DATA MANAGEMENT PLAN

Dissemination of data between partner groups. Regular meetings (phone/skype) will be held between collaborators (Stewart, Konstantinidis) to discuss on-going research and provide opportunities for integrative discussions. In addition to presenting research results at joint lab meetings between the PIs, 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 Biological Sciences at Georgia Tech (***Stewart is adjunct at Georgia Tech and retains access to computing resources***). This site currently provides a data archive for existing OMZ projects in the Stewart lab. In addition, during the project, team members will exchange data between labs and among external collaborators using Dropbox. Georgia Tech has a campus-wide license for Dropbox, providing unlimited, ITAR-compliant cloud storage space for each faculty member and his/her immediate collaborators.

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 dissolved organic matter (DOM). Biochemical data will include measurements of rates of nitrogen 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 ***and*** Montana State, with access via the project website 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/, with project ID number assigned pending initiation of the grant. 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 (omic) data and associated metadata will be a priority in this project. Our proposed metatranscriptome sequencing involves multiple runs on Illumina NovaSeq and PacBio 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. We will fully abide to the Minimum Information about a Genome Sequence and Metagenomic Sequence standards that have been recently established by the scientific community (MIGS and MIMS, respectively). 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 Montana State 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. All analysis results will be stored using internal data formats within the utilized software packages and common formats such as TIFF, ASCII, and Excel. Whenever possible, we will use portable data formats (such as ASCII files) to facilitate data exchange and subsequent analysis. Phylogenetic trees will be stored in Newick format, and any multivariate analysis data will be encoded both as Excel files and as R/Matlab binary images. ***All accession numbers will also be archived via BCO-DMO, which will serve as a central site for identifying omic datasets generated by this project.***

The proposed research may also generate cultured isolates of marine SAR11 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.

Plans for archiving and preservation of access. Digital data copies will be kept in the labs of the PIs for 10 years past the lifetime of the project. The anticipated data occupies a relatively small amount of space relative to our computing capabilities and storage resources; thus, long-term preservation will be easily accomplished by keeping several copies of the data on local computers at Georgia Tech and Montana State.