Questions

Importing Data into QIIME2

1. After demultiplexing, which sample has the lowest sequencing depth?

recip.460.WT.HC3.D14

2. What is the median sequence length?

5101.5

3. What is the median quality score at position 125?

38

4. If you are working on this tutorial alongside someone else, why does your plot look slightly different from your neighbors? If you aren't working alongside someone else, try running this command a few times and compare the results.

The plots look slightly different every time you run it because it was generated using a random sampling of 10,000 sequences without replacement.

Feature Table Summary

1. How many total features remain after denoising?

287

2. Which sample has the highest total count of features? How many sequences did that sample have prior to DADA2 denoising?

recip.539.ASO.PD4.D14 has the highest total count of features (4,996). Before denoising it had 5475 sequences.

3. How many samples have fewer than 4250 total features?

23

4. Which features are observed in at least 47 samples?

04c8be5a3a6ba2d70446812e99318905

ea2b0e4a93c24c6c3661cbe347f93b74

1ad289cd8f44e109fd95de0382c5b252

5. Which sample has the fewest features? How many does it have?

recip.460.WT.HC3.D49 has 347 total features.

Alpha Rarefaction and Selecting a Rarefaction Depth

1. Are all metadata columns represented in the visualization? If not, which columns were excluded and why?

Days post-transplant because it doesn't have categorical data.

2. Which metric shows saturation and stabilization of the diversity?

Shannon

3. Which mouse genetic background has higher diversity, based on the curve? Which has shallower sampling depth?

Wild type has higher diversity. Susceptible has shallower sampling depth.

4. What percentage of samples are lost if we set the rarefaction depth to 2500 sequences per sample?

48-4= 44 → (4/44)x100= 9.09%

5. Which mice did the missing samples come from?

Wildtype
recip.461.ASO.HC3.D49

recip.220.WT.OB1.D7

recip.460.WT.HC3.D14

recip.460.WT.HC3.D49

Diversity Analysis

1. Where did we get the value 2000 from? Why did we pick that?

2000 was chosen for the sampling depth based on the rarefaction curve since we can retain 47 of 48 high quality samples at this depth.

Alpha Diversity

1. Is there a difference in **evenness** between genotype? Is there a difference in **phylogenetic diversity** between genotype?

There is not a significantly significant difference between the two in either case.

2. Based on the group significance test, is there a difference in phylogenetic diversity by genotype? Is there a difference based on the donor?

There is a difference based on the donor.

Beta Diversity

1. Open the unweighted UniFrac emperor plot (core-metrics-results/unweighted_unifrac_emperor.qzv) first. Can you find separation in the data? If so, can you find a metadata factor that reflects the separation? What if you used weighted UniFrac distance (core-metrics-results/weighted_unifrac_emperor.qzv)?

You can clearly see separation in the data when using donor as a metadata factor in both plots but it's much clearer in **unweighted UniFrac.**

2. One of the major concerns in mouse studies is that sometimes differences in communities are due to natural variation in cages. Do you see clustering by cage?

There is not much clustering by cage aside from within their respective genotype.

---

1. Is there a significant effect of donor?

Yes.

2. From the metadata, we know that cage C31, C35, and C42 all house mice transplanted from one donor, and that cages C43, C44, and C49 are from the other. Is there a significant difference in the microbial communities between samples collected in cage C31 and C35? How about between C31 and C43? Do the results look the way you expect, based on the boxplots for donor?

There is a significant difference between C31 and C35 which share the same donor, and no difference between C31 and C43 which have different donors. These results are unexpected given the donor information.

Q. Is there a significant difference in variance for any of the cages?

There is similarity in significance patterns for PERMANOVA results among cages C31, C35, and C42.

1. If you adjust for donor in the adonis model, do you retain an effect of genotype? What percentage of the variation does genotype explain?

Even after adjusting for donor effects, genotype still contributes significantly to the variation observed in the microbial community composition, explaining approximately 4.145% of the total variance.

Taxonomic Classification

1. Find the feature, 07f183edd4e4d8aef1dcb2ab24dd7745. What is the taxonomic classification of this sequence? What's the confidence for the assignment?

This classification indicates that the organism belongs to the domain Bacteria, specifically the phylum Firmicutes, class Clostridia, order Clostridiales, and family Christensenellaceae. However, the genus and species are not identified in this classification.

98.36% Confidence

2. How many features are classified as g__Akkermansia?

Two

3. Use the tabulated representative sequences to look up these features. If you blast them against NCBI, do you get the same taxonomic identifier as you obtained with q2-feature-classifier?

Yes following some unknown bacterium partial 16S gene.

Taxonomy Barchart

Q. Visualize the data at level 2 (phylum level) and sort the samples by donor, then by genotype. Can you observe a consistent difference in phylum between the donors? Does this surprise you? Why or why not?

When sorted by donor, the mice with healthy donors seemed to have more alpha diversity, and uniquely include Actinobacteria and Verrucomicrobia, while mice with PD donors had microbiomes where only 3 phyla dominate. This does not surprise me as I was expecting to see dysbiosis marked by lower microbial diversity.

Differential abundance with ANCOM-BC

1. Are there more differentially abundant features between the donors or the mouse genotype? Did you expect this result based on the beta diversity?
2. Are there any features that are differentially abundant in both the donors and by genotype?
3. How do the bar plots for the combined formula ('donor + genotype') compare with the individual donor and mouse genotype bar plots? Are there more differentially abundant features in the individual plots or the combined?

Taxonomic Classification Again

1. Examine the enriched ASVs in the da_barplot_donor.qzv visualization. Are there any of these enriched ASVs that have differing taxonomic resolution in the dada2_rep_set_multi_taxonomy.qzv visualization?

Yes.

2. If so, which taxonomy provided better resolution?

04195686f2b70585790ec75320de0d6f has higher taxonomic resolution for the data set that was trained.

54f7ee881a58ad84fe3f81d76968b072 has higher resolution using the untrained data set.

3. Is this what we expect, based on what we learned about taxonomic classification, accuracy, and re-training earlier in the tutorial?

Yes, this outcome aligns with our understanding of taxonomic classification and classifier performance. The ASVs that are enriched may represent taxa that are more abundant or specific to certain conditions or treatments in our dataset. When comparing the taxonomic resolution between the two visualizations, we find differences in resolution for some ASVs. This discrepancy could be attributed to the effectiveness of the classifiers used in each visualization.

PCoA-based Analyses

1. Open the unweighted UniFrac emperor plot and color the samples by mouse id. Click on the "animations" tab and animate using the day_post_transplant as your gradient and mouse_id as your trajectory. Do you observe any clear temporal trends based on the PCoA?

Yes, there seems to be clustering among 536, 537, 538, 539, 546, and 547.

2. Can we visualize change over time without an animation? What happens if you color the plot by day_post_transplant? Do you see a difference based on the day? *Hint: Try changing the colormap to a sequential colormap like viridis*.

I can still see some clustering, but it's not as clear as the animation, to me.

3. Using the controls, look at variation in cage along PCs 1, 2, and 3. What kind of patterns do you see with time along each axis?

They are more volatile at first and with time become more stable.

Distance-based analysis

Q. Based on the volatility plot, does one donor change more over time than the other? What about by genotype? Cage?

Based on donor and genotype, it appears that the two groups have similar volatility. When sorted by cage, it appears that C42, C43 and C49 are more volatile than the others.

1. Is there a significant association between the genotype and temporal change?

Yes.

2. Which genotype is more stable (has lower variation)?

Susceptible.

3. Is there a temporal change associated with the donor? Did you expect or not expect this based on the volatility plot results?

Yes there is a temporal change associated with donor. I did expect this based on the volatility plot results.

4. Can you find an interaction between the donor and genotype

Based on the regression analysis there appears to be a significant interaction between the donor (pd_1) and the genotype (wild type).

Machine-learning classifiers for predicting sample characteristics

> What features appear to differentiate genotypes? What about donors? Are any ASVs specific to a single sample group?

There isn't a clear difference among genotypes that I can tell. However, there is a clear contrast between donors, as seen with the top half of the heatmap being colored on the left side of the map, while the bottom half are black. This shows that those ASVs are being found in varying frequencies in healthy donors but not in donors with PD.

Yes, there are many ASVs specific to a single sample group. They are the columns that have a black square for 3 treatments and any colored square for at least 1. On the far left side of the heatmap there are 4 such instances of unique ASVs belonging to only wild-type and healthy. They are as follows:

1e6th78e2244c.94019fc310880711d6

514b34f09e4acd2397d38273c630d338

8fe9306f9c749993b3ae825da7d495fa

1a285982a823e4cdbfa889bd75bad9f4