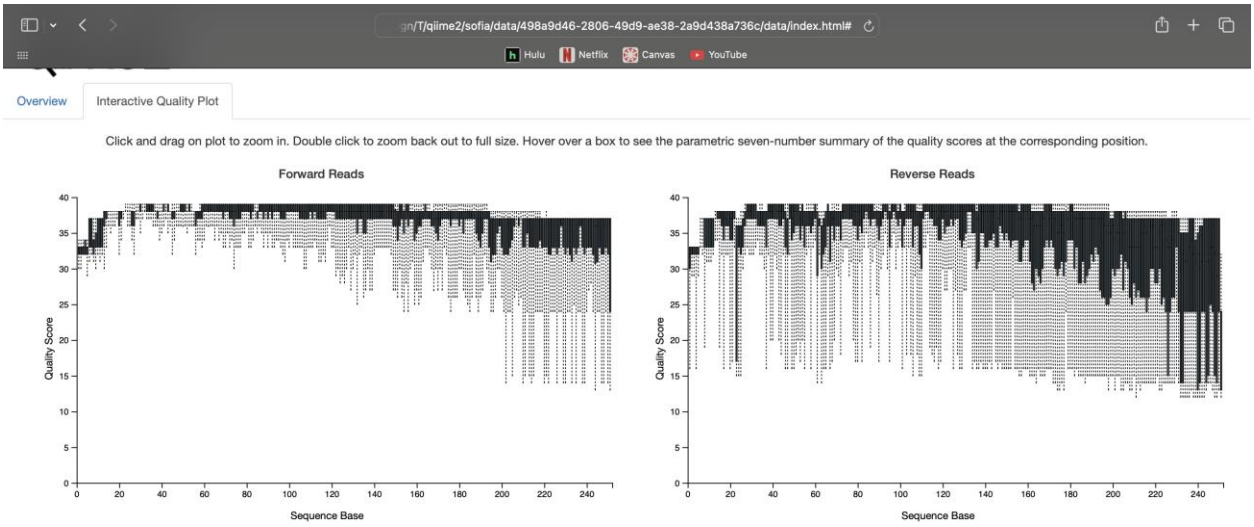
1) Examine the metadata file. What columns do you see that you think might be useful for later when doing alpha and beta diversity metrics? HINT: there are 3

The columns that might be useful for later when doing alpha and beta diversity metrics are the population category, sex category, and flock category.

2) Include a screenshot of your interactive quality plot. Based on this plot, what values would you choose for --p-trunc-len and --p-trim-left for both the forward and reverse reads? Why have you chosen those numbers? HINT: If you trim and truncate too much, you will lose too many of your reads, making your downstream analysis not useful. Think about it scientifically and only trim and truncate where the overall quality averages below 25-30.



Based on the plot, the values I would choose for –p-trunc-len forward is 240 and for the reverse is 200. The value for –ptrim-left forward is 0 and for reverse its 0. I chose these numbers because both graphs have good quality data for the left sequences so there is no need to trim. However, there is a need to truncate because some of the qualities drop below 30. For the forward reads it drops below 30 around 240 and for the reverse reads it's around 200.

3) Include a screenshot of the table summary from visualizing your table and a screenshot of the sequence length statistics from the rep-seqs file. Remember, we may eventually want to cut any samples with less than 10,000 reads. Do you see any in the interactive sample detail that might need to be cut? If so, which ones?

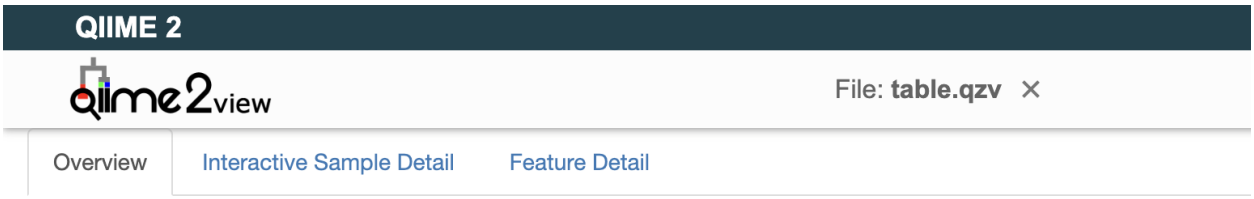


Table summary

Summary Statistic	Value
Number of samples	24
Number of features	5,948
Total frequency	526,012

There are 8 frequencies under 10,00: 84_S61_L001, 254_S69_L001, 168_S37_L001, 265_S133_L001, 100_S359_L001, 125_S13_L001, 104_S93_L001, 189_S23_L001.

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
5948	240	426	252.37	186	9.65

4) Once you have generated your taxonomy visualization, sort it by confidence. What are your top hits? What about if you sort by taxon? What hits do you see?

When you visualize the taxonomy file and sort it by confidence the top hits are:

k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__Myxococcales
k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Ellin329; f__; g__; s__

k__Bacteria; p__Bacteroidetes; c__Cytophagia; o__Cytophagales; f__Cytophagaceae; g__Spirosoma; s__

k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Acidithiobacillales; f__Acidithiobacillaceae; g__Acidithiobacillus; s__.

When you visualize the taxonomy file and sort it by taxon the top hits are:

k__Archaea; p__Crenarchaeota; c__Thaumarchaeota; o__Cenarchaeales; f__SAGMA-X; g__; s__

k__Archaea; p__Crenarchaeota; c__Thaumarchaeota; o__Nitrososphaerales; f__Nitrososphaeraceae; g__Candidatus Nitrososphaera; s__gargensis

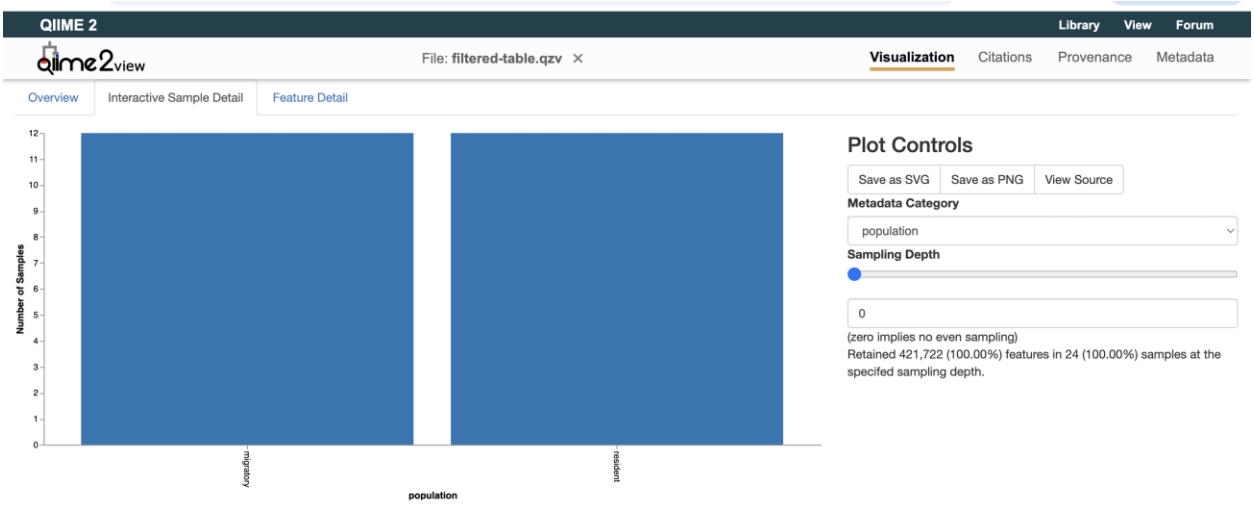
k__Archaea; p__Crenarchaeota; c__Thaumarchaeota; o__Nitrososphaerales; f__Nitrososphaeraceae; g__Candidatus Nitrososphaera; s__gargensis

k__Archaea; p__Euryarchaeota; c__Halobacteria; o__Halobacteriales; f__Halobacteriaceae

5) When you visualize level 3 of taxonomy, what level is this? Do you see any trends as you sort by various metadata categories?

Level 3 of taxonomy represents class. At level 3 taxonomy you can see all the species are from the Bacteria kingdom and the phylum/class differ from species to species. When the chart is sorted by population, we can see some samples have higher level of richness and evenness while others are dominated by few bacterial groups. When the chart is sorted by sex, we can see the compositions stay consistent between sexes with no major shift on either side. When the chart is sorted by flock, we can see the species compositions vary greatly between flocks.

6) After visualizing your filtered-table.qzv, what cutoff value will you use for generating alpha and beta diversity? Why? Include a screenshot of the interactive sample view to help justify your reasoning.



population

Sample ID	Frequency
106_S98_L001	48,643
13_S95_L001	43,345
18_S71_L001	29,814
174_S146_L001	27,837
212_S94_L001	27,406
4_S157_L001	26,985
309_S47_L001	25,639
364_S22_L001	25,451
78_S46_L001	25,288
128_S36_L001	20,229
366_S45_L001	18,684
307_S70_L001	17,596
163_S60_L001	14,482
245_S122_L001	12,809
61_S109_L001	12,596
385_S170_L001	11,151
254_S69_L001	8,285
168_S37_L001	7,538
84_S61_L001	7,273
265_S133_L001	4,432
100_S359_L001	2,609
125_S13_L001	2,097
104_S93_L001	1,128
189_S23_L001	405

Based on the data, the cutoff value for generating an alpha and beta diversity would be 10000. Once the data reaches the 10000 mark there seems to be a greater decline in frequencies. This implies that those data points have possible sequencing failure so, by choosing 10000 as the cutoff I keep 20 good samples and remove 8 moderate ones. By removing these samples, the strong statistical power of the data is kept

7) The first metric we will analyze is alpha diversity. In your own words, what is alpha diversity and what are the differences between the two types of alpha diversity we will analyze (Shannon and Observed features)?

Alpha diversity measures the diversity of species in a single sample/environment. This can show us the different types of species in an environment and how they are dispersed. A Shannon Alpha diversity measures the richness and evenness of the species in an environment. This differs from Observed features alpha diversity because it accounts for both the number of species in the environment and how evenly they are distributed. The Observed features alpha diversity measures the richness of species in an environment. It counts the number of different species in an environment. This differs from the Shannon alpha diversity because it doesn't consider distribution of the population.

8) S**ince you are looking at two metrics of diversity for 3 metadata categories, it would be helpful to make a table of the significant values. Are any of your comparisons significant? For one of the metadata data columns, there are actually 4 options. Include a screenshot of**

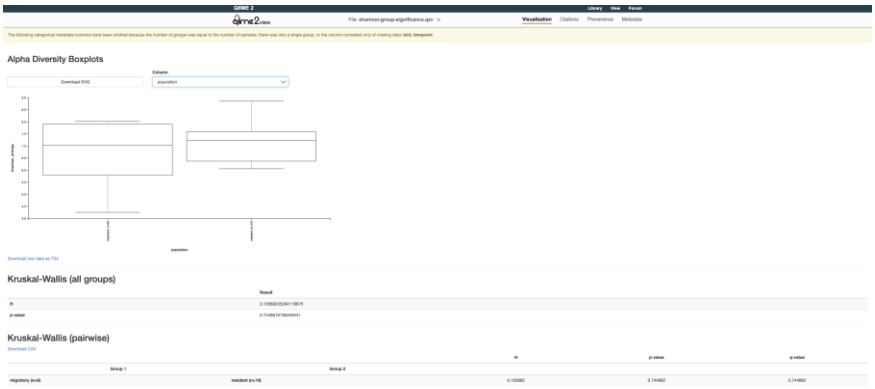
the pairwise comparisons for Shannon and Observed. Are there any significant comparisons? (HINT: look at the q value)

When comparing the q values for the Shannon file and observed features file there seems to be no significance between the categories. For the populations there is no significant difference between the results with both being greater than 0.05. For the sex category there is no significant difference with both values being greater than 0.05. Finally, for the flock category there is a difference between the two graphs with there being a significance for the Shannon data and no significance for the observed features.

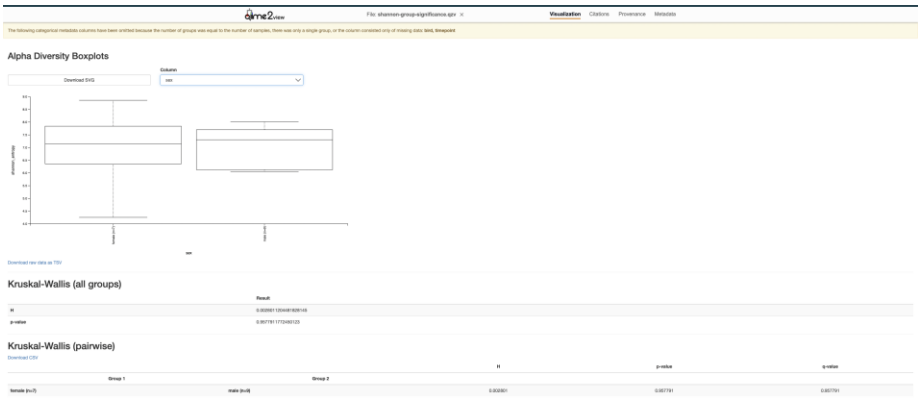
	H Shannon	P-Value Shannon	q-Value Shannon	H Observed	P-Value Observed	Q-Value Observed
Population	0.105882	0.744881	0.744882	0.294118	0.587594	0.587594
Sex	0.002801	0.957791	0.957791	0.473389	0.491432	0.49143
Flock	0.428571 3.750000 0.060000 3.84000	0.064078 0.052808 0.806496 0.050044	0.090960 0.806496 0.090960 0.090960	3.428571 3.750000 0.060000 0.960000	0.064078 0.052806 0.806496 0.327187	0.128155 0.806496 0.490780 0.557643

Shannon-group-signifigance

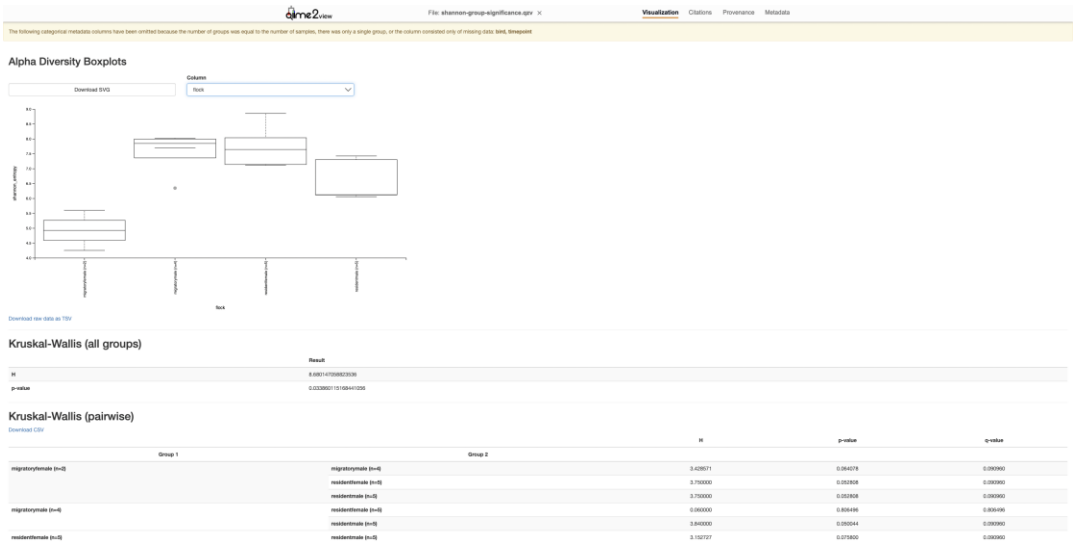
Based on the Shannon group significance there is no significance between the populations with the p value being g>0.05.



Based on the Shannon group significance there is no significance between the two sexes with the p-value being >0.05.

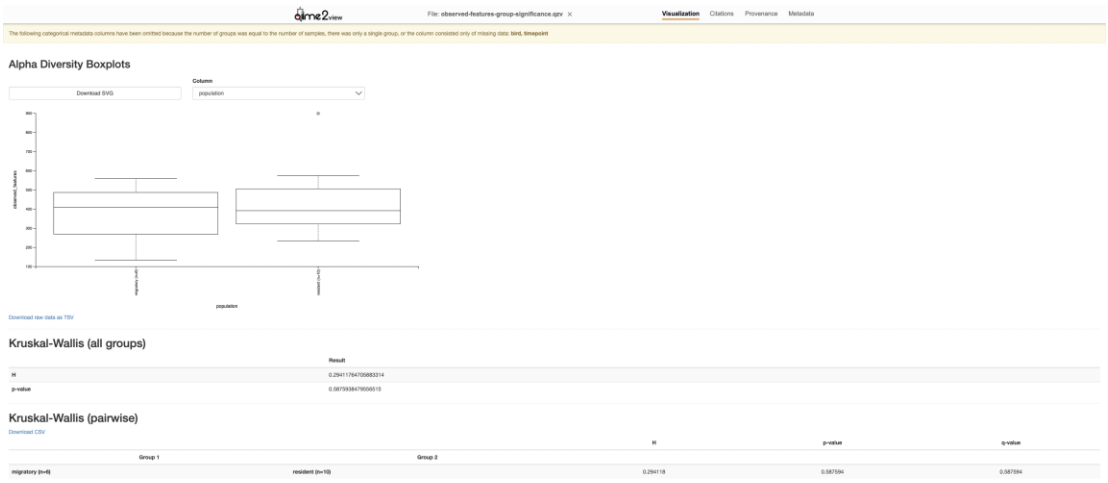


Based on the Shannon group significance there is a significance between the 4 flocks with the p-value being <0.05.

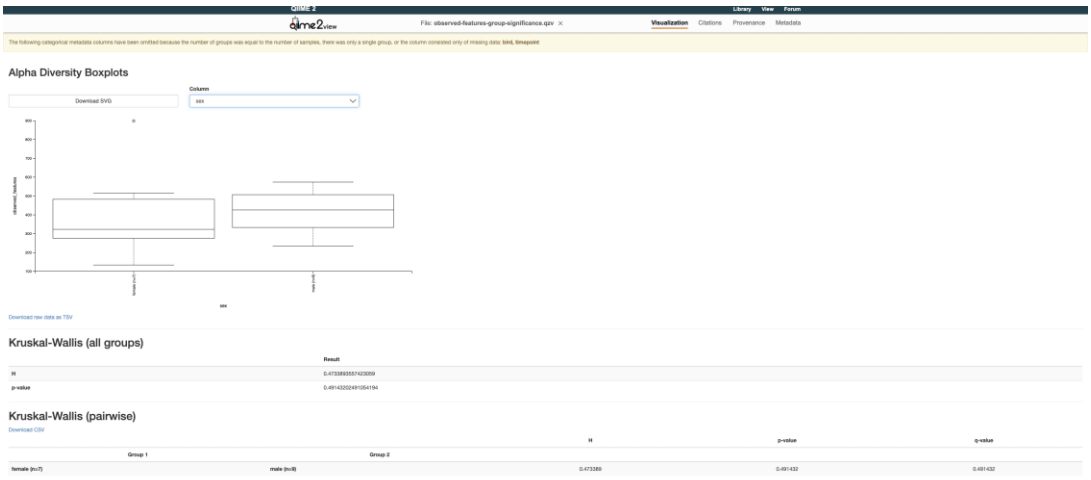


Observed-features-group-significance

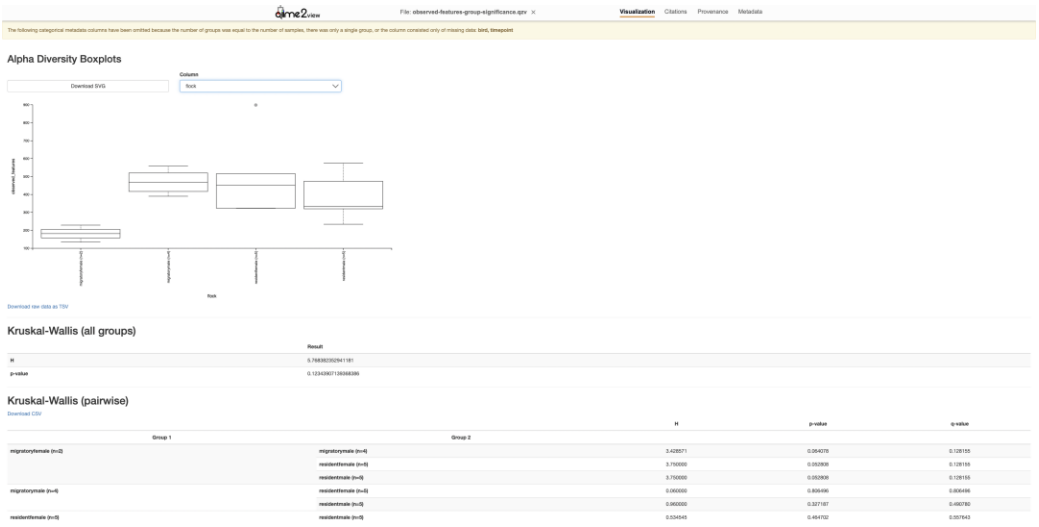
For the overserved features group significance table there is no significance between the two populations with the p-value being >0.05.



Based on the observed features group significance table there is no significance between the two sexes with the p-value being >0.05.



Based on the observed features group significance there is no significance with the overall group p-value being greater than 0.05.



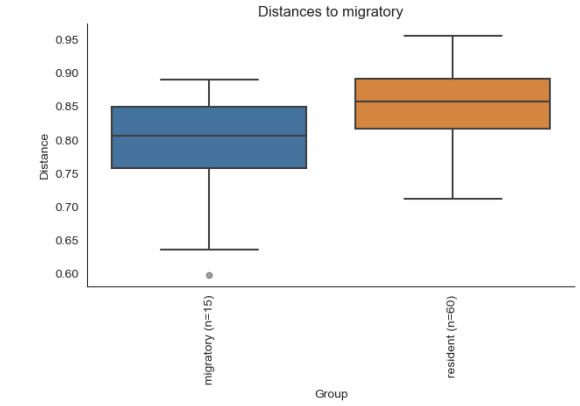
10) For beta diversity, you will need to create visualizations for Bray Curtis dissimilarity and Weighted Unifrac distance. This will require you to modify the beta-group-significance code. You should have one visualization for each metadata column you are interested in. Include a screenshot of each visualization. Is there any significance? Regardless of significance, how can you interpret these results (hint: what is beta diversity looking at?) How are Bray Curtis dissimilarity and Weighted Unifrac distance different?

These results can be interpreted regardless of significance by looking for patterns of similarity or dissimilarity between the microbial communities. For example, communities that are closely related will be grouped together while those that are different will be far apart.

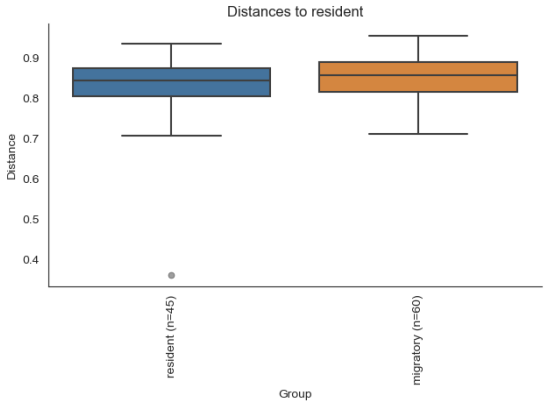
A Bray Curtis dissimilarity looks at the species present and how many are there. Its goal is to measure the abundance of specific taxa. A Weighted Unifrac distance looks at the species present, how many are there, and how closely they are related. Its goal is to measure the evolutionary relationship between two communities.

Group significance plots

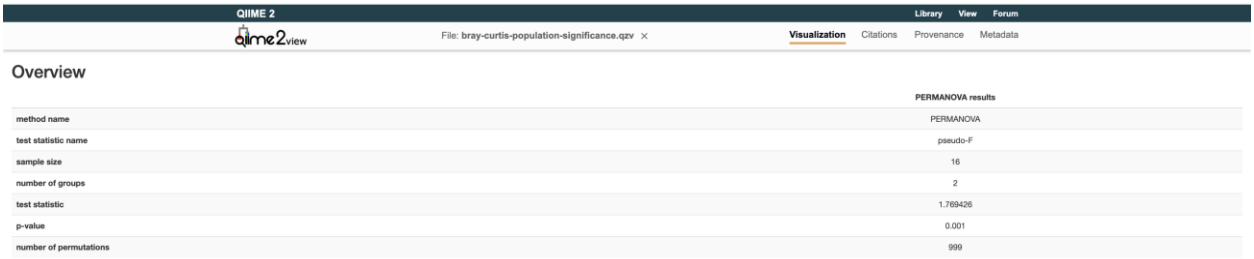
Download raw data as TSV



Download as PDF



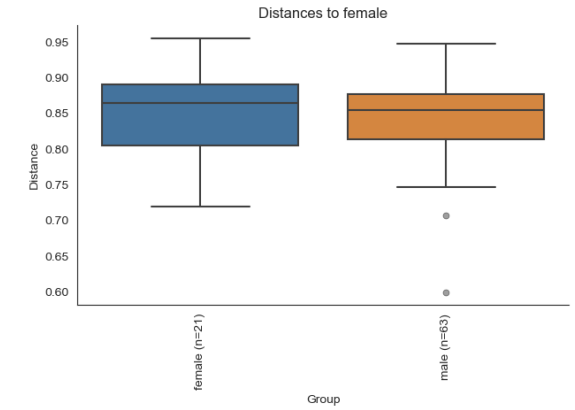
Download as PDF



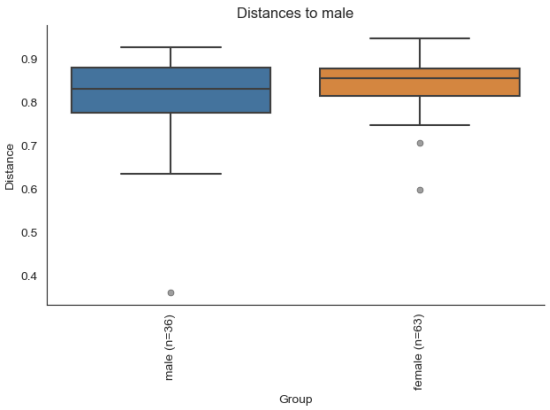
Using the bray-Curtis visualization for the population data there is significance with the p-value being less than 0.05.

Group significance plots

ownload raw data as TSV



Download as PDF



Download as PDF

QIIME 2 Library View Forum

qiime2 view File: bray-curtis-sex-significance.qzv × Visualization Citations Provenance Metadata

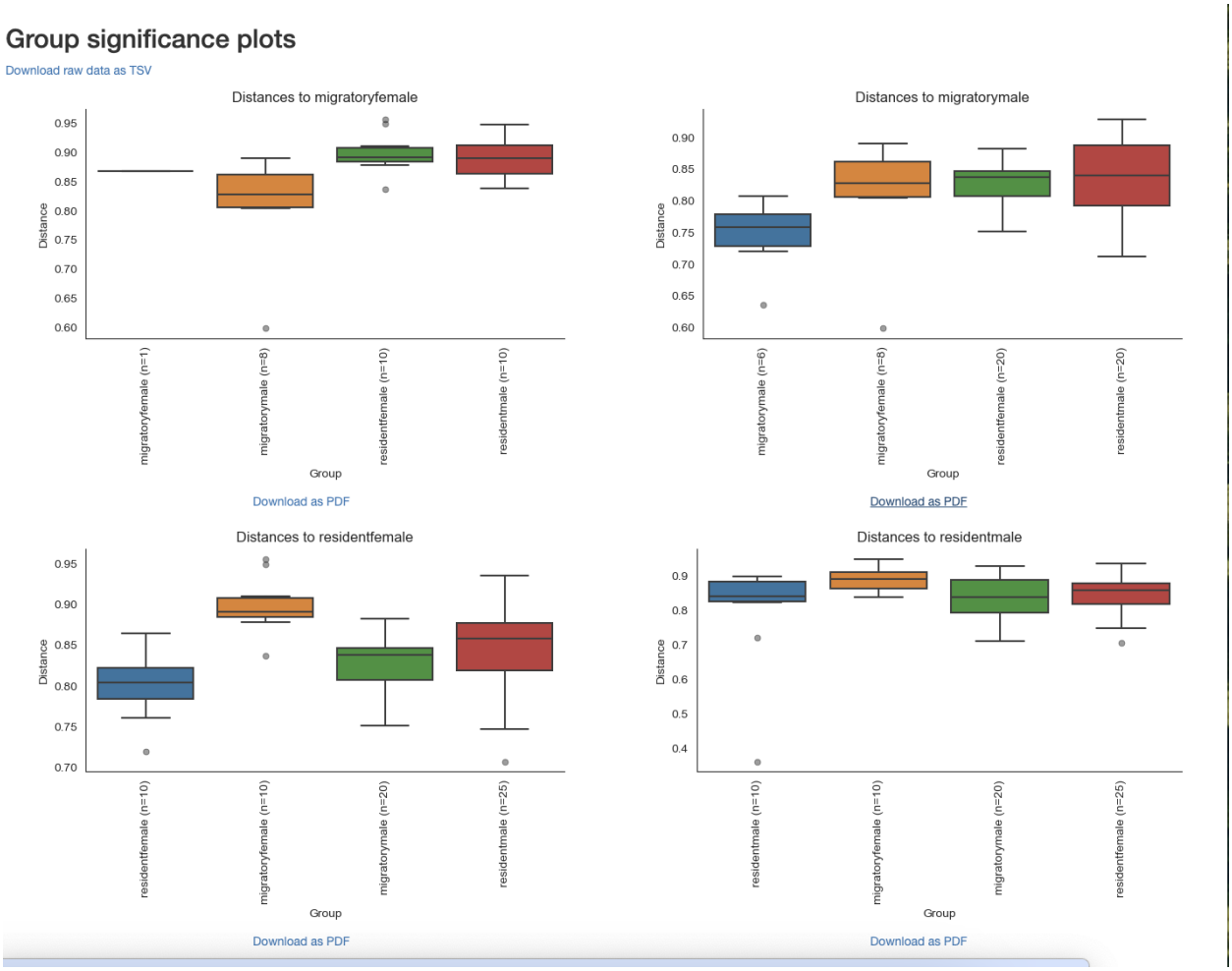
Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	2
test statistic	1.249632
p-value	0.103
number of permutations	999

Using the Bray-Curtis visualization for the sex data there is no significance with the P-value being greater than 0.05.

Group significance plots

Download raw data as TSV



QIIME 2 Library View Forum

File: bray-curtis-flock-significance.qzv

Visualization Citations Provenance Metadata

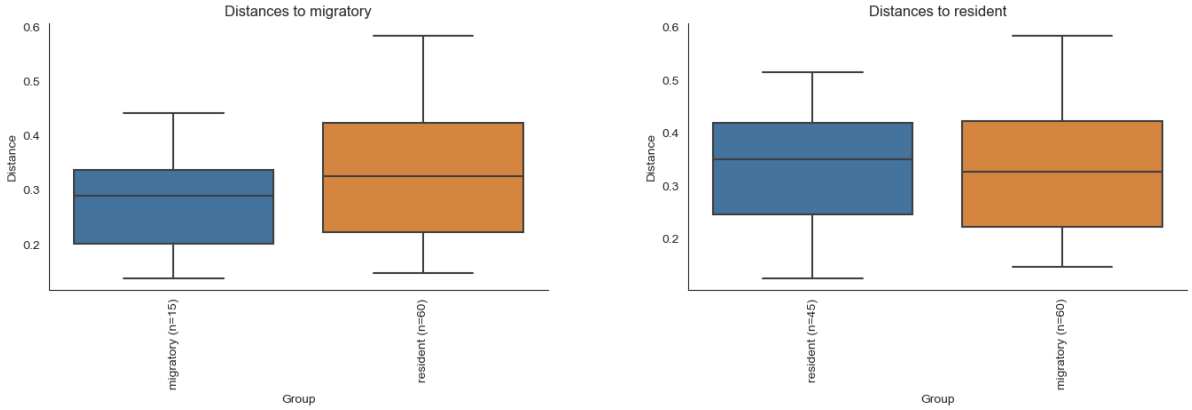
Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	4
test statistic	1.530614
p-value	0.002
number of permutations	999

Using the Bray-Curtis visualization for the flock data there is significance with the P-value being less than 0.05.

Group significance plots

Download raw data as TSV



Download as PDF Download as PDF

File: weighted-unifrac-population-significance.qzv

Visualization Citations Provenance Metadata

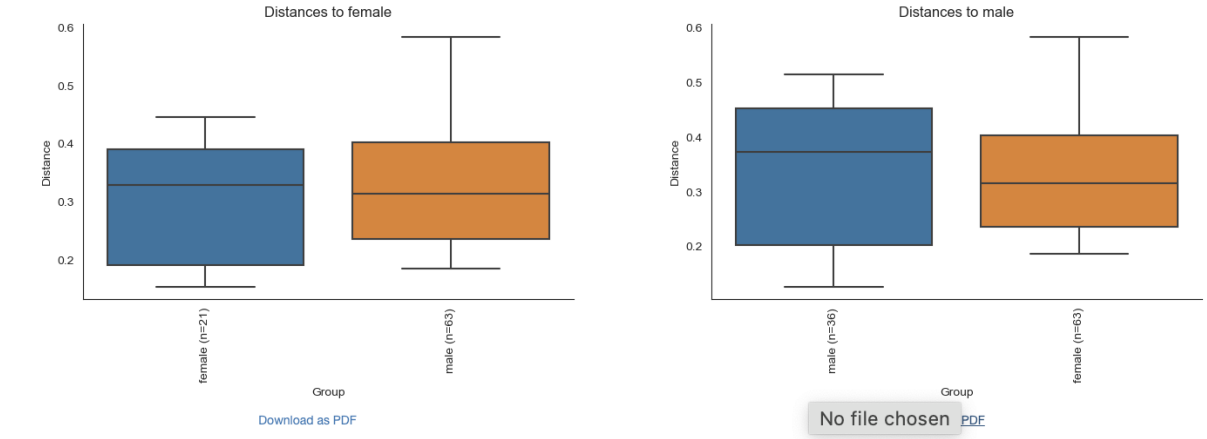
Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	2
test statistic	2.294866
p-value	0.038
number of permutations	999

Using the Weighted-Unifrac visualization for the population data there is significance with the P-value being less than 0.05.

Group significance plots

Download raw data as TSV



Download as PDF

No file chosen PDF

Pairwise permanova results

Download CSV

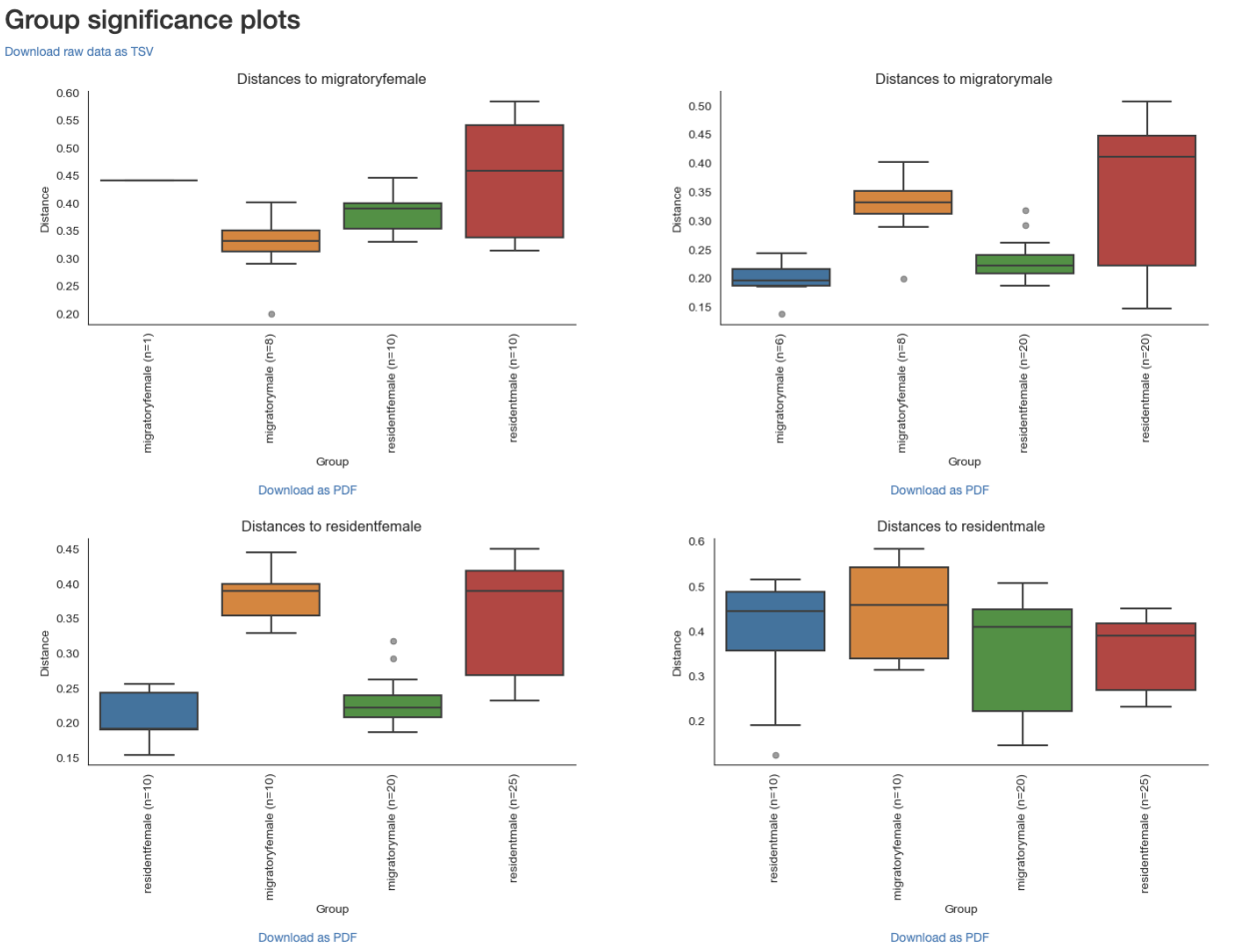
	Sample size	Permutations	pseudo-F	p-value	q-value

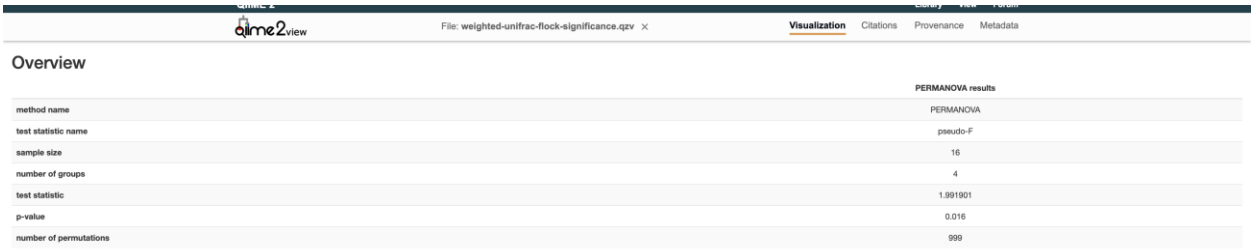
qiime2 view File: weighted-unifrac-sex-significance.qzv × Visualization Citations Provenance Metadata

Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	2
test statistic	0.99437
p-value	0.456
number of permutations	999

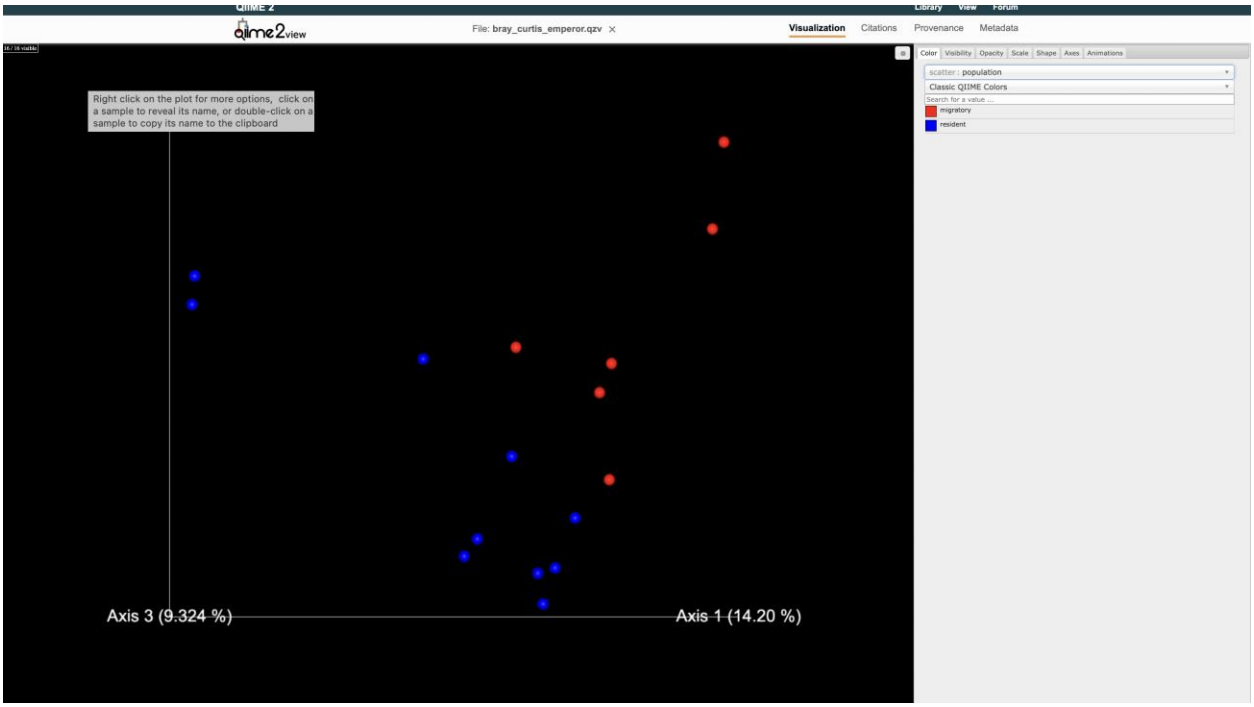
Using the Weighted-Unifrac visualization for the sex data there is no significance with the P-value being greater than 0.05.



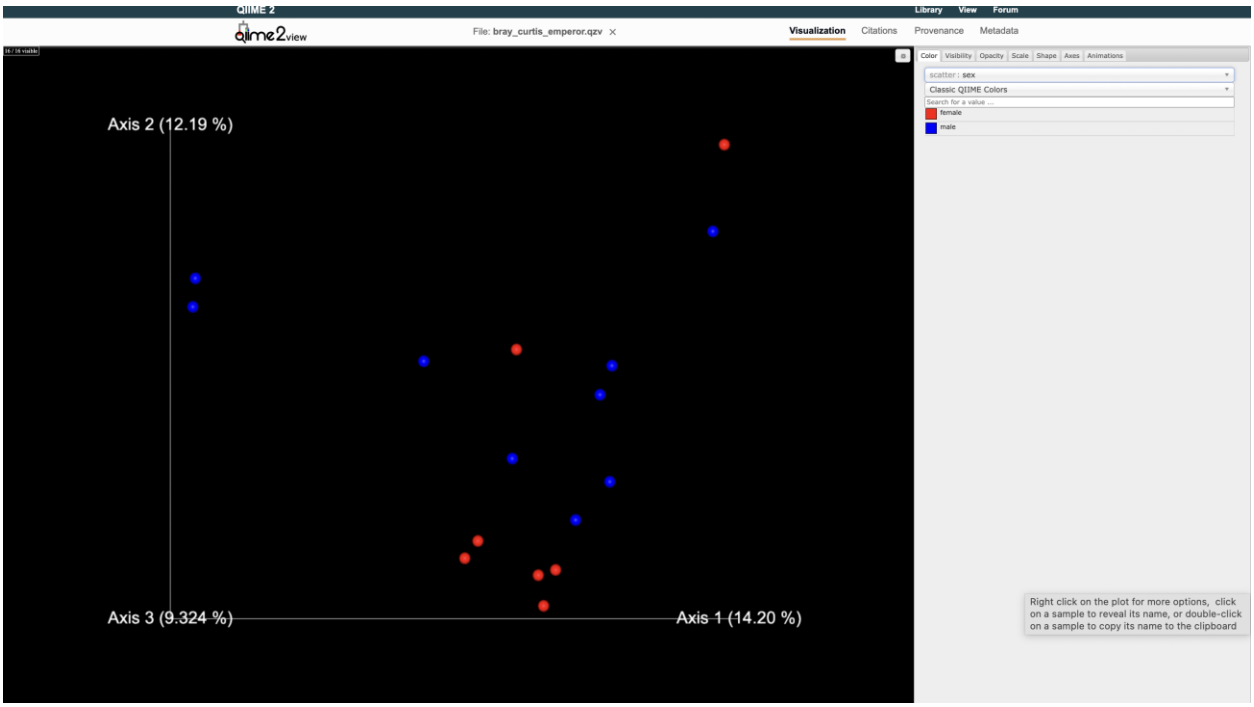


Using the Weighted-Unifrac visualization for the flock data there is significance with the P-value being less than 0.05.

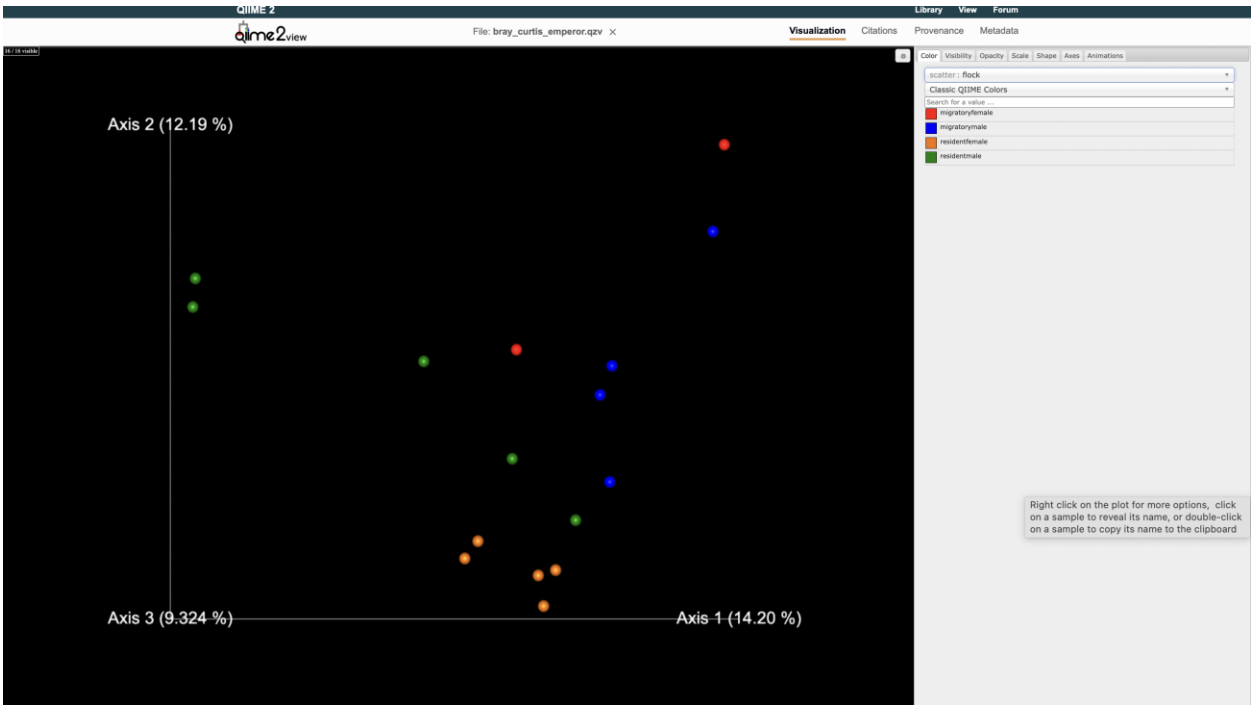
11) The core-metrics-phylogeny command generates a file called bray-curtis-emporer.qzv. Include 3 screenshots total (where the points are colored based on the metadata metrics). How do these results help you make sense of the results you got from question 10?



This graph shows the abundance of species in the populations migratory and resident. Based on the graph, the two species are not similar which will mean there is a significant difference between them. This matches the p value from the Bray-Curtis table.



This Bray-Curtis graph shows the relationship between the two sexes. Based on this graph the two sexes are somewhat similar because they are paired together on the graph. This matches the p-value from the graph which is greater than 0.05.

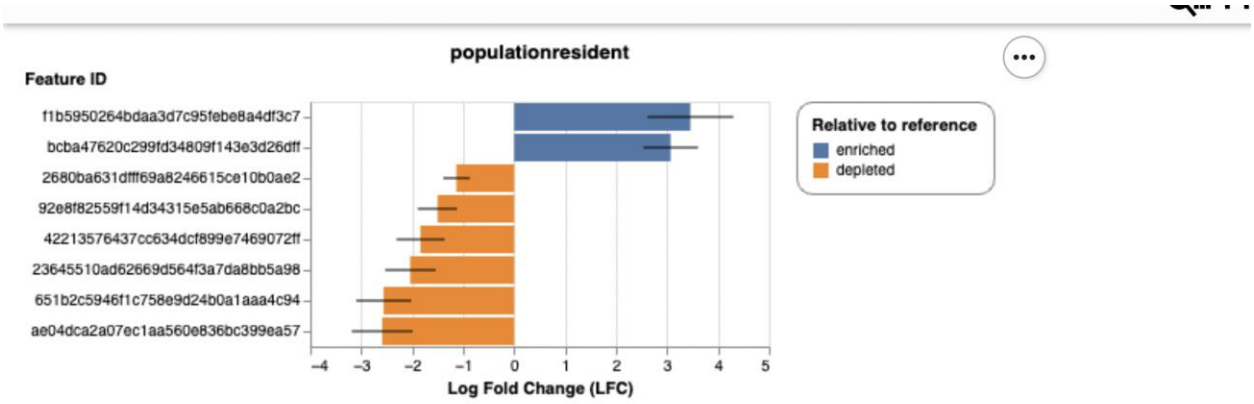


This graph shows the Bray-Curtis for the flock category. The flock category includes migratory female, migratory male, resident female, and resident male. Based on the graph, the resident female and migratory female are very different from one another; the resident male and resident female are like one another; the migratory male and resident male are like one another; and the migratory male and migratory female are like one another. Overall, many of the flock points are far apart from one another indicating a difference between the species. This matches the low p-value which indicates a significant difference between the flocks.

12) Using ANCOMBC, do you find any specific taxa that are differentially expressed with your three metadata categories? Please include a screenshot of any differentially expressed taxa and identify the species using the taxonomy.qzv file.

Using ANCOMBC for the population there was a graph generated for population resident. Using the taxonomy.qzv files the species' taxonomy can be identified.

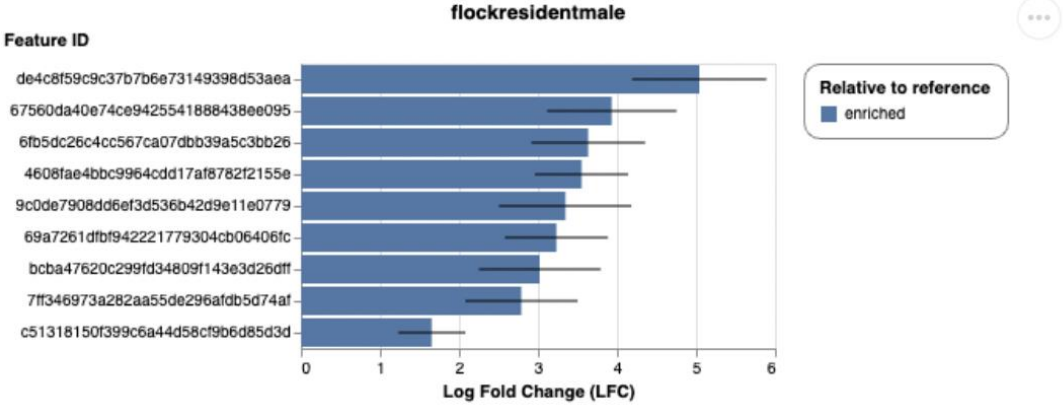
Feature ID	Taxonomy
42213576437cc634dcf899e7469072ff	s__marcusii

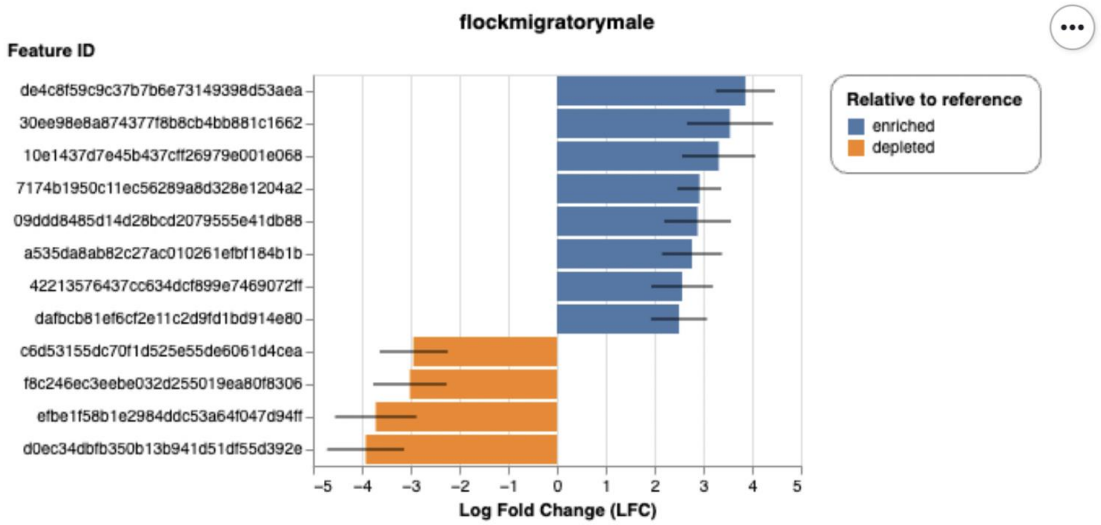


- Using ANCOMBC for the flock there were three graphs generated for flockmigratorymale, flockresidentfemale, and flock resident male.

Feature ID	Taxonomy
42213576437cc634dcf899e7469072ff	s__marcusii


flockresidentfemale
Feature ID
Relative to reference
enriched
depleted


flockresidentmale
Feature ID
de4c8f59c9c37b7b6e73149398d53aea
67560da40e74ce9425541888438ee095
6fb5dc26c4cc567ca07dbb39a5c3bb26
4608fae4bbc9964cdd17af8782f2155e
9c0de7908dd6ef3d536b42d9e11e0779
69a7261dfbf942221779304cb06406fc
bcba47620c299fd34809f143e3d26dff
7ff346973a282aa55de296afdb5d74af
c51318150f399c6a44d58cf9b6d85d3d
Relative to reference
enriched
0
1
2
3
4
5
6
Log Fold Change (LFC)



Using ANCOMBC for the sex there was a graph generated for sexmale.

Feature ID	Taxonomy
ee5cfac5c165c97e0a1475e8d9b4c297	s__pirum