Monterey Bay Aquarium
Research Institute

Prototyping a rapid bioinformatics pipeline for a miniature eDNA sequencer

Raksha Doddabele, Duke University

Mentors: Thom Maughan, Nathan Truelove, Francisco Chavez

Summer 2020

Keywords: environmental DNA (eDNA), sequence alignment, metabarcoding

ABSTRACT

Environmental DNA (eDNA) is a powerful and non-invasive tool to survey marine ecosystems. Using a rapid miniature sequencer like the Oxford Nanopore MinION, real-time analysis of the organisms present at a sampling location could be telemetered to shore. However, alignment of unknown reads to reference sequences can be a time-intensive process, and multiple alternatives to Basic Local Alignment Search Tool (BLAST) have been developed. In this study, we compared the speed and proportion of MinION eDNA reads classified to taxa of three sequence alignment programs: BLAST, Minimap2, and Kraken2. Minimap2, followed by MEGAN software for taxonomic classification, proved to be efficient while resolving a large proportion of reads to the genus level. When processed by this bioinformatics pipeline, MinION reads provided a suitable measure of taxa present at a sampling location. This pipeline utilized on MinION eDNA reads can yield a fast and accurate picture of the biodiversity in an ecosystem, paving the way for an entirely automated real-time sequencing analysis in the future, which will be useful in guiding conservation efforts.

INTRODUCTION

Marine organisms increasingly face a number of stressors, including ocean acidification and climate change (Sampaio & Rosa, 2020). Tracking and surveying marine populations is crucial to understanding the effects of these changes and guiding conservation measures. Environmental DNA, or eDNA, provides a non-invasive way to monitor marine populations and is emerging as a powerful alternative to other surveying methods, which can be costly, time-consuming, and even detrimental to the survey population (Murphy & Jenkins, 2010). Free-floating genetic material shed from an organism, including tissue, skin, and metabolic waste, can be collected from the water, amplified, and analyzed to determine the organism of origin. This makes eDNA a powerful tool for metabarcoding, or characterizing different species from a single water sample (Ruppert et al., 2019). By focusing on one variable mitochondrial gene in the eDNA samples, the different taxa present in a location can be determined. Furthermore, eDNA pairs well with automation; the Long Range Autonomous Underwater Vehicles (LRAUVs) with onboard Environmental Sample Processors (ESPs) in the Monterey Bay are able to collect and filter genetic material from the water. The next step in automation is to include a miniature DNA sequencer onboard.

The Oxford Nanopore MinION is a rapid miniature sequencer that is useful for metagenomics, but due to its real-time sequencing it is only about 95% accurate (Jain et al., 2016). This makes differentiating eDNA from individual species in similar taxa challenging. Other high-depth sequencers are more accurate but take much longer to sequence. Algorithms to improve accuracy and form consensus sequences by aligning similar sequences have been developed for the MinION, but this process can lose the diversity of similar species within genus or family. Traditionally, the Basic Local Alignment Search Tool (BLAST) algorithm is used to align reads to known sequences, but this process is slow and would use too much processing power on an onboard sequencer (Altschul et al., 1990). To place the MinION sequencer onboard the LRAUV in the future and perform real-time sequencing and taxon classification on the water, bioinformatics processing of sequences must be time and resource efficient.

In this study, we analyze the speed and proportion of sequences classified at the genus level of three alignment and taxonomic classification pipelines. We compare BLAST followed by MEGAN, which uses a lowest common ancestor algorithm to assign reads to taxa (Huson et al, 2016); Minimap2, a sequence alignment program which uses a chaining algorithm optimized for parallel processing units (Li, 2018), followed by MEGAN; and Kraken 2, which uses unique short sequences, or k-mers, to align query sequences to a database and then maps them to the lowest common ancestor for taxonomic classification (Wood et al., 2019). These three sequence analysis pipelines were compared for speed and proportion of reads classified to the genus level. This comparison will help us determine which sequence alignment program is best suited for processing MinION reads efficiently on board an LRAUV, without sacrificing the taxon diversity present in the sampling location.

MATERIALS AND METHODS

FIELD SAMPLING

Water samples of 1L were collected and filtered by an Environmental Sample Processor onboard the LRAUV. Sampling occurred at three different stations: C1 at the head of the Monterey Canyon, M1, and M2. Sampling occurred at two different depths at each station: 10 meters and 150 meters. Water samples were filtered to remove waste and then stored on the LRAUV until retrieval.

DNA EXTRACTION AND SEQUENCING

The DNA was extracted from samples once brought to land. Then, DNA samples underwent PCR amplification of a highly variable mitochondrial gene for metabarcoding. The DNA was amplified using the primer sets for either the metazoan 12S ribosomal DNA gene (Machida et al., 2012), the eukaryotic 18S ribosomal DNA gene (Amaral-Zettler et al., 2009), or the Cytochrome c Oxidase subunit I (COI) gene (Leray et al., 2013). Paired end sequencing was performed by the Oxford Nanopore MinION, which has six cores and a 256 core GPU with 8 GB of RAM. For comparison to a high-depth sequencer, the Illumina MiSeq, which is highly accurate but takes upwards of 48 hours to sequence, was used. MinION sequencing of the six samples yielded 220,876 total DNA reads at 10

meters at C1, 102,314 reads at 150 meters at C1, 247,119 reads at 10 meters at M1, 66,409 reads at 150 meters at M1, 261,493 reads at 10 meters at M2, and 124, 884 reads at 150 meters at M2.

SEQUENCE CLUSTERING AND POLISHING

Reads were clustered and polished on an Amazon Web Services (AWS) Linux server with sixteen parallel CPUs with around 113 GB of memory and one NVIDIA Tesla M60 GPU with 2,048 parallel processing cores and 8 GB of memory. The potential for multithreading across these processors allows us to accomplish time-intensive processes like genomic alignments much faster. Therefore, a bioinformatics pipeline that is optimized for parallel processing would be ideal. If required, clustering and polishing of Oxford Nanopore MinION reads were performed in an NGSpeciesID Conda environment, which clusters and forms consensus of Oxford Nanopore reads (Prost et al., 2020). Oxford Nanopore's Medaka algorithm was used to create consensus sequences from similar raw Nanopore reads, along with the Spoa algorithm to implement partial order alignment of consensus sequences. An abundance ratio of .001 was used to allow for more unique clusters of sequences.

ALIGNMENT AND TAXONOMIC CLASSIFICATION

Alignment was also performed on the AWS server. Reads were queried to a custom reference database mapping known sequences for 12S, 18S, and COI primers to NCBI accession IDs and aligned using BLAST, Minimap2, or Kraken2. The speed of each alignment program was timed and averaged for each of the six samples. BLAST was installed and run on a Docker container. Reads were mapped to reference sequences with minimum 85% identity. Minimap2 was installed through Conda and run in the NGSpeciesID environment. The parameters were set for Oxford Nanopore genomic reads. Kraken2 provided taxonomic names while BLAST and Minimap2 provided NCBI accession IDs for reads, including multiple accession IDs for certain reads. After alignment through BLAST or Minimap2, reads were processed with MEGAN software to match accession IDs to the NCBI's taxonomy database using a lowest common ancestor (LCA) algorithm, which

assigns reads to taxa based on the sequence's level of conservation (Huson et al., 2016). The assigned taxonomy output was then processed using a custom R (R Core Team, 2020) script developed in RStudio (RStudio Team, 2020) and the package phyloseq (McMurdie & Holmes, 2013).

RESULTS

RUNTIME

Figure 1 shows the average runtime of BLAST, Kraken2, and Minimap2 on six samples. Analysis was run on the 16 CPUS on the AWS server. BLAST took the greatest amount of time to align reads to references sequences, averaging 1348.55 seconds or about 22 minutes. Minimap2 averaged 170.77 seconds and Kraken2 performed alignment the fastest, averaging 1.08 seconds. Only Minimap2 and Kraken2 were fast enough to feasibly perform real-time analysis onboard an LRAUV. Furthermore, at peak memory usage, Minimap2 used a maximum of about 5 GB of memory while Kraken2 and BLAST used more memory. Note that there would be around 8 GB of memory on an embedded board placed on an LRAUV.

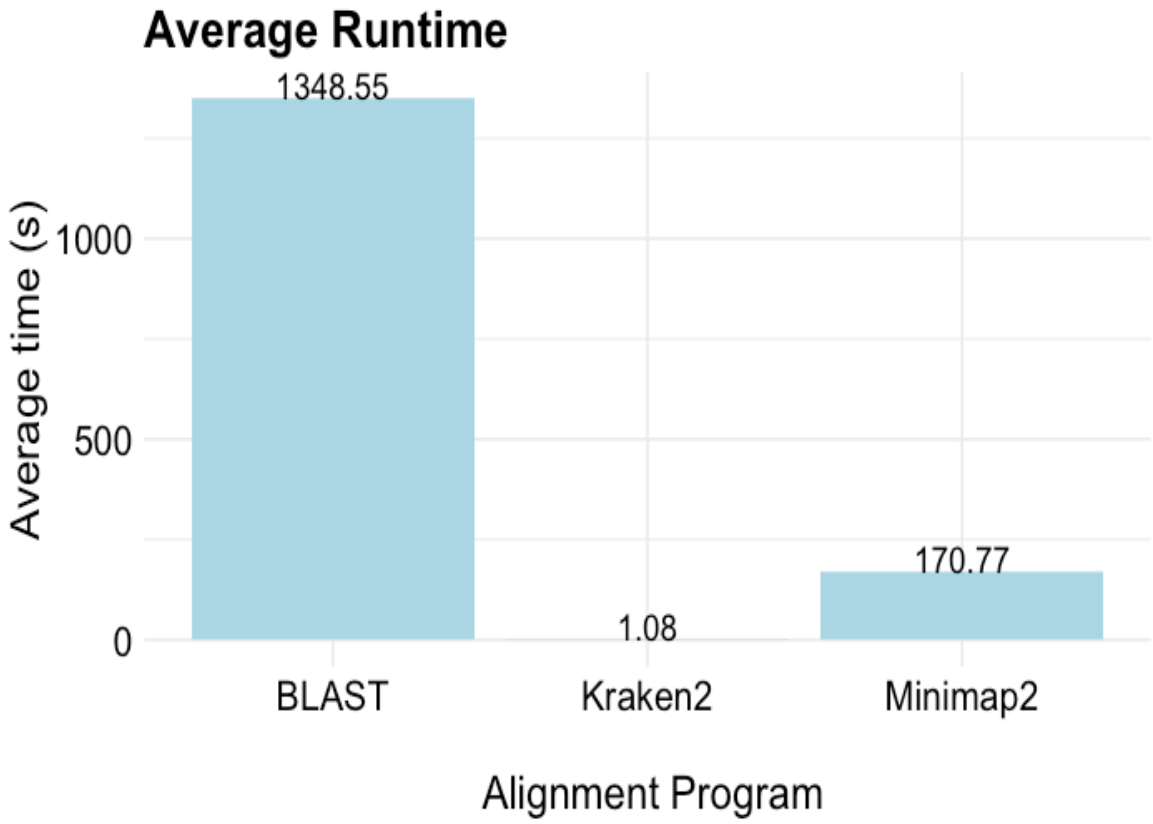

Figure 1. Comparison of the average runtime of each alignment program on six samples from C1, M1, and M2 sampling stations.

PROPORTION READS CLASSIFIED

The proportion of total reads classified to any taxon was calculated for each of the six samples when processed with BLAST followed by MEGAN, Kraken2, or minimap2 followed by MEGAN and plotted as a boxplot (Figure 2). BLAST/MEGAN yielded the lowest proportion of classified reads, averaging 0.33. Kraken2 classified an average proportion of 0.60 reads to a taxon, followed by Minimap2/MEGAN which classified an average proportion of 0.524 reads to a taxon. However, Kraken2's proportion of reads resolved to the genus level was poor, averaging 0.08 (Figure 3). Minimap2/MEGAN classified the greatest average proportion of reads to a genus at 0.34, followed by BLAST/MEGAN with an average proportion of 0.25. Therefore, the Minimap2/MEGAN pipeline can best classify similar eDNA sequences to a genus.

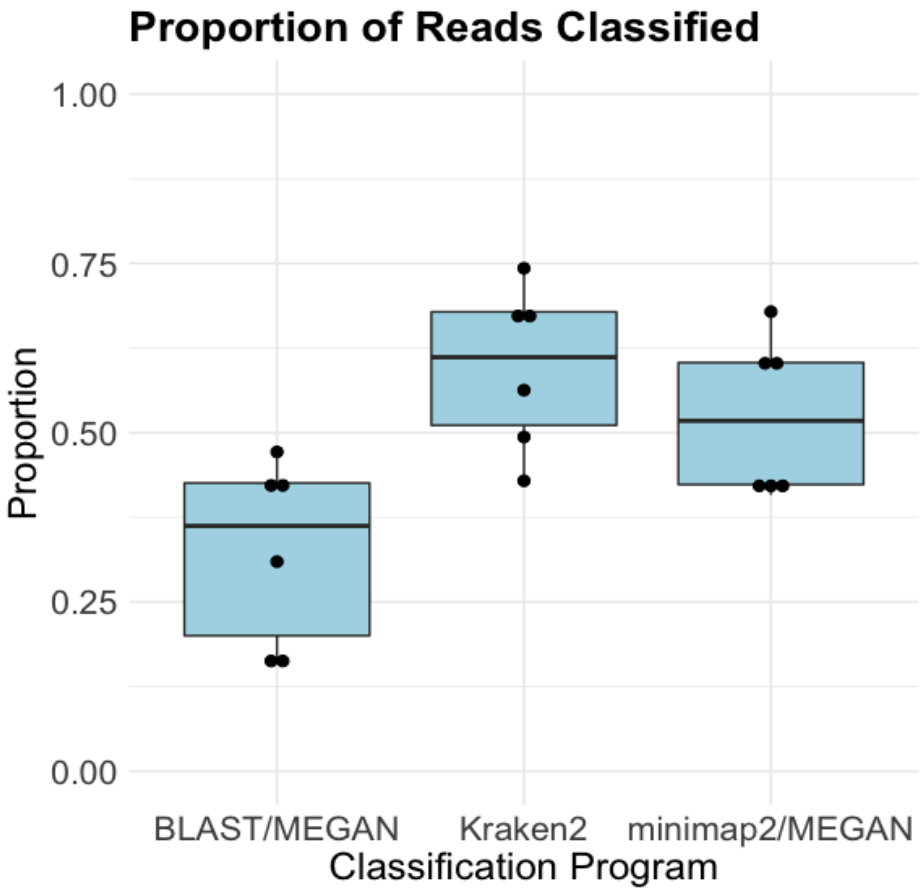

Figure 2. Boxplot comparing the distribution of proportion of reads classified by each classification program for six samples from C1, M1, and M2 sampling stations.

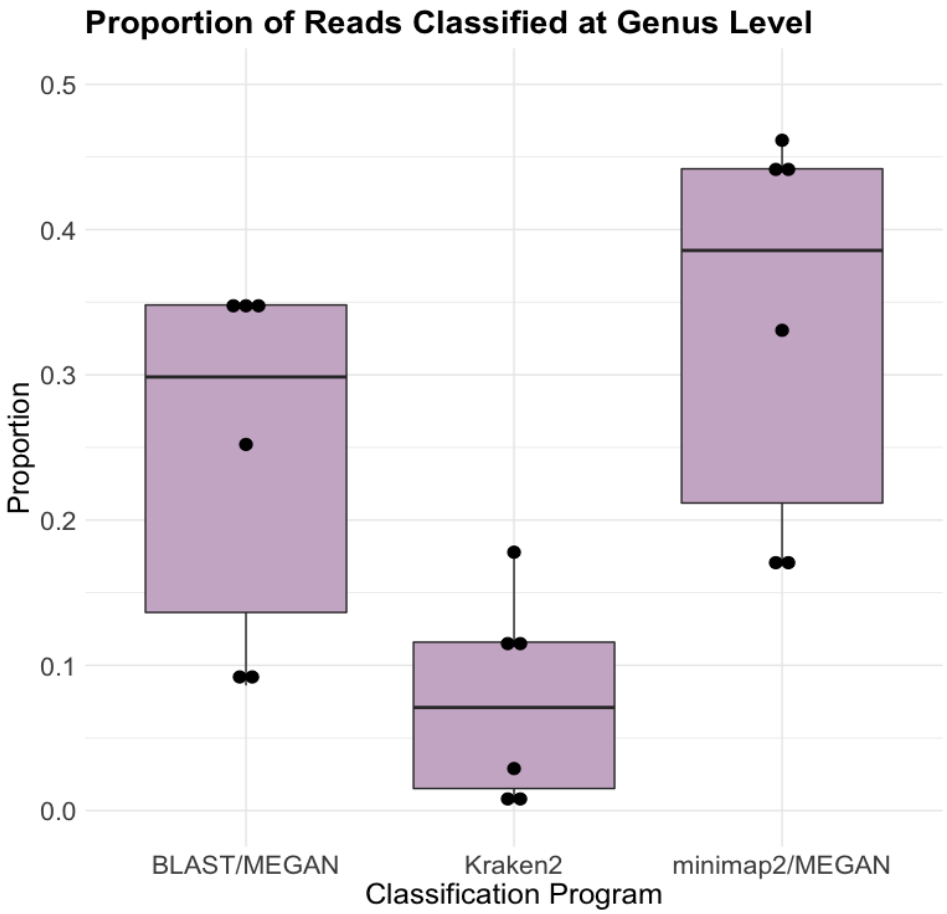

Figure 3. Boxplot comparing the distribution of proportion of reads classified to the genus level by each classification program for six samples from C1, M1, and M2 sampling stations.

GENUS DIVERSITY

The bioinformatics processing of MinION sequence reads with Minimap2 and MEGAN yielded a suitable spread of genus diversity when eDNA samples were amplified using 12S, 18S, and COI primers. The majority of reads were assigned to a genus. Figure 4 shows the thirteen most abundant genera at the C1 sampling location from 323,190 total reads. The majority of eDNA samples belonged to copepods, anchovies, and algae using this pipeline, which is predictable given the primer sets used. 0.39 proportion of the reads were unassigned to a genus. At the M1 sampling location, 313,528 total reads were processed and mostly assigned to copepods, anchovies, and algae, with a proportion of 0.27 unassigned reads (Figure 5). Notably, the presence of *Pseudo-nitzschia*, a diatom which is capable of

producing harmful algal blooms, was also observed at this location. At the M2 sampling locations, 386,377 total reads were processed and mostly assigned to copepods, anchovies, and algae, with a proportion of 0.42 unassigned reads (Figure 6). The most abundant genera we found were consistent with metazoans documented in the Monterey Bay National Marine Sanctuary (Burton & Lea, 2019). Throughout the six samples, a few striped dolphin and California sea lion reads were observed as well, showing that this tool can be used to search for the presence of marine mammals.

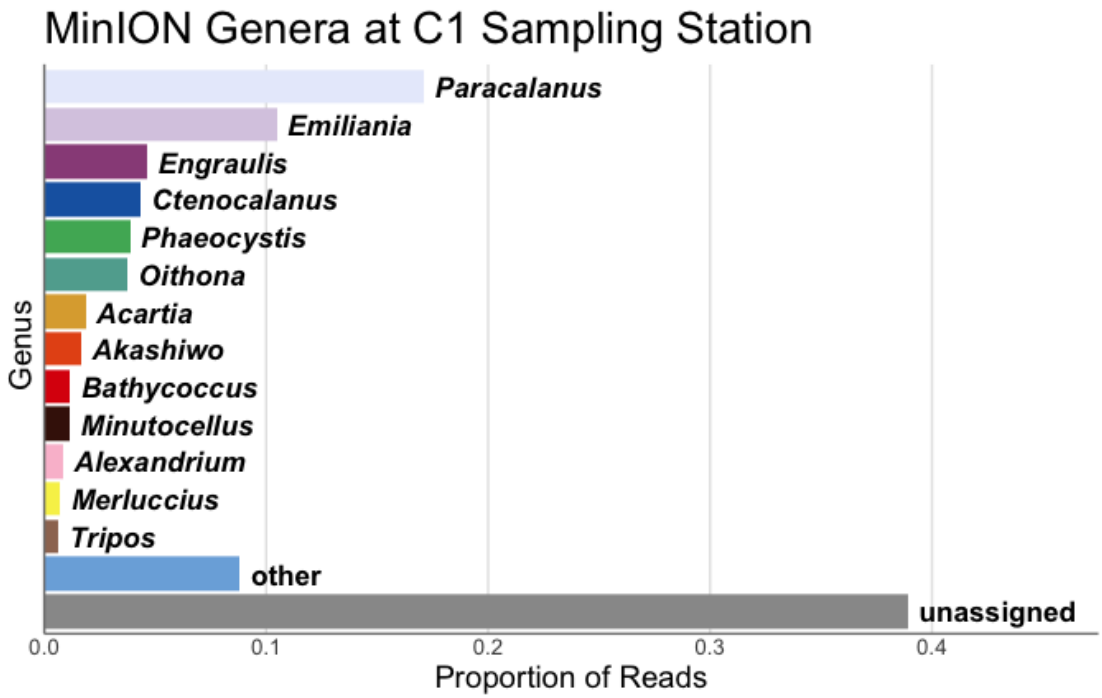

Figure 4. The proportion of MinION reads from two samples at the C1 location assigned to the 13 most abundant genera, as well as proportion of reads unassigned to a genus, using 12S, 18S, and COI primers. All other genus assignments are summed into the other category.

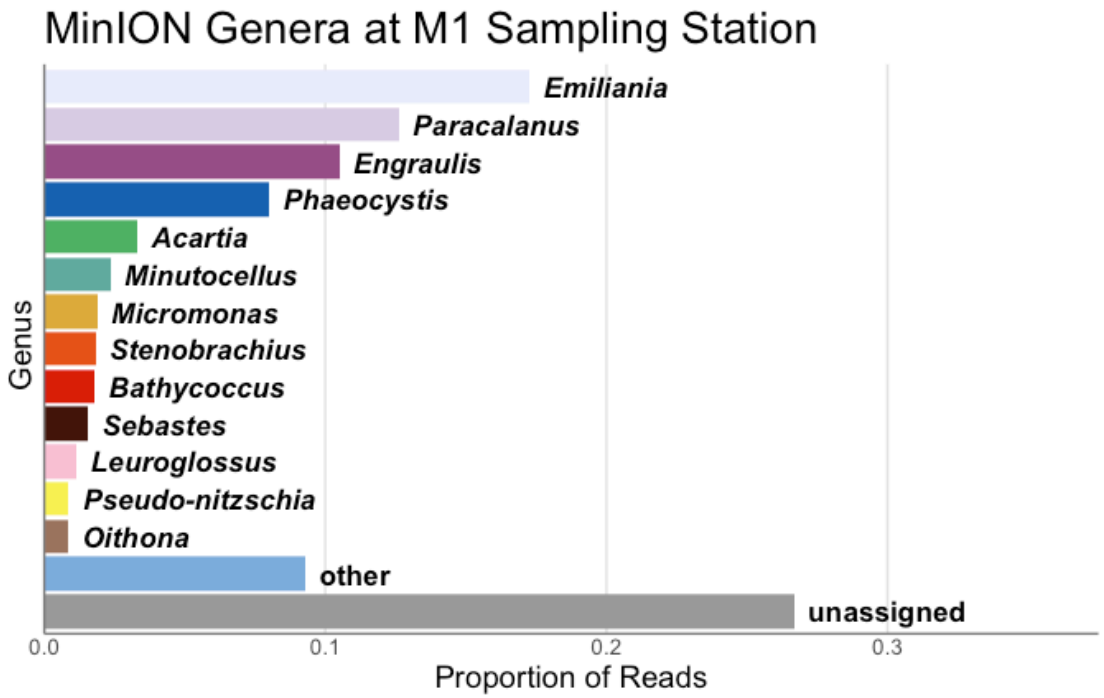

Figure 5. The proportion of MinION reads from two samples at the M1 location assigned to the 13 most abundant genera, as well as proportion of reads unassigned to a genus, using 12S, 18S, and COI primers. All other genus assignments are summed into the other category.

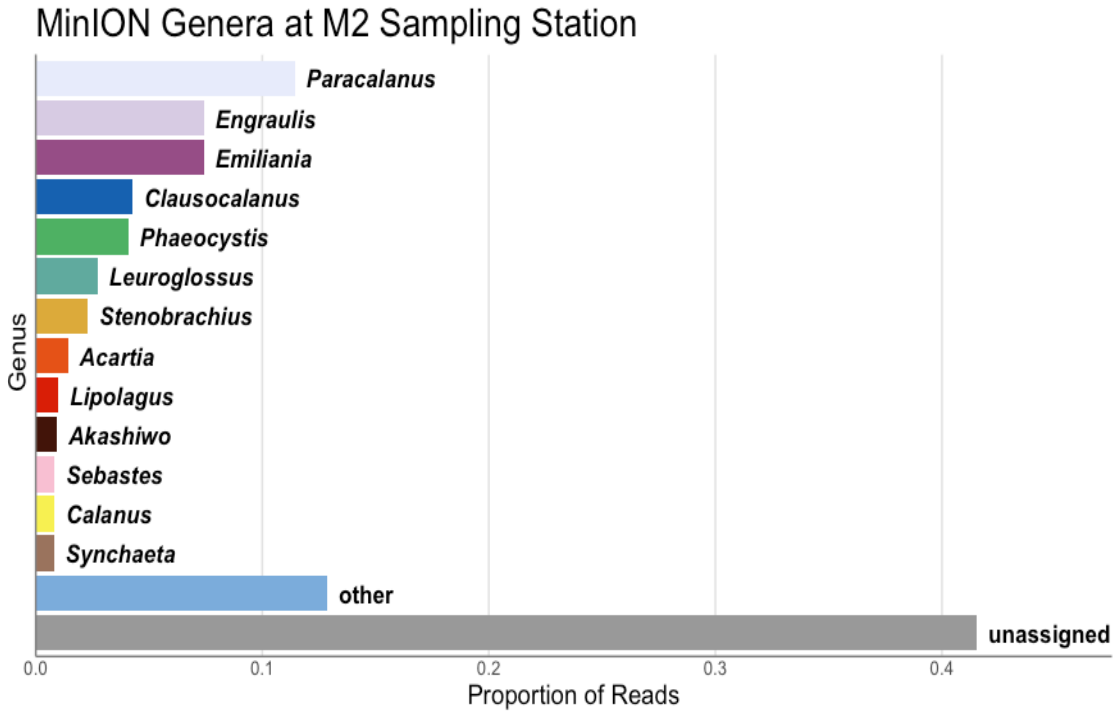

Figure 6. The proportion of MinION reads from two samples at the M2 location assigned to the 13 most abundant genera, as well as proportion of reads unassigned to a genus, using 12S, 18S, and COI primers. All other genus assignments are summed into the other category.

COMPARISON TO A HIGH-DEPTH SEQUENCER

The use of Minimap2 followed by MEGAN on 12,390,231 total reads amplified with 12S primers from the Illumina Miseq yielded a similar abundance of anchovies and rockfish as seen with the MinION reads. A proportion of 0.55 reads were unassigned to a genus, but a relatively large proportion of Mola Mola reads were observed, as well as possible human contamination. The MiSeq, while more accurate, takes much longer to sequence eDNA and gives a similar view of genus diversity as the MinION.

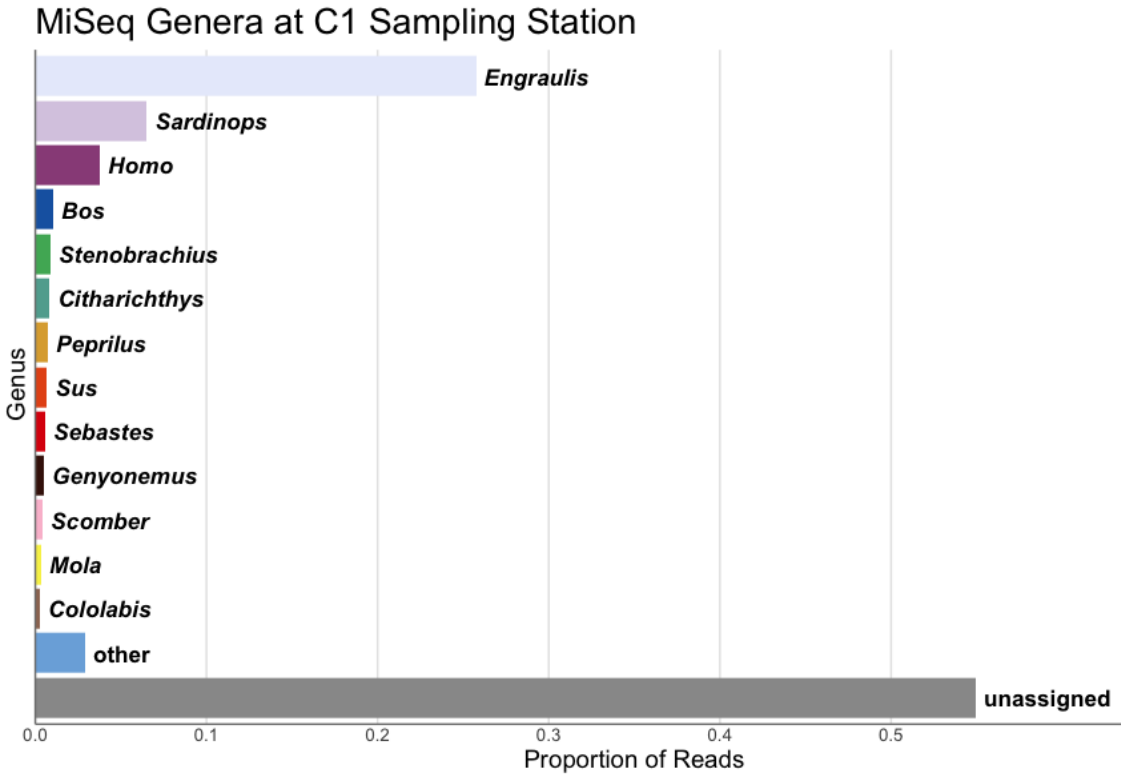

Figure 7. The proportion of Illumina Miseq reads from 120 samples at the C1 location assigned to the 13 most abundant genera, as well as proportion of reads unassigned to a genus, using only 12S metazoan primers. All other genus assignments are summed into the other category.

DISCUSSION

Minimap2 for reference sequence alignment followed by MEGAN for taxonomic classification proved to yield the highest proportion of reads classified to the genus level without sacrificing efficiency. Minimap2 also had conservative memory usage, which makes it apt to process sequencing reads in an embedded Linux system onboard an LRAUV. Minimap2's efficient detection of overlaps in inputted reads is due to its seeding and chaining algorithm, which runs much faster than other alignment programs (Li, 2018). It is also designed to handle long reads with an error rate up to 15%, which can handle the 5% error rate associated with Oxford Nanopore MinION reads. This makes Minimap2 a great tool for processing MinION sequencing results.

This bioinformatics pipeline for Oxford Nanopore MinION reads also accurately represents the genus diversity of a sampling location when compared to the deeper sequencing performed by the Illumina MiSeq, which is also more time-intensive. Minimap2's algorithms are designed to perform accurate read mapping, which explains the resolution to genus and conservation of genus diversity observed (Li, 2018). When the three primer sets, 12S, 18S, and COI, are all used on a sample, a range of animals, from microorganisms to megafauna, can be observed. We were able to observe *Pseudo-nitzschia* through this method, an organism which can produce a deadly neurotoxin in the Monterey Bay, as well as marine mammals which may be affected by this toxin. Minimap2's processing of eDNA reads from the MinION is a useful method to track and survey taxa present in the Monterey Bay, which can inform conservation efforts.

CONCLUSIONS/RECOMMENDATIONS

eDNA is rapidly becoming a powerful tool to survey populations that may be difficult to survey with traditional methods. The technology is improving fast to gain greater accuracy and taxa diversity of reads. A single water sample can yield

a sufficient idea of metazoan biodiversity in a location. This study's comparison of read mapping programs is one step closer to the goal of having the Oxford Nanopore MinION sequencer on board an LRAUV. This in-sea sequencer, with read mapping performed by the program Minimap2, can be utilized to quickly gain an idea of what organisms are present, including elusive marine mammals or invasive or toxic species. It could also be used to follow and track populations of animals without requiring us to ever leave land. Processing of eDNA reads with Minimap2 for mapping to reference sequences and MEGAN for taxonomic classification improves accuracy by forming consensus sequences without sacrificing genus diversity or efficiency. This bioinformatics pipeline serves as a powerful conservation tool to track biodiversity and individual marine populations without having to expend the time and energy to physically track and tag animals.

ACKNOWLEDGEMENTS

I would like to thank Thom Maughan, Nathan Truelove, and Francisco Chavez for their advice and support through this project. Some of the plotting code and data was provided by Markus Min. Thank you to the other members of the lab group, including my fellow interns Olivia Boerbitz and Kristina Samborski, for their feedback on this project and paper. Finally, I want to thank George Matsumoto, Megan Bassett, and Tatjana Ellis for maneuvering and supporting the first virtual MBARI internship program, as well as the David and Lucile Packard Foundation, the Dean and Helen Witter Family Fund, and the Rentschler Family Fund for funding my work.

References:

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Amaral-Zettler, L.A., McCliment, E.A., Ducklow, H.W. and Huse, S.M. (2009). A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. *PloS one, 4*(7), 6372.

Burton, E. J., & Lea, R. N. (2019). Annotated checklist of fishes from Monterey Bay National Marine Sanctuary with notes on extralimital species. *ZooKeys*, *887*, 1–119. https://doi.org/10.3897/zookeys.887.38024

Huson, D.H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., Ruscheweyh, H., & Tappu, R. (2016). MEGAN Community Edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput Biol*, *12*(6):e1004957. https://doi:10.1371/journal.pcbi.1004957

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239. https://doi.org/10.1186/s13059-016-1103-0

Leray, M., Yang, J.Y., Meyer, C.P., Mills, S.C., Agudelo, N., Ranwez, V., Boehm, J.T. and Machida, R.J. (2013). A new versatile primer set targeting a short fragment of the mitochondrial COI region for metabarcoding metazoan diversity: application for characterizing coral reef fish gut contents. *Frontiers in zoology, 10*(1), 34.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics, 34*, 3094-3100. doi:10.1093/bioinformatics/bty191

Machida, R. J., Kweskin, M., & Knowlton, N. (2012). PCR primers for metazoan mitochondrial 12S ribosomal DNA sequences. *PloS one*, *7*(4), e35887. https://doi.org/10.1371/journal.pone.0035887

McMurdie, P.J. & Holmes, S. (2013). phyloseq: An R package for reproducible interactive analysis and graphics of microbiome census data. *PLoS ONE,* 8(4):e61217

Murphy, H.M. & Jenkins, G.P. (2010). Observational methods used in marine spatial monitoring of fishes and associated habitats: a review. *Marine and Freshwater Research*, *61*(2), 236-252. https://doi.org/10.1071/MF09068

Prost, S., Sahlin, K., Lim, M. (2020). NGSpeciesID: DNA barcode and amplicon consensus generation from long-read sequencing data. *Authorea.* https://10.22541/au.160262406.62842291/v1

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston. http://www.rstudio.com/

Ruppert, K. M., Kline, R. J., & Rahman, M. S. (2019). Past, present, and future perspectives of environmental DNA (eDNA) metabarcoding: A systematic review in methods, monitoring, and applications of global eDNA. *Global Ecology and Conservation*, *17*, e00547. https://doi.org/10.1016/j.gecco.2019.e00547

Sampaio E., Rosa R. (2020) Climate Change, Multiple Stressors, and Responses of Marine Biota. In: Leal Filho W., Azul A.M., Brandli L., Özuyar P.G., Wall T. (eds) Climate Action. *Encyclopedia of the UN Sustainable Development Goals*. Springer, Cham. https://doi.org/10.1007/978-3-319-95885-9_90

Wood, D.E., Lu, J. & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biol, 20*, 257. https://doi.org/10.1186/s13059-019-1891-0