### RESULTS OF SEAL FECAL SAMPLE METABARCODING AT THE NEFSC MILFORD LAB

### 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.

***False Positive:*** A species or taxa identified in a sample that did not in fact occur in that sample

***False Negative:*** A failure to identify a species in a sample that was known to occur in that sample.

**Project overview**

The goal of this project was to develop a metabarcoding approach to marine mammal diet analysis that the NEFSC would be able to use in combination with other currently implemented techniques. Each diet analysis technique is thought to have its own benefits and drawbacks. For metabarcoding the benefits are that the approach does not rely on a prey items hard parts being consumed and identifiable. Additionally with a well designed technique a wide array of taxa can be identified at a fine taxonomic resolution. The drawback is that there are uncertainties about the degree to which the technique is quantifiable generally, with some levels of quantification likely not ever possible (e.g. for a species distinguishing 10 100g prey items from 1 kg prey item).

This project was originally intended to evaluate whether we could work with an outside group (Center for DNA barcoding/University of Guelph) to implement a metabarcoding approach to evaluating seal diets. In that initial pilot project, funded by the NE groundfish climate initiative, there were major problems with false negatives. That specific technique was deemed to be of little use for our purposes. A write up of those results is here (put in link).

This write up covers a second methodology implemented by Yuan Liu at the NEFSC Milford Laboratory. There were two components to this work. The first was a testing of the eDNA technique with a very wide array of morphologically identified taxa (23 orders of fish, 48 families, ≈75 species, with 6-18 taxa per sample). Those results are written up in detail here (). Overall that test of the approach was highly successful, with very minimal occurrences of false negatives or false positives. This was achieved by using more generalized primer sets than were previously trialed. One tradeoff in this primer set was that species level identification for some taxa was not possible. A second tradeoff was some less common taxa were not in the voucher sequence database for this gene region. This lack of vouchered material was particularly evident for more tropical or deepwater species. The lack of vouchered material was less of an issue for NE. U.S. continental shelf species. Additionally, nearly all if not all of the commercially important species in the NE US are in the database

This writeup is for the second component of this work, which involved working with seal diet samples. The makeup of these samples is as follows:

* 24 total samples
* 2 fecal samples from captive harbor seals only fed Atlantic herring
* 2 tissue samples from stomach contents
* 2 fecal samples from wild grey seals collected at the Head of the Meadow
* 9 fecal samples from bycaught Harbor Seals
* 9 fecal samples from bycaught Grey Seals

**Methodology:**

* Mitochondrial 12 S Gene, about 130-140 base pairs
* 1-5 replicates per sample. Replication was done to evaluate two different extraction kits and consistency of the results.
* The first step in the process was to homogenize the fecal sample. All replicates are based on different subsamples of this homogenate.

**Results from Captive animal:**

Two fecal samples were processed from a captive harbor seal that was only fed Atlantic herring (one sample was from a known individual, the second was not assigned to a specific individual). For each of these samples four replicates were run. For the first sample, averaged across the four replicates, 99.56% of the sequences corresponded to Atlantic herring, 0.4% of sequences to Atlantic menhaden (in 1 of 4 replicates), and 0.04% to silver hake (in 1 of 4 replicates). For the second sample 99.8% of the sequences were Atlantic herring and 0.1% were Atlantic cod and silver hake respectively (in 1 of 4 replicates). The total number of sequences for sample 1 in the 4 replicates were 2208, 1999, 2597, 8390. For sample 2 the numbers were 3084, 7550, 9752 and 10160.

Of note, these samples were previously run at the University of Guelph and yielded no sequence data. At the time the reason for this was uncertain, but we speculated that the failure may have been due to either the methodology used by the University of Guelph or a cleaning product (e.g. bleach) used at the facility that degraded the DNA. With the success of this run, a methodological issue seems like the most likely cause.

**Unidentified tissue from stomach samples**

Two unidentified samples of tissue material were pulled from stomachs and were run. The first had 99.997% Atlantic herring sequences in 1 replicate. The second had 76.8% Atlantic herring and 22.9% Atlantic cod sequences in one replicate.

**Fecal samples bycaught animals and wild animals**

***Read count***

There was substantial variability among samples in the number of reads obtained, but among replicates of an individual sample the variability in read counts was limited. The average read counts for each sample were as follows

* 10 samples had >10,000 reads on average per replicate
* 3 samples had 1000-9999 reads on average per replicate
* 2 samples had 100-999 reads on average per replicate
* 5 samples had <100 reads on average per replicate

These results likely point to a substantial amount of variability in the DNA degradation among samples. The highest read counts were in the two samples collected on the beach (Head of the Meadows).

**Species presence absence**

A figure showing the presence and absence of taxa by sample is below. Of note, taxa that were identified in <=50% of replicates for a sample are excluded from the figure. Taxa are categorized by the number of reads and the percentage of replicates they were identified in. The total number of replicates and the average read count per replicate is listed in the label for each sample.



**Proportion of sequences for a taxa by sample**

The figure below shows the average proportion of sequences across replicates that correspond to each taxa.



**Read counts and proportions across all samples excluding aquarium and tissue samples**



**REPLICATION PLOT**

The plots below show the data on replication. Note these are exploratory plots. Some more thought would need to be put into the statistics of this. For these plots individual samples are represented more than once if there were greater than 2 replicates of that sample. For example with a 3 replicate sample the plot shows: Replicate A vs B; Replicate A vs C, and Replicate B vs C. Some samples had 4 and 5 replicates done and in which case each replicate is represented 3 or 4 times in the plot respectively.

The first plot is the natural log of read counts for each taxa sampled.

The second plot is an evaluation of the differences in DNA extraction methodology, with comparisons of the New Fast Kit on the x-axis versus Old Kit on the y-axis. Each dot is an individual taxa.

The third plot is similar to the second, but only looking at total counts for a sample aggregated across all taxa. The Old extraction kit results had slightly more reads, about 20-30% on average.



These plots shows the consistency in the proportion of read counts for a taxa for among replicates. Again each replicate may be plotted multiple times. The plots are broken into categories based on the total number of reads for a sample across all taxa. That is, the first plot only includes data from replicate pairs if both replicates had >5000 reads across all taxa. For the second plot the number was 500.

Of note individual taxa/samples are represented multiple times here. For example, four beard rockling was absent in 1of 4 replicate in sample 18, but was a notable proportion of sequences in the other 3 replicates. This one absence accounts for the one dot on the x axis (0.42, 0) and the 2 on the y axis (0, 0.2 & 0, 0.22) circled in red.

As expected as the total number of reads is reduced (samples with low total read numbers usually indicate DNA degradation) there is less consistency in the results among replicates.

****

****

****