DATA MANAGEMENT PLAN

Primary investigator: Michael Rappé
Institution: University of Hawaii
Project: Population genomics and ecotypic divergence in SAR11 marine bacteria
NSF Division: OCE Solicitation Info: NSF OCE BIO PD 98-1650 Submission Date: 2/15/2015

The following Data Management Plan has been developed based on the recommendations of the NSF OCE Data and Sample policy (NSF 11-060) and the BCO-DMO Best Practices Guide.

DESCRIPTION OF DATA TYPES

Observational

1. **YSI 660 sonde data and associated event logs from small vessel operations:** data will include date, start and end time, latitude and longitude, air temperature, water temperature, tide, salinity, dissolved oxygen, total depth of water column, depth of sampling, pH, turbidity, fluorescence.
2. **Water samples and characteristics:** Seawater samples will be taken for (i) culturing experiments, and (ii) field observations. In addition to event log and YSI data, observational data will include: total volume sampled, total volume filtered for environmental DNA, volume and number of aliquots for cryopreservation, sample chemistry (macronutrients and DOC), chlorophyll a concentration, flow cytometrically-derived cellular concentrations of phytoplankton, *Synechococcus*, *Prochlorococcus*, and non-pigmented microbes, bacterial production, and primary production.

Experimental

1. **Environmental DNA sequencing:** Environmental DNA extraction and sequencing from filtered seawater samples will generate environmental DNA and Illumina sequence data from amplified SSU rRNA genes, amplified SAR11 petB genes, and unamplified metagenomes.
2. **Cultivation experiments:** product types will include isolated strains of marine microorganisms growing in the laboratory, cryopreserved stocks of isolates, and DNA extracted from individual strains.
3. **Isolate DNA sequencing:** DNA extraction and sequencing from isolated strains will generate isolate genomic DNA, Sanger-sequenced SSU rRNA genes and petB gene sequences, and Illumina-sequenced whole genomes.

Derived

1. **Whole genome sequences:** genome sequence data from individual isolates will be assembled into contigs and scaffolds and subsequently annotated to locate genes and putatively identify gene products via the Department of Energy Joint Genome Institute's Integrated Microbial Genomes on-line portal.

DATA AND METADATA FORMATS AND STANDARDS

Observational data will be stored in flat ASCII files, which can be read easily by different software packages. Illumina sequence data files will be stored as unprocessed qseq and fastaq data files. Metadata will be prepared in accordance with BCO-DMO conventions (i.e. using the BCO-DMO metadata forms) and will include detailed descriptions of collection and analysis procedures.

DATA STORAGE AND ACCESS DURING THE PROJECT

The investigators will store project data (including spreadsheets, ASCII files, images, and PDFs of scanned logs) on laboratory computers that are backed up daily via Apple Time Machine to an external hard drive. The computers and external hard drives are also backed up daily to an on-site server with RAID data mirroring maintained by the project PI. Laboratory notebooks and log spreadsheets will serve as hard copy backup. Raw Illumina sequence data files are immediately backed up to an external hard drive, on-site server with RAID data mirroring maintained by the PI, and off-site through the Hawaii Institute of Marine Biology's Evolutionary Core Facility. The products of processing and analyzing the Illumina DNA sequence data are backed up to an external hard drive and on-site server with RAID data mirroring maintained by the PI. Generating a workgroup on the in-house Rappé lab server will allow all project personnel to share and access files regardless of physical location.

MECHANISMS AND POLICIES FOR ACCESS, SHARING, RE-USE AND RE-DISTRIBUTION

DNA sequences will be deposited in the appropriate National Center for Biotechnology Information (NCBI) database (e.g. GenBank, Sequence Read Archive) upon submission of manuscripts. GenBank accession numbers and Sequence Read Archive project numbers will be provided to BCO-DMO in an Excel spreadsheet or .CSV file and metadata will be provided using the BCO-DMO Dataset Metadata submission form. Observational and non-DNA sequence-based data sets produced through this project will be made available through the BCO-DMO data system within two years from the date of collection. The project investigators will work with BCO-DMO data managers to make project data available online in compliance with the NSF OCE Sample and Data Policy. In addition to access through the NCBI, genome sequences and their annotations will be made publicly available through the Department of Energy Joint Genome Institute's Integrated Microbial Genomes on-line portal. Microbial strains and their products (e.g. extracted genomic DNA) will be made publically available directly from the laboratory PI upon submission of manuscripts or within two years, whichever arrives first. No public or private commercial strain repositories are yet able to cultivate and distribute SAR11 isolates. Preserved seawater and its products (e.g. environmental DNA) will be maintained as archive material for sharing with external users and made available upon request from the PI.

PLANS FOR ARCHIVING

BCO-DMO will ensure that project data are submitted to the appropriate national data archive. The PI will work with BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data. DNA sequence data will be archived through the NCBI, and linked with metadata through BCO-DMO.