Ben Caulfield
668 Bioinformatics Project

QIIME2 Mouse Parkinsons Questions.
Question blocks are in a colored spectrum to help differentiate.

After demultiplexing, which sample has the lowest sequencing depth?

Recip.469.WT.HC3.D14 if I'm reading the site right.

What is the median sequence length?

5101.5

What is the median quality score at position 125?

38

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.

Assuming that the demultiplexing has some random element that could lead to the program producing differing amounts of sequences which have different lengths.

How many total features remain after denoising?

287

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

Recipe.539.ASO.PD4, 5,463 -> 4,996

How many samples have fewer than 4250 total features?

Going by the feature detail tab it's almost easier to count which samples have *over* 4250 samples. Which would be 9, subtract that from the 48 total samples and we get 39 samples.

Which features are observed in at least 47 samples?

04c8be5a3a6ba2d70446812e99318905
ea2b0e4a93c24c6c3661cbe347f93b74
1ad289cd8f44e109fd95de0382c5b252

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

recip.460.WT.HC3.D49 at 347 features

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

Is this the 16,656 (8.50%) features in 48 (100.00%) samples at 347 features deep? Or the bit right after?16,356 (8.34%) features in 47 (97.92%).

Start by opening the alpha rarefaction visualization.

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

Nope, days_post_transplant is omitted due to being non-categorical data.

>Which metric shows saturation and stabilization of the diversity?

The observed features on the y-axis, once you hit roughly a slope of 0, you've hit your saturation point.

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

Mouse id #457 has the highest amount of diversity, while #537 has the shallowest depth.

Now, let's check the feature table summary.

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

33%? I couldn't find where in the tutorial to generate the table, and setting the minimum to 2500 had things getting weird.

Which mice did the missing samples come from?

The underperforming samples at 2500 are from mice #457, 469, 537 and 538

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

Its the value that gets us the least amount of sequence loss while getting the most sequences to use in diversity stuff. 2000 sequences/sample lets us keep 47/48 samples, which is a pretty darn good ratio.

Is there a difference in evenness between genotype? Is there a difference in phylogenetic diversity between genotype?

I'd say there is no significant difference in genotype- the wildtype minimum is a bit higher, but it's definitely not something I would say is something to talk home about. Same thing with phylogenetic diversity, though wildtype has a few more outliers.

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

No significant difference by genotype, though there is a significant difference based on the donor/ donor status.

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)?

There is a separation in the data, which is pretty cleanly depicted by the two donor statuses. The weighted distance graph is somewhat similar, though the healthy group is kind of splintered.

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?

Answer

Is there a significant effect of donor?

Yes, with a p-value of 0.001

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?

Theres a significant difference in both of the comparisons- the only cages which are for sure ***not*** significantly different are C43-44, C43-49 and C44-49. C31-43 looks more obvious than C31-35, the latter of which I could definitely assume insignificant if judging solely off of the graph.

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

The p-value is 0,235, meaning that there isn’t any significantly different variance values.

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 is significant when donor is adjusted for, but with an R^2 value of 0.041, the impact on the variation of the data is marginal.

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

It’s some bacterium within the Christensenellaceae family (0.984% confidence).

How many features are classified as g__Akkermansia?

There are **two** features with that genus classification, but they have the same species name?

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?

When blastedI got the same identifiers! However, the BLAST job took over two hours to get back to me which was weird.

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?

Yep, there's a lot more phyla of bacteria represented! It’s not that surprising though given how the two donor classes were already known to be significantly different though.

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

>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 a lot more in the donors, which sort of makes sense given how sort of “binary” the donor data is, whereas genotype the beta diversity is a bit spread out.

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

I do not believe so. The only feature abundant/seen in the only-genotype is 3017… which is not seen in the donor barplot.

>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?

It's somewhere in the middle, where the genotype only had one bar and the donor had a large number. I'd say it averages out.

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

>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, bespoke identifies the feature as E. coli whereas the normal taxonomy.qza only goes to family level.
54f7ee881a58ad84fe3f81d76968b072, Alistipes massiliensis vs family level

>If so, which taxonomy provided better resolution?

The bespoke taxanomy file.

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

Yes- we gave the classifier better data to train on.

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. The hc mouse donor group's trajectories move towards Axis 2 while the pd group stays mostly in the same spot.

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.

Yep, it's even better with something like reds as the palette to really drive in that shift over time, as we get to see a clear shift in color hue as time progresses.

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?

PC 1 the two groups stay pretty distinct from each other aside from one healthy group which spikes towards PD. PC 2 the two groups stay close, intercept before diverging. PC 3 the two groups are pretty intertwined for the duration of the data.

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

From how I interpret it, the healthy group changes more over time, while the PD group changes less. Genotype the two groups end up (visually) barely different then where they started from. Cage is pretty crazy though with some drastic changes, I'd say all of the cages changed substantially.

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

Just barely, at 0.044

Which genotype is more stable (has lower variation)?

Susceptible genotype seems to be more stable.

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

Yes, the hc group changes over time, which is pretty much what we saw from the volatility plot.

Can you find an interaction between the donor and genotype?

There is something significant, so I would assume there is an actual interaction there.

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.

I'd say it went pretty well, though wildtype+healthy was misjudged a bit.

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

79280… and 3017f… in particular seem to differentiate between genotypes, while what looks like a quarter to a third of the features seem to be particular to one donor group. A smaller group of features seem to stick to one sample group, like 7ce47… to W+PD, 1e6fb… to W+H, and 5bf22… to S+PD.