INSPIRE Track 1:Microbial Sulfur Metabolism And Its Potential For Transforming The Growth Of Epitaxial Solar Cell Absorbers

Data management plan:

Data Policy Compliance: The research products generated in the proposed effort include biological samples; genetic, genomic and transcriptomic data; chemical and geochemical measurements and materials characterization measurements. Here, we present a data management plan that is consistent with both the NSF Division of Ocean Sciences (NSF 11-060) and the Division of Materials Research (NSF 11-1) Sample and Data Policy. Accordingly, we will cooperate and collaborate with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) to archive any and all biological data as appropriate, and in a timely manner, adhering to the overall data sharing philosophy (section II).

In general, we will:

a) Provide both inventory metadata and primary data within specified time limits (section III, A), including any and all marine environmental and genomic data (as in section VI, D).
b) Address data sharing outcomes in annual and final reports (section IV).
c) Archive any and all biological samples collected during the proposed research expeditions at Harvard University and, after 2 years, make them freely accessible to other researchers upon request.
d) Consistent with the guidelines presented in the Division of Materials Research guidelines 11-1, which require that "*Investigators are expected to share with other researchers, at no more than incremental cost and within a reasonable time, the primary data, samples, physical collections and other supporting materials created or gathered in the course of work under NSF grants. Grantees are expected to encourage and facilitate such sharing*", any and all primary data, samples, physical specimens or collections, and any and all associated materials will be made available to other investigators upon request, and within a timely manner.

Pre-cruise: We will have no substantive activities prior to the expedition, and as such no relevant data (meaning, data that is of value to the broader community) will be generated.

Cruise: All metadata collected during the cruise (LAT, LONG, temperatures, data resulting from any CTD operations) will be archived in both the R2R and BCO-DMO repositories as appropriate. All samples collected during the cruise will be sub-sampled and photodocumented for delivery to the Ocean Genome Legacy (OGL), and facility funded to archive biological material for any and all investigators who require the materials. The biological samples will be preserved for genomic DNA as well as RNA and protein analyses. All contextual data (e.g. sample collection site, weights, etc) will be provided. We will work with the BCO-DMO to provide an indicator to other investigators that such samples are freely available through the OGL. This reduces the complexities of making requests from individual PIs.

All other data (e.g. sequence data, etc) will be collected and archived as described below.

Post-cruise management of products and data: The specific products and data that will be collected in the proposed effort are listed in Table 1. They will be managed as follows:

1. 16S rRNA and mitochondrial CO1 sequence data will be stored indefinitely on Girguis laboratory computers, as well as on the Harvard Research Computing Odyssey Cluster, which offers over 2.5 Petabytes of raw storage for its users and includes both daily checkpoints and off-site backups of stored data. In addition, we will ensure final archive by depositing the data in the Short Read Archive (NCBI) and GenBank (NCBI) within 2 years of acquisition. Associated metadata for every sample will also be submitted to the BCO-DMO.
2. Within two years of acquisition, assembled and annotated symbiont transcriptomic sequences will be made publicly available as well as archived through the genome repositories GenBank (NCBI), as well as the Joint Genome Institute's (Department of Energy) Integrated Microbial Genomes and Metagenomes (IMG/M) system. Metadata associated with the genomes will be stored on IMG/M and the BCO-DMO. Assembled transcriptomic sequences that are identified as derived from the eukaryotic, mollusc hosts will be made publicly available as a GenBank (NCBI) Whole Genome Shotgun project for use by other researchers. In addition, all data will be stored indefinitely on Girguis laboratory computers, as well as on the Harvard Odyssey Cluster (see above for details).
3. Geochemical data from the lab incubations will be stored indefinitely on Girguis laboratory computers and stored on the Harvard Odyssey Cluster. Final archiving will be ensured by depositing these data into the BCO-DMO within two years of acquisition.

Table 1: Description of the data that will be collected as part of the proposed research effort.

Type of data	*Brief description*
1. Population and diversity data	Microbial 16S rRNA genes collected *in situ* or during lab incubations and treatments.
2. Microbial host gene expression data	Transcriptomic sequences from *in situ* collections and experiments.
3. Geochemical measurements	Measurements of ion and volatile concentrations during lab incubations.
4. Materials characterizations	Extensive materials analyses (as presented in proposal), collected from natural hydrothermal sulfides and materials produced during the bioreactor incubations.