DATA MANAGEMENT PLAN

Data acquisition and quality control. Before each field collection (cruise), the research team will meet (in person or via teleconference) with ship support coordinators to plan the field objectives, sampling strategy, and to prepare for experimental analyses of field-collected specimens.

Our proposed research activities will generate detailed physiochemical, biochemical, and meta-omic sequence data from field collections and laboratory experiments. Physiochemical data will include shipboard CTD-based measurements of conductivity, temperature, depth, fluorescence, and dissolved oxygen, as well as measurements of nitrogen (nitrate, nitrite, ammonia) and sulfur (sulfide, thiosulfate). Biochemical data will include measurements of rates of key nitrogen and sulfur transformations. Analytical checks using measurements of external standards and no-specimen controls will be done before and after field collections and experiments. Chemical data will also be submitted to a rigorous quality control procedure before, during, and after acquisition. Conventional chemical measurements will include calibrations before and after the measurements to check for analytical consistency as well as quality control checks (measurements of accuracy) using certified reference materials when possible and blanks for contamination controls. All 'omic data will be filtered using default protocols associated with the sequencing platform. Additional quality filtering will be imposed using protocols established in the Stewart lab, or via downstream analytical processing (e.g., chimera checks).

Dissemination of datasets to publicly accessible data repositories. Shipboard underway data will be deposited by vessel operators at http://www.rvdata.us as soon after a cruise as possible. Processed physical, chemical, and rate data and associated files and observations will be archived for long-term storage on servers at Georgia Tech, with access via http://omz.biology.gatech.edu/ with access upon request. These files will also be archived and managed by the Biological and Chemical Oceanography Data Management Office (BCO-DMO) and the data sets will be available online from the BCO-DMO data system (http://bco-dmo.org/data/). Processed data and associated read-me files will be submitted to BCO-DMO within a year of generation. Project status reports will also be archived in BCO-DMO on a regular basis.

Rapid dissemination of sequence data and associated metadata will be a priority in this project. Our proposed sequencing involves 3 full-plate 150X150 paired end sequencing runs on the HiSeq2000 over 5 years (150 bp paired-end technology will be available in Q3 of 2011). At current specifications, a single paired-end run on the HiSeq2000 generates up to ~80 million sequence reads per flow cell lane (though the company advertises 300 million), which equates to ~560 (84 Gbp) per run (252 Gbp total). We also anticipate generating ~5 Gbp of sequence during rRNA gene amplicon sequencing using the Illumina MiSeq instrument in the Stewart lab, as well as 1-5 Mbp of Sanger sequencing-based data (marker gene clone library analyses). Following quality assurance filtering (automatic plus manual filtering with custom scripts), Illumina data (uncompressed .srf files, or compressed Fastq files) will be archived in the NCBI Sequence Read Archive (SRA), assigned a BioProject ID, and made publicly accessible **within one year of generation**. All submissions will be annotated with detailed environmental descriptions (project description, lat/long, date, habitat type, depth) and with brief summaries of associated physiochemical and nutrient conditions. These submissions will be appended with instructions on how to access the full meta-datasets, including CTD data (as Excel files), either directly from the PI or from research vessel websites. Additionally, .pdf copies of all protocols used in the generation of sequence data (if not prohibited by manufacturer copyright restrictions) will be linked to the data submissions, either directly or via instructions for accessing copies on the project website (http://omz.biology.gatech.edu/) or PI website (fjstewart.org). Sanger-based sequences will be submitted to NCBI's GenBank and annotated with associated domain characterizations and metadata as above.

The proposed research may also generate cultured isolates of marine bacteria. If so, subcultures and supporting descriptions of growth conditions will be made available to colleagues upon request. If obtained in pure culture, select strains may also be deposited in the American Type Culture Collection (ATCC), along with full descriptions of isolation conditions, phenotype, and genotype (e.g., NCBI accession numbers), according to ATCC guidelines.