DATA MANAGEMENT PLAN

The main objective of this proposal is to determine the phylogenetic and functional properties of microbial communities in the deep euphotic zone by applying next generation short read sequence technology to generate metagenomic sequence libraries from a suite of samples from the South Atlantic Ocean.

Data management issues fall into three main areas.

1) **Raw sequence data:** As described in the proposed work, we expect to generate ~50 Gbp of short read (50bp) sequence data from each of ten samples from the South Atlantic. It is highly beneficial for the data to be archived and publicly available in this format so that other researchers can apply different assembly methods or perform assembly-independent queries on the full dataset in the future. Unfortunately the archive of record for this type of data, the sequence read archive (formerly the short read archive) at NCBI, has recently announced they will no longer be accepting submissions due to budget constraints (http://trace.ncbi.nlm.nih.gov/Traces/home/). In its absence, the Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis (CAMERA) may the most appropriate archive for this data, and we will work with them as it is generated to determine if they will be able to host this volume of raw data.

2) **Assembled and annotated sequence data:** Assembled and annotated contigs from the meta-genomes will be deposited into NCBI's GenBank database where they will be publicly available. Other commonly used genomic tool resources (i.e. CAMERA, IMG/M) are routinely updated with new information from GenBank and thus the assembled data will also be accessible through these sites.

3) **Ancillary data:** The CoFeMUG cruise is already registered with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) (see http://osprey.bcodmo.org/project.cfm?id=34&flag=view) as is the publicly available data we have generated from this cruise to date (metaproteomics and 16S rRNA clone libraries). We will continue to upload data and metadata to this site as additional datasets are finalized. For this proposal in particular, BCO-DMO may not be the best archive for the primary sequence data, but we will deposit a dataset record with them to alert the community that the metagenomic libraries exist, communicate the metadata, and point to the location where the data itself can be obtained.