CAREER: DO EXTRAORDINARILY LONG ENZYME LIFETIMES DRIVE MICROBIAL COMMUNITY ASSEMBLY IN DEEP SUBSURFACE SEDIMENTS?

DATA POLICY COMPLIANCE

All project members will comply with the data management and dissemination policies described in the NSF Award and Administration Guide (AAG, Chapter VI.D.4) and the Division of Ocean Sciences (OCE) Sample and Data Policy, publication 17-037.

Additionally, publication in journals will require compliance with common journal policies on archiving raw sequence data and derived products (e.g. metagenome assembled genomes) in relevant archives, such as the Sequence Read Archive.

Finally, this project will involve human research, in the form of research on the effects of our educational program with students in East Tennessee Freedom School. The Human Research Protection Program, promulgated by the University of Tennessee Office of Research Integrity and Assurance, controls handling of sensitive human subjects data. We will design research methods to comply with this policy, and will obtain Institutional Review Board approval prior to beginning our work with human subjects.

PRE-CRUISE PLANNING

Since the research group is entirely at the University of Tennessee, cruise planning will be done in person within my research group. We will invite other relevant research groups who would like to occupy empty berths as a cruise of opportunity, since we expect to collect far more mass of sediments than we can use. We will coordinate with these groups, as well as with the Sikuliaq captain, via teleconference. This planning will culminate in a Science Implementation Plan prior to each cruise departure, to be agreed on by all PIs and the ship captain.

The goal of each cruise is to collect marine sediments via Jumbo Piston Corer, so little data will be collected shipboard. Cruise events will be recorded via the R2R event logger (if available) as well as on paper logs, to be scanned and made available as PDF documents. Cruise reports will be prepared and promulgated via the Steen Lab website.

DESCRIPTION OF DATA TYPES

Samples

Sediment samples will be collected from four sites in the Gulf of Alaska via Jumbo Piston Corer and Slow Corer, as described in the Project Description. The Jumbo Piston Corer returns a maximum of 30 meters of 10.16 cm diameter core. From those sediments, we expect to generate

1. **Porewater geochemical data**, as described in the Project Description. File types: .csv ASCII files
2. **Event log**, including event numbers, start and end dates/times, and locations as GPS coordinates. File types: R2R event logger files.
3. **Cruise underway data** that is collected by default (e.g. temperature, salinity, etc). File types: .csv ASCII files.
4. **Enzyme activities** of extracellular enzymes in sediments will be measured shortly after returning to land, as fluorescence measurements from a Tecan fluorescence plate reader. Data type: instrument raw files, derived .xlsx and .csv ASCII files.
5. **Shotgun metagenome data** to be sequenced by an outside facility. Data types: .fastq ASCII files, as well as derived files (assemblies, metagenome assembled genomes) as .fasta ASCII files. Accession numbers to be provided to BCO-DMO
6. **16S rRNA gene amplicon data** to be sequenced by an outside facility. Data types: .fastq and .fasta ASCII files. Accession numbers to be provided to BCO-DMO

Experimental data

1. **Enzyme activities** of heterologously expressed enzymes as well as manipulated sediments (e.g., melting temperature data) as fluorescence measurements from a Tecan fluorescence plate reader. Data type: instrument raw files, derived .xlsx and .csv ASCII files.
2. **Model results** of enzyme lifetimes and geochemical rates as .csv ASCII files.
3. **Bioinformatic predictions** of enzyme structure, as ASCII files.

Code

All data analyses will be done, to the extent possible, in human-readable code formats such as R scripts or jupyter notebooks.

DATA AND METADATA FORMATS AND STANDARDS

To the extent possible, all data will be stored as flat ASCII files, including measurements we collect by hand (e.g. geochemical measurements, model results) and nucleic acid sequence data (.fastq and .fasta files). A major exception is fluorescence data, which is created by the proprietary Tecan software as .xlsx files, but which we will immediately convert to .csv files and archive as such. Metadata will be prepared in accordance with BCO-DMO conventions and using BCO-DMO metadata forms and will include detailed descriptions of collections and analysis procedures. To the extent possible, the raw data, code used to analyze raw data, and results will be mixed in human-readable jupyter notebooks. All data files will be associated with a data dictionary including clear descriptions of data collection and analysis procedures and the meaning of each column in flat data files.

DATA STORAGE AND ACCESS DURING THE PROJECT

During the project, data will be stored on, as appropriate, one or more of the following:

- The Steen Lab's Google Drive shared folder,
- The Steen Lab server "public" hard drive space, which can be accessed by all lab members, or
- The Steen Lab group project space on ISAAC, UTK's high performance computing cluster (physically hosted on the 3 Pb lustre/haven DDN SFA14K file server)

The drive space and ISAAC space are managed by Google and University of Tennessee, respectively. The Steen Lab server's data storage system is under RAID level 6 control, which can withstand 2 drove failures without data loss. When possible, data will be duplicated across these systems.

Data size will be dominated by the four NextSeq runs, which will produce a total of ~800 Gb output. Since the Google Drive storage space is (theoretically) unlimited, the Steen Lab's server has 80 Tb total space, and the ISAAC file server has 3 Pb space, data volume will not be large compared to our storage facilities.

The exception will be human subjects data. These data will be "small" (no more than 1 Gb of notes and hand-entered survey data) and will be stored securely according to protocols to be defined according to IRB requirements, e.g. on ISAAC's "secure" enclave (meant for secure data such as human subjects data).

MECHANISMS AND POLICIES FOR ACCESS, SHARING, RE-USE, AND RE-DISTRIBUTION

The Steen Lab is committed to open data policies. We commit to making all data and code publicly available as soon as possible via lab members' individual GitHub pages (with the exception of human subjects data, as described above). Lab members are encouraged to keep all data analysis projects under Git control, and to regularly push those repositories to public GitHub repositories.

Immediately after completion of the research cruises, underway data and metadata will be submitted to the R2R project. Sequence data are too large to be kept under git control or to be posted to GitHub. In this case, data will be posted publicly as soon as possible to the Sequence Read Archive, and accession numbers will be added to the git repositories in order to unambiguously link data to analyses.

All data will be published under Creative Commons CC-BY license, which allows free reuse with attribution. We expect that other microbial ecologists, as well as possibly structural biologists or biochemists, will be interested in using these data in meta-analyses, as we have previously done with similar data.

Data will be archived at BCO-DMO as soon as possible after acquisition. All data will be made publicly available within two years of acquisition.

PLANS FOR ARCHIVING

Data will be archived over the long term with BCO-DMO when possible, in consultation with BCO-DMO managers. Raw sequence reads will be archived at Sequence Read Archive. Assembled metagenomic data and metagenome-assembled genomes will be archived at JGI.

ROLES AND RESPONSIBILITIES

PI Steen is the sole investigator and is responsible for all data management and data management plan compliance. All project members will be trained on appropriate data management to ensure that proper data management practices are followed in all phases of data collection, analysis, and promulgation.