Data Management Plan

Types of data. This project data will consist of RNAseq and metabolomics datasets consisting of NMR and chromatographic mass spectral data. RNAseq data will be derived from co-cultures of marine phytoplankton and bacteria, with metadata detailing culture conditions and sampling methods. Metabolomics data will be generated during analysis of co-culture experiments and from metabolite verification, substrate utilization assays, and unknown compound identification.

Data and Metadata Standards and Release. Nucleic acid sequence data will subjected to two QC pipelines: removal of low quality reads and removal of rRNA sequences. Sequence data will be formatted in NCBI-approved formats for submission as a BioProject to the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) and the iMicrobe cyberinfrastructure at the University of Arizona (http://imicrobe.us). Relevant research context metadata associated with the RNA-seq data will be deposited according to the Genomic Standards Consortium specifications for the "Minimum Information about a Genome Sequence" (MIGS). Analytical laboratory data from NMR or mass spectrometry will be archived in native format on servers at UGA or WHOI, and derived data products will be distributed in tabular format. Mass spectral data will be processed by MAVEN and the quantity of each compound will be exported to Matlab or R. All data will be stored in formats (e.g., spreadsheets, PDF files) that are readily accessible by common software. Reference spectra for metabolites will be deposited in the MetaboLights and Metabolomics Workbench databases. Data will be made available at the time of the first publication that uses the dataset, or by request from an interested researcher, or within 2 years of collection.

Policies for Access and Sharing
This project will conform to NSF standards for data access and sharing. Metadata will be available immediately after data are archived, and data files will be openly and freely available on the web within two years from date of collection if not sooner. Primary (raw) data will also be archived along with the finalized data, and links to web-accessible files will be provided in the data set metadata. RNA concentrates will be stored for at least two years in a frozen sample repository in the Moran lab. These materials will be made available to researchers upon request.

Data Sharing
DNA Sequence Data: The RNAseq data will be deposited both at the NCBI SRA and, in a more accessible form, at the iMicrobe project at the University of Arizona. iMicrobe is a data commons for microbial datasets that enables both access and analysis. Project information, sequence data and assemblies, metadata, and other associated files are publicly available. The Moran lab has previously deposited genomic and metagenomic data in iMicrobe.

Metabolomics Data: All of the raw and primary data generated in this project will be deposited, along with complete metadata, on the NIH-funded Metabolomics Workbench at UCSD and the MetaboLights repository. The Metabolomics Workbench supports metabolomics data from all types of organisms and techiniques, including ^{13}C labeling, fraction identification, and metabolic flux measurements. We have previously deposited datasets to the Metabolomics Workbench, which is shared automatically with the MetabolomeXchange international data aggregation site. This ensures that not only will the data be properly archived and available to the public at no cost, but that it will be easily available to researchers around the world. We will also deposit data in the MetaboLights repository (http://www.ebi.ac.uk/metabolights/), which is sponsored by EMBL-EBI in the United Kingdom. Copies of all the data will be maintained on local UGA data

storage through the Institute of Bioinformatics, and on the RAID server at the WHOI FT-MS facility for the duration of the project.

NMR Computational Scripts and Pulse Sequences: The Edison lab has a GitHub site where we will deposit MATLAB and other code needed to analyze the NMR metabolomics data that we generate. We use private GitHub trees for internal development, and we will post stable versions for no charge on a public site through GitHub. Pulse sequences developed for NMR studies will also be shared freely through this site, on NMR sites, or through Bruker software upgrades.