**Introduction**

The Philippines is the marine biodiversity capital of the world making it extremely important to protect. However, due to anthropogenic factors, the species living there are under threat of degradation and need immediate protective action ([Carpenter & Springer, 2005](#_ENREF_2)). The PIRE (Partnerships in International Research and Education) project is an educational program promoting outreach to students in hopes that the Philippines will continue to grow in marine research and education so the environment can prosper ([Carpenter, 2017](#_ENREF_3)). By studying the species in the Philippines and understanding what factors makes them vulnerable, scientists will then be able to understand how to protect the species there.

This project is a puzzle piece in the PIRE project and will focus on the effect of the type of habitat inhabited on demography inference in Philippine fishes. There has been little research done on how various habitats have affected populations over time, but there has been research done on how the habitats are used and/or affected. The habitats of interest for this project are reef-associated, estuarine/brackish, and near-shore semi-pelagic; each habitat is affected differently by humans or other impacts; therefore, each habitat should affect each population differently.

Coral reefs have the highest biodiversity of any of the three habitats, but coral bleaching which is caused by climate change, exploitation, sedimentation, pollution, and habitat modification threaten the effectiveness of this habitat ([Honda et al., 2013](#_ENREF_8); [Pratchett et al., 2011](#_ENREF_14)). Estuaries are dynamic, transitional habitats that connect the surface water from landmasses to oceans. It is vitally important for both fishes and humans, but because of large human populations established near estuaries, human impacts such as habitat modification, pollution, eutrophication, altered hydrodynamics, water extraction, and overexploitation greatly damage the important habitat ([Feyrer et al., 2021](#_ENREF_6)). Near-shore semi-pelagic habitats are defined as water depths shallower than 30 m but are not directly next to the coast and fish inhabiting spend some time on the seafloor and sometime in the water column. It is among the most productive habitats and is usually utilized as feeding and nursery grounds; however, habitat is mainly affected by overfishing, pollution, and habitat modification ([Seitz et al., 2013](#_ENREF_17); [Sichum et al., 2013](#_ENREF_18); [Wen et al., 2010](#_ENREF_19)). These three habitats are usually intertwined with no real barrier separating them and have similar threats that put them in danger, but the question prevails how do the populations that inhabit each habitat vary in their population histories?

This project hypothesizes that the type of habitat inhabited does have an effect on the population history of Philippine fishes. Between these three habitats, fourteen species will be analyzed as listed in Table 1.

|  |  |
| --- | --- |
| Habitat | Species |
| Reef-associated | *Corythoichthys haematopterus*  *Lethrinus variegatus*  *Plotosus lineatus*  *Stethojulis interrupta*  *Taeniamia biguttata*  *Taeniamia kagoshimanus*  *Taeniamia zosterophora* |
| Estuarine/brackish | *Ambassis urotaenia*  *Ambassis kopsii*  *Leiognathus leuciscus* |
| Coastal/semi-pelagic | *Atherinomorus duodecimalis*  *Herklotsichthys quadrimaculatus*  *Hypoatherina temminckii*  *Spratelloides gracilis* |

***Table 1****: List of species and their corresponding habitats*

To test the hypotheses, sub-sampling and DNA extraction will be done to the selected species. The extracted DNA will then be processed through the PIRE SSL (shotgun sequencing library) pipeline to develop a genome sequence to perform PSMC (pairwise sequentially Markovian coalescent) demographic analyses of the species to assess whether the type of habitat inhabited can be associated with historical population trends. The PSMC output can also test whether different habitats affect the differences between the genomes of different species. Those outputs can relate to effective population size, average effective population size, maximum, minimum, etc. Those outputs will be analyzed through R to produce statistical computing and graphics.

**Methods: Subsampling and DNA extraction**

One individual per species was selected and three libraries were started from each individual. The DNA was extracted in the lab at Old Dominion University using the Qiagen DNeasy blood tissue kit with the following modification: in the final step, the elutions were saved in separate vials. Elution 1 was saved into a different vial than elution 2, and so on. Each sample was subjected to gel electrophoresis and fluorescent quantification using Biodiem AccuClear kit Spectramax M3 plate reader.

The best elution was identified and working plates were created using an Eppendorf EP motion fluidics robot. A quantity of 16.67 micrograms of DNA was transferred for each library/specimen. The extracts went through two rounds of paramagnetic bead cleanup (Omega Bio-Tek) and then were processed through the KAPA HyperPlus kit, which ran at 1/4x for each reaction.

Personnel followed the manual to optimize the enzymatic digestion of the DNA targeting approximately 300-500 base pair fragments. The KAPA HyperPlus kit includes PCR and DNA ligation using iTru stubby adapters (Travis Glenn, University of Georgia). PCR with iTru primers is used to uniquely index each library. Then using fluorescent quantification again and advanced analytic fragment length analyzer, the samples were combined in equimolar amounts and subjected to a size selection (200-600 bp) step on a BluePipen pulse field electrophoresis machine. The library was quantified using a KAPA qPCR DNA quantification kit and was sent for sequencing at an Illumina NovaSeq 6000 S4 Flow Cell by NovoGene. 50 million read pairs per library were targeted.

**Methods: PIRE SSL**

Shotgun sequencing was used to randomly shear DNA into tiny fragments, clone them, and then sequence them creating “reads” that are output to a fq.gz file for downstream analysis. Before I could assemble the genome, quality control was necessary to ensure that the reads were ready for PSMC.

*Pre-Processing Section: Step 1. Quality assessment (FastQC/MultiQC)*

A baseline assessment of the quality and quantity of our data is necessary before starting any altering step and this is done by using FastQC and MultiQC. They are designed to give a brief assessment of the data and identify any problem areas. FastQC generates reports for each fq.gz file and MultiQC aggregates the information into one summary file. To check the quality of data, we ran an sbatch job by using the code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/Multi\_FASTQC.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta/shotgun\_raw\_fq" "fq.gz"

*Step 2. First Trim with fastp*

The first trimming removed low quality bases and “bad” reads; those can include reads that are too short, reads that have too many low-quality bases, 3’ ends of reads, and adaptor sequences. I ran an sbatch job to perform the trimming using the code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/runFASTP\_1st\_trim.sbatch shotgun\_raw\_fq fq\_fp1

*Step 3. Clumpify (De-duplicating)*

At this point, I needed to reduce data size to make the assembly portion more efficient and to minimize effects of sequencing errors. Leftover reads from PCR amplification during the library prep needed to be removed to prevent said sequencing errors. Clumpify sorted reads in the fq.gz files into “clumps” by ordering them by similarity and then removed the duplicates. I then ran the sbatch job by running this code:

bash /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/runCLUMPIFY\_r1r2\_array.bash fq\_fp1 fq\_fp1\_clmp /scratch/aethr001 3

*Step 4. Second Trim with fastp2*

For the second trim I had to remove the 5’ end of the reads if they have low quality bases. I ran the sbatch script to perform the second trim by using the code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/runFASTP\_2\_ssl.sbatch fq\_fp1\_clmp fq\_fp1\_clmp\_fp2

*Step 5. Removing any contamination (FastQ Screen)*

Contamination, non-target DNA, is a possibility during sample storage, extraction, and library prep. I ensured that contamination is removed as it can interfere with the assembly downstream. FastQ Screen searched our DNA against a database of other genomes to identify and remove extraneous DNA by running this code:

bash /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/runFQSCRN\_6.bash fq\_fp1\_clmp\_fp2 fq\_fp1\_clmp\_fp2\_fqscrn 6

After checking that there are no errors, I then proceeded to run Multiqc bash script separately using the code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/runMULTIQC.sbatch fq\_fp1\_clmp\_fp2\_fqscrn fastqc\_screen\_report

*Step 6. Repair fastq\_screen paired end files*

In this step, it is necessary to repair any read pairs that were separated during the previous steps and is to ensure the R1 and R2 fq.gz files are in the same order making assembly easier.  To run the script, this code was used:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/runREPAIR.sbatch fq\_fp1\_clmp\_fp2\_fqscrn fq\_fp1\_clmp\_fp2\_fqscrn\_repaired 40

Alongside the repair script, Fastqc-Multiqc must be ran separately using the code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_fq\_gz\_processing/Multi\_FASTQC.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta/fq\_fp1\_clmp\_fp2\_fqscrn\_repaired" "fq.gz"

*Genome Assembly Section: Step 1. Genome Properties*

The sequences at this point were ready to be assembled, however, generating data on the sequences is imperative before assembling the genome. Jellyfish is a script that can create a histogram file that can then be uploaded to GenomeScope v1.0 and GenomeScope v2.0 using this code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runJellyfish.sbatch "Sin" "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta/fq\_fp1\_clmp\_fp2\_fqscrn\_repaired"

The values of interest were heterozygosity, genome haploid length, and model fit and should be used to choose either the values derived from GenomeScope v1.0 or GenomeScope v2.0. Heterozygosity should be around 1% and model fit should be greater than 90%; the values for GenomeScope v2.0 aligned better with the preferred 1% and 90% so those values will be used in future steps.

*Step 2. Assemble the Genome Using SPAdes*

The estimated genome haploid length from GenomeScope v2.0 was then used to assemble the genome using SPAdes for each library and all the libraries combined using these codes:

#1st library

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runSPADEShimem\_R1R2\_noisolate.sbatch "aethr001" "Sin" "1" "decontam" "675000000" "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "fq\_fp1\_clmp\_fp2\_fqscrn\_repaired"

#2nd library

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runSPADEShimem\_R1R2\_noisolate.sbatch "aethr001" "Sin" "2" "decontam" "675000000" "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "fq\_fp1\_clmp\_fp2\_fqscrn\_repaired"

#3rd library

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runSPADEShimem\_R1R2\_noisolate.sbatch "aethr001" "Sin" "3" "decontam" "675000000" "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "fq\_fp1\_clmp\_fp2\_fqscrn\_repaired"

#all libraries combined

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runSPADEShimem\_R1R2\_noisolate.sbatch "aethr001" "Sin" "all\_3libs" "decontam" "675000000" "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "fq\_fp1\_clmp\_fp2\_fqscrn\_repaired"

SPAdes automatically runs QUAST which gives statistics on the genome noting number of contigs, size of largest contig, total length of assembly, N50, and L50.

*Step 3. Running BUSCO*

BUSCO tells how many expected genes were recovered by the assembly and code needs to ran for each library and combined libraries for both contigs and scaffolds using this code:

#1st library - contigs

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_Sin-CPnd-A\_decontam\_R1R2\_noIsolate" "contigs"

#2nd library -contigs

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_Sin-CPnd-B\_decontam\_R1R2\_noIsolate" "contigs"

#3rd library - contigs

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_Sin-CPnd-C\_decontam\_R1R2\_noIsolate" "contigs"

#all libraries - contigs

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_allLibs\_decontam\_R1R2\_noIsolate" "contigs"

#1st library -scaffolds

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_Sin-CPnd-A\_decontam\_R1R2\_noIsolate" "scaffolds"

#2nd library - scaffolds

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_Sin-CPnd-B\_decontam\_R1R2\_noIsolate" "scaffolds"

#3rd library - scaffolds

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_Sin-CPnd-C\_decontam\_R1R2\_noIsolate" "scaffolds"

#all libraries - scaffolds

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runBUSCO.sh "/home/e1garcia/shotgun\_PIRE/REUs/2022\_REU/Abby\_Ethridge/stethojulis\_interrupta" "SPAdes\_allLibs\_decontam\_R1R2\_noIsolate" "scaffolds"

*Step 5. Identify Best Assembly*

Selecting the best assembly for downstream analysis uses the results from QUAST and BUSCO. The metrics used in the determination is number of contigs, size of largest contig, total length of assembly, N50, L50, and BUSCO score.

*Step 6. Assemble Contaminated Data from the Best Library*

Finally, after selecting the best assembly, assembling the contaminated data from the selected library was done before putting the data through PSMC. It can be assembled with this code:

sbatch /home/e1garcia/shotgun\_PIRE/pire\_ssl\_data\_processing/scripts/runSPADEShimem\_R1R2\_noisolate.sbatch "e1garcia" "Sfa" "2" "contam" "635000000" "/home/e1garcia/shotgun\_PIRE/2022\_PIRE\_omics\_workshop/salarias\_fasciatus" "fq\_fp1\_clmp\_fp2\_fqscrn\_repaired"

**Methods: PSMC**

After processing through the PIRE SSL pipeline was completed, the samples were processed using PSMC to infer effective population size (Ne).

PSMC is built upon the concept of sequentially Markovian coalescence and the idea that a single organism’s genome carries a record of its lineage. Coalescence events are defined when two separate lineages within said genome fuse into a single ancestral lineage. Spotting these events requires a “backwards” view where base pairs are traced back to their common ancestors. A large count and frequency of coalescence events present in a genome is indicative of a small population, as smaller populations are more likely to share more recent ancestors, and vice versa ([Mather et al., 2020](#_ENREF_12)).

The Markov models in PSMC account for the amount of heterozygosity present in a genome. Multiple chunks of 150 base pairs each were processed and given a binary marker that indicates if a heterozygous locus is present within those 100 base pairs. Higher occurrences of “heterozygous-positive” base pair chunks are associated with larger populations, as individuals in larger populations tend to have a larger effective population and gene pool. This leads to a more diverse genome ([Mather et al., 2020](#_ENREF_12)).

PSMC is not without limitations. The process assumes neutral evolution, a set mutation rate, and nonrandom mating, which do not always occur in nature. Contemporary population sizes cannot be estimated as results from more recently than 10,000 years ago are unreliable. Additionally, a single run of PSMC only displays one possibility of demography based on the probability of that result’s coalescence timing and frequency. This can be amended by bootstrapping.

The following parameters were used by PSMC to estimate Ne: a manually set scaled mutation rate of 2.25x108 manually determined species-specific generation times, an automatically determined recombination rate, and an automatically determined piecewise constant population size ([Li & Durbin, 2011](#_ENREF_11)). A PERL -language script was used to filter out scaffolds that fell outside of the 20kb – 100kb length range. A modified version of the dDocent pipeline was used to map remaining shotgun reads to reference genomes. A samtools script calculated the mean depth of the mapped reads. A pipeline script modified from Harvard FAS Informatics tutorial, Applying PSMC to Neandertal data, & the PSMC documentation was used to call a consensus sequence from the mapped bam files. Only scaffolds ranging between one-third and two times the mean depth were used for this process ([Li & Durbin, 2011](#_ENREF_11)). The resulting scaffolds were converted to a PSMC format through an array script. The PSMC script was run applying both the re-formatted files and the species generation time as arguments. The PSMC was then run 100 more times via bootstrapping and combined into one graph by an additional script.

**Methods: R**

RStudio v.4.0.1 (R Core Team, 2020) was used to test the validity of the null hypothesis for this study using statistical analysis. First, the packages tidyverse, janitor, lubridate, ggpubr, and ggplot were installed using the following lines of code: The PSMC data was then read in using a tibble and piping (%>%) command. A plot was then created using a ggplot command, specifying the type of graph, x variable, and y variable. In the case of distribution data, a linear regression model was used to fit the quantitative data using an lm command. ANOVA was then implemented to determine the differences between statistically significant groups using the same lm command. The statistical significance was determined by having a p value lower than 0.05, at which point is considered statistically significant, and the null hypothesis is rejected. The response variables were the length of distribution and fixed effects were population size.

We then test assumptions of normality, equal variance among groups, independence, and linearity for regression using a Q-Q plot.

**Results: Shotgun Sequencing of *Stethojulius interrupta***

Several libraries were assembled to generate the genome of *S. interrupta.* However, only the best library was used to create the final genome assembly. Best quality was determined using 5 factors, ranked below in most important to least important.

1. BUSCO- gives the percentage of expected genes observed in the assembly.
   1. Larger is better
2. N50- gives the length of the smaller contig from the set of contigs needed to reach half of the assembly
   1. Larger is better
3. Genome size completeness-gives length of assembly divided by estimated genome length
   1. Larger is better
4. L50- number of contigs needed to reach half of assembly
   1. Smaller is better
5. Largest contig- length of largest contigs
   1. Larger is better

The following shows the Quast and BUSCO results of all of the libraries assembled:

Table

Description automatically generated

***Table 2****. Quast and BUSCO results of 6 assembled libraries for S. interrupta*

The most influential factor to a high-quality assembly is the BUSCO score. Based on Table 2, the library assembly with the highest BUSCO score was scaffolds library B, with a percentage of 65%. However, scaffolds library A had a similar BUSCO score, which was 64%. Since both are comparable, the N50 score was also used. Scaffolds library B had a lower score than scaffolds library A, which makes scaffolds library A higher quality. Genome size completeness was the exact same for both libraries. The L50 score was also slightly more favorable for scaffolds library A. While the largest contig metric was better for scaffolds library B, this score is the least influential in determining assembly quality. For these reasons in that order, Scaffolds Library A was the highest quality genome assembly. This assembly was pushed forward into the next data analysis stage, which was PSMC.

**Results: PSMC**

The Pairwise Sequentially Markovian Coalescent results are plotted using effective population size (Ne) over time, dating back to 1,000,000 years ago. A default mutation rate (∞) of 2.25x10-8 was used across all species, while estimated generation time (g) varied with each species. Bootstrapping bands are used as confidence intervals for our demographic history based on resampling chunks of the data. For this study, 14 species underwent PSMC. The plots below show an example of one species of each habitat grouping, and *S. interrupta*.

Below is a plot with bootstrapping bands of the PSMC results of *Stethojulis interrupta.*

A picture containing chart

Description automatically generated

***Figure 1.*** *Effective population size of S. interrupta over the last 1,000,000 years, with an estimated generation time of 2.89.*

Below is a plot with bootstrapping bands of the PSMC results of *Herklotsichthys quadrimaculatus.*

*Chart

Description automatically generated*

***Figure 2.*** *Effective population size of H. quadrimaculatus within the last 1,000,000 years, with an estimated generation time of 1.3.*

Below is a plot with bootstrapping bands of the PSMC results of *Ambassis kopsii.*

*Chart, waterfall chart

Description automatically generated*

***Figure 3.*** *Effective population size of A. kopsii within the last 1,000,000 years, with an estimated generation time of 2.53.*

As seen in all the PSMC and bootstrapping plots (Figures 1-3), confidence intervals decrease as time plotted becomes more recent. The more ancient the sample, the higher the confidence interval. *A. kopsii* generally had a higher confidence interval in more recent years than the other species; in the same manner, *S. interrupta* had a lower confidence interval in more recent years, as seen with the bootstrapping bands.

**Results: Habitat Usage Trends of the 14 Species**

In terms of habitat usage, the species in this study were grouped into 3 categories: estuarine/brackish, near-shore semi-pelagic, and reef associated species. The plot below shows this demographic history and effective population size both by species and habitat usage type. The red lines indicate all species that are estuarine/brackish species (3/14). Green lines indicate near-shore semi pelagic species (4/14) and blue lines indicate reef associated species (7/14).

Chart, histogram

Description automatically generated

***Figure 5.*** *Effective population size of all 14 species over the last 2.5 million years, grouped by habitat usage.*

Figure 5 shows that estuarine/brackish fishes generally had a lower effective population size, with the maximum of one of the species at roughly 1.5e+0.6 individuals. The maximum population of a near shore semi-pelagic species capped at 2.5e+0.6 individuals, around the same time that the maximum population size of a reef associated species, which reached around 2.6e+0.6 individuals. Overall, most of the species had a maximum effective population below 0.5e+0.6 individuals, except for 5 species: 2 estuarine, 1 near-shore semi-pelagic, and 2 reef associated species.

This study also looks at specific time points and effective population sizes. The table below specifies the estimated population size of all the species in this study at the following timestamps: 10,000 years ago; 25,000 years ago; 100,000 years ago; 500,000 years ago.

Table

Description automatically generated

***Table 3.*** *Estimated population size of all 14 species studied at specific time points.*

Chart, bar chart

Description automatically generated

***Figure 6.*** *Estimated population size of all 14 species studied at specific time points.*

*Chart, bar chart

Description automatically generated*

***Figure 7.*** *Theta0 values for each species.*

Table 3 and Figure 6 show that near-shore semi-pelagic species *H. temminckii* has a noticeably higher effective population at 10,000 years ago and 25,0000 years ago than other species. Similarly, the effective population size for *H. temminckii* is estimated to be at 2,493,790 at both time points. The species *S. interrupta* (reef associated) accounts for most of the individuals estimated at 100,000 years ago, with an effective population of 2,549,284. The effective population size of this species also did not change at 10,000 years ago and 25,000 years ago: both are measured to be at 1,816,778 individuals.

Theta0 is a metric for the long-term average of population size over time. Figure 7 indicates that *S. interrupta* (reef associated)has the highest theta0 value, followed by *H. temminckii* (near-shore semi-pelagic), followed by *L. leuciscus* (estuarine/brackish).

**Results: Hypothesis Testing**

The following table shows the values required for hypothesis testing to either accept or reject the null hypothesis, which states that habitat types have no effect on the population history of the different species studied. The alternative hypothesis states that the different habitat types do have an effect on the population history of the different species studied. The statistical significance was determined by having a p value lower than 0.05, at which point is considered statistically significant, and the null hypothesis is rejected. The F-statistic measures dispersion of the data values from the mean. Larger values represent greater dispersion. Variance is the square of the standard deviation. The higher the F-value, the lower the p-value. The R2 value measures the goodness of fit, or more specifically, the proportion of variance in the dependent variable that can be explained by the independent variable. The higher the R2 value, the higher the correspondence between the independent variable (habitat type) and dependent variable (population).

|  |  |  |  |
| --- | --- | --- | --- |
| Ne | p-value | F-statistic | R2 |
| Ne 10,000 | 0.7589 | 0.2829 | 0.04892 |
| Ne 25,000 | 0.7614 | 0.2795 | 0.04836 |
| Ne 100,000 | 0.8326 | 0.1862 | 0.03275 |
| Ne 500,000 | 0.3773 | 1.066 | 0.1624 |
| theta0 | 0.7813 | 0.2524 | 0.04388 |

***Table 4.*** *The effective population at the specified time points, theta0, p-value, F-statistic, and R2 value*

Table 4 shows that none of the data have a p-value lower than 0.05, so none of the data is statistically significant. From here, we can now accept the null hypothesis that habitat usage has no effect on population history. The F-statistic for Ne 500,000 has a high value at 1.066, which states that this has the greatest variation between sample means relative to the variation within the samples. All the R2 values are lower than 0.5, which means that there is little correlation between habitat type and all effective population sizes at different time points.

**Discussion**

*Reef Habitats*

As stated previously, coral reefs have the highest marine biodiversity of all three habitats studied. Figure 7 shows that *S. interrupta*, a reef associated species, had the highest long term population size over time and the highest population size at 100,000 years ago. *S. interrupta* also had an unchanging population between 10,000 years ago and 25,000 years ago. We have seen through Figure 5 that reef associated species can have the potential to have extremely large populations.

The highest levels of marine biodiversity associated with habitats lies in coral reefs. This is likely due to recent efforts in coral reef restoration because it is recognized for concentration of biological activity, fisheries and tourism, coastal protection, and geological processes ([Jaap, 2000](#_ENREF_9)). In a recent paper on coral reef restoration, it is found that only the establishment of large-scale nurseries and transplantation actions, together with conventional management tools, will be able to cope with extensive reef degradation on the global scale ([Rinkevich, 2008](#_ENREF_15)).

*S. interrupta* has a stable and large population. Stable populations are genetically favorable for several reasons. First, population size affects the probability of extinction, genetic diversity, and adaptive potential ([Boorman & Levitt, 1973](#_ENREF_1)). It is important in identifying trends to give inferences on management, as suggested by r and k strategists ([Parry, 1981](#_ENREF_13)). The stable population size of *S. interrupta* is approximately around its carrying capacity (k). This can be attributed to its high, minimum population doubling time less than 15 months, its preliminary K or fecundity ([Froese et al., 2017](#_ENREF_7)).

*Estuarine/Brackish Habitats*

Estuarine/brackish habitats are the transitional types of habitats for marine life. From the results, estuarine/brackish fishes generally had a lower effective population size but has the potential to have very high effective population sizes, as seen from its maximum effective population and high theta0 value.  *L. leuciscus* is the one species of the 14 studied that had a high average population size over the last 1,000,000 years. Unlike *S. interrupta* however, the estimated population size of L. *leuciscus* is not stable over the specified time points we studied.  The population of *L. leuciscus* was highest 25,000 years ago and much lower 10,000 years ago and 100,000 years ago.

Estuarine/brackish fishes may generally have a lower effective population compared to reef associated species and semi-pelagic species due to its reliance on freshwater flow. A study done in 1995 shows that estuarine habitats in San Francisco showed a decline of some estuarine populations due to decreased river inflow due to drought and increased freshwater diversion [Jassby et al., 1995](#_ENREF_6)). The same study also indicates that diversion of water for export from or consumption within the estuary can have a direct effect on population abundance independent of its effect on the 2% bottom salinity position. This metric is important for understanding the supply of phytoplankton and phytoplankton-derived detritus from local production and river loading; benthic macroinvertebrates (mollusks); mysids and shrimp; larval fish survival; and the abundance of planktivorous, piscivorous, and bottom-foraging fish ([Jassby et al., 1995](#_ENREF_10)). Similar problems regarding freshwater flow can be affecting the estuarine associated fish species populations.

*Near-shore Semi-pelagic Habitats*

As stated previously, near-shore semi-pelagic habitats is among the most productive habitats and is usually utilized as feeding and nursery grounds but is most affected by overfishing and similar human interferences. From our results, near-shore pelagic habitats are not as likely to reach high maximum effective populations, unlike reef-associated or estuarine/brackish species.

Near-shore semi-pelagic habitats include water columns that act as a productive area for marine life, including activities relating to feeding and nursery grounds. As a nursery ground is a large benefit to this type of habitat, it makes sense that the effective population of this habitat does not get as high as the other habitat types studied. Near-shore semi-pelagic habitats exist as a temporary ground for marine life. Near-shore semi-pelagic species are also most affected by human pressures such as overfishing. A study performed in 1987 shows that available evidence suggests that small pelagic fishes in the Philippines are very heavily exploited or overfished ([Dalzell & Ganaden, 1987](#_ENREF_5)). Small pelagic fish have high human value, which explains its overfishing. Virtually all the landings of small pelagic fishes are destined for human consumption either fresh or as diced, smoked, fermented, and canned products. Some components of the small pelagic catch such as round scads are a staple diet of lower income groups in Philippine society ([Ronquillo, 1975](#_ENREF_16)).

**Conclusion**

The shotgun sequencing pipeline used to generate the genome assembly of *S. interrupta* can be used for further studies as a reference genome assembly. The rest of the species included in this study had a reference genome to start, and so the first step of this project was to generate the assembly which can be added to a reference genome bank. This study is an example of how coverage whole genome sequencing can be conducted with as little as 1 individual, and how it can be applied to larger topics. Pairwise Sequential Markovian Coalescence was the second portion of this study, and as seen in the results, it is not without limitations. PSMC is far more accurate the further back we can trace the species, and much less accurate within 25,000 to 10,000 years. However, it is a valuable tool to see how our independent variable of habitat usage can have an effect on population history. Our statistical analysis shows that none of our data values are statistically significant, and so habitat usage has no effect on population history.

Further objectives that can relate to this study are to compare more species in the three distinct habitat categories as well as add other aquatic habitats in the Philippines such as freshwater. It should also be considered to use another program that can infer demographic history should be used for more recent time periods. Better estimates of generation time would make effective population estimates much more accurate. As mentioned before generation time is specific to each species and is derived by diving time by number of generations. Research suggests a combination of body-mass and phylogeny should be used when estimating generation time ([Cooke et al., 2018](#_ENREF_4)). To a broader effect, the PIRE Project has a focus on general speciation and its decrease to human interference, so a more recent study looking at population could be very useful. Regardless of the outcome, it would be beneficial to look at how different human pressures such as coral bleaching, overfishing, and pollution related habitat degradation has affected marine life.

**Works cited**

Boorman, S. A., & Levitt, P. R. (1973). Group selection on the boundary of a stable population. *Theoretical Population Biology*, *4*(1), 85-128. <https://doi.org/https://doi.org/10.1016/0040-5809(73)90007-5>

Carpenter, K. E., & Springer, V. G. (2005). The center of the center of marine shore fish biodiversity: the Philippine Islands. *Environmental Biology of Fishes*, *72*(4), 467-480. <https://doi.org/10.1007/s10641-004-3154-4>

Carpenter, K. e. a. (2017). PIRE Proposal.

Cooke, R. S. C., Gilbert, T. C., Riordan, P., & Mallon, D. (2018). Improving generation length estimates for the IUCN Red List. *PLOS ONE*, *13*(1), e0191770. <https://doi.org/10.1371/journal.pone.0191770>

Dalzell, P., & Ganaden, R. (1987). The overfishing of small pelagic fish stocks in the Philippines. *RAPA Rep*, *10*, 249-256.

Feyrer, F., Young, M. J., Huntsman, B. M., & Brown, L. R. (2021). Disentangling Stationary and Dynamic Estuarine Fish Habitat to Inform Conservation: Species-Specific Responses to Physical Habitat and Water Quality in San Francisco Estuary. *Marine and Coastal Fisheries*, *13*(5), 548-563. <https://doi.org/https://doi.org/10.1002/mcf2.10183>

Froese, R., Demirel, N., Coro, G., Kleisner, K. M., & Winker, H. (2017). Estimating fisheries reference points from catch and resilience. *Fish and Fisheries*, *18*(3), 506-526. <https://doi.org/https://doi.org/10.1111/faf.12190>

Honda, K., Nakamura, Y., Nakaoka, M., Uy, W. H., & Fortes, M. D. (2013). Habitat Use by Fishes in Coral Reefs, Seagrass Beds and Mangrove Habitats in the Philippines. *PLOS ONE*, *8*(8), e65735. <https://doi.org/10.1371/journal.pone.0065735>

Jaap, W. C. (2000). Coral reef restoration. *Ecological Engineering*, *15*(3), 345-364. <https://doi.org/https://doi.org/10.1016/S0925-8574(00)00085-9>

Jassby, A. D., Kimmerer, W. J., Monismith, S. G., Armor, C., Cloern, J. E., Powell, T. M., Schubel, J. R., & Vendlinski, T. J. (1995). Isohaline Position as a Habitat Indicator for Estuarine Populations. *Ecological Applications*, *5*(1), 272-289. <https://doi.org/https://doi.org/10.2307/1942069>

Li, H., & Durbin, R. (2011). Inference of human population history from individual whole-genome sequences. *Nature*, *475*(7357), 493-496. <https://doi.org/10.1038/nature10231>

Mather, N., Traves, S. M., & Ho, S. Y. W. (2020). A practical introduction to sequentially Markovian coalescent methods for estimating demographic history from genomic data. *Ecology and Evolution*, *10*(1), 579-589. <https://doi.org/https://doi.org/10.1002/ece3.5888>

Parry, G. D. (1981). The Meanings of r- and K-Selection. *Oecologia*, *48*(2), 260-264. <http://www.jstor.org/stable/4216304>

Pratchett, M. S., Hoey, A. S., Wilson, S. K., Messmer, V., & Graham, N. A. J. (2011). Changes in Biodiversity and Functioning of Reef Fish Assemblages following Coral Bleaching and Coral Loss. *Diversity*, *3*(3), 424-452. <https://doi.org/https://doi.org/10.3390/d3030424>

Rinkevich, B. (2008). Management of coral reefs: We have gone wrong when neglecting active reef restoration. *Marine Pollution Bulletin*, *56*(11), 1821-1824. <https://doi.org/https://doi.org/10.1016/j.marpolbul.2008.08.014>

Ronquillo, I. (1975). A review of the roundscad fishery in the Philippines. *Philippine Journal of Fisheries*, *2*(1-2), 86-126.

Seitz, R. D., Wennhage, H., Bergström, U., Lipcius, R. N., & Ysebaert, T. (2013). Ecological value of coastal habitats for commercially and ecologically important species. *ICES Journal of Marine Science*, *71*(3), 648-665. <https://doi.org/10.1093/icesjms/fst152>

Sichum, S., Tantichodok, P., & Jutagate, T. (2013). DIVERSITY AND ASSEMBLAGE PATTERNS OF JUVENILE AND SMALL SIZED FISHES IN THE NEARSHORE HABITATS OF THE GULF OF THAILAND. *Raffles Bulletin of Zoology*, *61*(2).

Wen, C. K.-C., Pratchett, M. S., Shao, K.-T., Kan, K.-P., & Chan, B. K. K. (2010). Effects of habitat modification on coastal fish assemblages. *Journal of Fish Biology*, *77*(7), 1674-1687. <https://doi.org/https://doi.org/10.1111/j.1095-8649.2010.02809.x>