**Report on Metabarcoding Positive controls**

Yuan Lui-Milford Lab

David Richardson-Narragansett Lab

Chris Orphanides-Narragansett 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.

**Current Project:** This project is an evaluation of metabarcoding of samples that are a mixture of known taxa. In an ideal situation all taxa included in the sample would be identified (i.e. no false negatives), no taxa would be identified that were not in fact included in the sample (i.e. no false positives) and all taxa would to identified at the species level (i.e. species level identification resolution).

**Positive control samples:** The positive controls were generated from ethanol preserved larval fish samples. There were a number of advantages in working with larval fish for this project.

* We had ready access to >50 taxa of larval fish across the major orders of ray-finned fish (Actinopterygii) in the marine environment in our region. The samples were put together in winter 2021 when NOAA cruises had not been sampling for ≈1 year and special sample requests were not being taken for upcoming cruises. Developing as broad a range of taxa based on adult material would not have been possible.
* With larval fish we could ensure consistency in the preservation (and hopefully DNA quality) of the individuals provided as positive controls. Many samples were a mixture of taxa from the same collection event and were preserved simultaneously in the same jar. We believe differences in preservation and DNA quality should not be a factor in the results.
* It was possible to retain all of the individuals used for analysis for future work. In all cases larval eyeballs served as the tissue sample, with the remainder of the larva being saved.

The major downside of working with larval fish is that for some taxa it is not possible to morphologically identify the individual at the species level.

We sought to generate positive control samples that were a mixture of taxa common to the northeast U.S. continental shelf, and positive control samples that were a more diverse set of taxa including tropical/subtropical, mesopelagic, and deep benthic species.

We provided 15 positive controls, with 7 being unique and 4x2 being replicate samples (i.e. one larval eyeball went in one sample and the second eyeball went in the second sample). In total the positive controls contained 23 orders of fish, 48 families, ≈75 species, with 6-18 taxa per sample. A total of 140 sample/taxa combinations were generated.

**Sample Results**

Of the 140 sample/taxa combinations we were able to identify sequences corresponding to 137 of them (97.8%). 3 samples were false negatives (taxa in the samples with no sequence). An additional 12 sequences were false positives (sequences that did not correspond to a taxa placed in the vial).



Histogram of true positives, false positives and false negatives by read count

Read counts of true positives ranged from 20-172,814. Read counts of false positives ranged from 10-289. For true positives 134, 125 and 116 had read counts above 200, 500, and 1000 respectively.

All three false negatives were in replicate samples. Two false negatives were from the same individual (Ophidiformes) that was not sequenced in each replicate sample. The third false negative was present in one replicate sample but not the other. Multiple possibilities exist to explain these false negatives. One possibility is that the eyeballs never actually made it into the sample at Narragansett that was eventually sent to the genetics lab at Milford.

Of the 12 false positives one had a closest match in Genbank to a worm species. Three samples were 100% matches to taxa not included in the true positives (Gadus morhua, Thunnus sp., Enchyelyopus cimbrius). These species occur in the plankton samples that were the source of the material for the positive controls. The contamination for these likely occurred during the compilation of the samples in Narragansett. The remaining eight were species that were found in other positive control samples; multiple points of contamination are possible in these instances.



For each sample the number of correct counts and 1 bp different counts

**Sequencing accuracy**

For many specimens included in the samples there were multiple unique sequences that matched, typically separated by a single base pair (out of 138). These additional sequences had in aggregate about 1/78th of the read counts of the main sequence for that taxa. The low frequency sequences likely correspond to a sequencing error.

**Replicates read counts**

Read counts for each taxa in replicate samples were correlated, as were the read proportions (taxa read count/total sample read count). Replicate samples were left and right eyeballs and should have been approximately matched in size, though further size standardization did not take place



**Species resolution and database coverage.**

We were able to match taxa used in the positive controls to specific sequences in all cases. However, there were multiple categories of matching:

1. A 100% match in the database to only a single western Atlantic species. Good database coverage of all of related taxa. Technique will give species resolution data for that taxa in the western Atlantic.
   1. As an example silver hake *Merluccius bilinearis* sequences matched *M. bilinearis* in the database with the only other congener in the region *M. albidus* having a different sequence.
   2. Other taxa in this category included: butterfish (Peprilus triacanthus), dolphinfish, windowpane, fourspot flounder, summer flounder, Xyrichthys novacula, Northern searobin, Atlantic herring, Northern sandlance, Pollock
2. A 100% match in the database, but other closely related taxa are also a 100% match.
   1. As an example, *Urophycis chuss*, *U. regia*, *U. tenuis* all have same sequence. The technique will not differentiate among these species of hake.
   2. Other examples:
      1. Yellowtail flounder-Winter flounder
      2. Sailfish-White marlin
      3. Auxis rochei-A. thazard
      4. Sebastes species
3. A 100 % match in the database but closely related taxa are not in the database so we do not know if these also match.
   1. *Ceratoscopelus* (Myctophidae) is a genus with a couple of species. One is in the database *C. maderensis* (and a perfect match to our sample), but others have not been sequenced.
4. A perfect (or >99%) match in the database is lacking and the species is not in the database. Coverage of the taxa needs to be improved.
   1. A variety of species fit into this category including mesopelagics, more suptropical species, and smaller non-commercially important benthic species (e.g. callionymids)