QIIME 2 Questions

Importing data into QIIME 2

1. After demultiplexing, which sample has the lowest sequencing depth?
   Recip.460.WT.HC3.D14 was found to have the lowest sequencing depth of 4237.
2. What is the median sequence length?
   5101.5
3. What is the median quality score at position 125?
   At position 125, the median score was 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 different as the training and testing data used to run the model is split unevenly.

Sequence quality control and feature table¶

1. How many total features remain after denoising?
   287 features.
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 highest total count of features (4996). Prior to DADA2 it had 5475.
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 only 347 features.

Generating a phylogenetic tree for diversity analysis

Start by opening the alpha rarefaction visualization.

1. Are all metadata columns represented in the visualization? If not, which columns were excluded and why?
   The redundant metadata columns where the column excluded.
2. Which metric shows saturation and stabilization of the diversity?
   The Shannon metric.

3. Which mouse genetic background has higher diversity, based on the curve? Which has shallower sampling depth?
   Mouse_id 547 seems to posses the highest diversity. Mouse_id 456 has shallowest sampling depth.

Now, let's check the feature table summary.

4. What percentage of samples are lost if we set the rarefaction depth to 2500 sequences per sample?
   8.33%
5. Which mice did the missing samples come from?
   Mouse_id 457,469,537,and 538.

Diversity analysis

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

The value came from the rarefaction plot. It was picked as it minimizes the number of samples lost.

Alpha diversity

1. Is there a difference in **evenness** between genotype? Is there a difference in **phylogenetic diversity** between genotype?
   The susceptible genotype and the wildtype genotypes have relatively the same evenness, it is mostly the same for phylogenetic diversity and genotype as well, wildtype has a value around 7.
2. Based on the group significance test, is there a difference in phylogenetic diversity by genotype? Is there a difference based on the donor?
   Examining the group significance test, the phylogenetic diversity by genotype the values are found to between 7-6.5, showing little variation. For donor, hc_1 has a value of around 7 on the box plot and pd_1 is approximately 6.1.

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`)?
   The separation in the data is potentially due to genotype (susceptible vs. wildtype).
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?
   In this study, it does not appear that there is clustering by cage primarily because there are some susceptible mice that are considered healthy and vice versa.
3. Is there a significant effect of donor?
   The is more variation between hc_1 and pd_1 in the unweighted unifrac than the weighted plot. Between the two types of plots, it could be said that donor does hold a noticeable effect.

4. 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?
   Comparing C31 and C35, the differences among the boxplots is relatively small with the most noticeable difference being present in the plot regarding distance to C44. Between C31 and C43 the difference is quite drastic, which is expected as they are from different donors.

5. Is there a significant difference in variance for any of the cages?
   The most significant variance between the cages is primarily C49 and C31 in Distances to C49 and C43 and C31 in Distance to C43.
6. If you adjust for donor in the adonis model, do you retain an effect of genotype? What percentage of the variation does genotype explain?
   The effect of genotype appears to be more insignificant compared to donor as its R2 value its smaller than that of donor. Genotype explains around 4 % of the variation.

Taxonomic classification

1. Find the feature, `07f183edd4e4d8aef1dcb2ab24dd7745`. What is the taxonomic classification of this sequence? What's the confidence for the assignment?
   k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Christensenellaceae; g__; s__. Confidence: 0.9836881157645692
2. How many features are classified as `g__Akkermansia`?
   Two features.
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?
   Upon blasting against NCBI, you might not be able to acquire the same taxonomic identifies because of ambiguity that cause certain sequence to be deemed unclassified.

   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?
   Majority of the samples sorted by donor and genotype have a prevalence of phylum Bacteroidetes and Firmicutes, with some outliers possessing Verrucomicrobia. Differences between the donors is noticeable but most have Bacteroidetes and Firmicutes.

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?
   Between the donors as donor had more effect of the results than genotype.
2. Are there any features that are differentially abundant in both the donors and by genotype?
   Feature known as ac5402de1ddf427
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?
   There are more differentially abundant features present in the donorpd_1 plot.

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?
   04195686f2b70585790ec75320de0d6f appears to have differing taxonomic resolution in the dada2 visualiaztion, as well as 54f7ee881a58ad84fe3f81d76968b072.
2. If so, which taxonomy provided better resolution?
   The bespoke_taxonomy.qza has better resolution as it is more descriptive.
3. Is this what we expect, based on what we learned about taxonomic classification, accuracy, and re-training earlier in the tutorial?
   This is expected as certain classifiers get further refined, the better resolution will come from more developed models.

Longitudinal analysis

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? 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?
   The separation of the data caused by PCoA, is further shown via the animation as most grouped data stay segregated, with the exception being 468 as possesses data points in both groups.
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.*
   Yes, when colored by day_post_transplant, the data points are colored corresponding to the days elapsed from the transplant. For instance, day 1 is red, day 2 is blue, etc.

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?

In Axis 1, most of the cage_ids are relatively level with little change, with the exception of C42. In Axis 2, C31 and C35 being to drop beginning at the 28th day. In Axis 3, most cage_ids are level again except for C42 which awkwardly dips at day 14 and rebounds back to its original value at around day 21.

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

Based on the unifrac volatility plot, it appears that the donors are diverging from one another at the same rate. Genotype seems to converge then being to diverge at around day 43. When observing cage_ids, majority of cage_id values seem to increase over time except for C43 and C49 which are seen to decrease.

1. Is there a significant association between the genotype and temporal change?
   There does not appear to be much association between genotype and temporal change as not much change is observed as days progress.

2. Which genotype is more stable (has lower variation)?
   It appears that the susceptible genotype has less variation compared to wildtype.
3. Is there a temporal change associated with the donor? Did you expect or not expect this based on the volatility plot results?
   The temporal change should be the same for the donors compared based on the results of volatility plot.
4. Can you find an interaction between the donor and genotype?
   It can be stated that donor and genotype possess and inverse relationship.

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?

Features 1ad289cd8, ea2b0e4a9, and 04c8be5a3 can potentially differentiate genotypes a well as donors, for they have high frequency in the 4 categories being analyzed. 64d0a182 and 7ce470a3 seem to be specific to just wild type and PD.