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?*

A screenshot of a graph

Description automatically generated

I chose 20-185 for the forward strand because before 20 and after 185, the quality score decreases significantly. I chose 150 for the reverse because after this point, the quality score decreases dramatically.

1. *How would you modify the code above to truncate and trim in your desired way?*

I changed:

--p-trim-left 0 \

--p-trunc-len 120 \

Into:

--p-trim-left-f 20 \

--p-trunc-len-f 185 \

--p-trim-left-r 0 \

--p-trunc-len-r 150 \

Nothing was taken off of the 5’ end on the reverse strand. However, on the forward strand, 20 bp were taken off the 5’ end, and the strand was cut again at 185 bp.

1. *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?*

qiime dada2 denoise-paired \

--i-demultiplexed-seqs demux.qza \

--p-trim-left-f 0 \

--p-trunc-len-f 185 \

--p-trim-left-r 0 \

--p-trunc-len-r 150 \

--o-representative-sequences rep-seqs-.qza \

--o-table table-.qza \

--o-denoising-stats stats-.qza

By changing the output name, I skipped the renaming step. The qiime metadata tabulate needs to be changed so that the name of the input files does not have .dada2.

1. *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?*

qiime feature-table summarize \

--i-table table.qza \

--o-visualization qiimetable.qzv \

--m-sample-metadata-file metadata.tsv

qiime feature-table tabulate-seqs \

--i-data rep-seqs.qza \

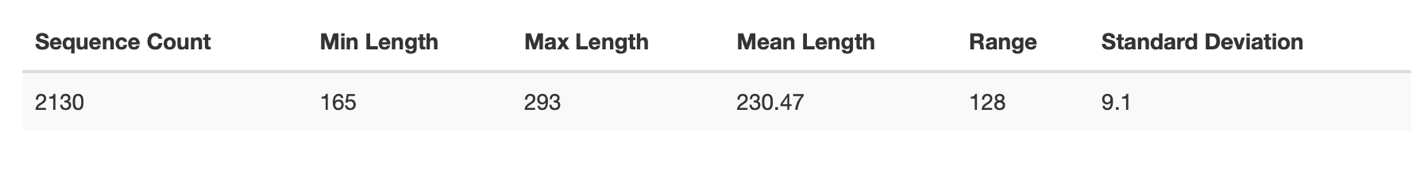
--o-visualization qiimerep-seqs.qzv

I changed the metadata file name from sample-metadata.tsv to metadata.tsv.

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

A screenshot of a graph

Description automatically generated



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

These three are my top hits:

k\_\_Bacteria; p\_\_Chloroflexi; c\_\_Ktedonobacteria; o\_\_JG30-KF-AS9; f\_\_; g\_\_; s\_\_

k\_\_Bacteria; p\_\_Chloroflexi; c\_\_Thermomicrobia; o\_\_Ellin6537; f\_\_; g\_\_; s\_\_

k\_\_Bacteria; p\_\_Proteobacteria; c\_\_Alphaproteobacteria; o\_\_Rickettsiales; f\_\_mitochondria

1. *What do you think this code is doing? Why do you think this is a necessary or important step?*

qiime taxa filter-table \

--i-table table-dada2.qza \

--i-taxonomy taxonomy.qza \

--p-exclude mitochondria,chloroplast \

--o-filtered-table table.qza

I think this code is excluding mitochondria and chloroplast species from the taxonomy. This will ensure that no mammalian or plant DNA is present in the bacterial reads. It is using the table from DADA2 and the taxonomy that I just made as its input. The output is making a new file called table.qza, which excludes mitochondria and chloroplasts, and can be used to make a taxa bar plot.

1. *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.*

These three are my new top hits:

k\_\_Bacteria;p\_\_Proteobacteria;c\_\_Betaproteobacteria;o\_\_Burkholderiales;f\_\_Oxalobacteraceae;g\_\_Janthinobacterium;s\_\_lividum

k\_\_Bacteria;p\_\_Proteobacteria;c\_\_Gammaproteobacteria;o\_\_Pseudomonadales;f\_\_Pseudomonadaceae;g\_\_Pseudomonas;s\_\_

k\_\_Bacteria;p\_\_Bacteroidetes;c\_\_Flavobacteriia;o\_\_Flavobacteriales;f\_\_Flavobacteriaceae;g\_\_Flavobacterium;s\_\_succinicans

I have differences in the bacterial species present because the sequences belonging to mitochondria and chloroplasts have been removed. This changes the frequency of each species.

1. *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.*

The top two phyla are Proteobacteria and Bacteroidetes. The five most abundant classes are Gammaproteobacteria, Betaproteobacteria, Alphaproteobacteria, Flavobacteriia, and Actinobacteria

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

1. *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 is used to describe the species diversity within a single sample. This can tell us the species richness of a sample. Beta diversity compares the species diversity of different samples. This value tells us how similar two different samples are in terms of diversity.

1. *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 qiimetable.qzv file helped me determine the sampling depth. After this sampling depth, the frequency drops dramatically from 1,226 to 864, and shortly down to 523. Stopping at 1,226 will yield the best results. Twenty samples were retained, four were lost.

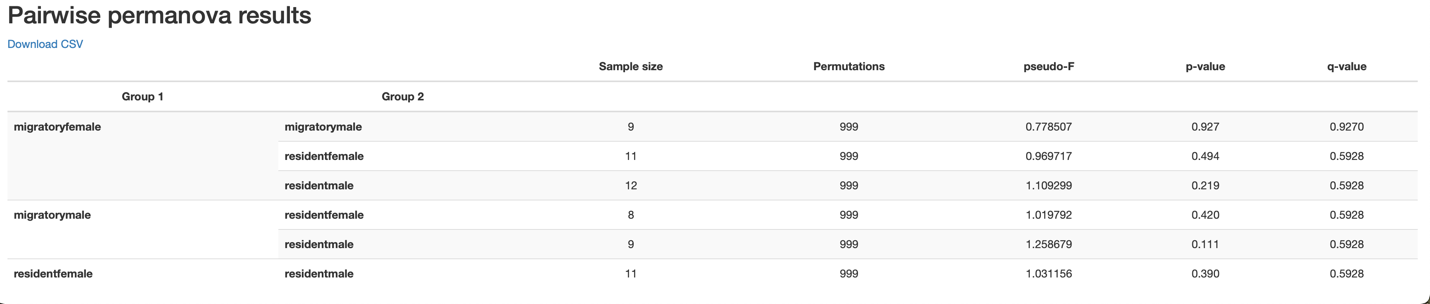
1. *For alpha diversity, you need to create visualizations for Shannon diversity and Observed features. This will require you to modify the alpha-group-significance code. For which metadata values were graphs generated? Were any of those comparisons significant? How do you know whether they were or were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

Population, sex, and flock were used from the metadata file. None of the comparisons were significant because the p value was greater than 0.005.

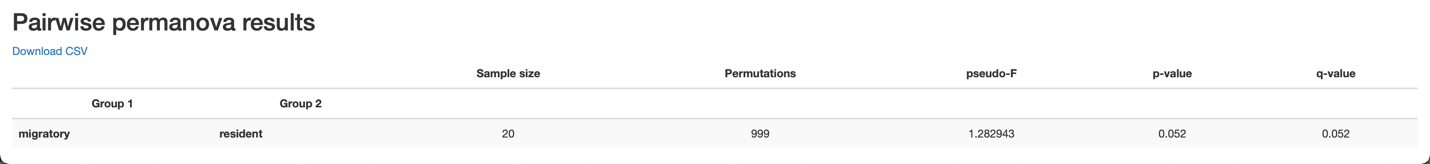
The observed features measure the species richness only. This measures how many species are present in the samples. The Shannon diversity metrics takes in account the evenness, or relative abundance, of each species. A higher Shannon diversity value means a more diverse community. For example, if a sample was composed of 99% one species, and the remaining 1% was split between 100 different species, the observed features measure may be larger than the Shannon diversity value.

1. *For beta diversity, you will need to create visualizations for Bray Curtis dissimilarity. This will require your 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?)*

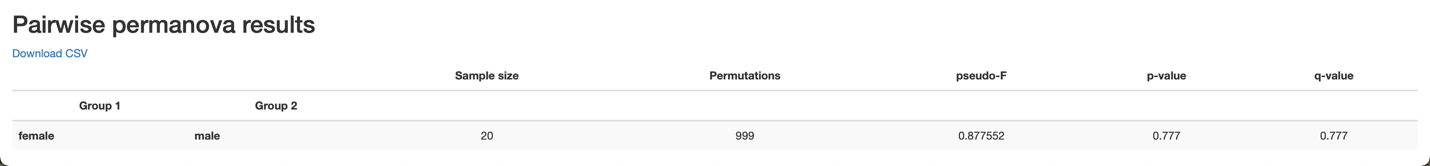
There are no significant differences in the beta diversity of bacteria in these samples based on flock, population, or sex. The closest significant value was in the population, where the p value was 0.0052. This makes sense because the results are showing the difference in bacterial microbiome diversity between different birds from different places. It would make sense, then, that different populations of birds have the most different microbiomes of the three metrics. However, birds of the same flock, regardless of sex, have similar microbiomes, as indicated by the p and q values.



*Flock Visualization*



*Population Visualization*



*Sex Visualization*

1. *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?*

These results help visualize the tables shown above. Because there is not a lot of clustering based on these three parameters (flock, population, sex), the microbiomes between these birds are not dissimilar enough to be statistically significant. As seen above, the p value for population is almost significant. In the population graph, there is some clustering of birds from the same population, but it is not very pronounced, which supports the p value calculated above. The other two, flock and sex, are not very clustered, supporting the idea that the difference in microbiome diversity are not statistically significant.

