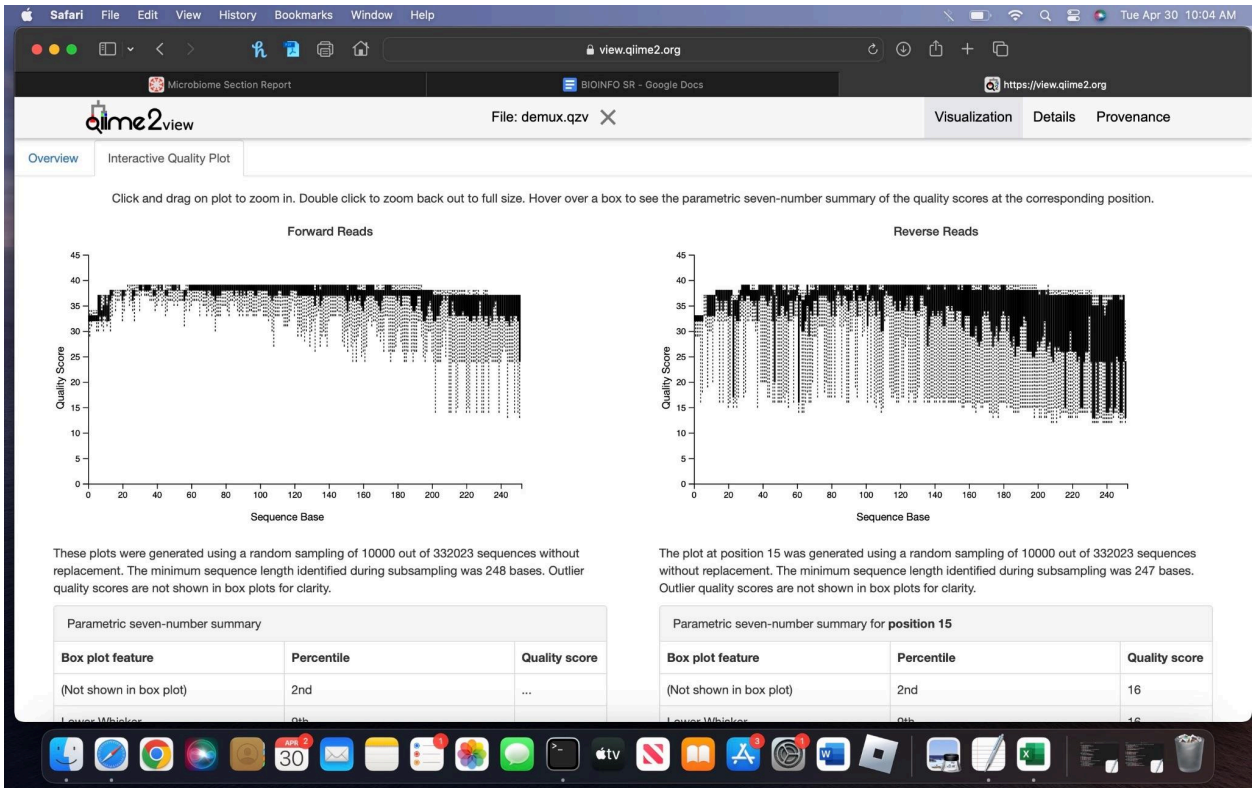
Andrea Villanueva
BIOL4810-001
3 May 2024
Dr. Van Laar

Microbiome Section Report

1) 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?



-
-
- The numbers I chose for the forward reads were 15 to 195, since the quality of the reads seem to drop before and after those areas. For the reverse reads, I chose 28 and 160 for the same reasons.

2) How would you modify the code above to truncate and trim in your desired way?

- We can modify the code to truncate and trim the reads specific to what areas of the reads we want to analyze. Since the interactive quality plot shows us a histogram or the quality, as well as statistics (such as the quality scores for each box plot), we can see where the score starts to drop, or areas we might not want to analyze later on. By moving the cursor to the areas where the data starts to

fall, the numbers change in the text to show what section we are looking at – this is how I decided on my numbers in the code.

3) In the tutorial, you had to `mv` the files to rename them to just `rep-seqs.qza`, `table.qza`, and `stats.qza`. How could you modify the above code to skip that step? How do you need to modify `qiime metadata tabulate` in order to account for the renamed files being generated?

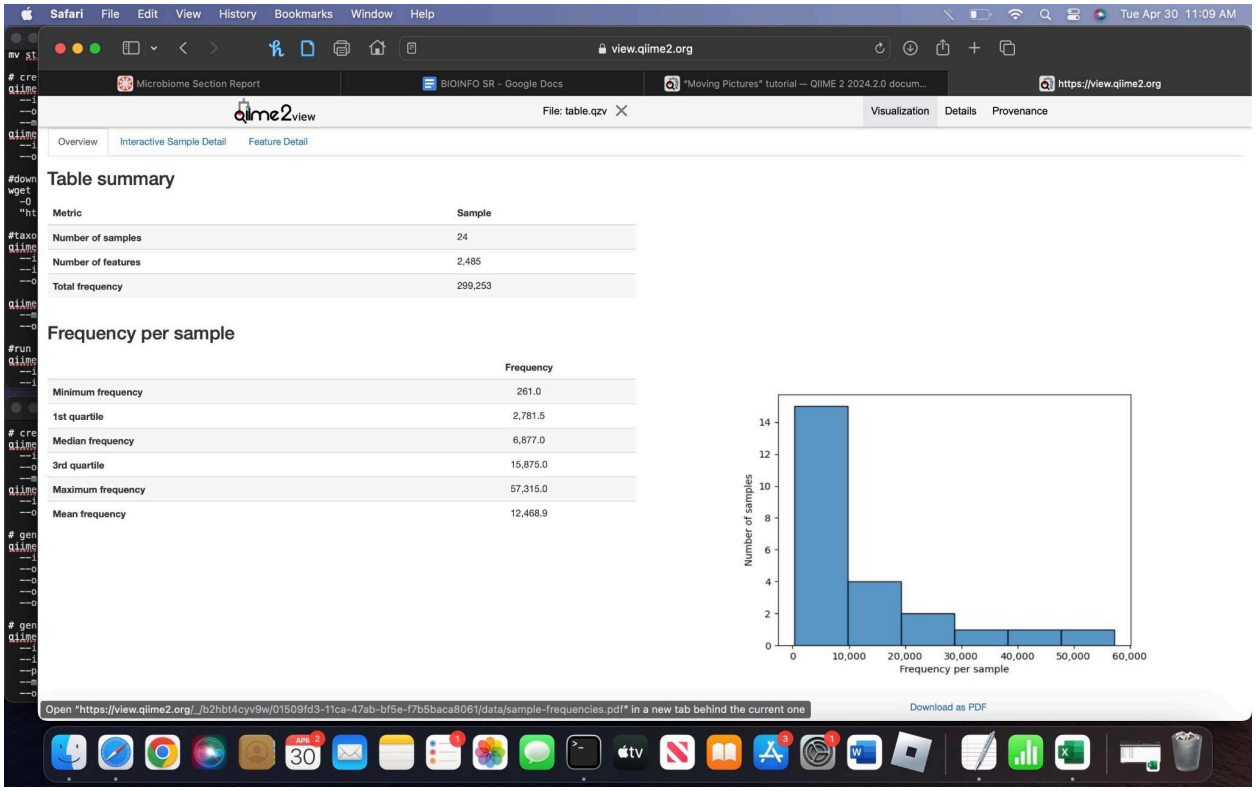
- We can add the code to rename the files along with the code for quality trimming for both the forward and reverse reads, using #mv rename files
- To modify the qiime metadata tabulate, we need to change the file names in the code from stats-dada2.qza to stats.qza, and stats-dada2.qzv to stats.qzv

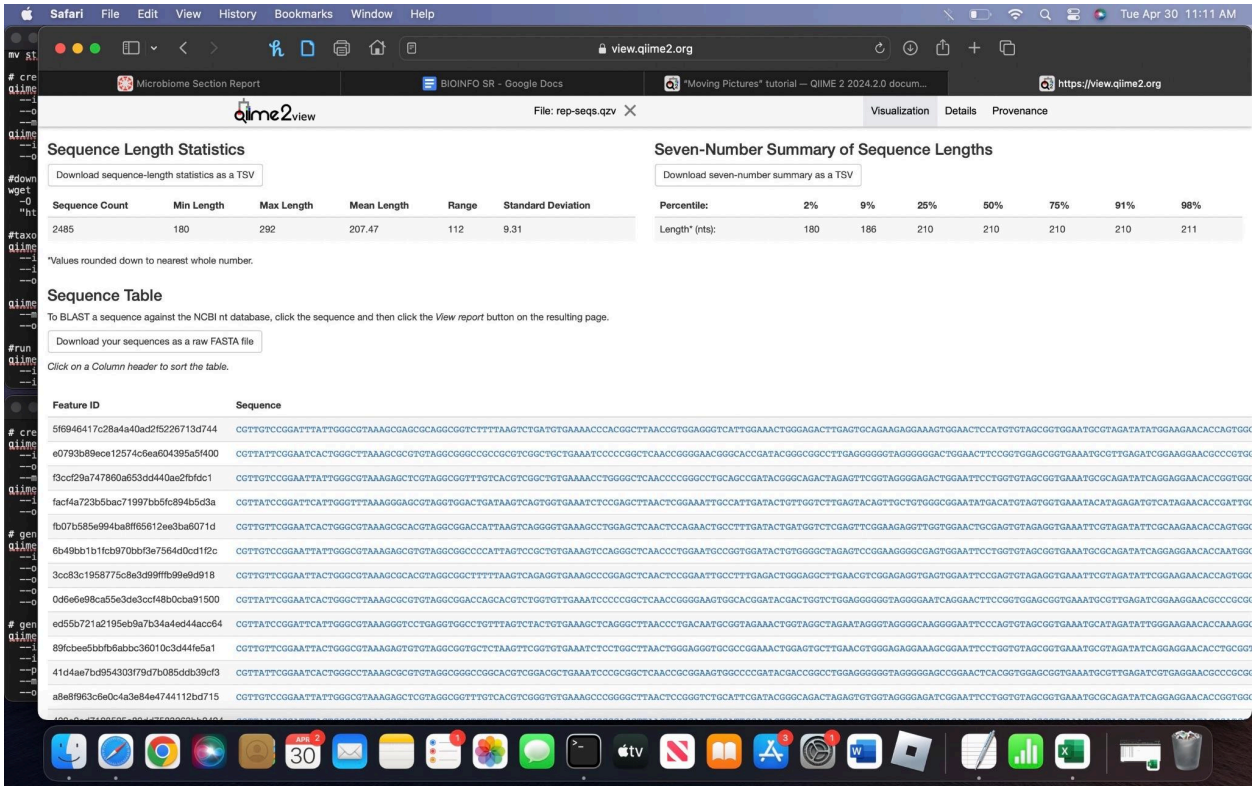
4) Your metadata file has a different name than that in the tutorial. How do you adjust your code in order to use the metadata file you have been given?

- We change the file name from the original tutorial code (sample-metadata.tsv) to what it's actually called, metadata.txt

5) Include a screenshot of the table summary from visualizing your table and a screenshot of the sequence length statistics from the rep-seqs file.

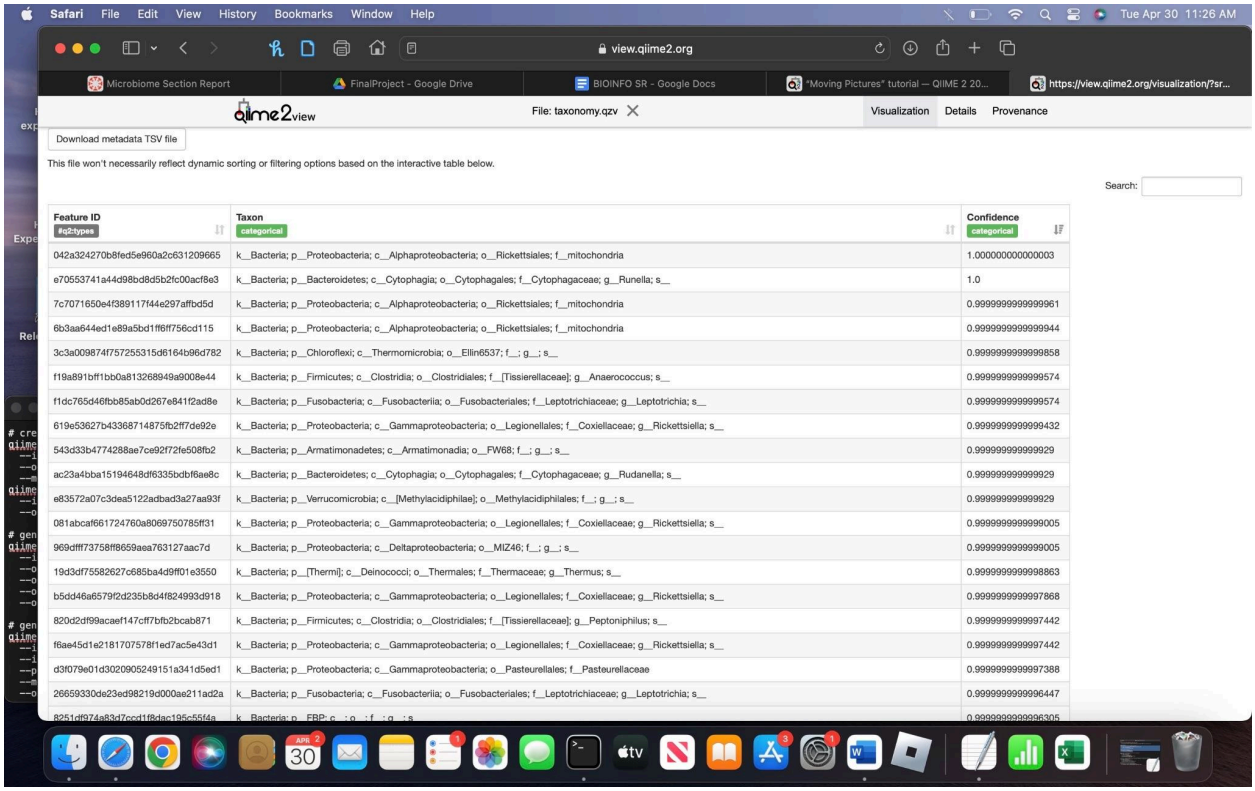
-





6) Jump down to taxonomy. Once you have generated your taxonomy visualization, sort it by confidence. What are your top hits?

- My top hits are in this screenshot:



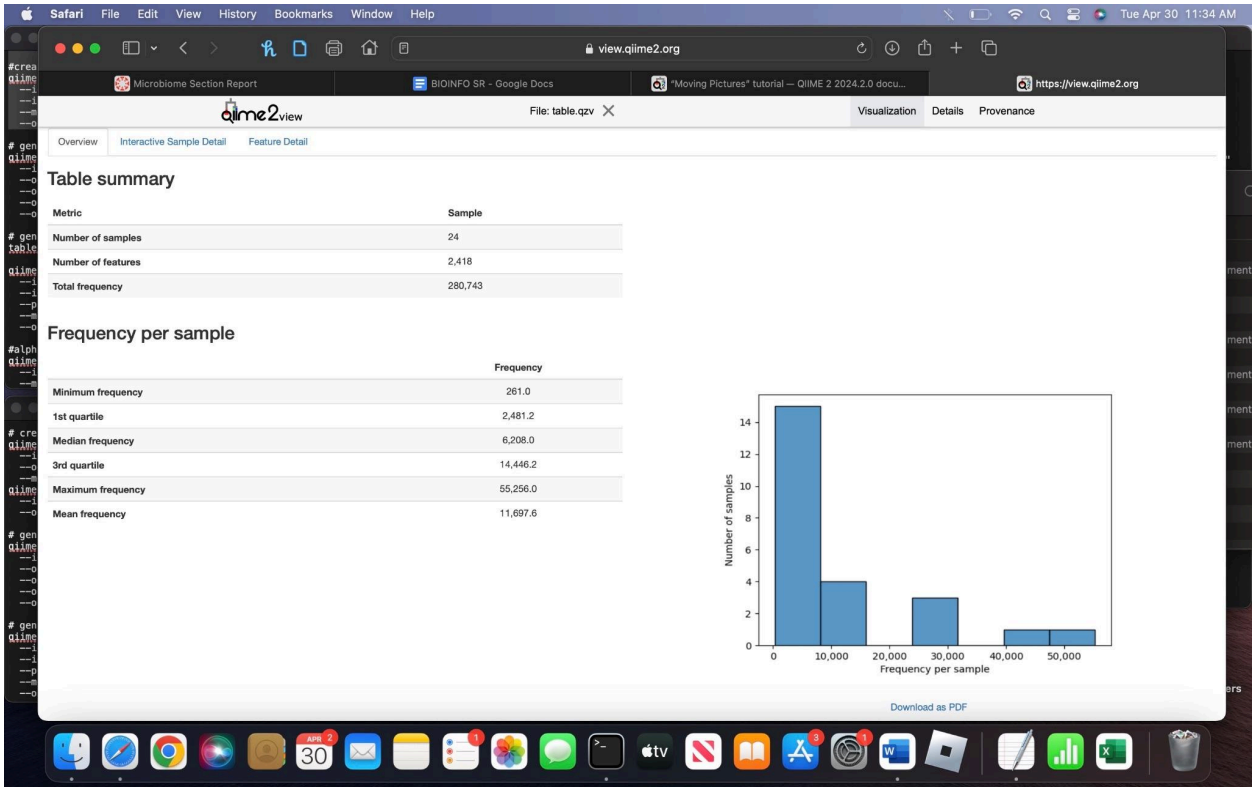
7) What do you think this code is doing? Why do you think this is a necessary or important step?

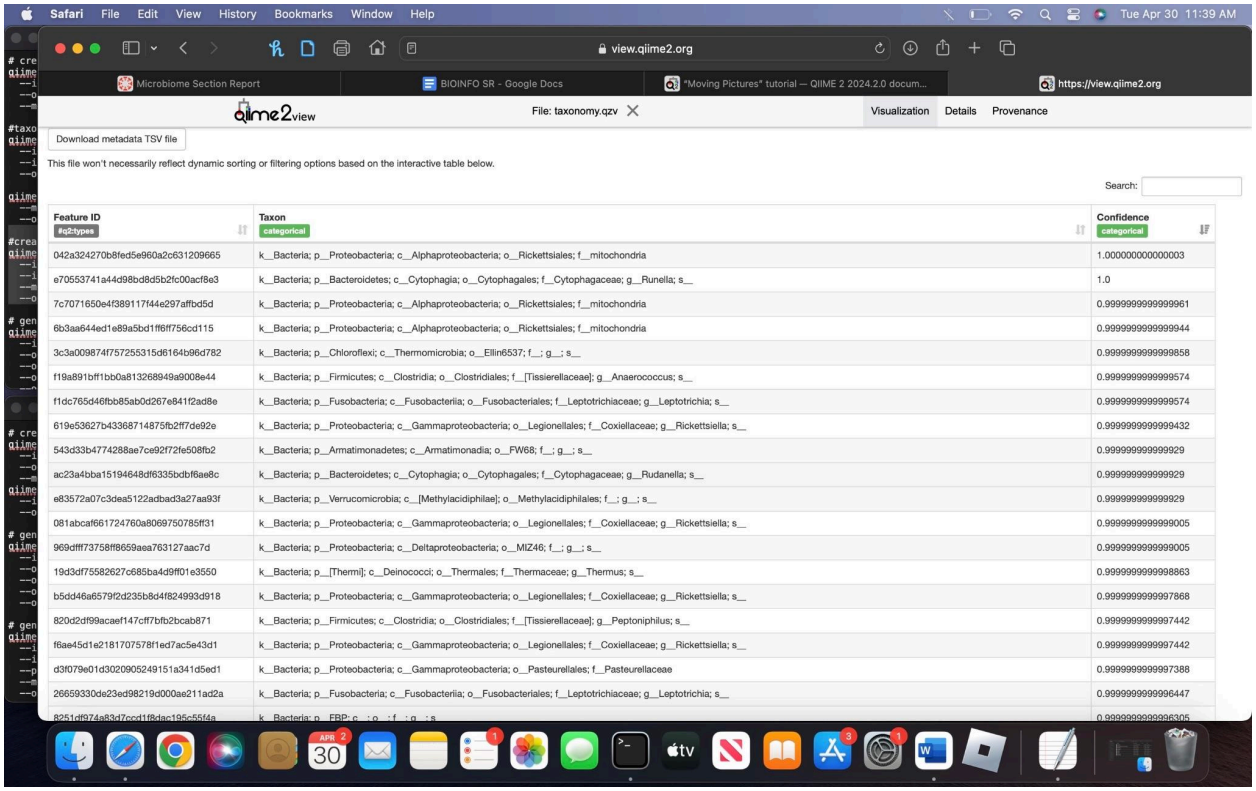
- This code is to exclude the mitochondria and chloroplasts to focus only on the bacteria. They are very similar because of the endosymbiont theory, which states that eukaryotic mitochondria and chloroplasts were taken up by prokaryotic cells.

8) Re-do your table visualization and re-do your taxonomy commands. Do you have any differences now in the hits with the highest confidence? Why or why not? Really think about what the code is doing.

- The mitochondria and chloroplasts are excluded, but not completely removed from the table. From the code, we can see that the table is filtered, but we can still see mitochondria because they're not excluded from the rep.seqs file.

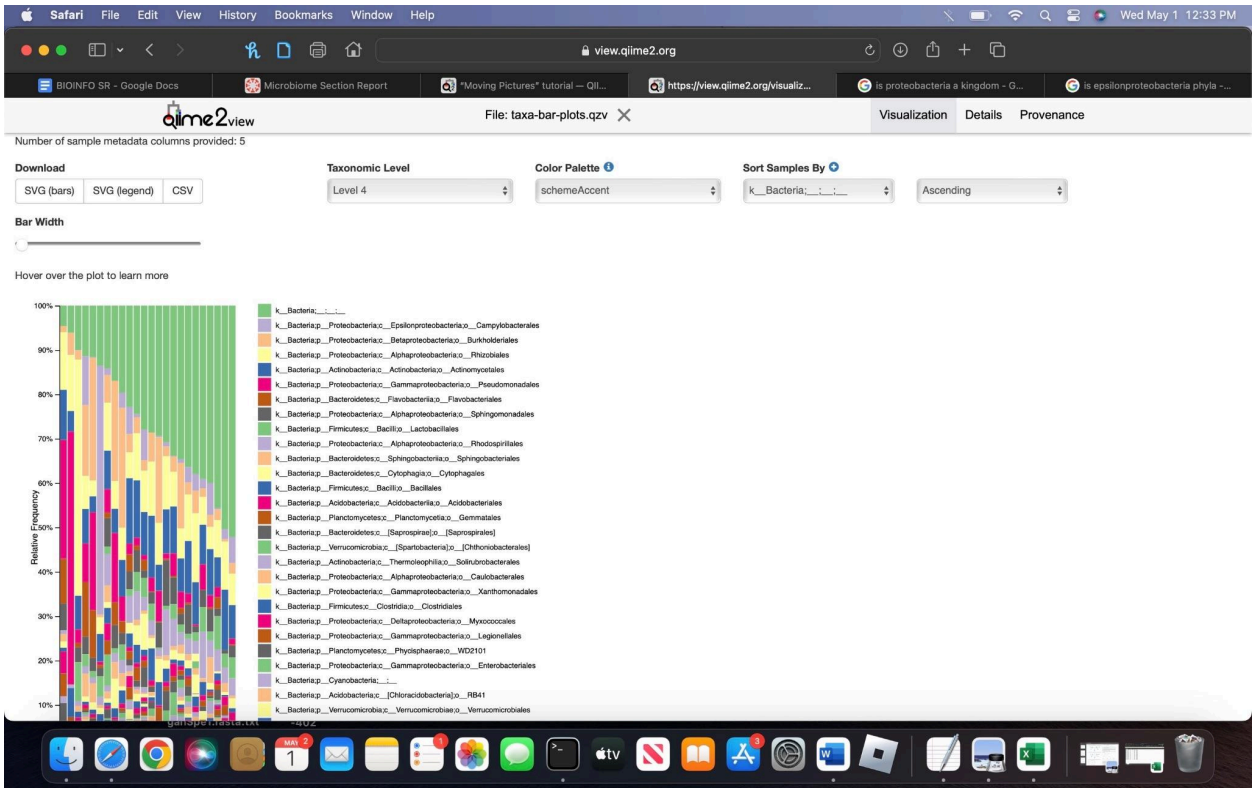
- My new table and taxonomy looks like this:





9) Looking at taxa bar plots, what are your top 2 phyla? Include a screenshot. What are the top 5 most abundant classes? Include a screenshot.

- Since Level 1 is the domain (bacteria), Level 2 would be Kingdom, and Level 3 would be Phyla. My top 2 phyla are Alphaproteobacteria and Epsilonproteobacteria. Since class is the next category, it would be Level 4, which shows the 5 most abundant classes as Campylobacterales, Burkholderiales, Rhizobiales, Actinomycetales, and Pseudomonadales.



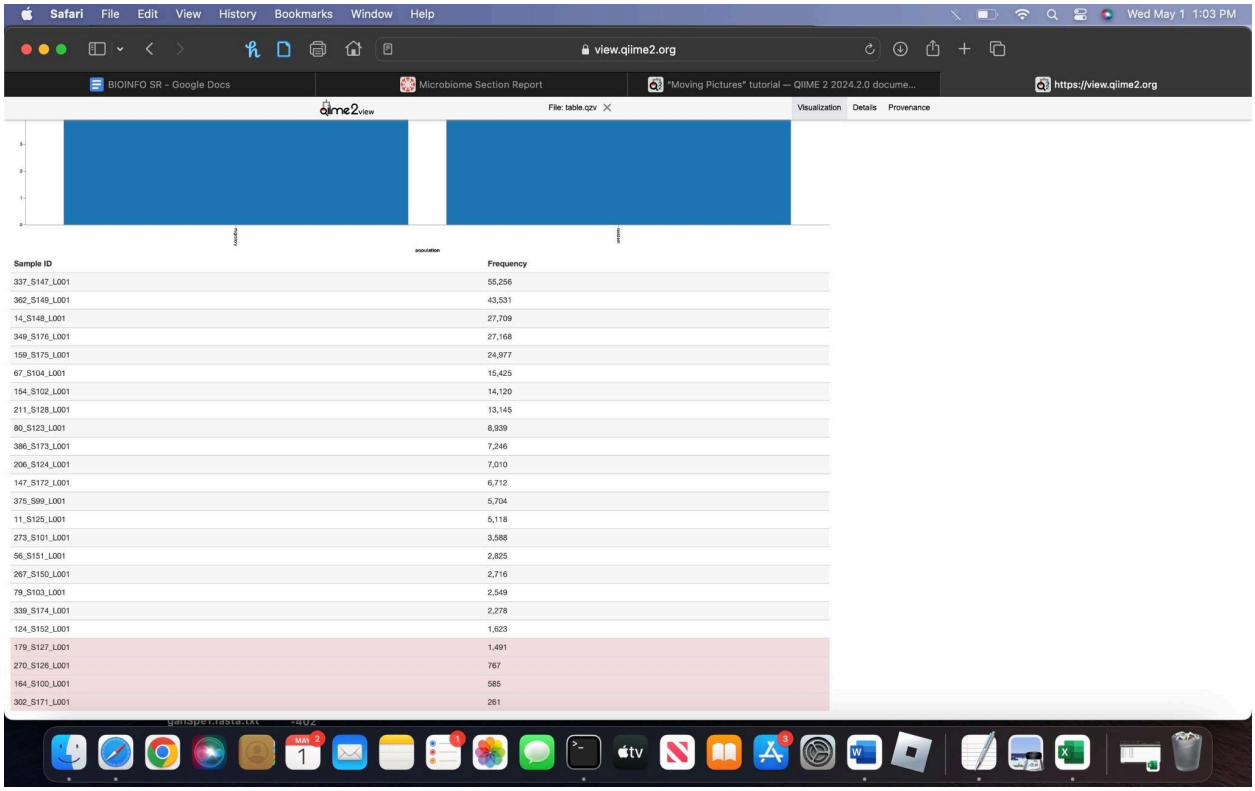
10) What is the difference between alpha and beta diversity? You will have to read outside resources to answer this question. Your response should be in your own words.

- Alpha diversity measures the community richness and evenness within the sample, while beta diversity focuses more on the dissimilarities of the sample, comparing the variability in the community.

11) Before you calculate your diversity metrics, you have to choose a sampling depth. What file previously generated will you use to help you determine what to choose? Defend your choice of sampling depth. How many samples do you retain and how many do you lose?

- The sampling depth I chose from my table.qsv file was 1623. I chose this specific depth because it retains 32,460 (11.56%) features in 20 (83.33%) samples. In

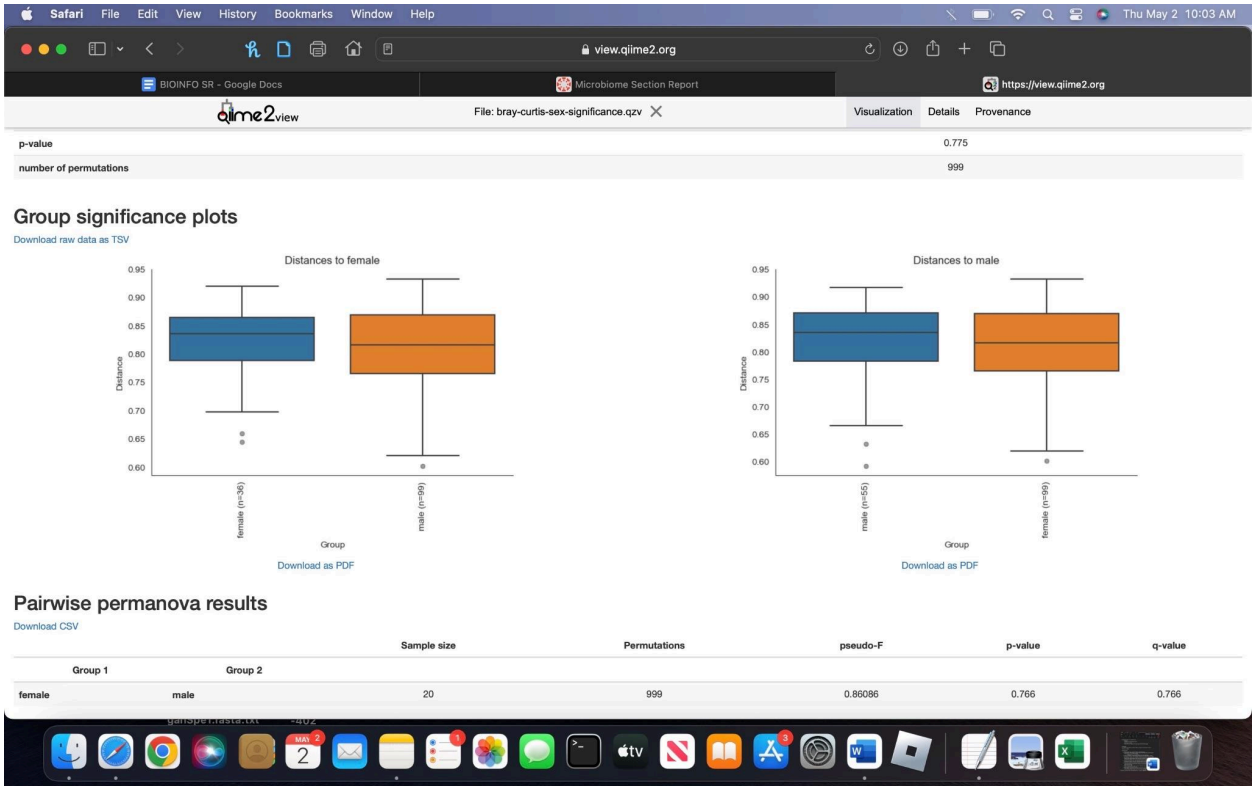
other words, it retains 20 samples with the loss of 4.



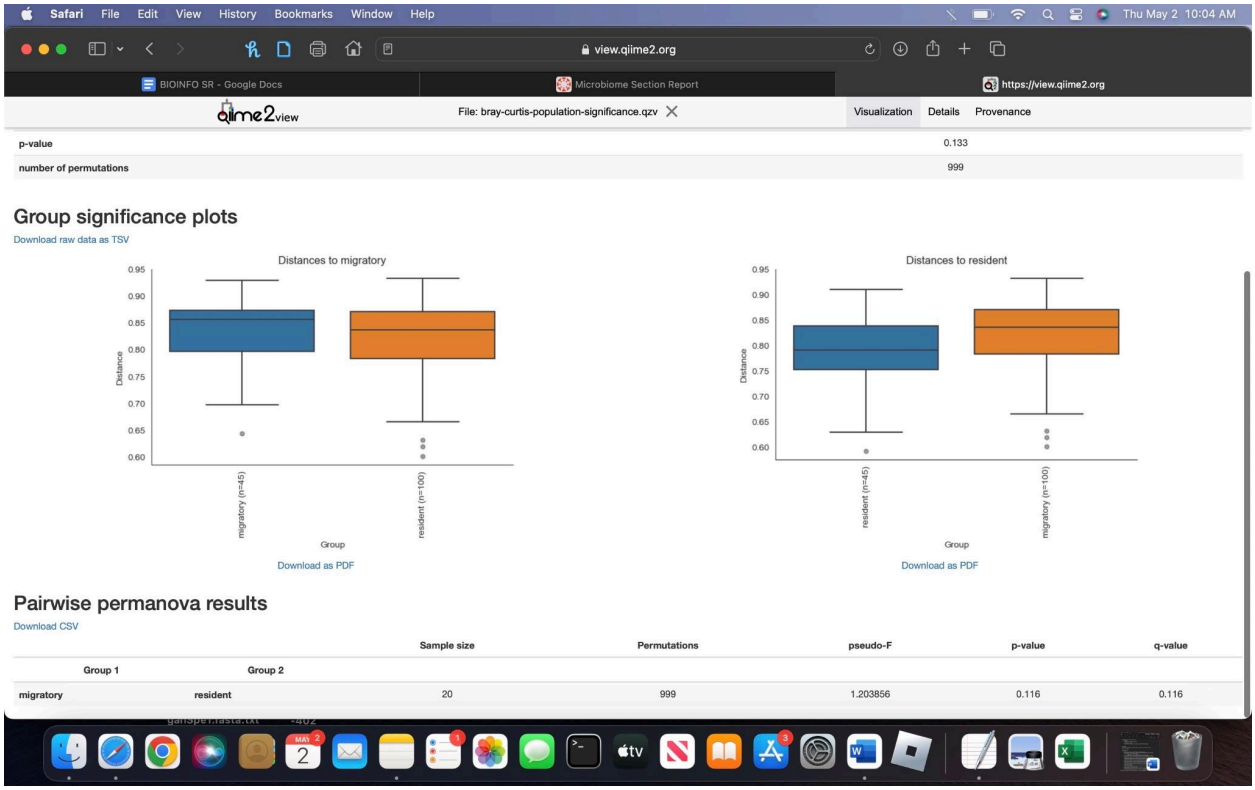
12) For alpha diversity, you need to create visualizations for Shannon diversity and Observed features. This will require you to modify the `alpha-group-significance` code. For which metadata values were graphs generated? Were any of those comparisons significant? How do you know whether they were or were not significant? Briefly describe what Shannon diversity and Observed features are measuring (less than 1 paragraph).

- While Shannon diversity measures the species richness and evenness, the observed features measures the specific features that we focus on in this sample set. When focusing on the population category, the q values were greater than 0.05, at 0.49 and 0.13 for the graphs (Shannon, observed features respectively), showing no significance between migratory and resident. Similarly for sex, the q values were also greater than 0.05 (at 0.07 and 0.25), also demonstrating no significance. For the flock, the q values were also insignificant, though the p value for migratory male-resident female was less than 0.05 for both graphs.

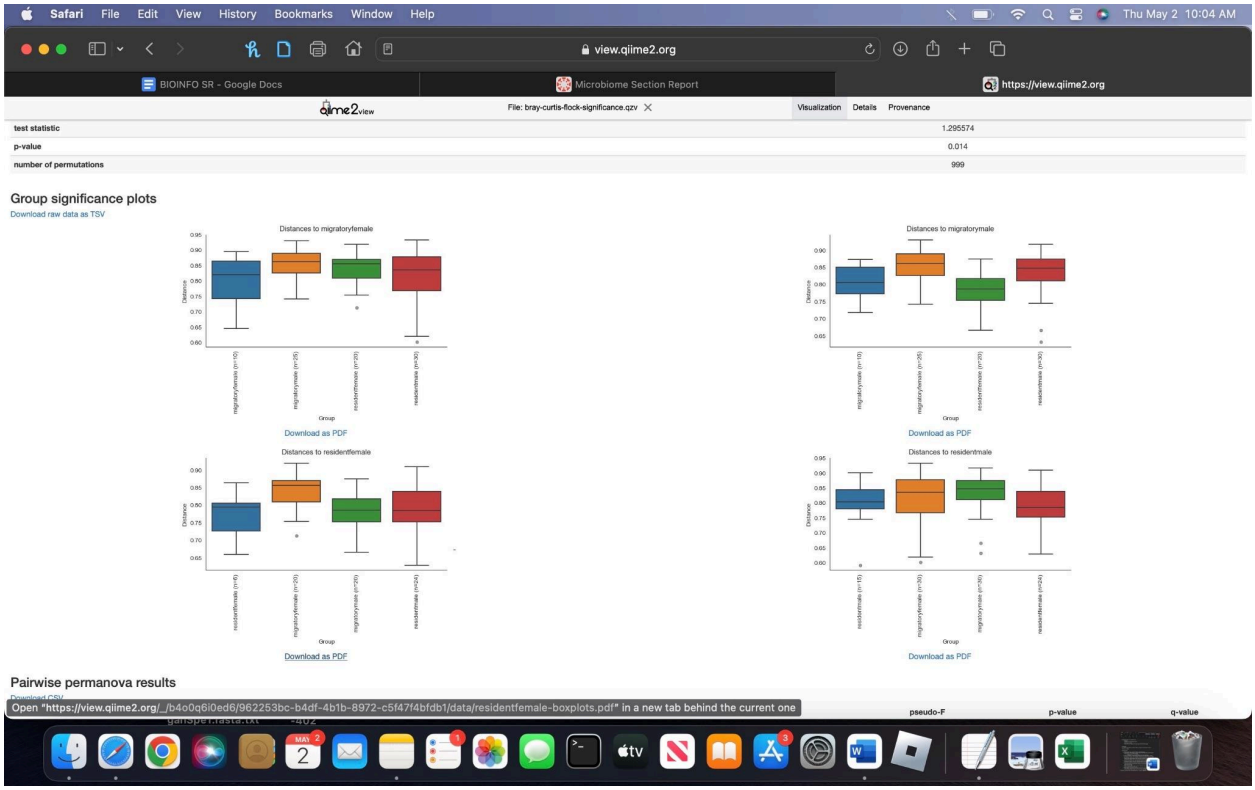
13) For beta diversity, you will need to create visualizations for Bray Curtis dissimilarity. This will require you to modify the `beta-group-significance` code. You should have one visualization for sex, one for population, and one for flock. Include a screenshot of each visualization. Is there any significance? Regardless of significance, how can you interpret these results (hint: what is beta diversity looking at?)

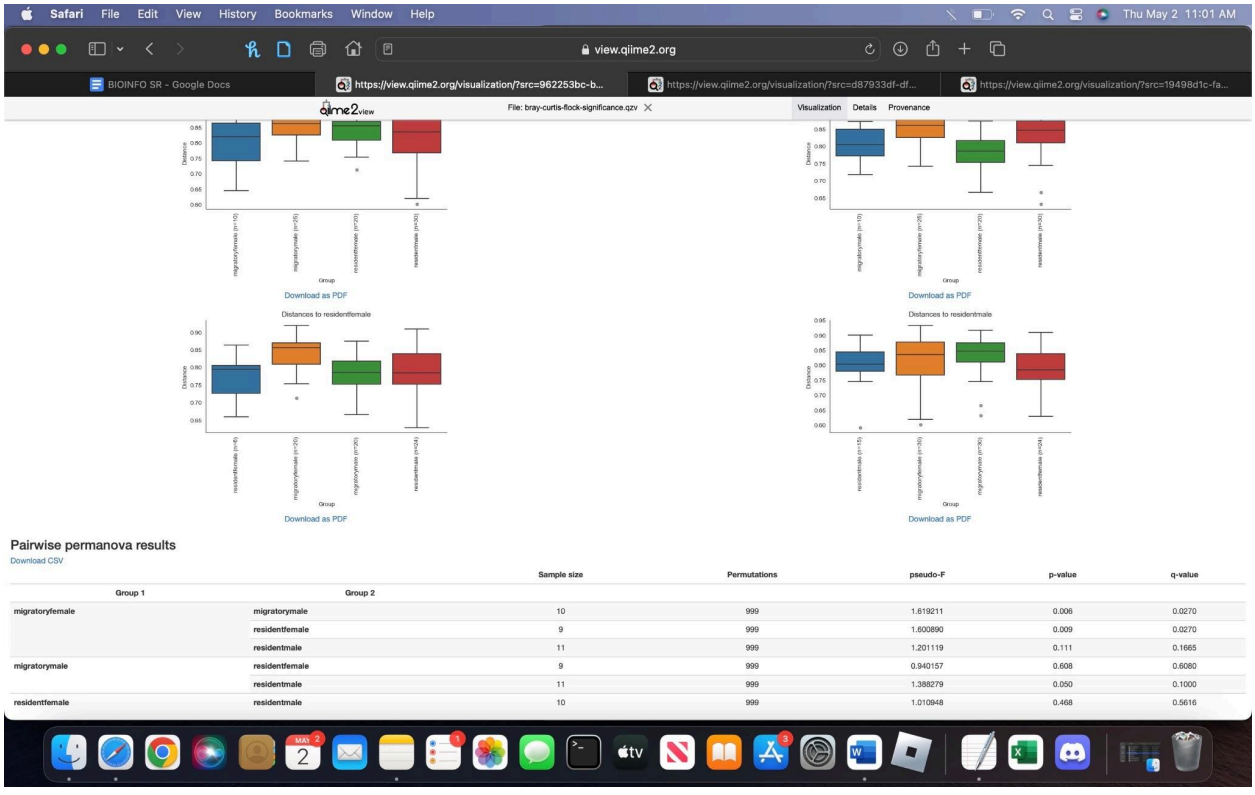


For the visualization for sex, it seems like there are not many dissimilarities between females and males, as seen in the fairly equal ranges for the boxplot, other than the outliers and the max/min. The average distances between the two also seem to be very similar to each other. Judging from the q value of 0.766, the differences are not statistically significant (> 0.05).



For the visualization for population, the boxplots for the resident and migratory categories are still fairly similar, but more dissimilar than the visualization for sex. The q value of 0.116 shows that it is still statistically insignificant, though the population graph seems to be closer to being significant (since the differences are slightly more emphasized).



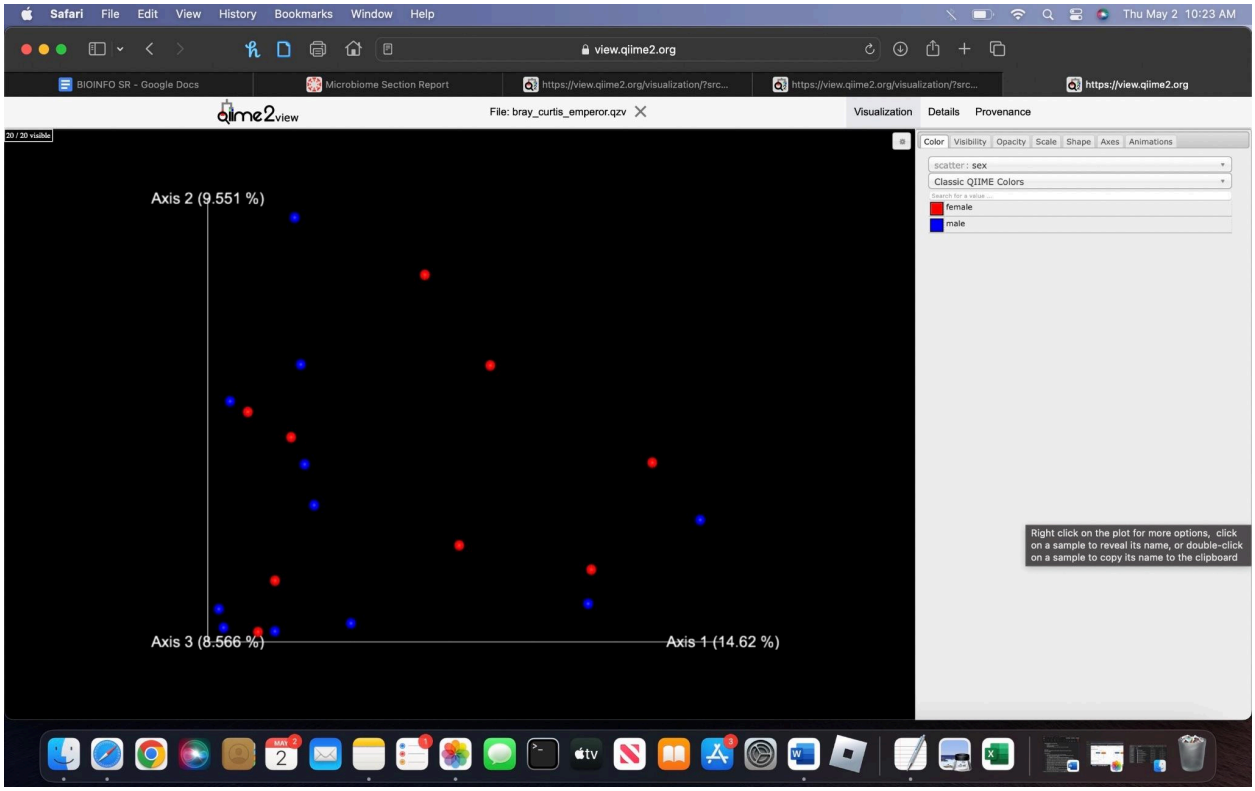


Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
migratoryfemale	migratorymale	10	999	1.619211	0.006	0.0270
	residentfemale	9	999	1.600890	0.009	0.0270
	residentmale	11	999	1.201119	0.111	0.1665
migratorymale	residentfemale	9	999	0.940157	0.608	0.6080
	residentmale	11	999	1.388279	0.050	0.1000
residentfemale	residentmale	10	999	1.010948	0.468	0.5616

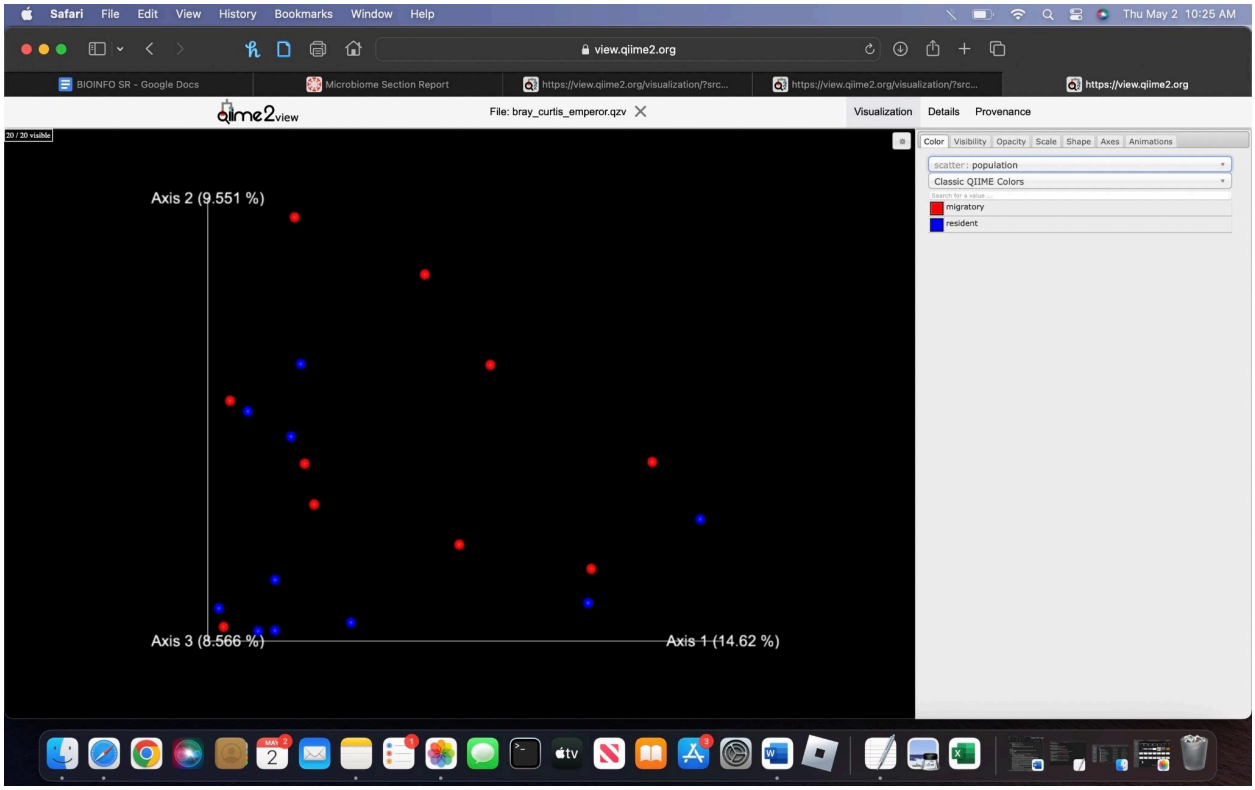
The flock significance shows that the distances between migratory female from both migratory male and resident female are the most significant dissimilarities,

with both the q values being less than 0.05, at 0.0270 for both migratory male and resident male.

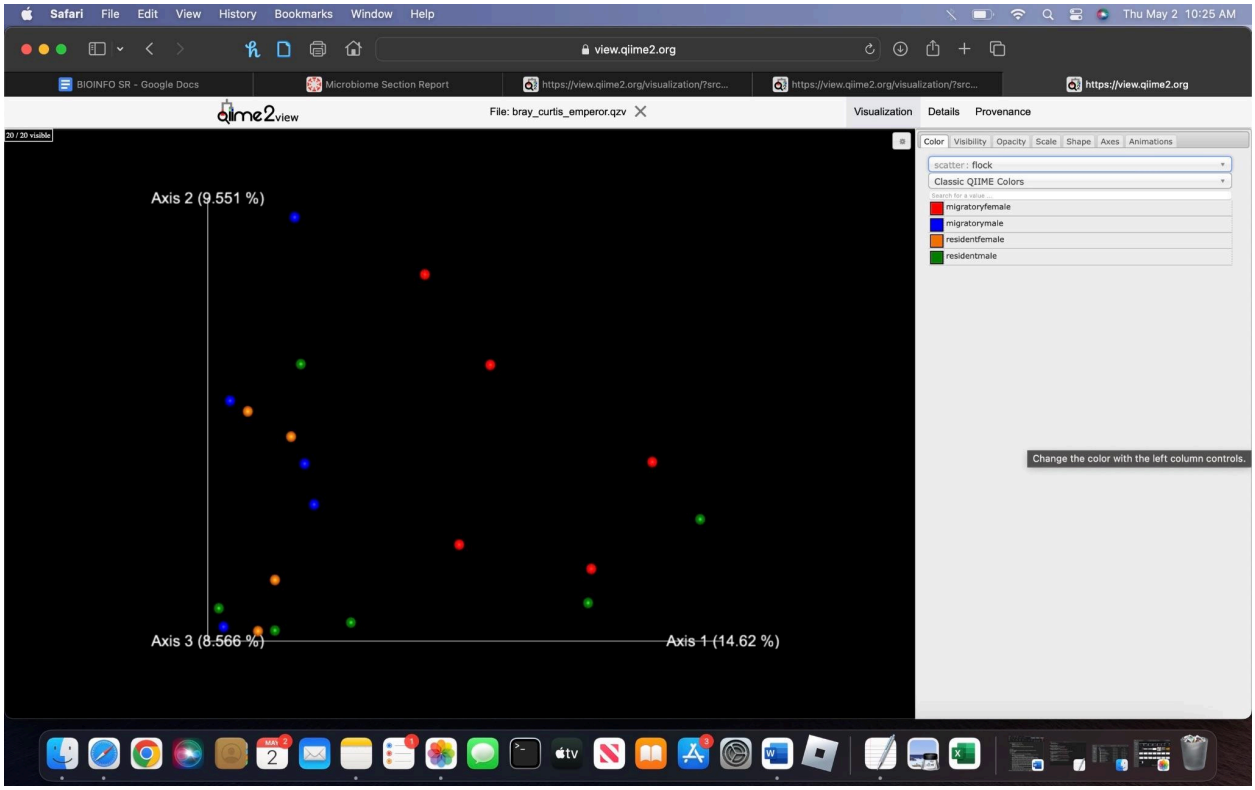
14) The `core-metrics-phylogeny` command generates a file called `bray-curtis-emporer.qzv.` Include 3 screenshots total (1 where the points are colored based on sex, one on population, one on flock). How do these results help you make sense of the results you got from question 13?



- This first visualization (based on sex) shows no real dissimilarities between the females and males, with both being fairly spread out throughout the graph. It helps visualize that there are no significant differences between the sexes.



- The second visualization (based on population) also doesn't show very many differences between migratory and residents, with the red points being similarly scattered throughout the plot as the blue points. Again, there is no significant difference, which correlates with the q values we saw in the boxplots for the previous question.



The third visualization (based on flock) shows more variety in the plot, focusing on 4 different categories. With the color key, we can see that there are differences in the positioning and grouping of each category, such as migratory males, migratory females and resident males all being somewhat spread out, and the resident females being more concentrated in one area of the plot. From the graph and the boxplot (from the previous question), we can see that the distances between migratory female (red) are the most dissimilar to the migratory male (blue) and resident female (orange).