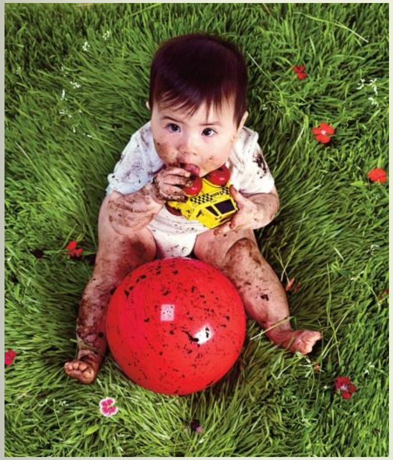
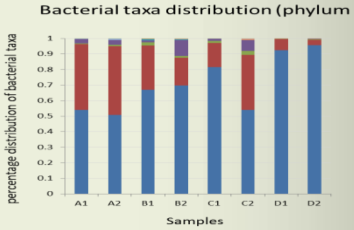
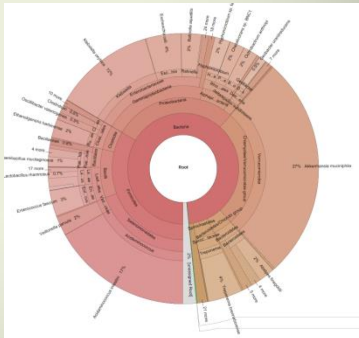
Bioinformatics analysis and Interpretation of Microbiome data




Bacterial taxa distribution (phylum
percentage distribution of bacterial taxa
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
A1
A2
B1
B2
C1
C2
D1
D2
Samples



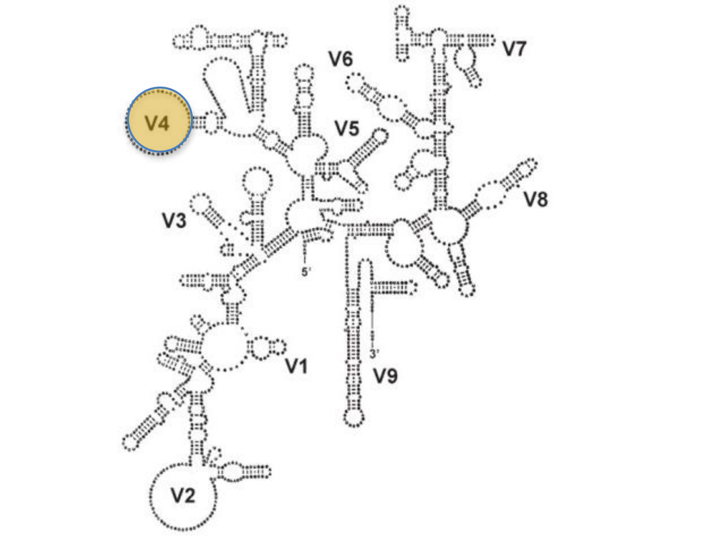
Ranjit Kumar

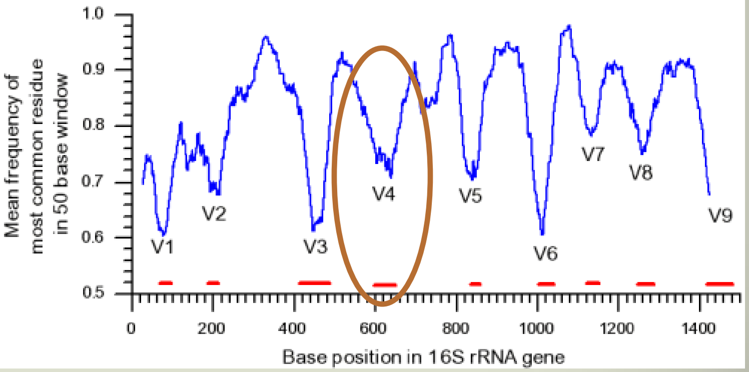
rkumar@uab.edu

UAB THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Studying microbiome : 16S rDNA gene sequencing

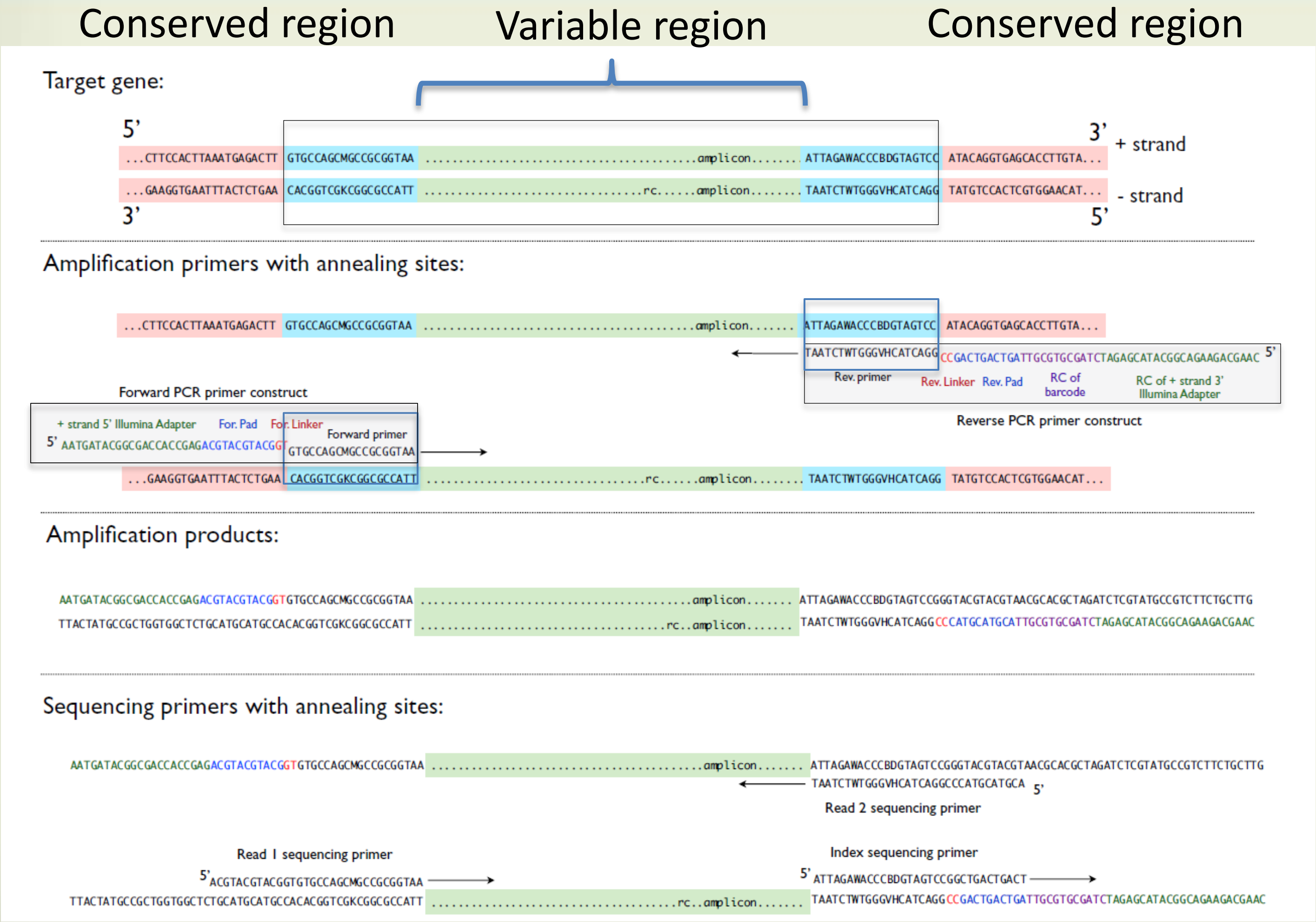
- 16S rRNA gene is found in all bacterial species
- Contains regions which are highly conserved and highly variable sequence.
- Variable sequence can be thought of as a molecular “fingerprint”. Can be used to identify bacterial genera and species.
- **Degenerate primers** are designed form the conserved region.
- Large public databases available for comparison.





□ RDP Ribosomal Database
□ Greengenes (1,049,116)
□ ARB-Silva

Primer design (V4 region)



Conserved region | Variable region | Conserved region

Target gene:

```
5'                                                                                                                      3'
...CTTCCACTTAAATGAGACTT GTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCC ATACAGGTGAGCACCTTGTA...   + strand
...GAAGGTGAATTTACTCTGAA CACGGTCGKCGGCGCCATT ...............................rc......amplicon....... TAATCTWTGGGVHCATCAGG TATGTCCACTCGTGGAACAT...   - strand
3'                                                                                                                      5'
```

Amplification primers with annealing sites:

```
...CTTCCACTTAAATGAGACTT GTGCCAGCMGCCGCGGTAA .......................................amplicon...... ATTAGAWACCCBDGTAGTCC ATACAGGTGAGCACCTTGTA...
                                                                                        <------- TAATCTWTGGGVHCATCAGGCCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC 5'
                                                                                                 Rev. primer     Rev. Linker  Rev. Pad  RC of barcode  RC of + strand 3' Illumina Adapter
                                                                                                                      Reverse PCR primer construct
Forward PCR primer construct
+ strand 5' Illumina Adapter  For. Pad  For. Linker  Forward primer
5' AATGATACGGCGACCACCGAGACGTACGTACGGT GTGCCAGCMGCCGCGGTAA ------->
...GAAGGTGAATTTACTCTGAA CACGGTCGKCGGCGCCATT ...............................rc......amplicon....... TAATCTWTGGGVHCATCAGG TATGTCCACTCGTGGAACAT...
```

Amplification products:

```
AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
TTACTATGCCGCTGGTGGCTCTGCATGCATGCCACACGGTCGKCGGCGCCATT ...................................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCCATGCATGCATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

Sequencing primers with annealing sites:

```
AATGATACGGCGACCACCGAGACGTACGTACGGTGTGCCAGCMGCCGCGGTAA .......................................amplicon....... ATTAGAWACCCBDGTAGTCCGGGTACGTACGTAACGCACGCTAGATCTCGTATGCCGTCTTCTGCTTG
                                                                                                  <------- TAATCTWTGGGVHCATCAGGCCCATGCATGCA 5'
                                                                                                           Read 2 sequencing primer

                 Read 1 sequencing primer                                                                  Index sequencing primer
             5' ACGTACGTACGGTGTGCCAGCMGCCGCGGTAA ------->                                                5' ATTAGAWACCCBDGTAGTCCGGCTGACTGACT ------->
TTACTATGCCGCTGGTGGCTCTGCATGCATGCCACACGGTCGKCGGCGCCATT ...................................rc..amplicon....... TAATCTWTGGGVHCATCAGGCCGACTGACTGATTGCGTGCGATCTAGAGCATACGGCAGAAGACGAAC
```

Proc Natl Acad Sci U S A. 2011 Mar 15;108 Suppl 1:4516-22. doi: 10.1073/pnas.1000080107. Epub 2010 Jun 3.

Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample.

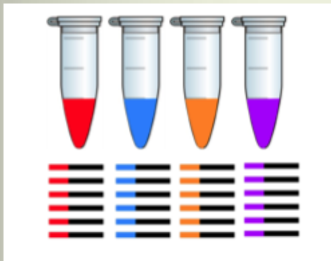
Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R.

Next Generation Sequencing

- ✓ Culture independent study
- ✓ Low quantity of sample needed (20-50ng DNA)
- ✓ High sequencing depth (identification of rare microbes).
- ✓ Multiplexing of many different samples in one run using indexes (around 90+ samples).

Example : Illumina MiSeq produces 20M reads/lane. Single run can be multiplexed to include around 90 samples.

Microbiome Analysis in Nutshell



Extract DNA and amplify 16S gene with barcoded primers

Next Generation sequencing using GAIIx

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

Assign reads to communities

- Sample De-multiplexing
- Quality Control
- Sequence clustering into OTUs (Operational Taxonomic Units)
- Pick representative sequences
- Assign Taxonomy
- phylogenetic tree

16S : Greengenes

OTU table

16S sequences

ANALYSIS

THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Species and OTUs

Species : A species is often defined as the largest group of organisms capable of interbreeding and producing fertile offspring. While in many cases this definition is adequate, the difficulty of defining species is known as the species problem. Source: Wikipedia

"No single definition has satisfied all naturalists; yet every naturalist knows vaguely what he means when he speaks of a species"

Charles Darwin,

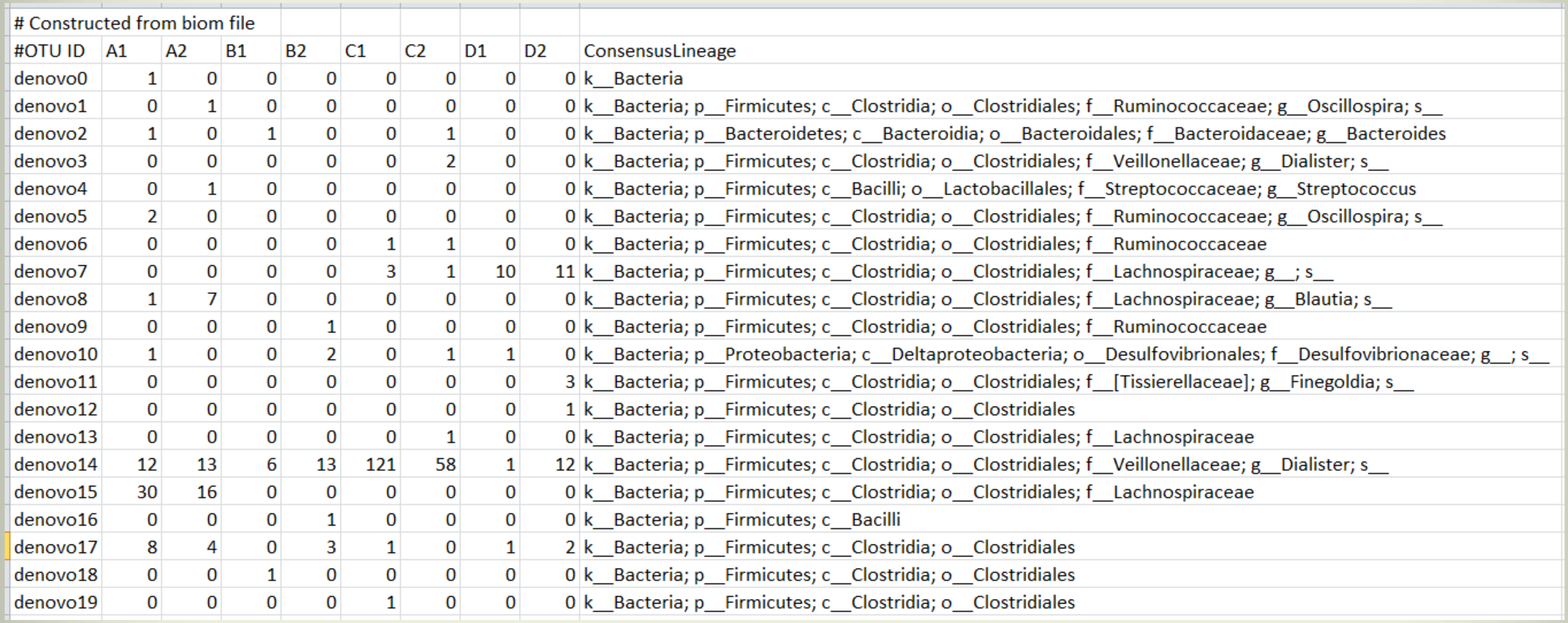
On the Origin of Species, 1859

OTUs (Operational Taxonomic unit) : An arbitrary definition of a taxonomic unit based on sequence divergence. Here OTUs are number of clusters of similar sequences. Generally, when 16S sequences are clustered at 97% identity ~ species.

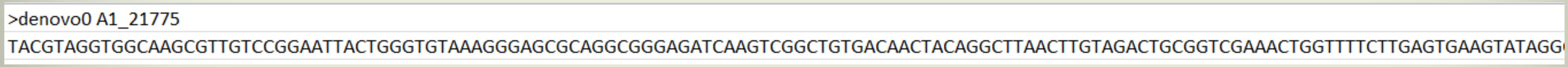
Taxonomy assignment

- Perfect 1 match to database
- Perfect multiple match to database
- No perfect match to database

OTU table

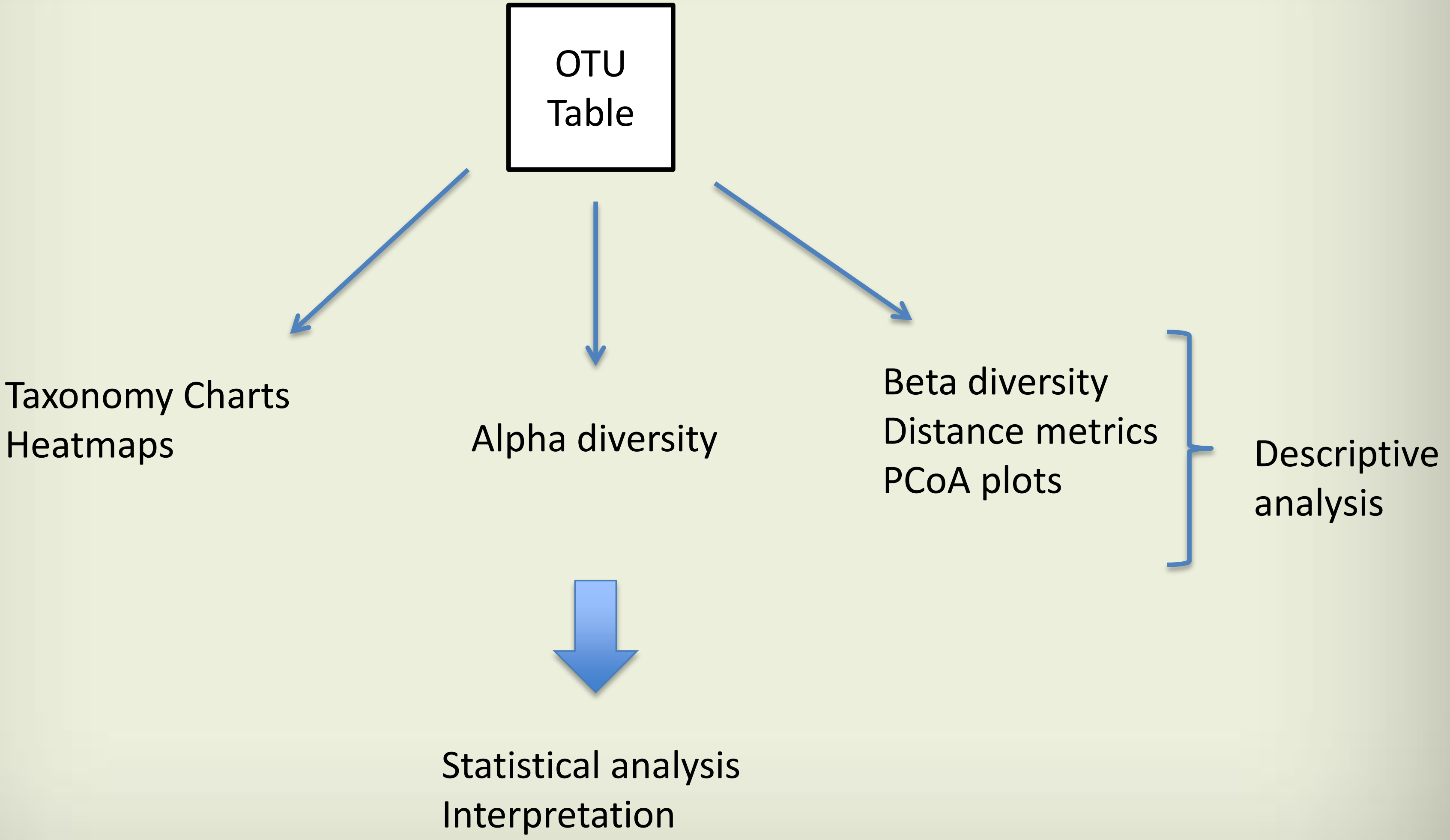


# Constructed from biom file									
#OTU ID	A1	A2	B1	B2	C1	C2	D1	D2	ConsensusLineage
denovo0	1	0	0	0	0	0	0	0	k__Bacteria
denovo1	0	1	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira; s__
denovo2	1	0	1	0	0	1	0	0	k__Bacteria; p__Bacteroidetes; c__Bacteroidia; o__Bacteroidales; f__Bacteroidaceae; g__Bacteroides
denovo3	0	0	0	0	0	2	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Dialister; s__
denovo4	0	1	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Streptococcaceae; g__Streptococcus
denovo5	2	0	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae; g__Oscillospira; s__
denovo6	0	0	0	0	1	1	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae
denovo7	0	0	0	0	3	1	10	11	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__; s__
denovo8	1	7	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae; g__Blautia; s__
denovo9	0	0	0	1	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Ruminococcaceae
denovo10	1	0	0	2	0	1	1	0	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Desulfovibrionales; f__Desulfovibrionaceae; g__; s__
denovo11	0	0	0	0	0	0	0	3	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__[Tissierellaceae]; g__Finegoldia; s__
denovo12	0	0	0	0	0	0	0	1	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
denovo13	0	0	0	0	0	1	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae
denovo14	12	13	6	13	121	58	1	12	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Veillonellaceae; g__Dialister; s__
denovo15	30	16	0	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Lachnospiraceae
denovo16	0	0	0	1	0	0	0	0	k__Bacteria; p__Firmicutes; c__Bacilli
denovo17	8	4	0	3	1	0	1	2	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
denovo18	0	0	1	0	0	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales
denovo19	0	0	0	0	1	0	0	0	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales



```
>denovo0 A1_21775
TACGTAGGTGGCAAGCGTTGTCCGGAATTACTGGGTGTAAAGGGAGCGCAGGCGGGAGATCAAGTCGGCTGTGACAACTACAGGCTTAACTTGTAGACTGCGGTCGAAACTGGTTTTCTTGAGTGAAGTATAGG
```

OTU table -> ANALYSIS



OTU Table

Taxonomy Charts
Heatmaps

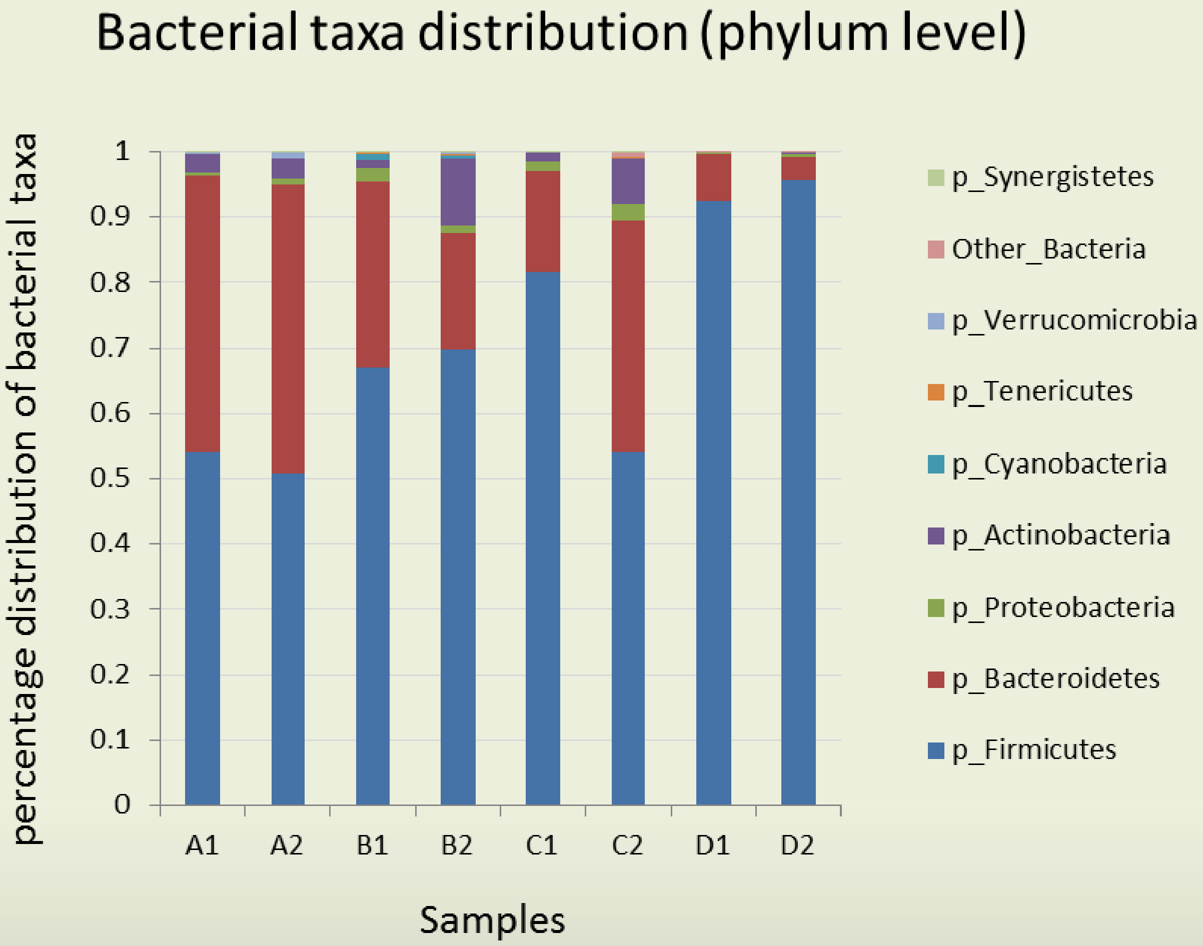
Alpha diversity

Beta diversity
Distance metrics
PCoA plots

Descriptive analysis

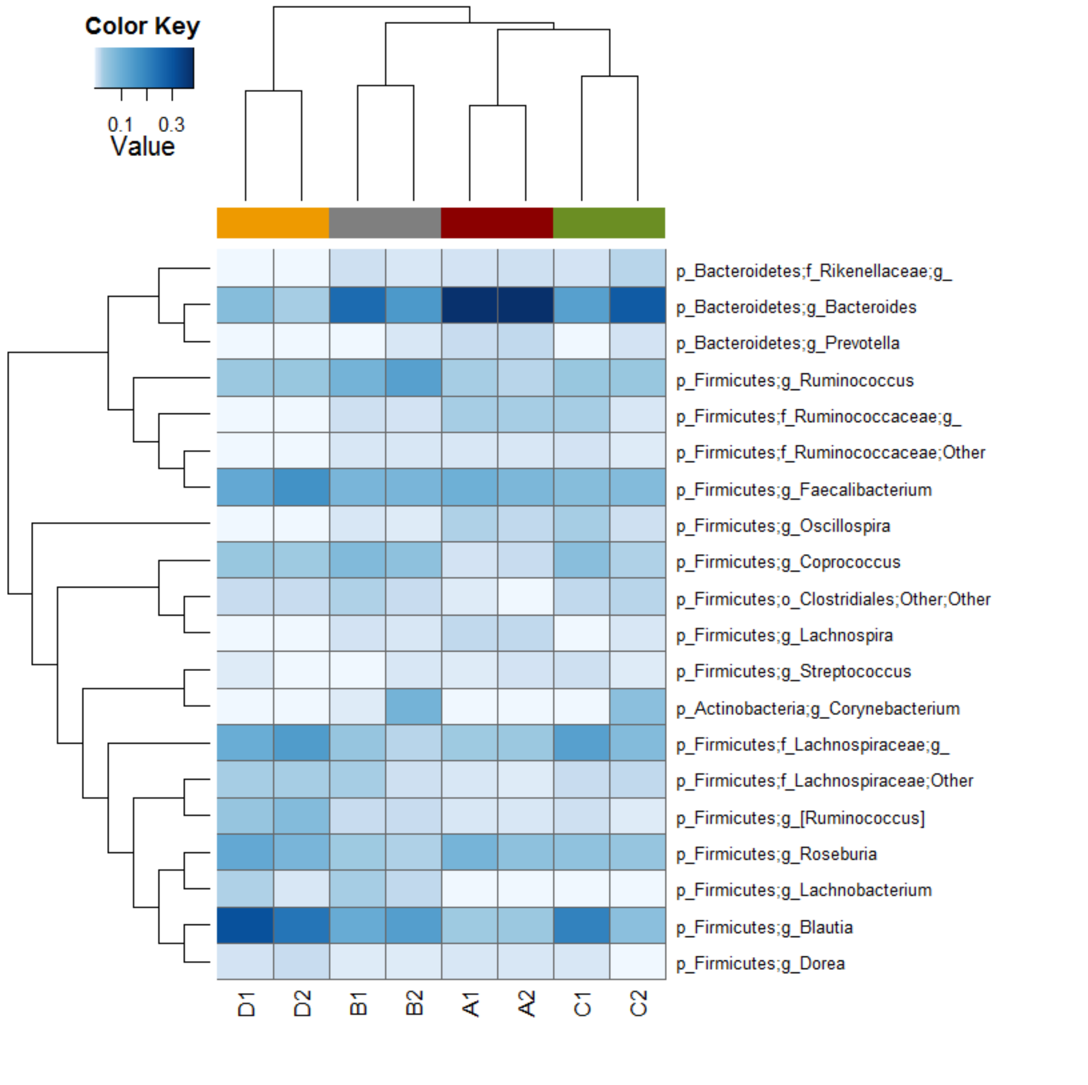
Statistical analysis
Interpretation

Taxa distribution (bar charts)



THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

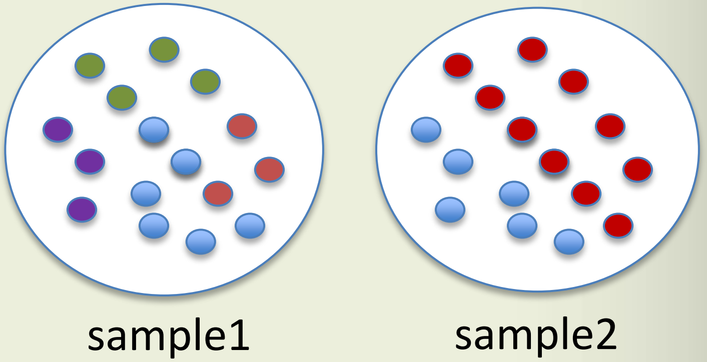
Taxa distribution (Heatmap)



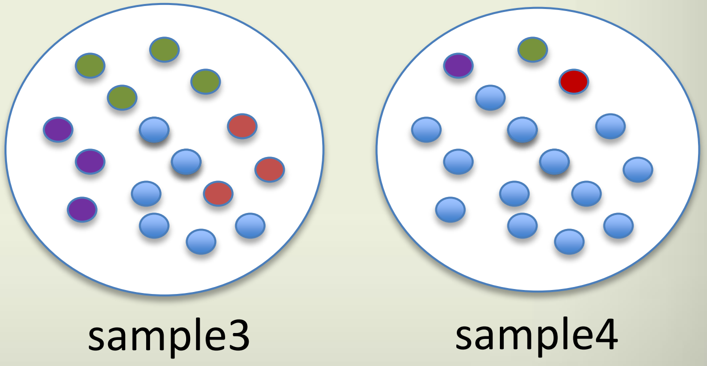
Alpha diversity

What is Alpha diversity - It is used to measure the diversity within a sample. It is calculated as a value for each sample. Different metrics were developed to calculate diversity in different ways.

Count of different microbes (OTU count)
Richness - Richness is a measure
of number of species present in a sample.

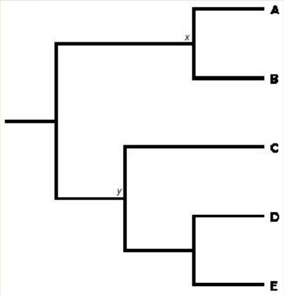


Distribution of different microbes
Evenness - Evenness is a measure of relative abundance of different species that make up the richness in that area



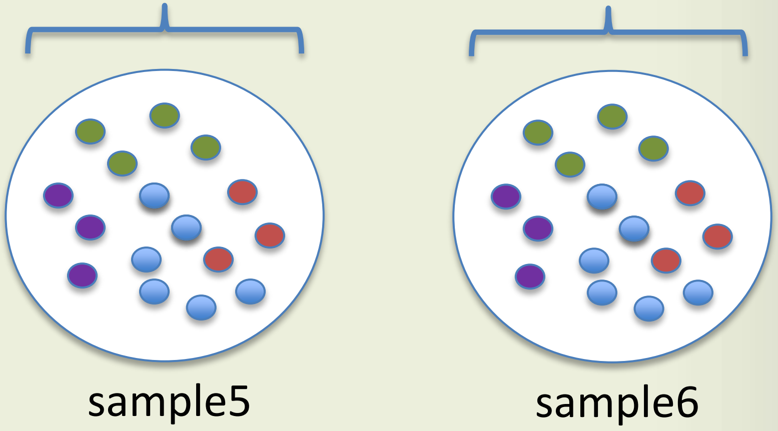
Alpha diversity

Phylogenetic relationship ??



All 4 belong to genus Streptococcus

3 belongs to Streptococcus
1 belong to Lactobacillus



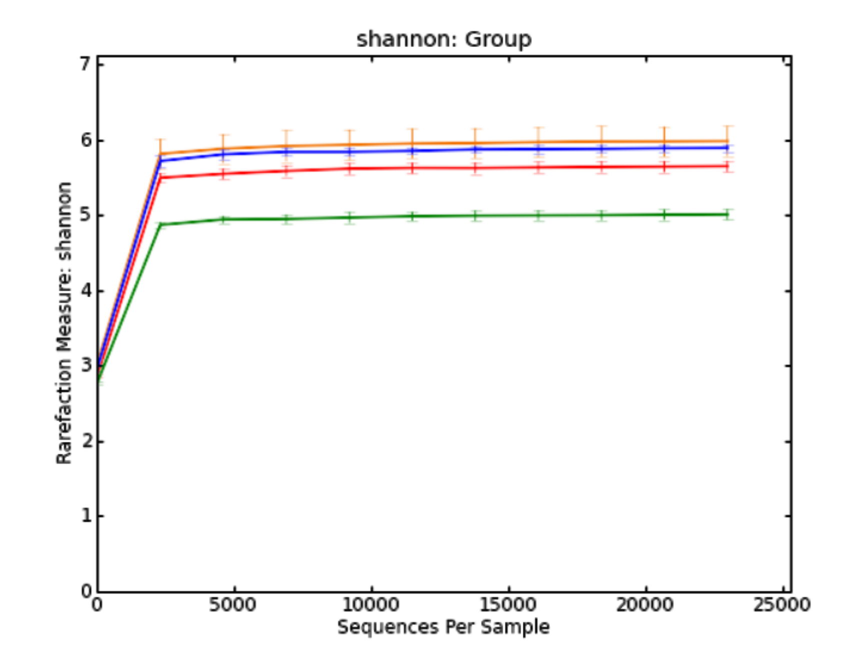
Commonly used diversity metrics

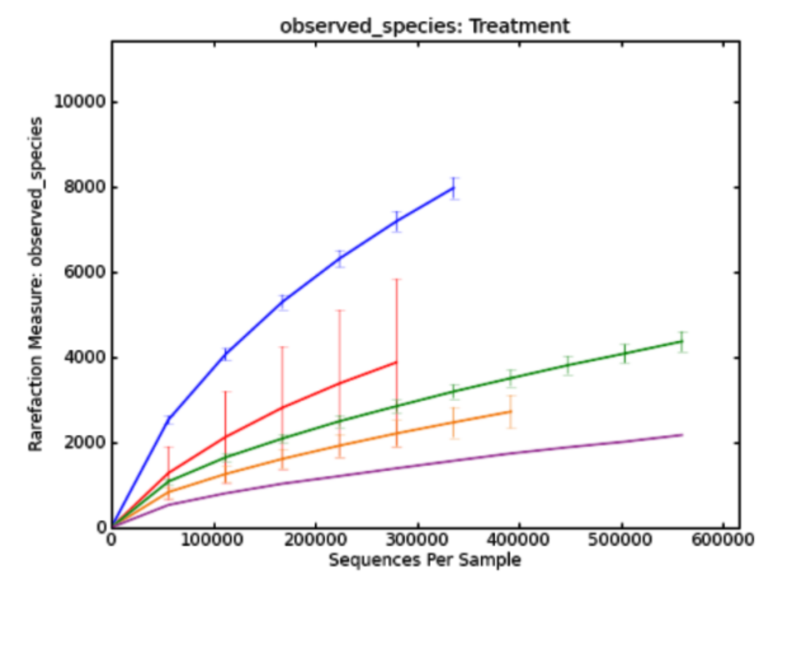
- Observed_species (measure richness only)
- Chao1 (measures richness and evenness)
- **Shannon** (measures richness and evenness)
- **Simpson** (measures richness and evenness)
- PD_whole_tree (includes phylogeny).

Alpha diversity – Rarefaction plot

Have you enough sequences to calculate the alpha diversity?
ANS : Take random subsample 10%, 20%, 30% …100% and calculate alpha diversity.

Rarefaction plot Rarefaction curve plots the number of individuals sampled versus the number of species.





THE UNIVERSITY OF ALABAMA AT BIRMINGHAM

Beta diversity

What is Beta diversity - It is a term for the comparison of samples to each other. . Beta diversity provides a measure of the **distance** or dissimilarity between each sample **pair**.

When more than two samples are used, the beta diversity is calculated for every pair of samples to generate a **distance/dissimilarity matrix**. A few of the commonly used beta diversity metrics are:

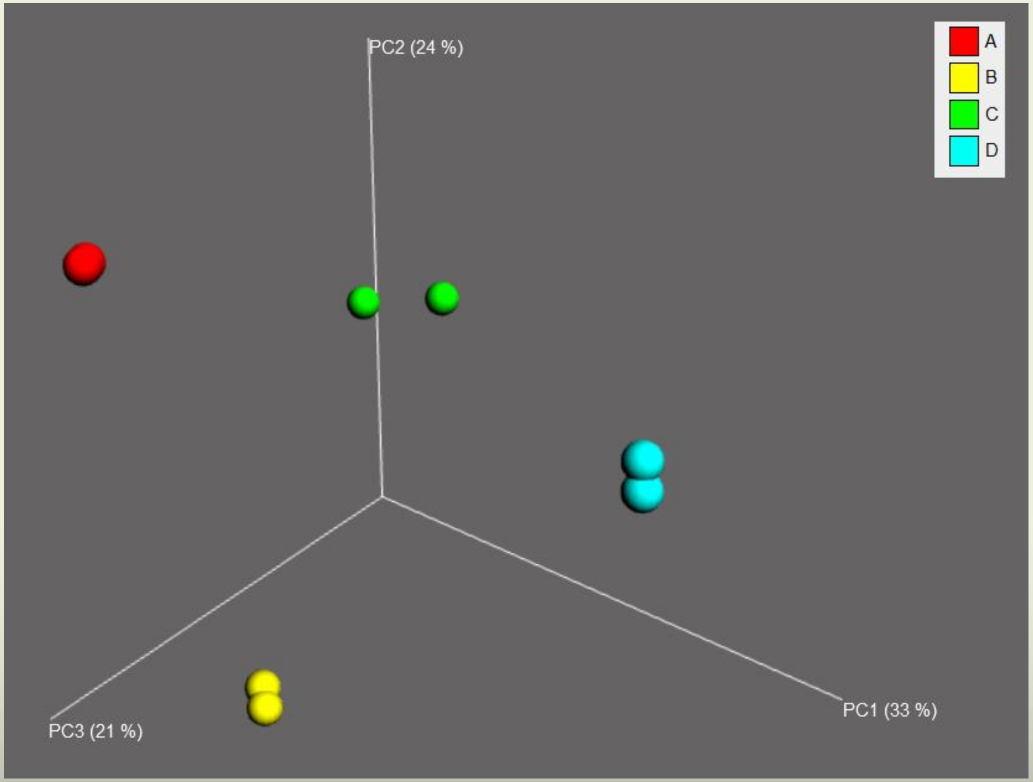
- **Bray-Curtis:** Non-phylogeny based method that takes abundance into account;
- **Un-weighted UniFrac**: Uses the presence and absence of OTUs and phylogeny
- **Weighted UniFrac:** Uses the abundance information of OTUs and phylogeny.

	A1	**A2**	**B1**	**B2**	**C1**	**C2**	**D1**	**D2**
A1	0.00	0.55	0.66	0.66	0.67	0.63	0.72	0.68
A2	0.55	0.00	0.66	0.65	0.67	0.63	0.72	0.68
B1	0.66	0.66	0.00	0.59	0.67	0.66	0.71	0.69
B2	0.66	0.65	0.59	0.00	0.68	0.64	0.71	0.68
C1	0.67	0.67	0.67	0.68	0.00	0.60	0.71	0.68
C2	0.63	0.63	0.66	0.64	0.60	0.00	0.72	0.68
D1	0.72	0.72	0.71	0.71	0.71	0.72	0.00	0.60
D2	0.68	0.68	0.69	0.68	0.68	0.68	0.60	0.00

beta div using unweighted unifrac

Principal Coordinate analysis plot (PCoA)

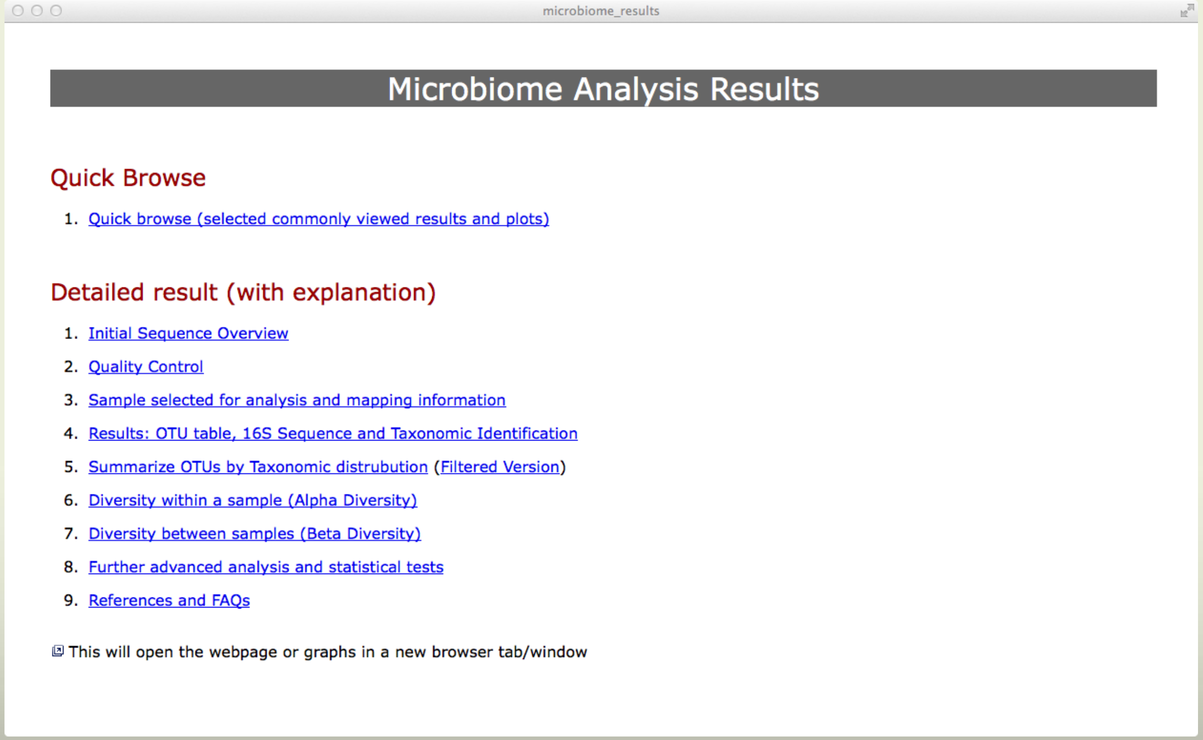
Principal Coordinates Analysis (PCoA) can be used for visualization of the data present in the beta diversity distance matrix in the form of 2-Dimensional or 3-Dimentional plots known as PCoA plots. PCoA transforms the distance matrix into a **new set of orthogonal axes** where the first axis (usually called **PC1**) can be used to explain the maximum amount of variation present in the dataset, followed by the second axis (**PC2**), and so on.



PCoA Plot

Microbiome analysis pipeline: QWRAP

- Code on Github : https://github.com/QWRAP/QWRAP
- Available on Cheaha cluster.
- Browse Live : A copy of example dataset is also available at https://dl.dropboxusercontent.com/u/428435/QWRAP/ANALYSIS/microbiome_report.html



microbiome_results

Microbiome Analysis Results

Quick Browse

1. Quick browse (selected commonly viewed results and plots)

Detailed result (with explanation)

1. Initial Sequence Overview
2. Quality Control
3. Sample selected for analysis and mapping information
4. Results: OTU table, 16S Sequence and Taxonomic Identification
5. Summarize OTUs by Taxonomic distrubution (Filtered Version)
6. Diversity within a sample (Alpha Diversity)
7. Diversity between samples (Beta Diversity)
8. Further advanced analysis and statistical tests
9. References and FAQs

⧉ This will open the webpage or graphs in a new browser tab/window

More about 16S rDNA sequencing

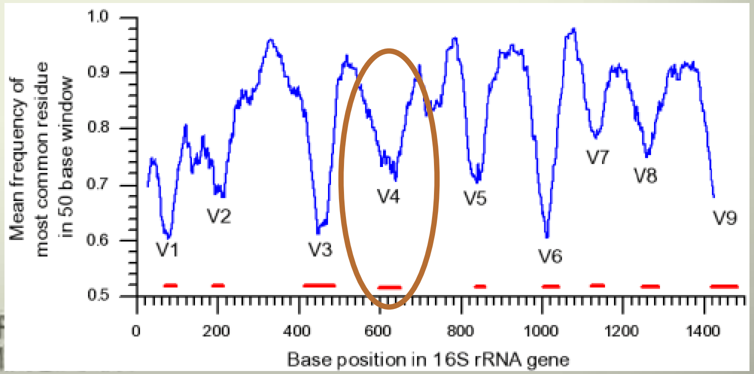
- Qualitative or Quantitative : 16S Microbiome analysis is Qualitative in nature

- Taxonomy resolution – Family / Genus /Species ?
  Depends on uniqueness of variable region

  -> Choice of 16S variable region : V4 or V3-V4 or V1-V3 or V6 etc.

  -> Length of variable region – 100/250/500 bases ?

- **V4** regions **250 bases** can provide resolution at **Genus** level (85%).



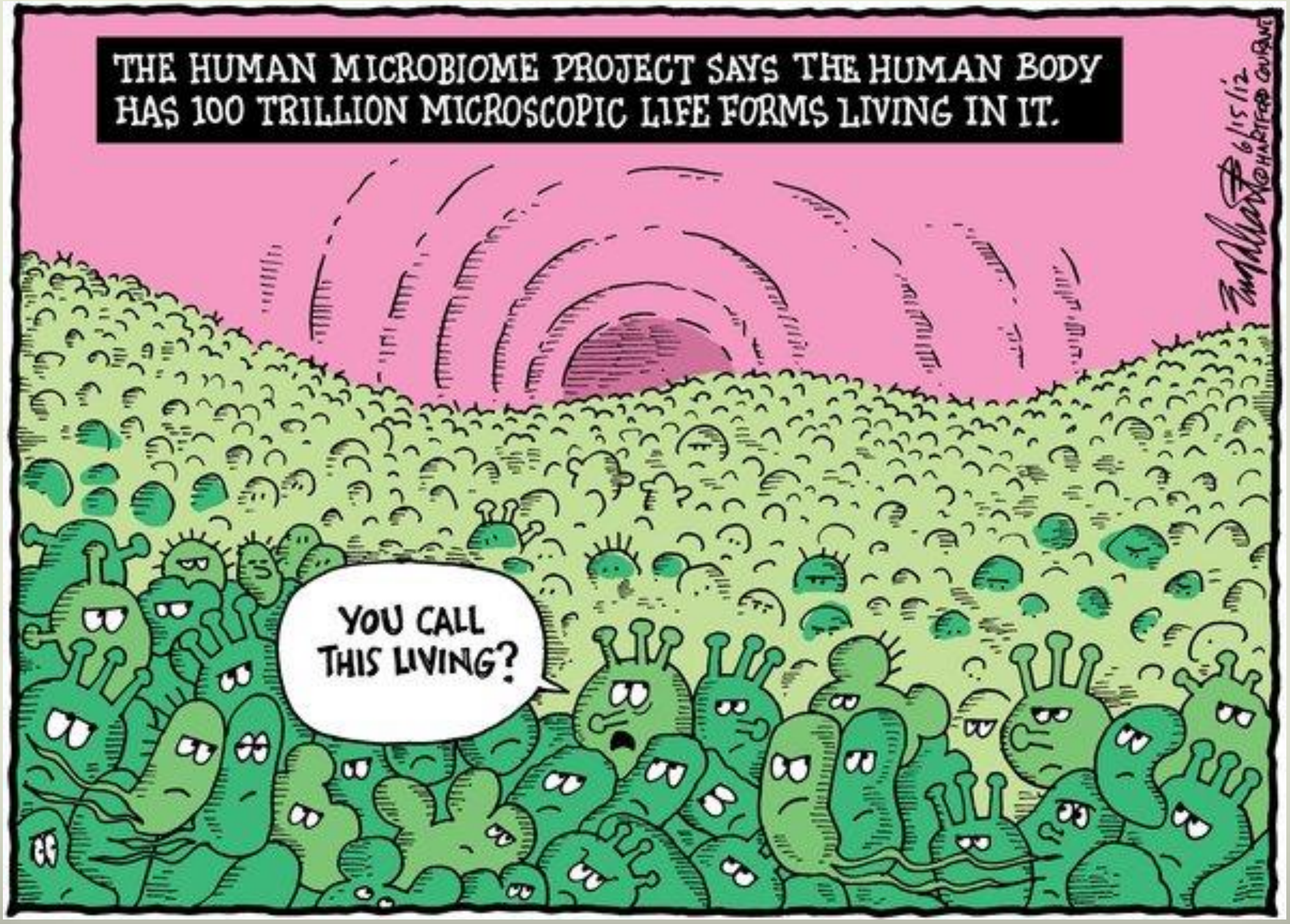
Statistics

- Differences in taxa at various levels of taxonomy?
- What are the top most OTUs (or species) ?
- What are the rare OTUs present?
- OTU correlation : Is there a correlation exists between OTUs and other attributes of sample like pH or other environmental conditions.

- Sample Size / Power Calculation
- Differences between 2 or more groups : T-test /ANOVA/PERMANOVA
- Multiple testing Correction
- Correlation with other attributes

- Graphs : Barplots, Heatmaps, PCoA plots, Diversity plots etc.

Thank You or Questions


THE HUMAN MICROBIOME PROJECT SAYS THE HUMAN BODY HAS 100 TRILLION MICROSCOPIC LIFE FORMS LIVING IN IT.
YOU CALL THIS LIVING?
6/15/12
©HARTFORD COURANT