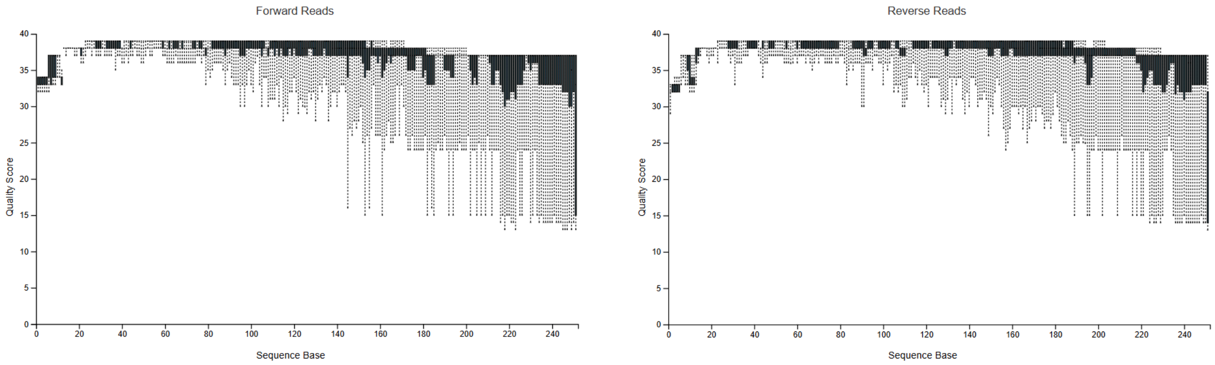


1)

2) For each data set, you will need to make a judgement call. For this data set, any samples with fewer than 10,000 reads will probably need to be cut. How many samples is that?

   No counts are under 10,000

3) What do you see if you filter by taxon? What is this file showing you?

   The filter is showing archaea species (taxonomy of each species) as well as the confidence intervals of each species listed.

4) 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?

The file shows different archaea spesies as well as some eukaryarchaea species intermingled in between. The archaea species may have the shared ability to amplify the 16sRNA.

5) 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?

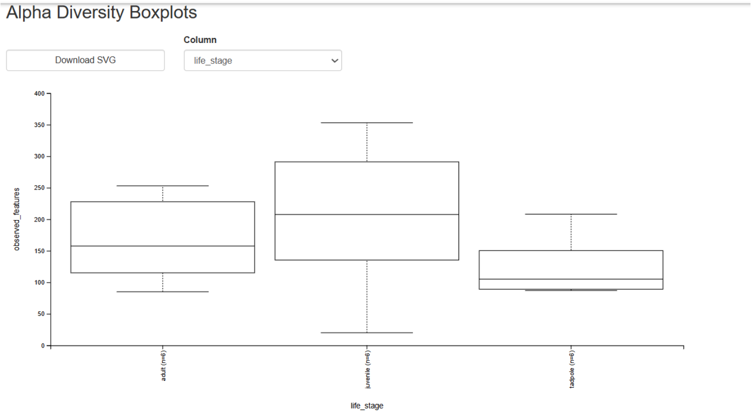
   I notice that the bar plots are all at 100%, denoting that all species are in the kingdom of bacteria. Clicking down to level 5, there are many more distinctions between all the different samples. This is due to the greater categorization of species. Once sorted by life stage, the most noticeable trend observed was the separation of adult, juvenile, and tadpoles in the samples.

6) The command above runs both alpha and beta diversity metrics. There are a number of each that are performed. We will take a look at 2 for alpha, observed features, and Shannon diversity. In lecture, we talked about Observed and Shannon. In your own words, summarize the briefly what each of these metrics is doing and how they are different.

   The observed features shows the different types of functions and (features) in the given sample size.

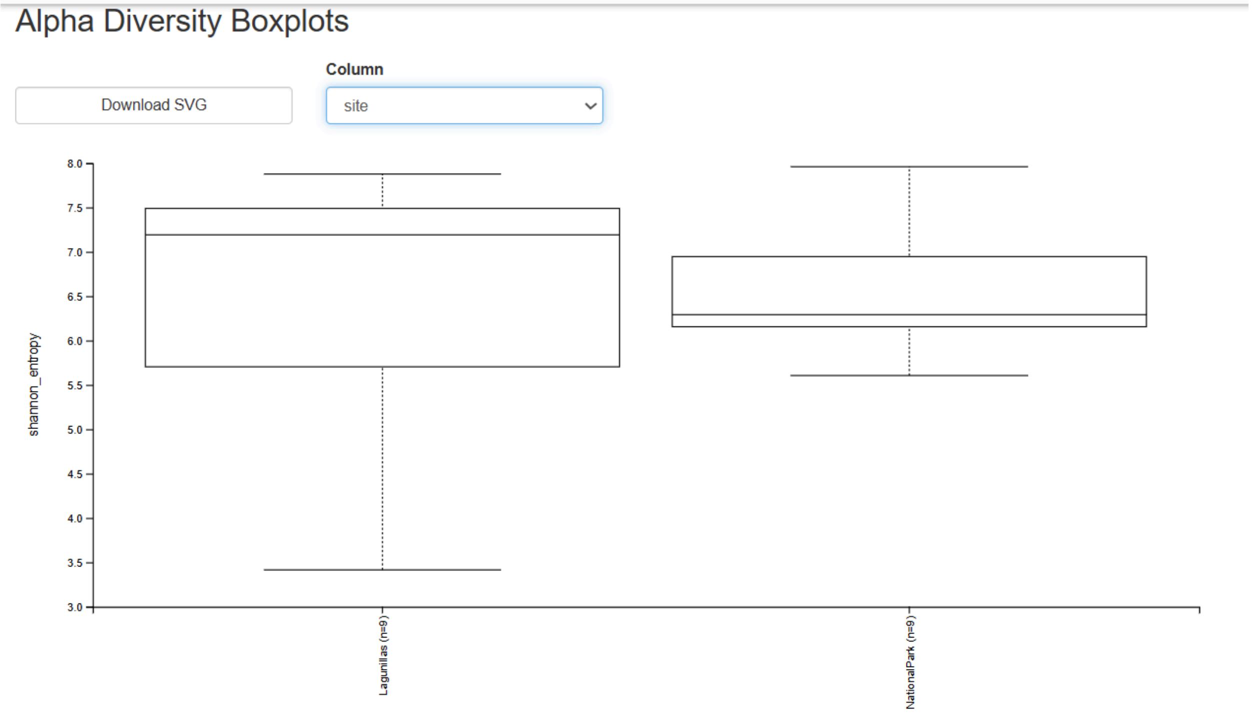
Shannon diversity dives deeper into the organziation across the species in terms of distribution of each sample, as well as the amount of species in each population.

7) 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?



Alpha Diversity Boxplots

Column: life_stage



Alpha Diversity Boxplots

Column: site

Both comparisons were not statistically significant due to the fact that both values are over 0.5

8) We can now look at beta diversity. How do alpha and beta diversity differ in what they are trying to tell you?

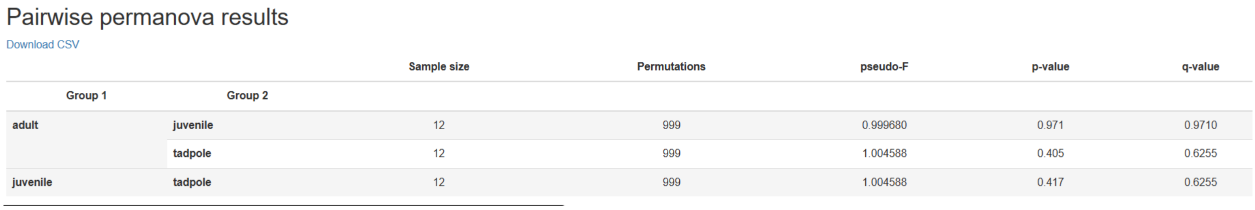
Alpha diversity focuses in on the diversity within one ecosystem, while beta diversity looks upon comparisons between two ecosystems.

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

You change all units in the commands by renaming all of the Bray Curtis to Weighted Unifrac.

10)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?

Bray Curtis life stage

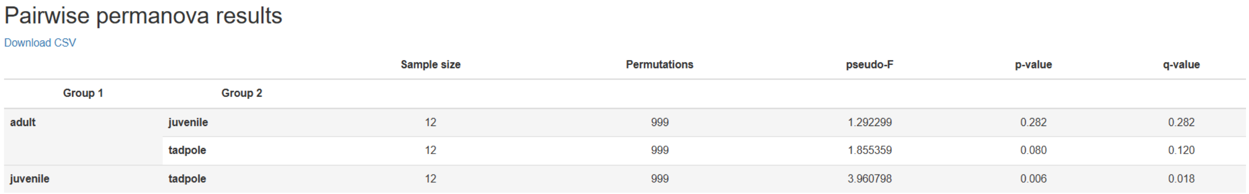


Pairwise permanova results

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
adult	juvenile	12	999	0.999680	0.971	0.9710
	tadpole	12	999	1.004588	0.405	0.6255
juvenile	tadpole	12	999	1.004588	0.417	0.6255

Weighted unifrac lifestage

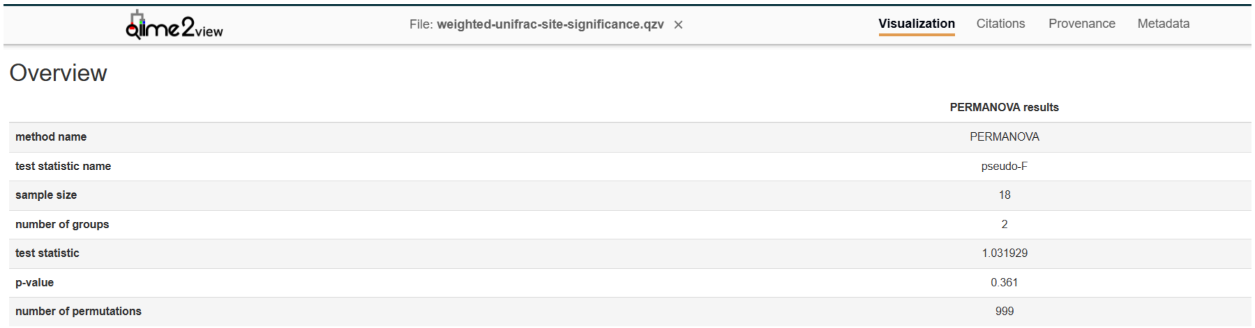


Pairwise permanova results

Download CSV

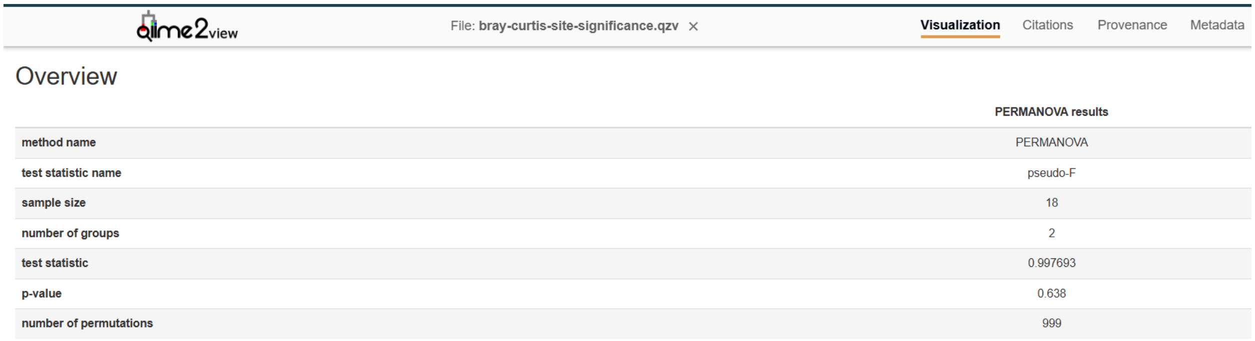
		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
adult	juvenile	12	999	1.292299	0.282	0.282
	tadpole	12	999	1.855359	0.080	0.120
juvenile	tadpole	12	999	3.960798	0.006	0.018

For the Bray Curtis sample, none of the groups show statistical significance, while the weighted unifrac sample shows all groups with statistical significance.



Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	18
number of groups	2
test statistic	1.031929
p-value	0.361
number of permutations	999

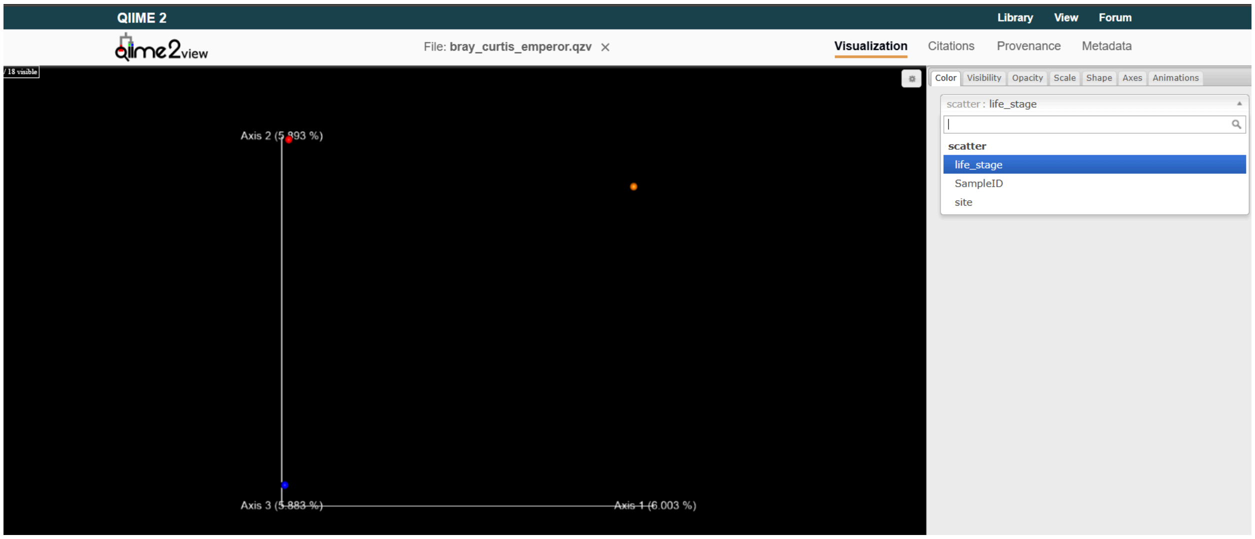


Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	18
number of groups	2
test statistic	0.997693
p-value	0.638
number of permutations	999

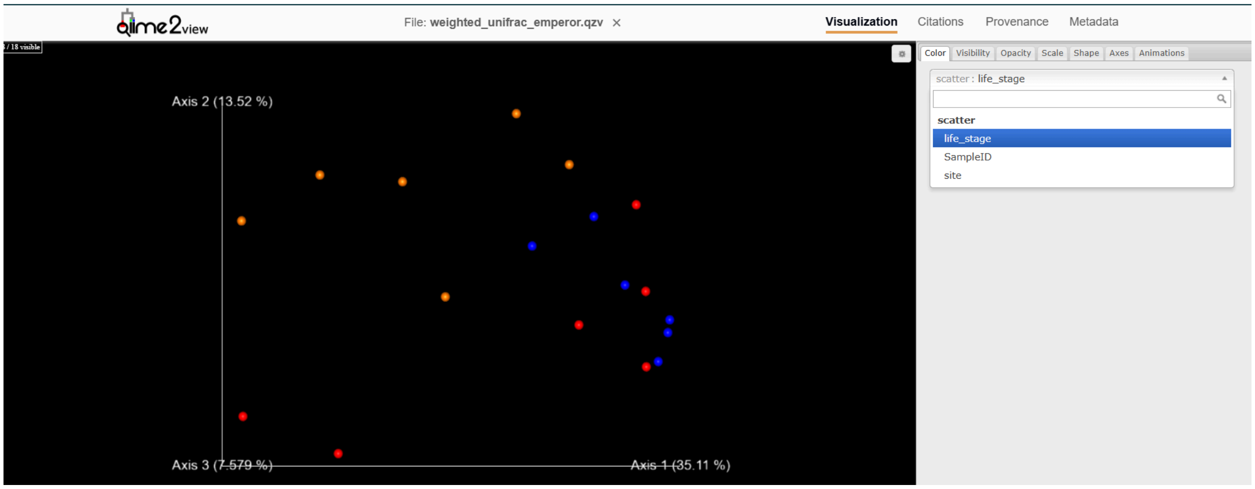
In the site samples, the opposite results occurred, with the unifrac yielding statistical significance (p value<0.5) and the Bray Curtis having no statistical significance.

11)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.



File: bray_curtis_emperor.qzv

scatter : life_stage



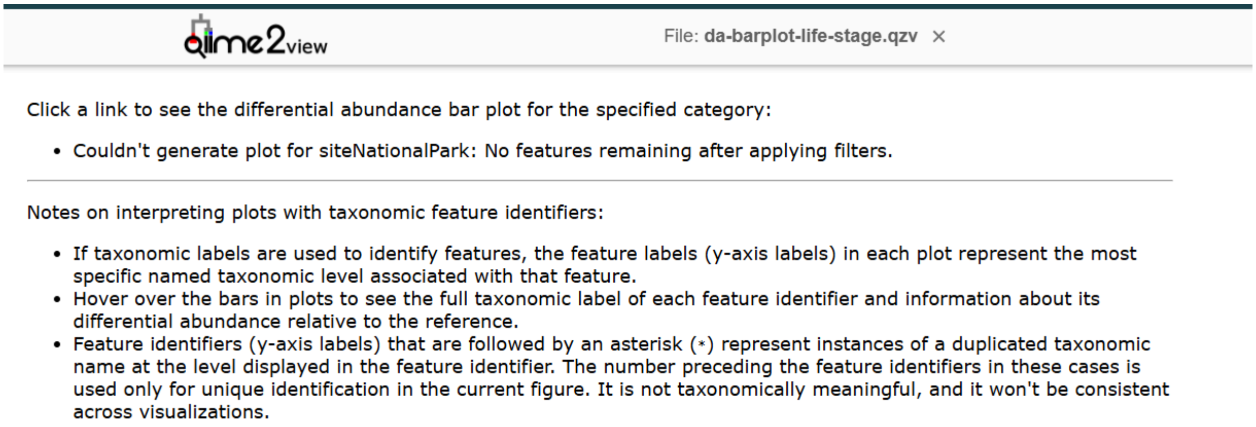
File: weighted_unifrac_emperor.qzv

scatter : life_stage

In terms of the Bray Curtis graph, the results showed close clustering of samples divided into their life stages far apart on the graph. This reflects the high statistical significance of each division while the Unifrac shows the complet opposite results, showing an intermingling of life stages across the graph, reflecting an equally opposite result in statistical significance of similarity.

12)f 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.

There was not enough of a statistical significance in the data to confirm a barplot for the samples, showing a lack of similarities across each sample.



Click a link to see the differential abundance bar plot for the specified category:

- Couldn't generate plot for siteNationalPark: No features remaining after applying filters.

Notes on interpreting plots with taxonomic feature identifiers:

- If taxonomic labels are used to identify features, the feature labels (y-axis labels) in each plot represent the most specific named taxonomic level associated with that feature.
- Hover over the bars in plots to see the full taxonomic label of each feature identifier and information about its differential abundance relative to the reference.
- Feature identifiers (y-axis labels) that are followed by an asterisk (*) represent instances of a duplicated taxonomic name at the level displayed in the feature identifier. The number preceding the feature identifiers in these cases is used only for unique identification in the current figure. It is not taxonomically meaningful, and it won't be consistent across visualizations.

There was a negative result for life stage as well, so it was not statistically