QIIME 2

PD Tutorial

Question

1. After demultiplexing, which sample has the lowest sequencing depth? recip.538.WT.PD4.D21
2. What is the median sequence length? 150 nts
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. I tried running several times, and the plot looked exactly the same for me every time.

Sequence quality control and feature table

Question

1. How many total features remain after denoising? 287
2. Which sample has the highest total count of features? recip.539.ASO.PD4.D14 How many sequences did that sample have prior to DADA2 denoising? 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? recip.460.WT.HC3.D49 How many does it have? 347

If you open the denoising summary, can you find the step where the sample with the fewest sequences fails?
recip.460.WT.HC3.D49 16327 9919 60.75 347 347 2.13

Alpha Rarefaction and Selecting a Rarefaction Depth

Question

Start by opening the alpha rarefaction visualization.

1. Are all metadata columns represented in the visualization? No. If not, which columns were excluded and why? Sample_name and days_post_transplant was excluded. These metadata columns have been omitted because they didn't contain categorical data, or the column consisted only of missing values: days_post_transplant.
2. Which metric shows saturation and stabilization of the diversity? Barcode shows saturation and stabilization of the diversity from both Shannon and observed_features
3. Which mouse genetic background has higher diversity, based on the curve? Wild type has higher diversity based on the curve.Which has shallower sampling depth? Susceptible has 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? 8.33%
5. Which mice did the missing samples come from? Wild type

Diversity analysis

Question

Where did we get the value 2000 from? We used the alpha rarefaction visualization and the sample summary visualization to pick a rarefaction depth. Why did we pick that? This will let us keep 47 of 48 high quality samples (discarding the one sample with sequencing depth below 1000 sequences/sample).

Alpha diversity

Question

1. Is there a difference in **evenness** between genotype? No. Is there a difference in **phylogenetic diversity** between genotype? No.
2. Based on the group significance test, is there a difference in phylogenetic diversity by genotype? No. Is there a difference based on the donor? Yes.

Beta diversity

Question

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? Donor and donor_status. What if you used weighted UniFrac distance (core-metrics-results/weighted_unifrac_emperor.qzv)? Donor and donor_status
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 unweighted_unifrac_emperor.qzv, C31, C35, and C42 are clustered together. C43, C44, and C49 are clustered together. The same is true for weighted_unifrac_emperor.qzv.

Question

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? For **weighted-unifrac-cage-significance, we got insignificant p-value** between samples collected in cage C31 and C35 (this might be due to the boxes are overlapping with each other). For **unweighted-unifrac-cage-significance, we got significant p-value** between samples collected in cage C31 and C35. How about between C31 and C43? For **weighted-unifrac-cage-significance, we got significant p-value** between samples collected in cage C31 and C43. For **unweighted-unifrac-cage-significance, we got significant p-value** between samples collected in cage C31 and C43. Do the results look the way you expect, based on the boxplots for donor? Yes, based on the boxplots for donor.

Question

Is there a significant difference in variance for any of the cages? The p-value is 0.235, so the difference in variance for any of the cages is insignificant.

Question

1. If you adjust for donor in the adonis model, do you retain an effect of genotype? The effect of genotype is significant. What percentage of the variation does genotype explain? 4.14%

Taxonomic classification

Question

1. Find the feature, 07f183edd4e4d8aef1dcb2ab24dd7745. What is the taxonomic classification of this sequence? k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Christensenellaceae; g__; s__ What's the confidence for the assignment? 0.9836881157645692
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? I did blastn in NCBI for the feature 07f183edd4e4d8aef1dcb2ab24dd7745, the top hit was Uncultured bacterium clone B4826 16S ribosomal RNA gene, partial sequence. I was not able to find the taxa, rather it says uncultured bacterium.

Taxonomy barchart

Question

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? Yes, this is surprising. Healthy donors have k__Bacteria;p__Verrucomicrobia and k__Bacteria;p__Actinobacteria, whereas this is missing in PD Donors. Similarly, susceptible genotype contain k__Bacteria;p__Verrucomicrobia and this is missing in wild type.

Differential abundance with ANCOM-BC

Question

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? There are more differentially abundant features in PD donors and there is only one feature for genotypewild. Did you expect this result based on the beta diversity? No.
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? The bar plots for the combined show less differentially abundant features than the individual donor bar plots, but show more features than the mouse genotype bar plots. For example, feature ac5402de1ddf427ab8d2b0a8a0a44f19 are enriched in ('donor + genotype') but depleted in individual donor bar plots. Are there more differentially abundant features in the individual plots or the combined? There are more differentially abundant features in the individual plots.

Taxonomic classification again

Question

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? 04195686f2b70585790ec75320de0d6f, 54f7ee881a58ad84fe3f81d76968b072
2. If so, which taxonomy provided better resolution?bespoke taxonomy for 04195686f2b70585790ec75320de0d6f and taxonomy for 54f7ee881a58ad84fe3f81d76968b072
3. Is this what we expect, based on what we learned about taxonomic classification, accuracy, and re-training earlier in the tutorial?

No, this is not what we expect based on what we learned about taxonomic classification, accuracy, and re-training earlier in the tutorial. Re-training should improve taxonomic resolution and accuracy.

Longitudinal analysis

PCoA-based analyses

Question

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
2. Can we visualize change over time without an animation? No. What happens if you color the plot by day_post_transplant? The samples are plotted based on the post transplantation days ie., 7, 14, 21, and 49 in ordination space. Do you see a difference based on the day? Yes. *Hint: Try changing the colormap to a sequential colormap like viridis.*

Question

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, the cage_id, C31, C35, and C42 are grouped together and C43, C44, and C49 are grouped together separately. In axis2, C31, C35, and C42 decrease with day_post_transplant and C43, C44, and C49 increase with day_post_transplant. In axis 3, it is same as axis 2, but the cage_id's are close to each other when compared to axis 2. Also, from metadata, C31, C35, and C42 are healthy and C43, C44, and C49 are PD.

Distance-based analysis

Question

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

Question

1. Is there a significant association between the genotype and temporal change? Yes.
2. Which genotype is more stable (has lower variation)? Wild type
3. Is there a temporal change associated with the donor? Yes. Did you expect or not expect this based on the volatility plot results? I expected this based on the volatility plot results.
4. Can you find an interaction between the donor and genotype? Yes, interaction between the donor and genotype is significant.

Machine-learning classifiers for predicting sample characteristics

Question

How did we do? Looks like we did pretty well! Just for fun, try predicting some of the other metadata columns to see how easily cage_id and other columns can be predicted. Tried for cage_id other than genotype and donor_status.

Question

What features appear to differentiate genotypes? What about donors? Are any ASVs specific to a single sample group? Although there wasn't a clear pattern in the barchart at the phylum level between donors or genotypes, we were still able to find ASVs which differentiated the genotypes using ANCOM-BC and Random Forest classification. There was no overlap between these ASVs in the donor and genetic background, supporting the hypothesis that the difference due to genotype is separate from the difference due to donor.