Data Management Plan

Products of the Research

The primary data products for this project will include **DNA and *cDNA pyrotag sequences*, quantitative PCR microbial abundance measurements, DNA sequences from Single-cell Amplified Genomes (SAGs) and estimates of microbial abundance by flow-cytometry**. Pyrotag 18S ribosomal cDNA will represent the metabolically active fraction of the planktonic microbes in dilution-based grazing studies conducted in Booth Bay (Maine) and Scripps Pier (California). It's expected that we will generate approximately several hundred thousand 'next gen' sequencing reads for characterizing the eukaryotic microbial assemblages in our grazing experiments. A smaller number of sequences (<5000) will be generated to aid in the development and improvement of both *Synechococcus* and *Synechococcus*-grazer qPCR primers. Deep sequencing of rRNA gene sequences will permit diversity estimation of the < 80 µm sized eukaryotic grazers and phytoplankton before, during, and after seasonal *Synechococcus* blooms. All DNA sequences will be curated and quality checked using established bioinformatic pipelines to remove low-quality sequences and potential PCR chimeras using established sequence-processing pipelines.

A suite of environmental parameters will be measured to aid in the interpretation of the primary data products described above. Basic ***environmental metadata*** will be collected (temperature, salinity, chlorophyll, major nutrient concentrations) in coordination with ongoing sampling at our study sites. In addition to the previous data types, ***microbial abundance*** data will be generated from our project (e.g., counts of bacteria, picoeukaryotes, heterotrophic and phototrophic protistan nanoplankton) and will serve as an important component for ground-truthing molecular data where possible. Abundance data will take the form of flow-cytograms and/or direct cell counts by microscopy. ***Digital photomicrographs*** resulting from this work will serve as a permanent visual record of the types of microorganisms in our samples.

Lab notebooks containing the raw data generated by this project, detailed descriptions of procedures and methodological approaches, deviations from protocols, specific equipment, and chemical reagents utilized for this project will be cataloged by all project participants. Periodic digital copies of laboratory notebooks and/or related digital media (e.g., spreadsheets and data summaries), will be exchanged between the labs to ensure that collaboratory research goals are met and to promote the exchange of ideas and best research practices. Additionally this exchange will promote data security and will prevent physical loss via unforeseen events. Notebooks and digital copies thereof will serve as permanent records of the project and will be available upon request from NSF program managers.

Plans for Preservation of Research Products

All DNA sequences generated by this project will be submitted to publically accessible databases such as GenBank: http://www.ncbi.nlm.nih.gov/genbank/ and CAMERA: http://camera.calit2.net/. GenBank frequently exchanges data with similar databanks in Europe (EMBL) and Japan (DDBJ) to ensure that data are available as widely as possible, thus insuring against data loss. DNA sequences will be available on public databases as soon as manuscripts are submitted. Countway has submitted, or has participated in projects that have submitted, more than 7,100 rRNA gene sequences to GenBank over the past 10 years. Environmental metadata will be included for all sequence submissions. Process-based data (e.g., rates of growth, grazing, and viral lysis) and microbial abundance data will be submitted to the National Oceanographic Data Center:

http://www.nodc.noaa.gov/ to comply with the '2004 NSF Division of Ocean Sciences Data and Sample Policy' for biological data http://www.nsf.gov/pubs/2004/nsf04004/nsf04004.pdf. We will work with staff at the Biological and Chemical Oceanography Data Management Office (BCO-DMO) to insure that all of our data are archived at the NOAA NODC facility and that DNA sequence and gene expression data are linked back from international databases to NODC data portals. Additionally, our data sets will be available online from the BCO-DMO data system (http://bco-dmo.org/data/) where they will be managed. We will work with BCO-DMO personnel to ensure that all data components are linked among the different databases including those hosting: **1)** DNA sequences and data generated from sequence analysis (e.g., alignments, phylogenetic trees, rarefaction curves, diversity estimates, and network analyses), **2)** process and rate measurements, **3)** environmental metadata, **4)** and microbial abundance and imagery data. Images will be submitted to Microbe Library http://www.microbelibrary.org/, a site created and hosted by ASM to generate a database of microbial images for use in undergraduate education and will also be submitted to Micro*Scope http://starcentral.mbl.edu/microscope/. All data will be archived as soon as it becomes available on storage devices located in each of the PIs labs and on their institutional servers, which are backed up both locally and remotely to ensure data security and prevent loss. DNA/RNA Samples in extracted form will be archived at -80 °C and divided among two freezers for safe-keeping as much as feasible. We are presently working to implement a lab-wide digital tracking program for logging samples collected during research programs.

Documentation and Sharing of Data and Samples

As indicated in the 'Products of the Research' section above, all aspects of our research will be documented in both physical (Lab Notebooks) and digital formats. The project will be highlighted on the websites of the Principle Investigators at Bigelow Laboratory for Ocean Sciences and Scripps Institute of Oceanography, where we will also provide links to digital data repositories following the NSF guidelines for the timeline of both data and sample release specified in the '2004 NSF Division of Ocean Sciences Data and Sample Policy'. Logs will be kept by PIs and project participants for all samples collected and analyses performed and a sample summary will be available online – this will ensure maximum transparency of both progress and data/sample availability. The PIs will work closely on NSF Annual Reports to provide updates to NSF program managers, and to the general public. The PIs will collaborate on data management during field deployments and back at their respective institutions. Archived and frozen (-80 °C) sub-samples of extracted nucleic acids and unextracted backup samples will be available to interested parties for their independent analysis following publication of results.

Curriculum Materials

A large amount of teaching and public lecture materials will result from the proposed research, primarily in the form of digital presentations and large-format research posters. These materials will be made available on the websites of each investigator for educational use following the guidelines from NSF on data release and fair use. The Bigelow PIs expect to incorporate results from the proposed research into their course lecture materials as part of the new affiliation between Bigelow Laboratory and Colby College. Additionally, PI Palenik will incorporate results from the project into his undergraduate and graduate teaching materials at UCSD and Scripps.