Data Management Plan

The following data management plan complies with NSF policy on the dissemination and sharing of research results and data as described in Grant Proposal Guide (Chapter II). Our plan will conform to NSF requirements by sharing data with other qualified researchers and making all data public within two years of collection.

1. Types of Data

The products produced by this project will be short-read Illumina DNA sequences, and physiological data on larval growth, feeding, survival, energetic content. Additionally, larval trajectory data (location, depth, and time) will be generated from the biophysical larval dispersal modeling effort.

2. Standards for Data and Metadata

Raw DNA sequence data will be stored as fastq files and assembled contigs as fasta files. Metagenomic and marker gene amplicon metadata will be carefully formatted according to the Genomic Standards Consortium recommendations (MIMS, MIMARKS, and the draft recommendations of MINSEQE, respectively).

Developmental and biochemical data and metadata will be documented by taking careful notes in laboratory notebooks (that remain in the PIs' labs at all times), which will refer to specific computer-generated data files by unique names. These metadata will describe experimental conditions not logged, the names and units of logged data, and the identity of missing value identifiers. Metadata forms in Ecological Metadata Language (EML) will be utilized to create and organize the overall project database for local use, and upon publication for increased ease of data dissemination.

Biophysical transport modeling will utilize digital annotated Jupyter notebooks to develop and store code. These notebooks contain a change log to track version control and change history. The output from the model will be stored as NetCDF files and contain the particle location (latitude and longitude), depth, and time and date, and can be customized to contain other biophysical parameters of interest. The NetCDF files will contain all necessary metadata.

3-5. Policies for Accessing, Sharing, and Archiving Data

We will publish the outcomes of the research in leading peer-reviewed research journals and have included funds in the proposal to offset the costs of Open Access publishing when possible and/or affordable. All of the data will be deposited in public repositories to provide unrestricted access to samples and data.

Roles and responsibilities. PI Marko and co-PI Moran will oversee data management for the project. Marko will be responsible for the genomic data, Moran the physiological data, and Wren and Kobayshi will be responsible for numerical modeling and environmental data.

DNA sequence data. Sanger and short-read Illumina DNA sequences will be stored on the PIs' desktop computers/external hard drives and backed up on a university server. All sequences will be deposited on NCBI databases (http://www.ncbi.nlm.nih.gov/guide/all/#databases_). DNA sequence data will also be maintained by the PIs in a variety of formats specific to software analysis (e.g. FASTA, VCF, ARLEQUIN, IMa, NEXUS) for more rapid dissemination to interested researchers.

Physiological and biochemical data. These data will be stored on the PIs' desktop computers, backed up on a university server, and uploaded to a dedicated project site on the Biological and Chemical Oceanography Data Management Office (BCO-DMO) server.

*Scripts and bioinformatics pipeline*s. Perl and R scripts, will either be published as supplementary information for a publication or on Github, depending on the policy of the journal.

Biophysical model output. Model output data will be stored as netCDF files on the PI's desktop computer, and backed up on PIFSC servers. These data can be uploaded to the NOAA Public Access to Research Results (PARR) site (https://www.ngdc.noaa.gov/parr.html) (or however NSF prefers. As long as it's publicly available I dont think NOAA cares if it's on NSF database or PARR system).

Biological samples. The PIs will also catalogue and maintain tissue samples, and DNA extractions for future research or loan. The PIs maintain sufficient space for long-term temperature controlled wet (ethanol) and frozen sample storage. Field collection data will be catalogued and uploaded to the project site on the BCO-DMO server.