Data Management Plan

1. Expected Data Types, Formats, and Security

Physical Samples: We expect to collect approximately 520 samples for microbial metabolomic analysis, including eelgrass seeds, roots, and plant exudates from all three objectives. We also expect to collect over 900 isolates through our microbial cultivation approach and associated experimental samples to measure their exo- and endo- metabolomes. All samples will be preserved according to best practices for each analytical method (e.g., preservation in DNA/RNA Shield for sequencing, flash frozen and stored at -80 °C for metabolomics, glycerol stocks for microbial isolates).

Sequencing Data: We will obtain full-length 16S rRNA amplicon data generated using a PacBio Sequel II. *Fasta* files consisting of circular consensus reads will be submitted to appropriate data repositories. We will also generate short and long-read metagenomic data using an Illumina NovaSeq and PacBio platforms. Raw, submitted files will be in a *fastq* format. Genomes generated from metagenomic and isolate data will be submitted as *genbank* files. All sequencing data will be generated at the DNA Technologies Core at UC Davis.

Metabolomics Data: Metabolomics data will be generated from eelgrass exudates and following sediment incubation experiments. All metabolomics data will be stored and deposited as *.mzXML* files.

Other data: We will also generate a number of other datasets including eelgrass physiological data and associated metadata. These data types will be stored and deposited to relevant repositories as .csv files.

Analytical Pipelines: We anticipate adapting bioinformatic pipelines to facilitate data analyses associated with the project. This includes processing and analyzing amplicon data, generating genomes from metagenomic samples, and mapping expression data.

Educational Materials: As part of this proposal's broader impacts, we aim to create a microbial symbiosis CURE aimed on training undergraduates in microbial ecology research. Products from this educational effort will include course learning outcomes, CURE-based lesson plans, and primary data generated by undergraduates.

2. Data Standards

The proposed research will produce a broad range of original data spanning from measurements of eelgrass physiology (e.g., photosynthetic efficiency, respiration, metabolic rates), to sequencing (metagenomics), and mass spectrometry results (e.g., metabolomics). The PIs (Sogin, Stachowicz, Eisen) will train all students, technicians, and postdocs in relevant methods. Datasets collected will include relevant metadata agreed upon by the PIs and based on the Minimum Information about any (x) Sequence (MIxS) standards for sequencing data (Yilmaz et al. 2011) and the Metabolomics Standards Initiative for mass spectrometry results (Spicer, Salek, and Steinbeck 2017). Each dataset will also include appropriate standards needed for best practices, including instrumentation used, methods applied, data processing information, sampling procedures and access restrictions. All project members will keep both physical and electronic field and laboratory notebooks. Physical notebooks will be scanned weekly and uploaded to an Open Electronic Lab Notebook (Open ELN) hosted on a project specific GitHub page. Open ELNs help to facilitate reproducible and open science that aims to share both outcomes and processes of scientific research. Furthermore, the open ELNs will

help to share activities across campuses to ensure all project members remain involved in daily research activities. All data and associated products (e.g., ELNs) will be stored electronically under a standardizing naming scheme decided upon by the PIs prior to the start of the project (e.g., incorporating date in file name).

3. Roles and Responsibilities

PIs Sogin, Stachowicz, and Eisen will be responsible for implementing the data management plan. Our data management plan will include ensuring data are accurately collected, processed, stored, and archived. All project personnel will be trained in best practices for collecting and managing data. PIs will ensure all new project personnel are onboarded each year to ensure consistency in data management. PIs Sogin will oversee the collection and storage of mass spectrometry data. PI Stachowicz will oversee the collection and storage of eelgrass physiology, experimental data and relevant metadata for each experiment. PI Eisen will oversee the collection and storage of sequencing data and microbial isolates. PIs will document compliance with the data management plan in the Annual Project Report.

4. Dissemination

We will work with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) to archive and make available metadata with links to other public repositories.
All raw physiological data and all metadata will be deposited into figshare and given a permanent DOI that can be cross linked with other data types (e.g., sequencing and mass spectrometry results). Raw sequencing reads and associated metagenomically assembled genomes will be submitted to NCBI's Short Read Archive or Genbank. Culture isolates will be deposited into US based culture collections such as atcc.org. Educational products will be submitted to CUREnet. All code used for data analysis will be either published along with the submitted manuscripts or made available on GitHub. All data generated from this proposal will be disseminated to the public as soon as possible either through publication (partially enabled by preprint servers) or within 2 years following data generation.

5. Data Sharing

We will practice FAIR Data Principles (Wilkinson et al. 2016). Specifically, all data generated from the proposed research will be **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable (see below). To achieve FAIR data practices, we will link data accession numbers to associated metadata and reproducible code that is well documented. These links will be made available to the public either through publications in peer-reviewed journals, on the project associated GitHub page, or through publication on our labs websites.

6. Archiving

Physical samples will be stored in a -80 °C freezer within the PI labs for at least 5 years following publication. All data will be deposited into public repositories. All data will be backed up on a shared cloud server (e.g., dropbox, google drive) where possible. When data size is limiting, data will be backed up on lab-servers. All physical lab notebooks will be retained in the lab for at least 5 years following project completion and open ELNs will be hosted on a project-specific github repository.