Lab 7: Post Lab

Figure 1. ERR59

My own sample ERR599031 from mesopelagic zone of Arabian Sea, and other samples from mesopelagic zone (ERR599031, ERR599078, ERR599090, ERR599142).

**1. Bin completeness**

The easiest bins to generate are often the ones that had the longest N50 and the lowest population diversity. Take a look at this file:

SAMPLES-MERGED/SUMMARY\_my\_bins/bins\_summary.txt.

1a. Which bin was the most complete?  
1b. What was the N50 of that bin?  
1c. What was the anvi'o-labeled taxonomy of this bin?

1d. Now find your bin folder in SAMPLES-MERGED/SUMMARY\_my\_bins/bin\_by\_bin/. Inside that folder is a file called SAMPLES-MERGED/SUMMARY\_my\_bins/bin\_by\_bin/Bin\_X/Bin\_X-Campbell\_et\_al-hmm-sequences.txt. Use 'less' to open that file. These are the single-copy universal genes (like ribosomal genes, for example) that anvi'o used to characterize the completeness of this bin. As a bonus, we can use these sequences to help figure out what this bin is. Copy the longest sequence in that file and BLAST it using blastn on the NCBI webpage. Report the e-value and description of the top hit. Does it match the taxonomy that anvi'o reports?

**2. Coverage across all bins for one sample**

Use the SAMPLES-MERGED/SUMMARY\_my\_bins/bins\_across\_samples/mean\_coverage.txt file to answer these questions.

2a. Which bin had the highest coverage across all samples, when looking only at the mapping of your own dataset (self-to-self)?

2b. What is the taxonomy of that bin, according to anvi'o and according to BLAST (see part 1 above)?

2c. Make a bar graph in which each bin is a different bar, and the bar height indicates the mean coverage. Call this 'Figure 2' and include a figure caption.

2d. What does it mean, biologically/ecologically speaking, for a bin to have high coverage?

**3. Coverage of one bin across samples**

Choose one bin that you're interested in. Use the file SAMPLES-MERGED/SUMMARY\_my\_bins/bin\_by\_bin/Bin\_X/Bin\_X-mean\_coverage.txt to answer these questions.

3a. Make a bar graph in which each sample is a different bar, and the bar height indicates the coverage of your bin in that sample. Call this "Figure 3" and include a figure caption.

3b. In which sample does your bin have the highest coverage? Second highest? What might this imply about the abundance of this microbe in different ecosystems?

**4. Variability of all bins for one sample**

Use the SAMPLES-MERGED/SUMMARY\_my\_bins/bins\_across\_samples/variability.txt file to answer these questions.

4a. Which bins had the highest and lowest variability (single nucleotide variants per kilobase pair (SNVs/kbp)) across all samples, when looking only at the mapping of your own dataset (self-to-self)?

4b. What is the taxonomy of those bins, according to anvi'o and BLAST (see above)?

4c. What do you think the variability might imply about the microbial populations represented by those bins? What might it imply about the selection pressure on those populations? (We'll talk more about this in class on Wednesday, so you might want to wait until after then...)

     \*Important caveat! If your sample had low coverage (i.e. less than 10), that may be skewing the variability results because if you don't have any reads mapping, there are no variants to count!