DATA MANAGEMENT PLAN

We will make our environmental, chemical, sequencing, and culturing data accessible within 2 years of collection through BCO-DMO and the use of our University affiliation. For any data stored in other repositories, we will provide links to our project page with BCO-DMO. We recognize that our data types will be diverse, spanning ecological data to DNA sequence information to cultures and we will use the appropriate archive. Further, much of our data will be analyzed through bioinformatic pipelines and we will document our commands and procedures.

1. Types of data

a. This project will collect a variety of information including: physical environmental characteristics (temperature, salinity, nutrients), DNA sequences of microbial population structure and function, notes and pictures on cultures, nutrient and isotope data, and metabolomic and proteomic mass spectral data.
b. We will collect this data with both observational (macrophyte and microbial assemblages) and field and lab experimental (e.g., isotope additions) techniques.
c. We will use of a variety of computer software packages to process the data, including basic manipulation in R. DNA sequence data will be analyzed through Qiime and the Anvi'o pipeline.

2. Data and Metadata Standards

Collection of all data types will be tracked with reproducibility and replicability as guiding principles, from swabs in the field to annotated sequences uploaded to public databases.

a. Metadata: For all collections will follow the established guidelines for the Genomic Standards Consortium for the Minimum Information for Marker Gene Sequences. All collections will have notes regarding method and location.
b. Molecular data: There are a multitude of safeguards built into our analysis pipeline to assure data quality of metabolomic results and DNA sequences. For sequencing data, we will implement all data quality assessments that are built into the Illumina next-generation sequencing protocol. After sequencing, we will impose additional data quality standards in the analysis pipelines of Qiime and the Anvi'o platform.
c. Culturing data: notes, pictures and DNA sequences will be archived.
d. No datasets will be covered by copyright.

3. Policies for access and sharing and provisions for appropriate protection/privacy

a. We expect to share with other researchers the primary data, and any samples or physical collections as appropriate.
b. PI's will share the data as it becomes available and is posted to the web sites. Data will be available within 2 years of the end of the project, unless it is part of a student thesis in which case it will be available upon submission of the thesis to the University. We will distribute large files by means such as rsync or cURL or, upon request, by exchange of hard drives for larger data transfers.

4. Plans for archiving and preservation of access

For all data types below, we will aim for persistent IDs through existing repositories or those associated with our publications.

a. Molecular data: All DNA sequence data is automatically archived and accessible when it is run at the Argonne National Lab or UChicago. We will further archive all curated sequence data with the NCBI. Environmental metadata associated with DNA samples will be deposited within the Global Environmental Sample Database (GESD) as part of the Earth Microbiome Project (EMP) and the NCBI. Metabolomic data will be converted to open source files (.mzXML) for archiving. Proteomic mass spectral data will be archived in original binary (.raw) and XML (.mzML) formats, along with peptide/protein identifications, in the public MassIVE repository (https://massive.ucsd.edu), which is a member of the international proteomeXchange consortium (http://www.proteomexchange.org) that facilitates proteomics data sharing.
b. Data on cultured microbes: We will work with the US Culture Collection Network (USCCN, https://usccn.org/) to archive the microbial taxa we preserve at -80.
c. Environmental Information: We will upload our data and data descriptions (metadata) to the KNB (http://knb.ecoinformatics.org/index.jsp), an NSF-funded program for sharing ecological and environmental data. Links to these data storage programs will be available on our websites. All isotope data and environmental data will be deposited at the Biological and Chemical Oceanography Data Management Office (BCO-DMO) based at WHOI. Pfister has archived data at both places previously.
d. Bioinformatic workflows: The code for our analyses will be accessible through GitHub.
e. We will utilize the UChicago Library Knowledge Base to maintain accessible copies of project related publications.
f. The original data as well as any published papers will be archived. All archival facilities have fully redundant off site back up of computer files.

Data Responsibilities of each PI

Pfister – Archive all sequencing and metabolomic data, metadata and bioinformatic workflows. Continue to deposit Washington coastal environmental data to BCO-DMO. Submit microbial cultures to USCCN. Maintain bioinformatic workflows on github.

Light– Submit microbial cultures to USCCN. Archive all bioinformatic workflows on github.

Waldbauer - Archive all isotope and proteomic data.

We will update our data management plans as technology continues to develop, observing best practices for open-science and sustained access to public resources.