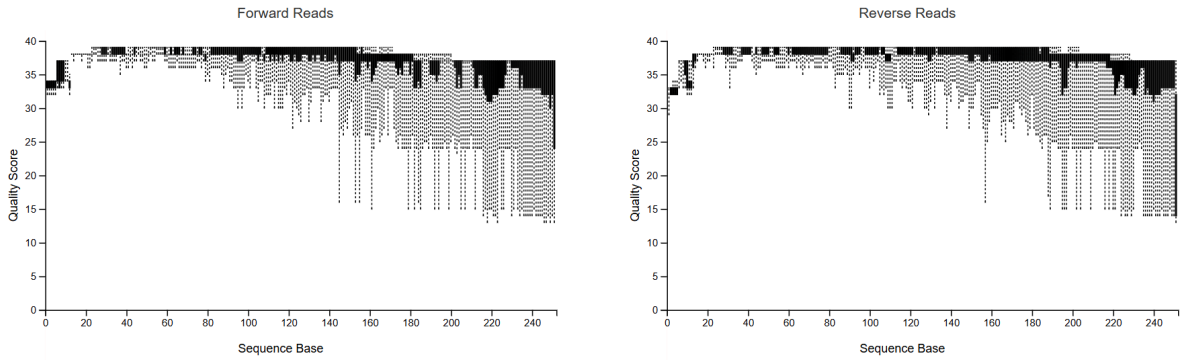
Isabella Fregoso
Proff. Van Laar
BIOL 4880-001
9 April 2025

Microbiome Analysis with qiime2

1. **Screenshot of demux.qzv visualization**



2. **Let's take a look at this file (taxonomy.qzv). What do you see if you filter by taxon? What is this file showing you?**

If I filter by taxon in order of decreasing, the "unassigned" taxa populates first. Then if I invert this in order of increasing, the Archaea populates first and then the rest that follows is various orders.

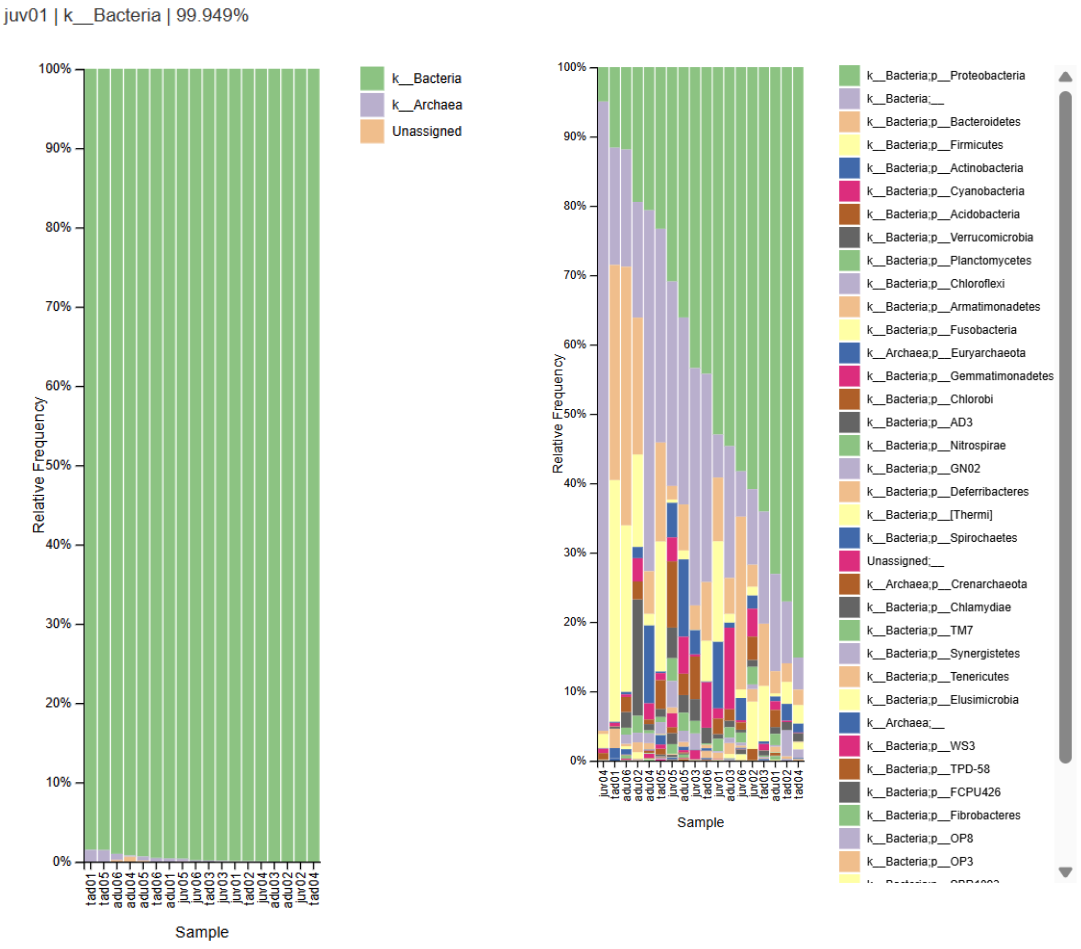
3. **Sometimes the file might have reads that match things other than Bacteria. Do you see that in your file? These are samples from frogs...what else, besides bacteria, could theoretically be amplified if you're amplifying 16S rRNA?**

I see types of archaea in the file. Theoretically if 16S rRNA is being amplified, I could see archaea and maybe even chloroplast.

4. **When you visualize the taxa-bar-plots.qzv file, what do you notice? How does changing the different levels change the visualization? Why do you think this is? When you sort the samples by life stage or site of collection, do you notice any trends?**

Changing the levels allows for more visualization of different bacteria's taxonomy. A more individual array of data is shown as the levels increase because it is focused on a specific bacterial species not just a domain. When sorted by life stage the colored bars rearrange themselves and are grouped based on the stages of the bacterial species. I notice the trend is that the big bars are organized toward the top of the bars while the shorter ones are at the bottom.

LEVEL 1 LEVEL 2



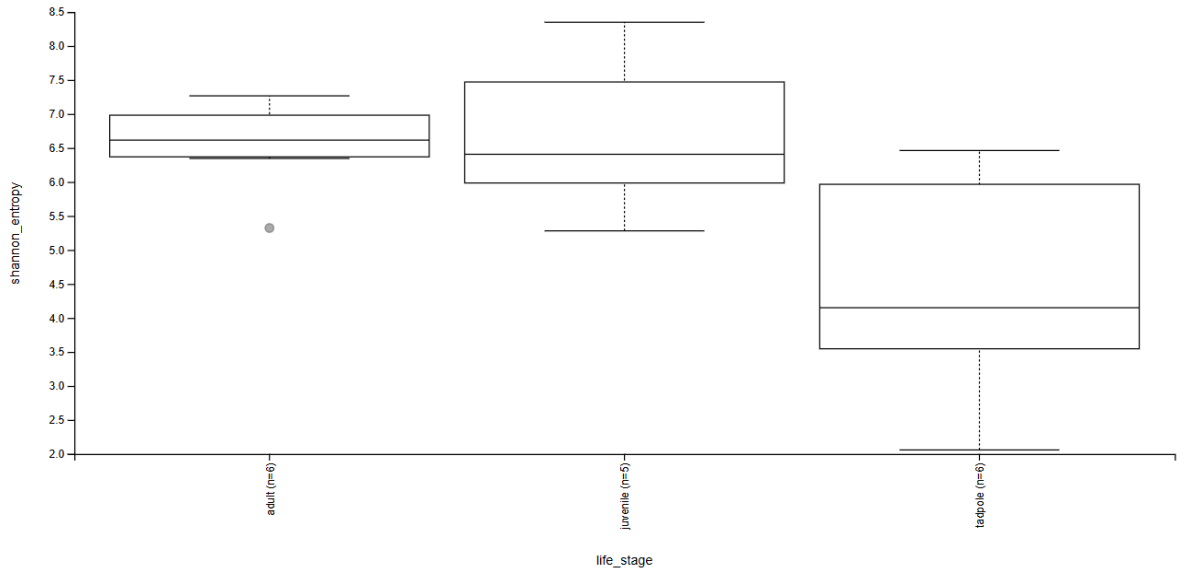
5. **In your own words, summarize the briefly what each of these metrics is doing and how they are different.**

The Shannon metrics measure diversity with an analysis on the statistics of the genomic data. It analyzes the number of species and the richness. The Observed feature focuses more on the number of features that are seen within the sample, and doesn't really pay attention to the richness.

6. **You will need to view these new .qzv files for alpha diversity. The boxplots can be viewed for diversity when comparing site of capture or life stage of the frog. Below the boxplots, you will see statistical tables for that analysis. If the p value is <u>< 0.05 there is a significant difference somewhere</u>. The table below that is a pairwise comparison which will show you which comparison is giving the significant difference. Take a screenshot of your Observed Features for life stage and Shannon for site of collection. Were any comparisons significant for any metric? If so, which ones?**

P value for the Shannon site of collection is lower than 0.05 at a value of 0.037 and 0.067 for the tadpoles (both adult and juvenile). There is also a significant difference in the Observed Features since the p values of the tadpoles (both adult and juvenile) are below 0.05 with a value of 0.01 and 0.028.

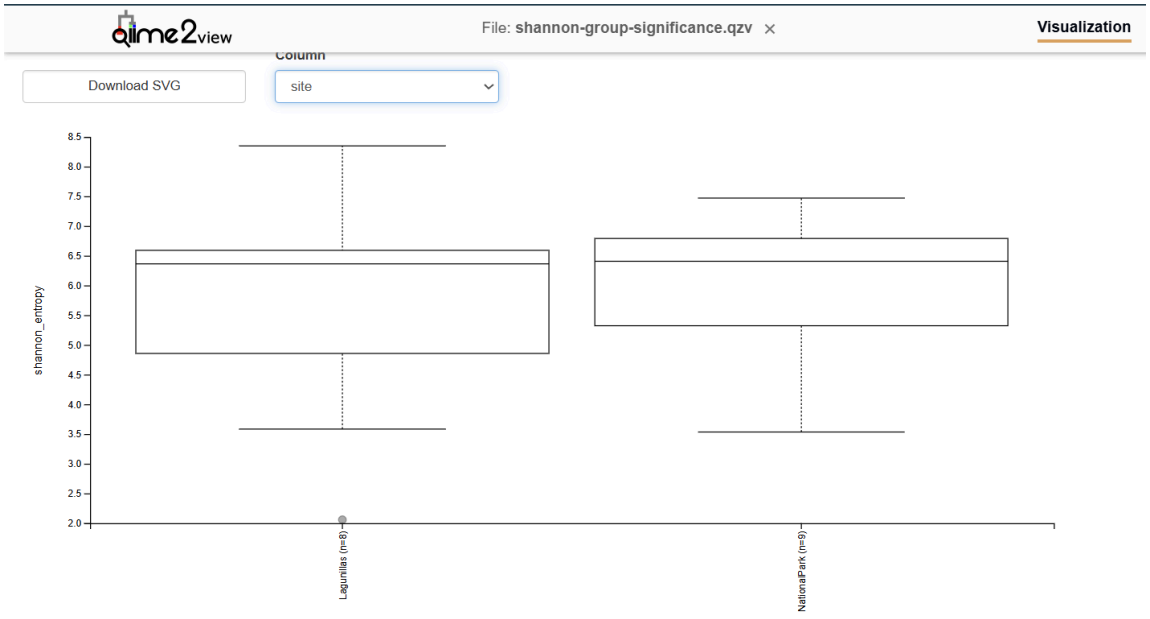
<u>Observed features- life stage</u>



Download CSV

		H	p-value	q-value
Group 1	Group 2			
adult (n=6)	juvenile (n=5)	0.000000	1.000000	1.000000
	tadpole (n=6)	6.564103	0.010406	0.031217
juvenile (n=5)	tadpole (n=6)	4.800000	0.028460	0.042690

Shannon- site of collection



Download CS

		H	p-value	q-value
Group 1	**Group 2**			
adult (n=6)	**juvenile (n=5)**	0.000000	1.000000	1.000000
	tadpole (n=6)	4.333333	0.037373	0.101834
juvenile (n=5)	**tadpole (n=6)**	3.333333	0.067889	0.101834

7. **We can now look at beta diversity. <u>How do alpha and beta diversity differ in what they are trying to tell you?</u> We can look at numerous beta diversity metrics output by QIIME2, but we will look specifically at Bray Curtis distance and Weighted Unifrac distance. We didn't talk about Weighted Unifrac, so you should read a little bit about what this metric does before <u>providing a brief explanation</u> about both Bray Curtis distance and Weighted Unifrac.**

Alpha and beta differ in the way that they measure sample population. Alpha diversity measures within a sample population and beta measures different samples and compares them together. Beta diversity is more of an assorted array of data. Bray Curtis and Weighted Unifrac are both beta diversity examples, but they do differ in what they analyze. Bray Curtis focuses on the abundance of species whereas Weighted Unifrac uses a phylogenetic tree to compare and consider the relatedness of the organisms.

8. **Here are the commands for visualizing the Bray Curtis data. How will you change these to look at the Weighted Unifrac?**

I could change the code for visualizing Bray Curtis data by adding a Weighted Unifrac command like this:

```
qiime diversity beta-group-significance \
--i-distance-matrix
diversity-metrics-results/weighted_unifrac_distance_matrix.qz
a \
--m-metadata-file metadata.txt \
--m-metadata-column life_stage \
--o-visualization
diversity-metrics-results/weighted-unifrac-distance-signifika
nce.qzv \
--p-pairwise
```

9. **For the pairwise comparison, we're actually looking for the q value, which is an adjust p value based on accounting for multiple comparisons. Do any life stages appear to have significantly different community composition based on either metric? Please include a screenshot of the table for both Bray Curtis and Weighted Unifrac. For the site comparison, we can just look at the p value. Do the sites differ in community composition?**

The q value from the Bray Curtis visualization is 0.173 for adult juveniles, 0.006 for adult tadpoles, and 0.042 for juvenile tadpoles.The life stages do have significantly different composition if the q value is <0.05. We can see that the adult tadpoles and juvenile tadpoles display significance based on the Bray Curtis metric. For the Weighted Unifrac, the q values are 0.502 for the adult juveniles, 0.160 for the adult tadpoles, and 0.066 for the juvenile tadpoles. Based on these values there are no significant differences in the community composition of the Weighted Unifrac.

Bray Curtis

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
adult	juvenile	11	999	1.208606	0.173	0.173
	tadpole	12	999	1.816888	0.002	0.006
juvenile	tadpole	11	999	1.633328	0.028	0.042

Weighted Unifrac

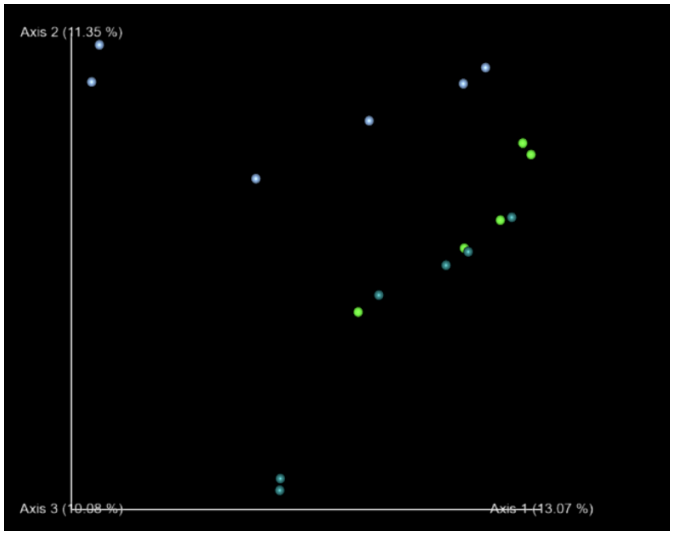
Download CSV

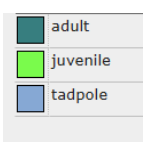
Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
adult	juvenile	11	999	0.928729	0.502	0.5020
	tadpole	12	999	1.705238	0.107	0.1605
juvenile	tadpole	11	999	2.427517	0.022	0.0660

10. **Once you have determined if there are any differences, you can view a plot of a Principle Components Analysis. This is basically a way to try and graphically represent community differences. You should view the Bray Curtis and Weighted Unifrac Emperor .qzv files. You can color code the individual points by site of collection or life stage. Do the points cluster like you would expect based on significance? For example, if you saw differences in life stage, do the various life stages seems to be close to one another (e.g., tadpole close to tadpole, but father from juvenile or adult). Include a screenshot of each Emperor plot (Bray Curtis and Weighted Unifrac) with the life stages color coded.**

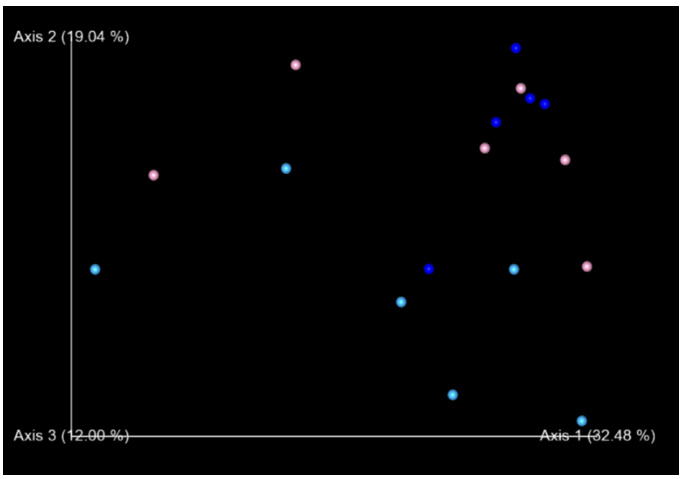
When I analyzed the p values for significance, I saw that there were no significant differences in community composition of the Weighted Unifrac. This is shown in the plot by the data points being more clustered together. The cluster refers to the similarity of the life stages. It makes sense that there are less clusters in the Bray Curtis metric since significant differences were seen.

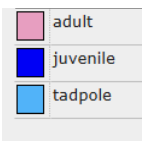
Bray Curtis



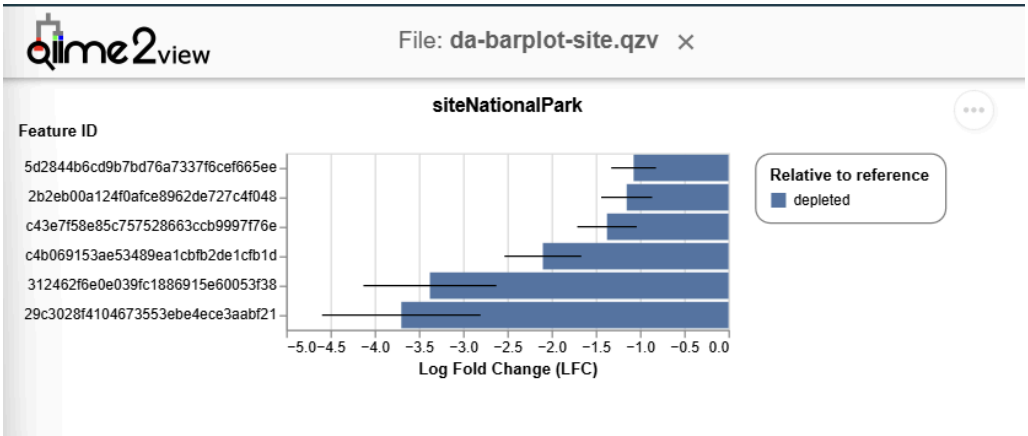


Weighted Unifrac





11. **If you find any that are differentially expressed, you can figure out what taxa they belong to by searching the taxonomy.qzv file for the alphanumeric code. Were there any differentially expressed taxa in the different sites? If so, provide at most 3 of them. You should repeat this analysis for the life stage as well.**



The alphanumeric code indicates for the bars (counting from the top down) 1,2, and 4 the taxa as follows :

1) k__Bacteria; p__Gemmatimonadetes; c__Gemmatimonadetes

2) k__Bacteria; p__Bacteroidetes; c__Flavobacteriia; o__Flavobacteriales; f__[Weeksellaceae]; g__Ornithobacterium; s__

4) k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Mycobacteriaceae; g__Mycobacterium; s__

<u>Life Stage Analysis</u>

The frequency and number of samples observed for each of the above codes are as follows.

	Taxa	*Freq.*	*Number of samples observed*
1	*k__Bacteria; p__Gemmatimonadetes; c__Gemmatimonadetes*	*11*	*2*
2	*k__Bacteria; p__Bacteroidetes; c__Flavobacteriia; o__Flavobacteriales; f__[Weeksellaceae]; g__Ornithobacterium; s__*	*16*	*2*
4	*k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales; f__Mycobacteriaceae; g__Mycobacterium; s__*	*76*	*5*