####### Commands used for the Qiime2 pipeline to generate custom classifiers

## The complete Diat.barcode database was filtered for both the 18S and the rbcL primer set (and removing rbcL sequences labeled “short”). Both resulting CSVs were filtered by length (mean +/- STDEV) in excel.

## for the sequences\_CSVs, only two columns of the species names and sequences were retained

# these files were then converted to FASTA files using the following command in jupyter lab:

**import os**

**!awk -F , '{print ">"$1"\n"$2}' “PATH/TO/18S/CSV” > diatbar\_18s\_all\_seqs.fasta**

**!awk -F , '{print ">"$1"\n"$2}' “PATH/TO/RBCL/CSV” > diatbar\_rbcl\_all\_seqs.fasta**

# these fasta files were then imported to qiime with the following command in bash.

**qiime tools import --type FeatureData[Sequence] --input-path [PATH/TO/FASTA] --output-path [PATH/TO/DATABASE/SEQUENCES]**

## for the taxonomy\_CSVs, only two columns of the species names and formatted taxonomy were retained

# these were converted to TSV files before importing to qiime with the following command in bash

**qiime tools import --type FeatureData[Taxonomy] --input-path [PATH/TO/FASTA] tsv --output-path [PATH/TO/DATABASE/TAXONOMY]**

## qiime custom classifiers were generated from these qza files using the following command

**qiime feature-classifier fit-classifier-naive-bayes --i-reference-reads [PATH/TO/DATABASE/SEQUENCES] --i-reference-taxonomy [PATH/TO/DATABASE/TAXONOMY] --o-classifier [PATH/TO/CLASSIFIER]**

# resulting classifiers were used in the qiime2\_lines\_for\_sample\_processing, and the qza files of both classifiers are uploaded to the repository