DATA MANAGEMENT PLAN

DATA POLICY COMPLIANCE

The project investigators will comply with the NSF OCE Data and Sample Policy and the NSF Award and Administration Guide (Chapter VI.D.4).

I. DATA PLANS

1. Data And Materials Produced

Laboratory experimental data: A number of different metrics will be used to quantify the physiology and ecology of evolved cell lines. This will generate a series of datasets reporting (1) growth rates, (2) photophysiology including chlorophyll, photosynthetic efficiency, and photosynthetic rates, (3) respirometry, (4) cellular stoichiometry, (5) grazing rates, and (6) palatability to higher trophic level consumers. These data and their metadata will be tabulated in .csv and .txt files as appropriate, and uploaded to BCO-DMO and/or Dryad data repositories.

Genomes and Transcriptomes: nucleic acid sequences will be compiled into .fasta files; unique genetic sequences will be uploaded to the NCBI repository, with accession numbers provided to BCO-DMO. Because the raw datasets will be large, they will be submitted to the iMicrobe platform (where the MMETSP transcriptomes are also available). Metadata will be provided to BCO-DMO.

Evolved cell lines: The evolution experiment will generate 405 independent lineages of *Ochromonas* across all experimental treatments and replicates. These lineages will be archived twice per year by cryopreserving a subset of culture in the Moeller Lab -80ºC freezer, and will be available to other researchers on request. During Year 3 of the proposed work, PI Moeller will contact the National Center for Marine Algae and Microbiota (NCMA) at Bigelow Labs to ask whether she can submit representative evolved lineages to their collection.

COBALT model output: The COBALT model simulates surface ocean biogeochemistry at the global scale and outputs abiotic (e.g., nutrient, DOC) and biotic (e.g., phytoplankton, microzooplankton, larger grazers) variables. Detailed logs of model runs will be kept, raw model data will be stored on local data servers in PI Moeller's lab, and key output data will be stored as Netcdf (.cf) files. Metadata and specific outputs (e.g., DCM depth, C export) will be uploaded to the BCO-DMO repository as manuscripts are published. Additional output from model runs will be available to members of the community on request.

Other model and statistical analysis code: Mathematical simulations (e.g., of the MOCHA model) and analyses of the above empirical datasets, and model validation incorporating both empirical and modeling results, will require statistical code (e.g., using the programs R, Mathematica, or MATLAB). After manuscripts have been accepted for publication, code will be made available either as a supplement (depending on the journal) or by deposition on the first author's GitHub page. On publication, data packages will be archived and assigned unique DOIs with Zenodo to ensure long-term accessibility.

2. Standards, Formats And Metadata

For all datasets, we will deposit data in non-proprietary, open data formats (e.g., .csv, .cf, .txt, .shp, ASCII, etc.) appropriate to the dataset in question (see above). Metadata preparation will follow BCO-DMO conventions, and will include detailed descriptions of experimental/data collection procedures, and data analysis pipelines. Code will be deposited in formats associated with the coding language, and metadata will be included noting the versions of all software and software packages used. Wherever possible, we will choose open-source software (e.g., R, Python, Julia), though MATLAB and Mathematica may be used for some eco-evolutionary models.

3. Accountability (Roles And Responsibilities)

As the project lead, PI Moeller is responsible for sharing and managing collected data across project participants; this will be facilitated by cloud storage solutions especially Google Suite products that are supplied by the University of California. All project data will be backed up on a shared project Google Drive hosted by the University of California, Santa Barbara; these shared drives are owned by the organization rather than individuals, and thus are retained even after individual employees separate from

the university. Google Office products also include automatic version control, so that records of access and edit history are maintained. The project team will meet at minimum quarterly to discuss project plans and progress, including data generation and archiving.

In general, we expect initial data generation and curation to be led by the project personnel that will be first-authors of publications. Personnel will be trained in lab/field notebook maintenance, including detailed note-taking and daily review of entries. All lab notebooks and datasheets will be scanned/photographed and uploaded to cloud databases for backup.

4. Data Sharing (Dissemination Methods)

At time of publication or within two years of data collection (whichever is soonest), all relevant data and model code will be deposited in BCO-DMO and Dryad. Dryad is supported by UCSB and the California Digital Library (CDL). It implements the latest best practices in data publication, citation, and archival, including assigning DOI identifiers to deposited datasets and making data discoverable through such services as the Thomson-Reuters Data Citation Index, Scopus, and Google Dataset Search. Storage and access are guaranteed by the UCSB Library for at least 10 years, with support provided to transition to longer-term storage if desired.

5. Policies For Data Sharing And Public Access

Upon publication, archival records of project data and code will be created from GitHub repositories and Zenodo DOIs will be generated. Data will also be deposited in Dryad (see above for details on storage and access guarantees). Data and software packages will be licensed under Creative Commons License CC BY-NC, allowing others to download, re-use, and adapt our data and code, but only for non-commercial purposes. (New work deriving from this project must acknowledge the data/software originators.) Prior to publication, data will be available by request to project PIs.

6. Protection of Data: Security and Integrity

Investigators will store project data on laboratory computers at UC Santa Barbara and UC Davis, with all data backed up to a shared Google Drive. Any hard-copy data (e.g., lab notebooks and data sheets) will be photographed and/or scanned, and PDF copies will also be saved to these data repositories. This cloud storage will be used for data sharing across project locations. PI Moeller also requires that all project participants routinely back up computer hard drives daily using Apple Time Machine or another onsite external hard drive, and weekly to an offsite hard drive.

We will create an environment of research integrity by explicitly discussing best practices for data collection and archiving when onboarding project personnel. All data acquisition protocols will be reviewed by at least two project personnel, and training of new personnel will involve close scrutiny of methodology and data handling. We will implement peer-reviewed spot checks of data acquisition and entry to ensure that data are accurately captured and transcribed.

7. Archiving, Storage And Preservation

By placing data in Dryad, NCBI, and GitHub repositories (as detailed above), PI Moeller and team will ensure that data are appropriately archived with appropriate and complete documentation, in addition to maintaining independent laboratory archives during her career. Hard-copy data (e.g., lab notebooks and data sheets) will be photographed and/or scanned, and PDF copies will also be saved to these data repositories. Data backups will be conducted as described above (Item 6: Protection of Data).

II. INTELLECTUAL PROPERTY PLANS

Intellectual property created by this project is limited to manuscripts and computer code, which will be handled according to the guidelines outlined above. In brief, data and software packages will be licensed under Creative Commons License CC BY-NC, allowing others to download, re-use, and adapt our data and code, but only for non-commercial purposes. Publications will also be licensed under Creative Commons License CC BY-NC at time of publication. The University of California advocates for open access publication and provides (1) partial defrayment of publication fees, and (2) databasing of published papers written by University of California authors.