Data Management Plan

"Collaborative Research: Dynamics of dissolved organic phosphorus production, composition and bioavailability along a natural marine phosphate gradient"

University of Hawaii at Manoa (UHM), Lamont-Doherty Earth Observatory of Columbia University, and Woods Hole Oceanographic Institution subcontract to PI Ruttenberg at UHM

We will conform to the Directorate for Geosciences' guidelines concerning Dissemination and Sharing of Research Results, NSF Data Sharing Policy, and the NSF Data Management Plan Requirements (http://www.nsf.gov/bfa/dias/policy/dmp.jsp and subsequent links therein).

We will also adhere to the Division of Ocean Sciences statement (May, 2011; available at http://http://www.nsf.gov/pubs/2011/nsf11060/nsf11060.pdf) concerning Dissemination and Sharing of Research Results.

All appropriate data products associated with this proposal will be submitted to BCO-DMO (http://www.bco-dmo.org/).

Here we provide a summary of the types of data we will collect in our project and our plans to preserve and allow sharing of these results.

Data Description:
Field samples (*in situ* and shipboard incubation) and laboratory culture samples will consist of filtrates (0.2 μm) and particulates. Geochemical data generated will consist of dissolved and solid phase carbon (C), nitrogen (N) and phosphorus (P) concentrations on whole water filtrates and on molecular weight segregates (MWS) generated via sequential ultrafiltration of 0.2 μm-filtered water. DOP bioavailability data will be generated via phosphohydrolytic enzyme incubations of DOP-MWS.

Biological data will consist of phytoplankton isolates and will derive from culture experiments, and consist RNA sequence reads, and derived data products from subsequent analysis.

Two types of mass spectral data will be generated. The first is chromatographic mass spectral (LCMS) data collected on the FT-ICR MS, which includes the ion chromatograms and the resulting lists of retention times and high resolution mass:charge values for individual compounds. Fragmentation spectra (MS/MS) are generated for selected ions in these analyses and will be stored on our data server. The second type of data is LCMS data collected on the triple-quadrupole (TSQ) mass spectrometer. These data include extracted ion chromatograms for selected precursor-fragment pairs as a function of retention time along the column. For solution phase DOP in MWS, untargeted metabolomics data will be generated. For particulate matter from culture experiments, targeted metabolomics data will be generated.

Mechanisms for Access and Data Sharing:
All data collected during this project will be stored on computers at the PIs' respective institutions, and backed up on central servers at those institutions. The PIs and postdoctoral researcher working on the project will have external hard drives on their computers to enable automatic backing up of all project data.

The mass spectrometry data will be stored in two locations. The Kujawinski lab is committed to making its data publicly available through MetaboLights, a repository established by the

European Bioinformatics Institute. The Kujawinski lab has been submitting data to MetaboLights since 2015 and these data have recently been re-used in a publication by Lawson et al. (2017), which is evidence of the utility of the lab's metabolomics data. The Kujawinski research group has also developed a boutique database for mass spectrometry data comparisons at WHOI (Longnecker et al., 2015). This database has been designed to facilitate storage of mass spectrometry data directly connected to the metadata associated with each sample. LC/FT-MS and LC-MS data are stored for at least one year after acquisition. Low-quality data due to analytical difficulties are transformed and the primary data is deleted. Chromatographic primary data and data from other analyses is stored for at least 5 years, but lists of retention times and compound concentrations can be stored indefinitely due to the reduced file size. By storing the raw data we can reprocess samples as new algorithms are developed and implemented. The 40 TB RAID server at the WHOI FT-MS facility is backed up nightly by the WHOI Computer and Information Services department.

Through the culture transcriptome sequencing and subsequent analysis, we will be generating and storing significant amounts of sequence data and the associated analytical files. The data will be stored and on two redundant 15T Raid 5 servers, which are backed up weekly to the Columbia server network. In this manner there is redundancy in preserving the raw data and the associated analytical files. Data analysis will be performed on a custom pipeline run via Dyhrman's XSEDE network access at the National Center for Genome Analysis on their Mason cluster. The Mason cluster at Indiana University (mason.iu.xsede.org) is a large memory computer cluster configured to support data-intensive, high-performance computing tasks. It is populated with genomics software intended for use by researchers. Any custom scripts will be made available via GitHub, or protocols.io. We have several strategies for data archival and accessibility. Transcriptome and sequences from cultures will be uploaded to the NCBI Gene Expression Omnibus (GEO) and/or the Short Read Database (SRA) with detailed metadata as appropriate, and the link will be archived with the project page at BCO-DMO. Derived data products (e.g. contig relative abundance, contig - metabolite linkages) will be included as supplementary material to publications when applicable. Dyhrman is actively exploring informatics and data management solutions to the unprecedented challenges presented by these large sequence datasets in an earth systems context. For example, she participated in the NSF workshop "Engaging the Ocean Science Omics Community in Envisioning Cyberinfrastructure to Archive, Access and Analyze Massively Parallel Environmental Omic Data". We will explore additional avenues for long-term archiving where there are appropriate metadata repositories, as alternatives become available.

Any unique axenic phytoplankton isolates used for the culture studies will be deposited with the NCMA, or provided upon request.

We are committed to making our results available to the scientific community as quickly as possible. We plan to present results at national and international meetings, and to regularly deposit our data files on a public database. The PIs are committed to publishing our results in widely available, high-quality scientific journals.

References added in the metabolomics section:

Lawson TN, Weber RJM, Jones MR, Chetwynd AJ, Rodrı́guez-Blanco G, Di Guida R, Viant MR, and Dunn WB (2017) msPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics. *Anal Chem* **89**, 2432-2439.

Longnecker K, Futrelle J, Coburn E, Kido Soule MC, and Kujawinski EB (2015) Environmental metabolomics: databases and tools for data analysis. *Mar Chem* **177, Part 2**, 366-373.