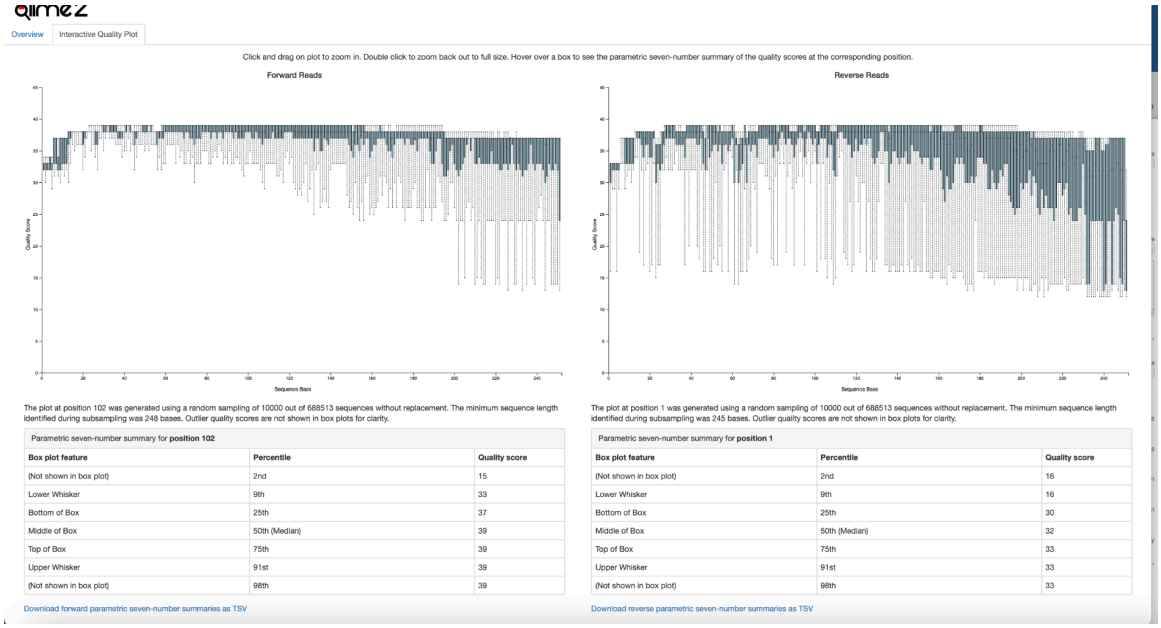
1) Include a screenshot of your interactive quality plot. Based on this plot, what values would you choose for --p-trunc-len and --p-trim-left for both the forward and reverse reads? Why have you chosen those numbers?



In terms of the values I picked, I looked at the table above, specifically the lower whisker, 9th percentile, the middle of box aka the 50th, and the bottom of box which is the 75th percentile. Then going through each data point, I made sure to exclude those that were less than 30 as an overall estimate, of over 30 being good quality according to the qimme2 guide. The values I picked were:

- trim-left-f **10**
- trunc-len-f **251**
- trim-left-r **5**
- trunc-len-r **230**

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

To modify the code to truncate and trim in your desired areas you would have to write the value or number after each line that you want to start at, and where you want to end at, depending on the forward or reverse reads. All in all, it would look like this:

```
qiime dada2 denoise-paired \
  --i-demultiplexed-seqs demux.qza \
  --p-trim-left-f 10 \
  --p-trunc-len-f 251 \
  --p-trim-left-r 5 \
  --p-trunc-len-r 230 \
  --o-representative-sequences rep-seqs.qza \
  --o-table table.qza \
  --o-denoising-stats stats.qza
```

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?

Instead of manually re-naming the files in the folder, and doing them one by one in the terminal as done in the tutorial one could change the code to something like this:

qiime metadata tabulate \
–m-input-file stats.qza \
–o-visulzation stats.qza

qiime feature-table summarize \
–i-table table.qza \
–o-visulization table.qza
–m-sample-metadata-file metadataWR.txt

qiime feature-table tabulate-seqs \
–i-data rep-seqs.qza
–o-visulatizion rep-seqs.qza

Essentially, removing the ‘dada2’ part from both lines of code.

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?

Since the metadata file has a different name we would just have to change the name in the code, rather than using exactly what it has in the tutorial:

qiime feature-table summarize \

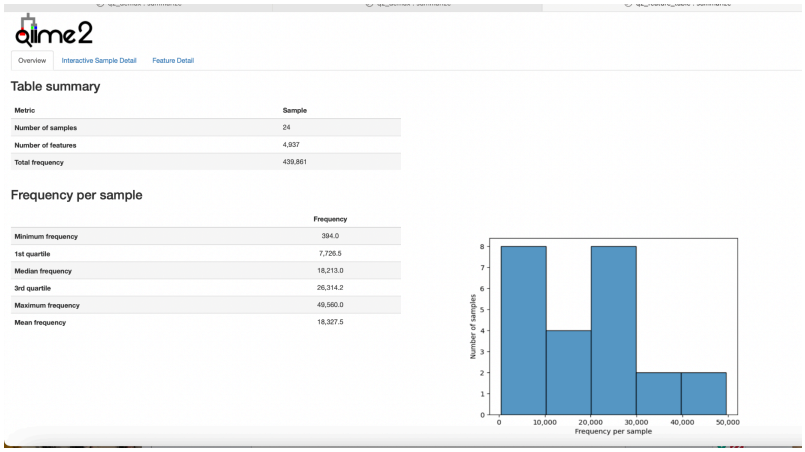
--i-table table.qza \

--o-visualization table.qzv \

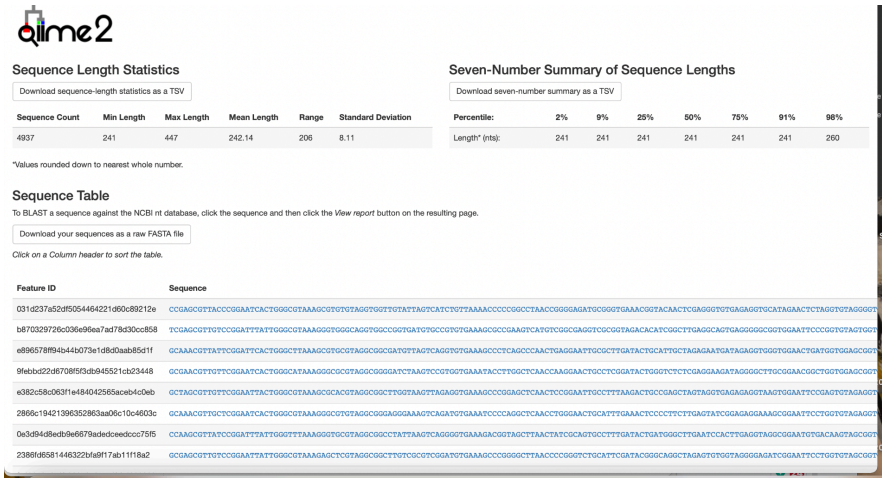
--m-sample-metadata-file metadataWR.txt

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.

Table Summary:

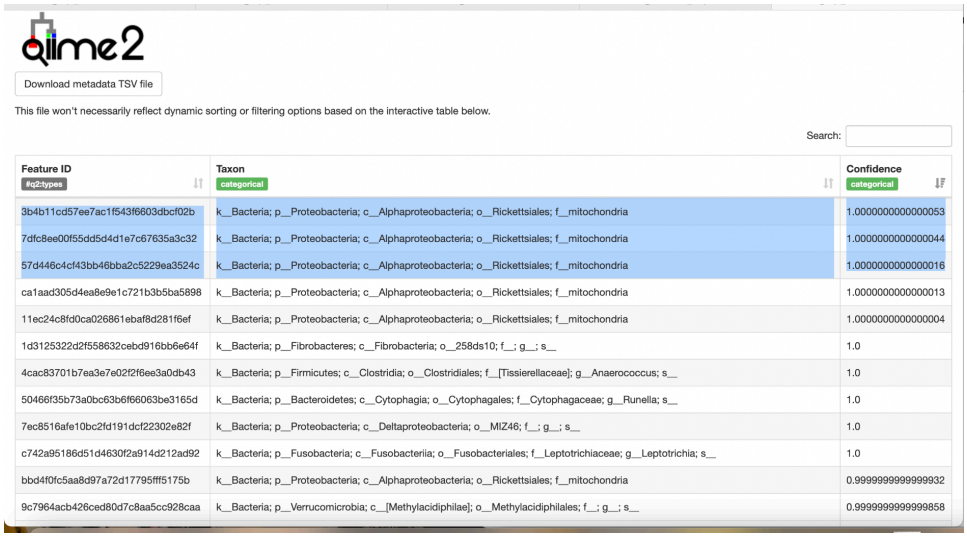


Statistics From rep-seqs file:



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

Screenshot of Top Hits



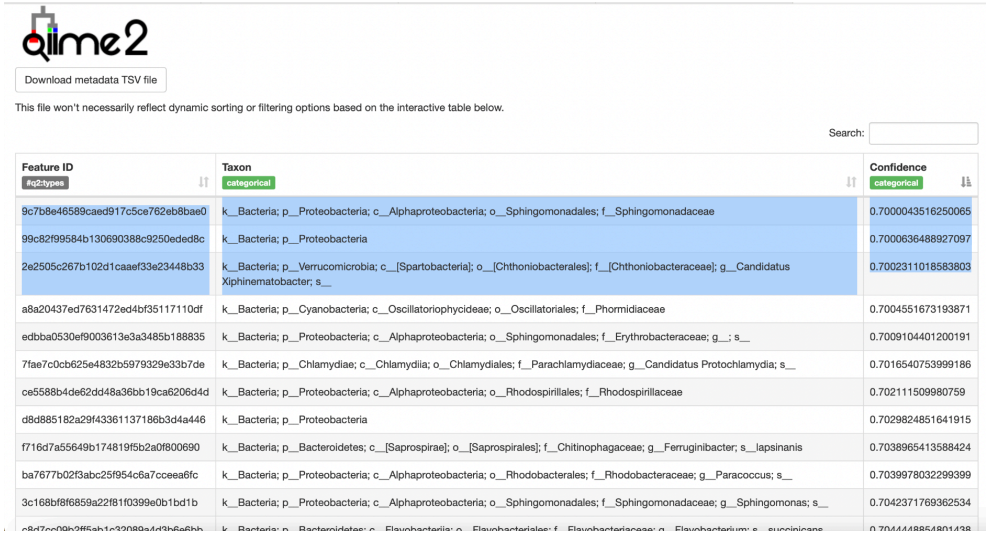
Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

Feature ID #q2:types	Taxon categorical	Confidence categorical
3b4b11cd57ee7ac1f543f6603dbcf02b	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000053
7dfc8ee00f55dd5d4d1e7c67635a3c32	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000044
57d446c4cf43bb46bba2c5229ea3524c	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000016
ca1aad305d4ea8e9e1c721b3b5ba5898	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000013
11ec24c8fd0ca026861ebaf8d281f6ef	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	1.0000000000000004
1d3125322d2f558632cebd916bb6e64f	k__Bacteria; p__Fibrobacteres; c__Fibrobacteria; o__258ds10; f__; g__; s__	1.0
4cac83701b7ea3e7e02f2f6ee3a0db43	k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__[Tissierellaceae]; g__Anaerococcus; s__	1.0
50466f35b73a0bc63b6f66063be3165d	k__Bacteria; p__Bacteroidetes; c__Cytophagia; o__Cytophagales; f__Cytophagaceae; g__Runella; s__	1.0
7ec8516afe10bc2fd191dcf22302e82f	k__Bacteria; p__Proteobacteria; c__Deltaproteobacteria; o__MIZ46; f__; g__; s__	1.0
c742a95186d51d4630f2a914d212ad92	k__Bacteria; p__Fusobacteria; c__Fusobacteriia; o__Fusobacteriales; f__Leptotrichiaceae; g__Leptotrichia; s__	1.0
bbd4f0fc5aa8d97a72d17795fff5175b	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rickettsiales; f__mitochondria	0.9999999999999932
9c7964acb426ced80d7c8aa5cc928caa	k__Bacteria; p__Verrucomicrobia; c__[Methylacidiphilae]; o__Methylacidiphilales; f__; g__; s__	0.9999999999999858

Screenshot of the reverse (bottom hits)



Download metadata TSV file

This file won't necessarily reflect dynamic sorting or filtering options based on the interactive table below.

Search:

Feature ID #q2:types	Taxon categorical	Confidence categorical
9c7b8e46589caed917c5ce762eb8bae0	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae	0.7000043516250065
99c82f99584b130690388c9250eded8c	k__Bacteria; p__Proteobacteria	0.7000636488927097
2e2505c267b102d1caaef33e23448b33	k__Bacteria; p__Verrucomicrobia; c__[Spartobacteria]; o__[Chthoniobacterales]; f__[Chthoniobacteraceae]; g__Candidatus Xiphinematobacter; s__	0.7002311018583803
a8a20437ed7631472ed4bf35117110df	k__Bacteria; p__Cyanobacteria; c__Oscillatoriophycideae; o__Oscillatoriales; f__Phormidiaceae	0.7004551673193871
edbba0530ef9003613e3a3485b188835	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Erythrobacteraceae; g__; s__	0.7009104401200191
7fae7c0cb625e4832b5979329e33b7de	k__Bacteria; p__Chlamydiae; c__Chlamydiia; o__Chlamydiales; f__Parachlamydiaceae; g__Candidatus Protochlamydia; s__	0.7016540753999186
ce5588b4de62dd48a36bb19ca6206d4d	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhodospirillales; f__Rhodospirillaceae	0.702111509980759
d8d885182a29f4336137186b3d4a446	k__Bacteria; p__Proteobacteria	0.7029824851641915
f716d7a55649b174819f5b2a0f800690	k__Bacteria; p__Bacteroidetes; c__[Saprospirae]; o__[Saprospirales]; f__Chitinophagaceae; g__Ferruginibacter; s__lapsinanis	0.7038965413588424
ba7677b02f3abc25f954c6a7cceea6fc	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Rhodobacterales; f__Rhodobacteraceae; g__Paracoccus; s__	0.7039978032299399
3c168bf8f6859a22f81f0399e0b1bd1b	k__Bacteria; p__Proteobacteria; c__Alphaproteobacteria; o__Sphingomonadales; f__Sphingomonadaceae; g__Sphingomonas; s__	0.7042371769362534

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

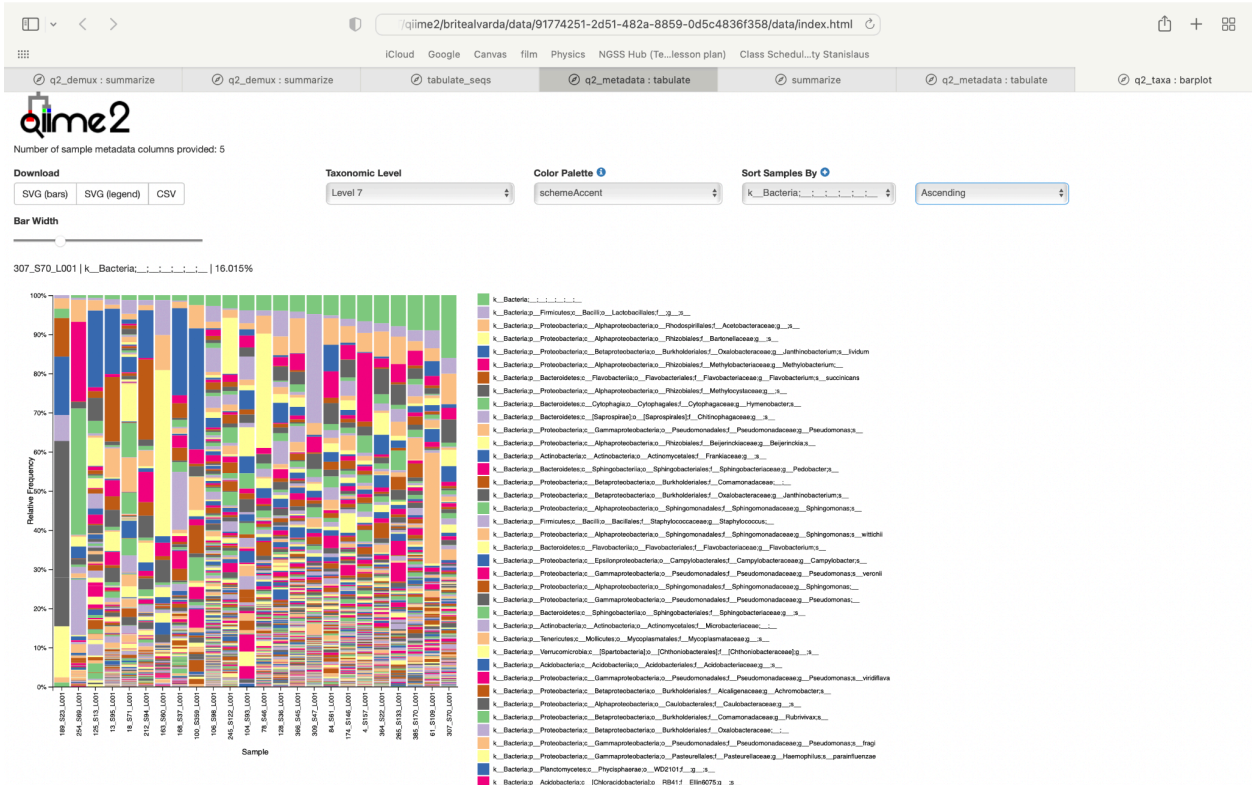
**qiime taxa filter-table **
These are what we are importing to be filtered
**--i-table table.qza **
**--i-taxonomy taxonomy.qza **
#this is what we wanted to be excluded from the feature table since bacteria came from mt/chloroplasts
**--p-exclude mitochondria,chloroplast **
#this says where the output file will be saved
--o-filtered-table table.qza

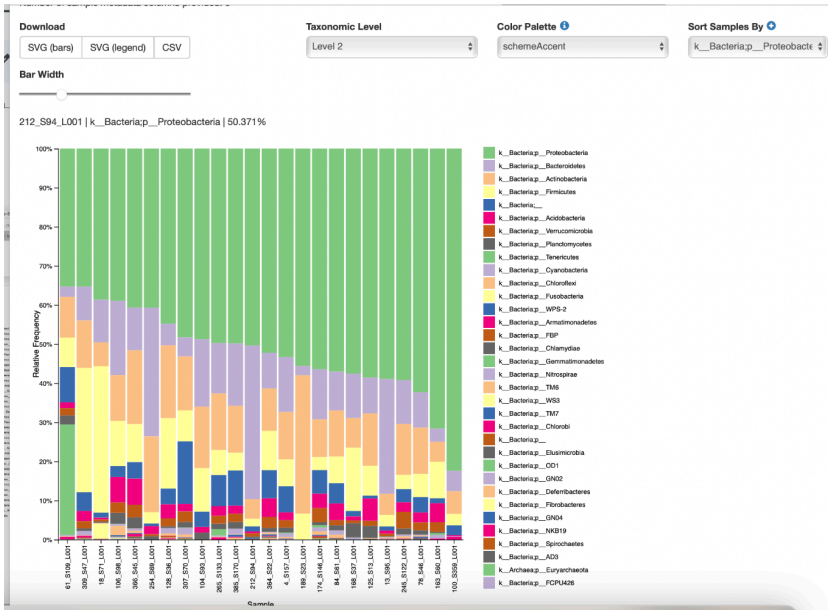
This is an important code because we want to limit any discrepancies in the data, for example, making sure to take out any mitochondrial or chloroplast data is important because it can skew the data. Specifically, since we are working with taxonomy it may remove any of the non-prokaryotic or bacterial strains from the list, or make the list more accurate.

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.

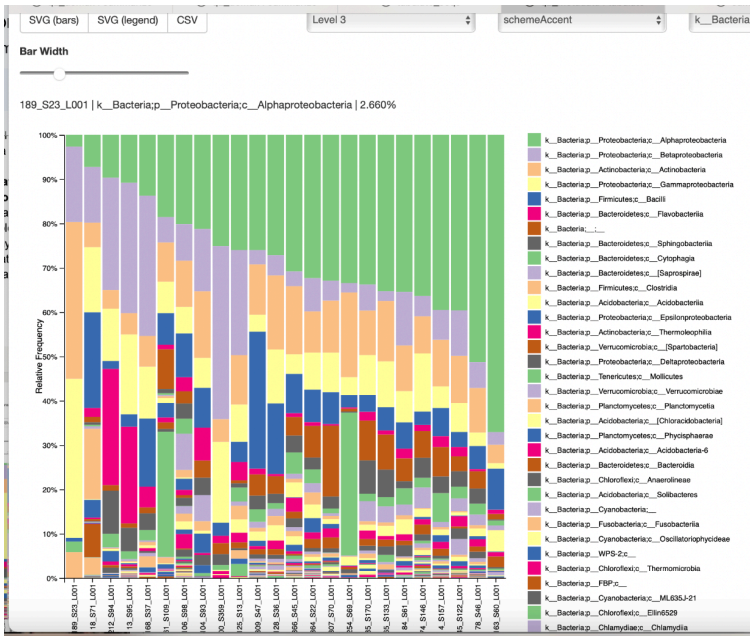
Originally I was thinking that there would be a change in the data, since we are re-generating the code, however re-looking at the code it makes sense that we technically aren't using the same file that we are filtering. After redoing both codes there is no change in the data. This is because the code is actually changing a different file that we are not re-loading to view.

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 two phyla were: Phyla Proteobacteria and Phyla Bacteroidetes



My top 5 classes were: Alphaproteobacteria, Betaproteobacteria, Actinobacteria, Gammaproteobacteria, and Bacilli

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 used to measure the amount of species within the community and their distribution within themselves. Beta diversity is similar but looks more at the different species samples and compares those within the community or data pool.

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 file that helped me was the table.qzv file to visualize the table data. Based on this data I had to find a value that would not lose too much of the sample or exclude a chunk of the samples since this would skew my results. So by looking through the data I landed on the value 8000. At first, I thought that a higher number would be better, however, when using something high like 25000, this would end up excluding so many data points altogether. According to the table data, there is a value of 128,000 (31.33%) features in 16 (66.67%) retained while losing 8 of the sample IDs.

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).

Starting with sex the q-values for Shannon and observed were greater than 0.05, meaning they were not significant. In terms of population, this compared the migratory/resident these values also had high q-values making them not significant. Lastly, in terms of flock, this still showed no significant diversity between the migratory and resident of the different groups based on the q-value. Shannon diversity is used to measure the abundance of the species and the evenness, while observed features is measuring the features given within the sample set, especially since this data set has a frequency greater than 0.

13) 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?)

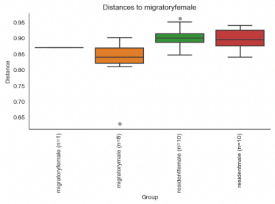
The beta diversity gives insight into how much the birds are different from the overall community. Rather than comparing how diverse each bird is individually like in alpha diversity, in beta we compare using the distance and dissimilatory data using the q-values and plots.

FLOCK: There is a significant difference between resident females and resident males, due to the very low q-value, showing their diversity. There is also a significant difference between the migratory male data compared to the resident data sets, implying again that there is high diversity. Lastly, comparing the migratory females with the migratory males, and the resident data sets, the q-value shows that there is overall no significant diversity between them.

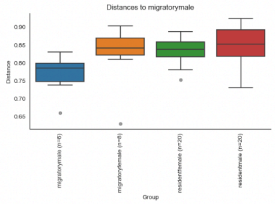
	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	4
test statistic	1.510462
p-value	0.002
number of permutations	999

Group significance plots

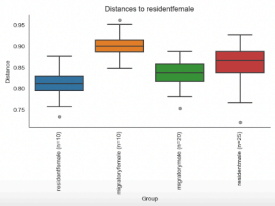
Download raw data as TSV

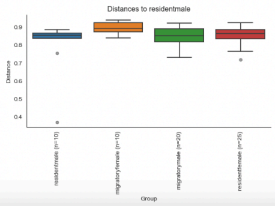


Download as PDF



Download as PDF





POPULATION: From this data set this is significant as the q-value is lower than 0.05, which implies that they have higher than normal diversity in the microbiome.

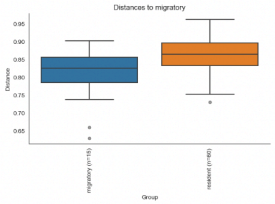
qiime2

Overview

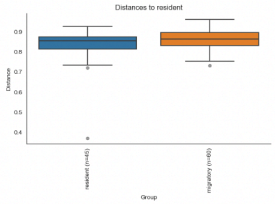
	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	2
test statistic	1.721353
p-value	0.002
number of permutations	999

Group significance plots

Download raw data as TSV



Download as PDF



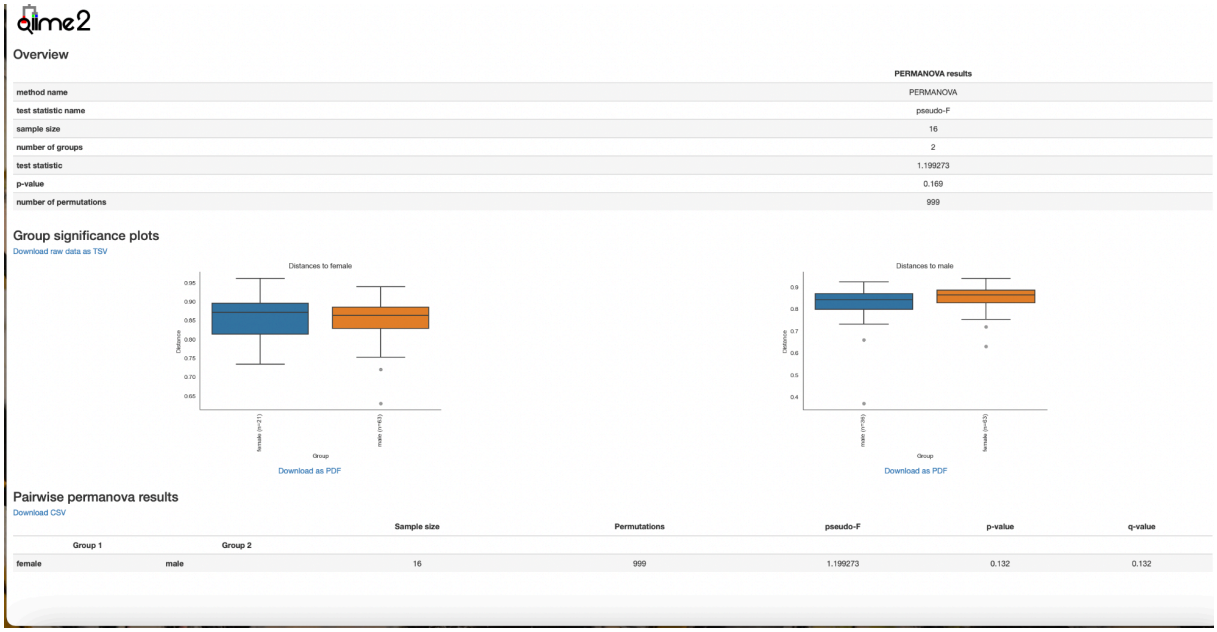
Download as PDF

Pairwise permanova results

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
migratory	resident	16	999	1.721353	0.001	0.001

SEX: This q-value shows that it is not significant, which implies that the data set between females and males is overall similar.



Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	16
number of groups	2
test statistic	1.199273
p-value	0.169
number of permutations	999

Group significance plots

Download raw data as TSV

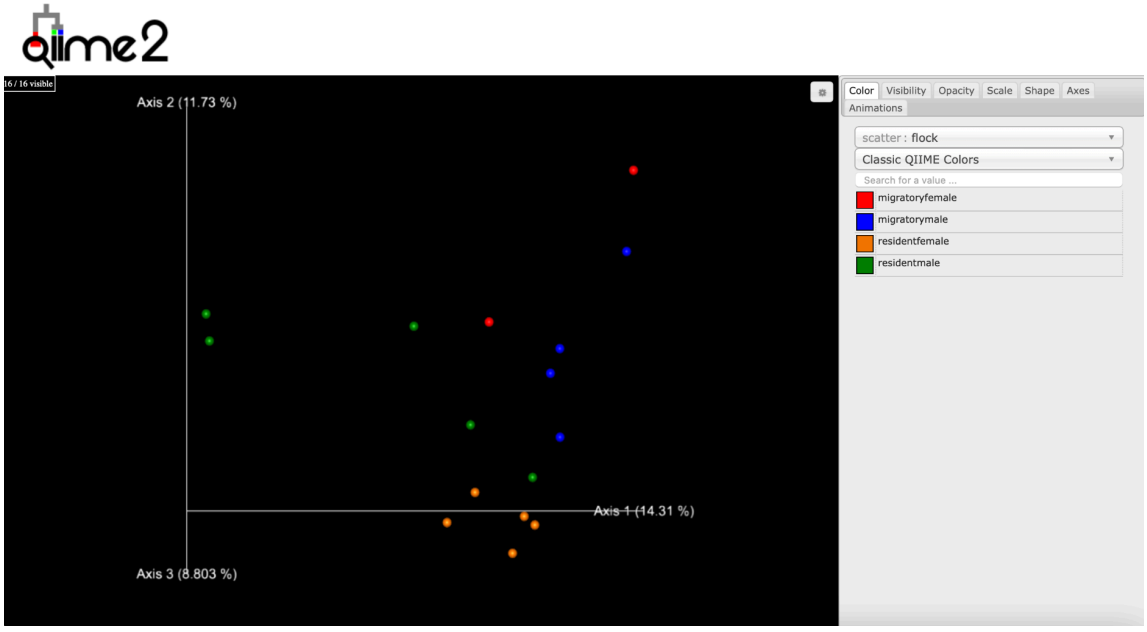
Pairwise permanova results

Download CSV

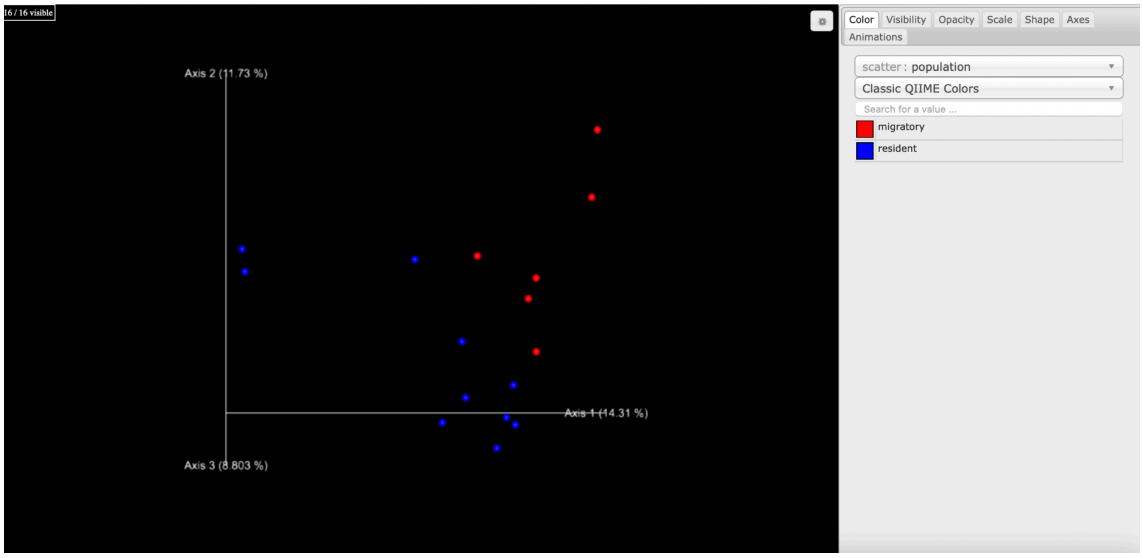
Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
female	male	16	999	1.199273	0.132	0.132

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?

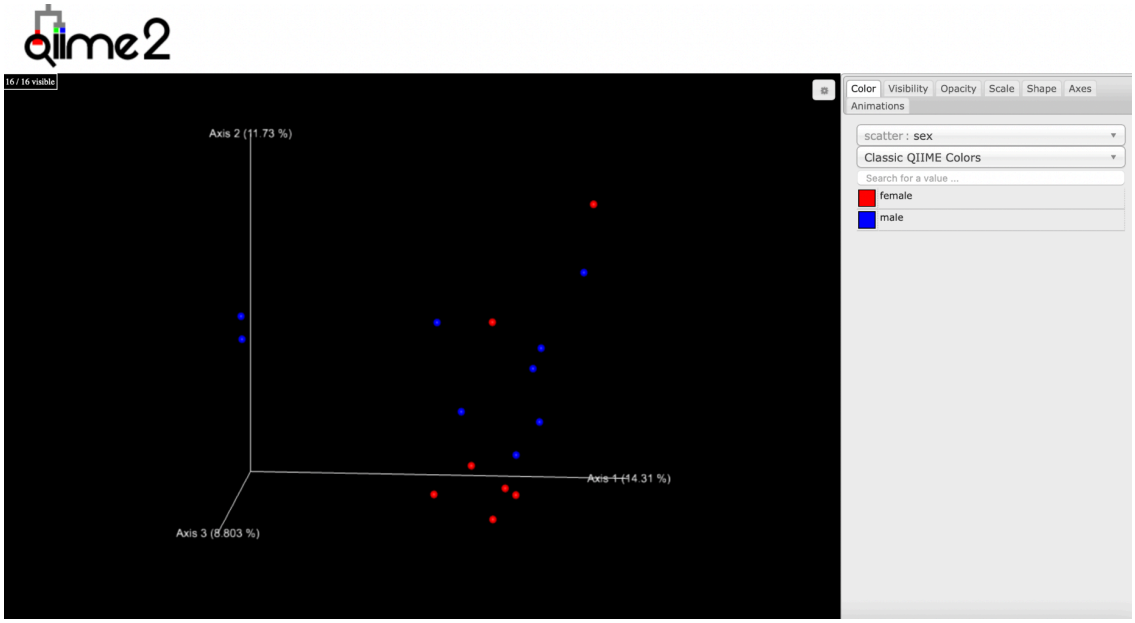
FLOCK:



POPULATION:



SEX:



These graphs based on the data help me make sense of the results from question 13 because the emperor makes the data into a graph that shows the distance between the same sample categories needed in the previous question.