Data management plan

Our data management plan is based on guidelines established by the National Science Board and the National Science Foundation and covers dissemination and sharing of materials and data that are expected to be collected as part of the research described in the project proposal. **This project will be registered with BCO-DMO and data will be available or linked through there.**

We intend to make our data as open access as possible in the shortest amount of time that is needed for securing publications.

A. Access, Sharing, and Intellectual Property Issues

Bioinformatics analyses will be performed using the TAMU High Performance Research Computing (HPRC). With few exceptions data will be publicly available. All code used in data analyses will be accessible via an anonymous read-only access. Any user will be able to obtain development versions as well as release versions of the code. The codes are to be publicly available under an open-source license (such as GPL). All contributions will be reviewed and approved, providing a means to detect violations.

B. DNA sequences sharing and availability

1. Metagenomic data:

Metagenomic reads (raw and post-quality control), assembled contigs, genes and annotations will be made available to the public in the metagenomic server MG-RAST (Argonne National Lab) and Department of Energy Joint Genome Institute Integrated Microbial Genomes (IMG) and Microbiome Samples databases, as appropriate, using standard formats and protocols for these types of data. The raw reads, including the quality scores, will also be submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).

2. PCR sequence data:

Sequences of SSU rRNA and viral genes PCR amplicons generated in this study will be made public as soon as available and curated, by posting in the National Center for Biotechnology Information (NCBI).

C. Physical samples long term storage

All field samples (water) will be labeled at the time of collection with location, sample type, and date of collection. Aliquots of 2 ml will be archived at -80˚C. All processed samples (virus concentrates, DNA extracts, organic carbon extracts) will be labeled at the time of preparation with sample type and date of preparation. These samples will be archived in freezers, refrigerators, or other appropriate storage facilities available in Labonté's laboratory. When feasible, duplicate and triplicate samples will be stored.

D. Physical data

1. Metadata

Metadata (salinity, temperature, conductivity, nutrients concentrations, etc.) will be measured and submitted along the DNA sequence data. The data will also be made available to the public in the Supplementary Information of the associated manuscript.

2. Grazing experiments, viral production, and microbial mortality experiment data

Microbial counts and nutrients analyses data will be organized in Excel spreadsheets. Data will be available to the collaborators via Google Documents and to the public in the Supplementary Information of the associated manuscript.

E. Data preservation, archiving, sharing and accessibility

The long-term data storage and archival needs of the project are being served by the High Performance Research Computing (HPRC). HPRC currently has over 11PB high performance storage running GPFS, including 3.5PB Tiered storage (10 TB flash storage, 1 PB disk storage, 2.5 PB tape storage) which provides short and long term archiving storage.

Data will also be backed up on external hard drives, as well as on the Amazon Cloud. Project data sharing among collaborators will be achieved via shared access to these drives, as well as through Google documents and other local and cloud solutions. A privacy statement will be provided to users of the HRPC project wiki and repository. That statement will explain that posts are public, which should be obvious to users. Code contributors will be made aware of the license.

Software packages associated with this project will be maintained in an SVN/GitHub repository, so all released packages are preserved to maintain historical context. This SVN/GitHub repository and associated wiki will be hosted on HPRC systems and supported long after the life of the grant. These servers are routinely backed up, protecting against system failure and vandalism. In conjunction with Texas A&M University's Information Technology (TAMU IT), the HPRC is planning to develop a University Data Repository to support researchers storing large data sets and in meeting their research data management plans. The HPRC is regularly updating its data management plan in concert with the needs of Texas A&M research community and NSF guidelines.

F. Access to publication results

Any results that will be displayed in charts, figures, or images of a publication will be made accessible as soon as possible to the public either through supplementary material (e.g. alignment used to generate phylogenetic tree) and information or by submission of the DNA sequence or metadata in the appropriate databases.

G. Curriculum and outreach materials:

Syllabi for college-level classroom teaching will be assembled according to TAMU syllabus guidelines. Syllabi for outreach events will be tailored toward to the particular event and group.

Syllabi for outreach events will be shared on the research group's Drive and will be made available to external parties upon request. Outreach activities will also be promoted on the project website and blog (designed by students enrolled in a specially designed course at TAMUG). Key powerpoint slides and hand-outs will be made available through these media.