Data Management Plan

Data Policy Compliance: The goal of our data management plan is to ensure that all data and products are archived, well-documented, shared, and accessible for the long term. The data management plan presented here is compliant with the NSF Division of Ocean Sciences Sample and Data Policy (NSF 17-037). We will cooperate and collaborate with all appropriate data repositories, particularly the Biological and Chemical Oceanography Data Management Office (BCO-DMO; http://www.bco-dmo.org/), and adhere to the overall data sharing philosophy. Since the proposed work leverages samples from a research cruise (AT42-22) relevant data products will be archived with or be linked to the cruise data in the Marine Geoscience Data System (MGDS; http://www.marine-geo.org) and Rolling Deck to Repository (R2R; https://www.rvdata.us). Additionally, relevant samples were collected with the support of the NSF Science and Technology Center C-DEBI. Sharing and deposition of data products will also follow the C-DEBI data management plan.

Data Types Generated During this Effort: The project will produce: (1) metagenomic sequencing data, (2) metatranscriptomic sequencing data, (3) relevant code and processed data for downstream bioinformatic data analysis, (4) quantitative cell count (bacteria, archaea, microeukaryote, and virus) and protistan grazing data, and (5) open-source lesson plans for teaching R and Python in the context of microbiology and oceanography.

Data Standards: All code generated by the project will be annotated and shared publicly via GitHub pages: https://github.com/carleton-spacehogs and https://github.com/shu251. We will follow the metadata guides and references produced by the Marine Metadata Interoperability project (marinemetadata.org). All genomic sample collection will conform to the Genomic Standards Consortium's (GSC) minimum information about a genome sequence (MIGS) or the minimum information about a metagenome sequence (MIMS). We will also follow the principles articulated by Goldstein et al. (2003, Chemical Geology 202, 1-4) regarding open reporting of analytical metadata. Calibration standards, standard reference materials, internal standards used to quantify analytical precision, and other items analyzed as part of good analytical practice will be fully described in relevant publications. Ultimately, the PIs are responsible for the appropriate level of QA/QC that is consistent with best practices in their respective field. In some cases, there is no established protocol, and as such the PIs will endeavor to provide the data in a manner that enables other investigators to easily access the final calibrated and/or processed data, as well as replicate the conditions under which the data were collected.

Data Access and Sharing: All data will be made publicly available no later than two years after collection (NSF policy). To be shared publicly, data will be provided to BCO-DMO and linked to other relevant cruise metadata (MGDS & R2R), protocols (protocols.io), large data files (Zenodo) code (GitHub). The raw, assembled, and annotated sequences will be made publicly available as well as archived through the genome repositories GenBank (NCBI; http://www.ncbi.nlm.nih.gov/) and Joint Genome Institute Integrated Microbial Genomes (JGI/IMG; http://img.jgi.doe.gov/) server. Both of these databases are federally supported and stable and can be linked through BCO-DMO with Digital Object Identifiers (DOIs). We will use a master sample/metadata spreadsheet and database that will allow us to rapidly cross-reference samples to facilitate our publication efforts. This will also assist others who might be interested in these samples or data, including marine microbiologists and geochemists. Moreover, if data products are yielded by third parties integral to the successful completion of the proposed project, we will ensure that these too are made publicly available along the same timeline, via BCO-DMO.

Developed lesson plans for the R and Python programming languages will be made publicly available to other students and educators on Github, protocols.io, and ReadTheDocs. Both PI Hu and

Anderson have previously used these platforms to deliver open-access computational resources. Alongside code, lesson plans include out to implement curricula with existing microbiology and oceanography topics.

Data Archiving: As noted above, all data will be deposited into publicly accessible archives, including BCO-DMO, JGI IMG, and NCBI. Final archive will be ensured by depositing this data into the BCO-DMO within two years of acquisition, and where stable links to data housed in other repositories (i.e., NCBI, GitHub, or Zenodo) can be made. Stable links, such as DOIs, will enable the data to be findable for the long term. A guiding principle for data archiving and sharing is to prioritize the Findability, Accessibility, Interoperability, and Reusability (FAIR data practices; Wilkinson *et al.* 2016, *Scientific Data*) of all data types. The PIs will upload all relevant data types and take the responsibility to complete archiving, which includes releasing data online via repositories such as GitHub, ensuring that all data submitted to BCO-DMO, JGI IMG, and NCBI have associated metadata, and that all methods to generate and access data products used in peer-reviewed literature are thoroughly described.

R. Anderson: For data storage, Anderson owns a total of 70TB of usable space on a new NetApp shelf (24 6TB drives), physically located on campus and dedicated to academic/research. All data created by this project will reside on this machine. All data generated by proposal research will be archived. The NetApp server is regularly backed up via Snapshot technology. The PI and appropriate technical staff will be responsible for the management of the data, including physical storage, archiving, and submission to public databases. Management of the NetApp server is done by members of the ITS staff and the computer science department technical director in consultation with the PI.

S.Hu & J. Huber: All experimental data (raw, processed, or in progress sequence data) at WHOI will be stored on laboratory computers and associated Network Accessed Storage devices that are backed up daily by WHOI's automated backup service. PIs will be responsible for managing this data and working with WHOI ITS staff to maintain the computer backup service and access.

Policies for access and sharing: The following specific steps will be taken to protect privacy, confidentiality, and security of the data:

- Access to the server used to store the project data will require both explicit authorizations from the PI as well as password identification by the user.
- Authorization to access the server storing the project data will only be granted to personnel that have a legitimate need to access it.
- No data with any identifying information will be provided to anyone without explicit authorization and a legitimate need to access the data.
- Appropriate technical measures to ensure security of the system will be taken. This includes, but is not limited to: restrictive file/directory permissions, security measures for databases behind firewalls, user authentication, patches updated regularly, etc.
- Any real data collected will be stored and shared in accordance with regulations set forth by the institution providing the data.
- Any datasets used in R and Python lesson plans will be acknowledged alongside lesson instructions.
- Lesson plans for R and Python developed as a part of this project will be publicly accessible

Status of Funding Support for Data Management: The PIs have access to institutional data storage facilities and have experience managing the types of data we will generate, and do not require additional funding for this activity. Federally funded data facilities exist for each of the data products we propose to generate. If there are nominal costs for these facilities to handle our data we will cover those costs from the awarded funds to ensure long-term data preservation and access.