Data Management Plan

We have read and agree to abide by NSF 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 and final reports.

Lead Investigator: Matthew B. Sullivan
Institution: Ohio State University
Collaborators: Bonnie B. Hurwitz (University of Arizona), Joshua Weitz (Georgia Tech), Tara Oceans and Malaspina Consortium collaborators
Project: Ecological impacts and drivers of double-stranded DNA viral communities in the global oceans
Solicitation Info: NSF PD 98-1650
Submission Date: Feb. 15, 2015

Project overview: Two global oceanographic research expeditions have been completed and data generation, curation and management has been centrally funded by the Consortium. This proposal seeks funding to support analyses of viral and relevant contextual data to look at ecological drivers of ocean viral community structure, as well as minimal data augmentation relevant towards these ends.

General data management and sharing overview: Since this project is largely built from data already collected and generated through two major oceanographic expeditions (Tara Oceans and Malaspina 2010), most of the data needed for the project are already archived and collaboratively shared with Consortium members through Consortium-level data archive and management efforts (an EU-funded Oceanomics project). Specifically, all raw data are hosted and available through the PANGEA database for meteorological, oceanographic, biochemical and plankton morphological data, and through the ENA repository for molecular data (http://www.ebi.ac.uk/ena/submit/tara-oceans-checklist). In additions to making our data available to the Consortium through these avenues, we intend to make all data that is annotated and curated through this project publicly available through 'iVirus'. The iVirus project represents a set of 38 apps and related public datasets dedicated to viral ecology recently developed by collaborator Bonnie Hurwitz and the NSF-funded iPlant team, https://de.iplantcollaborative.org/de/, and to which PI Sullivan and key personnel Roux have already contributed through other projects. Bonnie Hurwitz will be directly working with us to enable data accessibility and apps are made publicly available (see LOC-Hurwitz).

The few data generated by this project will be handled in the same manner: all the Single-cell Amplified Genomes data will be made publicly available through the expeditions websites and repositories, and curated annotated viral data extracted from these SAGs will be hosted separately on iVirus.

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). All curated data and tools / apps will be made available to the Consortium immediately and made available to the public in conjunction with the peer-reviewed publication that describe the materials.

Data archives: Viral metagenome and viral metaproteome raw data will all be stored in NSF-funded iPlant facilities through the iVirus portal. Viral metagenomes will also be submitted and archived at Metavir (http://metavir-meb.univ-bpclermont.fr/). 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, taxonomic networks, etc) will also be made available through iVirus. All records will be cross-referenced to sample data, and will follow the naming guideline of each expedition.