DATA MANAGEMENT PLAN

Primary investigator: Michael Rappé
Project: Illuminating the physiology, cellular characteristics, and ecogenomics of the first cultivated strain of the SAR86 lineage
Solicitation Info: NSF OCE BIO PD 98-1650 Submission Date: 16 August 2021

The following Data Management Plan has been developed based on the recommendations of the NSF OCE Data and Sample policy (NSF 17-037) and the BCO-DMO Best Practices Guide.

DESCRIPTION OF DATA TYPES

Observational

1. **Probe and sensor data, and associated event logs from vessel operations:** data will include date, start and end time, latitude and longitude of sampling stations, air temperature, water temperature, wind direction and velocity, 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) microbial ecology, and (ii) field observations. In addition to the event log and associated probe and sensor data, observational data will include: total volume sampled, total volume filtered for environmental nucleic acids, volume and number of aliquots for cryopreservation, sample chemistry (macronutrients and DOC), chlorophyll *a* concentration, HPLC-derived pigment concentrations, 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-based DNA sequence data from amplified 16S rRNA genes and unamplified metagenomes.
2. **Controlled experimentation with strains.** Product types will include cellular enumeration, images of cell morphology via electron microscopy and epifluorescence microscopy, rates of leucine incorporation into cells, oxygen concentrations, dissolved organic carbon concentrations, particulate carbon, nitrogen, and phosphorous, cellular RNA, and growth rates.
3. **Cultivation experiments to isolate strains.** Product types will include isolated strains of marine microorganisms growing in the laboratory, cryopreserved stocks of isolates, and DNA extracted from individual strains.

Derived

1. **Environmental DNA sequencing.** Product types will include taxon distribution tables (amplified 16S rRNA genes), metagenome-assembled genomes, and quantification of genome abundance via read recruitment from unassembled metagenomes.
2. **Controlled experimentation with strains.** Product types will include cell volume (and associated changes in cell volume over time), elemental composition of cells, and rates of metabolism under a variety of growth conditions measured via a variety of means (oxygen, dissolved organic carbon, leucine incorporation). Transcriptomes will be sequenced from extracted RNA.

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 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 allows all project personnel to share and access files regardless of physical location.

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

Nucleic acid sequences will be deposited in the appropriate National Center for Biotechnology Information (NCBI) database (e.g. GenBank, Sequence Read Archive) upon submission of manuscripts or within two years, whichever arrives first. 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 derived annotations will be made publicly available through the Department of Energy Joint Genome Institute's Integrated Microbial Genomes on-line portal.

Public access to stains. We will provide multiple avenues for scientists to access strain HIMB1674 and any others isolated over the course of this study, either upon submission of manuscripts or within two years of isolation, whichever arrives first. First, all microbial strains and their products (e.g. extracted nucleic acids) will be made publicly available directly from the PI. Second, we will work with the American Type Culture Collection (ATCC) to house and distribute viable cells of our strains directly to the public. At a minimum, this will include cryopreserved stocks of strains, but may also include growing cultures if this proves feasible by the ATCC. Lastly, if taxonomic descriptions are generated for any strains, we will follow the requirements mandated by the International Journal of Systematic and Evolutionary Microbiology and the International Committee on Systematics of Prokaryotes and deposit the proposed type strain in two recognized culture collections in two different countries.

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.