DATA MANAGEMENT PLAN

Dissemination of data between partner groups. Regular meetings (phone/skype) will be held between collaborators (Stewart, Glass, Thamdrup, Girguis labs) to discuss on-going research and provide opportunities for integrative discussions. In addition to presenting research results at joint lab meetings between the PI and Co-I, students and post-docs involved in this project will be expected to present at national and international meetings and to submit research manuscripts to peer-reviewed journals. To avoid problems associated with e-mailing large datasets, raw data, data summaries, and presentation files will be made accessible to project team members via a project-specific website maintained on servers in the School of Biology: http://omz.biology.gatech.edu/ This site currently provides a data archive for existing OMZ projects in the Stewart lab.

Data acquisition and quality control. Before each field collection/cruise, the research team will meet (in person or via teleconference) to plan cruise objectives, sampling strategy, and to prepare for experimental analyses of field-collected specimens. Data and observations obtained in the field will be archived after each collection and made accessible via the project website (above).

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 methane. Biochemical data will include measurements of rates of nitrogen and methane 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 at http://www.bco-dmo.org/project/. 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 metatranscriptome sequencing involves multiple runs on Illumina HiSeq and MiSeq instruments, ultimately generating hundreds of gigabases of sequence over the three-year period. Rapid dissemination of these data to the broader community will be a priority in our project - all sequence data will be made publicly accessible within one year of generation. Following automatic quality assurance filtering on the Illumina system, demultiplexed raw sequencing data with combined quality scores (FASTQ format) will be archived and stored on servers at Georgia Tech and for public access in the Sequence Read Archive (SRA) at the NCBI. SRA data will be assigned a single BioProject identifier with linked metadata. Our submissions will be annotated with detailed descriptions of the sampled environment or experimental treatments (project description, lat/long, date, habitat type, hydrography, chemical measurements, etc), including brief summaries of any associated physiochemical variables. 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 PIs website. 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 n-damo 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.