DATA MANAGEMENT PLAN

Data Description

The proposed project will generate many types of data and resources. Data types include sequence data (Illumina sequencing), environmental data (temperature), settlement/survival measurements of larval stages for four species, and a collection of genotyped individuals from natural populations. Raw FASTQ files (containing base calls and their associated quality scores) from high throughput sequencing will be made freely available through NCBI Sequence Read Archive (SRA) along with their respective metadata. Assembled genomes and associated individual genotype files will be shared as a NCBI BioProject for each species. Data from sequencing will be primarily analyzed with publicly available software and pipelines (R, Freebayes, etc.). Custom scripts and analysis pipelines will be made available through GitHub. Data analysis and interpretation will be disseminated through peer-reviewed publications. Environmental data and processed alignments, polymorphisms, and mapped loci will be made available by the Biological and Chemical Oceanography Data Management Office (BCO-DMO) and Dryad for publications. All processed data (identification and quantification of genes in each sample; lists of loci inferred to be under selection and associated statistics) will be included in the database submission as well as published manuscripts. Temperature data collected by deployed loggers and species identified in field samples will be made freely available through BCO-DMO.

Data and Metadata Standards

Raw data from high throughput sequencing will be transferred from the sequencing core platform at NYUAD or an external provider to the PI's lab for analysis and storage. All data will be stored on hard-drives in the lab and office of the PI as well as back-up servers at UNC Charlotte and NYUAD. Sequence data will be annotated with essential information for each sample, which will include description of the sample (e.g., location of origin for population, date), method for library generation, and the sequencing method. These data tags for description will be included with data submitted to SRA and BCO-DMO. Nucleic acids extracted from collected animals will be labeled with precise site of origin (latitude, longitude), date of collection, and species and archived in the Reitzel and Burt laboratories. Field temperature data will be stored in flat ASCII files, which can be read easily by different software packages. Field data will include date, time, latitude, longitude, and temperature, as appropriate. Metadata will be prepared in accordance with BCO-DMO using the BCO-DMO metadata forms and will include detailed descriptions of collection and analysis procedures.

Data Storage and Access During the Project

All raw and minimally processed sequence data (FASTQ files) and temperature data will be backed up immediately on hard drives in the laboratory and university servers. As a safe-guard, raw sequencing data will be uploaded to NYUAD's off-site archival tape storage system immediately after generation. These data will also be uploaded to SRA and BCO-DMO. All other digital data (e.g., images) will be similarly stored on hard drives and the university network through shared drives on the network. Laboratory notebooks are stored in the lab. Use of electronic notebooks will be encouraged for those interested.

Data Access, Sharing, Re-Use, and Re-Distribution

All data will be made available through NCBI GenBank, NCBI SRA, BCO-DMO, and the PI's website as appropriate. Data analysis pipelines for each publication will be described in dedicated Github repositories following community standards (e.g. Puritz *et al.*, 2016; Matz *et al.*, 2018). This will promote re-use of the data and code generated in the project and ensure reproducibility. Sequence-based datasets will be linked with the environmental data through BCO-DMO. GenBank accession numbers and SRA projects will be provided to BCO-DMO in an Excel spreadsheet and metadata will be provided using the BCO-DMO Dataset Metadata submission form. We will share data from our project with existing databases for specific taxonomic groups (e.g., Reef Genomics database) as part of a community effort for continued improvement to access to genomic resources through publicly available sites where appropriate. Data will also be shared via publication and presentations. Papers will be published in open access journals when possible using funds designated in the proposed budget. The project investigators will work with BCO-DMO data managers to make project data available online in compliance with the NSF OCE Sample and Data Policy. Data sets produced by the research team will be made available through the BCO-DMO data system (in addition to other aforementioned public repositories) within two-years from the date of collection or at manuscript submission, whichever is sooner. Data, samples, and other information collected under this project will be made publically available without restriction once submitted to the public repositories.

Plans for Archiving

Sequence data will be archived at NCBI GenBank or SRA, code will be archived on GitHub and processed data files (e.g., vcf files) will be deposited in Dryad. BCO-DMO ensures that the data are archived properly at the appropriate National Data Archive for long-term archive preservation. The PI will work with BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data.