Methods and Results Final Project

Juliane Mora

4/27/2022

# Methods

### Sample Collection

Sample collection took place across 24 locations around Narragansett Bay. Samples were collected during four seasons over the span of two years. To include a more robust sampling model, four replicates per sample were collected. Conditions throughout Narragansett Bay vary due to differences in freshwater input and human population density. This sampling scheme is designed to include locations that have varied dissolved oxygen, salinity, and nitrogen levels. Keeping this in mind, this sampling data will be used to assess biodiversity levels across Narragansett Bay.

At each sampling location, water was collected in four 1-liter samples of water for a total of 768 samples. Each of the water samples were filtered through glass sterile 1 μm filters (GF, borosilicate glass) and DNA was extracted from each of these filters using a Qiagen DNeasy kit. Three amplicons of each sample were generated using the primer TELE02. This primer is specifically developed for fish eDNA metabarcoding of fresh and saltwater fish. Samples were then sequenced by an Illumina MiSeq to computationally identify species. For the purpose of this project, a subset of 10 samples collected in the winter of 2019 were further analyzed in QIIME2 via HPC command line.

### Creating a metadata file

A metadata file was used to import important information about each sample into QIIME2. Included in the metadata file was sample ID, barcode, primer adapter, location, date of sample collection, and sample name. This information was important to include because it allows for differentiation between samples during further analysis. This metadata file was created in google docs and format verified by the extension Keemei. Keemei is used to check format compatibility so it can be used during QIIME2 analysis.

### Importing raw sequence files into QIIME2

The raw sequence files observed in this project were demultiplexed fastq.gz files containing 2 reads per sample. These fastq.gz files were passed through QIIME2 using specific QIIME2 commands for demultiplexed paired end readings. The output file created by QIIME2 contained all the paired end demultiplexed sequences for each of the 10 samples. From here further analysis was conducted.

### Creating a summary of demultiplex sequences

From the file created in the previous step, the demultiplexed paired end sequences were used to create a summary file. This file was created by taking the paired end sequences file and passing it through a visualization step. This is a QIIME2 specific code that takes .qza files and makes them into .qzv files so they can be viewed in QIIME2view.

### Denoising sequences

In order to remove sequence errors, the sequences were denoised. The process of denoising is done automatically in QIIME2 with a series of commands. The denoising method used was the QIIME2 DADA2 method. This process removed sequence errors from amplicon reads in order to identify the correct biological sequence. There are parameter options for denoising sequences. The parameters chosen for these samples were default for trimming (set to zero), and truncating left/right to 35 base pairs. This number was selected from the minimum sequence length observed during subsampling, which was 35 base pairs. The output of this code generated three files, a denoised sequence summary, denoising statistics, and representative sequence files.

### Creating visualizations

The final step after generating the various outputs served as a way to visualize the data. This was done by transforming the .qza files to .qzv files so they could be viewed in QIIME2view. The series of commands were used to create three different .qzv files. The first file generated was a feature table summary of the denoised sequences. This code used the denoised sequences table and metadata file to pass the sequences through a visualization step to create a feature table. The second file generated was a visualization of the denoising statistics. This code took the denoising statistics generated previously and passed it through a visualization step so it could be viewed in QIIME2view. Finally the third file generated was a visualization of representative sequences. This file was created by using the previous representative sequences file and passing it through a visualization step. This gave an output file that could be viewed in QIIME2view.

### Viewing visualizations in QIIME2view

QIIME2 view was used to visualize the files created in previous steps. This feature was used by retrieving the .qzv files and downloading them to a local computer drive. From there, the files were dragged and dropped into QIIME2view. The outputs were then saved to a local computer drive to be further analyzed.

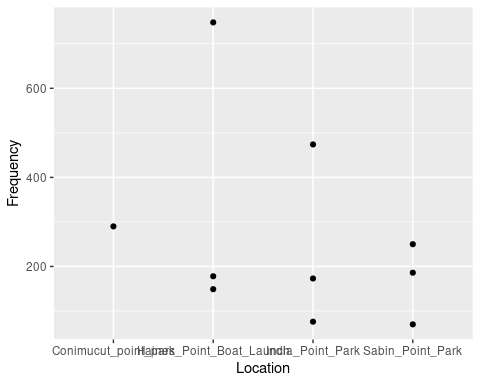
# Results

### Table 1 Metadata and Sequence Information

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Sample | Frequency | sequencecount | barcode | primer\_adapter | location |
| W-19-CP1 | 290 | 43153 | CGAGGCTGTAGATCGC | CTGTCTCTTATACACATCT | Conimucut\_point\_park |
| W-19-HP1 | 149 | 25336 | CTCTCTACTAGATCGC | CTGTCTCTTATACACATCT | Haines\_Point\_Boat\_Launch |
| W-19-HP2 | 748 | 46544 | CAGAGAGGTAGATCGC | CTGTCTCTTATACACATCT | Haines\_Point\_Boat\_Launch |
| W-19-HP3 | 178 | 27615 | GCTACGCTTAGATCGC | CTGTCTCTTATACACATCT | Haines\_Point\_Boat\_Launch |
| W-19-IPT1 | 76 | 20612 | TAAGGCGATAGATCGC | CTGTCTCTTATACACATCT | India\_Point\_Park |
| W-19-IPT2 | 173 | 22330 | CGTACTAGTAGATCGC | CTGTCTCTTATACACATCT | India\_Point\_Park |
| W-19-IPT3 | 474 | 43460 | AGGCAGAATAGATCGC | CTGTCTCTTATACACATCT | India\_Point\_Park |
| W-19-SP1 | 70 | 15681 | TCCTGAGCTAGATCGC | CTGTCTCTTATACACATCT | Sabin\_Point\_Park |
| W-19-SP2 | 186 | 28997 | GGACTCCTTAGATCGC | CTGTCTCTTATACACATCT | Sabin\_Point\_Park |
| W-19-SP3 | 250 | 40408 | TAGGCATGTAGATCGC | CTGTCTCTTATACACATCT | Sabin\_Point\_Park |

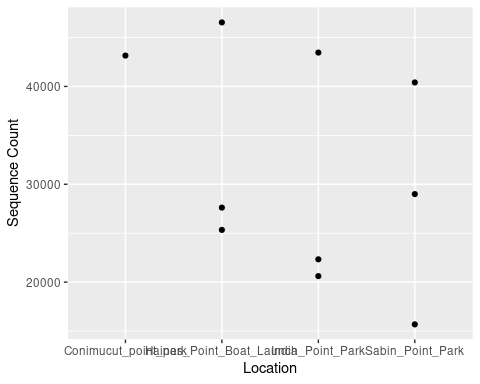
Table 1 includes information from the metadata file such as sample name, location, date collected and individual sample barcodes. The sample locations with the highest frequencies were collected from Haines Point boat launch and India Point Park. The sample with the lowest frequency was collected from Sabin Point Park.

### Figure 1 Location and Frequency



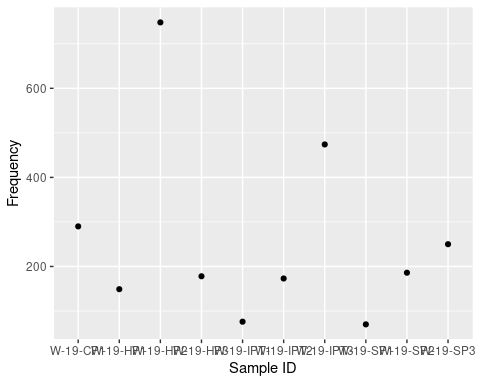
Haines Point boat launch has the highest frequency, followed by India Point Park. Haines Point boat launch is located near Barrington,RI, while India Point Park is located in Providence, RI. Conimucut Point Park has the lowest frequency, which is due to only one sample from that location being included. Conimucut Point Park is located is in Warwick, RI.

### Figure 2 Location and Sequence Count



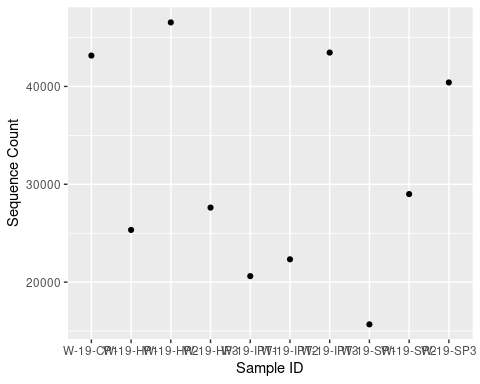
Haines Point boat launch has the highest sample sequence counts. India Point Park and Sabin Point Park are similar with sample sequence counts. The sample from Conimucut point park also has a high sequence count, but considering it is the only sample included from that location it remains lower than the other locations with multiple samples.

### Figure 3 Sample ID and Frequency



The sample with the highest frequency was W-19-HP2. This sample was collected from Haines Point Boat Launch. The samples with the lowest frequency are W-19-IPT1 and W-19-SP1. These samples were collected from India Point Park and Sabin Point Park, respectively.

### Figure 4 Sample ID and Sequence Count



The samples with the highest sequence count were W-19-CP1 and W-19-HP2. W-19-CP1 was sampled from Conimucut point park and W-19-HP2 was sampled from Haines Point boat launch.

The results of this subset of data shows a higher number of sequence counts and frequency in Haines Point boat launch followed by India Point Park. Conimucut Point Park has limited data due to only one sample included from this location, however the individual sample included has a high sequence count.The result expected from this analysis is, inconclusive due to a small sample size. Although some trends are visible in this subset, further analysis will need to be done in order to compare results across seasons and years including more locations.

The predicted results for this analysis points to higher biodiversity in southern areas of the bay with less human population density and more nutrient and dissolved oxygen availability, compared to northern areas of the bay. It is also expected to see increased biodiversity in warmer summer months than in colder winter months. It is also expected that eDNA sampling will generate more data about less frequently seen marine species overlooked in traditional methods of biodiversity sampling. In order to explore these possibilities, this analysis will be completed again with a larger sampling size over the span of two years.