NSF Data Management Plan

1. Expected Data

Our study will produce four large data sets: 1) raw cDNA transcriptome sequence reads for adult and larval stages of Atlantic silversides and a complete assembly of these reads into a single annotated reference transcriptome sequence for the species, 2) exon capture probe design based on the reference transcriptome, 3) raw DNA sequence reads from exon sequencing of experimental animals and inferred genotypes and allele frequencies at single nucleotide polymorphisms (SNPs), and 4) raw DNA sequence reads from exon sequencing of individuals collected from natural silverside populations and inferred genotypes and SNP allele frequencies.

2. Data Format

Raw sequence data will be in Fastq format, assembled reference sequence and exon-capture probes will be in Fasta format, and genotype and allele frequency data will be in tabular text format. The computational pipeline used to analyze the data will be written in regular text format and custom data processing scripts will be in Python or R language.

3. Access to Data, Data Sharing Practices and Archiving

We will work with Biological and Chemical Oceanography Data Management Office (BCO-DMO) staff to manage the data generated in the project and all data sets will be made available or linked to online from the BCO-DMO data system (http://bco-dmo.org/data/). Raw sequence reads will be archived digitally at the National Center for Biotechnology Information (NCBI) Sequence Read Archive (http://www.ncbi.nlm.nih. gov/sra) and the annotated transcriptome, capture probe sequences, individual genotypes and allele frequency estimates, as well as analysis scripts will be archived at the Dryad Digital Repository (http://datadryad.org/). All data will be digitally archived on these sites within 12 months and will be freely available for download.