**Case Study 8: Analysis of COI Amplicon Data**

Today, the data we are going to play with comes from a 2018 study by Jeunen et al., “Environmental DNA (eDNA) metabarcoding reveals strong discrimination among diverse marine habitats connected by water movement” (<https://doi.org/10.1111/1755-0998.12982>). Water samples were collected from four different zones, mudflats, rocky shore, sandy beach and open ocean in New Zealand. They collected amplicon sequence data from multiple marker sets to analyze different types of organisms within the zone but for our purposes, we will learn how to analyze just cytochrome oxidase subunit 1 (COI) metabarcoding data using established pipelines.

1. What part of the cell will you find COI?

2. Is this gene found in all domains of life? If not, which one(s)do you find it?

Most of the activity will be run as an interactive job on the server and then the later portion will be run using the interactive Rstudio interface through the Grace Portal. Before we get to far ahead, download the text file with the commands posted on Canvas to your personal computer. Open this text file in a text editing program like Text Edit or notepad.

Next, please login to the Grace environment using either Putty or Terminal. Once you are logged in, navigate to **YOUR** working directory. In your directory, make a new directory called “12\_amplicon”. Now, inside the new directory, run the following command:

cp /scratch/group/kitchen-group/MARB\_689\_Molecular\_Ecology/12\_amplicons/eDNA\_analysis\_qiime2.txt .

If you are successful, you will have copied over the file into your personal directory called 12\_amplicon. Now, we will start an interactive job as follows:

srun --time=02:00:00 --mem=7G --ntasks=4 --cpus-per-task=1 --pty bash

In this interactive job, you can directly submit commands to the node you’ve been allocated to without having to submit a batch script. This will allow us to see the output in real time.

On the command prompt, load the appropriate module required for today:

module load QIIME2/2024.10-Amplicon

Now, return to your text file. We are going to copy the first import data command. At first it will not look like anything is happening. When the job successfully finished, a green line of text will appear.

Now run the next two steps and make sure that it goes green each time. These two steps are importing a COI database of sequence extracted from NCBI.

3. This data has already been de-multiplexed. What does that mean?

Run the next command (qiime demux summarize). After it is finished, download this output file “single-end-demux.qzv” to your personal computer. Recall, you will need an intermediate tool to download the file locally.

4. On your web browser, go to <https://view.qiime2.org/>. Drag your file to the labeled area and see what it produces.

A. How many total reads are we analyzing?

B. How many reads on average does each sample have?

C. The interactive plot shows Phred scores across the sequences. What is a Phred score and how is it used to assess to data quality?

D. Would these be considered “good” reads? Why or why not?

Now go back to your shell window and run the cutadapt command.

5. Why do we need this step?

After completing cutadapt, run the next command for DADA2. This is the main step that will 1) further clean up the data (de-noising), 2) filters the reads based on Phred quality scores, 3) identifies amplicon sequence variants (ASV), and 4) removes chimeras. The output is a list of unique sequences (ASVs) and tabulated counts of reads for each ASV. This is one of the longer steps in the process.

By visualizing the output of the DADA2 step, we can see there are some sequences that dominate the data. Run the next three visualization steps and download those respective files. Return to your web browser to upload them to Qiime2 View, one at a time.

6. First upload the table.qzv. Look at the frequency per feature table on the first page. A feature in this tool is a unique sequence across all analyzed samples. Thus, the plot is showing the occurrence of each sequence across all samples. What is the lowest count of reads of a specific sequence (“feature”)?

What is the highest count of reads across all samples?

Now look at the feature detail tab and note which feature ID is observed most often below. This will be important for our next step.

7. Now upload rep-seqs.qzv to Qiime2 View. What does this table show?

Locate your ID from Question 6 above in the table. Click on the sequence which will take you directly to NCBI BLAST. Once there, click on the view report button that should populate the Blast results like we did in the first case study. What is taxonomic ID of the most abundant feature?

8. For the denoising-stats.qzv file, what can you take away from this step?

Next, we will assign taxonomy to our features using VSEARCH. This tool will use the database we provide.

9. The similarity cut-off is set at 97%. Is this appropriate for eukaryotes? What level of taxonomic groupings will this cut-off hit?

10. Look at what files were populated. Which file size is bigger, unmatched or clustered? Why might that be?

Let’s further filter our data. We will apply two filters. First, we will remove chlorophytes. Second, we will exclude samples that do not retain a certain read count used as the threshold in the original study. Third, we will remove singletons.

With the filtered and taxonomically assigned data, we will perform diversity analyses. First, we will make a phylogenetic tree that is required for phyloseq in R. Second, we will estimate alpha and beta diversity. Lastly, we will estimate rarefaction curves.

Download the rarefaction curve and visualize using Qiime2 View.

11. What do the curves tell us about our sampling effort?

Return to your web browser and start an interactive Rstudio session for 1 h on the Grace Portal.

12. What are the three alpha diversity metrics used here? What distinguishes Shannon vs. Simpson’s diversity? How does our interpretation of diversity change with these metrics?

13. In the principal coordinate analysis, what groups are more similar to each other?

14. Based on the bar plot, what Class is predominant in sandy beaches? What about open water?