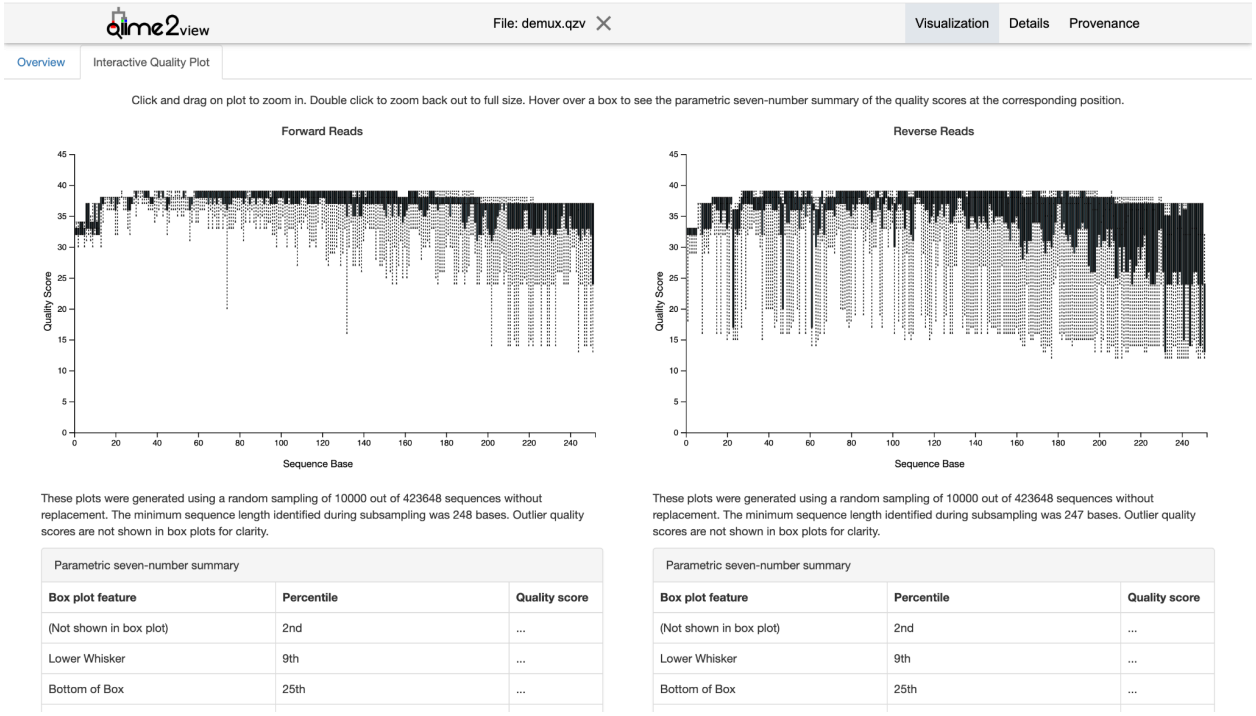
1) Include a screenshot of your interactive quality plot. Based on this plot, what values would you choose for `--p-trunc-len` and `--p-trim-left` for both the forward and reverse reads? Why have you chosen those numbers?

I chose these values based on where I saw the quality score dip. Generally the quality score remained high across the sequence, but I cut off just a short amount from the beginning and end. So, I chose the following values for both my forward and reverse reads quality plot, trim 25 and trunc 195.



These plots were generated using a random sampling of 10000 out of 423648 sequences without replacement. The minimum sequence length identified during subsampling was 248 bases. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary

Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	...
Lower Whisker	9th	...
Bottom of Box	25th	...

These plots were generated using a random sampling of 10000 out of 423648 sequences without replacement. The minimum sequence length identified during subsampling was 247 bases. Outlier quality scores are not shown in box plots for clarity.

Parametric seven-number summary

Box plot feature	Percentile	Quality score
(Not shown in box plot)	2nd	...
Lower Whisker	9th	...
Bottom of Box	25th	...

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
```

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

After each of the trim and trunc lines of code, I added in the values I mentioned above. The third and fourth lines have an f added to the end indicating that they correspond with the forward reads. So, after this f I added in where I wanted my forward reads to be trimmed and truncated.

--p-trim-left-f 25 \

--p-trunc-len-f 195 \

The fifth and sixth lines have an r added to the end indicating that they correspond with the reverse reads. So, after this r, I added in where I wanted my reverse reads to be trimmed and truncated.

--p-trim-left-r 25 \

--p-trunc-len-r 195 \

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 could simply delete the “dada2” portion of the code above to skip the mv step. This would result in the following code.

```
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.qza \
  --o-table table.qza \
  --o-denoising-stats stats.qza
```

To account for the renamed files, I need to make the same changes in my tabulate code. As above, I will simply delete the “dada2” portion resulting in the following code.

qiime metadata tabulate \

--m-input-file stats.qza \

--o-visualization 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?

When creating my FeatureTable and my FeatureData summaries I renamed the sample metadata (--m-sample-metadata-file) to match the file I have on my computer. This resulted in the following code.

qiime feature-table summarize \

--i-table table.qza \

--o-visualization table.qzv \

--m-sample-metadata-file metadata.txt

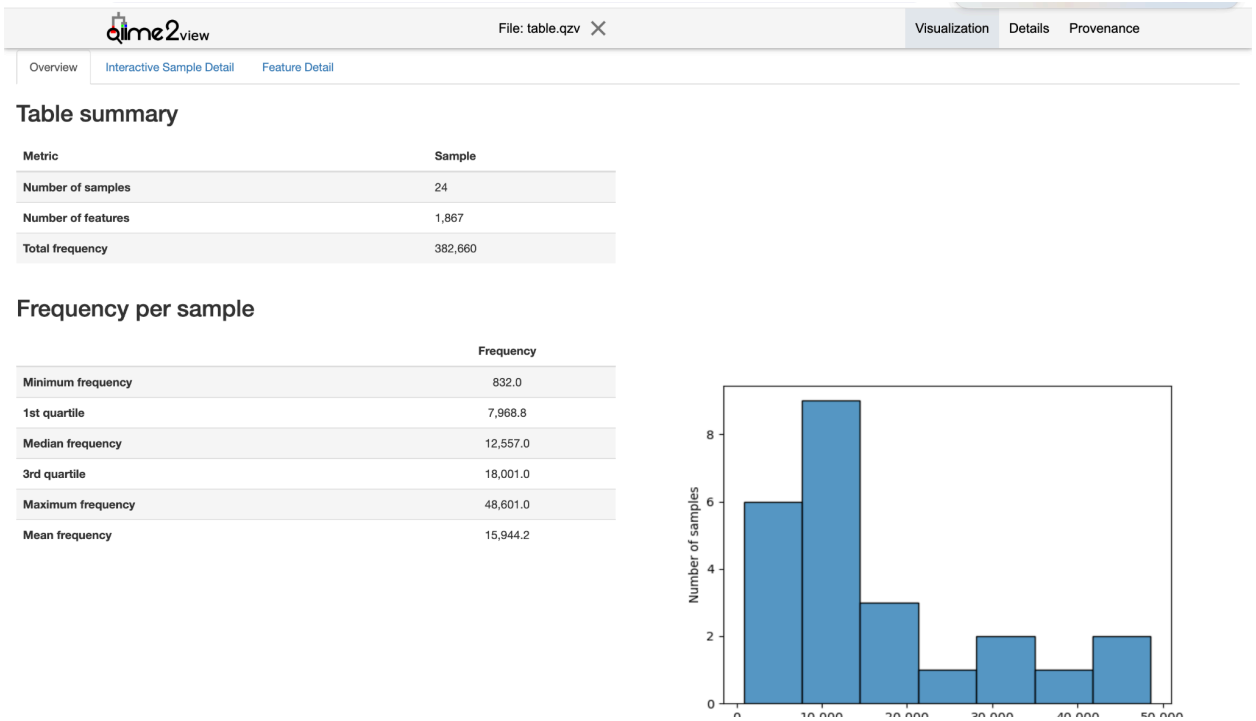
qiime feature-table tabulate-seqs \

--i-data rep-seqs.qza \

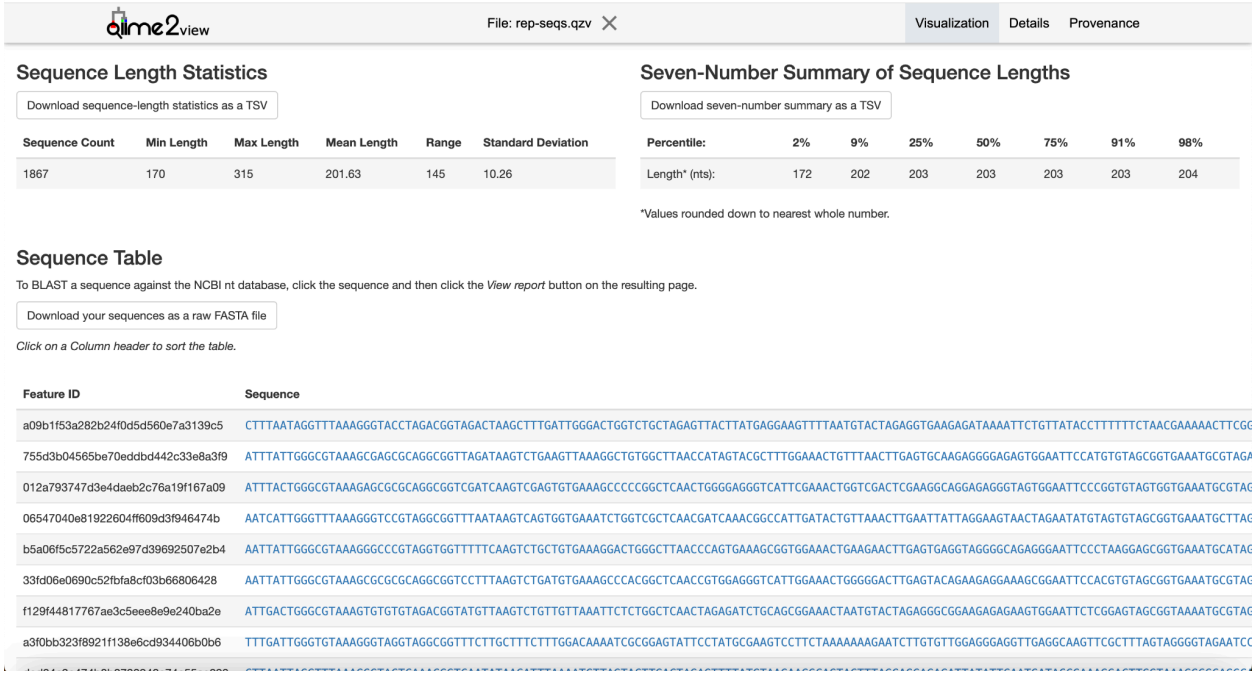
--o-visualization rep-seqs.qzv

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.

Table Summary:



Sequence length statistics:



qiime2view File: rep-seqs.qzv Visualization Details Provenance

Sequence Length Statistics

Download sequence-length statistics as a TSV

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
1867	170	315	201.63	145	10.26

Seven-Number Summary of Sequence Lengths

Download seven-number summary as a TSV

Percentile:	2%	9%	25%	50%	75%	91%	98%
Length* (nts):	172	202	203	203	203	203	204

*Values rounded down to nearest whole number.

Sequence Table

To BLAST a sequence against the NCBI nt database, click the sequence and then click the *View report* button on the resulting page.

Download your sequences as a raw FASTA file

Click on a Column header to sort the table.

Feature ID	Sequence
a09b1f53a282b24f0d5d560e7a3139c5	CTTTAATAGGTTTAAAGGGTACCTAGACGGTAGACTAAGCTTTGATTGGGACTGGTCTGCTAGAGTTACTTATGAGGAAGTTTTAATGTACTAGAGGTGAAGAGATAAAATTCTGTTATACCTTTTTTCTAACGAAAAACTTCGG
755d3b04565be70eddbd442c33e8a3f9	ATTTATTGGGCGTAAAGCGAGCGCAGGCGGTTAGATAAGTCTGAAGTTAAAGGCTGTGGCTTAACCATAGTACGCTTTGGAAACTGTTTAACTTGAGTGCAAGAGGGGAGAGTGGAATTCCATGTGTAGCGGTGAAATGCGTAGA
012a793747d3e4daeb2c76a19f167a09	ATTTACTGGGCGTAAAGAGCGCGCAGGCGGTCGATCAAGTCGAGTGTGAAAGCCCCCGGCTCAACTGGGGAGGGTCATTCGAAACTGGTCGACTCGAAGGCAGGAGAGGGTAGTGGAATTCCCGGTGTAGTGGTGAAATGCGTAG
06547040e81922604ff609d3f946474b	AATCATTGGGTTTAAAGGGTCCGTAGGCGGTTTAATAAGTCAGTGGTGAAATCTGGTCGCTCAACGATCAAACGGCCATTGATACTGTTAAACTTGAATTATTAGGAAGTAACTAGAATATGTAGTGTAGCGGTGAAATGCTTAG
b5a06f5c5722a562e97d39692507e2b4	AATTATTGGGCGTAAAGGGCCCGTAGGTGGTTTTTCAAGTCTGCTGTGAAAGGACTGGGCTTAACCCAGTGAAAGCGGTGGAAACTGAAGAACTTGAGTGAGGTAGGGGCAGAGGGAATTCCCTAAGGAGCGGTGAAATGCATAG
33fd06e0690c52fbfa8cf03b66806428	AATTATTGGGCGTAAAGCGCGCGCAGGCGGTCCTTTAAGTCTGATGTGAAAGCCCACGGCTCAACCGTGGAGGGTCATTGGAAACTGGGGGACTTGAGTACAGAAGAGGAAAGCGGAATTCCACGTGTAGCGGTGAAATGCGTAG
f129f44817767ae3c5eee8e9e240ba2e	ATTGACTGGGCGTAAAGTGTGTGTAGACGGTATGTTAAGTCTGTTGTTAAATTCTCTGGCTCAACTAGAGATCTGCAGCGGAAACTAATGTACTAGAGGGCGGAAGAGAGAAGTGGAATTCTCGGAGTAGCGGTAAAATGCGTAG
a3f0bb323f8921f138e6cd934406b0b6	TTTGATTGGGTGTAAAGGGTAGGTAGGCGGTTTCTTGCTTTCTTTGGACAAAATCGCGGAGTATTCCTATGCGAAGTCCTTCTAAAAAAAGAATCTTGTGTTGGAGGGAGGTTGAGGCAAGTTCGCTTTAGTAGGGGTAGAATCC

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

After sorting by confidence, these were the top hits I received:

Feature ID (#q2:types)	Taxon (categorical)	Confidence (categorical)
7151df6aa4b28f3d27bb2b921b64c9c9	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000069
f7f1ee400cd2dcc4a654cd0cd83a0817	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000044
07866dfe05e3d1765fa9e77b11938754	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000036
7dc53b99adac9ca03be81bcbd584efea	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	0.9999999999999999
8a99154da5740a95fd197367d688d36f	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	0.9999999999999977
7408b91d51d194ed958eea41460fd2b2	k__Bacteria; p__Actinobacteria; c__Actinobacteria; o__Actinomycetales	0.9999999999999645
eed0e0dba9ea2181edf62bb0e0e47685	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__MIZ46; f__; g__; s__	0.9999999999999432
41a6a91be42cfe95613d1406244918eb	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Pasteurellales; f__Pasteurellaceae	0.9999999999998316
f17ba65db7af0a6b89df3d2c733fd406	k__Bacteria; p__Bacteroidetes; c__Cytophagia; o__Cytophagales; f__Cytophagaceae; g__Rudanella; s__	0.9999999999998295
36a30fdc9a94221180c1a261076da5c0	k__Bacteria; p__Chloroflexi; c__Thermomicrobia; o__Ellin6537; f__; g__; s__	0.9999999999997726
63a33ba81c0f584e15e217b7a7c32800	k__Bacteria; p__Fusobacteria; c__Fusobacteriia; o__Fusobacteriales; f__Leptotrichiaceae; g__Leptotrichia; s__	0.9999999999995595
a63e4d209a11fba31c55e9b36d25e836	k__Bacteria; p__[Thermi]; c__Deinococci; o__Thermales; f__Thermaceae; g__Thermus; s__	0.9999999999991047
ab23bc3c57563727f7df6509bc29e818	k__Bacteria; p__FBP; c__; o__; f__; g__; s__	0.9999999999986926
2be4e916ed08f3db0a6cb0ed14beba95	k__Bacteria; p__Fibrobacteres; c__Fibrobacteria; o__258ds10; f__; g__; s__	0.9999999999985789
7bc161438ee22554539f9e5fcab1beb1	k__Bacteria; p__Proteobacteria; c__Gammaproteobacteria; o__Legionellales; f__Coxiellaceae; g__Rickettsiella; s__	0.9999999999982947

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
```

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

This code is removing any results that are coming from mitochondria or chloroplast DNA. This is not the DNA we are looking to analyze, so we need to filter it out. If we do not, it will throw off our results.

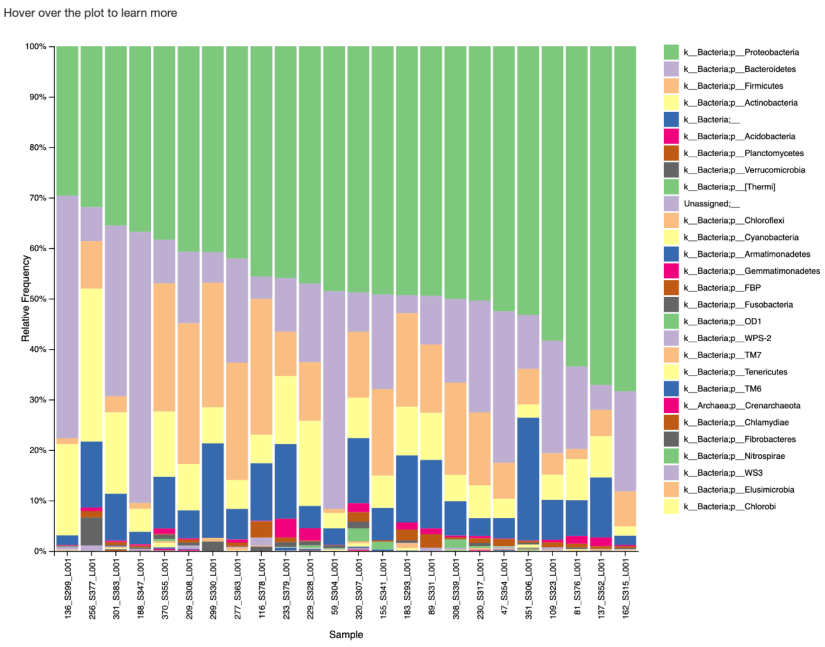
8) Re-do your table visualization and re-do your taxonomy commands. Do you have any differences now in the hits with the highest confidence? Why or why not? Really think about what the code is doing.

There are no differences in the hits with the highest confidence. This is because the code that we used to filter used the input files of table.qza dn taxonomy.qza while the original input files used to make the taxonomy were rep-seqs.qza and taxonomy.qza. The filtering made no change to the rep-seqs.qza file.

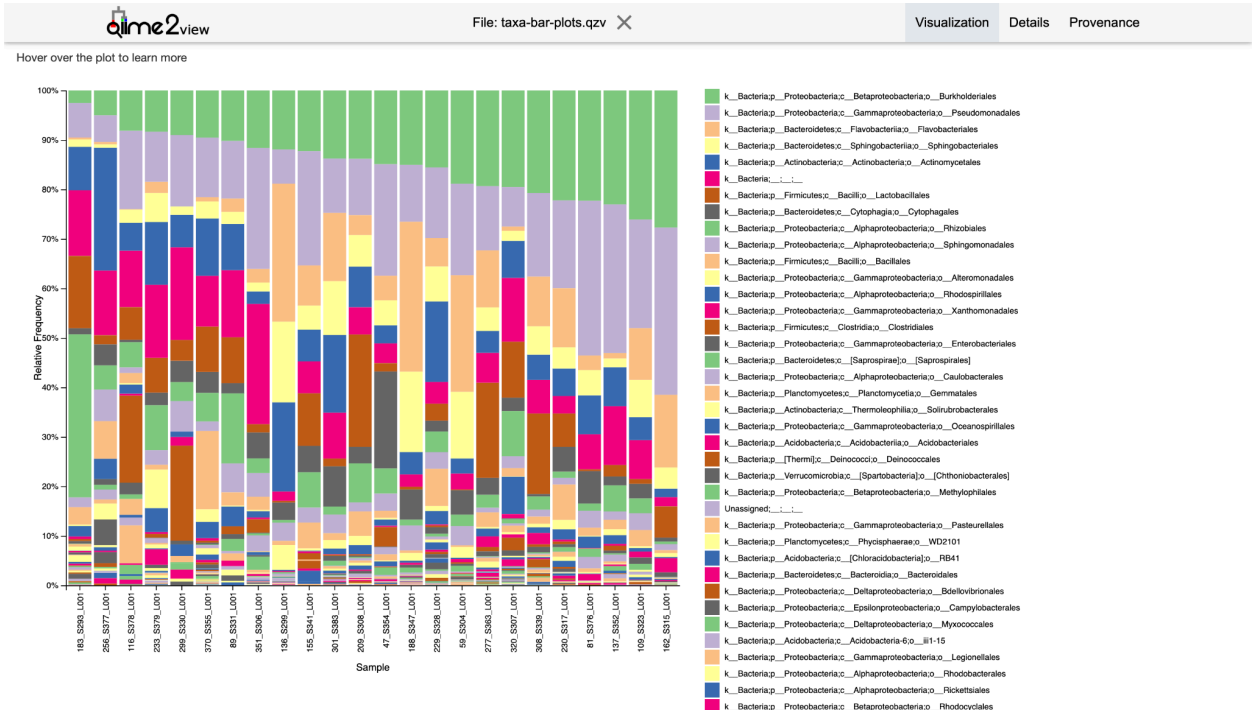
9) Looking at taxa bar plots, what are your top 2 phyla? Include a screenshot. What are the top 5 most abundant classes? Include a screenshot.

My top 2 phyla are Gammaproteobacteria and Betaproteobacteria. The top 5 most abundant classes are Burkholderiales, Pseudomonadales, Flavobacteriales, Sphingobacteriales, and Actinomycetales.

Phyla:



Classes:



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 is made up of two components: richness and evenness. Richness is simply the number of species present and evenness accounts for the amount of each species. For example, there could be 2 species present (richness) and there are 5 of each (evenness). On the other hand there could be 2 species present (same richness) but there is only 1 of one species and 30 of the other species (very different evenness). While alpha diversity describes the composition of a single community, beta diversity compares two separate communities. Beta diversity tells us if the species present in each community are the same or different. For example, there could be two communities each with the same richness and evenness. However, if the two communities are made up of the same species, then they will have low beta diversity. If they are made up of entirely different species then they will have lots of variability, and therefore high beta diversity.

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 table.qzv file will help me determine which sampling depth to choose. I am choosing 1,200. Past this number of reads there is a significant dropoff in read length.

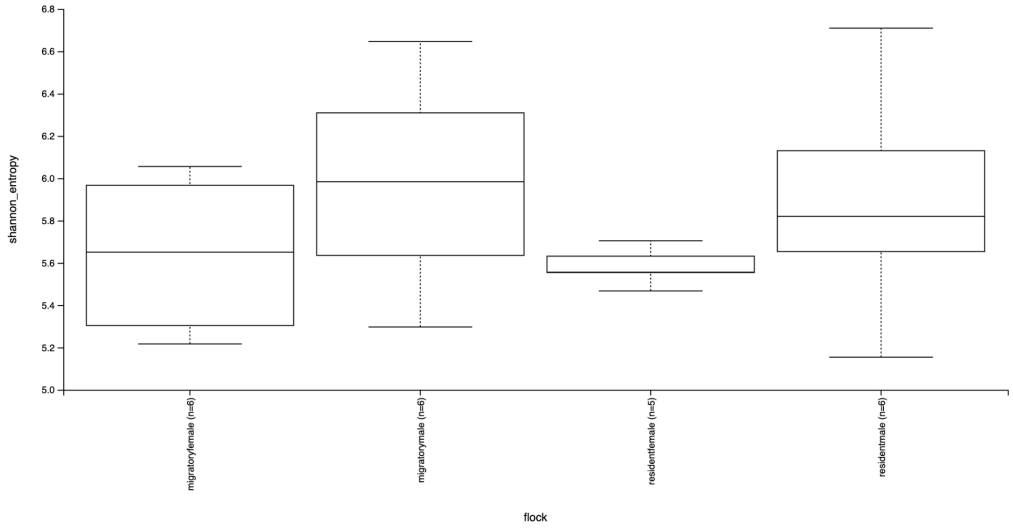
This number still includes most of the reads though. So, it is a good balance between including enough diversity while excluding low quality reads. This excludes 11 samples and retains 13. If I were to redo this I would choose a slightly lower value to retain a few more samples. However, once the code was already run once, it could not be changed. So, I decided to stick with 1,200 because it is acceptable.

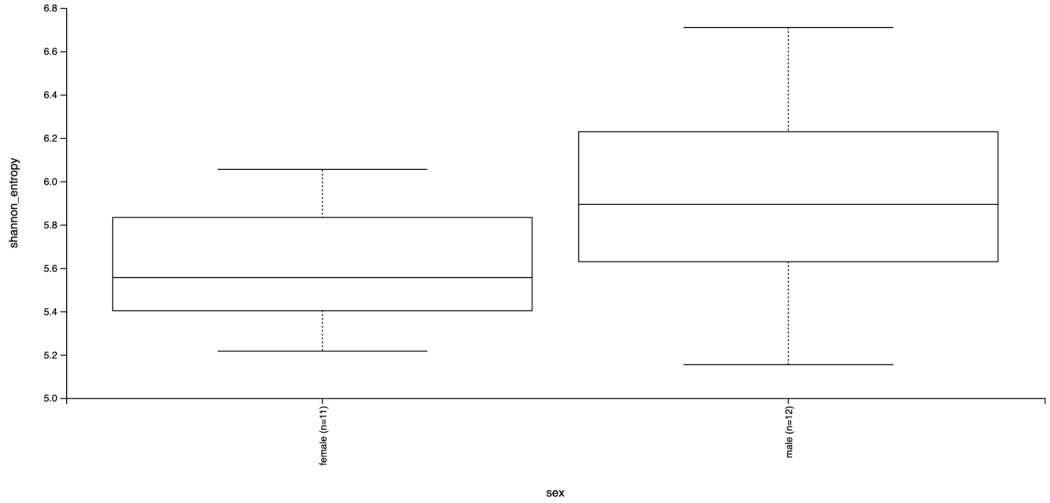
12) For alpha diversity, you need to create visualizations for Shannon diversity and Observed features. This will require you to modify the `alpha-group-significance` code. For which metadata values were graphs generated? Were any of those comparisons significant? How do you know whether they were or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

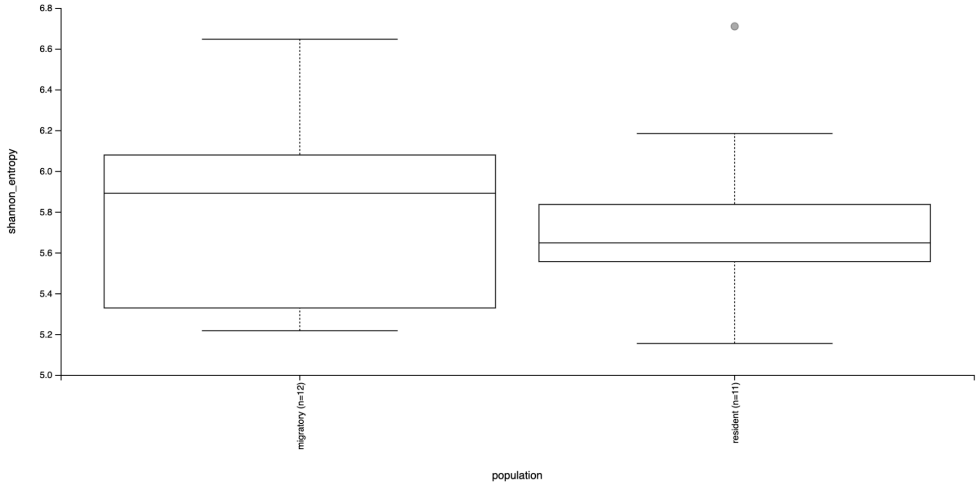
The graphs were generated for the population, sex, and flock column metadata values. For the overall groups, the comparison for sex was significant for Observed Features. I know this because the p-value of 0.02267407449864802 which is below 0.05. No other comparisons had a significant p-value. When examining the pairwise values, another significant comparison was present. For the Observed Features there was a significant difference between resident female 5 and resident male 6 (H=7.500000, p=0.006170, q=0.037019) for flock.

Shannon diversity is a measure of species richness and evenness. It is equal to the summation of the abundance of each species multiplied by the natural log of the abundance of each species. A high value is indicative of a diverse community, while a low value means the opposite.The Observed Features measurement is similar, but does not take evenness into account. According to the moving pictures tutorial, Shannon diversity is a quantitative measurement of richness while Observed Features is a qualitative measurement of richness. In this case, for both Shannon diversity and Observed Features, the diversity of the microbiome of birds in different categories is being measured.

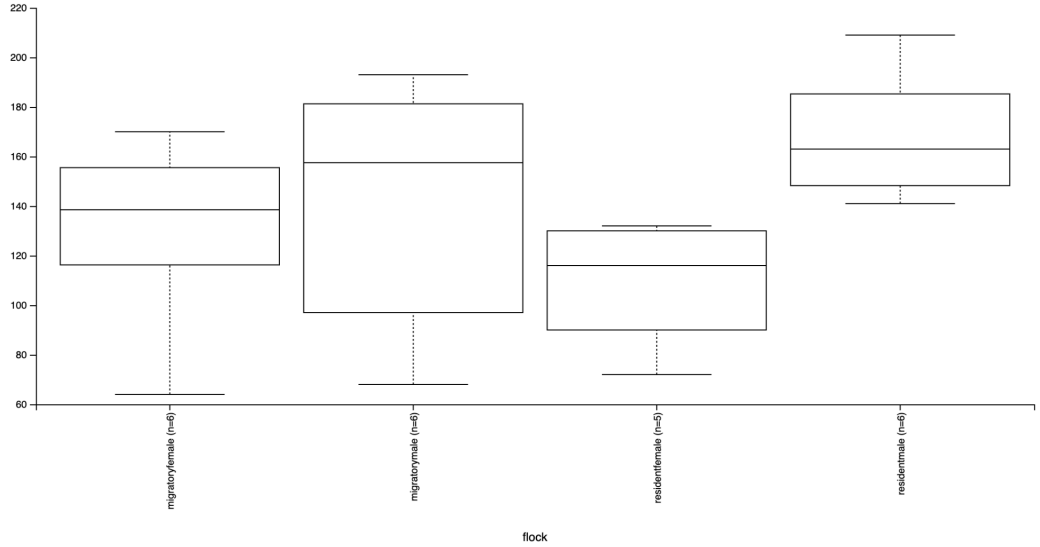
Shannon diversity:

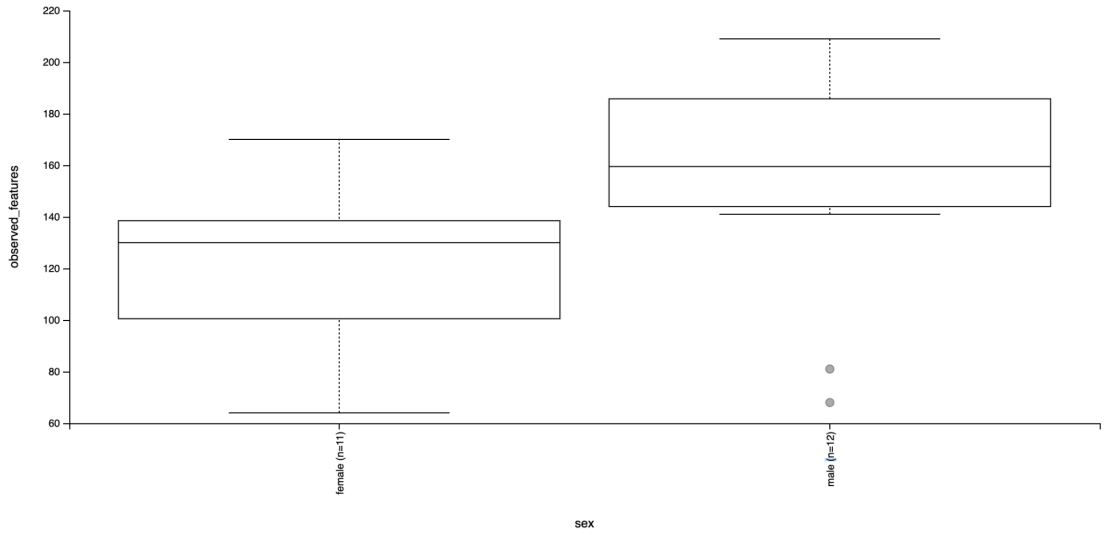


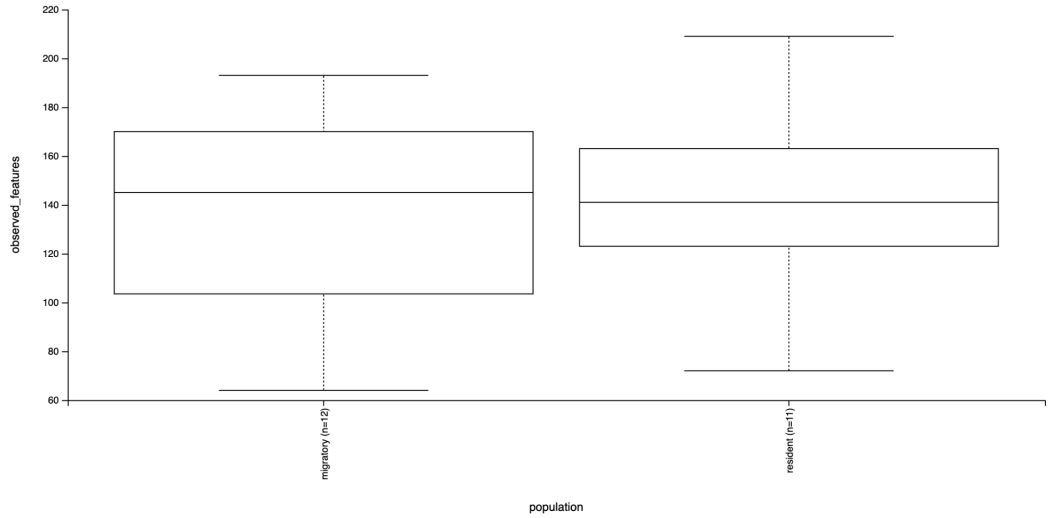




Observed Features:



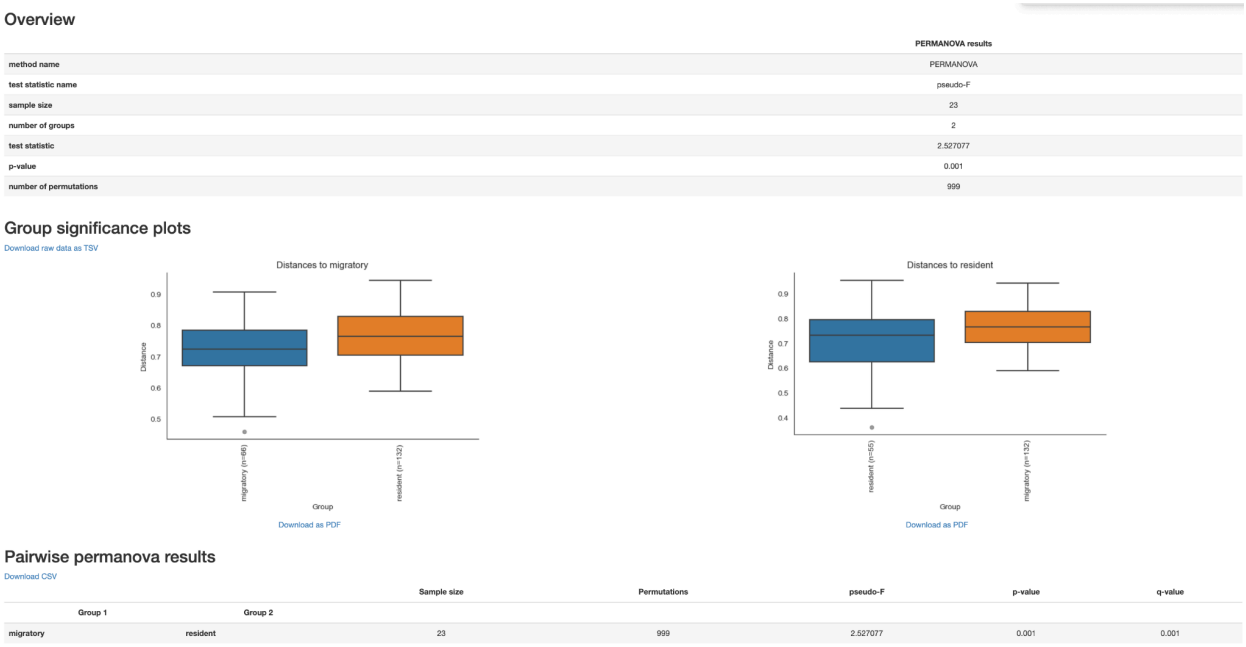




13) For beta diversity, you will need to create visualizations for Bray Curtis dissimilarity. This will require you to modify the `beta-group-significance` code. You should have one visualization for sex, one for population, and one for flock. Include a screenshot of each visualization. Is there any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

Population:

For population there is overall significance between the groups (p=0.001 which is less than 0.05). What this is telling me is that there is a notable difference between the microbiome composition of the groups between migratory vs resident birds.



Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	23
number of groups	2
test statistic	2.527077
p-value	0.001
number of permutations	999

Group significance plots

Download raw data as TSV

Download as PDF

Pairwise permanova results

Download CSV

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
migratory	resident	23	999	2.527077	0.001	0.001

Flock:

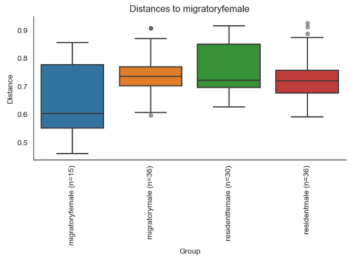
For flock there is overall significance between the groups (p=0.001 which is less than 0.05) and there is significance between each pairwise combination. As depicted in the screenshot, the p-values are all very low. More importantly, the q-values, which account for the data coming from groups, are all much lower than 0.05. What this is telling me is that there is a notable difference between the microbiome composition of the groups in terms of flock, which is a combination between sex and population (resident female, resident male, migratory female, or migratory male).

Overview

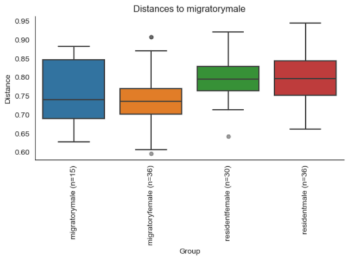
	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	23
number of groups	4
test statistic	2.687489
p-value	0.001
number of permutations	999

Group significance plots

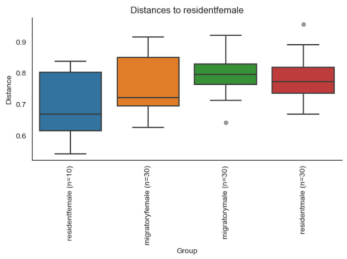
Download raw data as TSV



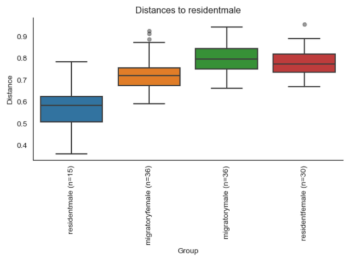
Download as PDF



Download as PDF



Download as PDF



Download as PDF

Pairwise permanova results

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	**Group 2**					
migratoryfemale	**migratorymale**	12	999	1.667307	0.028	0.028
	residentfemale	11	999	2.360236	0.002	0.004
	residentmale	12	999	3.462423	0.002	0.004
migratorymale	**residentfemale**	11	999	1.996501	0.004	0.006
	residentmale	12	999	3.350686	0.005	0.006
residentfemale	**residentmale**	11	999	3.736021	0.002	0.004

Note: There is no visualization for sex because of problems running the code

```
                There were some problems with the command:
 (1/3) Missing option '--m-metadata-file'.
 (2/3) Missing option '--o-visualization'. ("--output-dir" may also be used)
 (3/3) Got unexpected extra argument ( )
zsh: command not found: --m-metadata-file
zsh: command not found: --m-metadata-column
zsh: command not found: --o-visualization
(qiime2-amplicon-2024.2) allywatkins@Allys-MacBook-Pro qiime2-moving-pictures-tutorial %
(qiime2-amplicon-2024.2) allywatkins@Allys-MacBook-Pro qiime2-moving-pictures-tutorial % qiime diversity beta-group-significance \
  --i-distance-matrix core-metrics-results/bray_curtis_distance_matrix.qza \
  --m-metadata-file metadata2.txt \
  --m-metadata-column sex \
  --o-visualization core-metrics-results/bray_curtis_distance_matrix-sex-group-significance.qzv \
  --p-pairwise
Usage: qiime diversity beta-group-significance [OPTIONS]

  Determine whether groups of samples are significantly different from one
  another using a permutation-based statistical test.

Inputs:
  --i-distance-matrix ARTIFACT
    DistanceMatrix     Matrix of distances between pairs of samples.
                                                              [required]
Parameters:
  --m-metadata-file METADATA
  --m-metadata-column COLUMN  MetadataColumn[Categorical]
                      Categorical sample metadata column.     [required]
  --p-method TEXT Choices('permanova', 'anosim', 'permdisp')
                      The group significance test to be applied.
                                                [default: 'permanova']
  --p-pairwise / --p-no-pairwise
                      Perform pairwise tests between all pairs of groups in
                      addition to the test across all groups. This can be
                      very slow if there are a lot of groups in the metadata
                      column.                             [default: False]
  --p-permutations INTEGER
                      The number of permutations to be run when computing
                      p-values.                             [default: 999]
Outputs:
  --o-visualization VISUALIZATION
                                                              [required]
Miscellaneous:
  --output-dir PATH    Output unspecified results to a directory
  --verbose / --quiet  Display verbose output to stdout and/or stderr during
                       execution of this action. Or silence output if
                       execution is successful (silence is golden).
  --example-data PATH  Write example data and exit.
  --citations          Show citations and exit.
  --help               Show this message and exit.

                There were some problems with the command:
 (1/4) Missing option '--i-distance-matrix'.
 (2/4) Missing option '--m-metadata-file'.
 (3/4) Missing option '--o-visualization'. ("--output-dir" may also be used)
 (4/4) Got unexpected extra argument ( )
zsh: command not found: --i-distance-matrix
zsh: command not found: --o-visualization
(qiime2-amplicon-2024.2) allywatkins@Allys-MacBook-Pro qiime2-moving-pictures-tutorial % qiime diversity beta-group-significance \
  --i-distance-matrix core-metrics-results/bray_curtis_distance_matrix.qza \
  --m-metadata-file metadata2.txt \
  --m-metadata-column sex \
  --o-visualization core-metrics-results/bray_curtis_distance_matrix-se-group-significance.qzv \
  --p-pairwise
Usage: qiime diversity beta-group-significance [OPTIONS]
```

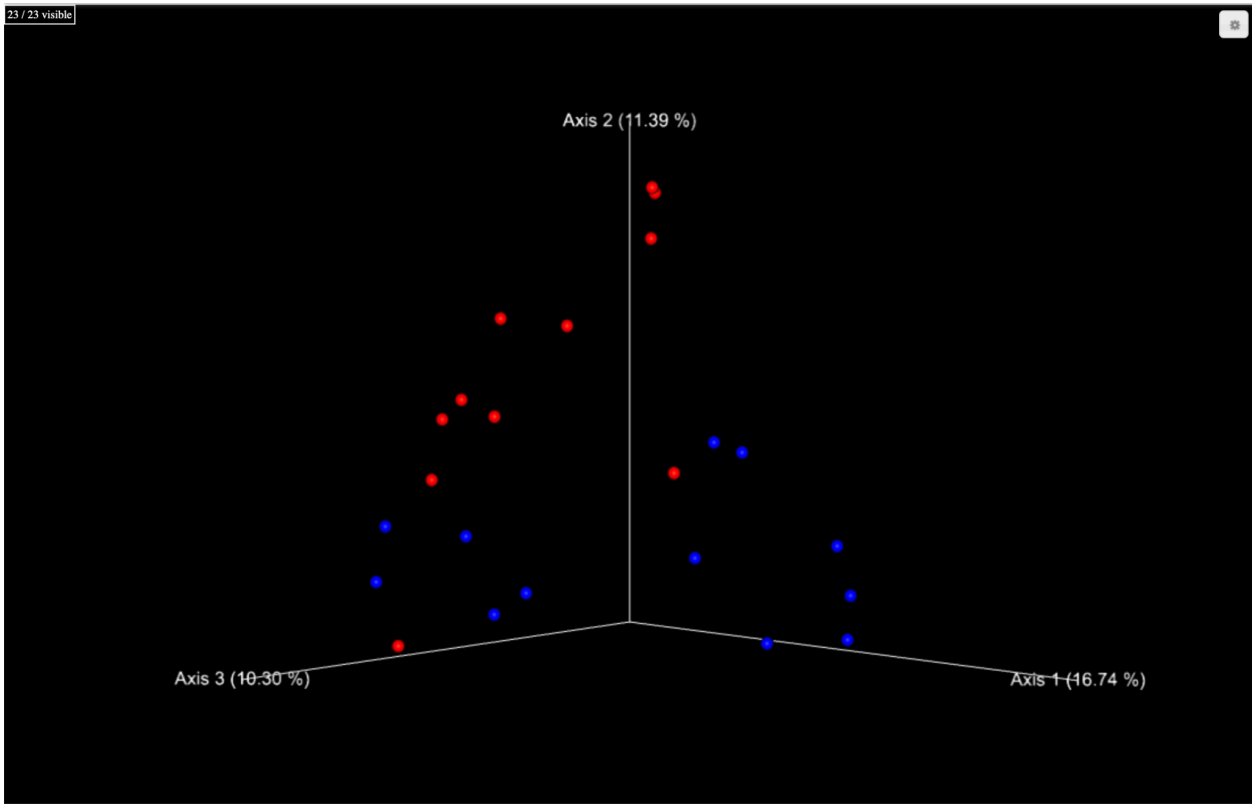
14) The `core-metrics-phylogeny` command generates a file called `bray-curtis-emporer.qzv.` Include 3 screenshots total (1 where the points are colored based on sex, one on population, one on flock). How do these results help you make sense of the results you got from question 13?

The closer together the samples are on the emperor plot, the more similar they are. If there is a lot of overlap between the samples, then they are likely not going to have a significant Bray Curtis dissimilarity. If there are distinct, spaced-out clusters of the samples, then there are probably significant differences between what is being compared.

Although I could not generate a Bray Curtis visualization in step 13 for sex, based on this emperor plot, there is likely significant community dissimilarity between the sexes. With the exception of one data point, the male and female samples are clearly in two separate groups without overlapping.

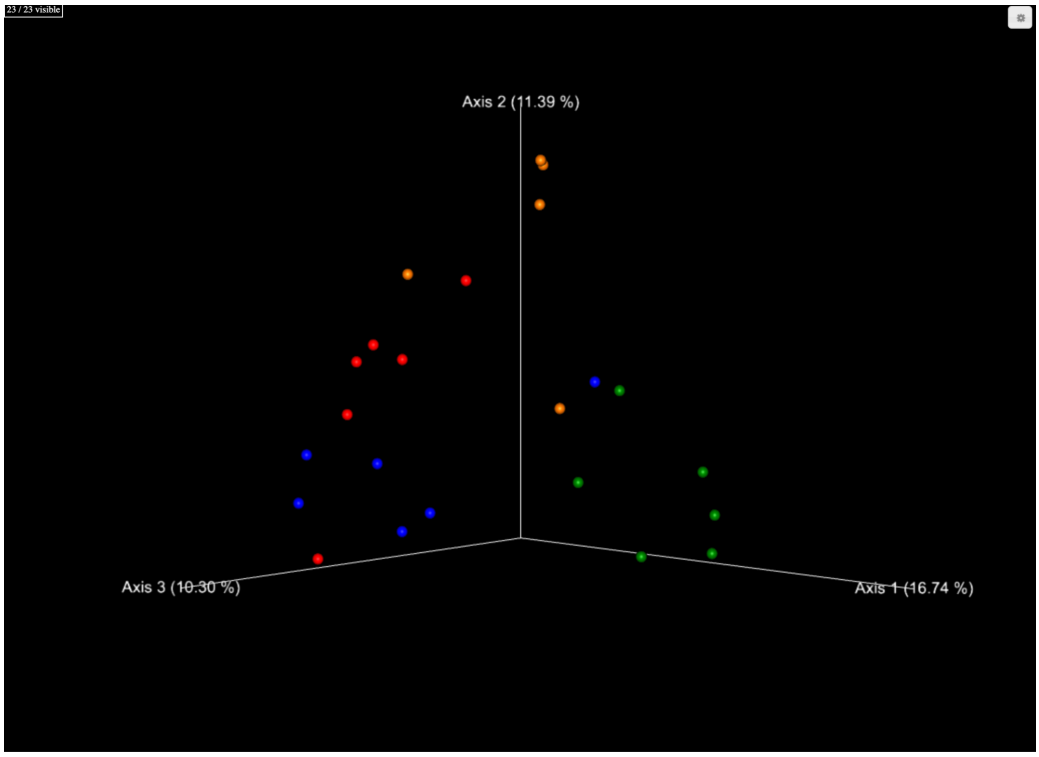
My work in step 13 indicated that there was significance between all of the groups based on flock and population. I do see this reflected in the emperor plot. With the exception of a few data points, there are well defined clumps of data based on color. Additionally, when I move the data around there is empty space between the clumps. This amount of distance between the data shows dissimilarity, confirming my conclusion in number 13.

Sex:



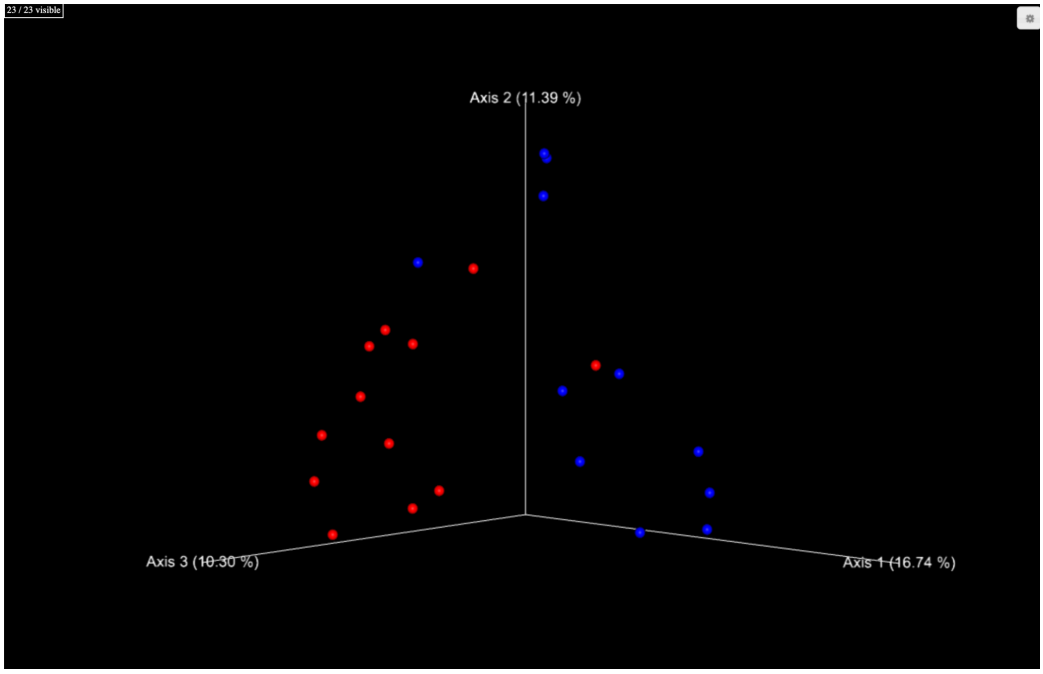
Blue is male and red is female

Flock:



Red is migratoryfemale, blue is migratorymale, orange is residentfemale, and green is residentmale

Population:



Red is migratory and blue and resident