Andrew Presley

Bio539 Final Project

Using R to Automate Breakdown of Class and Genus Level Diatom Barcoding Data into ASVs

Introduction and Background:

Diatoms account for nearly 40% of the annual carbon fixation, making them a major contributor to global carbon cycling, and an essential component of ocean systems (Nelson et. al, 1995). In order to understand diatoms’ function in the world’s oceans, we first have to understand them from a biochemical and ecological standpoint. Often, breaking down these facets of diatom life comes via the sequencing of their genome, or DNA barcoding. Barcoding provides a way to view a snapshot of existing diatom taxa in a given sample at a specific time, this allows us to see which species are dominating in certain environments or under certain nutrient limitations. The code generated here is intended to streamline the process of taking existing relative abundances from barcoding of diatom 18S V4 and break down a specific genus or class to the species or ASV level. The process of breaking down a class using higher resolution data in the Jenkins lab in the past required the use of an Excel sheet and hand done calculations. This R script preforms the same function but with only the input of only 2 .csv files, 2 variables and the desired taxonomical names for the ASVs. Simplifying this process will allow for more efficient generation of relative abundance plots for comparison of different species reactions to changes in variables such as iron concentrations. Processing and analyzing community dynamics is key to understanding diatom ecology in our global oceans, and this program should help our lab more efficiently work towards those goals.

Methods:

Data used as the example in the generation of this script was generated by barcoding of the 18S, specifically the V4 region of the 18S. This region is used as the standard for diatom barcoding and primers used are based on those developed by Zimmerman et.al. Barcoding of this region generally yields data that allows for the assignment of taxonomy to the genus or class level, but generally not the species level. However, when a specific genus or class stands out as having unique community dynamics (very high percentage of the population for example), it is advantageous to break down this genus or class to the species or Amplicon Sequence Variant (ASV) level. One way of achieving this is using the ITS region just outside the 18S, as a method of resolving the PN genus. This region has been shown in the literature to provide greater resolution to the PN than 18S alone can, as PN has extremely high similarities between species in the V4 region used for most diatoms (Hubbard et.al, 2008). Using this region PN can be broken out into species resulting in the counts seen in the ‘pn\_asv\_relabund\_osp.csv’ these relative counts can then be used to estimate the relative abundance of the species within the *Pseudo-nitzschia* fraction.

The R-script that was written for this project acts as a tool to merge the original relative abundances with these newly resolved counts within the PN genus. By taking the counts of each ASV and dividing them by the total