DATA MANAGEMENT PLAN

Four main types of data will be generated in the proposed activities: biological material, sequence data, physical and chemical data, and mass spectrometry-generated raw data. We are committed to making all data publicly available through peer-reviewed publications, public databases, the Biological and Chemical Oceanography Data Management Office (BCO-DMO).

Biological material: Growth experiments will be conducted to evaluate substrate ranges, growth rates, and enzyme kinetics associated with SUP05 nutrient transformations. Isolates studied in these experiments are preserved and will be made available upon request. Growth rates, cell densities and substrate concentrations and utilization rates are critical to the proposed study and future studies to understand the roles of SUP05 cells in carbon, nitrogen and sulfur cycling. These data will be available through peer-reviewed publications and public databases (such as GenBank and BCO-DMO).

Sequence Data: Protein sequence data will be deposited in the National Center for Biotechnology Information (NCBI) under accession numbers assigned by NCBI.

Physical and chemical data: The collection of physical and chemical data from the ETN will be coordinated with chief scientists Rocap and Keil (see letter of support) to ensure these data are made publicly available to the community through the Biological and Chemical Oceanography Data Management Office (BCO-DMO), hosted by Woods Hole Oceanographic Institution and the Ocean Observatory Initiative (OOI).

Mass Spectral data:

Proteomic data from cultures: Proteomic analysis of the proposed culture experiments will generate raw mass spectral files from a Waters QTOF LC-MS collected in data independent HRMS-MS^E acquisition mode. Data is temporarily stored on a connected 8 bay, 16 terabyte RAID Synology box housed in the MMRC and maintained by the School of Oceanography IT support team comprised of two staff. MS^E spectra are subsequently processed using Progenesis for Proteomics (Waters), a software package that matches peptides to MS^E spectra for subsequent searching against annotated genomes and relative concentrations are calculated based on an internal standard. Raw mass spectral files are backed up on lolo for permanent tape storage where they can be accessed upon request.

Proteomic data from environmental samples: For the proposed project the generated data consists primarily of raw mass spectra generated by a Thermo QExactive. After data acquisition from experimental runs, all raw data is stored on a 180 terabyte online SAN connected storage cluster that is owned and maintained by the Genome Sciences Department at the University of Washington. The Genome Sciences department has 9 IT support team professionals that help with data management, storage and access. The servers are backed up every day. External access to resources is provided by a pair of SSH gateways. Every aspect of the current computational pipeline is installed and operational on the cluster and made available to the public via the web. Acquired raw files are converted to publically readable mzXML files that are an open-format available to other researchers. Data generated from the proposed work will be stored on the

Genome Sciences server for a minimum of 3 years after conclusion of the award. Tape back-ups are also performed, making the data available indefinitely.

Metabolomics data: Mass spectral metabolomic data will be generated using either a Waters Xevo G2S QTOF LC-MS, Waters TQS triple quadrupole LS-MS or a Thermo QExactive LC-MS. These instruments are connected to our Synology drive for temporary data archiving and the UW data storage system lolo for permanent storage. There is currently no standard public repository for raw mass spectrometry files for metabolites. However, a minimum standard for data access will be to make these files publicly available. The Microbial Metabolomics Research Center at the University of Washington has this capability and will make all raw mass spectrometry files available upon publication.

Our goal will be to have an open access archive of our data and adequate computational space for manipulating data. File format standards for these data sets and electronic dissemination and preservation plans are as follows. The raw data and instrument settings from each mass-spec run are saved in the manufacturer's native file formats. For metabolomics analysis, the files will be converted to the open community standard mzXML file format. These data will be stored temporarily on our in house Synology drive and will also be accessible upon request through the collaboration file system described in the text below and in Figure 1. In addition, the metadata and tabular results will be made available through the collaboration file system in a timely fashion and deposited with federally funded clearing houses for electronic dissemination. Oceanographic data will be deposited within the Biological and Chemical Oceanography Data Management Office (BCO-DMO) according to best practices (http://bcodmo.org/resources). Morris and Ingalls have experience working with BCO-DMO to deposit data. Data will also be disseminated through publication of peer-reviewed articles.

Data collected under the project will be made available to the public with as few restrictions as possible. For data collected during the cruise, our policies will cohere with standing LTER policies regarding access. For experimental work, we plan for publication of most data with metadata after or in conjunction with primary publication of results, or at most two years after the completion of the study on the BCO-DMO website.

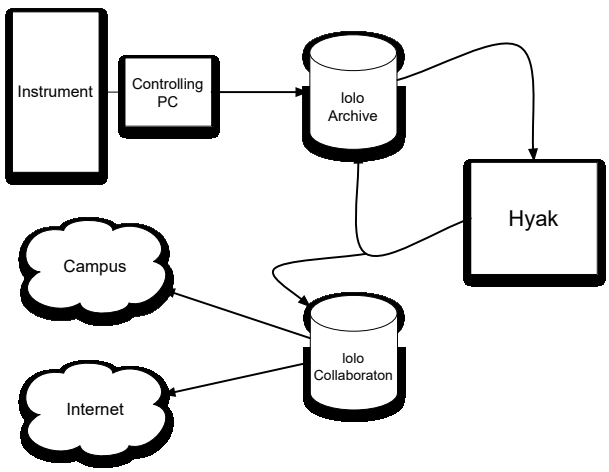

Figure 1: Workflow diagram illustrating how data generated using mass spectrometers in the Ingalls lab will be archived and placed in a collaborative workspace operated by the University of Washington, where it will be publically accessible.