**Report on Metabarcoding pilot project**

### Project:

### David Richardson and Chris Orphanides. A pilot project exploring the utility of next-generation DNA sequencing in fisheries research. FY 2018 Groundfish and Climate money

### Definitions:

### *DNA barcoding:* A means of providing species identifications using the sequence of a short standard fragment of a gene.

### *Sanger sequencing:* One long-established method of DNA sequencing that provides a single sequence for a single specimen. Most useful when a specific individual needs to be identified and tissue from that individual can be obtained without contamination from other individuals.

### *Next-generation sequencing:* An alternate high-throughput form of DNA sequencing in which many unique sequences can be developed from a single sample.

### *Metabarcoding:* A special case of DNA barcoding using Next generation sequencing in which many species can be identified from a single sample comprised of a mixture of DNA from many individuals.

### *Read count:* Metabarcoding can produce 100,000s of individual sequences. These are then each assigned to a species if they meet a certain threshold for quality and % match to a sequence in a database. The number of sequences that match an individual species is the read count. While this is a quantitative measure, the number should not be interpreted as corresponding directly to the amount of DNA of that taxa in that sample. Rather the guidance is to interpret it as a measure of the reliability of that identification.

### *PCR primers:* Short pieces of single-stranded DNA, usually around 20 nucleotides in length. Two primers are used in each PCR reaction, and they are designed so that they flank the target region (region that should be copied). Ideally, the sequence of a primer will allow them bind to opposite strands of the template DNA. Mismatches between primer and template sequences will negatively affect the PCR reaction.

### Project Goal:

### We sought to evaluate the cost and performance of using an outside group to perform metabarcoding on NEFSC samples. The ultimate goal is to be able to mail off a sample with little to no pre-processing and to have the data returned to us in a timely manner (1-2 months). This approach would eliminate the need for dedicated lab space, technicians, equipment and reagents and would be well suited to cases in which the amount of funding available and the project needs vary substantially year-to-year. Two sub-projects were initially identified to evaluate this technique:

### *Metabarcoding of seal feces:* The goal is to develop a list of species consumed by the seal that produced the scat. This approach to diet analysis would complement other techniques that are currently being used (hard part analysis, fatty acids, stable isotopes), and in particular would be useful for testing the hypothesis that some prey items are being underestimated due to the seal not consuming hard parts (i.e. belly-biting).

### *Metabarcoding of larval fishes:* The goal was to generate a species list for a mixed sample of larval fish eyeballs. In our offshore sampling, family level IDs are common for certain groups (e.g. Scorpionfish, Snappers, Lattern fish etc). We are interested in whether it would be possible to develop a cost-effective species list for these samples.

### Background of using the University of Guelph for Sanger Sequencing:

### The Oceans and Climate branch has been contracting with the University of Guelph for DNA barcoding of fishes using Sanger sequencing since 2009. We have submitted around 2500 samples, with about 90-95% of them producing species level IDs. Primarily this work has focused on identifying fish eggs and larvae, with the data going into a number of publications and assessment working papers.

### Often on the plates we have submitted we have added other types of samples to evaluate specific questions. We have evaluated whether there is a cryptic species of thorny skate in U.S. waters as was suggested in one paper (we saw no evidence of this). We have evaluated sand lance taxa on the northeast shelf which typically are not identified to species. We have also run a set of unidentified prey items from trawl survey fish with about an 80% success rate.

### Our success with the University of Guelph led us to evaluate whether we could also work with them on metabarcoding. In developing the contract we contacted other sequencing centers at regional universities. None of them were willing to provide the end-to-end service we were looking for (i.e. DNA extraction to bioinformatics).

### Cost:

### The invoice from the University of Guelph was for $9300 CAD. $7100 of this was for 2 plates (95 samples) of metabarcoding, and $2200 of this was for one plate (95 samples) of Sanger sequencing. The cost per metabarcoding sample was thus $37.37 CAD or approximately $28.20 USD given the current exchange rate.

### In this pilot project we used a high level of replication to evaluate the consistency of the results. For cost purposes, a sample is defined as one replicate. One fecal sample had 12 or 18 replicates so costs were scaled accordingly. Going forward the level of replication needed will determine the true cost of running a unique sample.

### Results of Larval fish project

### *Samples and approach*

### 23 samples were submitted.

### Each sample had two extraction replicates and each extraction replicate two PCR replicates.

### Samples were a mixture of eyeballs from 5-25 taxa of larval fish from a single net tow. The remainder of each fish was retained to allow further identification. Some fish we can ID to species level, others to only family level.

### Most samples were from the Slope Sea and Blake Plateau which have higher diversity and many more taxa identified at the family level. Northeast continental shelf samples of larval fish collected during EcoMon are typically identified at the species level morphologically and thus there is less utility in using this technique on them.

### *Results*

### All 23 samples provided data, though 2 out of 92 of the PCR replicates failed to provide data.

### On average each sample provided 157,000 reads that met the threshold for sequence quality and a match to a fish in the database.

### 118 taxa were identified, most at the species level, but a few at a coarser taxonomic resolution.

### Replication: Of 192 Sample/Taxa combinations 166 showed up in all 4 of the PCR/Extraction replicates (average reads 21000). Those that did not appear in each replicate typically had low read counts (<200). The analysis provided to us used a cutoff of 50 read counts; below that a species was not reported. These results indicate that the cases of poor replication were species at the threshold of detectability in the sample, rather than major changes replicate to replicate.

### Looking back at a subset of the vials of fish there appeared to be little problem with false positives. Those species that were sequenced had matches in the vials.

### There seemed to be a major problem with false negatives. Between 20-40% of the taxa that were submitted in a sample of ≈20 taxa did not have a matching sequence. Given that all of these samples have undergone the same exact collection and preservation (i.e. always been in the same jar/vial) the false negatives seem attributable to the technique and not the underlying sample. Additionally, the false negatives qualitatively did not correspond to the size of the eyeball submitted (i.e. many large fish with large eyes were still false negatives).

### The approach was far from quantitative. Read counts varied by a couple orders of magnitude with fish that had similar size eyeballs.

### *Conclusions on larval fish work*

### The high level of false negatives is a major concern and is limit on the utility of this technique as implemented. This level of false negatives is likely due to primer design. These primers amplify a 184 bp portion of the COI gene. In our Sanger sequencing these primers are used in a second round of sequencing when the first round of sequencing with a 650bp fragment fails. They do work generally for that purpose, but their utility for metabarcoding seems to be marginal.

### The consistency across replicates was good. In the future I would not envision having to implement the same level of replication.

### There are some applications in our ichthyoplankton work where a species list would be beneficial. However, the work we have done on individual taxa (red hake, bluefin tuna, etc) benefits most from the one sample/one sequence approach of Sanger sequencing.

### The larval fish project did provide important context for evaluating other results using the same approach (i.e. see below).

### Results of Seal fecal sample project

### *Notes on samples and approach*

### Four types of samples were submitted

### Fecal samples from the colon of bycaught gray seals and harbor seals that were collected during necropsies (N=6)

### Fecal samples of gray seals collected on the beach (n=2)

### Fecal samples of captive seals only fed Atlantic herring (n=2)

### A positive control of tissue from 5 species that are in ethanol in vials at the Narragansett lab

### The fecal samples of the bycaught animals were frozen upon collection. The others were put in ethanol. All samples were shipped in ethanol.

### A subsample (about a kidney bean size) of each fecal samples was transferred from the original sample container to a vial at the Narragansett lab for shipment. Most samples this represented <5% of the material available)

### High levels of replication were pursued for this approach. 2 fecal subsamples (bycaught animals, captive animals) or 3 fecal subsamples (beach collections) were sent for each original fecal collection. Each subsample had 3 extractions and 2 PCR replicates. The replication was thus 3x3x2=18 or 2x3x2=12 for a single field/lab collection.

### We specifically targeted fish taxa in our contract; no attempt was made to pursue a technique that also identified invertebrates.

### *Results of sequencing*

### The two Captive seals fed only herring yielded no sequence data. Among the possibilities for this are 1) the primers are a poor match for herring (see note on larval fish project above), 2) cleaning products (e.g. bleach) could have been used in the facility and caused issues.

### The positive control was problematic in terms of false negatives. 5 species were included in it. Haddock never sequenced. The myctophid only sequenced in one replicate. Sand lance read counts were very low. The other two species were consistent with high read counts. One of 6 replicates of the positive control had 100 reads of white shark that are unexplained.

### The top 6 species in terms of read counts and number of times identified across samples were: red hake, silver hake, cunner, longhorn sculpin, yellowtail flounder, ocean pout. Some species we expected (sand lance) were absent. Considering the results of the larval study, the results of the captive seals, and the positive control, it seems almost certain that there are false negatives including potentially major prey items.

### One sample had a harbor porpoise listed as prey of a gray seal. While this predator-prey action is documented, it seems highly likely that this is a contamination problem (false positive) as harbor porpoises are processed in the same work space.

### Read counts in general were far lower and less consistent for the fecal samples than the larval fish samples. Some samples had high read numbers across all replicates, with the same three species identified. Other replicates had very few read numbers, with values approaching the detection limit.

### A couple reported IDs were highly unlikely. In the sample with a lot of longhorn sculpin reads there were also arctic sculpin reads (1.5% the number of reads as longhorn sculpin). This is likely an issue that could be addressed on the bioinformatics end. While I have not examined the raw sequences, it is possible that they were only a poor match to arctic sculpin. The default threshold is 95% to call a species match. For other work I typically see >99% matches in all sequences. Changing this threshold would lower the number of reads likely but could eliminate these issues.

### *General Impressions*

### Encouraging aspects

### It is possible to get DNA suitable for metabarcoding from bycaught and beach collected samples with little specialized protocols in terms of preservation and collection.

### From the standpoint of time-demands for NEFSC staff, this is very easy and the results are pretty quick (late November sample shipment to data available January 20th). This 2 month period included the holidays.

### Many of the results made sense in terms of the species mix and had reasonable replication and read counts.

### Factors that need work to make this a useful technique

### As with the larval fish work there is evidence that false negatives are a notable issue.

### We are dealing with degraded DNA so the read counts were lower. We also likely need to evaluate thresholds and sequence quality cutoffs more closely rather than relying on the defaults.

### We had a near certain case of contamination (harbor porpoise DNA). The lab is used for multiple types of samples. Lab protocols will need to be considered if this type of work is to be pursued.

### Unlike for larval fish, replication on these samples seems to be needed. The degree of replication directly impacts cost: 1 sample ($30), 4 replicates ($120), 8 replicates ($240). Further work should help us understand the degree of replication.

### General Approach moving forward

### Enough material exists with these samples to support any direct comparisons that could be considered.

### In terms of working with the University of Guelph Center for DNA barcoding, the first issue to address concerns which primers to use. There are alternative primer sets that have been used for other studies that we can evaluate. There are various factors to consider in primer choice including: how universal are they for the target taxa, do they allow species-level IDs (or just lower taxanomic resolution IDs), are there established databases to compare the sequence data to.

### One potential way forward is to contract the University of Guelph to rerun these samples with a new set of primers and to compare the results. I would imagine the DNA is available to do this. The cost for this approach would likely be low and could potentially fit within the FY2020 NEFSC allocation for the NOAA Genomics Strategic Initiative. Other approaches in the short term are more costly.