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 graph of a graph of a graph

Description automatically generated with medium confidence

Both the forward and reverse reads have low quality reads on both sides of the quality plot, as expected. For the forward reads, the quality before 20 and after 190 on the x-axis is low which is why I trimmed both ends off at those values. For the reverse reads, the quality before 40 and after 160 is low, so I chose these two numbers as my values for the trimming step.

**For questions 2 and 3:** Because these are paired-end reads, you will have to modify the dada2 code in order to perform the quality trimming on both the forward and reverse reads. You will not do the deblur. You will need to adjust this code to account for --p-trunc-len and --p-trim-left for both the forward and reverse reads. The basics of the code you need to change are here.

qiime dada2 denoise-paired \  
  --i-demultiplexed-seqs demux.qza \  
  --p-trim-left-f  \  
  --p-trunc-len-f  \  
  --p-trim-left-r  \  
  --p-trunc-len-r  \  
  --o-representative-sequences rep-seqs-dada2.qza \  
  --o-table table-dada2.qza \  
  --o-denoising-stats stats-dada2.qza

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

To modify the code, I added values after the trim commands.

qiime dada2 denoise-paired \

--i-demultiplexed-seqs demux.qza \

--p-trim-left-f 20 \

--p-trim-left-r 40 \

--p-trunc-len-f 190 \

--p-trunc-len-r 160 \

--o-representative-sequences rep-seqs-dada2.qza \

--o-table table-dada2.qza \

--o-denoising-stats stats-dada2.qza

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?**

I could modify the last two lines as follows:

qiime dada2 denoise-paired \

--i-demultiplexed-seqs demux.qza \

--p-trim-left-f 20 \

--p-trim-left-r 40 \

--p-trunc-len-f 190 \

--p-trunc-len-r 160 \

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

--o-table table.qza \

--o-denoising-stats stats-dada2.qza

Here, I changed the output names of the files to avoid using the mv command.

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?**

In my code, I changed sample-metadata.tsv to metadata.tsv:

(qiime2-amplicon-2024.2) jairtorres@Jairs-MacBook-Pro MicrobiomeQiime2 % qiime feature-table summarize \

--i-table table.qza \

--o-visualization table.qzv \

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

qiime feature-table tabulate-seqs \

--i-data rep-seqs.qza \

--o-visualization rep-seqs.qzv

1. **A screenshot of a phone

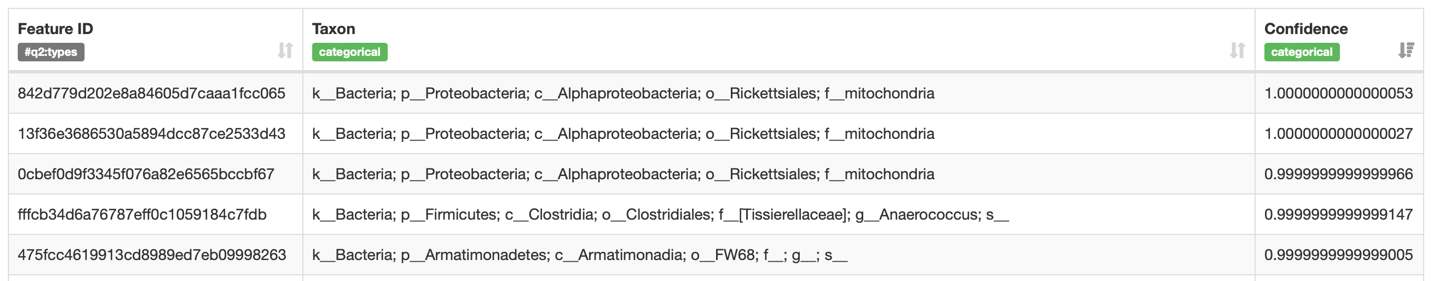
   Description automatically generatedInclude a screenshot of the table summary from visualizing your table and a screenshot of the sequence length statistics from the rep-seqs file.**

A screenshot of a statistics

Description automatically generated

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

My top three hits are members of order Rickettsiales. These hits are followed by members of order Clostridiales.



**For question 7**: Run this code

qiime taxa filter-table \  
  --i-table table.qza \  
  --i-taxonomy taxonomy.qza \  
  --p-exclude mitochondria,chloroplast \  
  --o-filtered-table table.qza

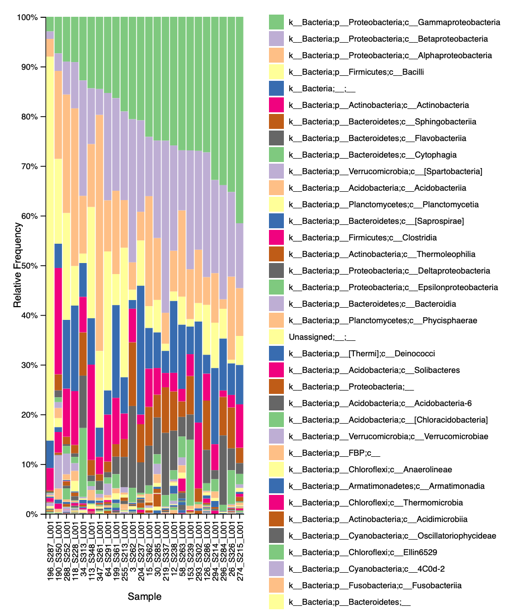
1. **What do you think this code is doing? Why do you think this is a necessary or important step?**

This code is removing the mitochondrial and chloroplasts from the table which is important because bacteria do not contain these organelles. These sequences may appear because of contamination from the eukaryotic host that the samples were taken from. Also, according to endosymbiotic theory, mitochondria and chloroplasts have a bacterial origin and have similar DNA.

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.**

No, there were no differences. My top three hits still had mitochondrial sequences. This makes sense since the above code only modified the table file, not the sequence file.

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 2 phyla are Proteobacteria and Firmicutes. The top 5 most abundant classes were Gammaproteobacteria, Betaproteobacteria, Alphaproteobacteria, Bacilli, and Actinobacteria.

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 a measure of diversity within one community while beta diversity is a measure that compares the diversity between different communities. Also, beta diversity measures distance with a higher value indicating more different diversities.

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?**

I am using the table.qzv file to determine the sequencing depth to use. Using the interactive detail, I found that 845 was a good value since not too many data points are lost. Specifically, 11 samples are retained while 1 is lost. In other words, most samples are retained.

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).**

The Shannon diversity graphs were generated for the population column in the metadata. Unlike the Observed features, Shannon diversity measures the evenness of species in a sample in addition to richness. While the Shannon diversity index measures the diversity of a single community, the subsequent Kruskal-Wallis test compares the alpha diversities for two groups. In this case, the null hypothesis is that the alpha diversities of migratory and resident birds are the same. Since the p-value of 0.758289 was larger than 0.05, we fail to reject the null hypothesis. There is not enough evidence to say that the alpha diversities of migratory and resident birds are different.

A white rectangular object with black text

Description automatically generated

A white rectangular object with a black stripe

Description automatically generated

For observed features, again graphs were generated for the population column in the metadata. Unlike the Shannon diversity index, observed features only measures species richness, that is, the total number of species in samples. Here, the p-value of 0.644 was larger than 0.05 which means that there is no significant difference between the alpha diversities of migratory and resident groups.

A diagram of a diagram

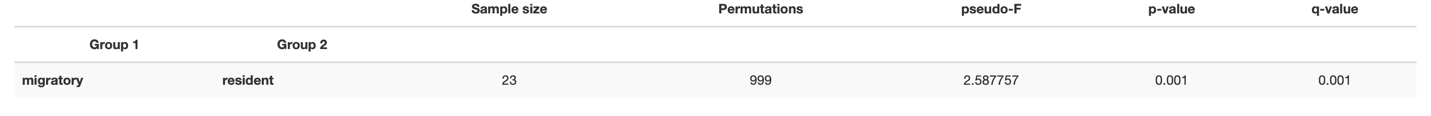
Description automatically generated with medium confidenceA white rectangular object with a white background

Description automatically generated

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?)**

Without even considering significance, beta diversity can be interpreted by how close the value is to 0 or 1, with 0 indicating highly similar communities and 1 indicating highly dissimilar communities.

The beta diversities of male and female birds were significantly different (0.028 < 0.05). The beta diversities for migratory and resident birds were also significantly different (0.001 < 0.05). Finally, the beta diversities for all flocks were also significantly different (p < 0.05).

A diagram of a bar chart

Description automatically generated with medium confidenceA comparison of a diagram

Description automatically generated with medium confidence

A group of colored boxes

Description automatically generatedA screenshot of a white list

Description automatically generated

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?**

Because beta diversities were significantly different between sexes, population, and flocks, we should expect clustering of the subgroups within each of these categories and this is exactly what we observe. When the dots are colored by sex, the red and blue dots seem to form different groups on the graph with very few overlapping points. The red and blue dots also seem clustered by color in the population graph. In the flock graph, each color representing the flock that the sample was taken from clusters with other dots of the same color.

A white panel with black lines

Description automatically generated with medium confidenceA screen shot of a computer

Description automatically generated

A close-up of a white panel

Description automatically generatedA screen shot of a graph

Description automatically generated

**A screenshot of a phone

Description automatically generatedA screen shot of a computer

Description automatically generated**

**How to format your code:**

Please submit a document of all the code you used. You do not need to include the output of the code, just the code itself. For each chunk of code, you should include an explanation of what that code is doing.

Here is an example:

#the function of this code is to generate a visualization of my table file.  
#it uses my metadata and the table artifact I generated earlier to create that visualization.  
#the visualization will include information that will allow me to calculate my sampling depth.  
qiime feature-table summarize \

--i-table table.qza \

--o-visualization table.qzv \

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