Data Management plan

1. Types of data

Genetic samples: Genetic samples include preserved tissue and extracted DNA.

Genetic data: Genetic data include DNA sequences and SNP genotypes

Field data: The fieldwork will generate data on the spatial location, growth, survival, and fecundity of *Bugula neritina* in the Northern Gulf of Mexico.

Computer code: This includes R code and model outputs from statistical analysis of the data collected in the field and lab, as well as scripts used to create .vcf files and filter SNP data.

2. Standards for data and metadata format and content

Metadata associated with this research will include information on sites, quadrat locations, sample collection date, time, location, and experimental treatments. All sequence data will include necessary and detailed metadata, including a description of the sample, library, and sequencing method. DNA sequence data will be stored in .fasta and .FASTQ files. SNP genotype data will be stored in .vcf files (variant call format) and summarized in .csv files (comma-separated values) (rows as sample identification, columns as SNP identification).

Statistical modelling work will be done in R. We intend to follow the Best Practices for Scientific Computing (Wilson et al. 2014). This entails writing annotated code that other people can later understand, using a version control system, embedding documentation, and reviewing code with collaborators. R Code will be integrated with a version control system such as GitHub, and stored on GitHub. Files will also be backed up to university servers.

3. Policies for access to and sharing data

All data will be stored permanently and backed up on a solid state hard drives at Florida State University. Following NSF policies, we will make all project data, including metadata files, genetic sequence/genotype data, data from field and lab experiments, R code and model outputs, available on publically accessible servers within two years of data collection. The exceptions will be data related to graduate thesis projects, which will not be made available until 12 months following graduation. We intend to publish our computer code on a data archive such as Dryad, regardless of whether it is required by the journal.

4. Polices and provisions for re-use, redistribution, and production of derivatives

All users will have open and free access to our data within two years of collection, unless otherwise embargoed to meet the needs of graduate thesis preparation. Although not required for access, we will encourage all users to agree to acknowledgement access to our data and make contact with the lead PIs in the spirit of effective collaboration.

5. Archiving and access to data

We will deposit data into the Biological & Chemical Oceanographic Data Management Office (BCO-DMO), to be made publically accessible within two years. We will work towards assigning DOI values as manuscripts are submitted.

Genetic samples will be deposited into -80ºC and -20ºC freezers at Florida State University.

Unadulterated DNA sequences and SNP genotypes will be deposited in public databases such as Dryad Digital Repository, Genomic Observatories Metadatabase (GeOMe), or NCBI Sequence Read Archive.

R Code will be stored on GitHub.

Wilson G, Aruliah DA, Brown CT, Chue Hong NP, Davis M, Guy RT, et al. (2014) Best Practices for Scientific Computing. PLoS Biol 12(1): e1001745. doi:10.1371/journal.pbio.1001745