Data Management Plan

Overview: PIs Morris, Hennon and Dyhrman have a track record of working together on phytoplankton, heterotroph and CO_2 responses. If funded, data collection and quality control will follow rigorous internal guidelines. The research proposed herein will be carried out and published in peer-reviewed journals in a timely manner. Prior to publication, all data will be stored on multiple computers and routinely backed up both locally and at an off-site server. After publication, all data necessary to confirm our results will be archived in an accessible manner. Most data generated will be archived at BCO-DMO, where it will be freely available to the public. This includes primary research data in flat-format (.csv) spreadsheet form; field physicochemical data; statistical analyses and full instructions on how they were generated, and in what software package; and any software that is written for this project along with instructions on how to install it, compile it, run it, and so forth. Bioinformatic and statistical workflows will be provided as .txt files in a format that can be copied and pasted into an appropriate command prompt to regenerate all of our results from publically available datasets. Links to sequence reads will be provided via NCBI GEO and or the SRA. Protocols will be additionally be deposited at git-hub as appropriate. All primary research documents such as laboratory notebooks will be permanently stored in our laboratories.

Data release: Prior to publication, all appropriate field data products associated with the sequencing treatments will be submitted to BCO-DMO as highlighted above with a special link to the sequence archive. Transcriptome and metatranscriptome sequences from cultures and from both field and incubation samples will be uploaded to the NCBI Gene Expression Omnibus (GEO) and or the Short Read Database (SRA) as appropriate and linked to both BCO-DMO and a single bio project number. Submission to GEO will include annotations and differential expression data from the metatranscriptome comparisons, as well as a link to the raw data in the SRA.

Data archiving: If awarded, upon receipt of the award we will also contact the Biological & Chemical Oceanography Data Management Office (BCO-DMO) to register our project. As soon as field sampling is completed we will submit all data collected to BCO-DMO for archiving for archiving in a searchable project format. We will keep NSF abreast of our compliance with data management through our annual reports and all data will be made available as expeditiously as possible. Through the culture and field metagenome and metatranscriptome sequencing and subsequent analysis, we will be generating and storing significant amounts of sequence data and the associated analytical files. The data will be stored on two redundant 15TB Raid 5 servers, which are backed up weekly, to the Columbia server network. In this manner there is redundancy in preserving the raw data and the associated analytical files. Data analysis will be performed on a custom pipeline run via Dyhrman's NSF supported XSEDE network access at the National Center for Genome Analysis on their Mason cluster. Mason at Indiana University is a large memory computer cluster configured to support data-intensive, high-performance computing tasks. It is populated with genomics software intended for use by researchers using genome assembly software (particularly software suitable for assembly of data from next-generation sequencers as proposed here) and other genome and transcriptome analysis. We have several strategies for data archival. Raw data will be included as supplementary material to these publications when applicable, and will be available to the public upon publication through BCO-DMO as well. Further data will be archived with GEO and the SRA as appropriate with a link to BCO-DMO metadata.

All strains and cultures used will be permanently cryopreserved in liquid nitrogen. The Morris lab has been very successful in cryopreserving and resurrecting both bacteria and microbial eukaryotes, including the fastidious organisms common to the open ocean. We have an

automated liquid nitrogen vapor storage freezer with more than enough capacity for the samples described here. These samples will be preserved in triplicate, documented in an easily searchable database, and will be available to qualified researchers upon request through a query to the PIs.