DATA MANAGEMENT PLAN

The PIs agree to support the NSF data management and data dissemination policies as described in the NSF Proposal & *Award Policies & Procedures Guide* (Chapter XI.D.4), NSF Geosciences OCE Approved data and Sample Repositories (www.nsf.gov/geo/oce/oce-data-sample-repository-list.jsp), and to comply with the NSF Biological Sciences Guidance: www.nsf.gov/bio/pubs/BIODMP_Guidance.pdf). The PIs are committed to observing the FAIR Guiding Principles for scientific data management and stewardship, ensuring proper curation of resulting data and deposition in public databases in a timely manner. Outlined below are the types of data and samples that will be collected and a description of how they will be managed.

The proposed research will generate lipidomic, geochemical, microvideography, amplicon sequencing, and meta-transcriptomic data from sampling throughout the water column, deck-board amendment experiments, and laboratory-based experiments. Before the cruise, the science party will develop a sampling plan that includes samples to be collected, shipboard and shore lab processing, sample storage, and sample shipment. During the cruise, we will keep a detailed log of samples collected.

Description of Data, format, and Standards

1. Raw lipidomic data: We will store lipidomic data in the MetaboLights repository (https://www.ebi.ac.uk/metabolights/) and GNPS-MassIVE (https://gnps.ucsd.edu/ProteoSAFe/static/gnps-splash.jsp).
2. Molecular data: All RNA Next-seq data (fastq files) will be uploaded to NCBI GenBank BioProjects. Quality controlled sequences will be stored in the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra).
3. Physiological, biochemical, and rate data: Nutrients (i.e. DOC/N/P, POC/N/P), enzyme activity, cell counts, microzooplankton counts, grazing and growth rates, and grazing behavior data will be uploaded to Biological and Chemical Oceanography Data Management Office (BCO-DMO; https://www.bco-dmo.org/) along with finalized lipid concentrations and metatranscriptomic and amplicon count tables.
4. Video Data: Video and image data produced through behavioral observations in this project will be in the form of HSMIS (High-Speed Microscale Imaging System) recordings, including: (1) time series of image files generated from high-magnification high-speed imaging of free-swimming ciliates and predator-prey interactions in cell culture flasks, and (2) time series of flow velocity vector fields generated from high-magnification, high-speed, time-resolved microscale particle image velocimetry (□PIV) measurements of the flow fields imposed by free-swimming ciliates. We expect the total data volume from HSMIS recordings will be from several hundred GB to a few TB. The acquired data will be initially stored in hard drives of local PCs and in portable hard drives as backup storages. The initial data will be stored at our laboratories at the WHOI and copied nightly through an EMC-networker based backup system at WHOI. Videos supporting publications will be deposited on Figshare (https://figshare.com).

Accountability

PI Edwards will be in charge of establishing cloud-computing and hard-drive based data storage at UC Berkeley, and will work with Co-PIs Johnson and Jiang to share and backup all project data on shared Google Life Science (cloud) account, and to archive these data in various repositories (see below). Since cloud computing resources at Berkeley and Google will be associated with a specific project server, anyone involved with the project will be able to take charge of data uploading and archival if needed. Before depositing or storing, these data will be carefully curated and linked with metadata, with all of the above members checking for accuracy. PI Edwards will be in charge of storing and depositing the lipidomic, molecular, and cruise station data, while Co-PI Johnson will store and deposit grazing and physiological data, and Co-PI Jiang will store and archive all video data (see below for details).

Data Sharing and dissemination

Code written to analyze and annotate the lipidomes and metatranscriptomes will be posted on the Edwards lab GitHub along with scripts developed for statistical analysis and modeling of the data. All raw reads will be deposited at NCBI, and alignments, phylogenetic trees, and annotation pipelines from this project will be made available through FigShare and GitHub. Phylogenetic tree data will be submitted to TreeBase database (www.treebase.org). Data from this project will be used to write articles for peer-reviewed scientific journals during the three-year lifetime of the project. Details of the analytical methods will be reported in these publications, and links to raw and processed data will be referenced. A data analysis workshop will also be developed which will use the meta'omic and geochemical datasets in a tutorial aimed at teaching students how to analyze various types of data in an integrated manner. The tutorial will be packaged in a virtual environment which will be posted to GitHub.

Data Archiving

Molecular Data: All RNA data (fastq files) will be deposited (archived) in the National Center for Biotechnology Information (NCBI; https://www.ncbi.nlm.nih.gov), within their Sequence Read Archive (SRA) database. All NCBI sequence data will be linked to a BioProject with associated metadata, and these will be specific for each isolate or strain of ciliate sequenced.

Image and Video Data: All image and videography data will be deposited on Dryad Digital Repository (https://datadryad.org) for archiving.