DATA MANAGEMENT PLAN

1. Types of data and samples

With this project we will generate field-based data and results from bioinformatics analyses. For field samples we anticipate the following general types of data: (1) Bacterial and viral abundance (2) electron micrographs of prokaryotes and viruses (3) estimates of the percentage of infected cells (4) metagenomic DNA sequences. Data from analyses will be generated according to established analytical protocols, with methodologies for each procedure included with the accompanying metadata. Samples and subsamples will be physically archived in appropriate locations (*e.g.,* freezers, refrigerators) until analysis. Types of samples will include: aldehyde-preserved and frozen crustal fluids, cell and virus concentrates for subsequent elution and fractionation, and nucleic acid extracts.

2. Data and Metadata Standards

Data quality will be assured through analysis of replicate samples, and proper accounting of standards, and controls. Data will be archived in multiple locations, including hard copies, laboratory computers, local servers and cloud-based servers. Our project will be registered with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) established through the Ocean Carbon and Biogeochemistry program at the Woods Hole Oceanographic Institution. We will follow the appropriate metadata standards outlined in the document "Data Management and Guidelines Manual" from the BCO-DMO project office. We will be generating a large amount of experimental and sequence data in this project that is not adequately served though BCO-DMO. DNA sequence data will be deposited into the GenBank public archive in accordance with the NSF Sample and Data Policy. Information on how to access the data (database location and accession numbers) will be posted with other project data in BCO-DMO. Both metadata and primary data will be submitted to the appropriate public data repositories (BCO-DMO, GenBank). In addition, to increase accessibility to project data and the dissemination of our research findings, we have budgeted funds for publications deriving from this project to appear as open-access articles in peer-reviewed journals.

3. Data Sharing, Reuse and Redistribution Policies

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. Biological data will also be maintained on computers in the PIs lab with backups on a portable external hard drive. Data will be made publicly available via GenBank and BCO-DMO as soon as possible after collection, but within 1 year of collection, and all project data will be made publicly available within six months of the project end date. The original data collector/ creator/ principal investigator does not retain the right to use the data before opening it up to wider use. All data will be made public after publication or at the end of the grant period. The processed fastq sequences will be submitted to the NCBI SRA (Short Read Archive).

Assemblies, function and taxonomic annotations will be uploaded to the IMG and the datadryad data repository (http://datadryad.org/)

The statistical models and the machine learning algorithm will be hosted on the project's website as well as to github (github.com), a web-based repository for source code. The implementation of the data management plan will be carried out by a graduate student with support and supervision from Mahdi Belcaid.

4. Policies and provisions for re-use, re-distribution

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. The data acquired and preserved in the context of this proposal will be further governed by the University of Hawaii's policies pertaining to intellectual property, record retention, and data management.

5. Plans for archiving and preservation of access

Data submitted to BCO-DMO are maintained in perpetuity by archiving at the National Oceanographic Data Center (NODC). Sequence data will be archived by submission to any or all of the following government-supported databases: National Center for Biotechnology Information, iMicrobe, and the Joint Genome Institute. The processed fastq sequences will be submitted to the NCBI SRA (Short Read Archive). Assemblies, function and taxonomic annotations will be uploaded to the IMG and the datadryad data repository (http://datadryad.org/). The statistical models and the machine learning algorithm will be hosted on the project's website as well as to github (github.com), a web-based repository for source code. The implementation of the data management plan will be carried out by a graduate student with support and supervision from Mahdi Belcaid.