Plan Overview

A Data Management Plan created using DMPTool

Title: EAGER: High-throughput, culture-independent technique identifying cyanobacteria infections to improve understanding of carbon biogeochemical cycling

Creator: Sarah Preheim

Affiliation: Johns Hopkins University (jhu.edu)

Principal Investigator: Sarah Preheim

Data Manager: Sarah Preheim

Funder: National Science Foundation (nsf.gov)

Funding opportunity number: PD 98-1650

Grant: https://www.nsf.gov/awardsearch/showAward?AWD_ID=1820652

Template: BCO-DMO NSF OCE: Biological and Chemical Oceanography

Last modified: 12-07-2017

EAGER: High-throughput, culture-independent technique identifying cyanobacteria infections to improve understanding of carbon biogeochemical cycling

Data Policy Compliance

Identify any published data policies with which the project will comply, including the NSF OCE Data and Sample Policy as well as other policies that may be relevant if the project is part of a large coordinated research program (e.g. GEOTRACES).

The project investigators will comply with the data management and dissemination policies described in the *NSF Award and Administration Guide* (AAG, Chapter VI.D.4) and the *NSF Division of Ocean Sciences Sample and Data Policy*.

Pre-Cruise Planning

If the proposed project involves a research cruise, describe the cruise plans. (Skip this section if it is not relevant to your proposal.) Consider the following questions: (1) How will pre-cruise planning be coordinated? (e.g. email, teleconference, workshop) (2) What types of sampling instruments will be deployed on the cruise? (3) How will the cruise event log be recorded? (e.g. the Rolling Deck to Repository (R2R) event logger application, an Excel spreadsheet, or paper logs) (4) Will you prepare a cruise report?

Not applicable.

Description of Data Types

Provide a description of the types of data to be produced during the project. Identify the types of data, samples, physical collections, software, derived models, curriculum materials, and other materials to be produced in the course of the project. Include a description of the location of collection, collection methods and instruments, expected dates or duration of collection. If you will be using existing datasets, state this and include how you will obtain them.

The project will produce several observational and experimental datasets, described below. In addition to the datasets described below, computational and educational resources produced by the project, including scripts, data and images, will be made available for public use through the PI's Github account and on the Preheim Lab website (http://preheimlab.johnshopkins.edu/). Observational data will be collected from the Chesapeake Bay from March 2018-March 2019.

Observational Datasets:

1. *Environmental measurements:* Environmental data will be collected using a YSI EXO1 sonde; data will include standard environmental measurements such as pressure, temperature, salinity, oxygen, turbidity, ORP, pH and conductivity. Chlorophyll and FDOM measurements collected at the SERC sample site by the Smithsonian National Museum of Natural History SI Physical Monitoring Network (nmnhmp.riocean.com) will also be recorded. File types: Excel spreadsheet and ASCII files. Repository: BCO-DMO.

2. *Sampling event log:* Sampling event log; will include unique sample numbers, start/end dates, start/end times, GPS coordinates of locations, replicate number, weather information, tidal events, and notes about sampling. Sampling event logs will be recorded on paper log sheets or with a computer. File types: Excel file converted to .csv; scanned PDFs. Repository: BCO-DMO

Experimental Datasets:

1. *Microscope Images:* Images collected from fluorescence microscopy will be stored on the Maryland Advanced Research Computing Center (MARCC) as .jpeg or .tiff images and backed up to a server. An image inventory will be created associating the images with the samples. Repository: BCO-DMO

2. *Plaque assay results*: Cyanobacteria isolates, infecting phage and infection networks will be generated from plaque assays. Strains will be archived as described below. Infection results will be stored in an excel spreadsheet. Repository: BCO-DMO.

3. *Genetic sequencing:* DNA sequences from samples collected at the sampling site will include 16S rRNA gene and viral marker gene amplicon data, as well as shotgun metagenomic sequencing of various components of the microbial community, including the viral and bacterial communities. EpicPCR data of associated virus and host amplicons will also be generated. 16S rRNA gene and viral marker gene data will be sequenced from isolates and plaques during plaque assays. Sequencing will be performed at the JHU Genetic Resource Core Facility, Deep Sequencing and Microarray Core Facility or University of Maryland Institute for Genome Sciences. File types: Quality (.fastq) and sequence data (.fasta) files. Repository: NCBI Short-read archive (SRA); accession numbers to be provided to BCO-DMO.

Data and Metadata Formats and Standards

Identify the formats and standards to be used for data and metadata formatting and content. Where existing standards are absent or deemed inadequate, these formats and contents should be documented along with any proposed solutions or remedies. Consider the

following questions: (1) Which file formats will be used to store your data? (2) What type of contextual details (metadata) will you document and how? (3) Are there specific data or metadata standards that you will be adhering to? (4) Will you be using or creating a data dictionary, code list, or glossary? (5) What types of quality control will be used? How will data quality be assessed and flagged?

Field observation data will be stored in flat ASCII files, which can be read easily by different software packages. Field data will include date, time, latitude, longitude and depth, as appropriate. Quality flags will be assigned according to the ODS IODE Quality Flag scheme when appropriate (IOC Manuals and Guides, 54, volume 3; http://www.iode.org/mg54_3). 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.

Standard operating procedures (SOP) are in place to ensure sampling quality during preparation of the sampling equipment, instrument calibration, sample collection and storage. SOPs are in place for laboratory work for DNA extraction, diversity surveys and sample sequencing. Positive and negative controls are required for each project, including field blanks, DNA extraction negatives, positive and negative polymerase chain reaction (PCR) controls and replicates. Samples are processed randomly to minimize the impact of batch effects. SOPs are in place for sequence analysis. All sequence data undergoes stringent quality filtering prior to analysis. Standard pipelines will be comprised of all of the processing steps in detail, including the script used to process the data.

Data Storage and Access During the Project

Describe how project data will be stored, accessed, and shared among project participants during the course of the project. Consider the following: (1) How will data be shared among project participants during the data collection and analysis phases? (e.g. web page, shared network drive) (2) How/where will data be stored and backed-up? (3) If data volumes will be significant, what is the estimated total file size?

The investigators will store project data (including sequences data, spreadsheets, ASCII files, images, and PDFs of scanned logs) on personal computers that are backed up by the Johns Hopkins University's central IT organization through CrashPlan to an WSE server and an off-site Cloud server. Additionally, computers will also have personal backups, such as Apple Time Machine to an onsite external hard drive. The Principal Investigator (PI) has also established an account with Maryland Advanced Research Computing Center of for data storage, computing and sharing among project investigators.

Mechanisms and Policies for Access, Sharing, Re-Use, and Re-Distribution

---

Describe mechanisms for data access and sharing, and describe any related policies and provisions for re-use, re-distribution, and the production of derivatives. Include provisions for appropriate protections of privacy, confidentiality, security, intellectual property, or other rights or requirements. Consider the following: (1) When will data be made publicly available and how? Identify the data repositories you plan to use to make data available. (2) Are the data sensitive in nature (e.g. endangered species concerns, potential patentability)? If so, is public access inappropriate and how will access be provided? (e.g. formal consent agreements, restricted access) (3) Will any permission restrictions (such as an embargo period) need to be placed on the data? If so, what are the reasons and what is the duration of the embargo? (4) Who holds intellectual property rights to the data and how might this affect data access? (5) Who is likely to be interested in re-using the data? What are the foreseeable re-uses of the data?

DNA sequences will be deposited in the National Center for Biotechnology Information (NCBI) database GenBank upon submission of manuscripts, but no more than 2 years after sample collection. GenBank accession numbers will be provided to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) in an Excel spreadsheet or .CSV file and metadata will be provided using the BCO-DMO Dataset Metadata submission form. Data sets produced will be made available through the BCO-DMO data system within two-years from the date of collection. Scripts used to process epicPCR data (bash and/or python scripts) will be made publically available through the PI's Github account and the Preheim Lab website. 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. Data, samples, and other information collected under this project can be made publically available without restriction once submitted to the public repositories. Data produced by this project may be of interest to biological oceanographers interested in the influence of viruses on microbial communities. We will adhere to and promote the standards, policies, and provisions for data and metadata submission, access, re-use, distribution, and ownership as prescribed by the BCO-DMO Terms of Use (http://www.bco-dmo.org/terms-use).

Plans for Archiving

Describe the plans for long-term archiving of data, samples, and other research products, and for preservation of access to them. Consider the following: (1) What is your long-term strategy for maintaining, curating, and archiving the data? (2) What archive(s) have you identified as a place to deposit data and other research products?

BCO-DMO will also 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. Characterized cyanobacteria and phage will be archived in glycerol stocks for 5 years after publication of our results at -80 °C. We will work with with various culture collections (Provasoli-Guillard National Center for Marine Algae and Microbiota and Culture Collection of Algae at the University of Texas) to archive strains and associated phage for distribution to the public. DNA, gylcerols and unfiltered water will be preserved for 5 years after publication of any manuscript.

Roles and Responsibilities

Describe the roles and responsibilities of all parties with respect to the management of the data. Consider the following: (1) If there are multiple investigators involved, what are the data management responsibilities of each person? (2) Who will be the lead or primary person responsible for ultimately ensuring compliance with the Data Management Plan?

Dr. Preheim will be responsible for sharing data among the project participants in a timely fashion. Dr. Sakowski will be responsible for sample collection, molecular biology, sequencing, culturing and will submit the resulting sequences to the National Center for Biotechnology Information's (NCBI) GenBank database. Dr. Preheim will coordinate the overall data management and data sharing and will submit the project data, including GenBank accession numbers, and metadata to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) who will be responsible for forwarding these data and metadata to the appropriate national archive.
