"Navigating the Genetic Sea: A Bioinformatic Analysis of eDNA for Fish Species Identification"

Flora Coden
Master of Advanced Studies – Marine Biodiversity and Conservation
Scripps Institution of Oceanography, University of California San Diego
Capstone Research Project – June 14, 2024

Capstone Advisory Committee

Dr. Nastassia Patin, Scripps Institution of Oceanography, University of California San Diego

Dr. Ronald Burton, Scripps Institution of Oceanography, University of California San Diego

Dr. Jeff Bowman, Scripps Institution of Oceanography, University of California San Diego

CAC Chair Signature: [redacted] 6/13/2024
Nastassia Patin (Date)

CAC Member Signature: [redacted] 6/11/2024
Ronald Burton (Date)

CAC Member Signature: [redacted] 6/12/2024
Jeff Bowman (Date

Abstract

Environmental DNA (eDNA) has become an important tool for marine biodiversity monitoring, offering a non-invasive alternative to traditional sampling methods. This bioinformatic analysis of eDNA samples collected from 24 stations between the years 2014-2016, aims to compare eDNA sampling with more conventional fish larvae sampling methods in assessing teleost fish diversity within the California Current. Specifically, the project determines where there are overlapping findings between the eDNA and corresponding fish larvae samples from the same locations sampled at the same time. The project also analyzes potential correlations between certain environmental and physical variables and the eDNA samples. Due to a limited number of overall fish reads resulting from the eDNA samples, a PCR optimization experiment was also conducted as part of the project to test variables that may increase fish reads. Overall findings from the study demonstrate that a rigorous data analysis was able to be conducted from the sample set, yet future efforts to improve the amount of fish reads would be beneficial. The data show that eDNA is best used in conjunction with other, more traditional methods to gain a full scope of the fish community composition within an ecosystem.

Introduction

For many years, marine researchers have been conducting surveys of aquatic ecosystems. These observations led to knowledge about the intricacies of marine species' ranges, behaviors, and lifespans. Net tows to collect and visually identify marine species have been a survey method in use for many years and are still integral to certain types of marine biology research implemented today. In the early 1900s, marine acoustic survey methods were introduced, initially used to map the bathymetry of the seafloor, and later to map entire schools of fish.[1] In the mid-1900s, the modern SCUBA diving regulator was invented to allow divers to stay underwater for longer periods of time. There are various methods of marine species surveying, but most involve either seeing the animal directly (either in person or via camera) or hearing the animal with a hydrophone or similar device.

Environmental DNA in marine research, is a lesser-known survey method to detect and identify species that have shed their DNA within a water sample, and simply requires collection of water. Collection of environmental DNA (eDNA) is an option that doesn't involve interaction with the animal, thus making it non-invasive. eDNA technology can be used for soil, sand, or in this case, ocean water samples. eDNA sampling has become increasingly popular in recent years due to its proven success in identifying presence and/or absence of species, and its certain advantages over other survey/detection alternatives. eDNA sampling can detect rare or shy species that are otherwise difficult to survey, as well as species present in low numbers. The technology is also more cost effective than most other survey methods. In addition, though not a perfect

[1] New York Sea Grant. (n.d.). *ACOUSTICS UNPACKED*. Cornell University. http://www2.dnr.cornell.edu/acoustics/History.html

technology, it can be more accurate in certain circumstances due to the reduction of human error in identification of species.[2]

This report will outline the bioinformatic analysis that was conducted on eDNA samples collected from the National Oceanic and Atmospheric Administration (NOAA) California Cooperative Oceanic Fisheries Investigations (CalCOFI) Genomics Project (NCOG): a partnership between NOAA, J. Craig Venter Institute and Scripps Institution of Oceanography (SIO). This report will also provide a high-level overview of the eDNA sampling process for this sample set, although that portion was not within the scope of this research project.

The samples used for the analysis, hereafter referred to as "NCOG eDNA samples", were collected from twelve CalCOFI cruises spanning 2014-2016. CalCOFI has been conducting ocean observations within the California Current since 1949, but only added eDNA sampling in 2014 as part of the NCOG data time series.[3]

Research Objectives

First, this project aims to understand relative abundance of fish families of the eDNA samples. Relative abundances of the samples were analyzed in relation to variables such as depth of sample, depth of sea floor, season, etc.

Second, while eDNA is no longer a newly discovered technology, there are still unresolved questions about its use. For this reason, "ground-truthing" the findings of eDNA with other sampling methods is of interest. The main research question from this project is to determine where overlap occurs between the NCOG eDNA samples and the larvae and egg samples taken from the same cruises at the same locations. This research question was possible to analyze due to the fact that CalCOFI conducts different types of sampling, multiple times a year along the same transect lines. The NCOG eDNA samples were collected from 24 stations, at which time egg and larvae samples were also collected, allowing for this comparison.

Other research questions of interest include analysis of environmental variables to determine if any strongly correlate with fish community composition (ex. chlorophyll A levels, season, temperature), as well as physical variables that may affect fish assemblage composition (ex. bathymetry, sample depth, ocean floor depth).

Lastly, due to the low percentage of teleost fish reads identified in the NCOG eDNA samples, a PCR optimization experiment was conducted to test which methods, if any, may increase percentage of fish reads during future sample processing. The variables tested were an increase

[2] Petit-Marty , N., Casas, L., & Saborido-Rey, F. (2023). State-of-the-art of data analyses in environmental DNA approaches towards its applicability to sustainable fisheries management. Frontiers in Marine Science, 10. https://doi.org/10.3389/fmars.2023.1061530

[3] CalCOFI. (2024, May). https://calcofi.org/

in the annealing temperature from 60ºC to 62ºC and the size selection of samples prior to sequencing, i.e., the physical removal of the off-target amplification band.

Methods

Water Collection

Marine eDNA analysis begins with collection of water samples. There are many protocols for water collection, one commonly used method is collection by research vessel. This method uses a conductivity, temperature, depth device (CTD), equipped with a rosette of Niskin bottles that open on either end and can close at the chosen depths. The CTD measures how conductivity and temperature of the water column change in relation to depth,[4] while also measuring salinity, fluorescence (a Chlorophyll A proxy), and oxygen. The CTD is lowered down to the depth/s of choice, and the Niskin bottles are closed remotely via a computer onboard the ship. For the NCOG eDNA samples, the depths of the samples ranged from surface to 515 meters deep. Once the CTD is brought back aboard the research vessel, water from each target depth is transferred from the Niskin bottles and filtered onto a membrane to collect the biological material. This project used .2 µm Sterivex filter cartridges, however filter pore sizes may vary depending on the target DNA molecule size. The filters are then stored at -80º (or similar) until DNA extraction can occur.

Laboratory

Generally, most methods of DNA extraction involve the following steps: lysing the sample to break apart the cell membrane; separating the DNA from unwanted material in the cell; binding the DNA together; washing any proteins or contaminants from the solution containing the bound DNA; and eluting the DNA to obtain it in its pure form.[5] The Macherey-Nagel NucleoMag Plant Kit for DNA purification was used for extraction of the NCOG eDNA samples.[6] Automated instrumentation for pipetting allow approximately 80 samples to be processed at a time using this method. Once the DNA has been extracted, a polymerase chain reaction (PCR) must be conducted to amplify target segments of the DNA.[7] In this case, the 12S ribosomal RNA MiFish gene was the target of the PCR, which has been proven to be successful in targeting teleost fish species.[8] The 12S gene has a slow rate of evolution, allowing any small differences in the gene to

---

[4]NOAA Office of Ocean Exploration and Research. (n.d.). *What does "CTD" stand for?* Ocean Exploration. https://oceanexplorer.noaa.gov/facts/ctd.html#:~:text=By%20measuring%20the%20conductivity%20of,the%20temperature%20and%20the%20salinity.

[5] *DNA Purification*. Promega. (2024). https://www.promega.com/resources/guides/nucleic-acid-analysis/dna-purification/#:~:text=There%20are%20five%20basic%20steps,to%20a%20purification%20matrix%2C%204)

[6] Rabines, A., Lampe, R., & Allen, A. E. (2020). Sterivex DNA extraction V.2 . *Protocols.Io*. https://doi.org/dx.doi.org/10.17504/protocols.io.bc2hiyb6

[7]Khehra, N., Padda, I. S., & Swift, C. J. (2023). Polymerase Chain Reaction (PCR). *National Library of Medicine*.

[8]Miya, M., Sato, Y., Fukunaga, T., Sado, T., Poulsen, J. Y., Sato, K., Minamoto, T., Yamamoto, S., Yamanaka, H., Araki, H., Kondoh, M., & Iwasaki, W. (2015). MiFish, a set of universal PCR primers for metabarcoding

be used to identify different species. Reagents for this PCR included 2X Phusion Master Mix, 10 µm Forward Primer (MiFish), 10 µm Reverse Primer (MiFish), nuclease-free water, rAlbumin, and the template DNA. Each reaction had a total volume of 20 µL.

Table 1: Multidisciplinary University Research Initiative (MURI) MiFish Primer Set[8]

Primer Name	Direction	Sequence	Target Amplification Size (base pairs)
MiFish-U-F_mod	Forward	GCCGGTAAAACTCGTGCCAGC	163-185
MiFish-U-R	Reverse	CATAGTGGGGTATCTAATCCCAGTTTG	163-185

Table 2: MURI protocol thermocycler conditions used for the NCOG eDNA samples

Step	Temp	Time	No of cycles
Initial Denaturation	98	30 s	1
Denaturation	98	10 s	35
Annealing	60	30 s	
Final Extension	72	10 min	1
Hold	4	Infinity	1

A thermocycler is used to conduct the PCR, heating up the extracted DNA to separate the two strands in a process called denaturation. Next, MiFish primers are used to bind to a hypervariable region of the 12S rRNA gene, which contains sufficient information to identify teleost fishes to taxonomic family, genus and species.[8] This occurs during the annealing process. Next, nucleotides are joined together by a thermal-stable DNA polymerase in the extension process to complete two new double strands of DNA. This process is cycled 35 times to result in exponential amplification.

environmental DNA from fishes: Detection of more than 230 subtropical marine species. *Royal Society Open Science*, *2*(7). https://doi.org/10.1098/rsos.150088

A second PCR is conducted to attach indices to each of the samples. This process essentially adds "barcodes" to each sample, to keep the data from each sample separate when they are pooled during the sequencing step.

Once the PCR is complete, gel-electrophoresis is conducted, to determine the results of the PCR amplification. DNA is negatively charged, thus when an electric current is run through the gel, the DNA will migrate towards the positive side. Smaller molecules of DNA will travel further through the gel, thus separating DNA based on fragment size. Using a fluorescent dye to stain double stranded DNA, results are visualized with UV illumination.

The NCOG eDNA samples were then sent offsite to be sequenced, and the raw reads were returned to be analyzed *in silico*.

PCR Optimization Experiment

Much off-target amplification was seen within the NCOG eDNA samples, potentially due to the small filter pore sizes and/or PCR conditions used. This optimization experiment was conducted as a result of the bioinformatic analysis results. The goal of the optimization experiment was to enhance the efficiency and accuracy of the 12S MiFish primers and next-generation sequencing in generating fish-specific sequences, for potential future use of the NCOG samples. The experiment was designed to test three different PCR conditions, in addition to library size selection. The PCR conditions tested included the original MURI protocol (Table 2), Monterey Bay Aquarium Research Institute (MBARI) protocol with 2X Platinum SuperFi II Taq at 60°C annealing temperature, and MBARI protocol with 2X Platinum SuperFi II Taq at 62°C annealing temperature (Table 3). The Platinum SuperFi II Taq was designed to have high fidelity and increased resistance to PCR inhibitors.[9]

Table 3: MBARI protocol thermocycler conditions

Step	Temp	Time	No of cycles
Initial Denat	98	30 s	1
Denaturation	98	10 s	38
Annealing	60 OR 62	10 s	
Extension	72	30 s	
Final Extension	72	5 min	1
Hold	4	infinity	1

A total of ten samples were chosen for this project from the CalCOFI Intercalibration Experiment, conducted in October, 2022. Each group of ten samples was subjected to all three

[9] Thermo Fisher Scientific . (2024). *Platinum SuperFi II DNA polymerase-high-fidelity PCR enzyme*. Platinum SuperFi II DNA Polymerase-High-Fidelity PCR Enzyme. https://www.thermofisher.com/us/en/home/life-science/pcr/pcr-enzymes-master-mixes/platinum-high-fidelity-pcr-enzyme.html

PCR conditions as well as two pooling conditions (size selected and not size selected). Ultimately, the results of the three PCR test cases and the library size selection would be compared to one another using the MiFish 12S marker gene, to determine which one yielded the highest ratio of vertebrate to bacterial sequences.

Bioinformatics

The bulk of this project was a bioinformatic analysis of the existing NCOG eDNA samples that had already been processed in the lab using the aforementioned methods. The raw sequences were received from UC Davis Genome Center, which used an Illumina MiSeq instrument (2 x 250 bp PE chemistry). The sequences were received demultiplexed, meaning there was one file per sample, and underwent the following process to identify species from the raw files.

Trim Raw Reads

The raw reads were received post-sequencing in paired-end format, with one forward read (R1) and one reverse read (R2) per sequence. When merged, the R1 and R2 reads make one single, paired-end sequence.[10] The first order of events was to trim the primers and adaptors from the raw reads. Primers are specific to the target gene and are used during the amplification process while adapters are specific to the Illumina sequencing process. Sequence quality was also assessed before and after trimming, using the programs FastQC and MultiQC. Whereas FastQC runs a quality analysis on individual reads, MultiQC aggregates the output of FastQC into one comprehensive quality report.[11]

[10] Azenta Life Sciences. (2023). *Raw Data Frequently Asked Questions*. GENEWIZ. https://web.genewiz.com/raw-data-faqs

[11] StabuWIKI. (2017). *FASTQC and MultiQC*. FASTQC and MultiQC - wiki. https://stab.st-andrews.ac.uk/wiki/index.php/FASTQC_and_MultiQC

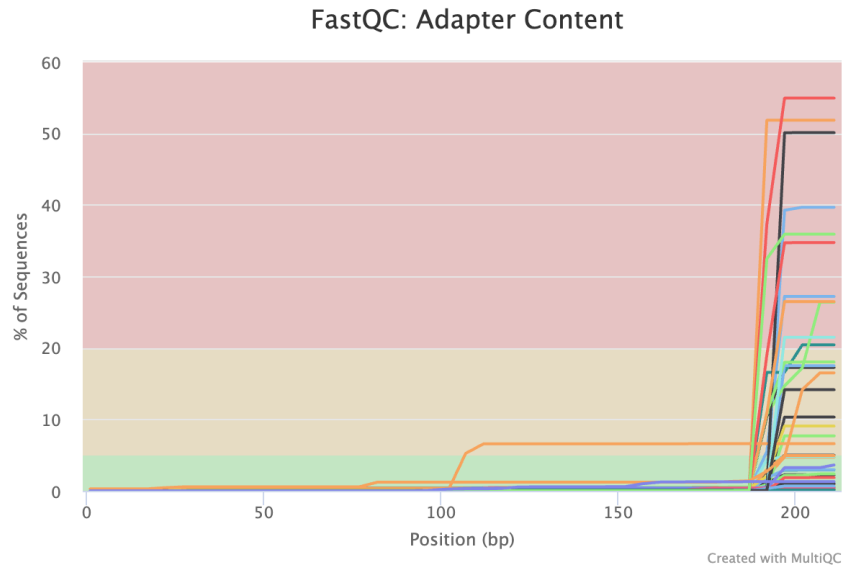


Figure 1: Adapter content plot created from FastQC, using MultiQC pre-trimming

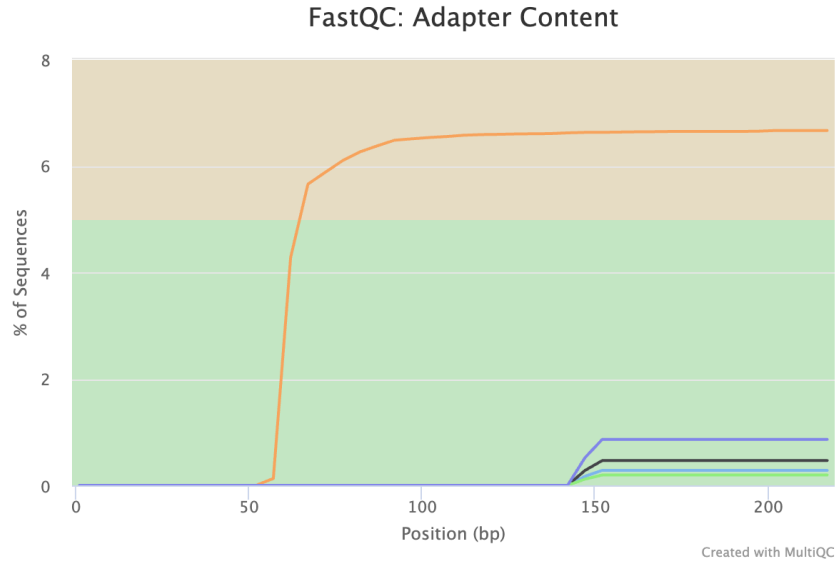


Figure 2: Adapter content plot created from FastQC, post-trimming with Atropos

The samples were trimmed to remove adapter and primer content using both Cutadapt and Atropos. Both programs are similar in output and were compared against each other to determine which more effectively trimmed the samples.

Denoise and Generate Amplicon Sequence Variants (ASVs)

The data undergoes a denoising process of grouping reads into ASVs. For this step, the program called DADA2 is used within the R programming platform. DADA2 has a few important steps, one of which is sample inference, in which the algorithm uses the error rates it has calculated in

order to determine 'true sequence variants.'[12] After that process, the reads are merged by aligning the paired forward and reverse reads and constructing the full-length sequence. Another important step of the DADA2 process was size selection, in order to ensure that we only retained reads within the length of our target sequence. MiFish primers, similar to other primers, tend to amplify reads that are shorter than and/or longer than the target read length. More often, the longer bacterial reads are amplified, which we attempt to both filter out during the PCR stage as well as the bioinformatic analysis. For fish DNA sequences, the lengths are most often between 163 and 185 base pairs, so we set these as the parameters, and filter out reads that have more or fewer base pairs. This process was especially important for this project, since the .22um filter that was used resulted in more microbial DNA reads than fish, so those reads were filtered out.

Assign Taxonomy

The use of eDNA as a survey method only works as well as the reference database and its sequences. For this project, a reference database that is specific to fishes within the California current was used. As part of development of this reference database, tissue was sequenced from 597 species using the MiFish 12S primers, ultimately adding 252 species to GenBank's existing 550 California Current Large Marine Ecosystem fish sequences.[13]

Plotting and Analyzing Data

Phyloseq is a program in R that is used to analyze and display ASV sequence data visually via plots and graphs. Phyloseq is used after the DADA2 process, once ASVs have been configured and taxonomy has been assigned.[14] Similar to the previously mentioned bioinformatic methods, Phyloseq was originally developed for microbiome analysis, but can be adapted for fish DNA sequence analysis. Phyloseq was used to obtain an overview of the different fish families in the samples, as well as calculate and plot diversity metrics using the Shannon and Simpson diversity indices.

A Principle Coordinate Analysis (PCoA) was also created using the Bray Curtis dissimilarity metric. The PCoA aggregates samples by common level of dissimilarity of the samples. Various variables were overlain onto these plots by color and shape to determine if there were correlations.

In addition, Phyloseq was used to calculate permutational ANOVAs (PERMANOVAs), to directly determine correlations between variables using an associated p-value. As opposed to the multivariate ANOVA test, PERMANOVA can be used when the data does not meet the ANOVA assumptions, as is often the case with genetic data. PERMANOVA calculates the variation of a group with a distance matrix and compares the variation between multiple groups

---

[12] Callahan, B. (n.d.). *DADA2 Pipeline Tutorial (1.16)*. Dada2 Pipeline Tutorial (1.16). https://benjjneb.github.io/dada2/tutorial.html

[13] Gold, Z., Curd, E. E., Goodwin, K. D., Choi, E. S., Frable, B. W., Thompson, A. R., Walker, H. J. Jr., Burton, R. S., Kacev, D., Martz, L. D., & Barber, P. H. (2021). Improving metabarcoding taxonomic assignment: A case study of fishes in a large marine ecosystem. *Molecular Ecology Resources*, 21, 2546–2564. https://doi.org/10.1111/1755-0998.13450

[14] An, A. (n.d.). Phyloseq: Explore microbiome profiles using R. https://joey711.github.io/phyloseq/

to the variation within a group.[15] The end result of the PERMANOVA test is a p-value, which determines significance of the test. A p-value less than .05 is considered significant and would be cause to reject the null hypothesis (that the variation is the same of all groups). A PERMANOVA was used in this research to compare the samples with various environmental variables from the metadata.

Results

Relative Abundance

This project had an interest in taxonomic composition of fishes. Variables such as depth of sample, depth of seafloor and season were plotted with the taxonomic composition to identify any correlations. Figure 3 shows family composition relative abundance of fishes, by depth of water sample. The composition of fish families by sample depth appears to be similar at each depth grouping. This may indicate that eDNA sampling from one depth, has the ability to identify fish families across different depths (within 500 meters). More research using larger sample sizes would be required to confirm this finding.

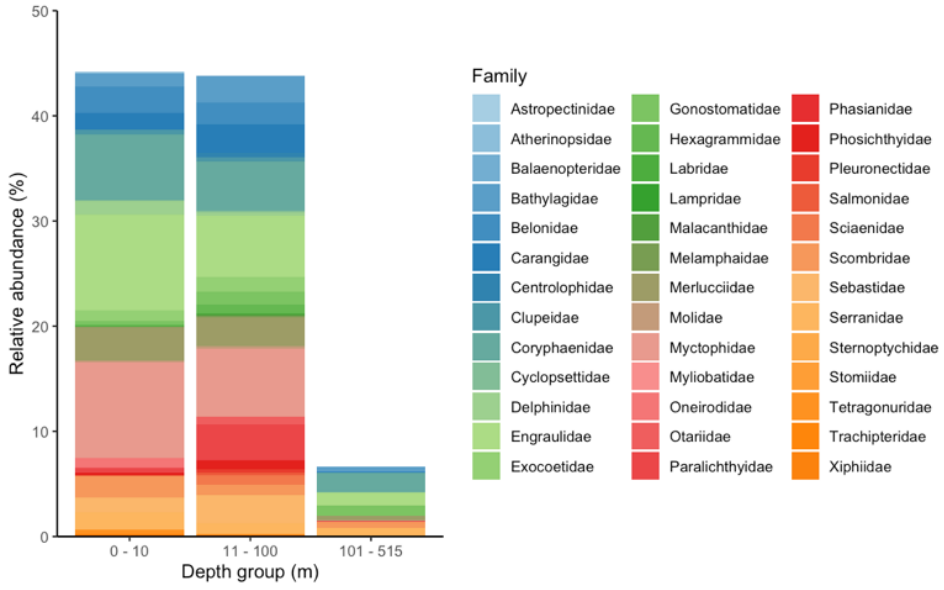


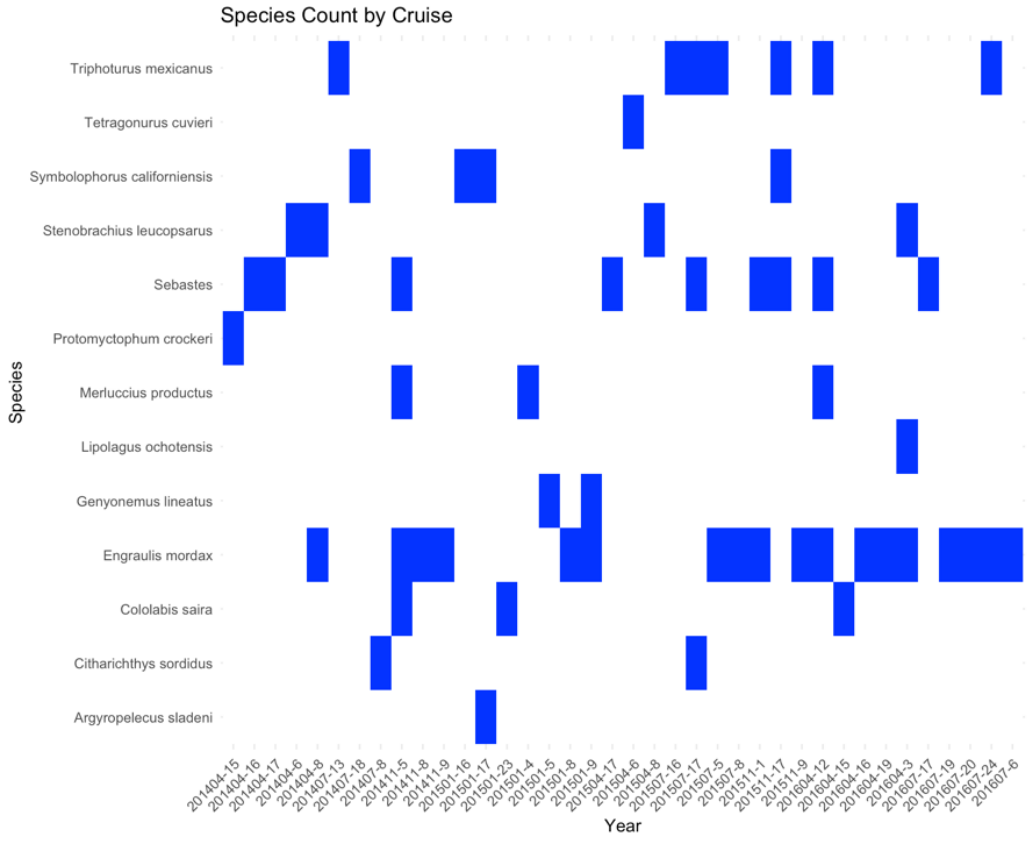
Figure 3: Relative abundance of families by depth grouping (from surface to ~500 meters). n= 66 for 0-10m, 58 for 11-100m, 9 for 101-515m)

[15] Bakker, J. D. (2024, January 3). *Permanova*. Applied Multivariate Statistics in R. https://uw.pressbooks.pub/appliedmultivariatestatistics/chapter/permanova/

Egg vs. Larvae vs. eDNA Overlap

The CalCOFI egg samples had less species variation (eight different species), yet four of them overlapped with both the larval and eDNA sample data. Specifically, Northern Anchovy (*Engraulis mordax)*, Pacific saury (*Cololabis saira)*, North Pacific hake (*Merluccius productus)* and Yellowtail amberjack *(Seriola lalandi).* The idea was raised by one Capstone Committee Member that eggs shed less DNA than larvae and adult individuals, so if DNA from an egg was to be "picked up" in the eDNA, it would likely be from the entire egg, not sloughed off DNA.[16] Based on the sample sizes and number of species identified in the larvae and egg data, the larvae had the more robust dataset for comparison, so that was used instead of the egg data to analyze overlap. Overlap was defined as a species that was identified in the eDNA and the larval samples at the same station during the same cruise.

By cruise and station, there were 56 instances of overlap between eDNA and larvae samples. Within the data samples, there were 12 species that overlapped and one family that was not identified to species level. Figure 4 shows these 56 instances of overlap by cruise and station.



[16] Ostberg, C. O., & Chase, D. M. (2021). Ontogeny of edna shedding during early development in Chinook salmon (*oncorhynchus tshawytscha*). *Environmental DNA*, *4*(2), 339–348. https://doi.org/10.1002/edn3.258

Figure 4: Fish eDNA and larvae species overlap by cruise and station. Each point on the x-axis represents both a cruise by month and year, and a CalCOFI sampling station

The data were also plotted by station, to visualize which stations had more overlap in species between eDNA and larvae (Figure 5). Some stations had more overlap than others, prompting additional questions such as: Were certain species overlapping more often due to PCR amplification bias? Did bathymetry or depth of the station contribute to more overlap in species identification?

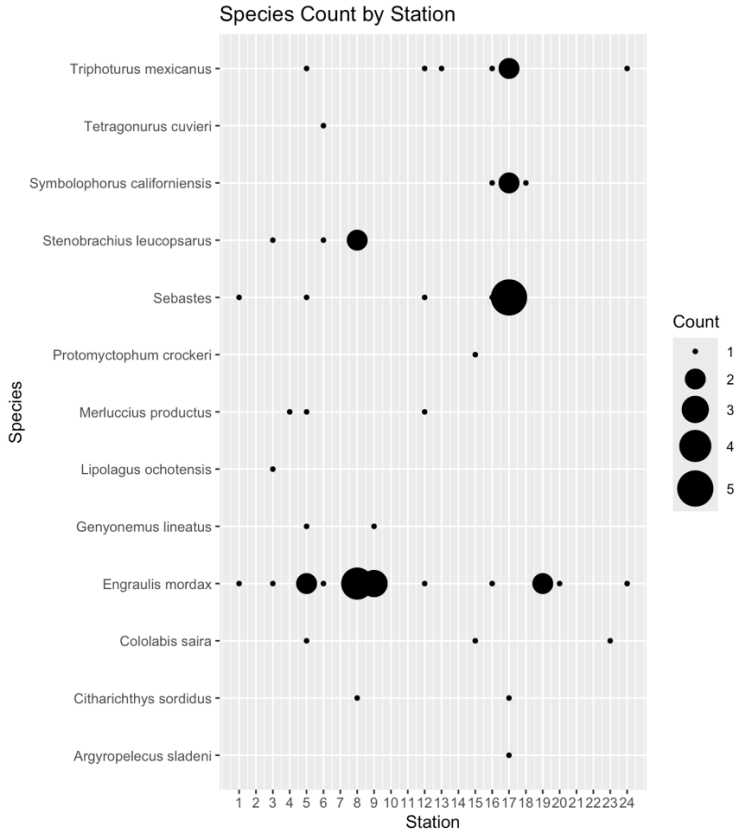


Figure 5: eDNA and larvae overlapping species (overlap defined as species identified at the same station during the same cruise). Plot shows only species by station, in order to determine which stations had more overlap.

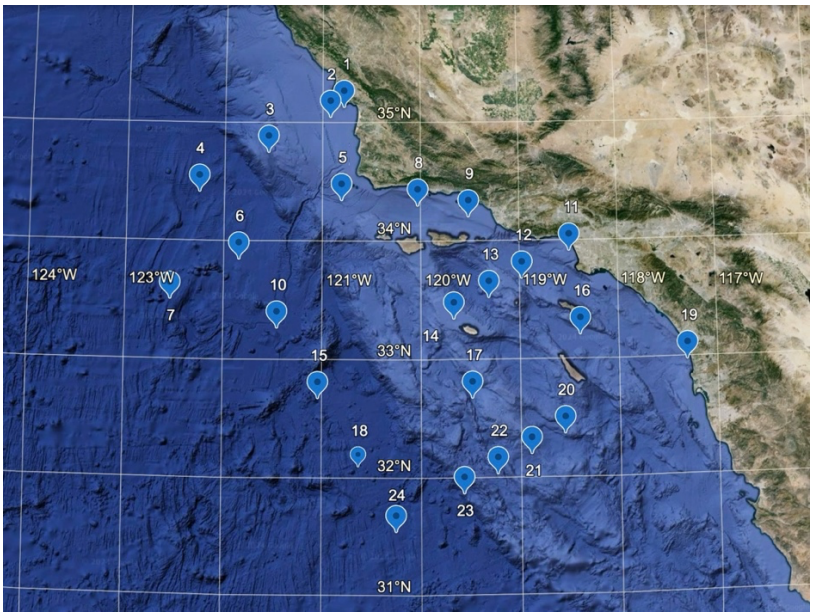


Figure 6: Geographical locations of eDNA sampling stations

Stations five and 17 had many overlapping observations and were both located in the "slope" region based on their bottom depth. Station five had seven overlapping observations of six total species, and Station 17 had 11 overlapping observations of five total species. Focusing on Northern anchovy (*Engraulis mordax*) in particular, since it overlaps many times in these samples, we see that Station eight had four overlapping observations of *Engraulis mordax*, two in 2014 and two in 2015. Station nine had three overlapping observations of *Engraulis mordax,*one in 2014 and two in 2015. Overall, the NCOG eDNA data showed seven instances of overlap by station of Northern anchovy in 2016, four in 2014 and six in 2015.

While these findings in particular may not reveal a unique pattern, these data show that eDNA may be useful in identifying fluctuations for this important species in the California Current. CalCOFI has been observing and sampling fishes off the California coast since 1949, one of which has been Northern anchovy. A 2013 assessment of coastal pelagic species in the California Current calculated overall risk related to climate change for Northern anchovy as a combination of exposure and sensitivity to climate change. Northern anchovy scored approximately .8 out of 1 for sensitivity and approximately .715 on exposure, where sensitivity is defined as 1/species range area, and exposure was estimated based on the *"magnitude of expected change in a multivariate climatic index, given the current observed distribution of each species within the California Current ecosystem."*[17] The elevated risk level of Northern anchovy makes it an interesting species to track over a longer time scale, to determine if the levels of identification are shifting in either direction. Northern anchovy is also an indicator of healthy ecosystems due to the important role it plays in the food chain for other fish, marine mammals

[17] Harvey, C.J., N. Garfield, E.L. Hazen and G.D. Williams (eds.). 2014. The California Current Integrated Ecosystem Assessment: Phase III Report. Available from http://www.noaa.gov/iea/CCIEA-Report/index.

and birds[18]. The fishing industry has an interest in this species as well for its importance as prey to many commercially caught fish.

Environmental and Physical Variable Correlations

Another research question involved testing the effects of environmental variables on fish communities observed within the eDNA samples. A general lack of clustering in the PCoA plots created from this sample set suggests that the variables chlorophyll A, season and depth do not strongly influence the similarities/dissimilarities among samples.

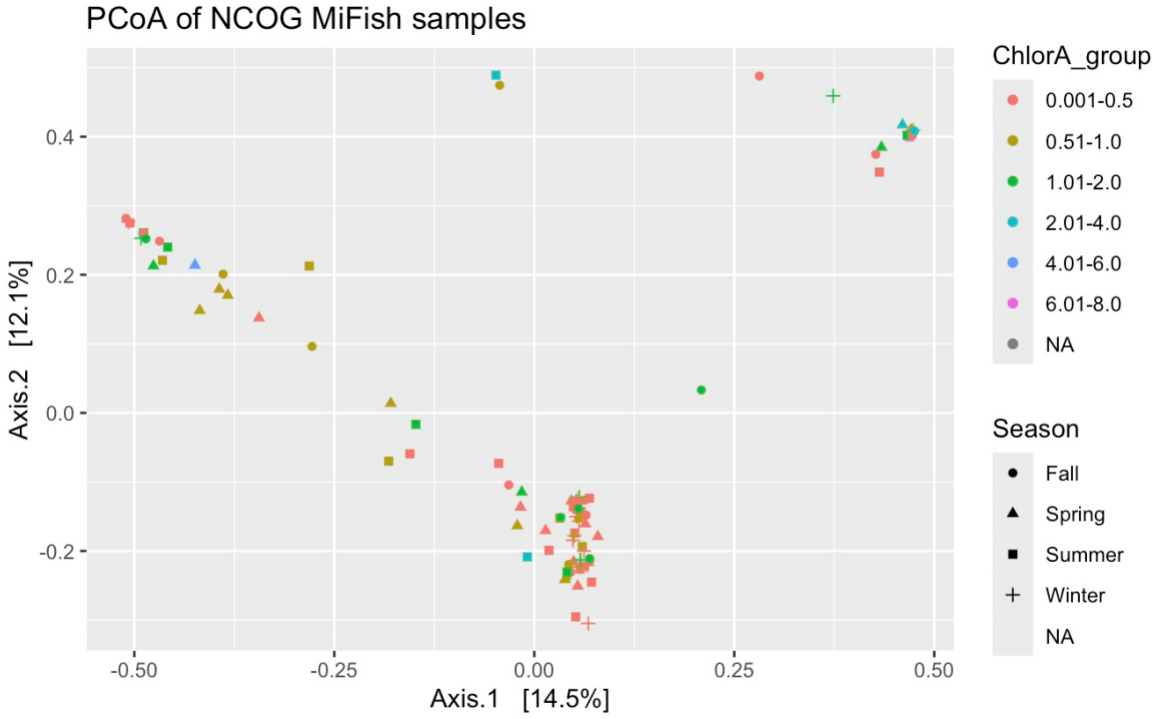


Figure 7: Principal Coordinates Analysis of NCOG eDNA samples categorized by season and chlorophyll A groupings

Additional plots assessing temperature and depth groupings do not show strong correlations either. Figure 8 shows a PCoA categorized by depth of seafloor at the location of the sample. The samples are categorized by three onshore stations (0-200m), 14 slope stations (200-3,000m), and seven offshore stations (>3,000m), to equal 24 stations. Within this graph the offshore samples appear to be closely related, however there are still confounding samples in the slope

[18] NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION | U.S. DEPARTMENT OF COMMERCE. (2024). *Northern Anchovy*. NOAA Fisheries. https://www.fisheries.noaa.gov/species/northern-anchovy

categories that make it difficult to draw a clear conclusion. The unbalanced sample sizes for each category could also contribute to the lack of pattern.

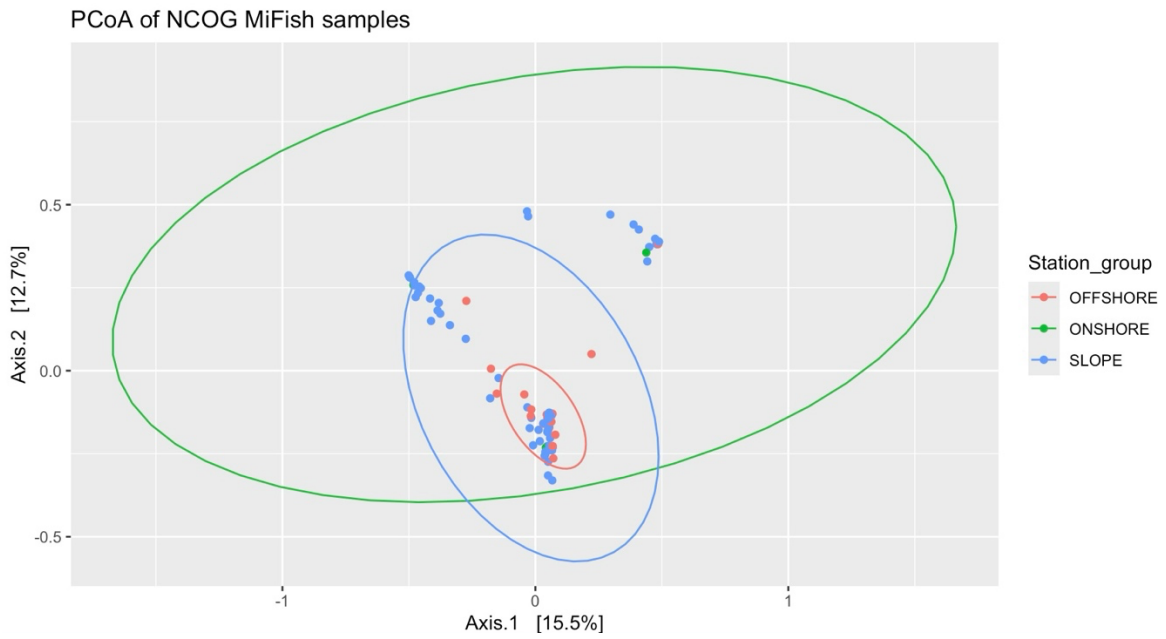


Figure 8: PCoA of NCOG MiFish eDNA samples by season and chlorophyll A groupings. The tighter ellipse around the offshore samples shows that they may be more tightly constrained.

The PERMANOVA in Figure 9 was run to determine correlations between the NCOG eDNA samples, season, and temperature. The p-value of .046 indicates a potential significant correlation between the samples and the season. The p-value for sample depth grouping in meters (1-10, 11-100, 101-515) is .409, showing no significant correlation, similar to chlorophyll A groupings which had a p-value of .928. The stations were also grouped by bottom depth to identify any patterns related to that physical variable, however a PERMANOVA for that variable returned a p-value of .74, indicating no significant correlation.

```
adonis2(formula = dist ~ Season * T_degC, data = ps.perm.data, permutations = 999)
              Df SumOfSqs      R2      F Pr(>F)
Season         3    1.812 0.03523 1.3240  0.046 *
T_degC         1    0.415 0.00807 0.9101  0.540
Season:T_degC  3    1.312 0.02550 0.9586  0.588
Residual     105   47.899 0.93120
Total        112   51.438 1.00000
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 9: A PERMANOVA from the Vegan Library in R showing a statistically significant correlation for season

PCR Optimization Experiment

The PCR optimization experiment tested three test cases: First, the protocol that was used for the NCOG eDNA samples (MURI protocol); second the MBARI protocol at both 60º and 62º; third

all three PCR conditions plus an extra step for size selection, where the off-target amplified band is physically removed prior to sequencing.

Preliminary results, after sequencing, showed that the size selection step had the largest impact on retained (merged) percentage of teleost fish reads. While percentage of merged reads increased, the overall number of distinct fish taxa did not, therefore the size selection did not have an effect of overall diversity. In addition, the MBARI protocol at 62º annealing temperature improved the percentage of merged reads, however it also did not contribute to improved diversity of fish taxa.

Limitations of eDNA

Species Abundance

While eDNA has been proven to be an important resource used for surveying marine species, there are limitations to the technology. One limitation is that as of now, there is not a rigorous way to link eDNA sample results to number of individuals, since the amount of DNA in the water cannot be directly correlated to number of organisms. However, research is being conducted to make progress in this area. In terms of eDNA being an indicator for abundance of an organism, Shelton et al. found that "both eDNA- and seine-derived abundance indices reflect the seasonal migration of salmon."[19] A comparison of the two methods found that at the population level, the two "provide virtually identical quantitative information", whereas they were less correlated at the site scale.[18]

There has also been research linking direct techniques that may better allow metabarcoding to estimate quantities of organisms. One approach models the process of generating ASVs to estimate starting DNA proportions of many taxa, rather than relying on data transformation later on in the process. Using these inputs, the research calibrates the model using mock communities in order to provide an estimate of proportions of DNA per taxa.[20]

Presence/Absence Determinations

Another limitation is that non-detection of a species' DNA does not unequivocally mean that species was not present. The species' DNA concentration may be under the detection threshold due to degradation within the water sample or simply less sloughing from the target species, or it may not have been amplified during the PCR process. Conversely, the presence of a species' DNA in the results does not always mean that species was present and alive in that water sample.

---

[19] Shelton, A. O., Kelly, R. P., O'Donnell, J. L., Park, L., Schwenke, P., Greene, C., Henderson, R. A., & Beamer, E. M. (2019). Environmental DNA provides quantitative estimates of a threatened salmon species. *Biological Conservation*, *237*, 383–391. https://doi.org/10.1016/j.biocon.2019.07.003

[20] Shelton AO, Gold ZJ, Jensen AJ, D Agnese E, Andruszkiewicz Allan E, Van Cise A, Gallego R, Ramón-Laca A, Garber-Yonts M, Parsons K, Kelly RP. Toward quantitative metabarcoding. Ecology. 2023 Feb;104(2):e3906. doi: 10.1002/ecy.3906. Epub 2022 Dec 21. PMID: 36320096.

There may be cases where DNA is excreted in the feces of a predator that had eaten elsewhere, it could have been carried from currents, or it could be deceased.[21]

Off-Target Amplification

There is also the issue of off-target amplification, which can occur as part of the sampling process, as well as due to the primers used during PCR. Choosing which filter pore size to use during the water filtration process may narrow down species detection. Filter pore sizes of .45-.7 microns have shown success in collecting fish DNA, although they are still small enough to capture bacteria in as well.[22] This can cause off-target amplification that must be filtered out during the bioinformatic step.

Primers can also amplify off-target DNA sequences. In the case of fish, 12S MiFish primers are regularly used, which have the tendency to amplify bacterial DNA that partially align with the primers. These bacterial sequences can be filtered out during the bioinformatic process, yet create more "noise" in the data. With a higher proportion of bacterial DNA sequences than fish, the bacteria are more likely to become amplified, potentially causing data results that prioritize microbes instead of fish.

Further Research

Low Percentages of Target (Fish-Specific) Reads

One of the drawbacks with this data analysis were the low proportions of fish reads overall in the dataset. Efforts are made during the water sampling process to collect an adequate amount of water with an effective filter for the target gene, as well as during the PCR step to conduct a PCR protocol that results in high proportions of the target gene. However, results can differ based on many variables in this process, and ways to increase percentages of target reads is an ongoing point of research.

Reads in this process were filtered out at the bioinformatic step by targeting the gene length. If a longer gene length that that of the target gene is used to filter the data allowing more reads to merge, there may be more reads to work with resulting in higher abundance, but also more off-target reads. A balance must be found between keeping reads that may be very close matches to the target gene, while not keeping many reads that are off-target.

This dataset, in particular, had fairly low percentages of reads that merged, i.e. reads at the right length for fish. Across the samples, approximately 10.8% of the reads merged from the filtered data per sample. Table 4 shows the input reads, the amount kept after initial quality score

---

[21] Roussel, J., Paillisson, J., Tréguier, A., & Petit, E. (2015). The downside of eDNA as a survey tool in water bodies. *Journal of Applied Ecology*, *52*(4), 823–826. https://doi.org/10.1111/1365-2664.12428

[22] Shu, L., Ludwig, A., & Peng, Z. (2020). Standards for methods utilizing environmental DNA for detection of fish species. *Genes*, *11*(3), 296. https://doi.org/10.3390/genes11030296

filtering, the amounts kept for forward and reverse reads after the program removed likely errors, and finally those that merged.

Table 4: Subset of samples (1-20) showing how many reads were the original inputs, how many were filtered, denoised, and merged

	input	filtered	denoisedF	denoisedR	merged
FISH_001_S1_L001_R1_001-trimmed.fastq	35654	33225	32830	32911	665
FISH_002_S2_L001_R1_001-trimmed.fastq	35996	33660	33400	33317	3542
FISH_003_S3_L001_R1_001-trimmed.fastq	43622	39392	39076	38959	4529
FISH_004_S4_L001_R1_001-trimmed.fastq	71461	67879	67238	67209	1136
FISH_005_S5_L001_R1_001-trimmed.fastq	40916	38546	38237	38271	3676
FISH_006_S6_L001_R1_001-trimmed.fastq	11745	11018	10887	10909	1294
FISH_007_S7_L001_R1_001-trimmed.fastq	40172	37524	37200	37215	3053
FISH_008_S8_L001_R1_001-trimmed.fastq	59185	55269	54894	55005	6223
FISH_009_S9_L001_R1_001-trimmed.fastq	47718	44523	44276	44215	4923
FISH_010_S10_L001_R1_001-trimmed.fastq	45087	42122	41936	41965	15151
FISH_011_S11_L001_R1_001-trimmed.fastq	43034	40675	40365	40426	980
FISH_012_S12_L001_R1_001-trimmed.fastq	46259	43313	43000	43021	1422
FISH_013_S13_L001_R1_001-trimmed.fastq	37113	34694	34425	34390	2363
FISH_014_S14_L001_R1_001-trimmed.fastq	7565	7053	6954	6962	413
FISH_015_S15_L001_R1_001-trimmed.fastq	83380	79067	78703	78555	115
FISH_016_S16_L001_R1_001-trimmed.fastq	59065	55736	55548	55511	22
FISH_017_S17_L001_R1_001-trimmed.fastq	38525	35749	35467	35568	3089
FISH_018_S18_L001_R1_001-trimmed.fastq	31697	30007	29813	29823	1144
FISH_019_S19_L001_R1_001-trimmed.fastq	43632	41285	41075	41003	1094
FISH_020_S20_L001_R1_001-trimmed.fastq	22240	21085	20938	20934	29

Another variable that can affect target reads are the PCR conditions, specifically the annealing temperature. In this case of many bacterial reads and not many fish reads, increasing the annealing temperature may help rid the data of the bacteria during the PCR stage, yet as with setting the target gene lengths, there must be a balance. As annealing temperature changes, it makes PCR more or less forgiving of mismatches in primer sequences. As the PCR temperature is lowered, it increases overall amplification (even off-target). Whereas increasing temperature reduces PCR yield in that the species that may not be direct matches but are close enough to become amplified, are no longer viable, however the results will be more specific.[23]

Filter Pore Size

Even before the samples becomes processed in the lab, there are various ways to optimize fish sequences, as opposed to bacteria. One of which is larger pore sizes in the filters used during water filtration. A paper published in NCBI that conducted a literature review of 168 papers on

[23] Lubelsky, Yoav. (2018). Re: WHY does increasing the annealing temperature make PCR more specific? . Retrieved from: https://www.researchgate.net/post/WHY-does-increasing-the-annealing-temperature-make-PCR-more-specific/5ac387dadc332dc8d756a833/citation/download.

eDNA, found that the most commonly used pore size for fish eDNA specifically is .45 μm.[22] However, the most common sized particles of fish DNA molecules in water are reported to be between 1 and 10 μm, so a larger filter may still be effective while having the advantage of filtering out more bacterial DNA particles.

The NCOG MiFish eDNA samples were filtered using .22 μm pore sizes, due to the fact that the original target of the eDNA was bacteria. However, while there was off-target amplification of bacteria when processing these samples for fish, they still yielded fish data that could be utilized for analysis.

A recent SIO cruise experimented with 3 μm filter pore sizes for 21 samples taken between San Diego and Morro Bay. After processing of the samples is completed, comparison can be made with smaller pore sizes to determine which yielded more fish DNA sequences.

Droplet Digital PCR (ddPCR)

Droplet Digital PCR (ddPCR) has certain advantages over qPCR, especially for detection of rare organisms, or ones that slough lower levels of DNA into the target substrate.[24] ddPCR distributes the sample into multiple individual reactions, leaving one or more of the target molecules in each one. This allows for detection of a positive match, seeing as how the comparison is not being made against any other molecules.[25] Based on presence or absence of a fluorescent signal post-partitioning, during "end-point PCR cycling", the absolute number of molecules can be calculated.[17] By providing absolute quantification directly and not relying on standard curves, this method efficiently measures the target DNA of the sample. Precision of this method may also be higher than qPCR, due to the partitioning of the sample enabling measurements of small differences in the copy numbers of the target DNA sequences among samples, as well as removal of the potential for amplification efficiency bias of qPCR.[26]

Conclusion

This analysis of the 2014-2016 NCOG eDNA samples found that 24 of the species (53%) identified in the eDNA samples, overlapped with the larvae samples, across all stations and cruises. By station and cruise, there were 56 instances of overlap between the NCOG eDNA samples and the CalCOFI larvae samples. 21 of the species identified in the eDNA samples were unique, in that they were not also found in the larvae samples. Within the unique species

---

[24] Mauvisseau, Q., Davy-Bowker, J., Bulling, M. *et al.* Combining ddPCR and environmental DNA to improve detection capabilities of a critically endangered freshwater invertebrate. *Sci Rep* **9**, 14064 (2019). https://doi.org/10.1038/s41598-019-50571-9

[25] https://www.qiagen.com/us/applications/digital-pcr?cmpid=CM_PCR_dPCR_Traffic_0123_SEA_GA_NA&gad_source=1&gclid=CjwKCAjwr7ayBhAPEiwA6EIGxKKrvsr7PDRj0xcH_gpWKEZGXzAYWOJ4UGoU6Lcum7ub_geqnxMNMxoCtcYQAvD_BwE

[26] *Droplet digital PCR (ddpcr) technology*. Bio-Rad. (2024). https://www.bio-rad.com/en-us/life-science/learning-center/introduction-to-digital-pcr/what-is-droplet-digital-pcr

identified only in the eDNA samples were Risso's dolphin (*Grampus griseus*), humpback whale (*Megaptera novaeangliae*) and common mola (*Mola mola)*. 159 species and families were found only in the larvae DNA samples, including many in the Sebastes family that were identified to species (whereas they were only identified to family level in the eDNA samples). Approximately 15% of the species identified in the larvae samples overlapped with the eDNA samples (across all stations and cruise dates). These data show that these two survey methods may be best used as complementary methods, since there were species identified in the eDNA samples that were not present in the larvae samples, and vice versa.

None of the environmental variables analyzed were found to strongly influence NCOG eDNA fish taxonomic composition, with the potential exception of season. The relative abundance by depth of sample plot shows similar family composition across different depth grouping. More research and a greater sample sized data set would be required to further confirm and analyze these findings.

While the NCOG eDNA time series samples were not originally targeting fish DNA and were therefore filtered with a pore size to target bacteria and microorganisms, this project has shown that the samples yielded sufficient fish DNA for a rigorous analysis. With the results of the PCR optimization experiment, efforts can be made to potentially improve the fish reads of these samples. Continuing to process these NCOG samples for fish species could show important trends in fish communities, especially during these times of intense changes within the ocean; ocean warming, ocean acidification, overfishing. By further utilizing this consistent dataset, comparative analyses of fish community compositions can be conducted utilizing larger sample sizes over greater temporal scales, and identification of patterns can occur in correlation with environmental and physical variables.