

Data Management Plan

We have read and agree to abide by NSF-wide and BioOc-specific guidelines and award conditions for scientific conduct and data management. Results, data, and collections will be made available to qualified researchers upon request, provided that the quantities and time requirements for compliance do not compromise our research objectives. We will also address data sharing issues in annual reports.

Project overview: This proposal seeks funding to infer virus-host interactions in a community context, thereby quantifying the role of functional pairs in disproportionately influencing cellular lysis and nutrient regeneration. Theoretical models and informatics analysis will be combined with *in vitro* experiments and *in situ* studies conducted at the Bigelow Laboratory for Ocean Sciences – a coastal oceanographic research station based in East Boothbay, Maine – and at the BATS site – an open-ocean oceanographic observatory off-shore of Bermuda in the North Atlantic Ocean.

Data policy compliance: 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: Pre-cruise planning will be done via teleconferencing and a planning workshop. Detailed plans for station locations, instrument deployment, water sampling strategy, and water sample allocation will be written up as a science implementation plan for the cruise. The actual sampling events will be recorded on paper logs (scanned into PDF documents) and/or in a digital event log using the Rolling Deck to Repository (R2R) event logger application.

Description of data types: The project will produce several observational and experimental datasets, described in the list below. In addition to the datasets described below, educational resources produced by the project, including data, images and video will be made available for public use on the UTK website and (as appropriate) open-access locations on the internet (e.g., these can range from YouTube for video to *protocols.io* for techniques and associated control data, *etc.*). As required by NSF, all oceanographic data will be deposited in the Biological and Chemical Oceanographic Data Management Office (BCO-DMO) by the end of this project (even if that occurs before publication). In addition, all sequence data, while archived elsewhere (see below), will be described and a link to the data provided through BCO-DMO. It is also our intent to direct share data with the BIOS-SCOPE research group as soon as it is available (see letter from Program Director Dr. Craig Carlson).

Observational Datasets (anticipated) w/repositories of BCO-DMO & R2R :

- 1. CTD and Niskin bottle data:** CTD data collected using a SeaBird SBE CTD package; processing to be done using SeaBird's SeaSave software; data will include standard environmental measurements (such as pressure, temperature, salinity, fluorescence). File types: Raw (.con, .hdr, .hex, .bl) and processed and .cnv, .asc, .bt1) ASCII files.
- 2. Event log:** Cruise scientific sampling event log; will include event numbers, start/end dates, times & locations of instrument deployments. Will be recorded using the R2R event logger (if available) and on paper log sheets. File types: Excel file converted to .csv; scanned PDFs. Our intent is to create a digital event log for the entire cruise (without completely abandoning the historical approaches).
- 3. Cruise underway data:** Routine underway data collected along the ship's track (including meteorological data, sea surface temperature, salinity, fluorescence, ADCP). Will be collected by the shipboard instrumentation. File types: .csv ASCII files.
- 4. Sampling logs and images:** We will collect CTD data every 4 hours once we have reached stations (BATS) and maintain in a Lagrangian manner. Samples for environmental data (nutrients, temperature, virus and bacterial abundances, chlorophyll, *etc.*) will be collected and logged into a network shared database. CTD numbers, locations, depths, dates, and times will be recorded by hand onto log sheets as a redundant backup to electronic versions. Information from logs will be transferred into an Excel spreadsheet. File types: Excel files of sampling logs; images (.jpg files).

Experimental Datasets:

1. **Genetic sequencing:** Sequencing data will be deposited in the Short Read Archive (SRA) at the National Center for Biotechnology Information in its raw format for open access once it has been validated and checked. Subsequently assemblies generated by this effort will be added back to the NCBI or released on other online platforms (MG-RAST, iVirus, *etc.*). We intend to make all data that is annotated and curated through this project publicly available through ‘iVirus’. PI Sullivan and PI Weitz have already contributed methods and data at iVirus through other projects. Locally, we now routinely work with data at this scale and have automated back-up procedures at multiple levels (private compute cluster, university high-performance compute cluster).
2. **Cross-infection data:** Data from the efficiency of plating experiments with mock communities will be recorded in spreadsheets that will be made available as a project through iVirus, as well as a supplemental document in any resulting publication.

Simulation datasets - InVirT: Source code for InVirT software and simulations will be released as open-source code using Creative Commons licenses. All conditions for fully reproducible simulations will be included as part of permanently archived software and simulation releases with publications, using a permanent doi via the zenodo.org archive (see doi:10.5281/zenodo.61196 for an example of prior Weitz lab release of code and simulation datasets for analyzing a model of phytoplankton mortality).

Data storage, access and archives: Metagenome and metatranscriptome raw data will be stored and processed locally by Sullivan (details in facilities description), with raw and processed data archived via the iVirus portal. Curated viral genomes and their associated annotation (e.g. predicted host, detection in microbial metagenomes, *etc.*) will be made available as a separate database at iVirus, in both fasta and genbank file formats so that a user can easily use these as new references. Other curated data products (e.g. co-occurrence networks, association networks, time-series, *etc.*) will also be made available through iVirus. All records will be cross-referenced to sample data, and clear data description files will be made available through iVirus with the data files. The original data will be posted to BCO-DMO whenever possible, i.e., subject to data size and type restrictions. Submission of “big data” will be done in a summary format, with links at BCO-DMO to the long-term deposition site on iVirus.

Theoretical Models: Theoretical models developed as part of this grant will be published with complete information necessary to recapitulate the theory. Theoretical models are collections of mathematical analyses and arguments that are best described in a standard publication.

Software: We will release all software supported in this project via open-source licenses. Weitz group regularly maintains a github site – <http://weitzgroup.github.io>. The Sullivan group maintains a lab code repository – <https://bitbucket.org/MAVERICLab/>. We will also distribute software and scripts in the CyVerse Discovery Environment, via the iVirus platform. We will announce code updates to the VerveNet discussion forums, which includes opportunity to innovate protocols as a community through **protocols.io** to maximize scientific openness and transparency. Both experimental and informatics protocols are contributed there, e.g., Sullivan’s iVirus apps and workflows are documented through protocols.io.

Long-term Archiving and Distribution of Data: All means of dissemination of data will be maintained, at least, for the lifetime of the project. Deposition to BCO-DMO, NCBI’s SRA, GitHub and iVirus ensures archival dissemination over longer (5-10 year) time frames.

Roles and Responsibilities: Each PI will be responsible for sharing his/her subset of data among the project participants in a timely fashion, i.e., simulation models and cross-infection inference (Weitz), cruise data (Wilhelm), and molecular biology and sequence data (Sullivan). PI Weitz, will coordinate the overall data management and sharing process and will submit the project data in coordination w/Co-PIs, 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.