QIIME2 Microbiome Analysis

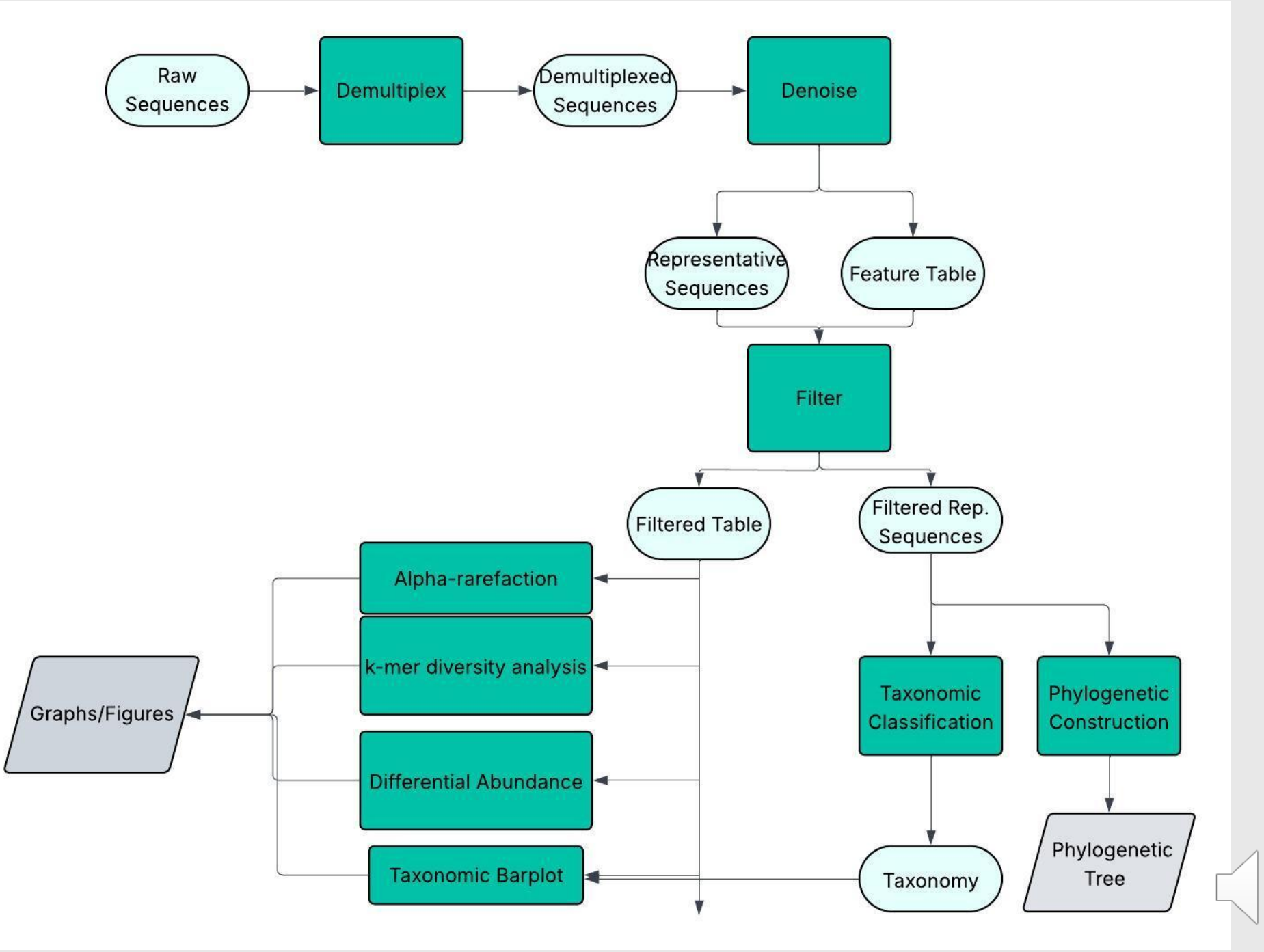
Bioinformatic Pipeline

Kayla Royce and Julia Murray

Background Information

- Data Sampling:
  - 20 samples
    - 2 pond locations
    - 2 treatments (duckweed vs pond water)
    - 5 replicates of each treatment
  - **manifest.tsv**: file containing locations of fastQ files
  - **metadata.tsv**: file containing information regarding identification of samples
- Microbiome Data Format:
  - Illumina HiSeq 2500
  - 250 base-pair, paired-end reads
  - Bacterial 16s rRNA
    - amplified 16S V4 region
    - 515f-806r barcodes
- **Goal**: Determine taxonomic differences between the two treatments and two locations

Methods


Raw Sequences
Demultiplex
Demultiplexed Sequences
Denoise
Representative Sequences
Feature Table
Filter
Filtered Table
Filtered Rep. Sequences
Alpha-rarefaction
k-mer diversity analysis
Differential Abundance
Taxonomic Barplot
Graphs/Figures
Taxonomic Classification
Phylogenetic Construction
Taxonomy
Phylogenetic Tree

Methods

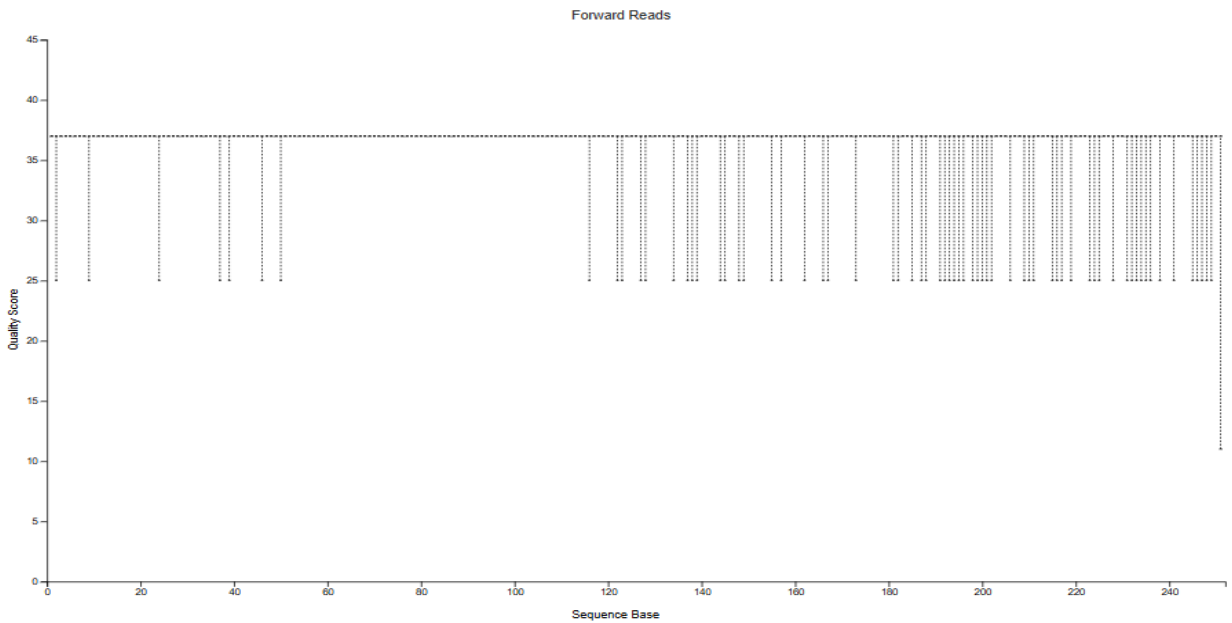
```
source activate qiime-2-amplicon-2024.5
```

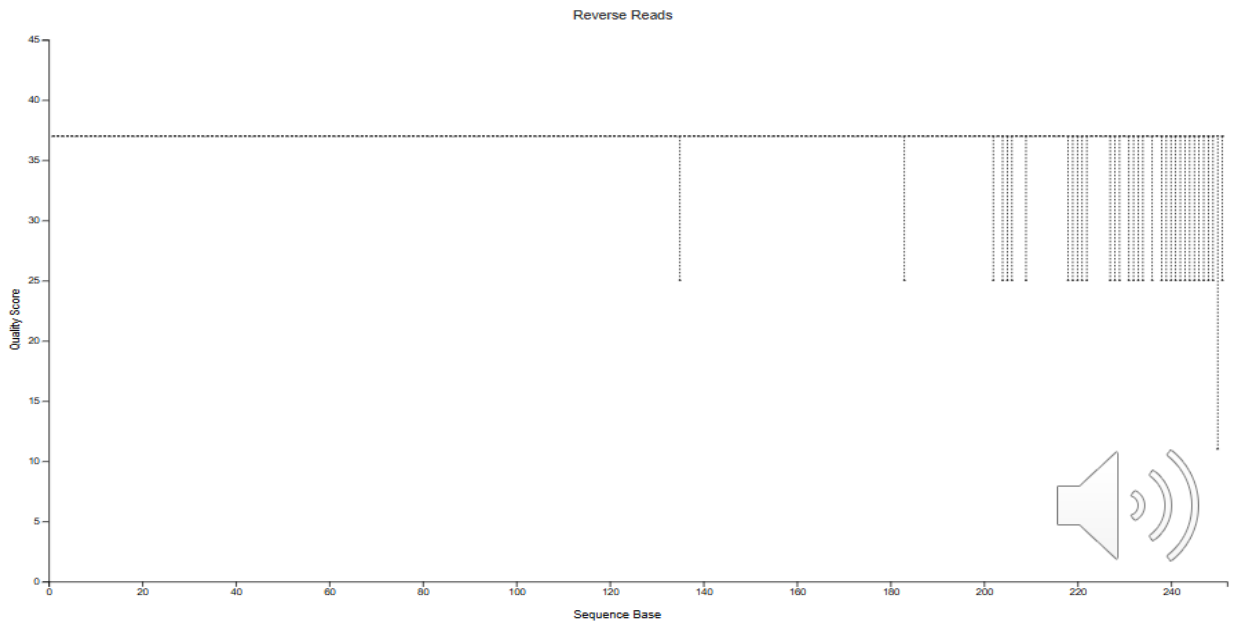
Importing Data

- `cp` manifest.tsv and metadata.tsv from /tmp
  - Already demultiplexed

Denoising Prep

- `demux summarize` converts demux.qza into qzv file
  - Used to determine where to denoise data





Methods (Upstream Analysis)

Denoising Data

- `dada2 denoise-paired` used to denoise data
  - forward reads trimmed at 220 bases
  - reverse reads trimmed at 200 bases
- `metadata tabulate` generates QIIME2 visualization of denoised data including feature IDs, sequence, and their counts
  - used to determine where to filter samples
- `tools export` used to export ASV representative sequences in BLAST-able file

Methods (Upstream Analysis)

Filtering

- `feature-table filter-samples`
  - removes samples with less than 1000 reads
  - Sample ODR-3-3 removed due to having 0 reads
- `feature-table summarize-plus`
  - summarizes the filtered ASV feature table with metadata information
- `feature-table tabulate-seqs`
  - creates compiled table of all ASV sequences and their frequency data
- `feature-table filter-features`
  - filters the feature table so all features must be present in a minimum of 25% of the samples
- `feature-table filter-seqs`
  - filters the ASV representative sequences to match those in the feature table
- `feature-table summarize-plus`
  - creates visualization of the filtered feature table

Methods (Upstream Analysis)

Training Classifier

- Classifier used: `wget -O silva-138-99-seqs.qza` and `wget -O silva-138-99-tax.qza`
- `feature-classifier extract reads` filters the classifier for specific primer sequences
  - Forward primer: GTGCCAGCMGCCGCGGTAA
  - Reverse primer: GGACTACHVGGGTWTCTAAT
- `feature-classifier fit-classifier-naive-bayes` trains custom classifier using the previously filtered reference sequences and the taxonomic classifier

Taxonomic Classification

- `feature-classifier classify-sklearn` assigns taxonomy to samples using the custom trained classifier
- `feature-table tabulate-seqs` visualizes ASV sequences into feature table with taxonomic information

Methods (Upstream Analysis)

Feature Table in QIIME2 View:

	Frequency	# of Samples Observed In
c721b4c609340cf459a6af3e02e5b7e5	5,690	14
6b3fd25486c30de6f3c9624143517c86	4,239	11
3bf5c259415e62364aea22cdf05c4933	3,713	11
67e3716aa0b883dd4e929e7a315ca0e1	3,622	12
fc011ad9c9a7e56a70e2e0a7418a33c5	2,488	12
3af48669bd795a4f010caf3d114f59ac	2,137	11
5287059f9ecee411ef4a8714c69cd60f	1,863	10

Taxonomic Classification in QIIME2 View:

Feature ID ▼	Sequence Length	Taxon: 0	Frequency	No. of Samples Observed In
fc2307d3001a3b10f17ce7	407	d__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__Mitochondria; g__Mitochondria	1,049.0	10

Methods (Upstream Analysis)

Phylogenetic Tree Construction

- `phylogeny align-to-tree-mafft-fasttree` aligns the features in feature table and creates a rooted tree for phylogenetic tree construction
- while loop to create "`itol.txt`" : file with node IDs and assigned genus and species
- Upload "`rooted_tree.qza`" and "`itol.txt`" (node labels) to iTOL for phylogenetic tree

```
LABELS
SEPARATOR COMMA

DATA
6ea7f755987885961e603405bf352d6b,Phacus paraorbicularis
a66c20f220d4270e4c7e540cd6f5e71d,Emticicia sediminis
8f493c1143a685f13d2315be3a19507a,Monodopsis sp.
5287059f9ecee411ef4a8714c69cd60f,Flavobacterium cheonhonense
3d7825c893a6aa3d7eb33ce6fa97905f,Raphidocelis subcapitata
```

Methods (Downstream Analysis)

K-mer based diversity analysis

- `conda activate q2-boots-amplicon-2025.4` activates QIIME2 environment boots kmer-diversity commands
- `boots kmer-diversity` computes k-mer based diversity metrics to avoid bias from taxonomic assignment

Alpha-rarefaction plot

- `diversity alpha-rarefaction` shows if selected sequencing depth contains majority of the species present

Taxonomic Bar-plot

- `taxa barplot` shows taxonomic composition and relative abundance for each sample type

Methods (Downstream Analysis)

Differential Abundance

- `feature-table filter-samples` filters features to compare duckweed and water samples
- `taxa collapse` collapses ASVs into species-level taxonomy (level 7)
- `composition ancombc` performs ANCOM-BC testing to identify significantly different species-level taxa across sample types
- `composition da-barplot` visualizes results of ANCOM-BC analysis with significance threshold of 0.001

Results:
Pond Microbiome Metadata (Downstream Analysis)

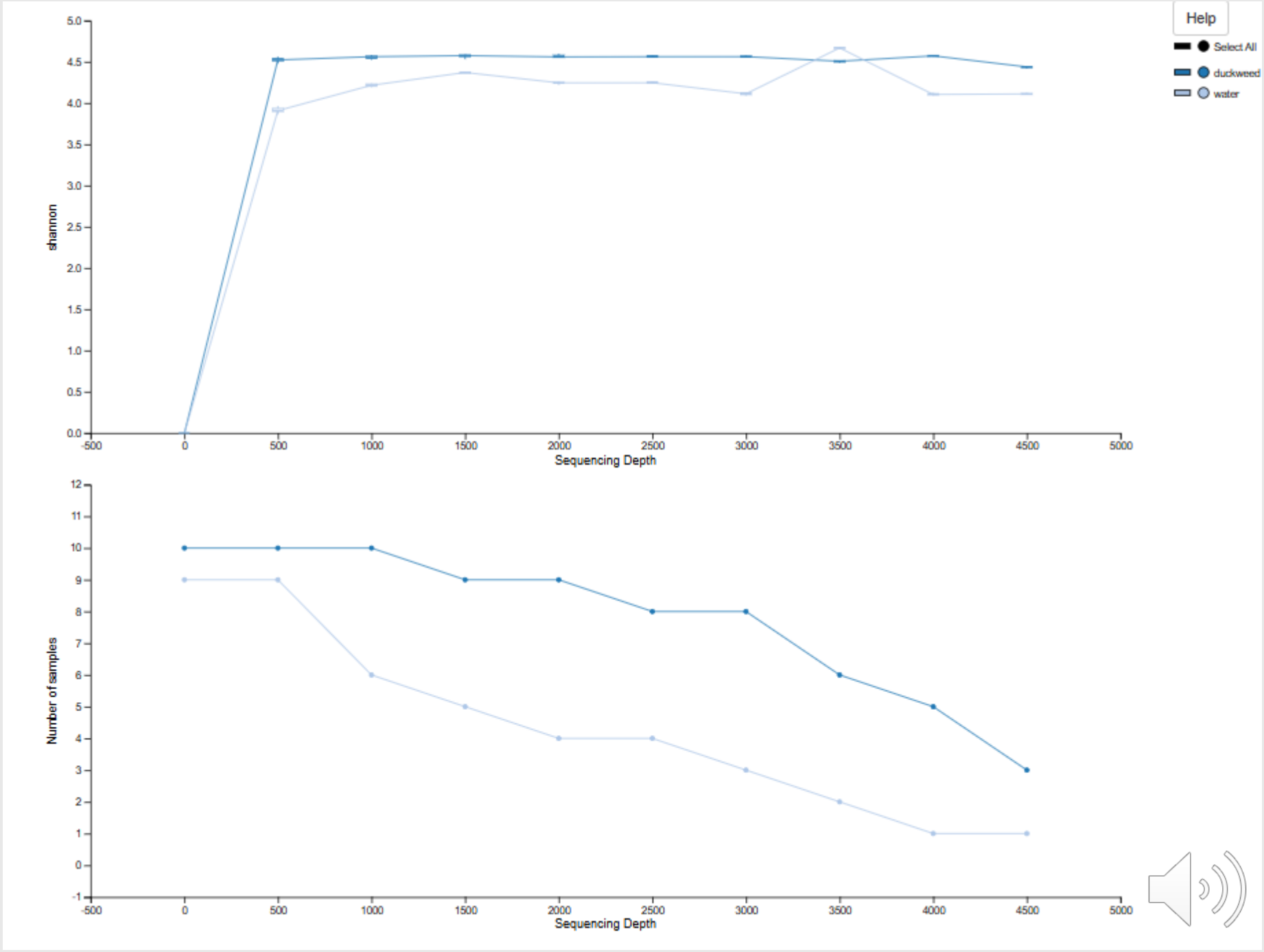
- A QIIME 2 metadata file that describes our microbiome sequencing samples
- **sampleid**: Unique identifier for each sample (required by QIIME 2).
- **sample_type**: Either "water" or "duckweed", indicating what was sampled.
- **sub_location**: Either "pond_2" or "pond_3" — different locations sites.
- **replicate**: Numbers 1 through 5 indicating replicates.
- 20 total samples:
  - 10 water samples (5 each from pond_2 and pond_3)
  - 10 duckweed samples (also 5 each from pond_2 and pond_3).

sampleid	sample_type	sub_location	replicate
#q2:types	categorical	categorical	numeric
ODR-2-1	water	pond_2	1
ODR-2-2	water	pond_2	2
ODR-2-3	water	pond_2	3
ODR-2-4	water	pond_2	4
ODR-2-5	water	pond_2	5
ODR-3-1	water	pond_3	1
ODR-3-2	water	pond_3	2
ODR-3-3	water	pond_3	3
ODR-3-4	water	pond_3	4
ODR-3-5	water	pond_3	5
ODR-2-1-DW	duckweed	pond_2	1
ODR-2-2-DW	duckweed	pond_2	2
ODR-2-3-DW	duckweed	pond_2	3
ODR-2-4-DW	duckweed	pond_2	4
ODR-2-5-DW	duckweed	pond_2	5
ODR-3-1-DW	duckweed	pond_3	1
ODR-3-2-DW	duckweed	pond_3	2
ODR-3-3-DW	duckweed	pond_3	3
ODR-3-4-DW	duckweed	pond_3	4
ODR-3-5-DW	duckweed	pond_3	5

Results:

Alpha-Rarefaction Plot

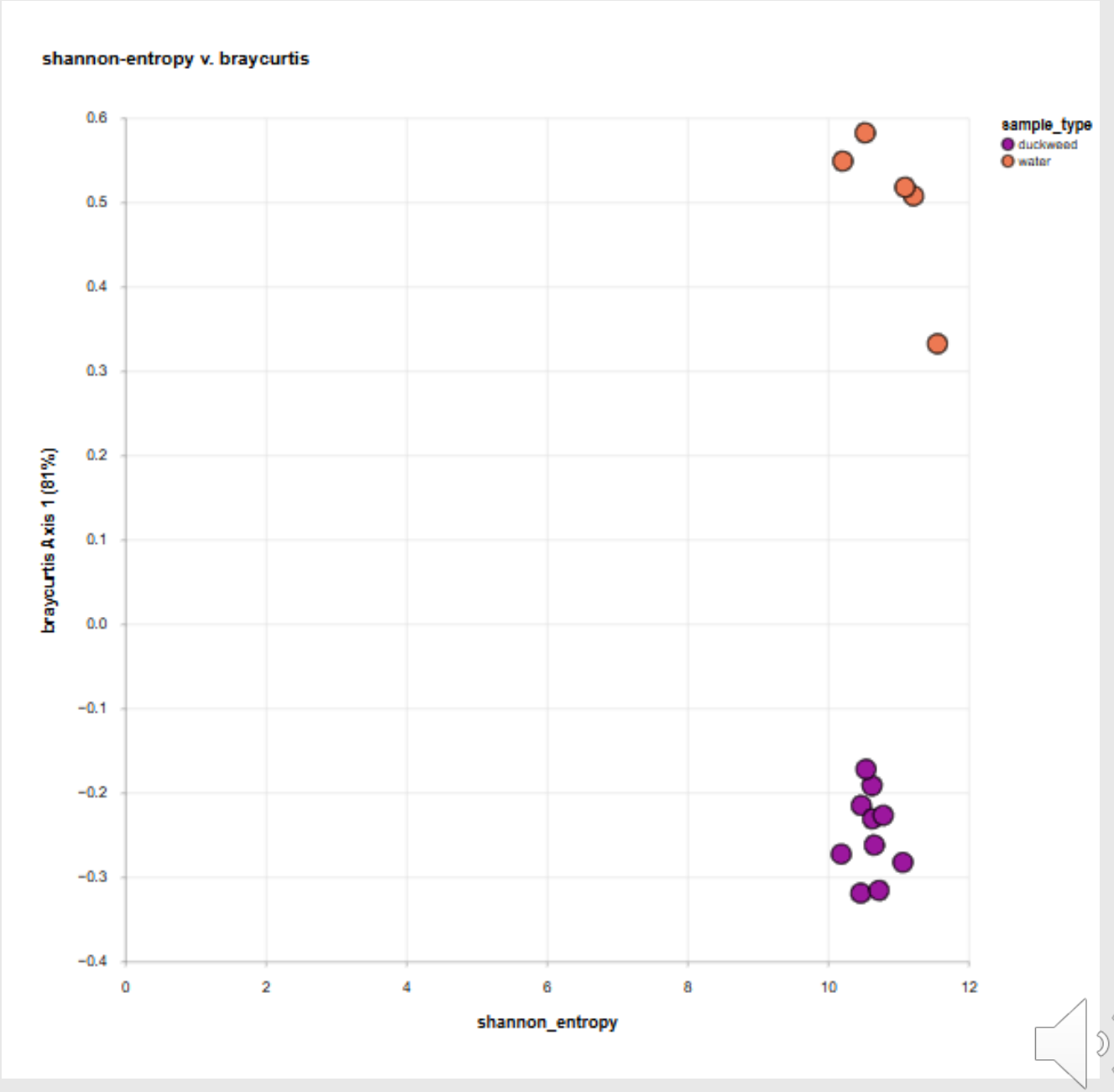
- **Top Graph:** Duckweed samples have higher Shannon index (Y-axis) = greater microbial diversity
  - Plateau = most diversity and captured
- **Bottom Graph**: Duckweed samples maintain higher retention across depths
  - As depth increases, fewer samples meet depth threshold



Results:

Shannon Entropy (alpha) v. Braycurtis (beta) Graph

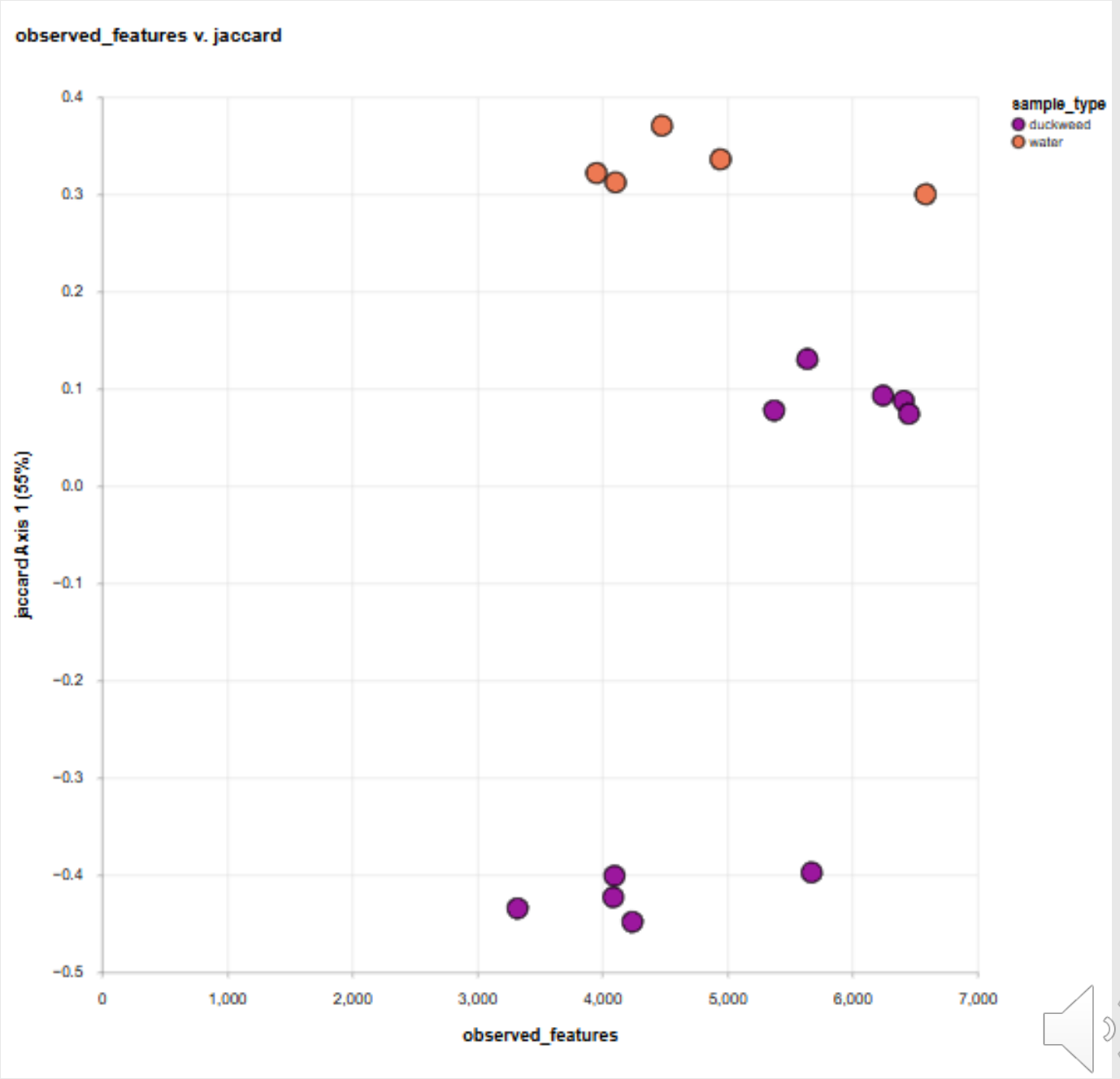
- Most informative
- **Water**: less diverse per sample but compositionally distinct
- **Duckweed**: more diverse within each sample and form distinct group
  - Rich and consistent microbial community



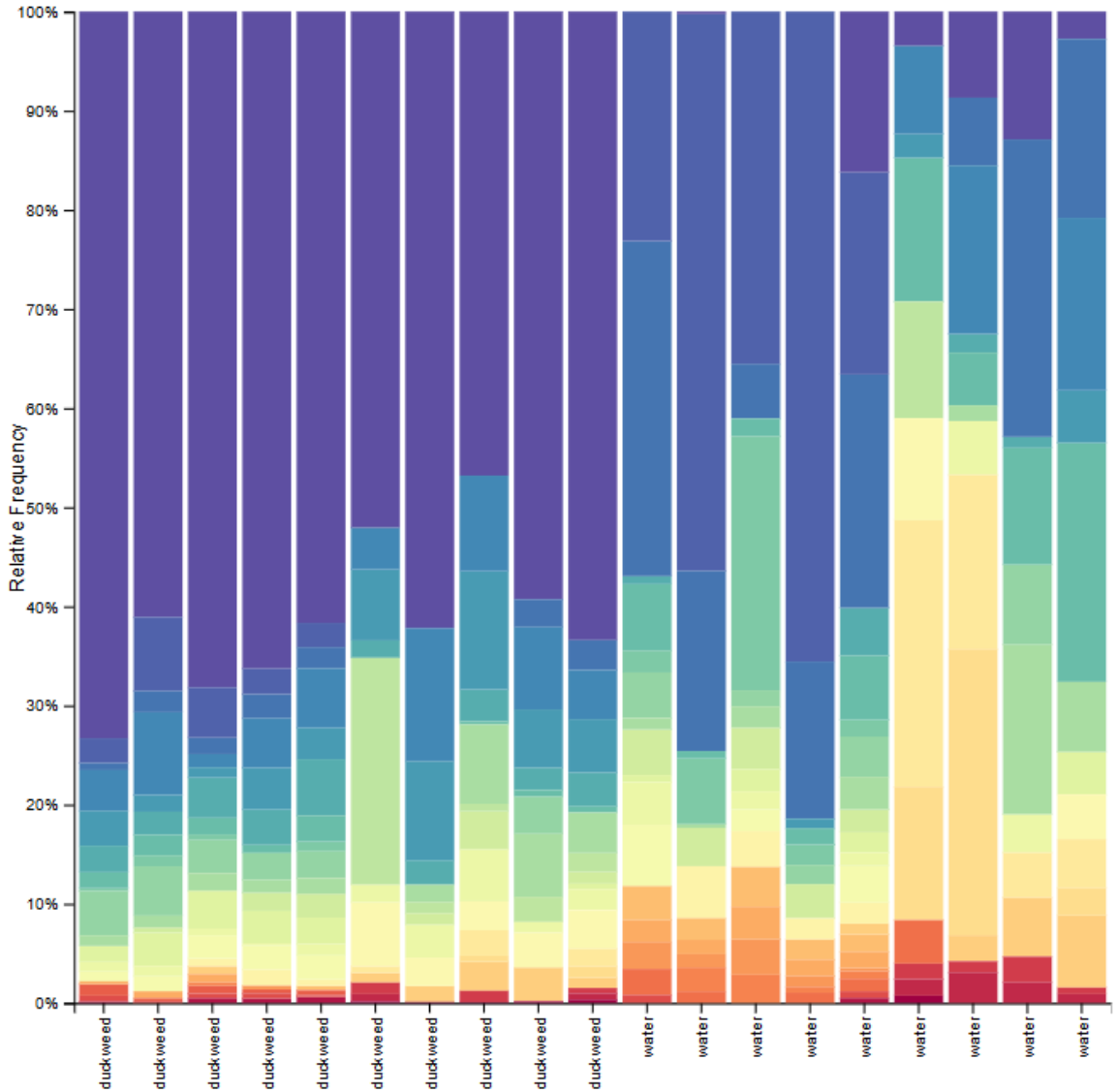
Results:

Observed Features (alpha) v. Jaccard (beta)

- **Observed Features:** richness
- **Jaccard: pr**esence/absence
- **Water:** Richer in taxa count = more observed features
  - Two groups: intra-group variation (ex. location of sample)
- **Duckweed:** more variable in richness (some samples have fewer taxa)

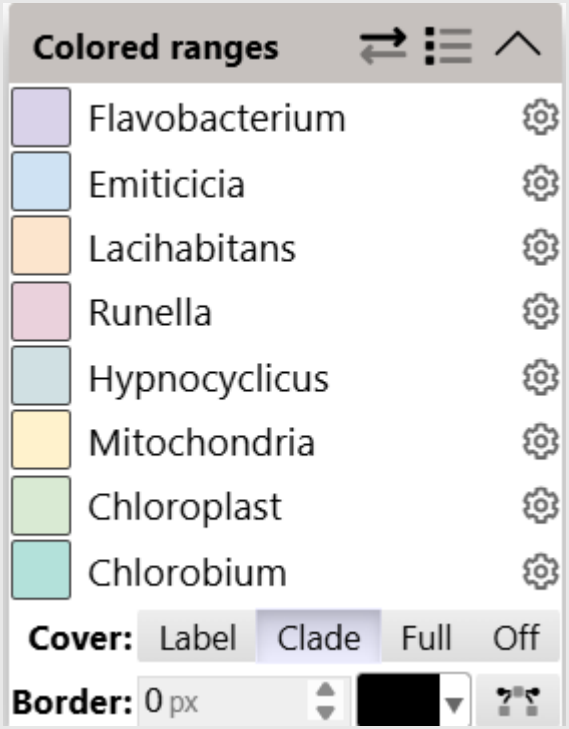


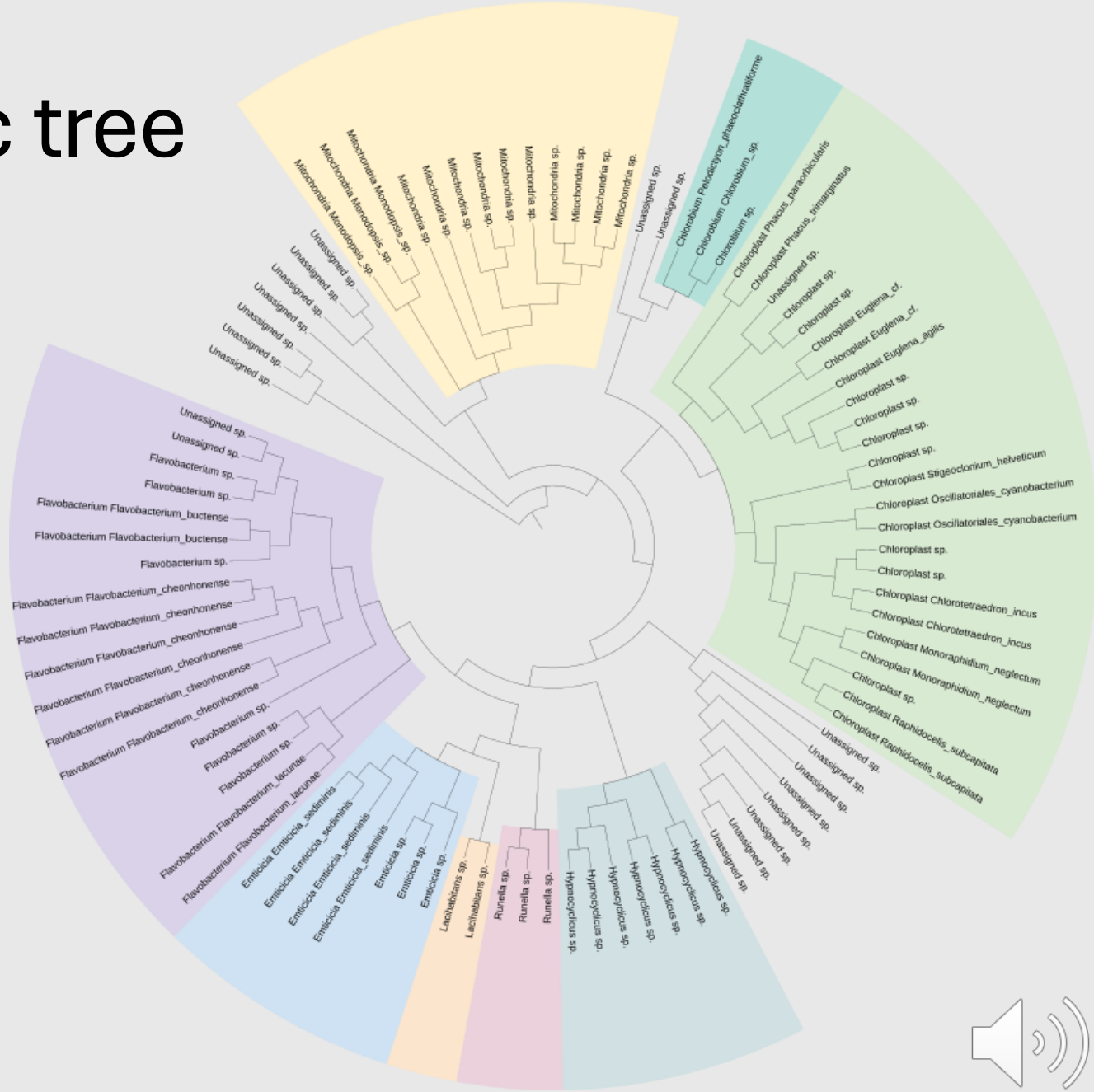
Results: Taxonomic Bar Plot



d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__Mitochondria;g__Mitochondria;__
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium;s__Flavobacterium_cheonhonense
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;__
d__Bacteria;p__Fusobacteriota;c__Fusobacteriia;o__Fusobacteriales;f__Leptotrichiaceae;g__Hypnocyclicus;__
d__Bacteria;p__Firmicutes;c__Clostridia;o__Peptostreptococcales-Tissierellales;__;__;__
Unassigned;__;__;__;__;__;__
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium;__
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Euglena_cf.
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;__;__;__
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Emticicia;s__Emticicia_sediminis
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium;s__Flavobacterium_buctense
d__Bacteria;__;__;__;__;__;__
d__Bacteria;p__Proteobacteria;c__Alphaproteobacteria;o__Rickettsiales;f__Mitochondria;g__Mitochondria;s__Monodopsis_sp.
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Runella;__
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Chlorotetraedron_incus
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Lacihabitans;__
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Monoraphidium_neglectum
d__Bacteria;p__Bacteroidota;c__Chlorobia;o__Chlorobiales;f__Chlorobiaceae;g__Chlorobium;s__Pelodictyon_phaeoclathratiforme
d__Bacteria;p__Bacteroidota;c__Chlorobia;o__Chlorobiales;f__Chlorobiaceae;__;__
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Flavobacteriales;f__Flavobacteriaceae;g__Flavobacterium;s__Flavobacterium_lacunae
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Raphidocelis_subcapitata
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Oscillatoriales_cyanobacterium
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Phacus_trimarginatus
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Phacus_paraorbicularis
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;__;__;__;__
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Stigeoclonium_helveticum
d__Bacteria;p__Bacteroidota;c__Bacteroidia;o__Cytophagales;f__Spirosomaceae;g__Emticicia;__
d__Bacteria;p__Bacteroidota;c__Chlorobia;o__Chlorobiales;f__Chlorobiaceae;g__Chlorobium;__
d__Bacteria;p__Bacteroidota;c__Chlorobia;o__Chlorobiales;f__Chlorobiaceae;g__Chlorobium;s__Chlorobium_sp.
d__Bacteria;p__Cyanobacteria;c__Cyanobacteriia;o__Chloroplast;f__Chloroplast;g__Chloroplast;s__Euglena_agilis
d__Bacteria;p__Proteobacteria;c__Gammaproteobacteria;o__Burkholderiales;f__Comamonadaceae;__;__

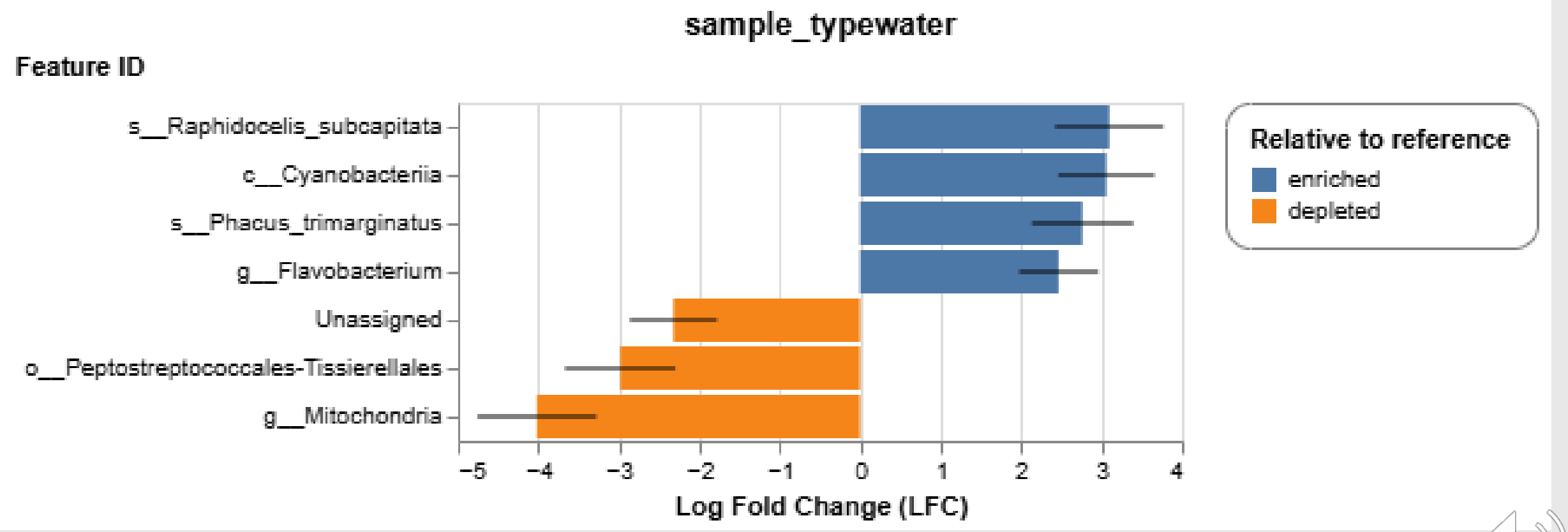
Results: phylogenetic tree





Results: Differential Abundance

Genus Ancombc



Thanks!