Exploring Native Microbiota Lab Manual

BIO398, Winter 2024

Lab 8 : Analysis of qiime2 output files, preparation for informal reports

---

Overview:

In the last 2 and a half weeks, we have been working on some basic command-line literacy and I have also demonstrated how to demultiplex and analyze amplicon sequencing data. As we look towards our final presentations and your organism of interest (OOI) report, it's important to spend some time to explore your dataset and identify patterns you wish to follow up on.

qiime2 generates wonderful interactive visualizations. I will have, at minimum, provided you with the raw output from the ASV pipeline (https://github.com/jcmcnch/eASV-pipeline-for-515Y-926R). If you have given me your **sample-metadata.tsv** and **samples-to-keep.tsv** file, I will have also given you subsetted barplots (to exclude samples with low sequence counts) and diversity plots (that include both metrics and 3D visualizations). These will form the basis of today's exploration activities.

Important note about qiime2 files and command-line instructions:

- "**.qzv**" files are visualizations ("v" for "visualization") that need to be opened on a special website (https://view.qiime2.org/), and
- "**.qza**" files are archives ("a" for "archive"). The latter are basically zip files and should be openable in any old archive manager but you might have to select "open with".
- Some command-line instructions discussed below are available on github here: https://github.com/stfx-microeco-lab/nativemicrobiota

---

Suggestion for keeping track of your explorations:

It is a good idea to make a word/google doc/something similar to keep track of all the different visualizations you end up generating along with notes about interesting patterns you observe. You can either take screenshots or export the plots as svgs (screenshots are probably easier).

Recording other types of metadata:

Today is also a good opportunity to record additional metadata for your samples from stored water/sediment. However, *I recommend you do the analysis steps below first since we will be presenting an informal report on Friday and this lab is your opportunity to get it ready. We will have time later to measure these variables in the remaining two labs.*

Suggested steps for today's lab:

1. If you haven't already, prepare your **sample-metadata.tsv** and **samples-to-keep.tsv** files:
    1. You should inspect your 16S **stats.tsv** file to determine which, if any, samples to exclude. Include *only* those samples that merit further analysis in your **samples-to-keep.tsv** file.
    2. Your **sample-metadata.tsv** file should contain any categorical or numeric data that you might find useful to interpret your data. It must be saved in a tab-separated file (plaintext).
    3. When you have a draft, please show the files to me first, then once I say everything looks good then please email them to me.

Some tips for exploring qiime2 interactive plots:

- Don't forget to resize the bars by dragging the slider (they start off very thin).
- Try exploring at different taxonomic levels. Higher levels might give you a sense of overall diversity and provide more specific information but can be overwhelming. Sometimes somewhere in the middle can give you a "bird's eye view" of what is different between your samples.
- Sort bars by your metadata to more easily visualize patterns.
- Remember all the subsets - they can help reduce the complexity and focus in on specific functional types (e.g. phototrophs).
- Note that in some cases, you will have very few or no sequences for some subsets or samples. This could either be due to the characteristics of your sample (i.e. low 18S fraction, no chloroplasts) or due to low overall sequencing counts. If you see a bar with a single taxon, that probably indicates low sequencing counts.
- Take taxonomic identifications with a grain of salt - they are just what the ASV matches best to in the database, and are not a definitive species identification. See below for a way in which you can extract ASV sequences and get more information by BLASTing.

2. **Inspect your barplots (see tips above)**. Remember there are various subsets, including the following (as long as you have sent me the above files, if not, you will just have 16S and 18S):
    1. 16S:
        1. Everything
        2. Archaea

3. Cyanobacteria
4. Cyanobacteria + Chloroplast 16S
5. Chloroplast 16S
6. Bacteria only

2. 18S:
    1. Everything
    2. Metazoa
    3. Everything but metazoa
    4. The above 3 categories but with different databases (we use both PR2 and SILVA)

Some notes on diversity metrics:

- Diversity metrics usually refer to one the following things:
    - How many taxa are present ("richness"; higher = more species)
    - How evenly distributed abundances of species in a sample is ("evenness"; higher = more even)
    - Other metrics that describe evenness in slightly different terms like "entropy" (higher entropy = less predictable, more even)
    - How dissimilar samples are from one another ("distance")
- The 3D plots are a reduction of our high-dimensional data into a more human-interpretable format. Don't worry too much about the underlying mathematics, but understand what it's telling you.

3. **Inspect your diversity metrics**:
    1. Play about with the "Emperor" plots ("**.qzv**" files) - drag them around, label the dots according to your metadata, etc.
        1. For your interest, read about the differences between "Jaccard" and "Bray-Curtis" diversity metrics.
    2. Check out the metrics stored in the "**.qza**" files by opening them in LibreOffice Calc/Excel. Make some simple barplots if you see any patterns you want to record.
        1. You should also try to have a general idea what these metrics mean.
4. **Inspect the raw ASV tables**:
    1. These will be found as "tsv" files in the original archive I sent you, and should be opened in LibreOffice Calc/Excel for easier exploration (may need to select "open with" if you're using Excel).

2. Try sorting the columns to see which ASVs are dominant in which sample.
3. Some questions to ponder:
    1. How even are the samples (in terms of diversity)?
    2. Do you notice any potential contamination in your samples based on comparisons with your negative controls?
    3. Are there any ASVs shared in all your samples?

5. For ASVs that are of particular interest to you, I suggest extracting those sequences and running a BLAST, as the taxonomic identifiers are not definitive. This can be done one of two ways:
    1. The crude way: just open up the ASV fasta file in a text browser and search for the ASV id of interest (one at a time)
    2. The sophisticated way: use **seqtk** to subset the ASV fasta file (many sequences at a time; see instructions on github : https://github.com/stfx-microeco-lab/nativemicrobiota/tree/main/extracting-asv-sequences
    3. In both cases, you'll paste the ASV sequences into NCBI BLAST as before, and you can try asking some of the following questions:
        1. How similar is your ASV to others previously reported?
        2. What kind of study was it - cultivation-dependent or cultivation-independent?
6. If you get through all this and are interested to explore more, check out the box below for some suggestions.

Options for further exploration:

- Search for other commands you might want to run in the qiime2 documentation: https://docs.qiime2.org/2024.2/plugins/
    - One that might be fun for differential abundance analysis: https://docs.qiime2.org/2024.2/plugins/available/composition/ancom/
    - Or if you wanted to make phylogenetic trees: https://docs.qiime2.org/2024.2/plugins/available/phylogeny/
- Do a BLAST versus a different database on NCBI (ask me how).
- Read papers that are from organisms identified either with a keyword search or a BLAST search (see above):
    - Try to understand their physiology, environmental role, whether they are cultivated/uncultivated
- (advanced) do a BLAST versus a different database (GTDB, someone else's data) using command-line tools. This can be done on the server (ask me for help).
- (advanced) install conda / the qiime2 pipeline on your computer (follow github instructions).