Data Management Plan

I. Types of data, samples, physical collections, software, derived models, curriculum materials, and other materials to be produced in the course of the project;

Both oceanographic and laboratory-based sequencing will be conducted. Records will comprise oceanographic data from CTD casts, flow cytometry analysis, quantitative PCR of rRNA gene copies for 6-10 gene assays, and V1-V2 16S rRNA gene library sequencing. Field data will be submitted to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) (http://www.bco-dmo.org/). Sequence data will be submitted to the National Institutes of Health's NCBI (http://www.ncbi.nlm.nih.gov/). Flow cytometry data will be shared in the Flow Repository (http://flowrepository.org/). Software scripts for analyzing flow cytometry data will be made available through a project website hosted at the co-PI's laboratory website, publications, and the request of other researchers. In addition, we will provide other researchers with samples generated during the project when possible. We will also generate classroom material as part of our Broader Impacts effort (PowerPoint presentations, Excel spreadsheets, Test templates, Handout materials etc.), curricula, and lesson plans that will be available through our laboratory website: http://baliga.systemsbiology.net/drupal/education/.

II. Data and Metadata Standards

Meta-data in the form of automated output from flow cytometry, qPCR, and sequencing sensors and log files will be stored in different formats (such as tab-delimited, fasta, fastq, BAM, xls, nexus, msf etc.) to ensure compatibility between different programs and databases. We will collect contextual details such as sampling time, horizontal and vertical location, oceanographic metadata, technology (e.g paired end vs. single end), parameters for 16S rRNA gene sequence analysis, alignments (e.g log files), and search databases along with other metadata standards. We will adapt meta data standards that are compliant with MIxS (Minimum Information About any (X) Sequence) standards (http://gensc.org/projects/mixs-gsc-project/). Flow cytometry data will be collected according to the Minimum Information about a Flow Cytometry Experiment (MIFlowCyt) standards. All education material developed in this project will be aligned to state and national standards.

III. Policies for access and sharing and provisions for appropriate protection/privacy

All raw and processed data in the form of web accessible databases, Python scripts, R objects, java packages and in other forms will be available on through indexed public websites and databases, our own website (http://baliga.systemsbiology.net/) and freely accessible through several mechanisms upon publication. Data privacy, confidentiality and security will be adjusted according to publication timeline.

IV. Policies and provisions for re-use, re-distribution

There will be no permission restrictions on data and they will be open for sharing and re-use upon publication.

V. Plans for archiving and Preservation of access

Data and research products will be archived to publically available databases and scientific journals. Samples remaining after analysis will be stored in -80 °C freezers while data collection

and publication are in progress. If samples remain valuable, they will be stored for longer periods following publication. Collected and generated data will be curated both automatically, manually and computationally before made available. Periodic data backups and submission to online repositories will be done according to suggested journal and archive guidelines. All acquired metadata and protocols will be documented and available in the form of references and research papers. Data preservation beyond the life of the project will be determined on case-by-case bases and will be backed up on tapes for long-term storage. We also publish all of our data on the co-PI's website: http://baliga.systemsbiology.net.