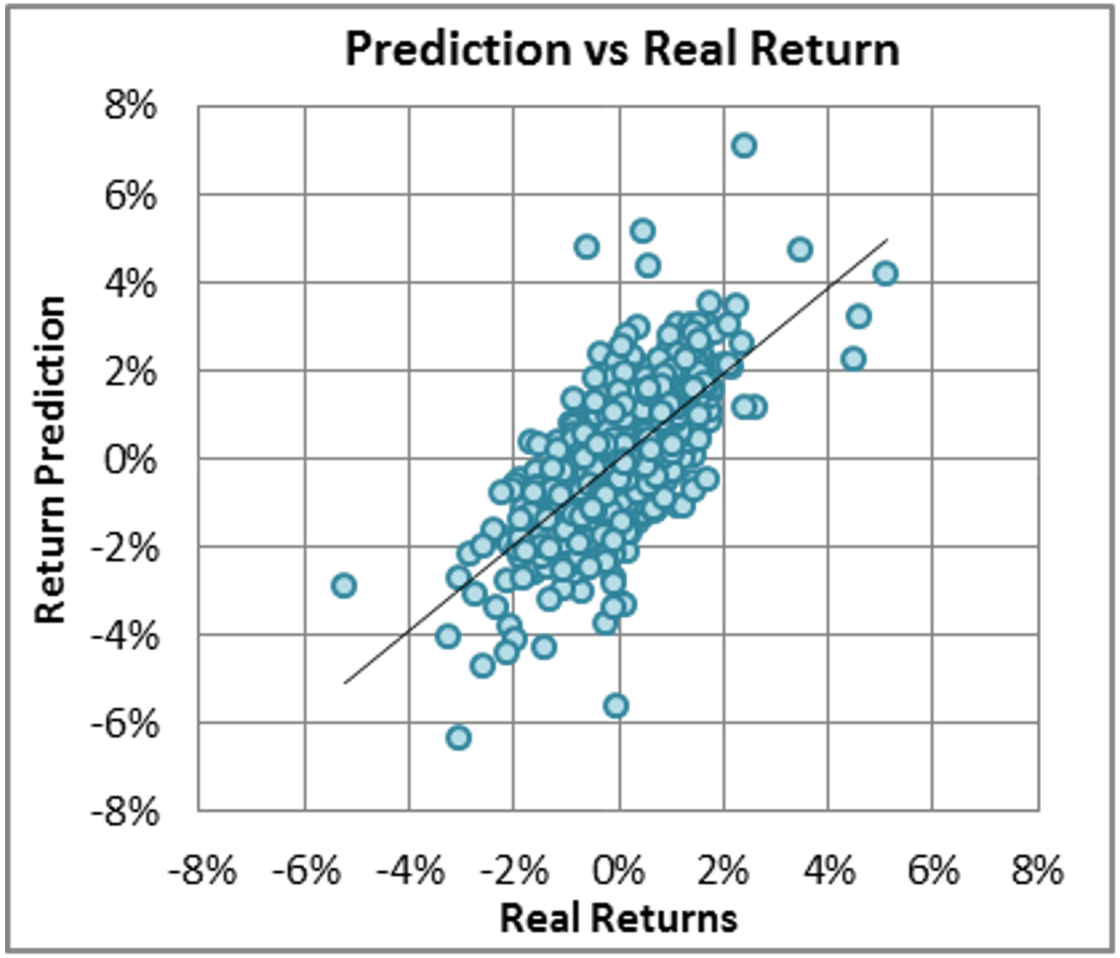
- ANCOM has very low FDR and comparable power to other methods, but sensitivity is decreased for smaller data sets

Can I predict metadata from microbial abundances?

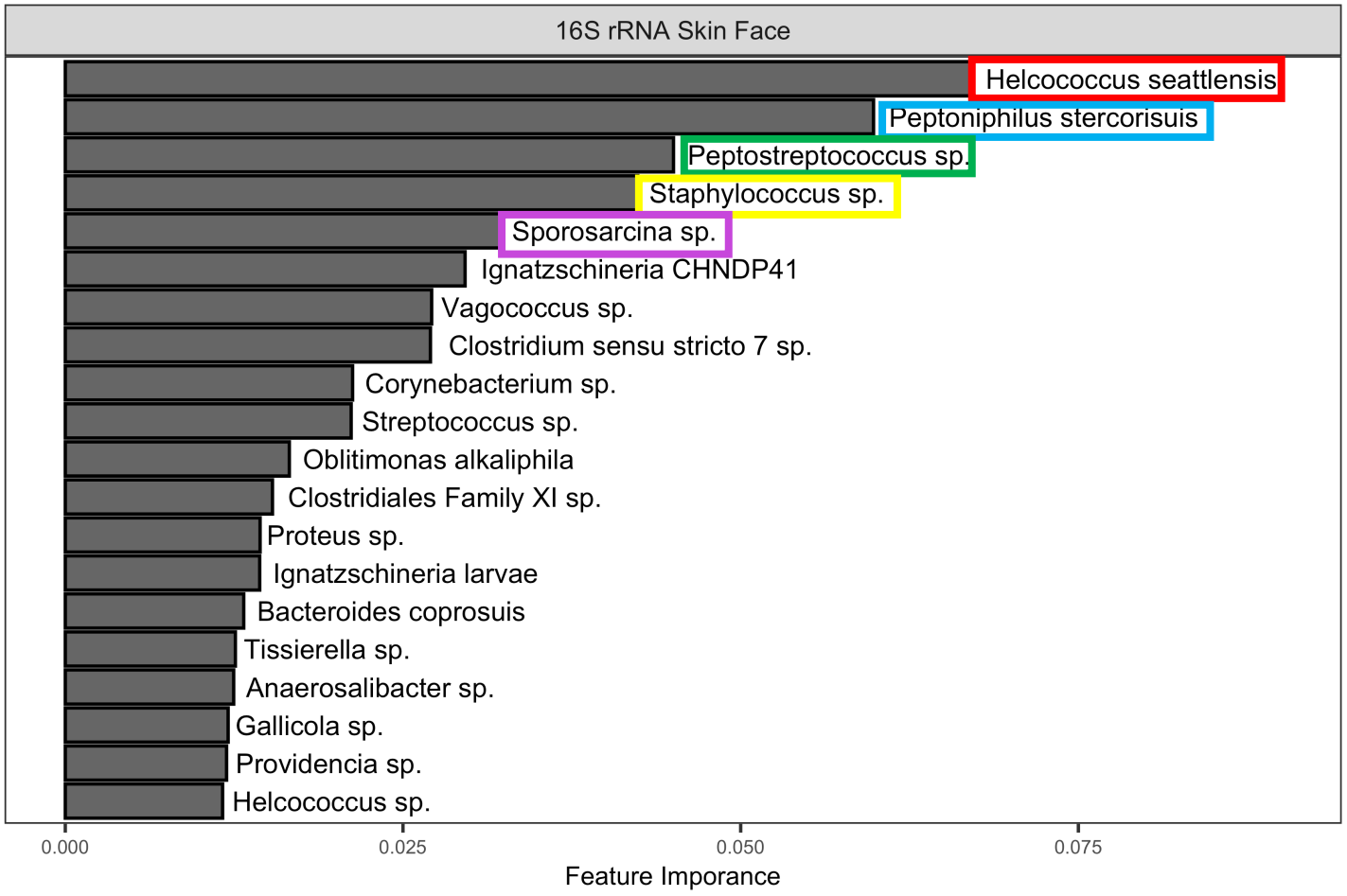


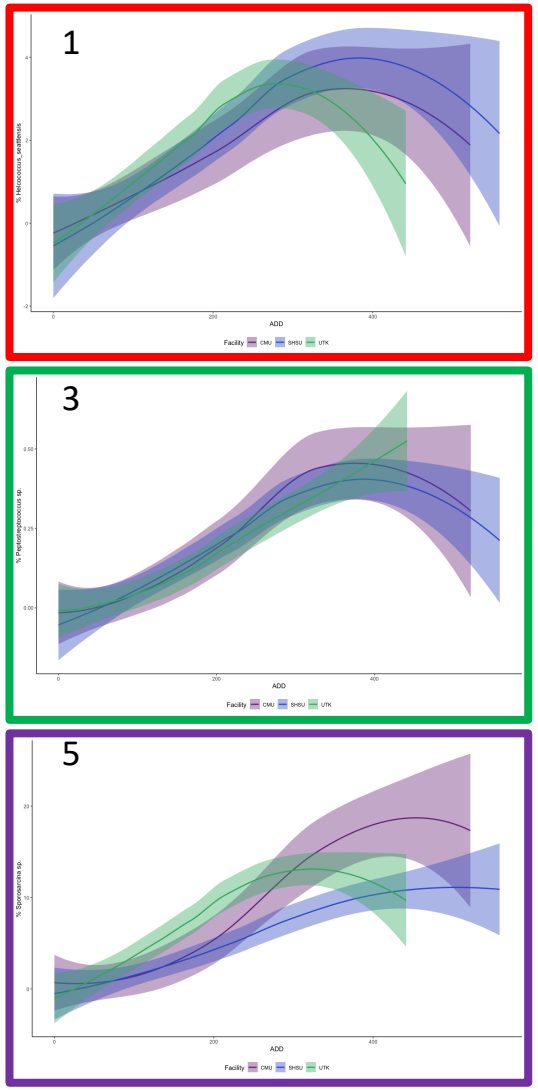
Supervised learning methods

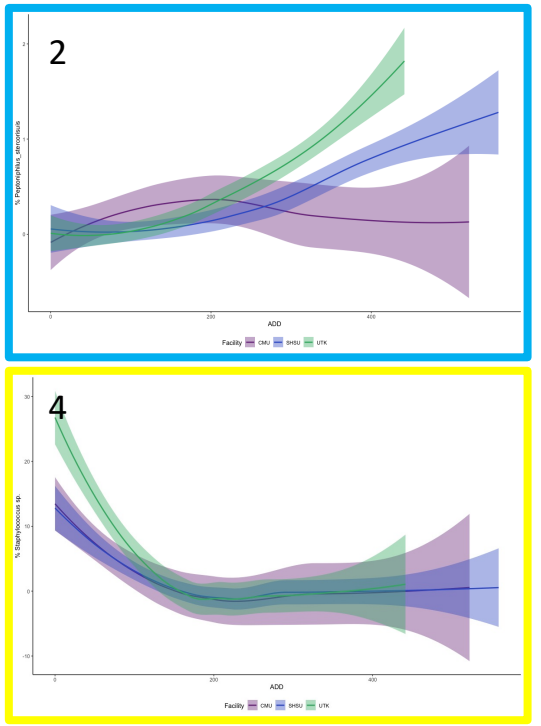
Regression vs Classification



https://quantdare.com/machine-learning-a-brief-breakdown/


16S rRNA Skin Face
Helcococcus seattlensis
Peptoniphilus stercorisuis
Peptostreptococcus sp.
Staphylococcus sp.
Sporosarcina sp.
Ignatzschineria CHNDP41
Vagococcus sp.
Clostridium sensu stricto 7 sp.
Corynebacterium sp.
Streptococcus sp.
Oblitimonas alkaliphila
Clostridiales Family XI sp.
Proteus sp.
Ignatzschineria larvae
Bacteroides coprosuis
Tissierella sp.
Anaerosalibacter sp.
Gallicola sp.
Providencia sp.
Helcococcus sp.
0.000
0.025
0.050
0.075
Feature Imporance


1
3
5
CMU
SHSU
UTK


2
4