

Advancing Environmental Genomics for Marine Mammal Characterization

Peter Thielen, D.Eng, ALM
 11100 Johns Hopkins Road, Building 201-486D
 phone: (240) 470-5979 email: peter.thielen@jhuapl.edu
 Award Number: N00014-21-1-2610

Additional Key Contributors: Hayley DeHart, Kristina Zudock

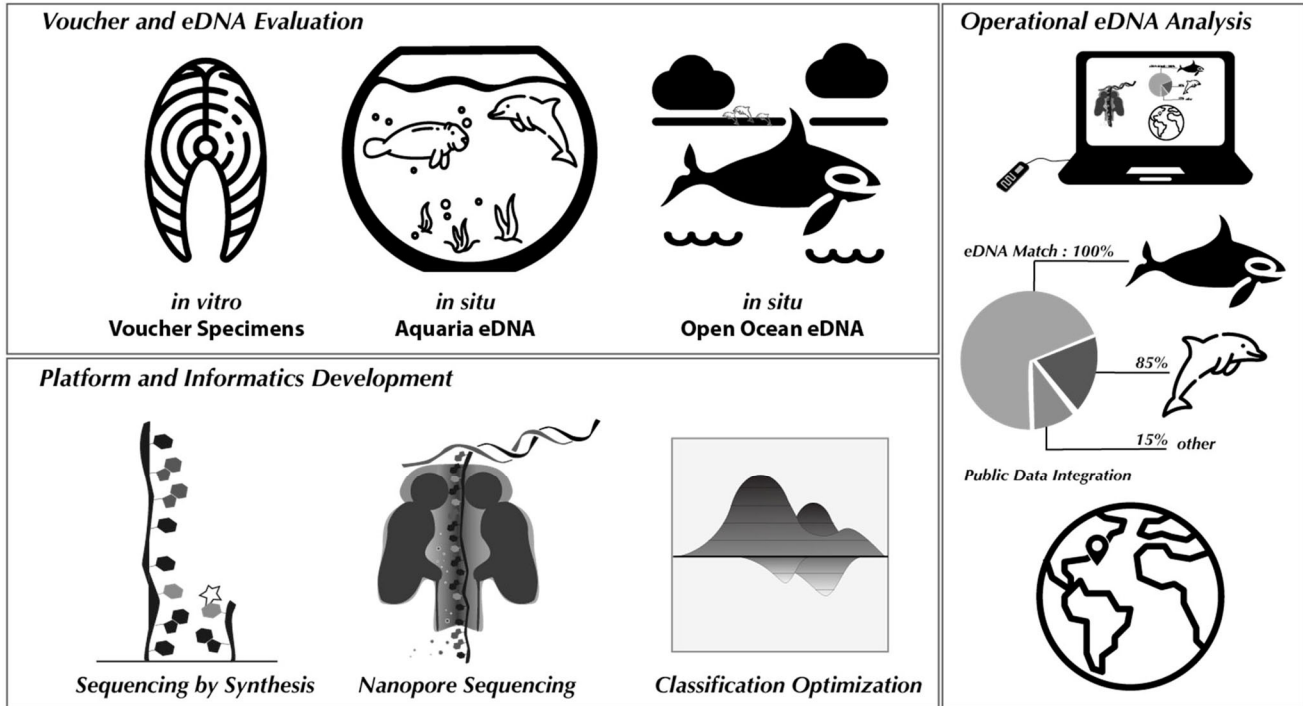


Table of Contents

Introduction	1
<i>Area 1: Advance marine mammal species identification using Oxford Nanopore sequencing</i>	1
<i>Area 2: Advance data processing and visualization to enable in-field eDNA analysis</i>	1
Area 1: Advance marine mammal species identification using Oxford Nanopore sequencing	2
<i>Verification and Comparison of Oxford Nanopore Sequencing Methods and Analysis</i>	2
<i>Advancing marine mammal mtDNA sequence data classification</i>	4
<i>Refinement of genomic classification pipeline for marine mammal eDNA</i>	6
<i>Area 1 Conclusion</i>	6
Area 2: Advance data processing and visualization to enable in-field eDNA analysis	7
<i>Development of new eDNA sample collection and analysis hardware</i>	7
<i>Field-based eDNA sequencing demonstrations</i>	8

<i>Aim 2 Conclusion</i>	12
Impacts and Applications	12
<i>Presentations</i>	12
<i>Publications</i>	14
<i>Engagements</i>	14
<i>Transitions</i>	15
<i>Interns</i>	15
Conclusions	15
Acknowledgements	15
References	16

Introduction

The goal of this effort is to develop an advanced operational genomic characterization capability for detection of marine mammal species using aquatic environmental DNA (eDNA). Throughout the period of performance, the research and development of hardware, software, and methods have supported every stage of the environmental DNA (eDNA) classification of marine mammals. During the first year of research, an emphasis was placed on development of the tools and methods needed to implement rapid eDNA classification of marine mammals, while the second phase of research emphasized optimization, sample collections, and increased testing of methods. The following report is divided into sections supporting the two main objectives of the research:

Area 1: Advance marine mammal species identification using Oxford Nanopore sequencing

In the past year, we have amassed a collection of eDNA samples from polar oceans that have allowed us to perform stress-testing on our Oxford Nanopore sequencing methods. We have compared these results to traditional lab-based Illumina approaches and have also compared these results to naturalist-collected metadata to confirm presence/absence of marine mammal sightings. Additionally, we have identified limitations to current marine mammal genetic markers and explored methods for more rapid identification using non-PCR based methods.

Area 2: Advance data processing and visualization to enable in-field eDNA analysis

In the initial year of research, we developed a suite of hardware and software to enable rapid field-based DNA sequencing and tested these in several operational environments. These efforts culminated in a field test conducted in conjunction with the Monterey Bay Aquarium Research Institute (MBARI). This past year emphasized optimized methods, increased analysis, and implementation of software and hardware updates to continue support of future field operations based on lessons learned from the previous year of field exercises.

This report will also highlight the overall impact and applications in which this research has contributed to and benefitted from. This research has contributed to multiple publications, external presentations and engagements, and research conducted can be utilized across a variety of additional eDNA research.

Area 1: Advance marine mammal species identification using Oxford Nanopore sequencing

Verification and Comparison of Oxford Nanopore Sequencing Methods and Analysis

Development of a rapid, field-capable Oxford Nanopore sequencing method is critical to enabling non-laboratory-based implementation of eDNA classification, especially in monitoring of marine mammals. As previously discussed in our annual report, our first year of research focused on initial development of such a method and initial validations in an aquarium environment. We also briefly demonstrated success of such methods in polar environments to identify marine mammals.

More recently, these rapid methods have been examined utilizing both Arctic and Southern Ocean samples, which are geographically isolated areas and have a wide array of distinct and diverse organisms, including marine mammals. Through partnerships with Lindblad Expeditions, we have collected over 100 samples across a wide array of marine environments and within sightings of multiple identified marine mammal species. Our previous report highlights some initial results in marine mammal identification from unknown polar samples. Not only can these samples serve as a repository for polar ocean biodiversity, but we can utilize these samples and the sightings associated with them to baseline our marine mammal classification abilities utilizing methods that were previously developed.

Several samples were subjected to our optimized Oxford Nanopore rapid sequencing method for amplicon sequencing classification. Upon examination of 12S MarVer1 regions in Arctic and Antarctic subsamples, we were able to correctly identify and validate several instances of known sightings and distributions of baleen whales, toothed whales, birds, seals, and sea lions (Fig. 1) when examining top species-level hits across multiple samples. In instances where incorrect species with no known distributions in the locations sampled (i.e., temperate ocean species in polar samples), we further examined potential causes for mis-classifications. Many mis-identifications could be traced to either a lack of representatives in our database or similar species classification, likely as a result of lack of differentiation for the 12S marker we used. These results demonstrate some of the first verifications of marine mammal sightings with eDNA correspondence at the species level. Future efforts should be aimed to further investigate classification metrics for false positives/false negatives as well as increase in marine mammal attribution and decrease in off-target amplifications. Most off-target amplifications have been attributed to bacterial contamination and with appropriate primer designs and controlled on-target amplification, we would likely be able to address this current problem in eDNA marine mammal sequencing.

Common Name	Latin Name	Antarctica	Arctic
Crabeater seal	<i>Lobodon carcinophagus</i>		
Guadalupe fur seal	<i>Arctocephalus townsendi</i>	genus	
Brown fur seal	<i>Arctocephalus pusillus</i>	genus	
Weddell seal	<i>Leptonychotes weddellii</i>		
Harp seal	<i>Pagophilus groenlandicus</i>		
New Zealand fur seal	<i>Arctocephalus forsteri</i>	genus	
Southern elephant seal	<i>Mirounga leonina</i>		
California sea lion	<i>Zalophus californianus</i>	family	
Walrus	<i>Odobenus rosmarus</i>		
Adelie penguin	<i>Pygoscelis adeliae</i>		
Gentoo penguin	<i>Pygoscelis papua</i>		
Baikal teal	<i>Anas formosa</i>	non-native	
Greater white-fronted goose	<i>Anser albifrons</i>	non-native	
Canada goose	<i>Branta canadensis</i>	non-native	
Gray whale	<i>Eschrichtius robustus</i>		
Common minke whale	<i>Balaenoptera acutorostrata</i>		
Fin whale	<i>Balaenoptera physalus</i>		
Beluga whale	<i>Delphinapterus leucas</i>		
Bowhead whale	<i>Balaena mysticetus</i>		
Humpback whale	<i>Megaptera novaeangliae</i>		
Gray's beaked whale	<i>Mesoplodon grayi</i>		
Strap-toothed whale	<i>Mesoplodon layardii</i>		
Atlantic white-sided dolphin	<i>Lagenorhynchus acutus</i>		
Pan-tropical spotted dolphin	<i>Stenella attenuata</i>		
Harbor porpoise	<i>Phocoena phocoena</i>		
Risso's dolphin	<i>Grampus griseus</i>		
Greater grisson	<i>Galictis vittata</i>	non-native	
Human	<i>Homo sapiens</i>		

Figure 1. Marine mammal polar ocean representation. Several species with known Arctic/Antarctic distributions were identified in appropriate location samples. Further investigations into likely mis-identified species (black outlines) have shown a combination of attribution to lack of representation in databases, close relative designations, and identification of anthropogenic introduction.

Polar ocean samples have also proven crucial to understanding overall biodiversity estimates in comparison of standard laboratory-based eDNA processing and our field-capable Oxford Nanopore protocols. Southern Ocean samples were processed for general eukaryotic diversity using the 18S rRNA region sequenced on both Illumina and Oxford platforms. Upon examination at a phyla-scale, high taxonomic resolution, there was no statistical difference in overall diversity estimates between the two methods (Fig. 2). Additionally, a wide array of different phyla were identified and consistent across both platforms, indicating that a wide array of general eukaryotic diversity exists in polar ocean eDNA samples. Further distinctions between the two platforms may exist at lower taxonomic levels, but these results would need to be further examined to identify both the limitations of the methods, as well as the classification ability of the 18S markers used for these comparisons. Most differences in classifications occur in fungi species, such as Chytridiomycota (Fig. 3), and future efforts should aim to deconvolute these differentiations in classifications between Illumina and Oxford data. As has been noted in previous research, Oxford Nanopore generally exhibits much higher rates of insertion/deletion sites (InDels) in sequencing data, compared to Illumina (Koren et al. 2012; Chen et al. 2020). If higher levels of InDels exist in fungi, this could account for the differentiation in the platforms. These results (along with additional field testing discussed in Aim 2) have shown that current Oxford Nanopore rapid amplicon sequencing for eDNA classifications are comparable to laboratory-

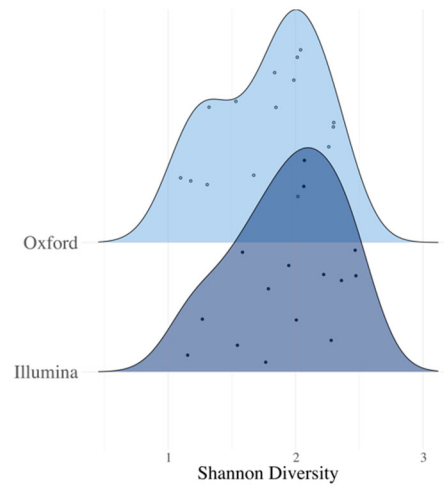


Figure 2. Marine biodiversity estimates in Illumina and Oxford data. At the phylum level, there is no statistically significant difference in overall diversity between field-capable Oxford methods and Illumina sequencing.

As has been noted in previous research, Oxford Nanopore generally exhibits much higher rates of insertion/deletion sites (InDels) in sequencing data, compared to Illumina (Koren et al. 2012; Chen et al. 2020). If higher levels of InDels exist in fungi, this could account for the differentiation in the platforms. These results (along with additional field testing discussed in Aim 2) have shown that current Oxford Nanopore rapid amplicon sequencing for eDNA classifications are comparable to laboratory-

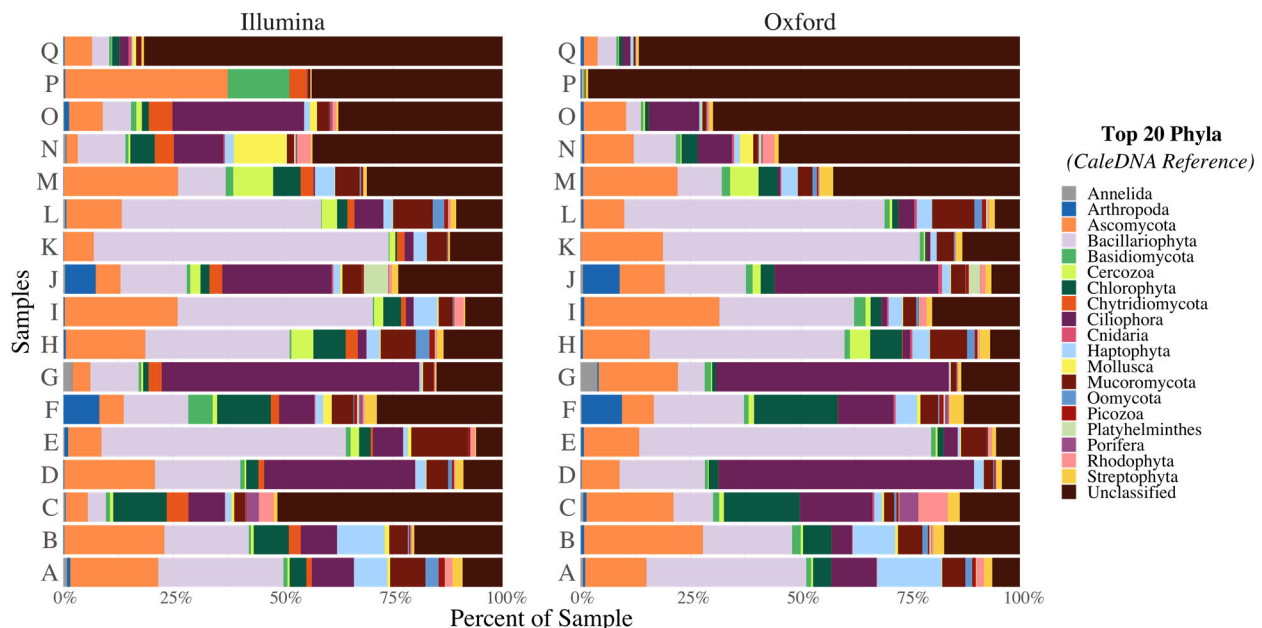


Figure 3. Marine biodiversity comparison. In Southern Ocean eDNA samples, there is a wide array of eukaryotic diversity that is shared amongst Oxford and Illumina data. Overall abundances in Ascomycota, Bacillariophyta, and Ciliophora are consistent between the two platforms. Largest differences in classifications come from fungi Chytridiomycota and Cercozoa. Further investigation is needed at a lower taxonomic resolution.

based Illumina sequencing methods, and represent a major milestone in future field-capabilities for eDNA sequencing and species classifications.

To further assess novel ways in which Oxford Nanopore sequencing methods could be used for marine mammal species identification with eDNA as the ultimate sample source, we performed experiments to determine if non-PCR-based methods could be used to identify marine mammals from tissue samples. Oxford Nanopore has recently developed their adaptive sampling method, which allows for enrichment of targeted sources on the sequencing platform itself, without any sample manipulation or enrichment (in the form of targeted PCR barcoding). To perform this, we prepared sequencing libraries of several marine mammal species (including baleen whales, toothed whales, seals, and sea lions) from tissue we have previously analyzed (previous annual report) using total genomic DNA, which includes non-specific nuclear DNA and species-specific mitochondrial DNA that we typically amplify for species-classifications. By providing the sequencing platform with targeted mitochondrial region sequences of interest, the sequencer is able to reject non-interest regions and sequence only regions matching to the provided sequences. The ability to not amplify mitochondrial regions would make the processing of samples even more rapid than current state-of-the-art methods, and would also limit biases in species representation caused by preferential amplification of different organisms.

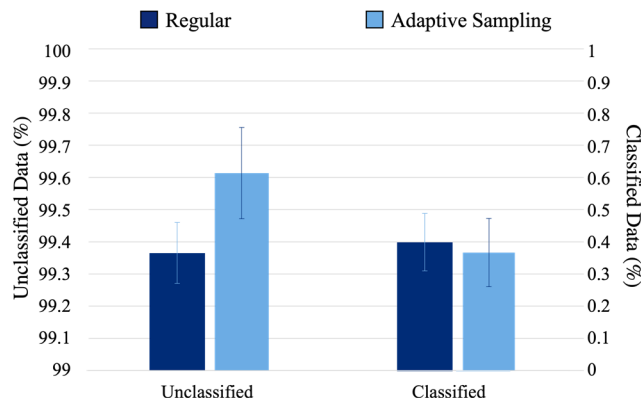


Figure 4. Non-PCR-based Oxford Nanopore classifications. To assess the effectiveness of adaptive sampling for non-PCR Oxford Nanopore sequencing and classification, we sequenced marine mammal tissue and identified the differences in % data classified using traditional and new methods. There was no significant difference in overall percentage of data classified using the adaptive sampling method, indicating unsuccessful enrichment.

suitable for marine mammal eDNA classification, as the ability to highly enrich mitochondrial DNA pre-sequencing via PCR or other methods is still a requirement for eDNA analysis. However, future efforts to utilize higher quality data (such as R10 flow cells or other adaptive sequencing methods) and further testing of samples and rapid library prep methods may show adaptive sampling as a potential candidate for eDNA studies.

Advancing marine mammal mtDNA sequence data classification

As discussed in our previous annual report, the dolphin habitat team at the National Aquarium has been invaluable in providing us access to eDNA samples for testing and troubleshooting our bioinformatics pipeline for marine mammal eDNA analysis. By sequencing eDNA from the dolphin exhibit pools and classifying data using our Kraken2 marine mammal mitochondrial database, we were able to identify

To examine the differences in mitochondrial representation in sequencing data, we first sequenced total genomic samples and then continued sequencing with adaptive enrichment. As expected, total genomic data sequencing only produced an average of 0.4% classified mitochondrial data. However, adaptive sampling did not enrich for mitochondrial DNA as hypothesized. Adaptive sampling actually caused an overall decrease in classified data, and additionally did not enrich for longer sequences (Fig. 4). Both methods were correctly able to classify the correct species in the classified data, but eDNA samples would likely contain much less marine mammal DNA present in samples, and correct species identifications would likely only come from a very small proportion of reads, which would be insufficient for most biodiversity purposes. These results suggest that, currently, non-PCR-based methods are not

and mitigate several issues with both our amplicons and our reference set. Initially, we used both the MarVer1 and 3 amplicons for marine mammal eDNA metabarcoding. However, when baselining this approach with dolphin eDNA, we found that the most abundant species was erroneously identified as *Stenella attenuata*. The correct species, *Tursiops truncatus*, was the second most abundant organism identified. After confirming with the aquarium and their detailed genealogical records that the dolphins at the National Aquarium are indeed all *Tursiops truncatus*, we were able to identify a problematic region in our *Stenella attenuata* reference genome which appeared to be overly represented in our sequencing data, specifically for MarVer1 (Fig. 5). After masking this stretch of the genome, we were able to mitigate this classification issue. Identification of this reference problem would have taken much longer if it were not for the resources provided by the National Aquarium. This issue also speaks to the need for better reference cleaning across NCBI's GenBank, as in recent years, many complete reference genomes have been found to have nontrivial amounts of contamination, leading to an increased likelihood of misclassifications, especially by K-mer-based classifiers such as Kraken2 (Steinegger and Salzberg 2020). We also found that the MarVer3 region was not specific to the species level (Fig. 6), and thus, time and resources can be saved by sequencing only the MarVer1 amplicon.

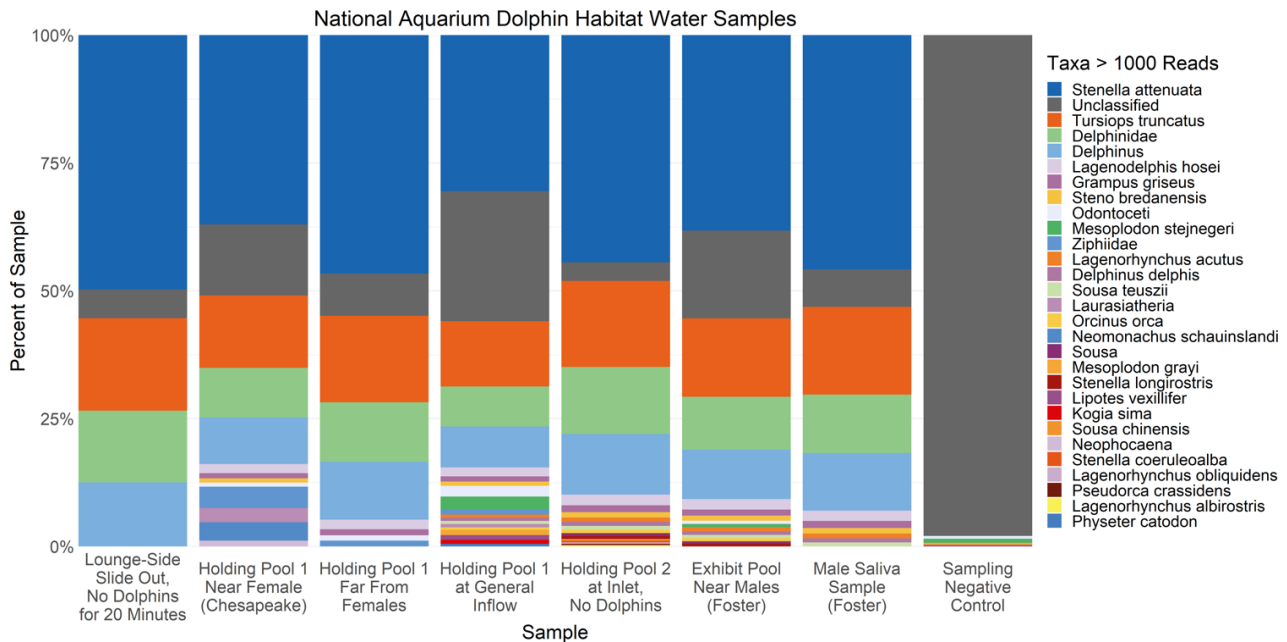


Figure 5. National Aquarium eDNA classifications. When examining the entire MarVer fragment (12S-16S) using a custom Kraken2 database with increased marine mammal mitochondrial genomes representation, we observed the majority of species-level calls classified to the closely related *Stenella attenuata*, with the correct *Tursiops truncatus* as the second most abundant species call. These results highlight the need for increased resolution in marine mammal mitochondrial markers for dolphin species.

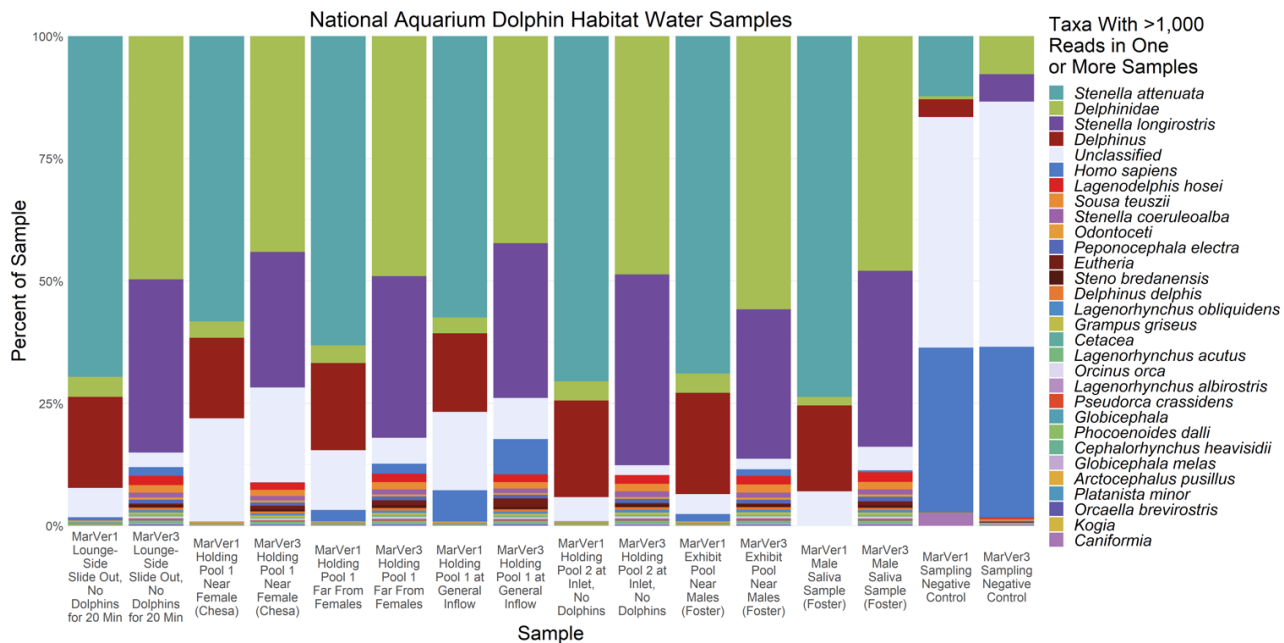


Figure 6. MarVer1/3 comparisons. When looking at individual, shorter MarVer1/3 segments, it is clear that MarVer3 is less capable of resolving dolphin species with Kraken2 classification methods, while MarVer1 produces a much higher abundance of species-level calls, even at the incorrect species. These results have informed MarVer1 as a more suitable short fragment marine mammal eDNA marker.

Refinement of genomic classification pipeline for marine mammal eDNA

The issues elucidated by the National Aquarium samples prompted an investigation into the standard genomic references used for organism classification and their performance with Kraken2 K-mer classification. We compared the classification outcomes of using a single reference per species database—comprised of complete mitogenomic references from NCBI’s RefSeq plus additional mitochondrial references provided by collaborator Phil Morin—to a multiple references per species database (GenBank), which included many partial, rather than complete, reference sequences. We found that the multiple (and partial) reference approach resulted in more accurate classifications for species which lack complete assembled mitogenomes or have high intra-species diversity, while the single, complete reference approach was better for increasing specificity and precision of the classifications. Specifically, GenBank as the reference database for Kraken2 resulted in more family-level calls, especially for Delphinidae, as well as more *Stenella attenuata* false positives, as we observed with the aquarium samples (Fig. 7). These results highlight the need for further examination into references used for marine mammal classifications at the species level, especially within Delphinidae.

Area 1 Conclusion

In support of the overall goals associated with advancing marine mammal species identification using Oxford Nanopore, we have made significant progress in developing novel eDNA methods and employing these methods on a wide array of marine mammal samples, controlled eDNA samples, and field-collected samples. While there is still much to be learned about limits of detection and classification abilities in marine mammals from eDNA samples, we believe our current methods are as well-suited to understanding marine mammals in eDNA as other laboratory-bound methods.

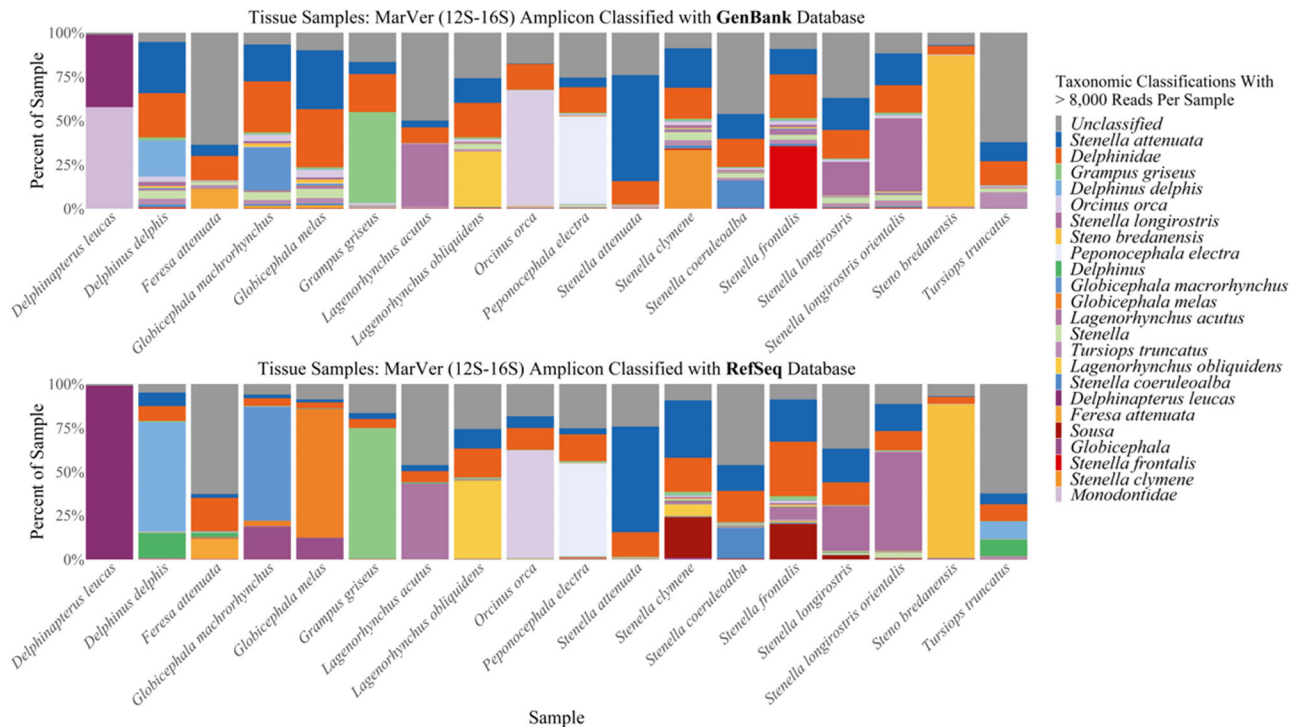


Figure 7. Comparison of marine mammal classifications with different databases. To aid in resolving species-level classifications of dolphin species, we examined using a custom RefSeq mitochondrial database and a larger, less-specific GenBank database with Kraken2 K-mer based analysis. A custom, marine mammal focused database produced much higher overall abundances at the species level for dolphin species, while the GenBank database had much higher taxonomic classifications (e.g. Delphinidae).

Area 2: Advance data processing and visualization to enable in-field eDNA analysis

Several opportunities for field-based sequencing efforts were conducted in support of this research. As discussed in our previous annual report, from April 6-15th, 2022, the JHU/APL team traveled to Monterey, CA to visit with collaborators at the Monterey Bay Aquarium Research Institute (MBARI) and conduct shipboard sequencing as a first demonstration of field sequencing capabilities. This visit also served to collaborate with expert eDNA scientists with similar interests in deployable eDNA sequencing technologies, such as the Environmental Sample Processor (ESP) platform, which can collect and preserve samples aboard a long-range autonomous underwater vehicle. MBARI has a long history as a pioneer in eDNA research, and collaborating with them has allowed us to observe their methods, advise on our own approaches, and collect valuable samples to demonstrate the ability to rapidly and reliably sequence marine mammal eDNA. Initial results were further investigated in this second year of research, and further information on the hardware and software updates are discussed below, along with increased analysis of data generated during this trip. A full manuscript on this effort was also drafted and has planned submission to the journal Environmental DNA in July/August 2023.

Development of new eDNA sample collection and analysis hardware

Aboard the *R/V Rachel Carson*, the Niskin Rosette-collected samples were sequenced using the JHU/APL developed pelican-case slotted with the MinION Mk1B and NVIDIA GPU compute board (Jetson Xavier 4 GB model) (Fig. 8). Users interface with the GPU via a touch screen, keyboard, and interchangeable standard mouse and candy bar keyboard with a trackpad. Samples were sequenced using the ARM-based deployment of the MinKNOW software, distributed by Oxford Nanopore. Basecalling was run via CUDA installed on the embedded compute board and MinKNOW's guppy deployment (GPU-based). The raw sequencing files were imported into JHU/APL's Basestack for use with MyTax

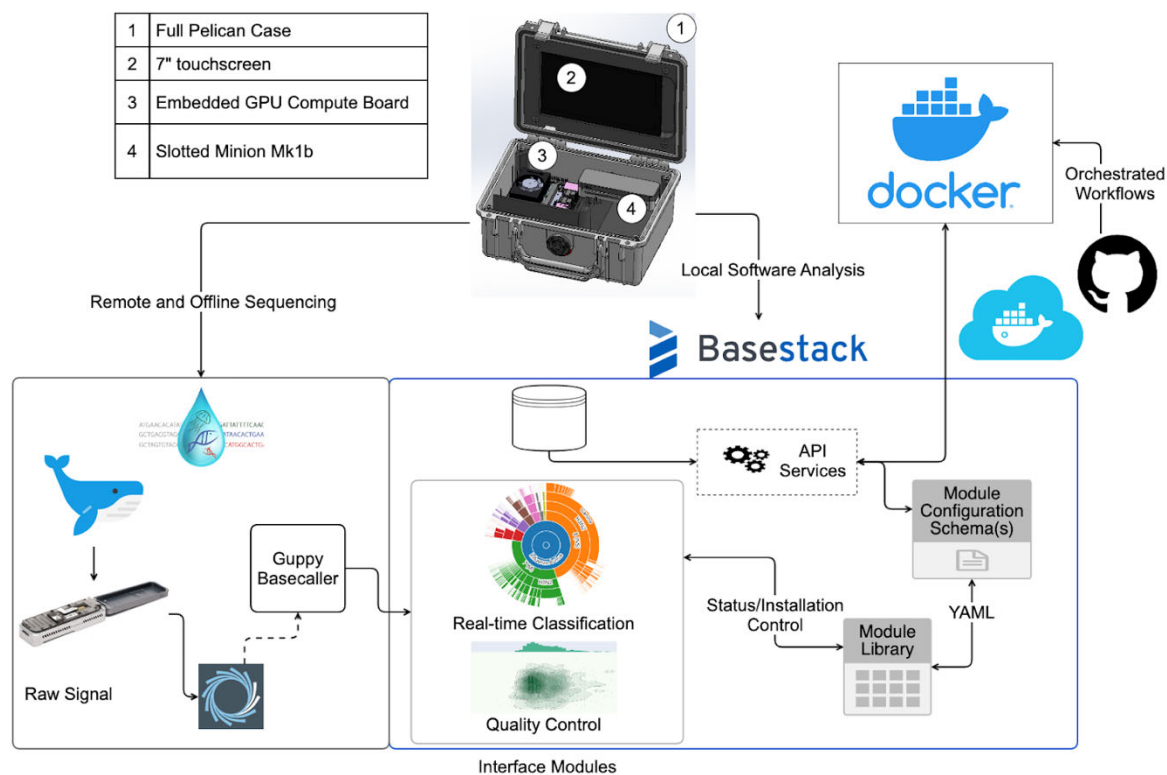


Figure 8. Hardware and Software Architecture for Sample Preparation through Sequencing Analysis. The overall architecture of the sequencing pelican case relies on NVIDIA compute board, Docker, and local software analysis using the JHU/APL-developed Basestack platform.

version 2, a dashboard for real-time metagenomics classifications using Kraken2. This updated software is now capable of analyzing multiple samples at a time and can watch demux directories, updating sample taxonomic analysis as more fastqs are generated. Results were examined via a variety of push-button, marine metagenomic-friendly plots integrated into MyTax’s analysis and visualization module. This module has also been updated to include pictures and common names for all marine mammal classifications found in a sample, in addition to the standard scientific name provided for all classifications, making marine mammal eDNA analysis accessible to a wider audience (Fig. 9).

Field-based eDNA sequencing demonstrations

After the field test aboard the *R/V Rachel Carson*, the Oxford Nanopore sequencer was restarted back on dry land and sequencing of the Niskin Rosette-collected samples was allowed to continue until the flowcell was spent (~24 hrs). While shipboard, basecalling was only performed in FAST mode—to enable rapid analysis—the complete run was basecalled in both FAST and high accuracy (HAC) modes. Finally, the Rosette samples were brought back to JHU/APL and were sequenced on an Illumina MiSeq for comparison.

Bioinformatics analysis was performed via Kraken2 using the CaleDNA 18S reference set as the database for classification. While the SILVA 18S reference set was also tested for bioinformatics analysis of these samples, though comprised of a much larger number of reference sequences, we found use of this database resulted in far more unclassified reads as compared to CaleDNA. This could be due to the CaleDNA references being specific to the region in which we sampled (California) as compared to the much more widespread SILVA references. We also investigated building a Kraken2 18S reference database from the PR2 reference set, as this is the largest repository of 18S reference sequences, and

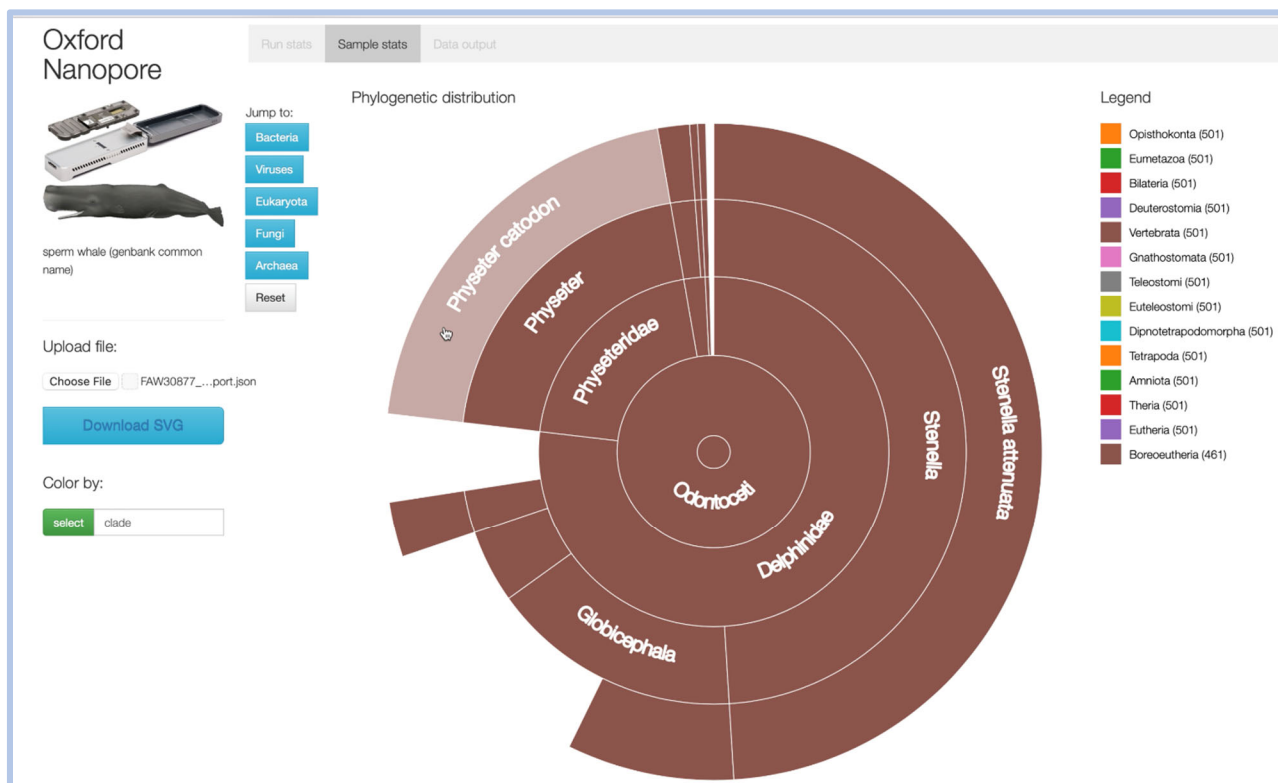


Figure 9. MyTax updates for marine mammal classification. Updates to the MyTax visualization module have resulted in common names and photos of marine mammals added to aid non-experts in visualizing their sequencing data, without having to examine cumbersome scientific names.

marine organisms make up a large portion of its references. However, building this database proved to be extremely computationally expensive and could not be completed within a reasonable amount of time (neither wall nor CPU time) on the JHU/APL compute cluster, much less the eDNA analysis case. Thus, targeting an efficient build of PR2 could be beneficial for future marine eDNA analysis work.

For all samples except 60 m, the shipboard FAST sequencing and basecalling of the samples exhibited slightly higher levels of reads unable to be classified than the high accuracy-basecalled complete ONT runs. However, for both runs, the vast majority of the reads were able to be classified, and often at least half of the reads were classified down to the family-level or better. This indicates that the shipboard sequencing (run for under an hour) with FAST basecalling is a fairly reliable proxy for the complete run with HAC basecalling. This finding validates the utility of eDNA for rapid, in-the-field marine monitoring.

Interestingly, for all samples, Illumina sequencing resulted in the highest number of unclassified reads (Fig. 10). This unexpected result is possibly due not to errors in the Illumina sequencing, but rather, sample degradation due to the time latency between collection and sequencing. While the Nanopore samples were all sequenced the same day they were collected, the Illumina samples were brought back to Maryland from Monterey and were prepped and sequenced many days later. Additionally, a long-read advantage can be ruled out, as each Illumina read pair covers the entire 18S amplicon.

Interestingly, the number of unclassified reads increased with sample depth (Fig. 10). As the sample quality was equivalent for all samples, this trend could be due to a decrease in references for organisms at greater depths. Thus, future efforts to sequence and assemble references for metabarcoding of more deep-water organisms could be invaluable for more informative for ecosystem and biodiversity monitoring via eDNA.

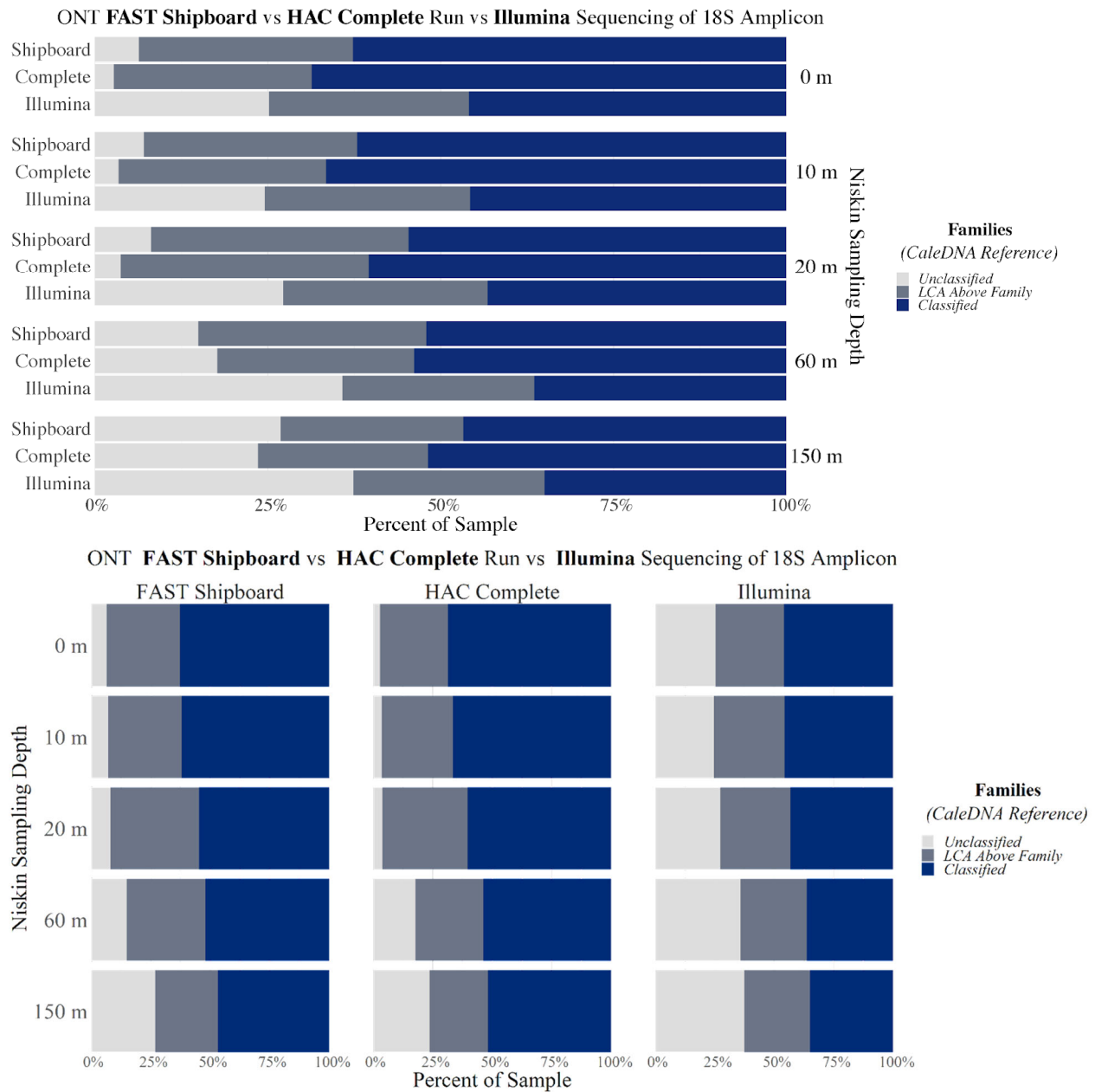


Figure 10. General classification abundances of shipboard, land-based, and lab-based sequencing methods. As Kraken2 outputs included all classifications at all taxonomic levels provided in the database, data were parsed to present all classifications at the family level and below, classifications at a lowest common ancestor (LCA) above the family level, and complete unclassified data.

Additionally, across the sample classifications of many organisms corresponded to known depth profile distributions in Monterey Bay. Calanoid copepod classifications in both Oxford and Illumina data showed ecologically relevant distributions. Surface dwelling *Paracalanus* were identified in the surface, 10, and 20 m samples, and were either not found or had much lower read counts in 60 m and 150 m

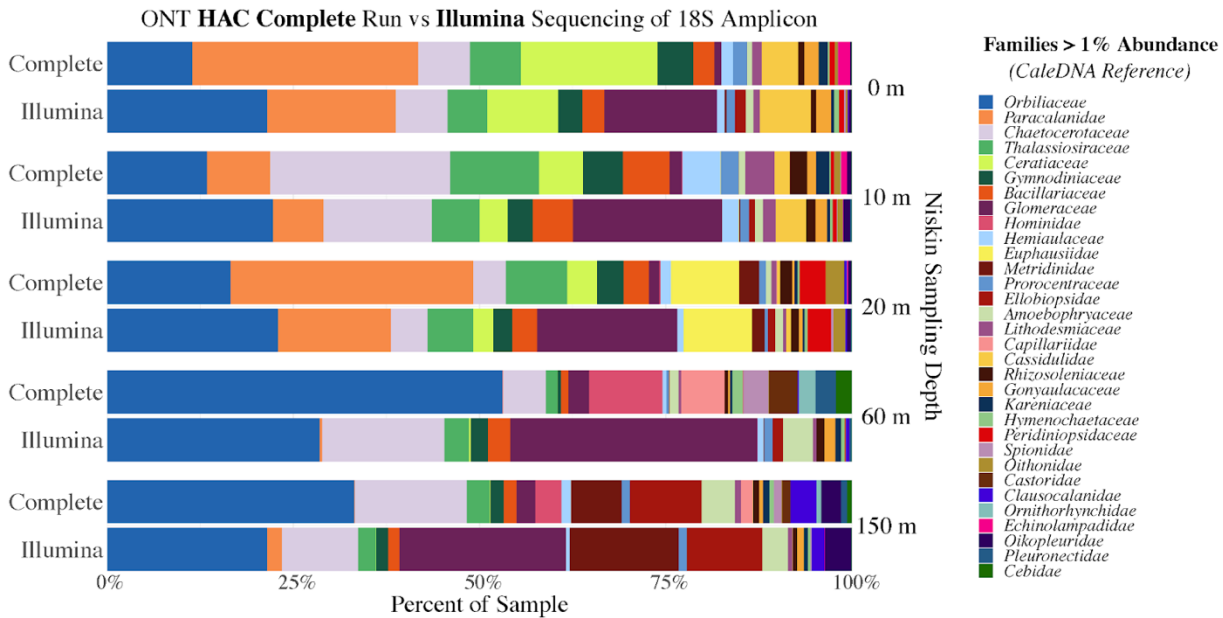


Figure 11. Comparison of Oxford and Illumina data. The complete, 17-hour Oxford sequencing data using the high accuracy (HAC) basecalling algorithm family-level classifications in comparison with the comparable Illumina sample.

samples. *Metridia longa* (likely mis-identified as *Metridia pacifica*), a vertically migrating species, were only identified in the 20 m and 150 m samples. Higher abundances of *Chaetoceros* diatoms and *Dinophyceae* dinoflagellates were also represented in surface to 20 m samples. *Siphonophorae* were identified across the 20 m-150 m, and the cnidarian *Oikopleuridae* were identified throughout samples, with highest read counts at 150 m. Additionally, identification of *Euphausiidae*, *Chaetoceros*, and *Pseudo-nitzschia* are all consistent with known distributions of these organisms within the Monterey Bay. The higher classifications in the Oxford samples appears to originate from the overall higher relative abundances of some of the most common taxa, including *Paracalanus*, *Chaetoceros*, *Dinophyceae*, and *Tripos* in near surface samples, while the 60 m and 150 m samples appeared to have higher classifications of fungi compared to Illumina data (Fig. 11).

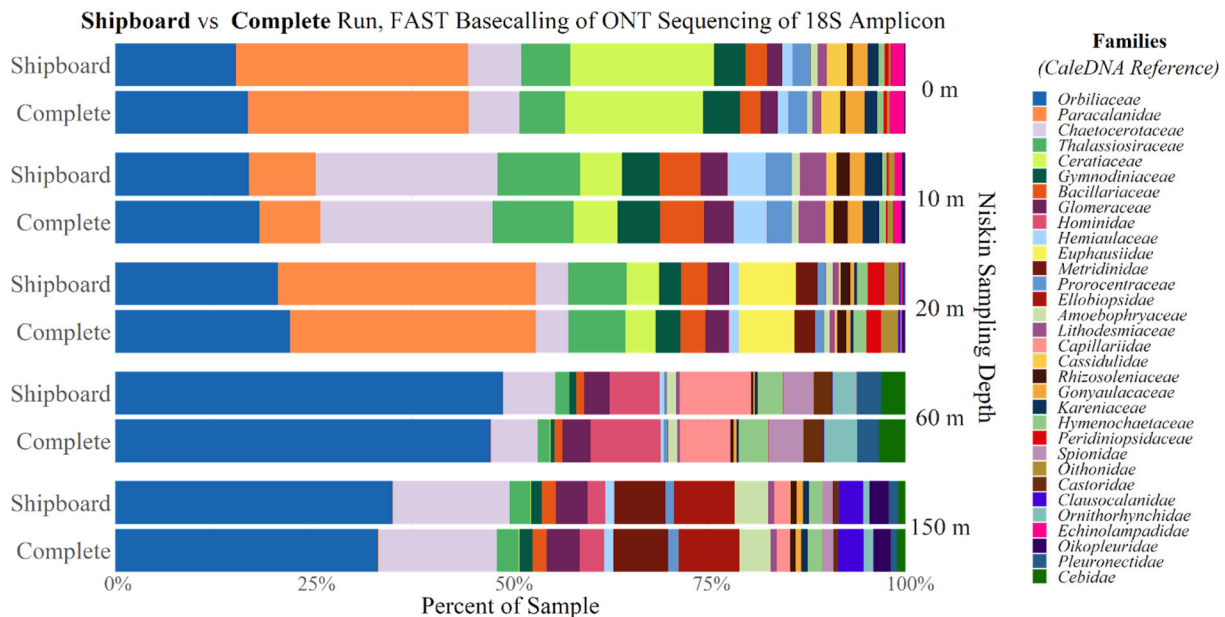


Figure 12. Comparison of shipboard and complete Oxford data. Comparison of family-level classification abundances of short (~one hour) and complete (~17 hours) Oxford data using fast basecalling.

Looking at the classifications for the shipboard vs complete ONT run, we again see that the shipboard subset very closely reflects the complete run, again reaffirming that rapid eDNA analysis can be used to reliably characterize organism presence and relative abundance in the field (Fig. 12).

Overall diversity estimates for the Monterey samples were consistent with previous 18S diversity estimates (Truelove et al. 2022) across all sequencing platforms, basecalling methods, and sequencing times when incorporating all raw read counts (Fig. 13). Expectedly, diversity decreased when a 1% abundance threshold was applied. Additionally, the Illumina data had slightly elevated diversity estimates compared to all Oxford data types, but there was a lack of statistical support for differences across any comparison made using either all raw read count diversity or 1% abundance cutoff diversity.

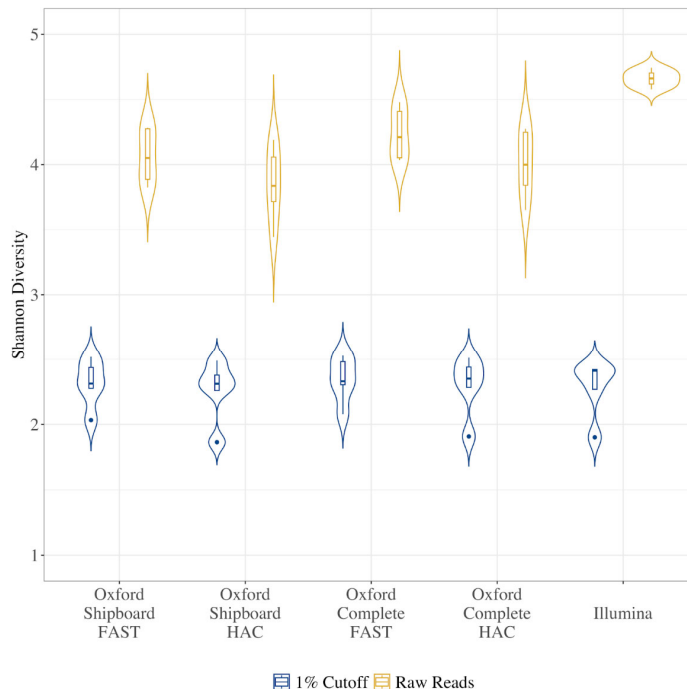


Figure 13. Comparison of shipboard and complete Oxford data. Comparison of family-level classification abundances of short (~one hour) and complete (~17 hours) Oxford data using fast basecalling.

Aim 2 Conclusion

Over the course of this research program, we have built several advanced methods to characterize eDNA in the field, which is critical to wide-scale adoption of eDNA as a monitoring tool. This work has involved advancements in hardware and software that has culminated in a wide body of products and research that can be leveraged in the future for additional optimization and transitions to additional researchers and future stakeholders in eDNA research and implementation.

Impacts and Applications

This research, data, and products produced, have had significant and wide-reaching impact across the entire field of eDNA research and marine mammal genomic field. In addition, this research has also had impacts on additional ONR eDNA programs, including the automation of eDNA sequencing projects (Platform Development for Autonomous DNA Sequencing, ONR Code 32 Navy). The summaries of impacts and applications are highlighted below:

Presentations

This research has helped support multiple internal and external presentations at an array of conferences and events, including the USGS 5th and 6th Annual Environmental DNA Technical Exchange Workshop and the 2nd National Workshop on Marine eDNA. A full list of presentations and content is below:

2021

- **University of Essex Seminar Series**
 - **Talk:** Transitioning Advances in Molecular Epidemiology to Ocean Observing: Lessons from COVID-19

- Presenter: Peter Thielen
- **5th annual Environmental DNA Technical Exchange Workshop**
 - **Talk:** Developing reliable analysis tools to enable field-portable eDNA classification
 - Presenter: Kristina K. Zudock
 - **Talk:** Standardizing time to filtration: on site filtration of aquatic eDNA samples
 - Presenter: Mark Gasser
 - **Poster Lightning Talk:** Marine mammal genetic resources, gaps, and a path towards real-time eDNA analysis
 - Presenter: Hayley M. DeHart
- **NOAA Emerging Technologies Workshop**
 - **Talk:** Developing a Fully Autonomous eDNA Sequencing Capability
 - Presenter: Peter Thielen

2022

- **6th annual Environmental DNA Technical Exchange Workshop**
 - **Talk:** Validation and refinement of rapid eDNA sequencing methods and analyses for marine mammal observation networks in polar oceans
 - Presenters: Hayley M. DeHart, Kristina K. Zudock
- **2nd National Workshop on Marine Environmental DNA**
 - **Talk:** The Path Towards *in situ* eDNA Analysis and its Applications
 - Presenters: Francisco Chavez (MBARI), Peter Thielen
 - **Poster:** Building trustworthy bioinformatics pipelines for real-time marine mammal eDNA characterization in low-resource environments
 - Presenter: Kristina K. Zudock
 - **Poster:** Establishing field-capable eDNA workflows for marine mammal characterization
 - Hayley M. DeHart
- **2021-2022 ASPIRE Showcase (JHU/APL high school intern poster session)**
 - **Poster:** Effects of ecological factors on environmental DNA (eDNA) sampling
 - Presenters: Dino Casciani, Supriya Shastri, Sophie Shimeall
- **2022 ASPIRE Summer Showcase (JHU/APL high school intern poster session)**
 - **Poster:** Comparison of DNA extraction methods for environmental DNA (eDNA) research
 - Presenters: Olive Lara, William Ross, Christopher Li
 - **Poster:** Genomic curation for improved marine mammal eDNA classification
 - Presenters: William Ross, Christopher Li, Olive Lara
- **University of Washington MURI Seminar Series**
 - **Talk:** eDNA Research, Development, and Engineering at JHU/APL
 - Presenter: Peter Thielen

2023

- **13th IEEE Integrated STEM Education Conference**
 - **Poster: Genomic Curation for Improved Marine Mammal eDNA Classification**
 - Presenters: William Ross, Christopher Li, Olive Lara
- **2022-2023 ASPIRE Showcase (JHU/APL high school intern poster session)**
 - **Poster:** Making bioinformatics marine mammal eDNA analysis accessible for all
 - Presenters: William Ross, Christopher Li

- **Poster:** Polar ocean biological monitoring and biodiversity with a comparison of DNA sequencing methods
 - Presenter: Olive Lara

Publications

DeHart HM, Gasser MT, Dixon J, Thielen P. 2023. An aquatic environmental DNA filtration system to maximize recovery potential and promote filtration approach standardization. *PeerJ*. 11:e15360. <https://doi.org/10.7717/peerj.15360>.

Thompson LR and Thielen P. 2023. Decoding dissolved information: environmental DNA sequencing at global scale to monitor a changing ocean. *Current Opinion in Biotechnology*. 81:102936. <https://doi.org/10.1016/j.copbio.2023.102936>.

Kelly RP, Lodge DM, Lee KN, Theroux S, Sepulveda AJ, Scholin C A, Craine JM, Andruszkiewicz Allan E, Nichols KM, Parsons KM, Goodwin KD, Gold Z, Chavez FP, Noble RT, Abbott CL, Baerwald MR, Naaum AM, Thielen PM, Simons AL, ... Weisberg SB (2023). Toward a national eDNA strategy for the United States. *Environmental DNA*, 00, 1– 10. <https://doi.org/10.1002/edn3.432>

Stein ED, Jerde CL, Allan EA, Sepulveda AJ, Abbot CL, Baerwald MR, Darling J, Goodwin KD, Meyer RS, Timmers M, Thielen PM. Critical Considerations for Communicating Environmental DNA Science. *In review at Environmental DNA*

DeHart HM, Truelove K, Merritt B, Zudock K, Cheezum W, Chavez F, Thielen P. *In Progr*. Rapid field-forward eDNA methods for sequencing and taxonomic assignment in marine and freshwater environments. *Environmental DNA*. *Target submission August 2023*.

Engagements

We have broadly discussed our work with both government, private, and commercial organizations. In many of these instances, our work has been referenced as an alternative approach to identify ecological change in open ocean waters that we have described in this document.

NOAA: We have collaborated with NOAA to assess current state and improvements in marine mammal genetic resources and analysis methods. We have worked with sequencing data provided to classify beaked whales. Collaborators include Phil Morin, Kelly Goodwin, and Kim Parsons.

Marine Mammal Commission: We discussed with the MMC to discuss marine mammal health and classify sequencing data from a mass stranding of Irrawaddy dolphins (*Orcaella brevirostris*). Discussions related to marine mammal health have since expanded, and we expect future engagements.

eDNA Workshop Organization: In attendance to the 2nd National Workshop on Marine eDNA, we began discussions with the steering committee and previous organizers to host or co-host the next workshop. We are now actively engaging to host the 3rd National Workshop on Marine eDNA in June 2024 in collaboration with the Smithsonian Institution National Museum for Natural History.

Capitol Hill Ocean Week: We attended Capitol Hill Ocean Week (CHOW) in June 2023 to engage with marine stakeholders and eDNA practitioners. This event also involved participation on a panel associated with eDNA, and led to follow up by organizations including NOAA, Scripps, USGS, and the Marine Mammal Commission, among others.

Monterey Bay Aquarium Research Institute (MBARI): As described in this document, we have visited MBARI to test our systems and have worked closely with them to analyze resulting data. They have become an excellent collaborator and have openly invited us to work with them as we prove out our newest technologies.

University of Washington (UW): Our team both hosted UW and presented findings outlined in this document to the University of Washington group recently awarded a Multi-University Research Initiative (MURI) to study marine mammal eDNA concepts. These engagements have resulted in excellent knowledge sharing across efforts, resulting in new insights regarding our collective methodological strengths and weaknesses.

Transitions

We have transitioned several aspects of this work to the community as well as to other projects within our own suite of eDNA work. Our environmental sampling pump (highlighted in our previous annual report) was published as an open-source piece of hardware that anyone can build and use, and we have provided multiple prototypes to collaborators in the field for use and feedback on future optimization. We have engaged with commercialization partners to produce versions of these eDNA collection pumps, and have additionally released the design as an open-hardware platform via the JHU/APL Office of Technology Transfer. Additionally, this research facilitated initial work with Lindblad Expeditions that allowed for sponsored research to build eDNA capabilities onboard Lindblad Expedition ships. These resources can now be utilized for future research, which will provide invaluable access to field collected samples and processing that is otherwise unavailable. Additionally, access to samples containing marine mammal eDNA can help future work further characterize eDNA for marine mammals specifically. We were able to successfully transition work initially developed under this research effort to multiple collaborators and users in the form of best practices and protocols that were further optimized over the course of this research.

Interns

This work would not have been possible without the support of dedicated interns, and this research supported the efforts of six high school interns and two college interns, who were able to provide critical contributions in the testing of molecular methods and analysis and development of software.

Conclusions

The work presented here, in addition to the wide range of engagements, publications, and collaborations, resulted in a substantial advancement of eDNA for marine mammal detection and classification, with the ultimate goals of creating field capable methods and technologies that can be leveraged by future work in this domain. We believe there is a substantial amount of progress made in the creation of hardware and software for field-based eDNA sequencing, and further research is needed in the classification abilities and confidence in marine mammal eDNA samples. Future work in these areas should be focused in the ground-truthing of marine mammal classifications, optimization of marine mammal-specific eDNA sequencing, and expansion of confidence metrics associated with sequencing data.

Acknowledgements

We would like to acknowledge the National Aquarium in Baltimore (Sylvia Rickett and Kimmy Barron), the Smithsonian Institution National Museum of Natural History (Chris Meyer, Michael McGowen), Monterey Bay Aquarium Research Institute (MBARI; Chris Scholin, Kobun Truelove, Francisco Chavez), National Geographic Pristine Seas program (Molly Timmers), National Oceanographic and

Atmospheric Organization (NOAA; Phil Morin), University of Washington (Ryan Kelly, Eily Allen), and Lindblad Expeditions (Ashley Knight). Special thanks to Rebecca Pugh and Deb Ellisor for their guidance on permitting to enable our receipt of tissue samples from the NOAA/NIST Marine Mammal Tissue Bank collection. We would also like to acknowledge the efforts of Tessa Byrum, Andrew Tang, William Ross III, Christopher Li, Dino Casciani, Supriya Shastry, Sophie Shmeall, and Olive Lara.

References

Chen Z, Erickson DL, Meng J. 2020. Benchmarking Long-Read Assemblers for Genomic Analyses of Bacterial Pathogens Using Oxford Nanopore Sequencing. *Int Journal of Mol Sci.* 21(23):9161. <https://doi.org/10.3390/ijms21239161>.

Koren S, Schatz M, Walen, B, et al. 2012. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol.* 30:693–700. <https://doi.org/10.1038/nbt.2280>.

Steinegger M, Salzberg SL. 2020. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* 21:115. <https://doi.org/10.1186/s13059-020-02023-1>.

Truelove NK, Patin NV, Min M, Pitz KJ, Preston CM, Yamahara KM, Zhang Y, Raanan BY, Kieft B, Hobson B, Thompson LR, Goodwin KD, Chavez FP. 2022. Expanding the temporal and spatial scales of environmental DNA research with autonomous sampling. *Environmental DNA*, 4, 972– 984. <https://doi.org/10.1002/edn3.299>.