DATA MANAGEMENT PLAN

The proposed work will generate digital images of eukaryotic microbes (with focus on the ciliates); HTS analyses of amplicons from communities; transcriptomes and genomes from single cells; transcriptomes from communities; and gene trees. The PIs are committed to proper curation of the resulting data and deposition in public databases in a timely manner.

Field work. We will adhere to the data-sharing principles of the Global Earth Observing System of Systems (earthobservations.org/geoss.php), which encourages full and open exchange of data and metadata with the least practicable cost and delay. All data from cruises will be stored at the website of the Biological and Chemical Oceanography Data Management Office (BCO-DMO). This is critical for capturing the metadata associated with field work, including temperature, salinity, depth and GPS. Our past collaborative data are already available there (www.bco-dmo.org/project/560529).

Products of Research. Our research will generate both raw and processed data, as well as a variety of images and associated metadata. These data will be used for the generation of reports to NSF as well as peer-reviewed articles and presentations. We will submit HTS data in two forms: raw reads, and OTUs/contigs edited for quality and used in analyses. Raw reads will be deposited into GenBank's Sequence Read Archive (trace.ncbi.nlm.nih.gov/Traces/sra/)or its equivalent and OTUs will be deposited as environmental sequences in the NCBI nucleotide database. We will also deposit 'omics resources on MG-RAST (http://www.mg-rast.org/). All biological and environmental information will be reported consistent with the Darwin Core glossary of terms as maintained on the GitHub repository (github.com/tdwg/dwc). Biogeographic information will be deposited with the Ocean Biogeographic Information System (OBIS www.iobis.org/). As manuscripts are submitted, raw molecular data will be deposited onto NCBI, and phylogenetic data will be deposited in the TreeBASE (treebase.org) database as well as the OpenTree curator (devtree.opentreeoflife.org/curator) to enhance synthesis with other analyses.

Data Formats and Metadata. Local data storage will occur primarily in spreadsheets as most of the programs and python scripts that we use are capable of creating comma-delimited ASCII flat files, which are compact, designed for simplicity, durability, accessibility and interoperability with the vast majority of data analysis/processing programs past and present. We will adopt the Ecological Markup Language (EML) metadata standard for creating and maintaining metadata associated with the environmental measurements in our project. Upon publication, abstracts and these data will be deposited at Dryad (datadryad.org).

Photodocumentation: We will use light microscopy to document the various morphospecies characterized for this study. Digital images will be archived and be publically accessible through species pages created for the encyclopedia of life (eol.org). Images will also be archived locally on redundant storage devices, and backed up onto google drives where we have unlimited storage.

Python scripts: Python scripts used in this project will be made available through github. Our collection is open to the public: https://github.com/Katzlab/. Scripts are annotated with explanations of their goals and major steps.