DATA MANAGEMENT PLAN

Data management and sharing are essential to the goals of the proposed research. The PI is committed to working with the Biological and Chemical Oceanography Data Management Office (BCO-DMO), an NSF-funded data repository, to archive and make all data sources publically available. The PI has corresponded with BCO-DMO and is assured that they will be able to assist with archiving and storing data generated from this proposed project. During years 2 and 3 of this project, the PI has budgeted salary time to organizing, managing, and publishing data from this project.

This project will produce large amounts of data in the form of: (i) coral exometabolomic data, (ii) genomic amplicon sequence data, (iii) metatranscriptomic data consisting of annotated assemblies and raw sequence reads (iii) picoplankton cell counts, and (iv) ancillary metadata from the field and aquaria experiments (e.g., latitude/longitude, temperature, PAR, etc.) and (v) peer-reviewed publications.

Raw data management, archival and storage

Data collection and analysis processes as well as contextual details will be documented in individuals' laboratory notebooks. All field and experimental metadata as well as picoplankton cell count data will be electronically recorded and managed by data type using the software Microsoft Excel, stored on laptop computers and backed up daily using external hard-drives (in the field) or to a remote location at WHOI.

The raw and processed next-generation sequencing data (amplicon genes and metranscriptomic reads) will be stored on a server located in Apprill's laboratory, which is backed-up nightly. Raw amplicon sequence data will be submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive depository (http://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi). This submission will include the raw fastq files, and metadata about the sequences that conforms to MIMARKS, the Genomic Standards Consortium (http://gensc.org). Accession numbers for all sequences will be made available in respective publications. Metatranscriptomic data, including raw sequence reads, annotations, and the associated metatranscriptomic metadata, will be submitted to the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena). ENA is a data repository that is capable of archiving the diverse data types (reads, assembly data, and annotated data) generated through sequencing. ENA's centralized archive structure and strong collaborations with NCBI and the DNA Data Bank of Japan (DDBJ) make it an ideal repository for sharing and linking the metatranscriptomic data with the amplicon sequence data stored within the Sequence Read Archive depository (NCBI). The targeted and untargeted metabolomic data will also be stored on the server in Apprill's laboratory. Raw experimental data, metadata, protocols and metabolite data will be submitted to the MetaboLights repository (http://www.ebi.ac.uk/metabolights/). All data within these repositories will be made available upon publication.

BCO-DMO archival and integration of project data

The PI will work with BCO-DMO to archive, integrate, and link all data from the genomic and metabolomics repositories and make it available for use.

Publications

Project results will be published in open access, peer reviewed publications with links available to the data (BCO-DMO, NCBI, ENA and MetaboLights).

Policies and provisions for re-use, re-distribution

There will be no embargo periods for political/commercial/patent reasons. Further, there will be no permission restrictions placed on the data. Biological data will be made available following collection and analysis. Data dissemination will be noted in the publications within the Materials and Methods section to inform the scientific community of the data availability and accessibility. All nucleic acid sequence data will be available through NCBI or ENA and metabolite data through MetaboLights. These data sources

are free of charge and open to the public. We will retain the right to hold data prior to publication only if a conflict of interest seems warranted.

The dissemination of the biological data to be collected for this proposed research will not restricted by any ethical or privacy issues, copyright concerns or restrictive licenses. As discussed above, all the data collected will be made readily available to the scientific community through various datacenters, published manuscripts in open access peer-reviewed journals, and upon request to the affiliated researchers.