Emma Hoover

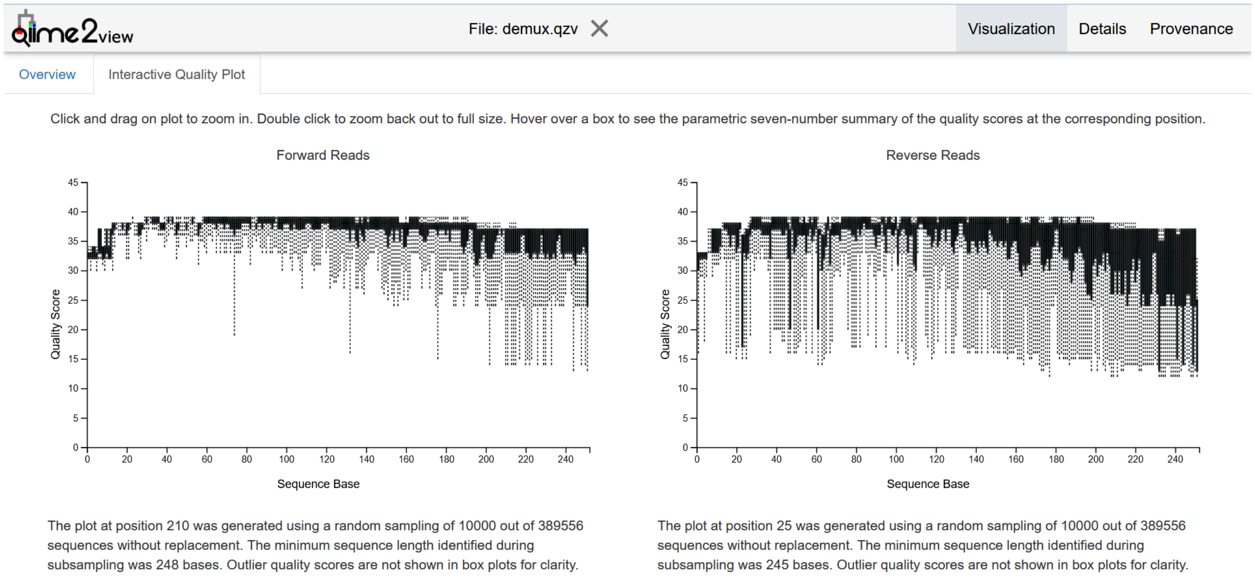
Dr. Van Laar

BIOL 4810

3 May 2024

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?**



Click and drag on plot to zoom in. Double click to zoom back out to full size. Hover over a box to see the parametric seven-number summary of the quality scores at the corresponding position.

The plot at position 210 was generated using a random sampling of 10000 out of 389556 sequences without replacement. The minimum sequence length identified during subsampling was 248 bases. Outlier quality scores are not shown in box plots for clarity.

The plot at position 25 was generated using a random sampling of 10000 out of 389556 sequences without replacement. The minimum sequence length identified during subsampling was 245 bases. Outlier quality scores are not shown in box plots for clarity.

Based on this plot, for the forward reads, I would choose the --p-trunc-len value to be 210 and the --p-trim-left value to be 25. For the reverse reads, I would choose the --p-trunc-len value to be 220 and the --p-trim-left value to be 25. I have chosen these --p-trim-left numbers because that seems to be where the reads begin to be their highest quality. As for the --p-trunc-len values, I chose those numbers because they seem to be where the reads begin to drop in quality. However, I did choose a slightly higher value than where they immediately drop off because I did not want to accidentally exclude too many reads.

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

I would modify the code above to truncate and trim in my desired way by putting:

```
--p-trim-left-f 25 \
--p-trunc-len-f 210 \
--p-trim-left-r 25 \
--p-trunc-len-r 220 \
```

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?**

I could modify the above code to skip that step by changing the original code from:

```
--o-representative-sequences rep-seqs-dada2.qza \
--o-table table-dada2.qza \
--o-denoising-stats stats-dada2.qza \
```

to:

```
--o-representative-sequences rep-seqs.qza \
--o-table table.qza \
--o-denoising-stats stats.qza \
```

By deleting the “-dada2” part, the output files will not include that in their name, so I will not have to rename the files. In addition, I would also need to modify the qiime metadata tabulate by changing the original code from:

```
--m-input-file stats-dada2.qza \
--o-visualization stats-dada2.qzv
```

to:

```
--m-input-file stats.qza \
--o-visualization stats.qzv
```

This way, the input file for the qiime metadata tabulate code has the same name as the output file I generated from the quality control 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?**

I would adjust my code in order to use the metadata file I have been given from:

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

to:

```
--m-sample-metadata-file 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.**

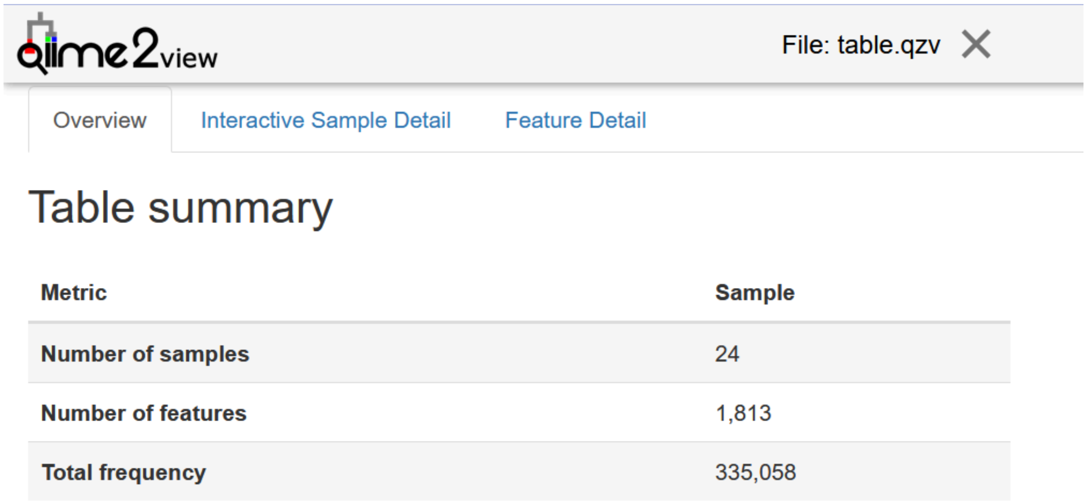
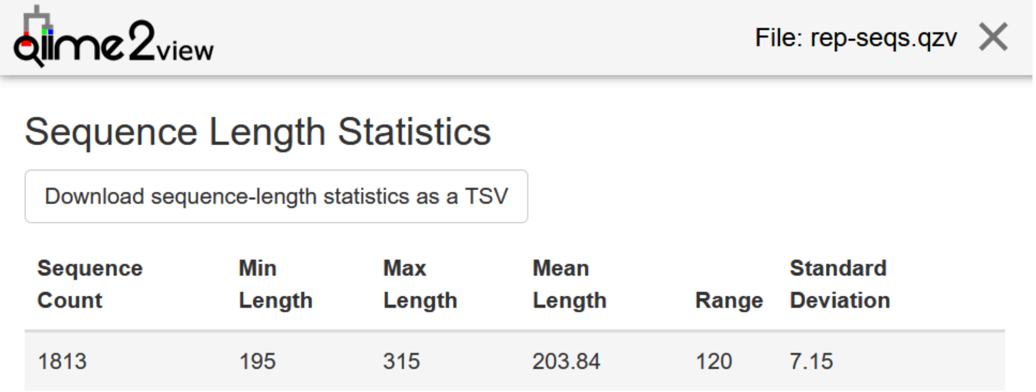


Table summary

Metric	Sample
Number of samples	24
Number of features	1,813
Total frequency	335,058



Sequence Length Statistics

Sequence Count	Min Length	Max Length	Mean Length	Range	Standard Deviation
1813	195	315	203.84	120	7.15

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

The top three hits are Rickettsiales mitochondria, the fourth hit belongs to the class Armatimonadia, and the fifth hit is from the genus Leptotrichia.

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

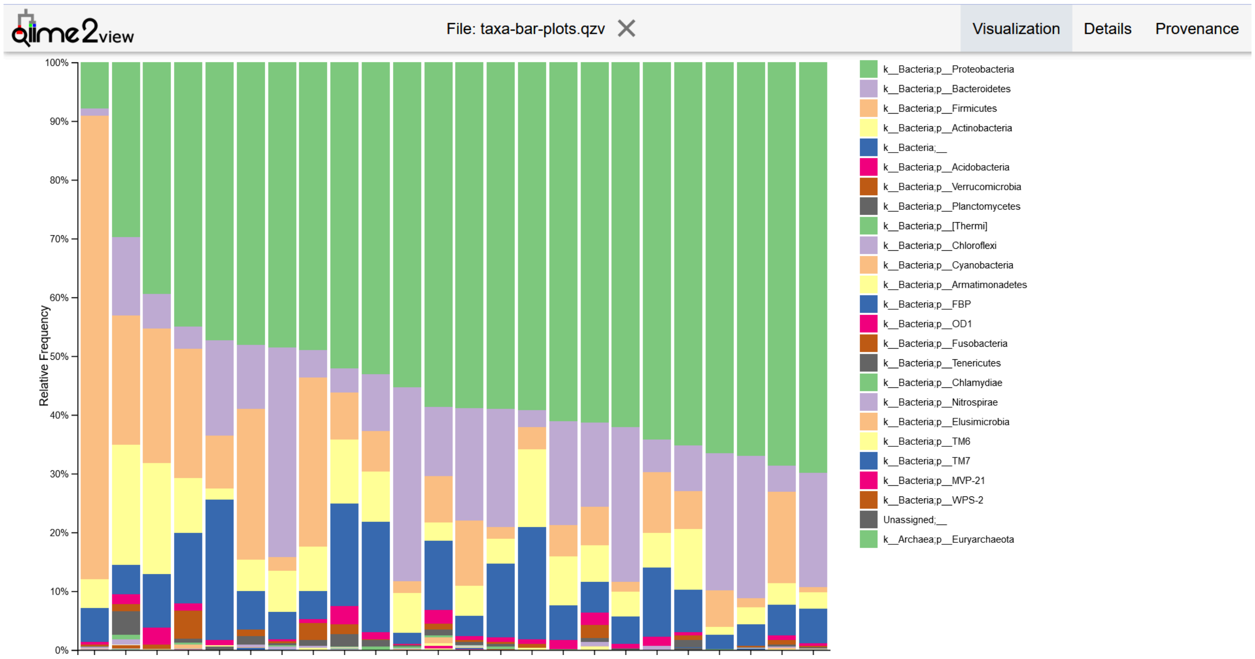
I think this code is filtering out the mitochondria and chloroplast reads from your genome in the table and taxonomy files I generated in previous steps. I think this is a necessary or important step because the mitochondria and chloroplasts have their own DNA.

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.**

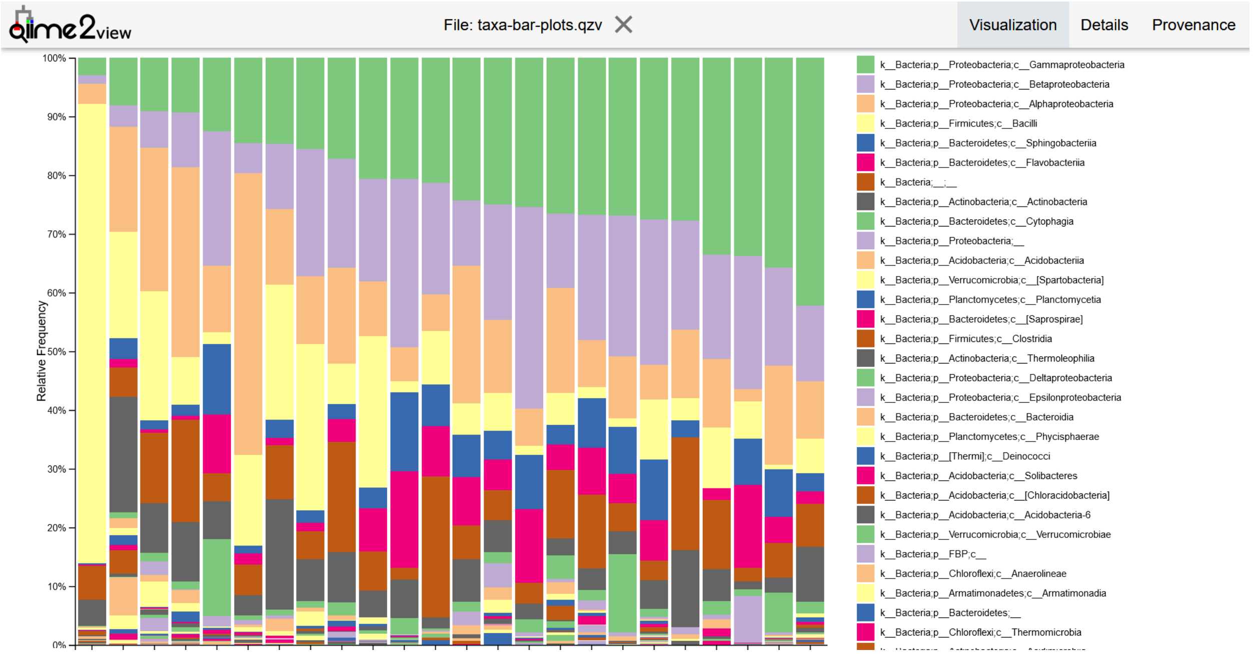
No, there are no differences in the hits with the highest confidence. Nothing has changed because even though I redid the table visualization and taxonomy commands, the input for the code is the table and taxonomy, not the rep-seqs.qza file. Since the table and taxonomy were made from the rep-seqs.qza file, nothing will change unless you filter out the mitochondrial and chloroplast sequences from 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.

Looking at taxa bar plots, my top 2 phyla are Proteobacteria and Bacteroidetes.



The top 5 most abundant classes are Gammaproteobacteria, Betaproteobacteria, Alphaproteobacteria, Bacilli, and Sphingobacteriia.



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 the measure of diversity within a community based on species richness, or the number of species, in that community. Beta diversity is a measure of diversity by comparing the composition of different communities, usually comparing how different the communities look from one another by measuring how the amount of species change between the communities.

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 I previously generated that will help me determine what to choose for the sampling depth is the table.qzv file. The sampling depth value I am going to use is 2,713 because the next sampling depths listed after that are 1,737, 798, and 765. There is a large drop in frequency from 2,713 and 1,737 which is why I chose 2,713 as my sampling depth and excluded anything below that. Using this sampling depth, I retain 21 samples and lose 3 samples.

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 metadata values that graphs were generated for were population, sex, and flock.

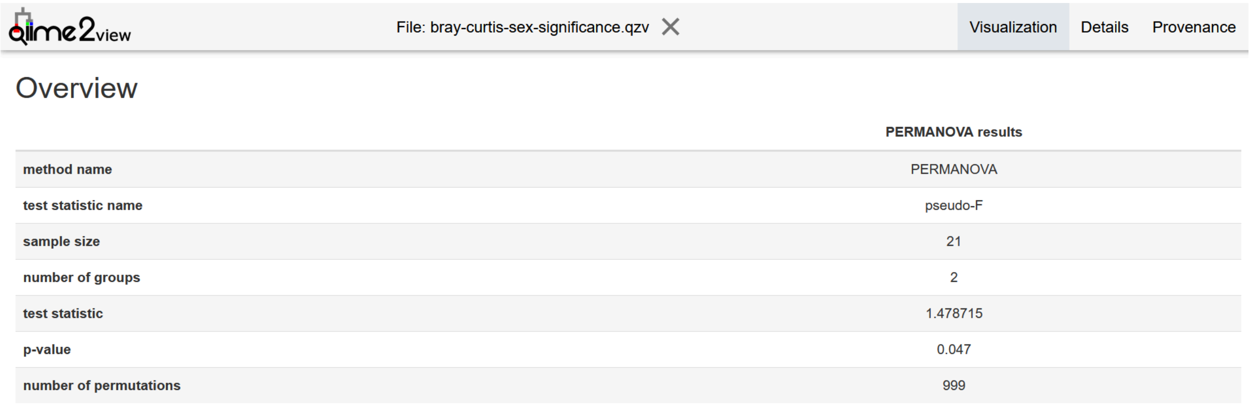
None

of those comparisons were significant. I know the comparisons were not significant because they did not have a q-value (adjusts the p-value for multiple comparisons) less than 0.05.

The Observed features test measures the richness of the sample (the amount of different features present in the sample). The Shannon diversity test measures the richness of the sample, but also takes into account the evenness of the features. This means it measures the diversity of the features by incorporating the difference in amounts of the features in the sample.

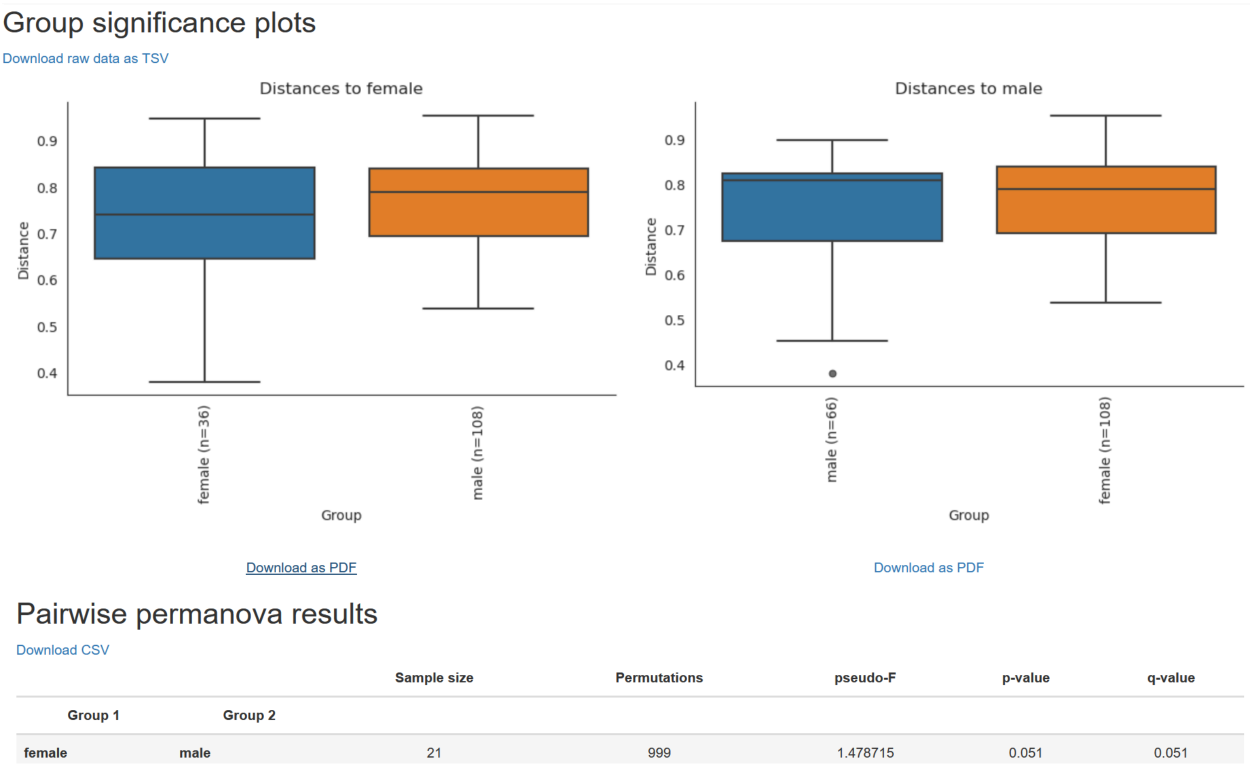
13) **For beta diversity, you will need to create visualizations for Bray Curtis dissimilarity. This will require you to modify the beta-group-significance code. You should have one visualization for sex, one for population, and one for flock. Include a screenshot of each visualization. Is there any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

Sex:



Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	21
number of groups	2
test statistic	1.478715
p-value	0.047
number of permutations	999



Group significance plots

Download raw data as TSV

Download as PDF

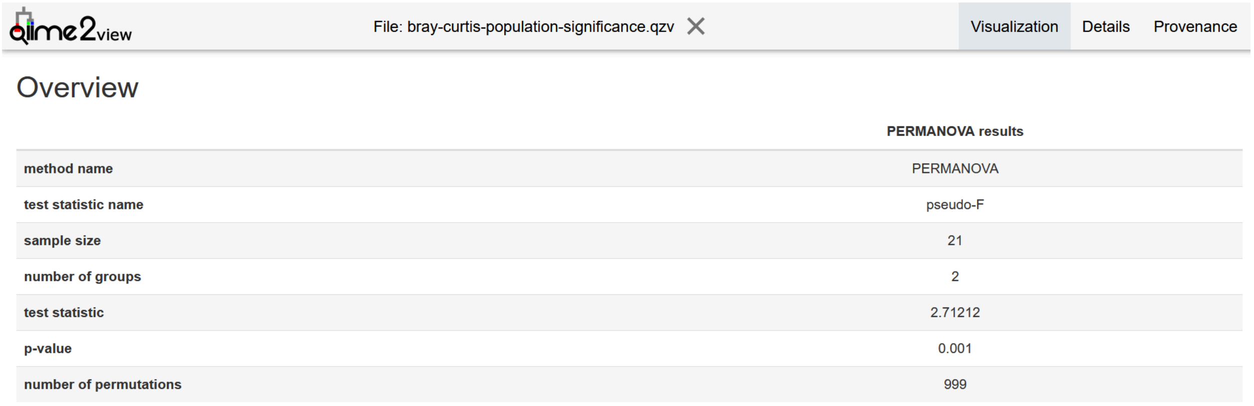
Download as PDF

Pairwise permanova results

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
female	male	21	999	1.478715	0.051	0.051

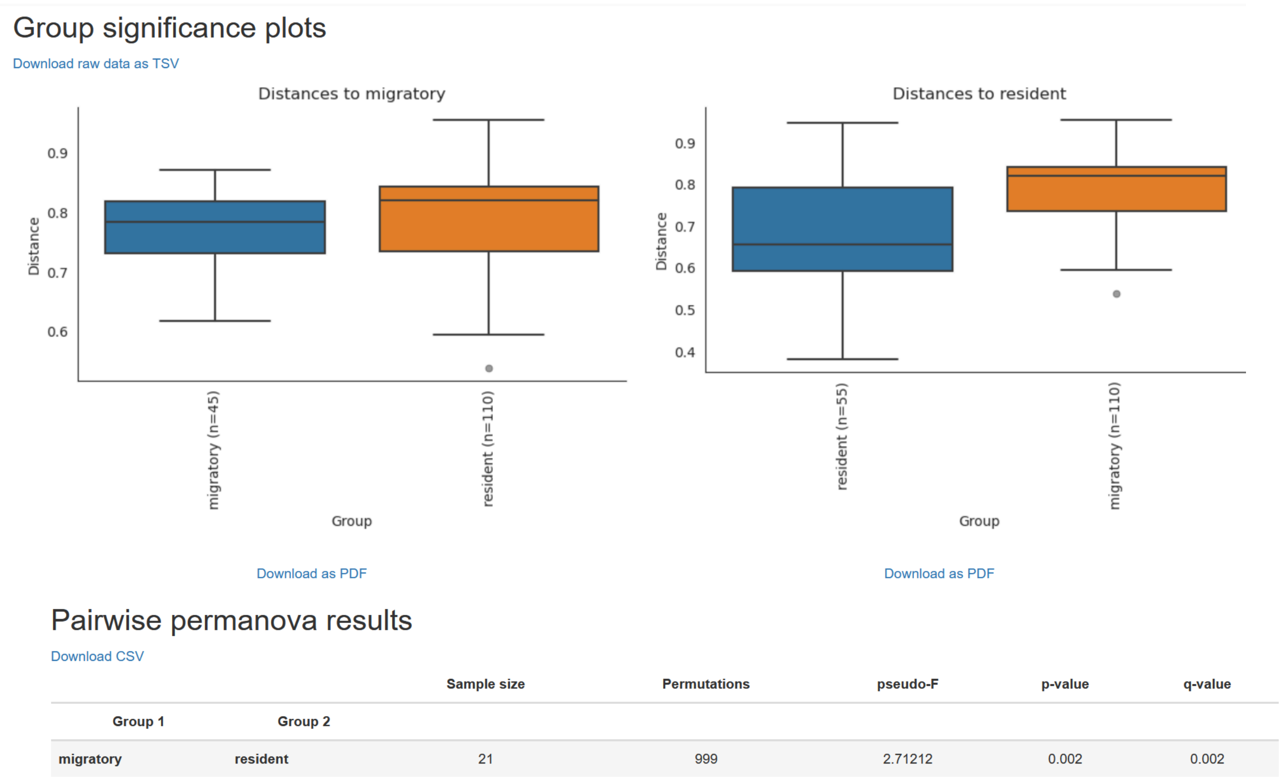
Population:



File: bray-curtis-population-significance.qzv

Overview

	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	21
number of groups	2
test statistic	2.71212
p-value	0.001
number of permutations	999



Group significance plots

Download raw data as TSV

Download as PDF

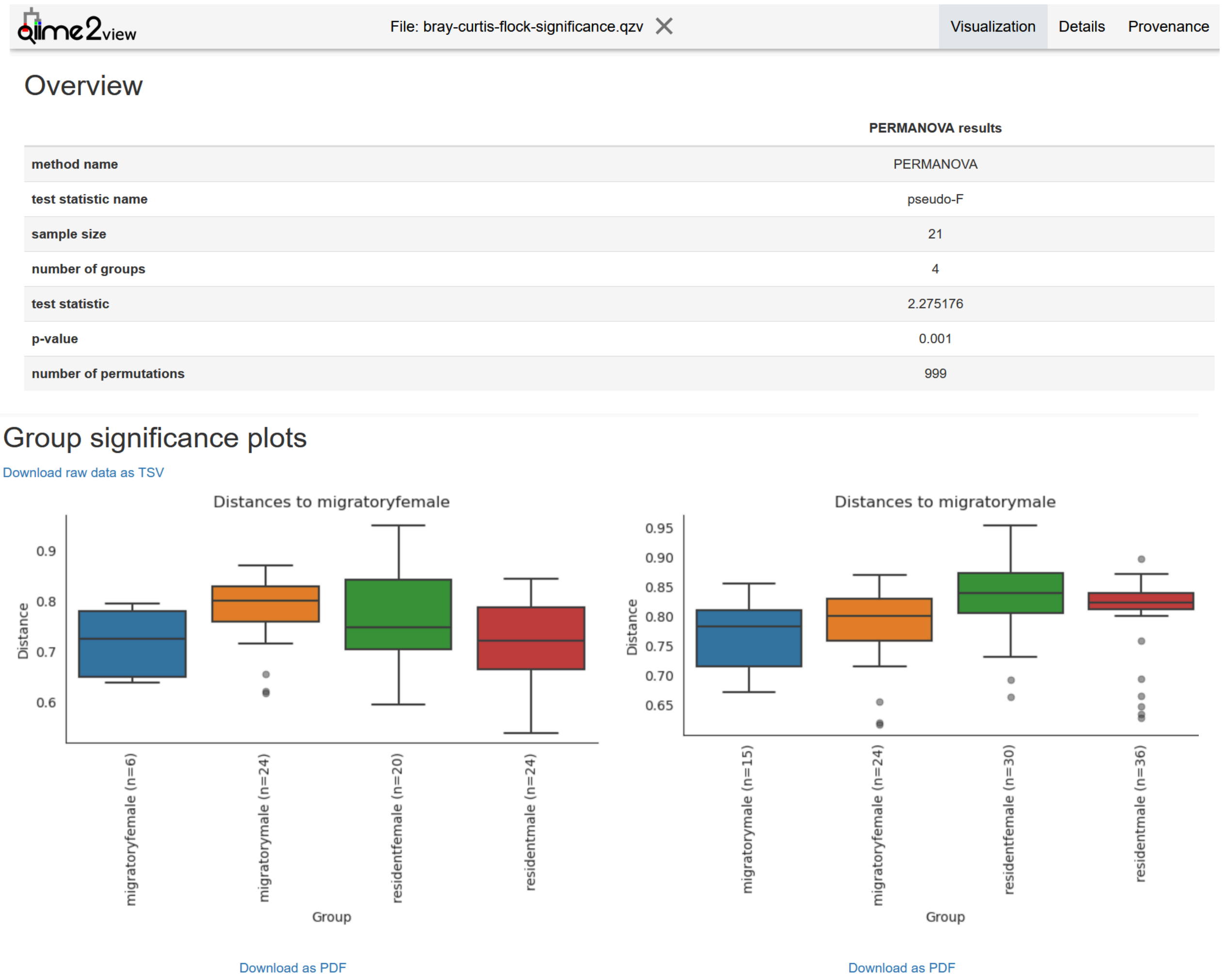
Download as PDF

Pairwise permanova results

Download CSV

		Sample size	Permutations	pseudo-F	p-value	q-value
Group 1	Group 2					
migratory	resident	21	999	2.71212	0.002	0.002

Flock:



File: bray-curtis-flock-significance.qzv ✕

Overview

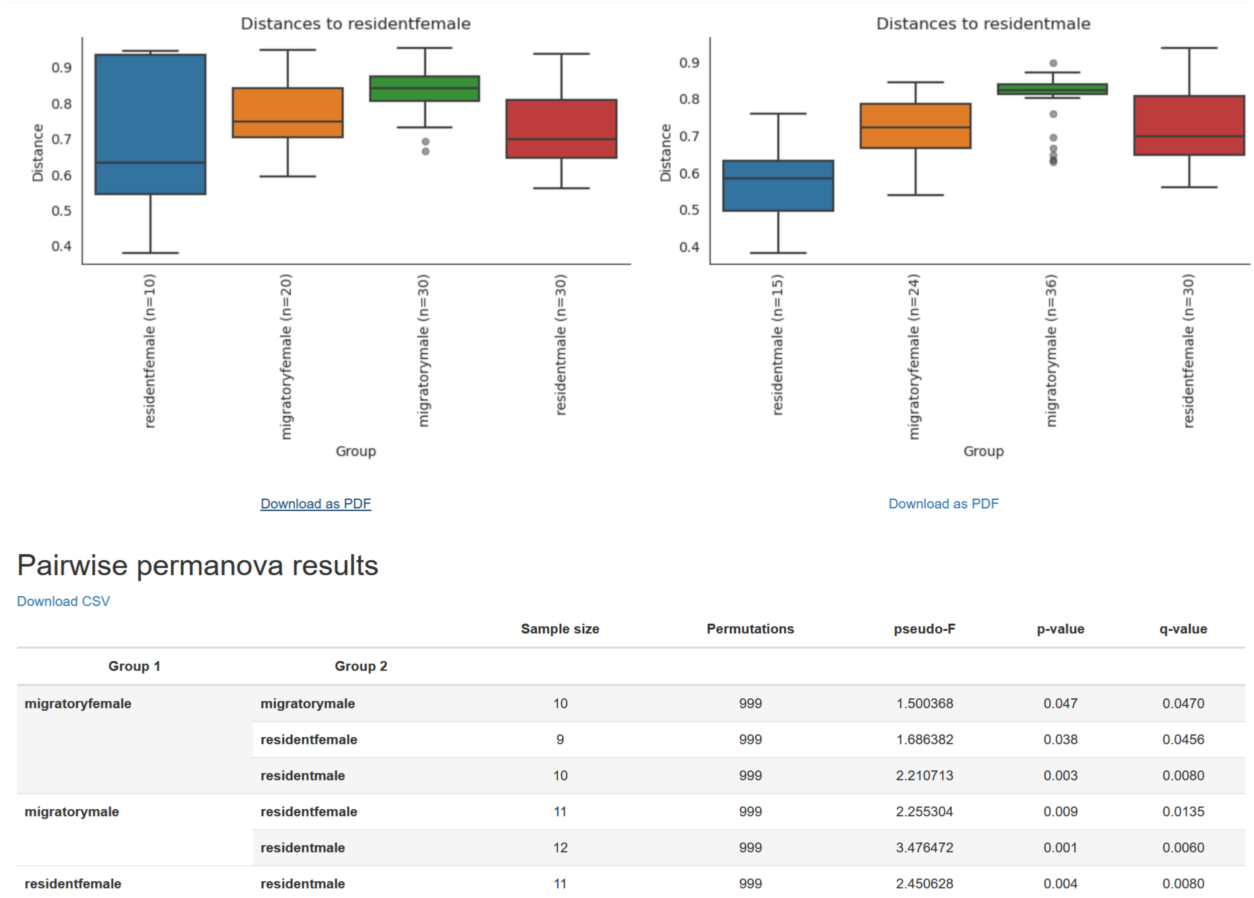
	PERMANOVA results
method name	PERMANOVA
test statistic name	pseudo-F
sample size	21
number of groups	4
test statistic	2.275176
p-value	0.001
number of permutations	999

Group significance plots

Download raw data as TSV

Download as PDF

Download as PDF



Pairwise permanova results

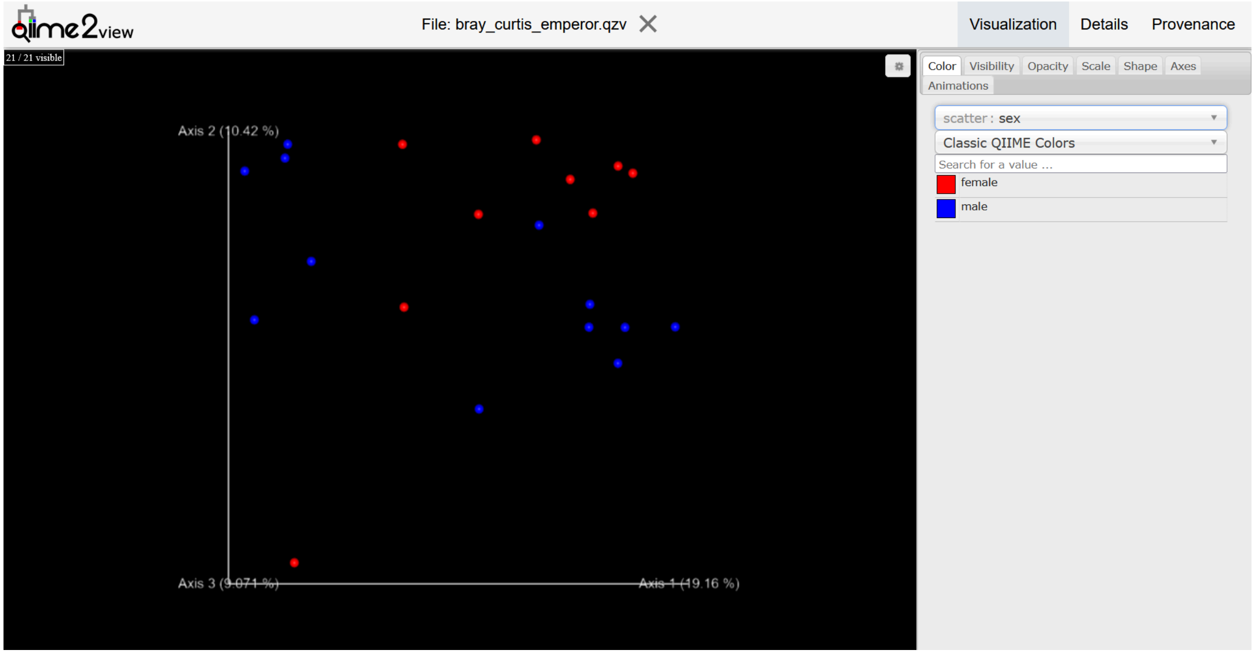
Download CSV

Group 1	Group 2	Sample size	Permutations	pseudo-F	p-value	q-value
migratoryfemale	migratorymale	10	999	1.500368	0.047	0.0470
	residentfemale	9	999	1.686382	0.038	0.0456
	residentmale	10	999	2.210713	0.003	0.0080
migratorymale	residentfemale	11	999	2.255304	0.009	0.0135
	residentmale	12	999	3.476472	0.001	0.0060
residentfemale	residentmale	11	999	2.450628	0.004	0.0080

There is significant diversity between the migratory population vs resident population, migratory female flock vs migratory male flock, migratory female flock vs resident female flock, migratory female flock vs resident male flock, migratory male flock vs resident female flock, migratory male flock vs resident male flock, and resident female flock vs resident male flock. Regardless of significance, in interpreting the results, I would say that for the populations, the migratory population is more similar to itself than the resident population because the distance is lower. I would say the same about the relationship of the resident population to itself. Therefore, there is diversity between the migratory population compared to the resident population. As for the flocks, I would say the migratory female is most similar to itself, the migratory male is most similar to itself, the resident female is most similar to the resident male, so those two groups have low diversity, and the resident male is most similar to itself.

14) The core-metrics-phylogeny command generates a file called bray-curtis-emperor.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?

Sex:



File: bray_curtis_emperor.qzv

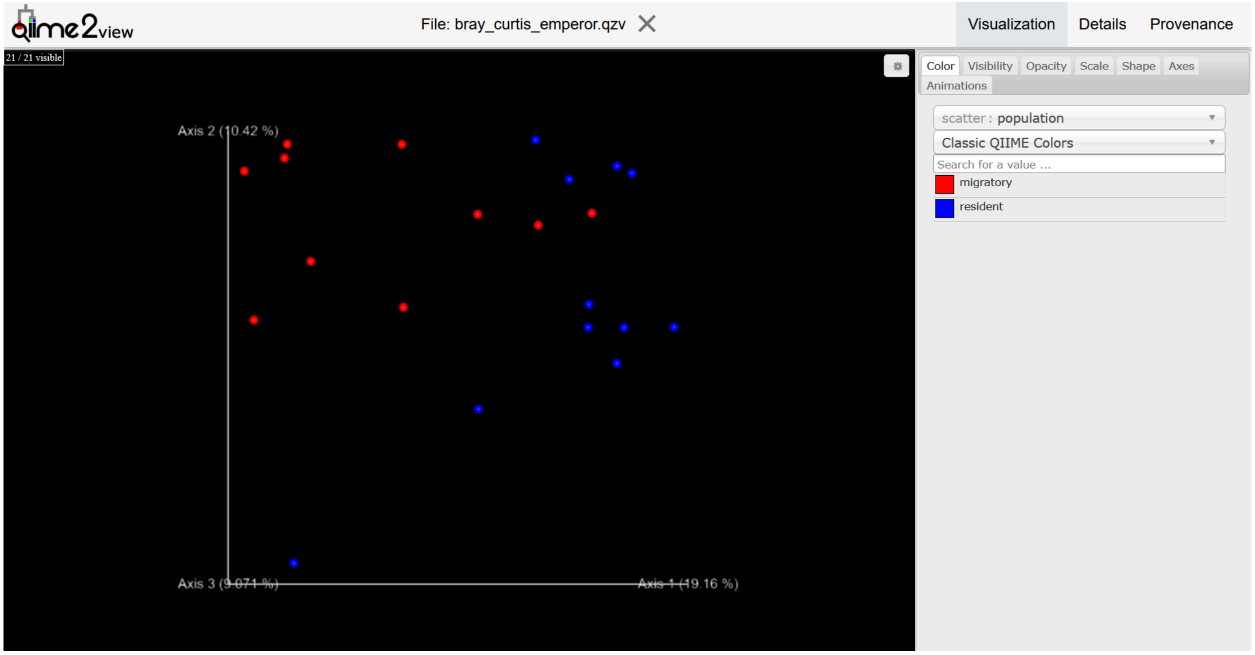
scatter : sex

Classic QIIME Colors

🟥 female

🟦 male

Population:



File: bray_curtis_emperor.qzv

scatter : population

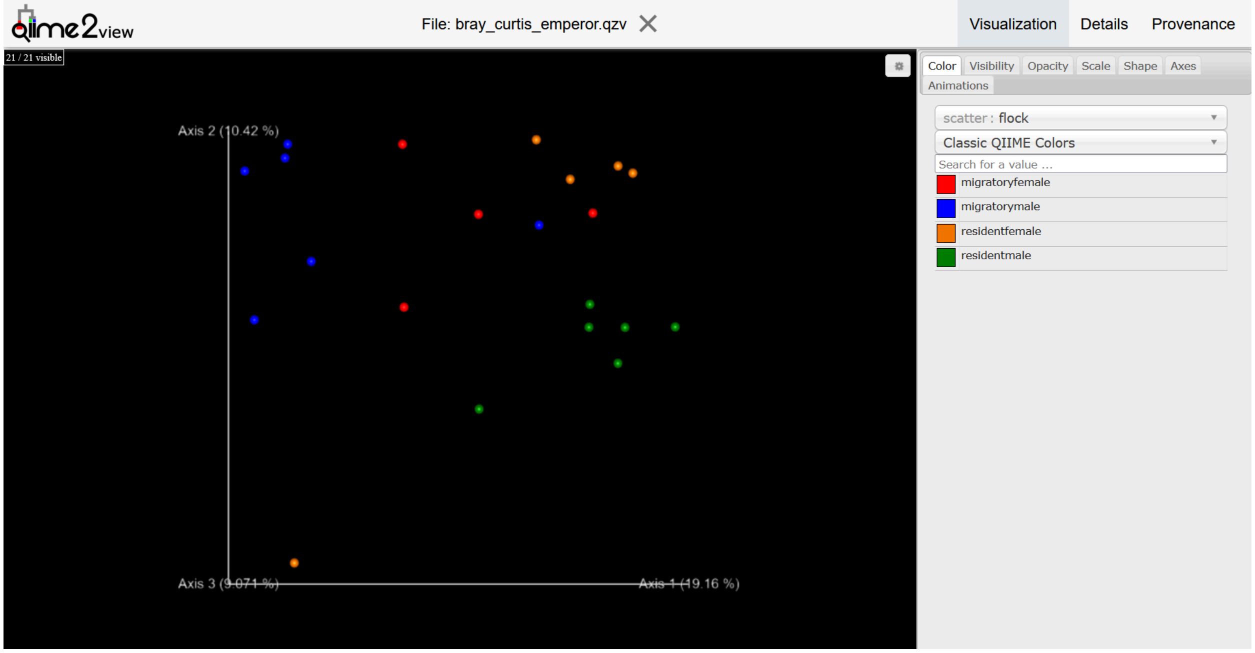
Classic QIIME Colors

Search for a value ...

🟥 migratory

🟦 resident

Flock:



File: bray_curtis_emperor.qzv

scatter : flock

Classic QIIME Colors

Search for a value ...

🟥 migratoryfemale

🟦 migratorymale

🟧 residentfemale

🟩 residentmale

These results help me make sense of the results I got from question 13 because they show where the individual features fall, so it is a more specific representation of how similar or diverse the communities are to one another.