

## Data Management Plan

The proposed project will generate several different kinds of data. The most important data for archival purposes are molecular in nature, but data from field experiments and isotope measurements will also be acquired. Most of these data will be deposited in public databases (see below), and we will abide by the documentation and sharing requirements and policies of those repositories.

1. Oceanographic data: Hydrographic data, measurements of dissolved inorganic nutrient concentrations, pigment measurements, phytoplankton microscopy, and underway flow cytometry data at the two field sites, Sargasso Sea (SS) and Subarctic North Atlantic (SNA), will be deposited with NODC as required, and also at the Biological and Chemical Oceanography Data Management Office (BCO-DMO) at Woods Hole.
2. Microarray data: these will be archived according to MIAME standards (Brazma et al. 2001; 2006) using the GEO (Gene Expression Omnibus) database at the US National Center for Biotechnology Information (NCBI; Edgar and Barrett, 2006).
3. Gene sequences from the environmental EST libraries: Currently we are formally collaborating with CAMERA, through a Moore Foundation funded project, in an effort to deploy two internal JCVI web resources, MetaDB and PhyloDB, on a public portal within the CAMERA framework. Through the research proposed here we will continue development of MetaDB, which coupled to PhyloDB, will provide an easily navigable interface (available through the CAMERA website) where users can search various metatranscriptomics databases and evaluate the expression level and phylogenetic distribution of transcripts that are orthologous to queries.. All of the data collected in the proposed project from SS and SNA along with newly developing metatranscriptomic datasets from other sites acquired by the PIs, will be made available through MetaDB. JCVI backs up all critical data nightly through incremental backups and weekly through complete backups, and tapes are stored off-site. IT support includes systems, network, web, and database services, and is provided by a staff of engineers.
4. Additional single gene sequences obtained from investigation of the diversity of functional genes of interest from the experimental sites and cultures will be archived according to standard procedures in GenBank (NCBI).
5. All gene expression data (EST and microarray) will be organized into a searchable website database that will be managed by JCVI and publicly available. Data will be indexed with experimental details related to each experiment and organized by expression level and other gene ID annotation information (e.g., GO, KEGG, KOG). The data will be searchable by a variety of BLAST and keyword search options. In addition to and prior to deposition in public databases, all EST and microarray gene expression data will be archived and stored within the JCVI/TIGR IT infrastructure.

Information such as location and initial conditions at the field sites will be included as metadata for the resulting molecular data. Experimental results from experiments (isotope tracer experiments, isotopic measurements on sorted samples) will be permanently archived on the PIs' computers, the GEO departmental server, and Princeton University automatic backup systems. The permanent repository for these data will be the resulting publications. In addition, interactive learning modules, phytoplankton taxon lists and outreach websites will be publically available and maintained on the GEO departmental servers.