Anacapa Pipeline

Zack Gold

Metabarcoding Sequence Assignment and Analysis

Zack Gold, Ph.D Student

Barber Lab Dept. of Ecology and Evolutionary Biology

EEB 234 Final Project

Introduction

DNA sequencing technologies are evolving rapidly and now permit the collection of 10-15 million DNA sequences simultaneously. This advance is revolutionizing biodiversity science, particularly through the study of environmental DNA (eDNA). eDNA is a mixture of dissociated, free floating genomic DNA that accumulates from organisms living in an environment (Taberlet *et al.* 2012). This eDNA can be amplified from the environment, sequenced, and then identified to the species level through DNA barcoding, providing a rapid, cost effective, and accurate tool for assessing and monitoring biodiversity (Taberlet *et al.* 2012). Importantly, eDNA methods only require small volumes of water (less than 3L) and standard microbiology filtering techniques, allowing for simple and rapid collection of samples even in remote locations (Miya *et al.* 2015; Port *et al.* 2016). As such, eDNA is an ideal method for comparative studies, such as those necessary to test the impacts of marine management practices on marine biodiversity.

Controlled studies show that eDNA is accurate, detecting over 93% of fish species from seawater in both contained environments like aquaria as well as natural marine ecosystems like coral reefs (Miya *et al.* 2015), out performing all traditional fish survey methods including observer transect surveys and trawling (Thomsen *et al.* 2012; Port *et al.* 2016; Valentini *et al.* 2016). A strong advantage of eDNA techniques is their ability to detect rare and cryptic species that are frequently missed with traditional fish biodiversity assessment methods (Willis 2001; Valentini *et al.* 2016), including the detection of both endangered species and recently introduced invasive species, which are of special concern in conservation management (Dejean *et al.* 2012; Taberlet *et al.* 2012; Valentini *et al.* 2016). Importantly, eDNA not only detects the presence of species, but can provide critical information on the abundance of fish species as well (Port *et al.* 2016; Valentini *et al.* 2016). For example, recent work in Monterey Bay found strong correlations between fish counts and abundance of species specific sequences in eDNA analyses (Port *et al.* 2016). Lastly, eDNA techniques are sensitive, distinguishing communities in marine coastal habitats including kelp forest, sea grass, and sandy bottoms as close as 60m apart, highlighting the power of eDNA techniques to capture fish species composition differences at the microhabitat level (Port *et al.* 2016). Combined, the above advantages demonstrate the utility of eDNA techniques, allowing for unprecedented rapid, reliable, and repeatable assessments of marine fish biodiversity across a wide range of taxa (Taberlet *et al.* 2012). Thus eDNA has the capacity to revolutionize biodiversity monitoring and fisheries science by facilitating large scale comparisons of marine biodiversity across environmental gradients, anthropogenic stressors, and marine management practices.

This project seeks to streamline and automate the Anacapa metabarcoding bioinformatics pipeline to allow for standardized and rapid analysis of eDNA sequences. The Anacapa pipeline takes raw sequences, a reference library, and a mapping file and out puts assigned taxonomy, relative abundance, alpha diversity figures, and beta diversity figures and analysis.

Overview of Anacapa Pipeline

The Anacapa pipeline has one main script that calls on a series of dependencies, input files, and a master config file.

Dependencies Required to Run the Anacapa Pipeline:
1) QIIME (Caporaso *et al.* 2010)
2) PEAR (Zhang *et al.* 2014)
3) Cutadapt (Martin 2011)
4) Fastx Tool Kit (Gordon & Hannon 2010)
5) Blast (Madden 2013)
6) Custom Python Scripts in Utilities Folder https://github.com/zjgold/eeb234/tree/master/eeb-177/eeb-174-final-project

The Anacapa pipeline can be broken into three steps:
1) Sequence Cleaning
2) Taxonomy Assignment
3) Data Analysis

Sequence Cleaning

During the sequence cleaning step raw sequences are modified and compiled into a assembled cleaned text file (.fasta) for downstream analysis.

1) Raw Sequence input in a fastq file

2) Rename fastq by user defined sample instead of Nextera Index adapters

3) PEAR (paired end read merger) software aligns matching forward and reverse sequences and outputs a paired fastq file (used for downstream analysis) and a discarded and unassigned forward and reverse read fastq files (ignored)
4) Concatenate paired fastq files to seperate from poor quality sequences

5) Use cutadapt to remove Illumina adapter sequences on ends of reads

6) Custom Primer Sort ScriptFastX toolkit to remove low quality and short reads

7) Convert fastq text file to fasta text file
8) Split fasta file by primer as taxonomy assignment is difficut with mulitple target regions

9) Output is then a cleaned fasta file with sequences from one primer target

Taxonomy Assignment

Cleaned paired sequences are then assigned taxonomy to determine which species were present in the eDNA samples. Sequences are clustered into operational taxonomic units or similar

sequences with a given identification. These sequences are then collapsed to a representative sequence and assigned a taxonomy from a reference library.

1) Swarm OTU clustering is used to collapse identical sequences into representative operational taxonomix units

2) Representative set of OTUs are generated with counts of each OTU class

3) Taxonomy Assignment of Reference Sequences

4) Taxonomy assignment of target sequences using BLAST with MEGAN and reference library taxonomy file

5) Generate BIOM Table using Qiime metabarcoding software program

6) Qiime script: filter out unassigned reads and rename samples with sample names

7) Python Decontaminate Script to remove sequences below PCR blank and field extraction blank thresholds

Data Analysis

Cleaned metabarcoding data is then used for data analysis. Current analysis involves comparing different taxa through plots of relative abundance, comparing species alpha diversity at each site as well as comparing species betadiversity between each site.

1) Summarize Taxa Through Plots to show bar graphs of relative abundance of species

2) Alpha Diversity rarefaction curves to show biodiversity across sites and sequence saturation curves

3) Beta Diversity NMDS to compare sites across specified environmental variable

4) PERMANOVA and PERMDISP to calculate stastics on beta diversity analysis

Anacapa Pipeline

Below are the copied contents of the config.sh file and the anacapa_pipeline.v1.sh scripts. Both these scripts work in full form on my personal desktop. However, there were difficulties operating the full pipeline on the virtual machine because of low computing power and an issues with the fastx dependency. Future work on the pipeline will allow it to be transferable to virtual machines. Please visit my git hub at https://github.com/zjgold/eeb234/tree/master/eeb-177/eeb-174-final-project to view the inputs and outputs of the Anacapa pipeline.

Config File

(This config file was formatted to work on Zack Gold's personal desktop not virtual machine) The config file is designed to run the anacapa_pipeline_v1.sh script with user input replacing the location of the dependencies and input files as downloaded from the github.

```
#### Configuration File for Anacapa Pipeline
#### Created March 2017 by Zack Gold for EEB 234
#### Last Modified 03/23/2017

#### Instructions
#### Step 1: Replace all file paths to match those on your computer
#### Step 2: Run Anacapa Pipeline

#### Directory with Scripts and Files needed for Anacapa Pipeline
UTIL_FOLDER_PATH=
"/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/utilites/"

#### Sample Names to Replace Raw Sequencing Names
SAMPLE_NAMES="RWZG001_S12_L002_ RWZG001_S12_L002_ RWZG002_S26_L002_
RWZG002_S26_L002_ RWZG003_S27_L002_ RWZG003_S27_L002_"

#### Directory of Raw Fastq files from Sequencer
FASTQ_DIK_PATH="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/dummy/"

#### Input Directory with user generated/downloaded NCBI Taxids,
#### Primer file,  QIIME Mapping File, and Unassigned Reference
#### Library Database
INPUT_FOLDER=:"/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/input/"

#### For Storing Quality Controlled and PEAR'd fastq
QC_FOLDER="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/qc/"

#### Config for PEAR program
PEAR_PATH="pear "
PEAR_OUT_PATH=
"/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/dummy/PEAR/"
PEAR_Q="30"
PEAR_T="100"
PEAR_J="10"
PEAR_Y="4G"

#### Config for Assembling PEAR Outputs
ASSEMBLED_FASTQ="moorea_assembled.fastq"
DISCARDED_FASTQ="moorea_discarded.fastq"
```

```
FWRD_FASTQ="moorea_unassembled_forward.fastq"
RVRSE_FASTQ="moorea_unassembled_reverse.fastq"


#### Config for Cutadapt program
CUTADAPT_PATH="/Users/zackgold/.local/bin/cutadapt"
FWRD_INDEX_PRIMER="TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGT"
RVRSE_INDEX_PRIMER="GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG"
CUT_FASTQ="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                        qc/moorea_assembled_cut.fastq"

#### Config to convert fastq to fasta file
FASTQ_2_FASTA="/Users/zackgold/.local/bin_2/fastq_to_fasta"
FASTA_ASSEMBLED="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                        qc/moorea_assembled_cut.fasta"

#### Config to print number of assembled reads
READS_NUM="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                        qc/moorea_assembled_cut_reads.txt"

#### Config to Split on Primer (must be done even if only 1 primer)
SPLIT_ON_PRIMER="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                        utilites/Split_on_Primer.py"
PRIMER_LIST="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                        input/PANEL_5_PRIMERS.txt"
SPLIT_M="8"
SPLIT_S="5"

#### Config for Pick OTU in Qiime
PICK_OTU_DIR="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                qiime/otu_98_swarm"
PICK_OTU_S="0.98"
CLUST_TYPE="swarm"

#### Config for Pick Rep Set
PICKED_OTUS="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                qiime/otu_98_swarm/MiFishUFR_no_n_otus.txt"
REP_SET="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                qiime/repset_98.fasta"

#### Config for Make BLAST Database
BLAST_DIR="/Users/zackgold/.local/bin/ncbi-blast-2.6.0+/"
BLAST_DB_INPUT="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                            input/blast_databases/sequence.fasta-4.txt"
```

```
BLAST_DB_TITLE="MiFish_Universal_12s"
BLAST_DB="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                          input/blast_databases/MiFish_Universal/12S"
TAXID_MAP="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
          input/ncbi_accessions_taxonomy_dump/nucl_gb.accession2taxid"

#### Config for best_taxon_picker.py
MEGAN_TAX="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                        input/mifish_rep_set_localsonly20-taxonomy.txt"
MEGAN_BEST_TAX="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
            input/mifish_rep_set_localsonly20-taxonomy_best_taxonomy.txt"
ASSIGNMENT_THRESHOLD="100"

#### Config for Blast Rep Set
BLASTED_REP_SET="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                            qiime/BLAST_repset_99.txt"

#### Config for Assigning Taxonomy

ASSIGNED_TAXONOMY="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                                qiime/assigned_taxonomy/"
REF_LIB="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
      input/blast_databases/MiFish_Universal/12S/Mifish_40k_no_n_nr.fasta"

#### Config for Splitting OTU Table based on project
OTU_TABLE_RAW="/Users/zackgold/Documents/UCLA_phd/Projects/a
                nacapa/qiime/assigned_taxonomy/mifish_megan_otu_table.biom"
MAP_TABLE="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
                                                input/moorea_1_seq_map.txt"
CATEGORY_TO_SPLIT="Project"
SPLIT_DIR="/Users/zackgold/Documents/UCLA_phd/Projects/
                                      anacapa/qiime/per_project_otu_tables"

#### Config for Collapse Samples
PROJECT_1_BIOM="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
    qiime/per_project_otu_tables/otu_table_mc2_w_tax__Project_Moorea__.biom"
PROJECT_1_BIOM_RENAMED="/Users/zackgold/Documents/UCLA_phd/Projects/anacapa/
qiime/per_project_otu_tables/otu_table_mc2_w_tax__Project_Moorea__renamed.biom"
SAMPLE_NAME="Sample_name"

#### Config for Filter out Unassigned Reads
PROJECT_1_BIOM_UN_REMOVED="/Users/zackgold/Documents/UCLA_phd/Projects/
    anacapa/qiime/per_project_otu_tables/
    otu_table_mc2_w_tax__Project_Moorea__renamed_unassigned_removed.biom"
```

```
UNASSIGNED="/Users/zackgold/Documents/UCLA_phd/Projects/
                    anacapa/qiime/per_project_otu_tables/unassigned.txt"

#### Config for decontaminate.py
PROJECT_1_BIOM_UN_REMOVED_TXT="/Users/zackgold/Documents/UCLA_phd/
      Projects/anacapa/qiime/per_project_otu_tables/
      otu_table_mc2_w_tax__Project_Moorea_r_unr.txt"
OTU_TABLE_TXT="'otu_table_mc2_w_tax__Project_Moorea_r_unr.txt'"
LABEL_OTU="'#OTU ID'"
LABEL_TAX="'taxonomy'"
CONTROL_1="'pcrblank'"
CONTROL_2="'negcontrol1'"
CONTROLS_ALL="config['CONTROL_1'],config['CONTROL_2']"
#### just add more controls if you have more blanks and make sure
####     to include them all in controlls_all
CLEANED_OTU="'cleaned_otu_table.csv'"
CLEANED_OTU_TSV="cleaned_otu_table.csv"
CLEANED_OTU_BIOM="/Users/paulbarber/Desktop/zjg_folder/201609_ZG/gz/
    pick_otu/moorea_mifish/per_project_otu_tables/
    otu_table_mc2_w_tax__Project_Moorea_r_unr_controls_removed.biom"
CONVERTED_BIOM="/Users/paulbarber/Desktop/zjg_folder/201609_ZG/gz/
  pick_otu/moorea_mifish/per_project_otu_tables/
  otu_table_mc2_w_tax__Project_Moorea_r_unr_controls_removed_converted.biom"


########## Config for Data Analysis

#### Config for Summarize taxa through plots
TAX_SUMMARY="/Users/zacharygold/Documents/UCLAPhD/Projects/
                             Moorea/Sequences/taxa_summary"

#### Config for Alpha diversity Rarefaction
RARE_DATA="moorare"
REP_SET_TRE="/Users/zacharygold/Documents/UCLAPhD/Projects/Moorea/
                       Dropbox/201609_ZG/MiFish/otu_default/rep_set.tre"

#### Config for Betadiversity Rarefaction
JACK_BDIV_OUT="jack_bdiv_even1000"

#### Config for Beta Diversity Statistics
DIST_TYPE="dist_bray_curtis,abund_jaccard,binary_dist_jaccard"
BETA_STATS_OUT="beta_div/"
BETA_DATA="/Users/zacharygold/Documents/UCLAPhD/Projects/Moorea/
        Sequences/beta_div/
```

```
        weighted_unifrac_otu_table_mc2_w_tax__Project_Moorea__cleaned.txt"
```

anacapa__pipeline__v1.sh

```
#!/bin/bash

#### Anacapa Pipeline v1

#### Metabarcoding bioinformatics pipeline to take raw sequence
#### data and output data analysis
#### Further description found in the README on github
# https://github.com/zjgold/eeb234/tree/master/eeb-177/eeb-174-final-project
#### Requires sister config file that houses paths to dependencies,
#### inputs, and outputs

#### Author: Zachary Gold
#### contact: zack.gold@ucla.edu

#### Activate Configuration File
. /Users/zackgold/Documents/UCLA_phd/Projects/anacapa/config_file.sh

----------------------------------------
# Sequence Cleaning

#### Rename Raw Sequence Reads
#### raw sequence reads have a strange naming convention,
#### replace this with user defined sample names for each sample.
for i in ${SAMPLE_NAMES}  ;
do
    echo $i
    #grabs forward sequence files and renames them
    sed -i '' "s/K00188/$i/g" ${FASTQ_DIK_PATH}$i"R1_001.fastq"
    #grabs reverse sequence files and renames them
    sed -i '' "s/K00188/$i/g" ${FASTQ_DIK_PATH}$i"R2_001.fastq"
done


#### Paired End reAd meRger
#### runs PEAR program on each pair of matched sequences, PEAR software
#### works to align forward and reverse sequences with given accuracy
#### parameters, and outputs an assigned, discarded, forward unassigned,
#### and reverse unassigned file.
for i in ${SAMPLE_NAMES}  ;
```

```
do
    echo $i
    ${PEAR_PATH} -f ${FASTQ_DIK_PATH}
    $i"R1_001.fastq" -r ${FASTQ_DIK_PATH}$i"R2_001.fastq"
-o ${PEAR_OUT_PATH}$i -q ${PEAR_Q}
-t ${PEAR_T} -j ${PEAR_J} -y ${PEAR_Y}

done

#### make a new directory to store quality controlled reads
mkdir ${QC_FOLDER}

#### concatenates all of the paired merged reads, assembled sequences
#### are the most important and stored in the QC folder
cat ${PEAR_OUT_PATH}*.assembled.fastq > ${QC_FOLDER}${ASSEMBLED_FASTQ}
cat ${PEAR_OUT_PATH}*.discarded.fastq > ${PEAR_OUT_PATH}${DISCARDED_FASTQ}
cat ${PEAR_OUT_PATH}*.unassembled.forward.fastq > ${PEAR_OUT_PATH}${FWRD_FASTQ}
cat ${PEAR_OUT_PATH}*.unassembled.reverse.fastq > ${PEAR_OUT_PATH}${RVRSE_FASTQ}

#### Cut Adapter
#### removes adapter sequences from the ends of the reads and filters
#### out small sequences  (<125 bp)
module load fastx_toolkit/0.0.13.2
${CUTADAPT_PATH} -a ${FWRD_INDEX_PRIMER} -g ${RVRSE_INDEX_PRIMER}
--minimum-length 125 ${QC_FOLDER}${ASSEMBLED_FASTQ} > ${CUT_FASTQ}

#### Fastaq to FASTA
#### Converts Fastq text file to to fasta text file
#### (one text form to another to be used for input into qiime)
${FASTQ_2_FASTA} -i ${CUT_FASTQ} -o ${FASTA_ASSEMBLED} -n -Q33

#### save total number of cleaned assembled reads
grep -c ">" ${FASTA_ASSEMBLED} > ${READS_NUM}


-----------------------------------------------------
#Assign Taxonomy
#### Split on primer
#### separates primers for different downstream taxonomic assignment,
#### each primer is split into its own fasta file
${SPLIT_ON_PRIMER} -f ${FASTA_ASSEMBLED} -p ${PRIMER_LIST}
-m ${SPLIT_M} -s ${SPLIT_S}

#### Pick OTUS
```

```
#### clusters sequences to representative OTUs which serve as identifiers
#### for related sequences set by the s threshold i.e. 98% similarity
macqiime pick_otus.py -i ${FASTA_ASSEMBLED} -o ${PICK_OTU_DIR}
-s ${PICK_OTU_S} -m ${CLUST_TYPE}

#### Picks a representative set of OTUs
#### picks a representative OTU sequence for each OTU and collapses
#### all matching sequences into a sequence count associated with each
#### sequence, i.e. all unassigned Great white shark sequences are
#### collapsed into 32,000 sequences assigned to OTU number 123456
macqiime pick_rep_set.py -i ${PICKED_OTUS} -f ${FASTA_ASSEMBLED}
-o ${REP_SET}

#### makeblast db
#### Make a reference database of specific sequences of interest
#### with assigned taxonomy
#### blast is a program that compares a set of sequences to online
#### published sequence reference database
#### sequence database of interest is generated separately by
#### downloading all vertebrate sequences from the NCBI website
#### makeblastdb assigns a taxonomy (i.e. Great White shark) to the
#### sequence
#### WARNING: Very slow
 ${BLAST_DIR}bin/makeblastdb -in ${BLAST_DB_INPUT} -dbtype nucl
 -title ${BLAST_DB_TITLE} -out ${BLAST_DB} -taxid_map ${TAXID_MAP}
 -parse_seqids

#### next step could not be automated as of 3/23/2017 due to technical
#### difficulties
#### open database in the MEGAN program and run least common ancestor
#### taxonomy assignment analysis to generate reference library with
#### assigned confidence values to each level of assigned taxonomy

#### remove low confidence assignments from Megan taxonomy
#### current threshold set at 100%
python ${UTIL_FOLDER_PATH}best_taxon_picker.py ${MEGAN_TAX}
${MEGAN_BEST_TAX} ${ASSIGNMENT_THRESHOLD}

#### now that there is a cleaned reference library with good
#### taxonomic assignments

#### blast repset
#### blasts the representative set of OTU sequences against the
####reference library to assign accurate taxonomy to each
```

```
#### OTU representative

#### generates a BIOM table which is essentially like a tsv file
#### with rows as OTU IDs and columns as individual samples with
#### number of sequences of each OTU
#### WARNING: Very slow
#### taxonomy at this step is ignored, MEGAN curated taxonomy is more accurate
${BLAST_DIR}bin/blastn -query ${REP_SET} -db ${BLAST_DB}
-evalue 0.0000001 -outfmt 0 -num_alignments 20 -out ${BLASTED_REP_SET}

#### assign taxonomy to biom table
#### appends curated MEGAN taxonomy to the biom table using blast
#### to match sequences
assign_taxonomy.py -i ${REP_SET} -r ${REF_LIB}  -t ${MEGAN_BEST_TAX}
-m blast -e 0.000001 -o ${ASSIGNED_TAXONOMY}

#### Split OTU table
#### Splits samples based on project using a user generated Map file
split_otu_table.py -i ${OTU_TABLE_RAW} -m ${MAP_TABLE}
-f ${CATEGORY_TO_SPLIT} -o ${SPLIT_DIR}


#### Collapse Samples
#### Renames Raw Sample Names to User defined site names based on
#### user generated map file
collapse_samples.py -b ${PROJECT_1_BIOM} -m ${MAP_TABLE}
--output_biom_fp ${PROJECT_1_BIOM_RENAMED} --output_mapping_fp omitted
--collapse_mode sum --collapse_fields ${SAMPLE_NAME}

#### Remove Unassigned Reads
#### Keeps OTUs with assigned taxonomy
filter_otus_from_otu_table.py -i ${PROJECT_1_BIOM_RENAMED}
-o ${PROJECT_1_BIOM_UN_REMOVED} -e ${UNASSIGNED}

#### Biom Convert w/ Taxonomy Labels (must make a new file, will not
#### replace a file correctly)
#### convert .biom to .tsv for python script
biom convert -i  ${PROJECT_1_BIOM_UN_REMOVED}
-o ${PROJECT_1_BIOM_UN_REMOVED_TXT}
--to-tsv --header-key taxonomy

#### Remove Contamination from Sequences
#### python script (works) to remove contaminated sequences
python ${UTIL_FOLDER_PATH}deconaminate.py
```

```
#### convert back to .BIOM
biom convert -i ${CLEANED_OTU_TSV}  -o ${CLEANED_OTU_BIOM}
--table-type="OTU table" --to- --process-obs-metadata taxonomy

#### convert to HDF5 type for QIIME
biom convert -i ${CLEANED_OTU_BIOM} --to-hdf5 --collapsed-samples
-o ${CONVERTED_BIOM}

-----------------------------------------
```

decontaminate.py

```
import sys
sys.path.append('/home/eeb177-student/miniconda3/lib/python3.5/site-packages/')
import numpy
import pandas as pd

alldata=pd.read_csv('otu_table_mc2_w_tax__Project_Moorea__renamed_text.txt',
    sep='\t', header=1)
labels=alldata[['#OTU ID','taxonomy']]
controls=alldata[['pcrblank','negcontrol1']]
data=alldata.drop(['#OTU ID','taxonomy','pcrblank','negcontrol1'],axis=1)
def decontaminate_column(data_column,control_column):
    column = data_column.copy()
    column.ix[column <= control_column] = 0
    return column

def decontaminate(dataframe,control):
    for control_name in control.columns.values.tolist():
        dataframe=dataframe.apply(lambda c: decontaminate_column(c,
        control[control_name]), axis=0)
    return dataframe
data=decontaminate(data,controls)
cleaned_otu_table = pd.concat([labels['#OTU ID'],
controls,data,labels['taxonomy']],axis=1)
with open('otu_table_mc2_w_tax__Project_Moorea__renamed_text.txt', 'r') as f:
    comment=f.readline()
with open('cleaned_otu_table.csv', 'w') as f:
    f.write(comment)
cleaned_otu_table.to_csv('cleaned_otu_table.csv', sep='\t',
    mode='a',index=False)
```

```
infile_name = "mifish_rep_set_localsonly20-taxonomy.txt"
outfile_name = "mifish_rep_set_localsonly20-taxonomy_best_taxonomy.txt"
```

```
threshold = float(100)

def best_taxon_picker(line,threshold):
    output = []
    splitted = line.split(';')
    for i in range(0,len(splitted)-1,2):
        taxon = splitted[i]
        match_level = splitted[i+1]
        if float(match_level.strip()) >= threshold:
            output.append(taxon+';'+match_level)
    otu_taxon = ";".join(output) + ";"
    return otu_taxon

with open(infile_name) as infile, open(outfile_name, 'w') as outfile:
    for line in infile:
        outfile.write(best_taxon_picker(line,threshold)+'\n')
```

Data Analysis

In the summer of 2016, we collected 42 eDNA samples from Mo'orea, French Polynesia. Mo'orea is a volcanic island surrounded by fringing reefs and home to a moderate diversity of coral reef fishes (Leray *et al.* 2012). This was an ideal location to test for the power of eDNA techniques as the Smithsonian BIOCODE project previously sequenced all known marine and terrestrial species around the island(Leray *et al.* 2012). In addition, we wanted to test for the spatial variability of eDNA signatures and whether we could accurately distinguish different coral reef habitats. TO do this, we collected samples from four tropical marine habitats (back reef, fore reef, pelagic, and high flow back reef). Eight of the 13 sites were along a transect moving away from shore toward the open ocean as seen in the image below. Physical oceanography data previously collected in Mo'orea indicates a strong north south movement of water which we expected to have a strong effect on the signature of fish communities obtained from eDNA data(Hench *et al.* 2008). To collect eDNA sequences, we used standard eDNA methods validated with both controlled aquaria studies and systematic field surveys to collect these water samples. I collected three 3 1L samples of seawater water in sterile collapsible bottles from each site using SCUBA. Pelagic and back reef samples were collected at 2m depth while fore reef samples were collected at 0m,10m,and 30m at the same site. We filtered the water samples in the field using 0.22 µm sterivex filters and a gravity filtration. Sterivex filters with eDNA were then extracted using the DNAeasy Qiagen Kit and then amplify the fish DNA using fish specific MiFish Universal Teloest primers with polymerase chain reaction (PCR) (Miya *et al.* 2015). Each eDNA sample was subsampled for three PCR reactions to quantify and control for potential false positives. We then pooled the PCR products for sequencing on an Illumina HiSeq. These raw sequences were then fed into the Anacapa pipeline for metabarcoding analysis.

These results are from a failed sequencing run conducted in October of 2016 and thus do not

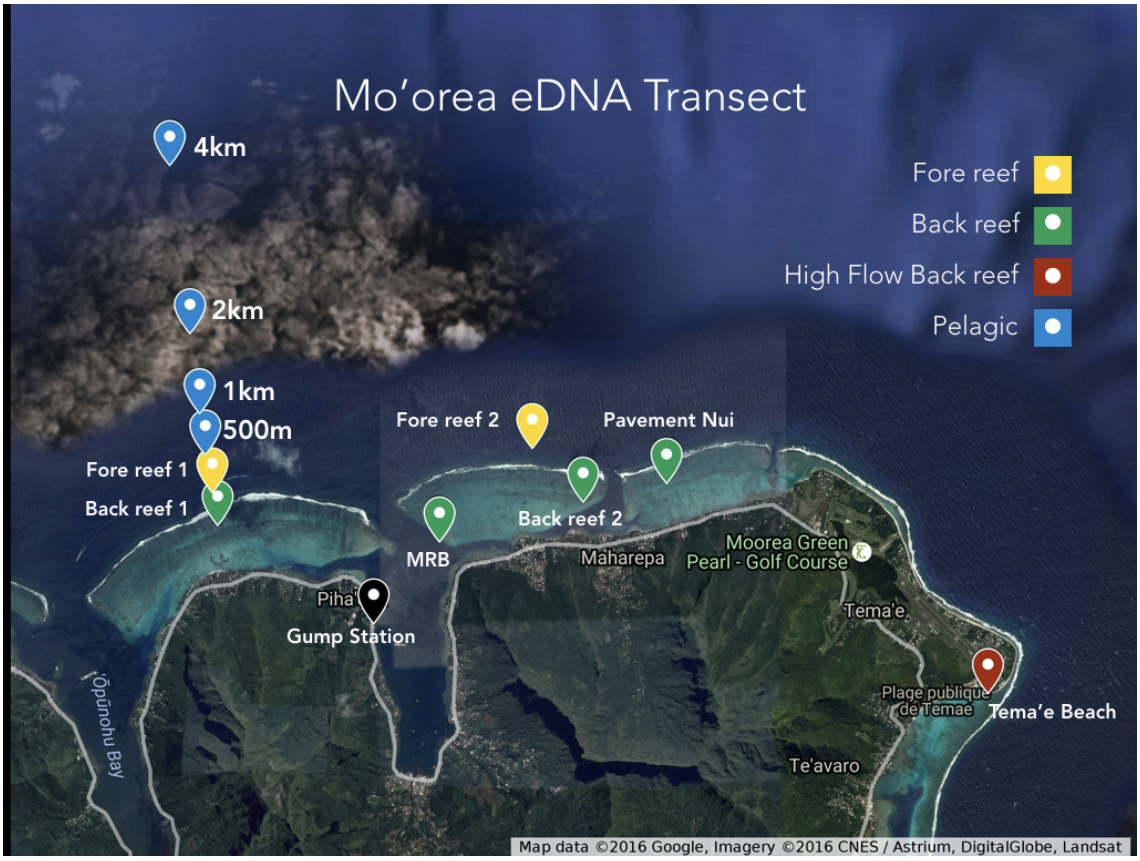


Figure 1: Moorea Sampling Design

reflect publishable results.

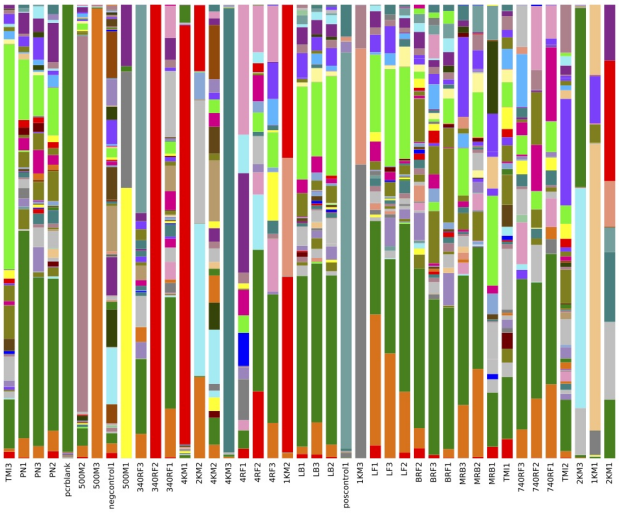


Figure 2: Taxa Summary Bar Plot: Demonstrating sites and relative abundance of different species found at each site.

```
#### Summarize Taxa through plots
#### generates a series of bar graphs at different taxonomic levels
#### showing relative species abundance at each site
summarize_taxa_through_plots.py -i ${CONVERTED_BIOM} -f -o ${TAX_SUMMARY}
-m ${MAP_TABLE} -p ${UTIL_FOLDER_PATH}species_parameter.txt
```

Alpha Rarefaction

```
#### Alpha Diversity Rarefaction
####  runs an alpha rarefaction curve analysis to show 1) if sequencing
#### depth was adequate i.e. the curve saturates at a given sequence depth
#### and 2) compare number of species found in each sample
alpha_rarefaction.py -i ${CONVERTED_BIOM} -m ${MAP_TABLE} -o ${RARE_DATA}
-p ${UTIL_FOLDER_PATH}alpha_params.txt -f -- min_rare_depth 1000
```

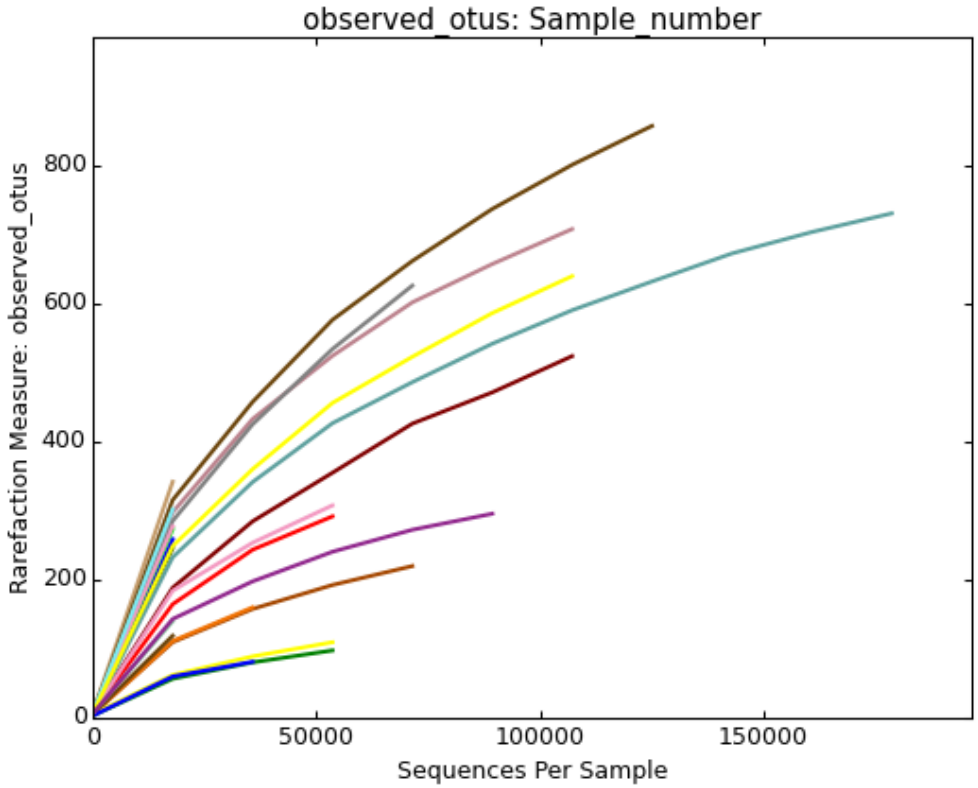


Figure 3: Alpha Rarefaction Plot: Number of OTUs found per number of sequences. Tailing off of the curve indicates saturation of sequencing depth. Height of tail indicates number of OTUs per sample. This example data set did not have high enough sequencing depth to reach saturation for the majority of samples.

Beta Diversity

Beta diversity was compared between sites to determine if eDNA fish communities were different between habitat type. Previous visual transect data has indicated that fish communities are varied between these habitat types (Leray *et al.* 2012). To compare beta diversity between sites I calculated bray curtis similarity distance matrices based on the abundance of reads of each species found at each site. I then ran a PERMANOVA on the distances to compare the similarity of habitat types. I then ran PERMADISP variance analyses to identify which of the habitats, if any, are signficantly different from eachother.

```
#### Beta Diversity Plot
#### generates beta-diversity plots comparing species distributions
#### between samples
#### Additional R generated plots in separate script
jackknifed_beta_diversity.py -i ${CONVERTED_BIOM} -m ${MAP_TABLE}
-o JACK_BDIV_OUT -e 1000


#### Stats analysis using permanova and permadisp
beta_diversity.py -i ${CONVERTED_BIOM} -m ${DIST_TYPE} -o ${BETA_STATS_OUT}

compare_categories.py --method permanova -i ${BETA_DATA}
-m ${MAP_TABLE} -c Habitat -o ${BETA_STATS_OUT} -n 999

compare_categories.py --method permdisp -i ${BETA_DATA}
-m ${MAP_TABLE} -c Habitat -o ${BETA_STATS_OUT} -n 999
```

PERMANOVA

method name PERMANOVA test statistic name pseudo-F sample size 42 number of groups 5 test statistic 3.251341471583491 p-value **0.001** number of permutations 999

Results of the PERMANOVA analysis indicate that fish communities are signficantly different across habitat types in Mo'orea.

PERMDISP

Analysis of Variance Table

Response: Distances Df Sum Sq Mean Sq F value Pr(>F)
Groups 4 0.36783 0.091956 3.5749 0.01459 * Residuals 37 0.95174 0.025723
— Signif. codes: 0 ‘’ ***0.001*** ‘’ *0.01* ‘’ 0.05 ‘.’ 0.1 ‘’ 1

Permutation test for homogeneity of multivariate dispersions Permutation: free Number of permutations: 999

Response: Distances Df Sum Sq Mean Sq F N.Perm Pr(>F)
Groups 4 0.36783 0.091956 3.5749 999 0.012 * Residuals 37 0.95174 0.025723
— Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Pairwise comparisons: (Observed p-value below diagonal, permuted p-value above diagonal) Backreef Backreef_flow Forereef Lab Pelagic Backreef 0.4620 **0.003** 0.8330 **0.001** Backreef_flow 0.470288 0.315000 0.827000 0.139 Forereef **0.0123** 0.308760 0.199000 0.574 Lab 0.838011 0.789045 0.225934 0.101 Pelagic **0.0012** 0.145567 0.540436 0.102084

The PERMDISP analysis indicate that the backreef fish communities are signficantly different from both forereef and pelagic fish communities. This suggests that eDNA has the power to distinguish different fish communities despite the movement of offshore fish eDNA. However, eDNA was unable to capture this difference at back reef sites exposed to high current velocities suggesting that physical oceanography is playing a role in the fate and transport of eDNA particles. These results suggest that current patterns and water movement have the potential to be important factors affecting the eDNA signatures captured from an environment. This has important ramifications for marine eDNA sampling design as samples inside or outside of a marine protected area could be receiving eDNA inoput from beyong the boundaries of the reserves. This could signficantly bias eDNA results and out understanding of the effects of reserves on fish abundance and biodiversity.

However, these results are from a sequencing run with severe pooling issues and under sequencing of samples. New sequencing results just arrived on March 24, 2017. Future analysis buidling off those conducted here will be used to identify if this same pattern holds with more accurate data.

Data Visualization

Here I visualized the beta diversity analyses using a heat map and Non-metric multidimensional scaling. I conducted two parallel analyses using the Bray-Curtis and Jaccard similarity matrices. Bray Curtis similarity matrices place a greater emphasis on differences in abundances of species while the Jaccard matrice place a greater emphasis on presence/absence of species. Both of these resuls act to provide complimentary views of the fish communities observed from eDNA analyses.

```
# R script to implement Qiime in R for data analysis and plotting

library(qiimer)
library(reshape2)
library(ggplot2)
library(ggrepel)
library(vegan)
```

```
## Loading required package: permute
```

```
## Loading required package: lattice
```

```
## This is vegan 2.4-2
```

```
library(wesanderson)
# I just want to give a shout out to karthik for being a true legend
# and fellow Wes Anderson afficiando, as well as to Gaurav for sharing
# with me this amazing package. Truly life altering package.

library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
# Read in cleaned biom table
biom_table <- read_qiime_otu_table("/home/eeb177-student/Desktop/eeb234/eeb-177/eeb-174-
#biom_table

#Read in mapping file
mor_map <- read_qiime_mapping_file("/home/eeb177-student/Desktop/eeb234/eeb-177/eeb-174-
head(mor_map)
```

```
##   SampleID Sample_name Lab_id Sample_number Project       Habitat Depth_m
## 1     TMI3     RWZG032  MO_38            38  Moorea Backreef_flow       1
## 2      PN1     RWZG003  MO_11            11  Moorea      Backreef       2
## 3      PN3     RWZG031  MO_37            37  Moorea      Backreef       2
## 4      PN2     RWZG017  MO_24            24  Moorea      Backreef       2
## 5 pcrblank     RWZG050  MO_55            55  Moorea           Lab       0
## 6    500M2     RWZG009  MO_17            17  Moorea       Pelagic       2
##        Transect
## 1 Backreef_flow
## 2      Backreef
## 3      Backreef
## 4      Backreef
## 5           Lab
## 6           500
```

```
# Create data frame from biom_table
data <- biom_table$counts
# Transpose data frame to R usable format
dataT <- t(data)
dataT <- as.data.frame(dataT)

# Map metadata to mapper data frame
mapper <- data.frame(mor_map$Habitat, mor_map$SampleID)
#append metadata to dataT
colnames(dataT)[0] <- "mor_map.SampleID"
dataT_lab <- dataT
dataT_lab$mor_map.SampleID <- biom_table$sample_ids
dataT_lab <- left_join(dataT_lab, mapper, by = "mor_map.SampleID")
```

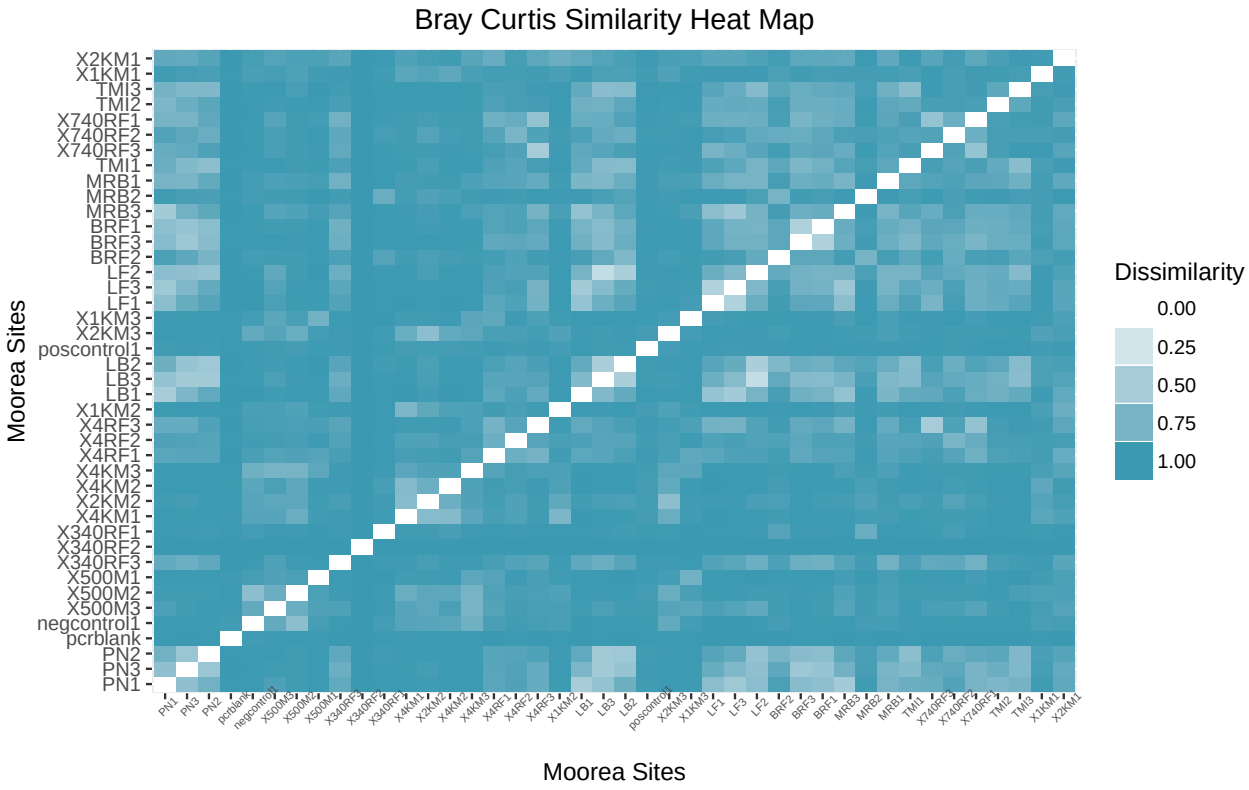
```
## Warning in left_join_impl(x, y, by$x, by$y, suffix$x, suffix$y): joining
## factor and character vector, coercing into character vector
```

```
# Calculate Bray Curtis Similarity Distance
bray_dist_dataT<-vegdist(dataT, method = "bray")
head(bray_dist_dataT)
```

```
## [1] 0.6119912 0.7710098 0.9977847 0.9977248 0.9352140 0.9810754
```

```
# Plot Bray Curtis Distance Map
melted_bray <- melt(as.matrix(bray_dist_dataT))
pal <- wes_palette("Zissou", 100, type = "continuous")
# "Can I call you Stevesies?""
ggplot(data = melted_bray, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() + labs(x= 'Moorea Sites', y= 'Moorea Sites') +
  scale_fill_gradient(low="white", high= pal[1], guide=guide_legend(title="Dissimilarit
  ggtitle('Bray Curtis Similarity Heat Map') +
  theme(plot.title = element_text(hjust = 0.5))+
```

```
  theme(axis.text.x = element_text(size=5, angle=45))
```



This is a heat map showing the Bray Curtis similarity between samples. Dark blue colors indicate dissimilar fish communities while lighter colors indicate more similar fish communities.

```
#  Bray Curtis NMDS Plot
## Create Distance Data Frame
gr.habitat <- dataT_lab$mor_map.Habitat
col.gr <- wes_palette(name = "Darjeeling2")
grp <- dataT_lab$mor_map.Habitat
test<-cmdscale(bray_dist_dataT)

data.scores_1 <- as.data.frame(scores(test))
#Using the scores function from vegan to extract the site scores and
# convert to a data.frame
data.scores_1$site <- rownames(data.scores_1)
# create a column of site names, from the rownames of data.scores
head(data.scores_1)
```

```
##                   Dim1        Dim2        site
## PN1         -0.3659208 -0.15403640         PN1
## PN3         -0.3723564  0.10402733         PN3
## PN2         -0.2894101  0.31158785         PN2
## pcrblank     0.1728298  0.02240680    pcrblank
## negcontrol1  0.3544329  0.01866297 negcontrol1
## X500M3       0.2557008 -0.09711367      X500M3
```

```
data.scores_1$grp <- grp  #  add the grp variable created earlier
head(data.scores_1)  #look at the data
```

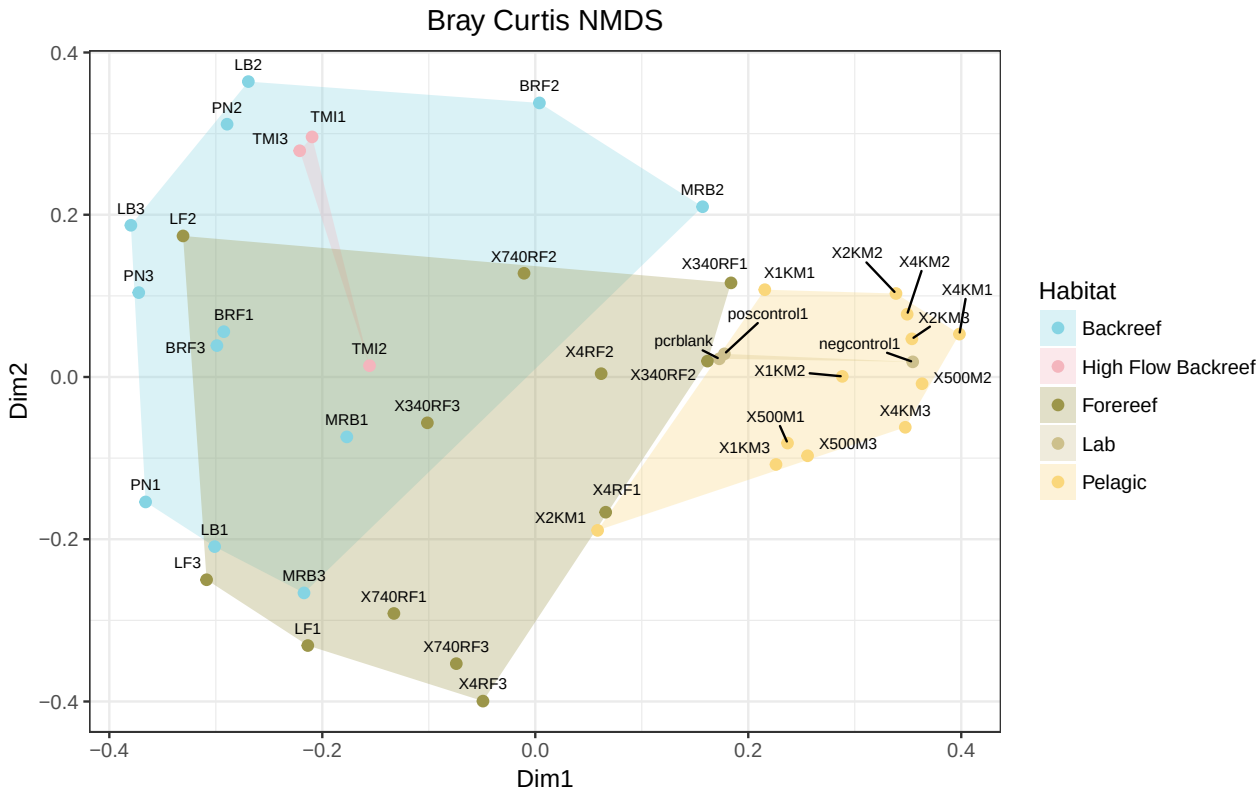
```
##                   Dim1        Dim2        site      grp
## PN1         -0.3659208 -0.15403640         PN1 Backreef
## PN3         -0.3723564  0.10402733         PN3 Backreef
## PN2         -0.2894101  0.31158785         PN2 Backreef
## pcrblank     0.1728298  0.02240680    pcrblank      Lab
## negcontrol1  0.3544329  0.01866297 negcontrol1      Lab
## X500M3       0.2557008 -0.09711367      X500M3  Pelagic
```

```
#### Calculate Shape Around Points
grp.backreef <- data.scores_1[data.scores_1$grp == "Backreef", ][chull(data.scores_1[dat
grp.fourreef <- data.scores_1[data.scores_1$grp == "Forereef", ][chull(data.scores_1[dat
grp.lab <- data.scores_1[data.scores_1$grp == "Lab", ][chull(data.scores_1[data.scores_1
grp.pelagic <- data.scores_1[data.scores_1$grp == "Pelagic", ][chull(data.scores_1[data.
grp.backflow <- data.scores_1[data.scores_1$grp == "Backreef_flow", ][chull(data.scores_

hull.data <- rbind(grp.backreef, grp.fourreef,grp.lab,grp.pelagic,grp.backflow)
```

```
#### Plot Bray Curtis NMDS
col.gr <- wes_palette(name = "Moonrise3")
# Palette Inspired by Ed Norton in short shorts
ggplot() +
```

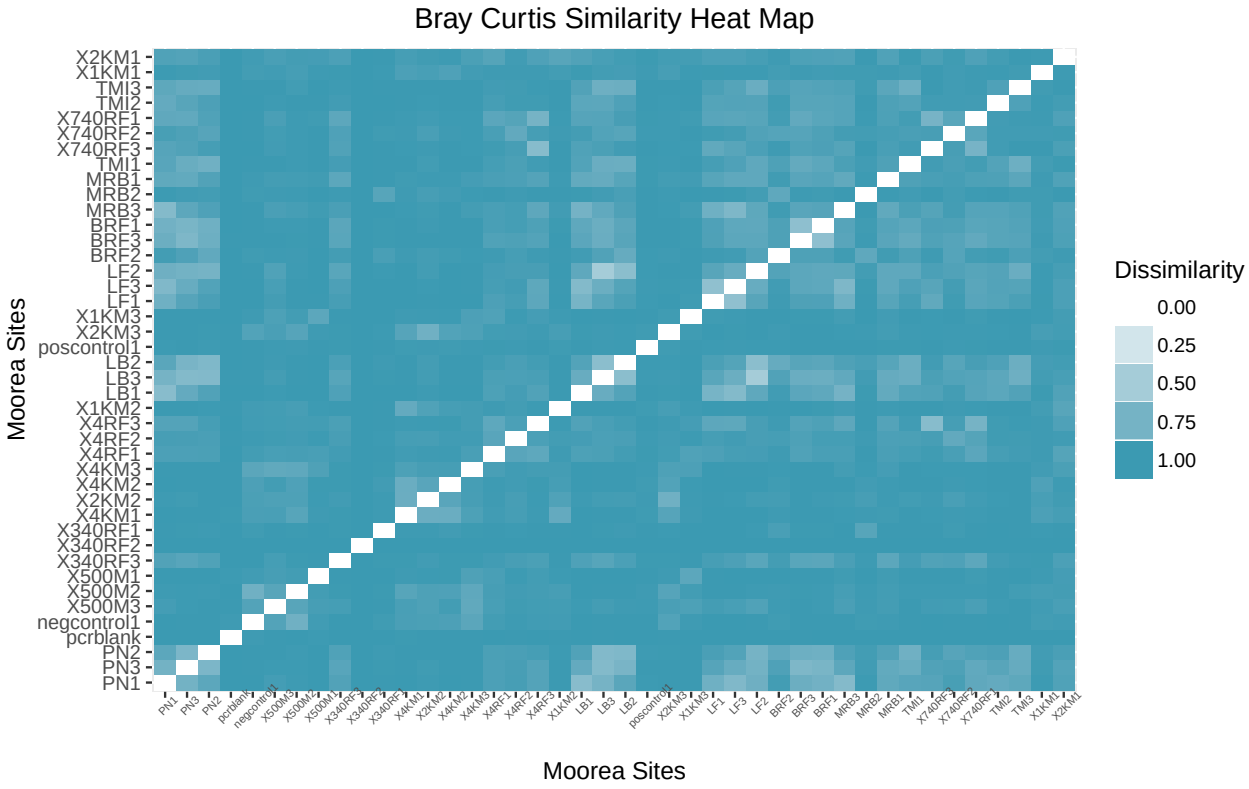
```
  geom_polygon(data=hull.data,aes(x=Dim1,y=Dim2,fill=grp,group=grp),alpha=0.30) + scale_
  geom_point(data=data.scores_1,aes(x=Dim1,y=Dim2, colour = grp ),size=2) + scale_colour
  geom_text_repel(data=data.scores_1,aes(x=Dim1,y=Dim2,label=site),size=2.5, nudge_y = 0
  theme_bw() + ggtitle('Bray Curtis NMDS') + theme(plot.title = element_text(hjust = 0.
```



The NMDS figure plots plots uses the Bray Curtis similarity distance matrix to plot points along the coordinate system. Points closer together indicate that they share more similar fish communities. The distribution of the points suggests that habitat type is having a moderate effect on the eDNA signature of fish communities.

```
#Jaccard Similarity
# Calculate Jaccard Similarity Distance
jaccard_dist_dataT<-vegdist(dataT, method = "jaccard")

# Plot Jaccard Distance Map
melted_jaccard <- melt(as.matrix(jaccard_dist_dataT))
pal <- wes_palette("Zissou", 100, type = "continuous")
ggplot(data = melted_jaccard, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  labs(x= 'Moorea Sites', y= 'Moorea Sites') +
  scale_fill_gradient(low="white", high= pal[1],
                      guide=guide_legend(title="Dissimilarity")) +
  ggtitle('Bray Curtis Similarity Heat Map') +
  theme(plot.title = element_text(hjust = 0.5))+
  theme(axis.text.x = element_text(size=5, angle=45))
```



This is a heat map showing the Jaccard similarity between samples. Dark blue colors indicate dissimilar fish communities whereas lighter colors indicate more similar fish communities.

```
#  Jaccard NMDS Plot
## Create Distance Data Frame
gr.habitat <- dataT_lab$mor_map.Habitat
col.gr <- wes_palette(name = "Darjeeling2")
grp <- dataT_lab$mor_map.Habitat
jaccard<-cmdscale(jaccard_dist_dataT)

data.scores <- as.data.frame(scores(jaccard))
#Using the scores function from vegan to extract the site scores
#and convert to a data.frame
data.scores$site <- rownames(data.scores)  # create a column of site names,
#from the rownames of data.scores
head(data.scores)
```

```
##                   Dim1        Dim2        site
## PN1         -0.3239400 -0.17886337         PN1
## PN3         -0.3347970  0.11401508         PN3
## PN2         -0.2596373  0.29968573         PN2
## pcrblank     0.1478859  0.01635679    pcrblank
## negcontrol1  0.2835542  0.03004619 negcontrol1
## X500M3       0.2099937 -0.04703458      X500M3
```

```
data.scores$grp <- grp  #  add the grp variable created earlier
head(data.scores)  #look at the data
```

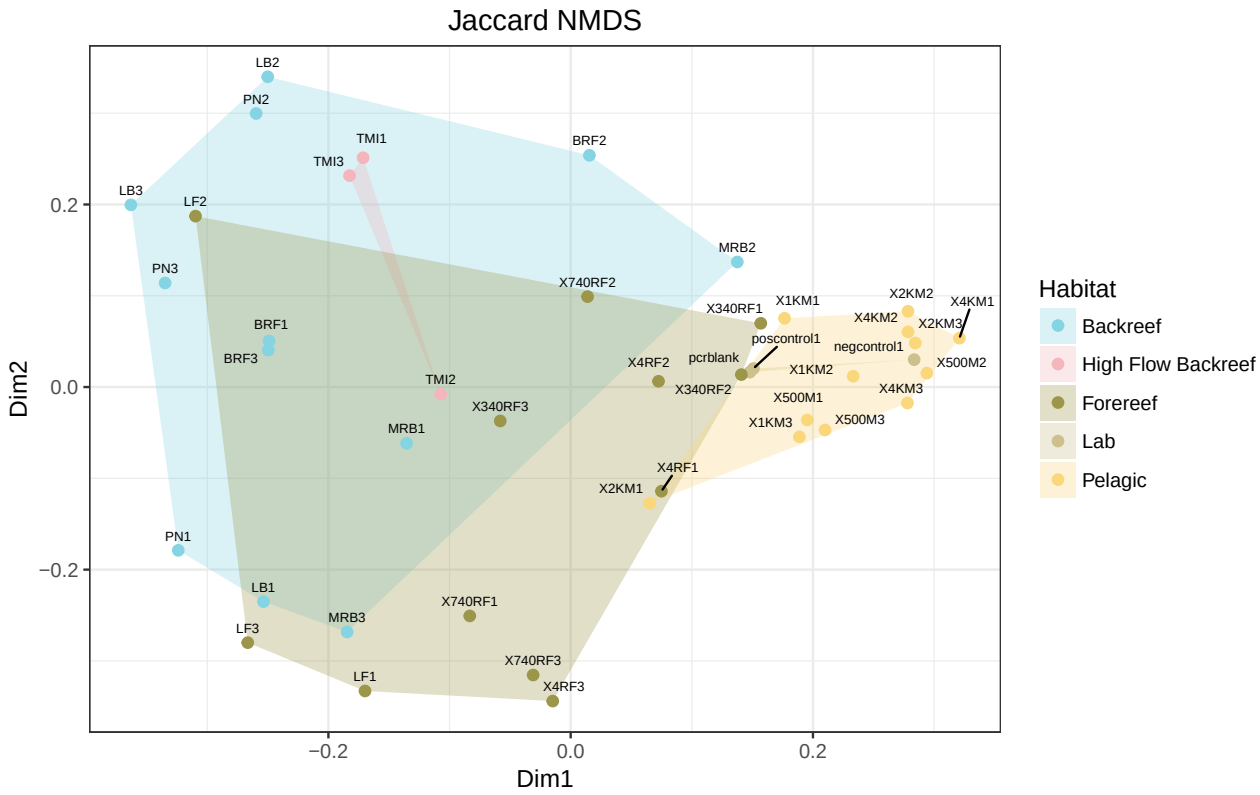
```
##                   Dim1        Dim2        site      grp
## PN1         -0.3239400 -0.17886337         PN1 Backreef
## PN3         -0.3347970  0.11401508         PN3 Backreef
## PN2         -0.2596373  0.29968573         PN2 Backreef
## pcrblank     0.1478859  0.01635679    pcrblank      Lab
## negcontrol1  0.2835542  0.03004619 negcontrol1      Lab
## X500M3       0.2099937 -0.04703458      X500M3  Pelagic
```

```
#### Calculate Shape Around Points
grp.backreef_j <- data.scores[data.scores$grp == "Backreef", ][chull(data.scores[data.sc
grp.fourreef_j <- data.scores[data.scores$grp == "Forereef", ][chull(data.scores[data.sc
grp.lab_j <- data.scores[data.scores$grp == "Lab", ][chull(data.scores[data.scores$grp =
grp.pelagic_j <- data.scores[data.scores$grp == "Pelagic", ][chull(data.scores[data.scor
grp.backflow_j <- data.scores[data.scores$grp == "Backreef_flow", ][chull(data.scores[da

hull.data_j <- rbind(grp.backreef_j, grp.fourreef_j,grp.lab_j,grp.pelagic_j,grp.backflow
```

```
# Plot Jaccard NMDS Plot
col.gr <- wes_palette(name = "Moonrise3")
ggplot() +
  geom_polygon(data=hull.data_j,aes(x=Dim1,y=Dim2,fill=grp,group=grp),
```

```
              alpha=0.30) +
  scale_fill_manual(values=col.gr, guide=guide_legend(title = "Habitat"),
                    labels=c("Backreef", "High Flow Backreef", "Forereef", "Lab", "Pelag
  geom_point(data=data.scores,aes(x=Dim1,y=Dim2, colour = grp ),size=2) +
  scale_colour_manual(values=col.gr, guide=guide_legend(title = "Habitat"),
                      labels=c("Backreef", "High Flow Backreef", "Forereef","Lab","Pelag
  geom_text_repel(data=data.scores,aes(x=Dim1,y=Dim2,label=site),size=2,
                  nudge_y = 0.01,box.padding = unit(0.2, "lines"),
                  point.padding = unit(.1, "lines")) +
  # add the site labels
  theme_bw() + ggtitle('Jaccard NMDS') +
  theme(plot.title = element_text(hjust = 0.5))
```



The NMDS figure plots plots uses the Jaccard similarity distance matrix to plot points along the coordinate system. Points closer together indicate that they share more similar fish communities. The distribution of the points suggests that habitat type is having a moderate effect on the eDNA signature of fish communities.

Conclusion

This project provides a bioinformatic package to take raw sequence data, clean sequences, assign taxonomy, and then analyze metabarcoding data to compare communites across

categoical variables. This prototype pipeline is both flexible and dynamic and allows for the processing and analysis of a wide range of metabarcoding data including eDNA. As I continue to conduct eDNA research during my PhD beyond this class, I hope to improve upon this Anacapa Pipeline and add additional and more robust data analysis features as well as improve the speed and usability of this pipeline.

Ultimately, eDNA has the potential to transform the way that marine ecosystems are monitored, giving resource managers the ability to survey and monitor biodiversity, in a simple, cost effective manner. Creating simple easy to use bioinformatics pipelines for eDNA analysis are crucial for adoption as an effective monitoring technique for marine ecosystems. This anacapa pipeline provides a first step in creating a simple to use eDNA bioinformatics pipeline.

References

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Peña, A.G., Goodrich, J.K., Gordon, J.I. & others. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature methods*, **7**, 335–336.

Dejean, T., Valentini, A., Miquel, C., Taberlet, P., Bellemain, E. & Miaud, C. (2012). Improved detection of an alien invasive species through environmental dna barcoding: The example of the american bullfrog lithobates catesbeianus. *Journal of applied ecology*, **49**, 953–959.

Gordon, A. & Hannon, G. (2010). Fastx-toolkit. *FASTQ/A short-reads preprocessing tools (unpublished) http://hannonlab. cshl. edu/fastx_toolkit*.

Hench, J.L., Leichter, J.J. & Monismith, S.G. (2008). Episodic circulation and exchange in a wave-driven coral reef and lagoon system. *Limnology and Oceanography*, **53**, 2681–2694.

Leray, M., Boehm, J., Mills, S.C. & Meyer, C. (2012). Moorea biocode barcode library as a tool for understanding predator–prey interactions: Insights into the diet of common predatory coral reef fishes. *Coral reefs*, **31**, 383–388.

Madden, T. (2013). The blast sequence analysis tool.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**, pp–10.

Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H. & others. (2015). MiFish, a set of universal pcr primers for metabarcoding environmental dna from fishes: Detection of more than 230 subtropical marine species. *Royal Society open science*, **2**, 150088.

Port, J.A., O'Donnell, J.L., Romero-Maraccini, O.C., Leary, P.R., Litvin, S.Y., Nickols, K.J., Yamahara, K.M. & Kelly, R.P. (2016). Assessing vertebrate biodiversity in a kelp forest ecosystem using environmental dna. *Molecular ecology*, **25**, 527–541.

Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012). Environmental dna. *Molecular ecology*, **21**, 1789–1793.

Thomsen, P.F., Kielgast, J., Iversen, L.L., Møller, P.R., Rasmussen, M. & Willerslev, E. (2012). Detection of a diverse marine fish fauna using environmental dna from seawater samples. *PLoS one*, **7**, e41732.

Valentini, A., Taberlet, P., Miaud, C., Civade, R., Herder, J., Thomsen, P.F., Bellemain, E., Besnard, A., Coissac, E., Boyer, F. & others. (2016). Next-generation monitoring of aquatic biodiversity using environmental dna metabarcoding. *Molecular Ecology*.

Willis, T.J. (2001). Visual census methods underestimate density and diversity of cryptic reef fishes. *Journal of Fish Biology*, **59**, 1408–1411.

Zhang, J., Kobert, K., Flouri, T. & Stamatakis, A. (2014). PEAR: A fast and accurate illumina paired-end reAd mergeR. *Bioinformatics*, **30**, 614–620.