Plan Overview

A Data Management Plan created using DMP Tool

Title: Investigating the environmental dynamics and physiology of recently identified & metabolically streamlined marine picocyanobacteria from an iron-limited ocean region

Creator: Garrett Sharpe

Affiliation: North Carolina State University (ncsu.edu)

Principal Investigator: Garrett Sharpe

Data Manager: Garrett Sharpe

Project Administrator: Garrett Sharpe

Funder: National Science Foundation (nsf.gov)

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

Start date: 07-01-2024

End date: 06-30-2026

Last modified: 07-08-2024

Copyright information:

Investigating the environmental dynamics and physiology of recently identified & metabolically streamlined marine picocyanobacteria from an iron-limited ocean region

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 investigator and collaborators will adhere to the data management policies outlined in the NSF Award and Administration Guide and the NSF Division of Ocean Sciences (OCE) 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?**

Pre-cruise preparation and planning will be done via telecommunication and email with Dr. Steven Hallam and the Line P coordinater Marie Roberts. A cruise research implementation plan will be created detailing station locations, water collection depths, water sampling strategy, and water sample allocation. All sampling events will be recorded via pre-prepared paper sampling forms and scanned into PDF documents.

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 a variety of observational and experimental datasets. Observational data will be collected on several research cruises along Line P in the Northeastern Pacific (May 2024 pre-funding, September 2024 and January 2025 with proposal funding).

Observational datasets:

1. **CTD data:** CTD data collected using a SeaBird CTD package; processing to be done using IOSShell software; data will include standard environmental measurements (such as pressure, temperature, salinity, fluorescence). File types: ASCI (.ctd and .che) and .csv. Repository: BCO-DMO and Water Properties Group website.
2. **Event log:** Cruise scientific sampling event log; log data includes event type, event numbers, start/end dates, times & locations of instrument deployments, and samples collected. Will be recorded on paper log sheets. File types: Excel file converted to .csv; scanned PDFs. Repository: BCO-DMO and Water Properties Group website.
3. ***Synechococcus* isolation log:** Incubations of 1.2 micron-filtered seawater in SN, PRO2, and PRO99 media from three depths at each major station along Line P. CTD cast and Niskin bottle numbers, date and time of incubation start, depth of inoculum collection, sample dilutions, and *Synechococcus* counts via flow cytometry at the end of the cruise will be recorded on log sheets by hand and transferred to an Excel form. File types: PDF of scanned sampling logs; Excel file of log data.
4. **Microbial sampling logs:** Marine microbial samples for DNA extraction will be collected by peristaltic pumping of CTD-collected seawater successively through a 3 micron filter (large size fraction) and 0.2 micron filter (small size fraction). CTD cast and Niskin bottle numbers, location of sampling, date and time of sampling, filter numbers, size

fraction, depth of sampling, volume filtered, and filtration time will be recorded on log sheets by hand and transferred to an Excel form. File types: PDF of scanned sampling logs; Excel file of log data. Repository: BCO-DMO.

Experimental datasets:

1. **Isolate strains and genomes:** *Synechococcus* strains isolated from Line P will be deposited into the Roscoff Culture Collection. Genomic sequencing generated from the Line P *Synechococcus* cultures will be used to assemble their genomes, and these genomes will be submitted to NCBI. File type: .fasta files. Repository: NCBI for genomes; Roscoff Culture Collection for *Synechococcus* strains; accession number of genomes provided to BCO-DMO.
2. **Line P metagenomes and metagenome-assembled genomes (MAGs):** DNA sequences from 0.22-3.0 micron size fraction of marine microbial community. Sample preparation and DNA sequencing will be performed at the Paerl lab at North Carolina State University. Metagenome-assembled genomes will be assembled from this DNA sequencing data. File types: Short read archive (.sra) and .fasta files. Repository: NCBI; accession numbers of raw metagenome data and MAG data provided to BCO-DMO.
3. **Variable iron incubation experiments:** measurements for the *Synechococcus* variable iron incubation experiments including fluorometric values, cell counts, cell size, calculated growth rates, etc. will be recorded in an excel table. Proteomic mass spectrometry data collected from late exponential stage during the experiments will be submitted to the Proteomics Identification Database (PRIDE) Archive. File types: Excel File of experimental conditions, .raw file of raw spectra data from proteomes. Repository: PRIDE Archive for proteomic data; excel files and PRIDE project number 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 and include date, time, latitude, longitude, cast number, and depth. Quality flags will be assigned according to the ODS IODE Quality Flag scheme (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 sampling equipment, collection, and storage. SOPs are in place for laboratory work DNA and protein extraction, experimental incubations, and sample sequencing and processing. Samples are processed randomly to minimize the impact of batch effects. SOPSs are in place for DNA sequence and mass spectra analysis. All sequence and mass spectra data undergoes stringent quality filtering prior to analysis, and standard pipelines will be comprised of all of the processing steps in data, including scrips used for data processing.

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 investigator will store project data (including sequence data, spreadsheets, ACII files, and PDFs of scanned logs) on laboratory computers that are backed up by the University's central IT organization. The Principal Investigator (PI) has also

establed an account with North Carolina State University's High-Performing Computing Services for data storage and sharing among project investigators. Personal computers will be backed up on an onsite external hard drive as well as via Dropbox Cloud Services, which will be additionally used to share data with collaborators.

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?**

Immediately after completion of the research cruise, metadata will be submitted to the Water Properties Group website. DNA sequences (genomes, metagenome-assembled genomes, and metagenome samples) will be deposited in the National Center for Biotechnology Information (NCBI) Genbank database upon submission of manuscripts. Genbank accession numbers and related metadata 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 online Submission Tool. Data sets produced by the science party 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. Data, samples, and other information collected under this project can be made publicly available without restriction once submitted to the public repositories.

Data produced by this project may be of interest to biological and chemical oceanographers and climate scientists interested in the role of cyanobacteria in global biogeochemistry and its relation to the global climate system. 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 ensure that project data are submitted to the appropriate national data archive. The PI will work with the Water Properties Group and BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data. Characterized *Synechococcus* strains will be archived in glycerol stocks for 5 years after publication of our results at -80°C. We will work with the Roscoff Culture Collection to archive these strains for distribution to the public. DNA, unprocessed environmental bacterial samples, and glycerol preserved 1.2 micron filtered 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. Sharpe will be responsible for sharing data among the project participants in a timely fashion. Dr. Sharpe will also be responsible for sample collection, molecular biology extraction, library prep, and sequencing, culturing, and experimental design and implementation, and will submit the resulting sequences to o the National Center for Biotechnology Information's (NCBI) GenBank database and mass spectra to the Protein Identification Database (PRIDE) archive. Dr. Sharpe will coordinate the overall data management and data sharing and will submit the project data, including GenBank accession numbers and PRIDE project number, 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.

Planned Research Outputs

Dataset - "Isolate genomes, metagenomes, and metagenome-assembled genomes (MAGs)"

All DNA sequencing data (isolate genomes, environmental metagenomes, and metagenome-assembled genomes) will be made publicly available on NCBI for other researchers to use to answer questions about seasonal behavior of the 0.2-3.0 micron size fraction of the microbial community. Isolate genomes and metagenome-assembled genome will also be available and provide a number of new high-latitude bacterial genomes for study and comparison with strains from other regions. Metadata will include environmental sampling context for all sequencing data.

Dataset - "Proteomic data from variable iron incubation experiments"

Proteomic data from low and high iron incubation experiments of *Synechococcus* strains will be publicly available on the Protein Identification Database Archive (PRIDE) for comparison with proteomes from other cyanobacterial strains. Metadata will include experimental variables measured during experiment.

Workflow - "Workflow scripts for DNA and protein sequence data processing and analysis"

Bioinformatic scripts for processing and analyzing metagenomic, genomic, and proteomic data will be generated and shared publicly for modification and/or use by others on GitHub.

Physical object - "Northeast Pacific Line P Synechococcus isolates"

Cultured axenic and clonal strains of *Synechococcus* collected and isolated from the Line P cruise will be submitted to the Roscoff Culture Collection to be made available to researchers interested in studying high-latitude *Synechococcus* strains from a High Nutrient Low Chlorophyll region. Metadata conforming to the Roscoff Culture Collection Standards will be used, where taxonomic identity, growth conditions (media type, light, temp, etc.), location of sampling and isolation, etc. will be recorded and submitted along with the physical strains.

Planned research output details

Title	Type	Anticipated release date	Initial access level	Intended repository(ies)	Anticipated file size	License	Metadata standard(s)	May contain sensitive data?	May contain PII?
Isolate genomes, metagenomes, and metagenome-assem ...	Dataset	2026-05-31	Open	NCBI	270 GB	Creative Commons Attribution Share Alike 4.0 International	Minimum Information about any (x) Sequence (MIxS)	No	No
Proteomic data from variable iron incubation exper ...	Dataset	2026-05-31	Open	EMBL-EBI	150 GB	Creative Commons Attribution Share Alike 4.0 International	MIBBI (Minimum Information for Biological and Biomedical Investigations)	No	No
Workflow scripts for DNA and protein sequence data ...	Workflow	2026-05-31	Open	GitHub	20 MB	Creative Commons Attribution Share Alike 4.0 International	None specified	No	No
Northeast Pacific Line P Synechococcus isolates	Physical object	2026-05-31	Open	Roscoff Culture Collection		Creative Commons Attribution Share Alike 4.0 International	None specified	No	No