Erin Horkan

Parkinson's Mouse Tutorial

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 sequence count, 150nts length

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 plot was generated using random sampling of the dataset without replacement. Every time the plot is generated a different random sample will be selected leading to slight differences in the plots.

Sequence quality control and feature table

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, 5475 sequences before denoising

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, 347 features

If you open the denoising summary, can you find the step where the sample with the fewest sequences fails?

2021

Alpha Rarefaction and Selecting a Rarefaction Depth

Start by opening the alpha rarefaction visualization.

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

   Days_post_transplant was excluded because it did not contain categorical data

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

   Diversity as a function of sampling depth (rarefaction curve)

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

Wild type had higher diversity, and shallower 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?

   About 10% (4 of 48)

5. Which mice did the missing samples come from?

   Wild type, 1 ASO

Diversity Analysis

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

2000 is the rarefaction depth (sequences/sample) this was an informed but still slightly arbitrary selection. 2000 keeps 47/48 samples and gets rid of 1 sample with sequencing depth below 1000 seq/sample which is the general best estimate for high biomass samples.

Alpha Diversity

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

   There is not a significant difference in evenness between genotypes though there is a greater spread than the phylogenetic diversity plots.

   There is not a difference of phylogenetic diversity between genotype.

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

   No for genotype, yes for donor (p-value 0.05)

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)?

   Yes, cage-ID, Donor, and mouse ID all seem to have some separation in the unweighted dataset. Possibly less separation in the weighted dataset but the separation is still evident.

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?

   Yes there is some clustering by cage.

3. Is there a significant effect of donor?

   Yes

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?

   There is not a significant difference between C31, and C35. There is between C31 and C43. This makes sense as there was a significant difference between the donors.

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

   No

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?

   Genotype explains 4% of the variation due to donors, while donor accounts for 30% of the variation due to genotype.

Taxonomic Classification

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

Bacteria, Clostridiales christensenellaceae, confidence 98%

2. How many features are classified as g__Akkermansia?

   2

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?

   Sure

Taxonomy Barchart

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?

Hc_1 had more phylum diversity with Verrucommicrobia, Actinobacteria present in most hc_1 samples but none of the pd_1 samples.

Genotype was less consistent, but about 1/3 of the susceptible genotype had Verrucomicrobia present while none of the wild type did.

This is not super surprising because there was a significant difference between the donors.

Differential Abundance with ANCOM-BC

Open the da-barplot visualizations for donor and genotype as the selected ANCOM-BC formula term.

1. Are there more differentially abundant features between the donors or the mouse genotype? Did you expect this result based on the beta diversity?

   There are way more differentially abundant features between the donors compared to mouse genotype. This is in line with the previous diversity metrics.

2. Are there any features that are differentially abundant in both the donors and by genotype?

   No.

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 when compared to the mouse genotype, but fewer compared to the individual donor bar plots.

Taxonomic Classification Again

Open up the dada2_rep_set_multi_taxonomy.qzv visualization and the da_barplot_donor.qzv visualization.

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?

   Enriched Bacteriodes had higher taxonomic resolution. Almost all were classified to at least the genus, most to the species. Other species of bacteria like clostridiales and gammaproteobacterial while enriched were classified to the family.

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

   Yes, taxonomic classification is dependent on available bacteria species genomes with which to compare datasets. Bacteroides is a more well study group of bacteria so the taxonomic classification is possible at a finer resolution.

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?

   The mice stay clustered around themselves except for mouse 468. There is more movement in the first few days after the transplant with a sharp decrease for mice 435, 436, and 456 at the end of the study.

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 this shows a more clear difference between days.

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?

   Initially there is an increase, as the experiment progresses, they slowly decrease.

Distance-based Analysis

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

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

   No.

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

   Susceptible is slightly more stable.

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

   Yes, the volatility plots and temporal analysis show that the microbiome in different genetic backgrounds changed differently over time.

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

   Donor has the greatest influence on the data. Genotype and Donor combined have a larger influence than just donor type.

Machine-learning classifiers for predicting sample characteristics

How did we do? Just for fun, try predicting some of the other metadata columns to see how easily cage_id and other columns can be predicted.

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

There are features that differentiate genotype, and donors.

These features are associated with HC but not in PD, similar differences are present for donor type.