Rebecca Venezia

**Updated Proposed Final Bioinformatics Project Plan**

**Questions/Predictions:**

1. How does sewage pollution from human areas impact the genetic diversity and the connectivity of marine organisms’ populations in coastal areas?

*Prediction: Fiddler crab populations around sewage outfalls will have lower genetic diversity and connectivity compared to the control populations.*

1. To understand the genes that are evolving with adaptation selection in marine populations from human sewage pollution?

*Prediction: Certain genes under selection will have similar variant frequencies when compared with the same treatment, so ones in areas with swage outfall with look the same and the control will look the same.*

This will be studied using mudflat fiddler crabs, since they are a good model organism to use for coastal environments. The reason they are easy to sample with high population numbers and are present in most salt marsh communities. The location of the sample sites will be at Corpus Christie because there are a large number of sewage sties (~70 sewage outfalls) with 1 to 26 million gallons of waste per day and control sites with no sewage that can be used. These sites are all in the same area, so their connectivity can be compared and have similar conditions. The human pollution in marine systems comes from sewage and runoff, which contains a large variety of substances that could impact marine life such as chemicals, substrates, oils, detergents, heavy metals, pesticides, fertilizers, road salt, organic waste, and hormones. Sewage will be used as human pollution for this study.

**Data:**

Samples were collected from mudflat fiddler crabs from 12 total sites with a mix from sewer outfalls and some from control sites with now sewage. The samples were taken at each site and then below and above each stie, with the distance changing depending on the size of the sewage outfall.

DNA samples will be prepared with RadSeq, which will focus on different locus in specific parts of the genes. The DNA will be extracted and pooled together, where the pooled sample will be sent off for Illumina sequencing. The data received from this will be in a fasta file format.

**Analysis plan:**

The analysis of this RadSeq data will be done in the dDocent pipeline to produce SNPS. The SNP will be filtered and graphed with programs in R to answer the questions of this study with the steps listed out below.

1. dDocent steps

* To produce SNPs that can tell the number of loci across the fiddler crab individuals.
  1. Quality Filtering
  2. De Novo Assembly
     1. Read Mapping
  3. SNP calling
  4. SNP Filtering

1. Call and Further SNP filtering with the command vcftools

* to find the loci that are outliers in the data and know which loci are being selected
* Use the filtering to find the best way to get the cleanest loci from the fiddler crab individuals

1. Outlier Detection

* 2 methods

1. Using BayeScan in Linux/R for the outliers to be detected and plotted in R
2. Using PCA adapt in R (package pcadapt) and then plotted in R to find the outlier loci and see population structure between the control and sewage groups
3. Calculation and plotting of Fst values to answer questions about the loci outliers in the data.

* Outflank will be used to calculate the Fst values and to create plots of the Fst values for visualization