Data Management Plan

1) **The types of data, samples and physical collections, and derived models produced in the course of the project.**
Data generated from this project will be of three forms: (1) **Genomic/transcriptomic sequence data** (2) **non-genomic data** (including lab, field, and modeling records) and 3) **model derivations** and data.

Transcriptomic data (including raw read files from the Illumina sequencing platform) will be stored as compressed FASTA/FASTQ files stored on redundant RAID storage devices. This will include raw data (read files) from RADseq, EpiRADseq, and RNAseq data collection proposed in the Project Description. Genomic sequence data will be maintained in direct association with paired metadata providing full details of experiments that gave rise to these data. These metadata for all samples for which genetic sequence data is generated for in the grant will accompany public data depositions (NCBI).

This proposal will also yield **non-genomic data**, including field, experimental disease transmission data, traits and modeling results. We will build the trait database in Excel spreadsheets that will be saved as comma-separated value (.csv) files and uploaded to a server daily. All datasets will be annotated with meta-data. The same procedure will be utilized for the data generated by our **modeling approaches**. Data collected from the each annual survey on the abundance and health condition [healthy, diseased (type of disease), injured] for each major reef-building coral species and four abundant octocoral species will be entered in Microsoft access or excel spreadsheets, organized, edited for errors and stored for later analyses. A Metadata file will be produced to have all relevant information on the methods and data collected readily available. **Mechanisms for access and sharing data**

Transcriptomic sequence data comprised mostly of raw read files from the Illumina sequencing platform will be stored as compressed FASTA/FASTQ files stored on redundant RAID storage devices. These RAID devices will include redundant drive striping and a reciprocal backup in house, and an off-site backup. In addition to raw data, we will also maintain copies of intermediate datasets generated during analyses (e.g., transcriptome assemblies) in FASTA form, or other appropriate standard filetype. To facilitate archiving, files will be highly compressed (e.g. bzip2) for mid-long term storage. Ultimately all raw files will be permanently archived in the NCBI Gene Expression Omnibus and Sequence Read Archive. The short-term data storage plan for the **non-genomic data** generated by the experiments and surveys will saved as metadata files and Excel spreadsheets (saved as .csv files) to an external drive and a UTA server that is backed up nightly. All modeling data will be saved to external drives daily.

In addition, appropriate field and experimental data (with links to genomic data) will also be deposited in The Biological and Chemical Oceanography Data Management Office (BCO-DMO) housed at Woods Hole, MA (http://bco-dmo.org).

2) **Policies and provisions for use of data**
The datasets, databases including raw and annotated sequences (as described above) will be available for download from the Mydlarz Lab webpage (http://www.uta.edu/biology/mydlarz/index.htm). Further promotion of the databases will occur through the Coral List Serve (http://coral.aoml.noaa.gov/mailman/listinfo/coral-list/), as well as through the sites of some of the other open access coral genomic and transcriptomic data, such as SymBioSys (http://sequoia.ucmerced.edu/SymBioSys/), Matz lab data (http://www.bio.utexas.edu/research/matz_lab/matzlab/Data.html) and Meyer lab (http://people.oregonstate.edu/~meyere/data.html). There will be no charge for the data.

3) **Plans for archiving data, samples, and other research products, and for preservation of access to them.**

Our server (coralimmunity.uta.edu) is hosted at the Arlington Regional Data Center University of Texas Arlington it is a 32 cluster system with 196GB of memory and 6TB of storage running RedHat. If funded we will upgrade the memory by 200GB and storage by 8-10TB. Currently it has installed: Trinity, TopHat, Cufflinks, Trimmomatic, R, Perl, BioPerl, Biopython, deconseq, blast and Rnannotator, Velvet and other packages. The website is UTA commons data (http://dspace.uta.edu/).

As part of the University of Texas system, we also have allocated space and essentially unrestricted access to the Lonestar Linux Cluster at Texas Advanced Computing Center (TACC). As a leading resource provider in the NSF XSEDE project, TACC is one of 11 centers across the country providing leadership-class computing resources to the national research community, including dedicated computational biology support. The cluster consists of 1,888 compute nodes, with two 6-Core processors per node, for a total of 22,656 cores. It is configured with 44 TB of total memory and 276TB of local disk space. The theoretical peak compute performance is 302 TFLOPS. In addition, Lonestar provides five nodes, each with six cores and 1TB of memory. We regularly use our allocated space on Lonestar for next gen sequence analysis, especially de novo transcriptome assembly. Therefore, there are many options for data storage and many resources available to safely store, backup and effectively analyze and share data. As PI, Mydlarz will oversee the data storage and sharing plan is properly executed.

In addition, appropriate field and experimental data (with links to genomic data) will also be deposited and archived in The Biological and Chemical Oceanography Data Management Office (BCO-DMO) housed at Woods Hole, MA (http://bco-dmo.org).