DATA MANAGEMENT PLAN

The collaborative PI's on this project are well aware of the needs for broad sharing of scientific information and will include multiple activities to provide unrestricted access for other researchers and the public as final results are available. All field data collected under this program will be made available as per NSF guidelines within 2 years of collection via published manuscripts, publicly available final reports to NSF, the BCO-MPO data management office (http://www.bco-dmo.org/ and eventual data archiving with NODC. As described below, subsets of the data are available almost immediately through open science efforts.

The Policy of Open Science

In 2013, the Nunn lab implemented technology to achieve our goal of being completely open science. We use an online platform (Evernote) for our lab notebooks so that the research we do is shared in real time. The Noble and Harvey labs also exclusively use online lab notebooks so that all analysis steps can be shared with collaborators instantly. Further, Dr. Nunn uploads published and unpublished mass spectrometry data within 1 week of collection onto Chorusproject.org for free public download, viewing, and analysis. Data and presentations resulting from this research from all PIs will be shared similarly.

Expected types of data, samples, software and curriculum materials:

Expected data types include raw genomic data from Illumina Hi-Seq, raw proteomic data from tandem MS data, amino acids data from GCMS, total bacterial counts, total and dissolved organic carbon, and additional chemical metrics (i.e., as collected in time course experiments).

Data: For the proposed project the majority of generated data consists primarily of raw mass spectra generated by a Thermo Fisher Scientific QExactive or TSQ mass spectrometer. Thermo Scientific MS instruments generate .raw files that are recognized by nearly all MS-centric software platforms. Furthermore, raw files will be converted to the standard, open mzML format, making the data accessible to all researchers. Further, all simplified organic and proteomic data will be provide as supplemental files in published manuscripts as tab delimited or .csv files (more details below).

Curriculum Materials: All curriculum-resources that can be provided via the internet will be available on http://www.environmentalproteomics.org/outreach.html.

Contextual Biological and Chemical data:

Harvey will oversee data management for the environmental data and related organic and chemical biomarker concentration measurements and work with the staff of the Biological and Chemical Oceanography Data Management Office (BCO-DMO) to help link the broad array of data generated during this project. Physical, chemical and biological data collected will eventually be archived at the designated National Data Centers (http://www.nodc.noaa.gov/) in different formats as needed for rapid dissemination and will be made available through different channels. BCO-DMO staff will provide additional assistance to coordinate interactions with other repositories that are natural locations for archival and access to molecular data (e.g., NCBI, GenBank, CAMERA). Harvey was on the data management design team for the recent Bering Sea Ecosystem Study (7 years, 54 PI's) and is familiar with organizational needs. Harvey also regularly submits data to his current BOEM project for public dissemination of organic contaminants and environmental data (http://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.nodc:0123220.

Omic data access and public release:

Metaproteomics: All mass spectrometry raw files collected through this project will be uploaded onto the https://chorusproject.org/ website under the publically available database and listed under the "NUNN LAB". Associated location and preparation details will be provided with the analysis files. All raw files can be accessed and downloaded by anyone at any time. All files will be uploaded within 1

week of collection. The Chorus Project was created for the sole purpose of providing research scientists and developers with the ability to store, analyze, and share their mass spectrometry data.
Metagenomics: Raw data from the Illumina Hi-Seq will be transferred to the labs of the PI and co-PIs where it will be used for analysis and archived. Within one month of acquiring the raw data it will be uploaded into the appropriate repository at NCBI (the Short Read Archive). Data will not be released immediately, but this will provide additional redundancy of storage. Data will be released once the results are published or within three months after the termination of the project. Metaproteome fasta databases generated from each site will be provided upon request (as these files are huge and can be recreated directly by translating the public release of the metagenome or metatranscripome).

Software access and distribution:

All software generated will be made freely available to the public and include step-by-step tutorials. Every aspect of the current computational proteomic pipeline (i.e., Comet, Crux, and Tide) is already available to the public via the SourceForge repository (http://cruxtoolkit.sourceforge.net/) and we will continue to maintain updates to software.

Archiving data, samples, and software

Data: All mass spectrometry data generated within the University of Washington Proteomics Resource is sent to the Nunn lab storage space within the department of Genome Sciences. All departmental systems backed up to tapes are regularly shipped off-site for 3rd party vaulted storage. Tape backups are done using a pair of tape libraries with a combined 22 drives and a local storage capacity of ~6 petabytes. The 180 terabyte online SAN connected storage cluster is owned and maintained by the Genome Sciences Department and is part of the high performance cluster.
The Genome Sciences department has 9 IT support team professionals available to help members of this department including Nunn and Noble. External access to resources is provided by a pair of SSH gateways. Genome Science-ITS managed servers are backed up every day.
Samples: All samples will be digested and analyzed using mass spectrometry. Remaining peptides will be stored in a -80ºC freezer for up to 7 years and available for further analysis if requested by outside labs.
Archiving Software and updates: Dr. Noble's lab will manage code development through the Subversion version control system. All versions of all source code files will be retained on Genome Sciences storage volumes and backed up as described above.
Software: All released versions of created software will be made available and retained publicly, through SourceForge or a similar public repository.

Period of data retention and data storage for preservation of access:

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. Since tape back-ups are performed on the aforementioned cluster, data will be available to the public indefinitely. In addition, all mass spectrometry raw files are now uploaded onto the chorusproject.org website under the publically available database and listed under the Nunn lab as environmental samples. All raw files can be accessed by anyone at anytime.

Newly discovered protein and peptides identified using the six-frame translational searches against tandem mass spectral data will also be deposited on the National Center for Biological Information website. Novel research and developed software will be made publically available and will be reported in peer- reviewed publications. Taxonomic and function-specific peptide sequences will be detailed in Evernote, on our website, and in peer-reviewed publications.