## Data Management Plan

A cruise is proposed on the *R/V Knorr* during which CTD casts will be made and standard chemical, biological, and physical oceanographic data will be collected via the CTD sensor package. Datasets will include but are not necessarily limited to: salinity, temperature, pressure, chlorophyll fluorescence, photosynthetically available radiation, and oxygen. Detailed plans for CTD station locations, instrument deployment, water sampling strategy and water sample allocation will be written up as a science implementation plan for the cruise. The actual sampling events will be recorded on the Rolling Deck to Repository (R2R) ELOG scientific event logger system. The Van Mooy lab was one of the very first users of this system on a recent cruise aboard the R/V *Knorr*. 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. In the case of our recent *Knorr* cruise, CTD data was published and freely available to the public in less than two months (http://mapservice.bco-dmo.org/mapserver/mapsol/ index.php?datasetId=14056). 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 the National Ocean Data Center (NODC) and/or the National Geophysical Data Center (NGDC) as appropriate for the data type. All data will be made available to the general public as expeditiously as possible.

Additional water column biogeochemical measurements made by the science party will be managed by the BCO-DMO and will be available online from the BCO-DMO data system (http://bco-dmo.org/data/). Full descriptions of methods and standards will be provided to BCODMO upon submission of datasets. The BCO-DMO will also archive all the data they manage at the appropriate national archive facility, such as NODC and NGDC. Experiments will also be conducted during both cruises. The results from these experiments will be made available through peer-reviewed, open-access publications. Raw data will be included as supplementary material to these publications when applicable, and will be available to the public upon publication. Samples will also be taken during the cruise. Most of these will be destroyed in the course of our planned analyses, but any excess pristine samples will be preserved and maintained for future analysis by interested parties for one-year after the project is completed.

Through the 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 a 15T Raid 5 server, which is backed up weekly, to the WHOI and to the Columbia server network. We have one large workstation for processing the sequencing files, are requesting a second because our current one is heavily used; in this way we will have a designated work station for the Post Doctoral Investigator (see budget justification). 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 through a computer workstation and a CLC Genomics workbench with updated databases and our own custom databases from prior work. These bioinformatics efforts are additionally supported by the Columbia Genome Center and the Columbia Center for Computational Biology and Bioinformatics. We have several strategies for data archival. All appropriate field data products associated with the sequencing treatments will be submitted to BCO-DMO (http://www.bco-dmo.org/). Metagenome and metatranscriptome sequences from both field and incubation samples will be uploaded to the Community

Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) database (http://camera.calit2.net/), so the community can compare our data with existing and newly emerging environmental datasets as well as NCBI through the Gene Expression Omnibus (GEO). Submission to GEO will include annotations and differential expression data from the metatranscriptome comparisons. In addition, we are actively exploring informatics and data management solutions to the unprecedented challenges presented by these large sequence datasets. We will explore avenues for long-term archiving the data at other venues where there are appropriate metadata repositories as alternatives become available.