DATA MANAGEMENT PLAN

We are committed to broad dissemination of our data and metadata during the timeframe of this project (within 2 years of data collection). All project personnel will be trained in data management procedures in accordance with NSF requirements.

Products of the research:
Our project will generate 1) physical samples of Atlantic silverside (*Menidia menidia*) collected for genomic analysis; 2) raw genomic sequence data and processed datasets; 3) phenotypic data; and 4) custom data processing scripts. Each type of data will be managed by the PI to maximize access, sharing and preservation.

1) *Physical samples.* We will archive remaining tissue samples and genomic DNA preparations in the Therkildsen Laboratory at Cornell University. Tissue samples will be frozen or preserved in ethanol and genomic DNA will be stored at -80˚C.
2) *Genomic datasets.* We will produce four large genomic datasets: 1) Two assembled and annotated reference genomes for the Atlantic silverside (from two different individuals); 2) Low-coverage whole-genome re-sequencing data from a total of 7 locations across the species range (650 individuals) that will be used for spatial analyses of adaptive differentiation and genome-wide association mapping (GWAS) of putative adaptive traits; c) restriction-site associated DNA (RAD) data for linkage mapping and QTL mapping; d) genotype data from sequence capture of the seasonal samples. For each dataset, we will archive publically both the raw sequencing reads and the processed datasets used for analysis.
3) *Phenotypic data.* We will collect phenotypic data for QTL mapping and GWAS from samples caught in the wild and samples reared under standardized conditions.
4) *Custom scripts.* We will develop custom scripts as necessary for data analysis.

Data format and storage:
All raw genomic sequence data from the genome assembly, whole-genome re-sequencing, RADseq and sequence capture will be in Fastq format. The genome assembly and annotation files will be in Fasta format and GFF format, respectively. Genotype likelihoods, allele frequency estimates, and called SNP genotypes (from the sequence capture data) will be in tabular text format. Individual sample metadata and phenotypic measurements will be in tabular text format and will be associated with detailed annotations about collection location and time to enable association with geospatial environmental data. Custom scripts for bioinformatics analyses and trait mapping will be in Python or R language and saved in plain text format. During the project all data will be stored with automatic daily back-up on PI Therkildsen's networked computing server and on external harddrives.

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/). All 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), while the genome assembly,

annotation and linkage map will be uploaded on the NCBI Genome Database. Phenotype, genotype and allele frequency data will be archived at the Dryad Digital Repository (http://datadryad.org/). Custom scripts and pipelines will be available on GitHub (https://github.com/). All data will be digitally archived on these sites within 2 years of data collection and will be freely available for download. We anticipate no sensitive or confidential data.