Data Management Plan

Data Policy Compliance

This project will comply with the NSF OCE Data and Sample Policy. The Biological and Chemical Oceanography Data Management Office (BCO-DMO) will be the primary place to serve our data. The genomic information will go into NCBI and linked with our data at BCO-DMO.

Description of Data Types

Genomics: This project will generate several TB of .fastq files from raw Illumina reads (MBD-BSseq, RNA-seq, EpiRADseq). Each Ilumina lane will produce two .fastq files (one for each paired end). These files will form the basis of subsequent analysis.

Phenotyping: This research will generate phenotypic data for several thousand larvae.

Ocean chemistry: Environmental data from the experiments will include pH, total alkalinity, dissolved inorganic carbon, and salinity.

Genetics samples: Genetic samples (Preserved tissues and DNA extractions) will be deposited into -80ºC and -20ºC chest freezers at our institutions and will be made available to researchers upon request.

Data and Metadata Formats and Standards

Metadata associated with this proposed research, including information on sites, experiments, and data collected (e.g., date, time, location, experimental treatments and maintenance, and environmental variables measured) will be documented for all data following BCO-DMO recommendations.

All raw Illumina data will include necessary and detailed metadata. While there is not a single common standard for short read sequence data, essential information including a description of the sample, library, sequencing method will be included in the SRA repository. Data tags will allow the data to be easily retrievable at NCBI.

Data Storage and Access During the Project

All data will be stored permanently and backed up on a 96 TB RAID array that Lotterhos has at Northeastern University. Within one month of acquiring the raw Illumina data it will be uploaded into the Short Read Archive at NCBI and linked to BCO-DMO. Raw data from DNA sequence platforms will be transferred to the lab of a collaborator (Roberts). In the Roberts lab, raw data is organized on a network attached storage (NAS) device with RAID redundancy. This NAS is open to the public (http://owl.fish.washington.edu/nightingales/). To make it easier for searching and discovery we also maintain a separate database including metadata and direct links to files (http://goo.gl/XxjTkW). Raw data from secondary procedures including mapping and genome feature analysis will be available in real-time on via online lab notebooks. Data will be in non-propriety formats such as tab-delimited text files. Limited analyzed data and workflows will also be made available via Galaxy and Cyverse as some analysis will take place on these platforms.

Mechanisms and Policies for Access, Sharing, Re-Use, and Re-Distribution

As described, raw data from short read sequencing and tiling array analysis will be uploaded to NCBI for archiving as well as providing access, and linked to BCO-DMO. All PIs are dedicated to conducting open research that is reproducible. Custom scripts and pipelines will be developed collaboratively and publically on GitHub, thus allowing transparency during the code development phase as well as access of completed code to other researchers.

Plans for Archiving

Within one month of acquiring the raw Illumina data it will be uploaded into the Short Read Archive at NCBI and linked to BCO-DMO. SQL databases will be used to organize the phenotyping and ocean chemistry data, which will also be deposited to BCO-DMO. For each publication, raw data and

reproducible pipelines (including scripts used to create publication figures) will be compressed and archived permanently at BCO-DMO. Within two years of data collection the primary data will be submitted to BCO-DMO following the Division of Ocean Sciences Sample and Data Policy.

Roles and Responsibilities

Lotterhos will be responsible for data management of the EpiRAD, RNAseq, and phenotyping data. Roberts will be responsible for the MBD-BSseq data. Ries will be responsible for the environmental data. Lotterhos will take the lead on ensuring that all project personnel comply with the data management plan.