DATA MANAGEMENT PLAN

Data description: The cruise is planned for the R/V Endeavor, and all PIs have experience as a UNOLS fleet Chief Scientist including associated data collection, calibration, archival and dissemination in accordance with NSF policy. At each station CTD casts will be made and standard chemical, biological, and physical oceanographic data (e.g. salinity, temperature, pressure, chlorophyll fluorescence, photosynthetically available radiation, oxygen etc.) will be collected via the CTD sensor package. Additionally at each station, biomass will be collected for downstream metatranscriptome analyses and phytoplankton community structure (samples for microscopy and molecular diversity analysis). Isolates of the numerically and/or biomass-dominant diatoms will also be generated for further experimental work in the lab. At each process station, incubation experiments using whole communities will be conducted. Each incubation experiment will consist of triplicate grow-out incubation treatments and a control where in situ communities are amended with a separate nutrient in each treatment (N, P, Si, deep water). From these experiments we will harvest biomass for metatranscriptome analysis, and related physiological and biogeochemical parameters. In short there are three major data types **1) field metadata, 2) sequence data, and 3) biological cultured isolates.**

Data release: 1) Plans for CTD station locations, instrument deployment, water sampling, and incubation related samples will be written up as a science implementation plan for the cruise. As feasible, the actual sampling events will be recorded on the Rolling Deck to Repository (R2R) ELOG scientific event logger system. This system greatly streamlines the uploading of CTD data to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) website, where it is immediately accessible to cruise participants and the broader scientific community. The original underway data will be contributed by the vessel operator to the UNOLS central data repository at http://www.rvdata.us/catalog/ managed R2R. Also, R2R will ensure that the original underway measurements will be archived permanently at National Ocean Data Center (NODC) and/or National Geophysical Data Center (NGDC) as appropriate for the data type. If R2R is not available then cruise metadata will be released to the Biological & Chemical Oceanography Data Management Office (BCO-DMO) as the data are published. **2)** All appropriate field data products associated with the sequencing treatments will be submitted to BCO-DMO (http://www.bco-dmo.org/) 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. 3**)** Novel cultured isolates will be a) submitted to anyone who places a request pending sufficient material and b) archived in the Provosoli-Guillard National Center for Marine Algae and Microbiota (NCMA—formerly CCMP).

Data archiving: **1)** All biological and chemical data collected in the field (e.g., nutrients, chl a, etc.) will be archived in the National Oceanographic Data Center (NDOC) database (www.nodc.noaa.gov). 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 from the CTD as project metadata to BCO-DMO for archiving as highlighted above. Physiological and community structure data will be assembled and organized in electronic spreadsheets and stored on local and backup servers, prior to submission to BCO-DMO. We will submit all data upon publication to BCO-DMO 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. **2)** Through the culture transcriptome and field 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 URI and 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 (mason.iu.xsede.org) 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. We are actively exploring informatics and data management solutions to the unprecedented challenges presented by these large sequence datasets in an earth systems context. For example, PIs Dyhrman and Jenkins both participated in the NSF workshop "Engaging the Ocean Science Omics Community in Envisioning Cyberinfrastructure to Archive, Access and Analyze Massively Parallel Environmental Omic Data" and Jenkins is on the steering committee of a new Research Coordination Network focused on data management of 'Omic data, that was born of this workshop. We will explore additional avenues for long-term archiving the data at other venues where there are appropriate metadata repositories as alternatives become available. **3)** Diatom species isolated from our field studies will be cultured in at least two labs as a bulwark against their loss. Further, cultures will be deposited in the Provosoli-Guillard National Center for Marine Algae and Microbiota (NCMA—formerly CCMP) whenever possible.