Data Management Plan

Introduction: Data management will be coordinated by PI Gifford and carried out by all participants. All data and metadata generated as part of this project will be appropriately stored, organized, and disseminated to current data management standards as detailed here. We will leverage existing systems wherever possible. All data will be made available in a timely fashion following NSF Division of Ocean Science's Sample and Data policies.

1. **Types of data to be generated:** Several types of data will be generated in the proposed activities. In the lab, types of data include biochemical, physiological and molecular measurements. Chemostat runs will generate cell density data, substrate utilization, environmental conditions. Field experiments will generate data on environmental conditions (temp, salinity, DOC concentrations). Bioassays will generate data on cell densities and DOC drawdown. The transcriptomic experiments and metatranscriptomes will generate data on RNA yields, RNA quality, RNA fragment distributions, and nucleic acid sequences.

2. **Standards used for data types:** Most physiological and metabolic data to be generated is relatively small scale and will be stored in spreadsheets. Cell concentrations will be monitored both by plate reader absorbance (data directly exported to spreadsheets) and flow cytometry (individual event records stored as .fcs file and total region events counts exported as .csv files). Given the relatively large number of chemostat runs proposed, we will create a template that incorporates all relevant measurements (conditions, cell density, environmental parameters) into a single .csv file, which will then be used to standardize the collection and storage of chemostat run results. DOC concentrations are exported from the Shimadzu analyzer as .csv files. We will conform to the metadata standards established by the BCO-DMO. As much as possible, data will be archived in ASCII format, which is the most flexible and readable over the long term, though some environmental time series data may be transferred to BCO-DMO in more native formats.

Raw sequence data from the transcriptomes and metatranscriptomes will be in fastq files. After quality control, read data will be maintained in fasta format. The sequence data will be processed via Galaxy's bioinformatic platform on the Gifford lab server as well as custom scripts and open source programs. Output from BLAST and Diamond homology searches will be in m8 tabular files. Mapping of transcriptomic reads to reference genomes will be in SAM and BAM file formats, and transcript tag counts output in .csv format. All sequence data will be maintained on the Gifford lab's server. The processing of individual samples through the lab's workflows will be tracked using Galaxy histories. Cellular transcript abundance counts arising from the quantitative transcriptomics work will be stored in a standardized .csv format.

3. **Investigator role**: Team members involved in the project will collect and store data individually (with the exception of the sequence data (raw and processed) which is centrally stored on the lab's server), but all data will be regularly transmitted to PI Gifford who will maintain the data during and after the grant. PI Gifford will maintain a copy of all data generated in case any key personnel depart from the project. All sequence data and downstream processing files will regularly be backed up to external storage.

4. **Data dissemination methods**: Data associated with publications will be submitted to online databases and archives as appropriate, and small data sets will be included in publications. Any requests for published data or data associated with a publication from the investigators will be honored. Small data sets such as metabolic or growth associated data will not be large and can be sent to the requesting individual via email, dropbox or similar service, or on a USB drive

depending on the needs of the requesting individual and size of the dataset. Any custom codes used in processing data will be made available on the PI Gifford's website. Data will be transferred to Biological and Chemical Oceanography Data Management Office (BCO-DMO) following processing, and public access will be granted to data following its publication or at most two years after its collection. Raw transcriptomic and metatranscriptomics reads will be deposited in NCBI's Sequence Read Archive (SRA). A central aspect of the proposed work is creating a standardized database of genome-wide, cellular transcript abundance which will be made publically available through the Gifford lab's server.

5. **Policies for data sharing, public access, and re-use**: We are committed to making all data types publicly available through peer-reviewed publications and public databases with as few restrictions as possible. As sequences and samples are analyzed, data will be processed, and raw and processed data will be uploaded to the Gifford lab server which is backed up monthly and can be made readily available to collaborators needing access to the data through the Galaxy web interface.

6. **Plan for archiving data, sample, software, and other research products**: As much as possible, data will be submitted to BCO-DMO as well as other publically available data archives (Genbank, IMG, etc) from which the public can freely access and use the data. Sequences obtained through Illumina high-throughput sequencing platforms will be stored on the Gifford lab's server and deposited in public sequence databases (NCBI SRA, IMG). Data that cannot be archived in public archives will be maintained by the PI using individual laboratory and university computer resources. Samples (RNA, DOC, metabolites, etc) will be maintained for several years in appropriate storage conditions. All other data will be archived at BCO-DMO in adherence with the Division of Ocean Sciences Sample and Data Policy. In all our efforts we will work with the BCO-DMO to archive the data and to ensure our metadata conform to their standards.