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, which is generally used for community analysis. 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. In our dataset, the outlier that we want to focus in on is the genus *Pseudo-nitzschia* (PN), which makes up a very large percent of our diatom taxa. One way of achieving this increase in resolution is using the ITS1 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 counts for each sample. These newly calculated ratios are then multiplied by the existing relative abundances of Pseudo-nitzschia, generating a new relative abundance for that specific ASV. The ASVs can be assigned taxonomy determined by maximum likelihood phylogenetic trees or taxonomical assignments from programs such as QIIME2. Finally, a stacked bar plot can be generated using the new, species level relative abundances to visualize the data for papers or presentations.

Results

A screenshot of a cell phone

Description automatically generatedOnce data is processed though ‘Final\_Project\_Script.R’ the resulting plot can be used to observe the results visually. One notable observation of the data used here, is that one of the most abundant ASVs within *Pseudo-nitzschia*, ASV2, was actually assigned to *Fragilariopsis* after closer examination. *Fragilariopsis* is known to sometimes ingroup with PN during 18S V4 analysis, and our deeper analysis revealed that this was indeed the case in our data. Using the newly assigned names we can see in Fig.1 that the newly assigned *Fragilariopsis* ASV

**Fig1.** Stacked bar plot representing relative abundances of diatoms including newly assigned ASVs previously grouped as a single genus, *Pseudo-nitzschia*. ASVs with no species name are unassigned ASVs within the Pseudo-nitzschia genus, representing a species or subspecies level resolution that did align closely enough to a single species to be granted taxonomy beyond genus level with current data.

(Fragilariopsis\_New) is a large portion of EX391, EX393 and EX395, but the actual *Pseudo-nitzschia* ASVs take over in the other samples with Pn turgidula being the most abundant. Some of the ASVs were clustered in trees in a way that made them difficult to confidently assign a species, resulting in them being represented in the tree as only “ASV”. With future work using the ITS region it is likely that these ASVs will be assigned taxonomy later on in our research as our trees become more detailed and contain more ITS references to compare against. While the data presented here isn’t necessarily ready for publishing, this data represents how this script can be used to process future data as we continue to increase resolution of our barcoding data.

Using this script, it will be easy to resolve more species to higher resolution in the future as we have done here with Pseudo-nitzschia. This script can also be used for Barcoding relative abundances outside of Diatoms as well, as long as the file formats outlined in the README on GitHub are followed correctly, any relative abundance data can be input into this script to generate a plot for visualization as well as a table of updated abundances that can be used for other analysis.

Some features could still stand to be added to this script: For example, a method of automating inputs and outputs to happen without the user having to run line by line would be ideal, perhaps through terminal. Additionally, a method of assigning names to each ASV without the user having to alter the base code of the script would be good but would potentially require dozens of variables that would potentially confuse users. Overall though, this script greatly reduces the time it takes to go from raw counts to a full, presentation-ready relative abundance plot.

References

Hubbard, Katherine A., et al. “Inter- and Intraspecific Community Structure Within the Diatom Genus Pseudo-Nitzschia (Bacillariophyceae)1.” *Journal of Phycology*, vol. 44, no. 3, 2008, pp. 637–49. *Wiley Online Library*, doi:[10.1111/j.1529-8817.2008.00518.x](https://doi.org/10.1111/j.1529-8817.2008.00518.x).

Nelson, David M., et al. “Production and Dissolution of Biogenic Silica in the Ocean: Revised Global Estimates, Comparison with Regional Data and Relationship to Biogenic Sedimentation.” *Global Biogeochemical Cycles*, vol. 9, no. 3, 1995, pp. 359–72. *Wiley Online Library*, doi:[10.1029/95GB01070](https://doi.org/10.1029/95GB01070).

Zimmermann, Jonas, et al. “Barcoding Diatoms: Evaluation of the V4 Subregion on the 18S RRNA Gene, Including New Primers and Protocols.” *Organisms Diversity & Evolution*, vol. 11, no. 3, July 2011, p. 173. *Springer Link*, doi:[10.1007/s13127-011-0050-6](https://doi.org/10.1007/s13127-011-0050-6).