

Data Management Plan

Types of data and samples. We will be generating (1) hydrographic data (conductivity, temperature, depth, fluorescence, oxygen sensors), and virus and picoplankton abundance estimates (epifluorescence microscopy and flow cytometry) at the two primary study locations (Station ALOHA and coastal California), (2) RNA and DNA metagenomes from 12 samples at those locations - two surface samples at ALOHA (summer & fall) and six stations at California, plus one DCM and one 200 m sample at ALOHA and one California station; (3) qPCR and RT-qPCR data on the abundance of select RNA and DNA viruses from 8 depths at 8 sampling times (ALOHA in summer & fall, and the 6 California stations), plus samples at one depth every 4 hours for 28 hours at ALOHA and station 76.7-51 at California; (4) experimental measurements of host range, burst size and latent period, ingestion rate, and decay rate for a suite of virus isolates; (5) targeted sequence data for new phytoplankton and virus isolates (18S rRNA for phytoplankton, DNA *po/B* for DNA viruses, and RdRp for RNA viruses); (6) phylogenies for phytoplankton strains used in host range analyses; (7) eco-evolutionary modeling of virus size structure.

We will share and archive data collected as part of this research project in compliance with the Division of Ocean Science Data and Sample Policy. All data from this project are considered within the public domain and the datasets deriving from the project will not be copyrighted. Hence, we do not anticipate intellectual property issues associated with the acquisition of the data.

Hydrographic data. Sensor data from hydrographic casts will be processed and archived on BCO-DMO and the National Centers for Environmental Information. Flow cytometry and microscopy counts of viruses and picoplankton will be archived as CSV files in BCO-DMO.

Metagenome data. Sequence data will be submitted to the Sequence Read Archive (SRA) and assembled genomes will be deposited in GenBank of the National Center for Biotechnology Information. An inventory of, and metadata associated with, the metagenomes will be archived at BCO-DMO project web site for this project.

Non-sequence experimental data. Measurements of virus host range, burst size and latent period, ingestion rates by different grazers, decay rates, and statistical analyses will all be provided as supplemental information in the associated publications and deposited to BCO-DMO at the project web site.

Targeted sequence data. Targeted sequence data will be deposited to GenBank, in accordance with the NSF Sample and Data Policy. Sequence

accession numbers and associated environmental data will be deposited to BCO-DMO, to provide an oceanographic record of the isolation locations of the different virus and phytoplankton strains.

Phylogenies. Phylogenies of phytoplankton strains used in host range analyses will be formatted and deposited to the Open Tree of Life, with metadata in accordance with MIAPA checklist for phylogenetic metadata.

Model code. All code used in statistical and dynamical models will be stored as annotated, fully reproducible R scripts. These scripts will be included as supplemental data in the resulting publications, and deposited to BCO-DMO.

Isolated strains. We will isolate new strains of *Proteomonas* hosts and *Florenciella* viruses. We will negotiate with a culture collection, e.g., National Center for Marine Algae and Microbiota (NCMA), American Type Culture Collection (ATCC), Culture Collection of Algae at the University of Texas at Austin (UTEX), or Roscoff Culture Collection (RCC) about accepting the cultures for maintenance. The hosts and viruses will be submitted as soon as practical (no later than the date on which their descriptions are published). Pure cultures not accepted for archiving will be maintained alive at UH for as long as feasible, but for no less than three years following description/publication. Frozen or lyophilized stocks of cells and viruses will be prepared as a genetic resource should the cultures fail.