COLLABORATIVE RESEARCH: Tracking fine-scale selection to temperature at the invasion front of a highly dispersive marine predator

Data Management Plan

Data Policy Compliance
The project investigators will comply with the data management and dissemination policies described in the *NSF Award and Administration Guide* (AAG, Chapter VI.D.4) and the *NSF Division of Ocean Sciences Sample and Data Policy*.

Description of Data Types
The project will produce several observational and experimental datasets, described in the list below.

Observational Datasets:

1. **Metadata associated with genetic sampling:** Data collected on genetic samples in the course of new sampling. Data will include standard population demography measurements for each individual (size and sex), and site-specific information (location, time of sampling, person who conducted sampling, type of tissue sample collected, preservation method). File type: .csv. Repository: Biological and Chemical Oceanography Data Management Office (BCO-DMO).
2. **Temperature logs:** Temperature logs from HOBO data loggers; will include time and temperature readings, logger locations, and logger serial number. Will be recorded using the HOBO's internal logging capability and retrieved from loggers using HOBOware software. File type: .csv. Repository: BCO-DMO.

Experimental Datasets:

1. **Genetic sequencing:** DNA sequences for genotyping, derived from both new sampling and historical samples. Sample preparation will be performed at the PI's lab in Woods Hole, MA following the development of a targeted genotyping panel; final sequencing will be conducted at UC Berkeley's genomics core. File types: .fastq files. Repository: NCBI; accession numbers to be provided to BCO-DMO.
2. **Transcriptomic sequencing:** cDNA sequences for a subset of ~50 samples collected during new sampling. Sample preparation will be performed at Tepolt's RNA lab in Woods Hole, MA; sequencing will be conducted at UC Berkeley's genomics core. File types: .fastq. Repository: NCBI; accession numbers to be provided to BCO-DMO.
3. **Partial genome sequencing:** Assembled partial genome sequences of flanking regions around balanced polymorphisms, derived from MinION sequencing conducted at Tepolt's lab in Woods Hole, MA. File types: .fast5, .fasta. Repository: NCBI; accession numbers to be provided to BCO-DMO.
4. **Model output:** Three-dimensional particle trajectories. File type: .nc (netcdf) files. The model output will be freely available to the community and public per request. We will also explore the possibility of archiving/releasing the model output at BCO-DMO.

Data and Metadata Formats and Standards
Field observation data will be stored in flat, comma-separated ASCII files, which can be read easily by different software packages. Field data will include date, time, latitude, and longitude, as appropriate. Genetic and genomic data will be stored in .fasta (processed data), .fastq / .fast5 (raw data) files, or .csv files (called genotypes), standard genetic input formats which can

be read easily by all major genetic software packages. Particle tracking model data will be stored in Netcdf format. Metadata will be prepared in accordance with BCO-DMO conventions (i.e. using the BCO-DMO metadata forms) and will include detailed descriptions of collection and analysis procedures.

Data Storage and Access During the Project
The investigators will store project data (including spreadsheets, ASCII files) on laboratory computers that are backed up daily to an onsite external hard drive, and weekly to an offsite hard drive. Because of large file size, raw sequencing data will be stored on onsite external hard drives and offsite on WHOI's computer cluster. Personal computers used in this project will be backed up daily to an onsite external hard drive, and weekly to an offsite hard drive.

Mechanisms and Policies for Access, Sharing, Re-Use, and Re-Distribution
DNA sequences will be deposited in the National Center for Biotechnology Information (NCBI) GenBank database upon submission of manuscripts. GenBank accession numbers will be provided to the BCO-DMO in a .csv file and metadata will be provided using the BCO-DMO Dataset Metadata submission form. Data sets produced by the project will be made available through the BCO-DMO data system within two years from the date of collection. In addition to BCO-DMO, the PI will use the Dryad Digital Repository to store processed genetic data and associated metadata, as this is a publicly-available repository widely used by researchers in genetics. The PI will work with BCO-DMO data managers to make all project data available online in compliance with the NSF OCE Sample and Data Policy. Data, samples, and other information collected under this project can be made publicly available without restriction once submitted to the public repositories. Historical samples remain the property of their original collectors and will not be included in this sharing, though sequence data derived from these samples will be shared. Data produced by this project may be of interest to marine and evolutionary biologists, invasion biologists, physical oceanographers, and marine resource managers. We will adhere to and promote the standards, policies, and provisions for data and metadata submission, access, re-use, distribution, and ownership as prescribed by the BCO-DMO Terms of Use (http://www.bco-dmo.org/terms-use).

Plans for Archiving
BCO-DMO will ensure that project data are submitted to the appropriate national data archive. NCBI and Dryad will provide a long-term, publicly-available archive for the genetic and genomic data; Dryad will link these processed data with relevant metadata. The PI will work with BCO-DMO to ensure all data are archived appropriately and that proper and complete documentation are archived along with the data.

Roles and Responsibilities
Tepolt, Zhang, and Grason will be responsible for all aspects of implementing and monitoring this data management plan. Zhang will ensure appropriate management of the modeling data, Grason will ensure appropriate management of the observational data, and Tepolt will ensure appropriate management of the genetic and genomic data. Tepolt will also take overall responsibility for the managing and sharing of data from this project. The PIs will take responsibility to ensure data quality, metadata generation, and data archiving. In addition, the PIs funded on this project will ensure training of students and postdocs engaged in data entry and capture, including quality, generation, storage, and preservation.