DATA MANAGEMENT PLAN

Data products produced through this project include sequence data generated by the Department of Energy's Joint Genome Institute (JGI) and the laboratory of Co-PI Rappé: environmental genomic DNA sequences (metagenomes), environmental RNA sequences (metatranscriptomes), genome sequences from isolated viral and microbial strains, and DNA sequences from the amplified genomes of single cells. All of the sequence data and products produced by the JGI will be derived from existing samples whose metadata is already (or in the process of being) associated with an existing project in BCO-DMO entitled "Microbiology and biogeochemistry of Juan de Fuca Ridge flank borehole fluids" (http://www.bco-dmo.org/project/635868). The Co-PIs will generate nucleic acid sequence data derived from new samples collected over the course of this project. While recognizing that this project will require its own BCO-DMO project, we will also associate the sequence data and metadata derived from this project with this existing BCO-DMO project such that all metadata and sequence data can be cross-referenced and accessed simultaneously. Our overall goal is to facilitate its broader accessibility and use.

The research groups of all three Co-PIs will continue presenting their research at local, national, and international meetings, and will continue to publish in respected peer-reviewed journals. Our goal is that all publications are open-access.

1. Description of Data Types

Experimental:
Environmental DNA and RNA sequencing: Environmental sequencing from filtered fluid samples will generate high throughput, massively parallel (Illumina) sequence data from single-amplified genomes (SAGs), metagenomes, metatranscriptomes, and small subunit ribosomal RNA gene amplicons. DNA sequence data will also be generated from genomic DNA extracts isolated from microbial and virus strains.

Physical, biogeochemical, and other associated data: Data associated with fluid samples include date, location and sample line, temperature, salinity, pH, the concentrations of inorganic ions, organic carbon, cell abundance, and dissolved gases (hydrogen, methane, oxygen, carbon monoxide). Note that it is not possible to make all measurements (e.g. dissolved gases) from the full suite of samples, so representative samples will be analyzed.

For the purposes of this proposal, we also include isolated and cryopreserved strains of microorganisms or viruses, or limited-diversity enrichments of identified constituents, as a data type.

Derived:
Whole genome sequences: Genome sequence data from SAGs and metagenomes will be co-assembled into contigs and scaffolds and subsequently annotated to locate genes and putatively identify gene products via the Department of Energy Joint Genome Institute protocols, and viewed through the Integrated Microbial Genomes on-line portal. Whole genome sequences from isolated microbial and viral strains will be assembled and subsequently annotated in a similar manner.

2. Data and Metadata Formats and Standards

Illumina sequence data files will be stored as unprocessed qseq and fastaq data files. Metadata will be prepared in accordance with BCO-DMO conventions (i.e. using the BCO-DMO metadata forms) and will include detailed descriptions of collection and analysis procedures.

3. Data Storage and Access During the Project

The investigators will store project data and working files on laboratory computers that are backed up daily to external hard drives. External hard drives are also backed up daily to an on-site server with RAID data mirroring maintained by the project Co-PI Rappé. Raw Illumina sequence data files will be immediately backed up to an external hard drive, on-site server with RAID data mirroring, and off-site through the University of Hawaii High Performance Computing Resources. The products of processing and analyzing the Illumina DNA sequence data are backed up to an external hard drive and on-site server with RAID data mirroring. Generating a workgroup on the in-house Rappé lab server allows all project personnel to share and access files regardless of physical location.

Any isolated microorganisms or enrichments will be maintained in batch culture until confirmation that they can be cryopreserved. Upon successful cryopreservation, batch cultures will be maintained on an as-needed basis. Cryopreserved stocks will be maintained in two separate storage facilities.

4. Mechanisms and Policies for Access, Sharing, Re-Use and Re-Distribution

The Department of Energy's Joint Genome Institute is generating most of the sequence data for this project. Their data access policy (http://jgi.doe.gov/data-and-tools/data-management-policy-practices-resources/) reflects their strong advocacy for the public access of sequence data and its products generated by their facility. In fact, for most sequencing projects, raw sequence data are submitted to NCBI's Sequence Read Archive (SRA) for public access as soon as practicable, generally within 30 days. In addition to access through the NCBI, genome sequences and their annotations are made publicly available through the Joint Genome Institute's Integrated Microbial Genomes and Metagenomes on-line portals. Additional products derived from raw sequence data (e.g. assembled genomes and annotations) will be deposited in the appropriate National Center for Biotechnology Information (NCBI) database (e.g. GenBank) upon submission of manuscripts. GenBank accession numbers and Sequence Read Archive project numbers will be provided to BCO-DMO in an Excel spreadsheet or .CSV file. The project investigators will work with BCO-DMO data managers to make project data available online in compliance with the NSF OCE Sample and Data Policy.

When isolated strains are used in any manner in a publication, the will be deposited in the American Type Culture Collection (ATCC) for general distribution. Isolates that do not appear in a publication by the end of the two-year project will be deposited in the ATCC at that time. Limited diversity enrichments will be distributed directly to colleagues by request in the event that they cannot be deposited with the ATCC or similar collection (e.g. DSMZ, JCM).

5. Plans for Archiving

BCO-DMO will ensure that project data are submitted to the appropriate national data archive. The Investigators will work with BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data. Sequence data will be archived through both the NCBI and JGI, and linked with metadata through BCO-DMO.