Data Management Plan

Data Policy Compliance - Project investigators will comply with NSF OCE Data and Sample Policies and work with BCO-DMO staff to contribute data and results to the BCO-DMO system.

Pre-Cruise Planning

Pre-cruise planning for the R/V *Blue Heron* will be coordinated via teleconference and messaging through the team collaboration tool Slack. A science implementation plan will be written up with sample locations and a detailed plan for sampling and sample processing. Updated sampling protocols will be available on protocols.io. Sampling events will be recorded on paper logs that will be scanned into PDF documents and in a digital event log. Soon after the completion of each cruise, the original underway data will be contributed by the vessel operator to the UNOLS central data repository http://www.rvdata.us/catalog/ managed by the Rolling Deck to Repository (R2R) project.

Planned cruises on the EPA's R/V *Lake Guardian* are a continuation of Coleman Lab sample collection initiated in 2012. Established sampling protocols will be available through protocols.io. Sampling events will be recorded on paper logs that will be scanned into PDF documents. Underway and CTD data will be maintained by the EPA's Great Lakes National Program Office; local copies will be maintained on UChicago's Box file sharing service.

Description of Data Types

Observational

- Data collected by the CTD/Rosette during cruises on the Blue Heron will be deposited in BCO-DMO and archived at the NOAA National Centers for Environmental Information (NCEI; formerly the National Oceanographic Data Center).
- Field experiments will generate biological and chemical samples as well as data for sample processing. A database will be maintained to track the location and disposition of each sample (e.g., DNA extracted, submitted for amplicon sequencing). Surplus samples from this project will be made available to the wider scientific community upon request, provided that it does not interfere with completing proposed aims.

Experimental

- Molecular sequence data will be deposited in NCBI's Sequence Read Archive (ncbi.nlm.nih.gov/sra)
- Flow cytometry FCS files will be deposited in FlowRepository (https://flowrepository.org/)
- Enrichment cultures of cyanobacteria will be maintained in culture in the Coleman Lab and available upon request.
- Genome sequences from cyanobacterial enrichment cultures will be deposited as Whole Genome Shotgun submissions (ncbi.nlm.nih.gov/genbank/wgs.submit/).
- Proteomic mass spectrometry data will be made publicly available via the Tranche data repository hosted by the University of Michigan (proteomecommons.org/tranche).

Derived

- CTD/rosette and water chemistry data from the Lake Guardian will be accessed through the EPA's GLENDA database (https://cdx.epa.gov/), maintained in a relational database in the Coleman Lab, and included as metadata for archival of associated data products.
- Partial genomes from single cell amplified genomes sequenced by the Joint Genome

Institute will be downloaded from IMG/M (https://img.jgi.doe.gov/cgi-bin/m/main.cgi) and compiled, along with enrichment culture sequences (this project) and other available cyanobacterial sequences, into a database made available through a project website.

- Software/analysis tools and curated data will be made available to the scientific community through a public GitHub page.
- Educational and outreach materials (e.g. lesson plans, etc.) will be made publicly available through a project website and through broader educational material repositories (e.g., www.learnnc.org and opened.creativecommons.org).

Data and Metadata Formats and Standards

All analyses will employ certified reference materials (CRMs) where available (NRC Institute for National Measurement Standards, www.nrc-cnrc.gc.ca/eng/services/inms/reference-materials.html) and, where CRMs are not available, other best practices will be employed such as internal standards or appropriate controls. Data will be formatted according to requested formats for archival database repositories. Specifically, sequence data will conform to NCBI's guidelines (www.ncbi.nlm.nih.gov/guide/howto/submit-sequence-data) and proteomics data submissions will conform to the Minimum Information About a Proteomics Experiment – Mass Spectrometry (MIAPE-MS) guidelines maintained by the Proteomics Standards Initiative of the Human Proteome Organization (www.psidev.info/miape/MIAPE_MS_2.24.pdf). Environmental data including metadata from research cruises will be submitted to BCO-DMO as tab-separated ASCII files (bco-dmo.org).

Data Storage and Access During the Project

Data will be shared among project participants during data collection and analysis phases using a shared folder on UChicago Box, a cloud-based storage service that provides unlimited free online space for storing and sharing files. A private Bitbucket repository will be used for sharing data, code, and results of specific analyses. Data stored in UChicago Box will be backed up on an external hard drive in the Coleman Lab. Large data files (e.g., DNA and RNA sequencing data) will be stored and analyzed in Coleman's Project Directory on the University of Chicago's Research Computer Center's Midway compute cluster. The RCC maintains snapshots for quick and easy data recovery as well as archival tape backups performed on a nightly basis and located in a different data center than the main storage system to safeguard against hardware failure.

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

Underway and CTD data will be made available immediately following completion of Blue Heron cruises. Additional data products will be made available following publication.

Plans for Archiving

R2R will ensure that the original underway measurements are archived permanently at the NOAA NCEI. BCO-DMO will also ensure that project data are submitted to the appropriate national data archive. PIs will work with R2R and BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data.