DATA MANAGEMENT, ANALYSES AND INTEGRATION

Buchan-Campagna

Data for the current project will be generated at UTK, where it will be backed up immediately to our central servers. Data will be managed among the research partners within a framework designed to integrate among measurements, disseminate results, and enable hypothesis testing. Our overall goal is to optimize the availability and utility of data to all members of the research team, facilitating communication and ultimately the publication of results. Both PIs will plan the overall sampling campaign. The actual sampling events and experiments will be recorded on paper logs (scanned into PDF documents), samples split and prepared for the required analyses.

Data analyses will involve comparisons within and between measurements and experimental data sets. Data analysis will be carried out in the framework of "reproducible research" (e.g. Mesirov 2010 Science 327: 415-416). While each individual lab that generates the various data streams will be responsible for maintaining records of data quality (standard curves, measures of analytical error, etc.), the collated data will also be screened for anomalies. Where possible, re-analyses of archived samples will be completed to check anomalous values. Possible outliers included in the final, submitted data set will be flagged to alert subsequent data users.

Raw data will be processed to the extent possible using script-based software environments, including the R Statistical Package and MATLAB (Some molecular data will likely need to be processed using non-script-based software packages, *i.e.,* The CLC Workbench). During year 3, the all-project database (including all chemical and physical analyses) will be sent with relevant metadata to the ***Biological and Chemical Oceanography Data Management*** (BCO-DMO) repository. With respect to distribution to the scientific community, sequence data will be deposited in the ***National Center for Biotechnology Information*** **(NCBI)** and the *Metagenomics Analysis Server* (MG-RAST) which will allow us to couple molecular and site specific data for use by other researchers. Taxonomic composition will be included in all documentation (*i.e.,* often as supplemental data sets to papers). After year 3 of the study, we will make residual nucleic acid data, metabolomics and lipidomics data publicly available upon request through a web/email interface. Scripts for data analysis will be published as supplemental information in peer-reviewed reports of our work, so that outside workers will have all the tools necessary to audit our data analysis procedures and recreate our findings. Prior to submission to any of the aforementioned repositories, all data will be collated and managed at UTK, which will provide a backup to the data storage systems.

Date Use, Privacy and Sharing Policies.
Any data and software generated will be open access and made available for educational, research and other non-profit purposes. Distribution will be via participant websites, and the BCO-DMO, NCBI, MG-RAST archives, as appropriate, where products will have associated metadata and unique DOIs. Hyperlinks to products, as well as how to cite them, will be noted on project websites, in publications, and in metadata. Intended and foreseeable users are marine microbial ecologists, organic geochemists, aquatic scientists and microbial physiologists.