COLLABORATIVE RESEARCH: DEFINING THE ROLE OF THE PAN GENOME IN EMILIANIA HUXLEYI ECOLOGY AND BIOGEOGRAPHY

DATA POLICY COMPLIANCE

The project investigators will comply with the data management and dissemination policies described in the NSF Award and Administration Guide (AAG, Chapter VI.D.4) and the NSF Division of Ocean Sciences Sample and Data Policy.

PRE-CRUISE PLANNING

N/A

DESCRIPTION OF DATA TYPES

1. **Genetic sequencing:** mRNA and DNA sequencing from *E. huxleyi* cultures grown in the lab will be collected. Sequencing will be performed at the Columbia Genome Center (New York, NY). All raw sequence data will be deposited in the short-read archive (SRA) through the NCBI. Assembled genomes will be deposited under the same project number through the NCBI's Assembly database. Assembled transcriptomes will be deposited on the NCBI's Transcriptome Shotgun Assembly (TSA) database. Associated annotation files will be uploaded with the genome and transcriptome files to the appropriate database. File types: Short-read archive (.sra), raw sequencing files (.fastq), assembled fasta files (.fasta), annotation files (.gff). Repository: NCBI; accession numbers will be provided to BCO-DMO.
2. **Physiological experiment data:** Physiological experiments carried out on 10 different *E. huxleyi* strains will be conducted in the lab. Physiological and chemical parameters of the cultures will be collected over time including: growth rate, cell size, Fv/Fm, chlorophyll, particulate organic carbon, particulate inorganic carbon, dissolved nitrogen and phosphorus, and particulate nitrogen and phosphorus. File types: csv files. Repository: BCO-DMO.
3. **Bioinformatic pipelines:** Analysis of all genetic sequencing data will be done through the construction of reproducible pipelines designed in Snakemake. These pipelines will be made public on GitHub and archived and provided a doi through Zenodo. Repository: Zenodo; doi will be provided to BCO-DMO.

DATA AND METADATA FORMATS AND STANDARDS

Sequencing data will be stored in .fastq files, which are standard for combining quality with sequencing information. Metadata will be prepared in accordance with BCO-DMO conventions (i.e. using the BCO-DMO metadata forms) and will include detailed descriptions of collection and analysis procedures.

DATA STORAGE AND ACCESS DURING THE PROJECT

The investigators will store project data (including csv files, raw sequencing data, data analysis products and PDFs of scanned logs and notebooks) on PI Alexander's laboratory server that is backed up by the WHOI's CIS service. PI Alexander also has established an account with WHOI's Google Drive program for data storage and sharing. This will facilitate data sharing between PIs. Personal computers in all laboratories are backed up daily using Apple Time Machine to an onsite external hard drive, and weekly to an offsite hard drive.

It is estimated that in total the sequencing data from this project will amount to about 1Tb of raw data. PI Alexander has 80Tb of data storage capacity as part of WHOI's high performance compute system and will be able to store this data as well as manipulate it.

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

All genomic sequences will be deposited in the National Center for Biotechnology Information (NCBI) database GenBank upon submission of manuscripts. GenBank accession numbers will be provided to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) .csv file and metadata will be provided using the BCO-DMO Dataset Metadata submission form. Data sets produced by the science party will be made available through the BCO-DMO data system within two-years from the date of collection. The project investigators will work with BCO-DMO data managers to make project data available online in compliance with the NSF OCE Sample and Data Policy. Data, samples, and other information collected under this project can be made publically available without restriction once submitted to the public repositories.

Data produced by this project may be of interest to chemical and biological oceanographers, and climate scientists interested in the role of biogeochemistry in the global climate system. We will adhere to and promote the standards, policies, and provisions for data and metadata submission, access, re-use, distribution, and ownership as prescribed by the BCO-DMO Terms of Use (http://www.bco-dmo.org/terms-use).

PLANS FOR ARCHIVING

BCO-DMO will also ensure that project data are submitted to the appropriate national data archive. The PI will work with BCO-DMO to ensure data are archived appropriately and that proper and complete documentation are archived along with the data.

ROLES AND RESPONSIBILITIES

Lead PI Alexander will lead the data management effort with support from co-PI Dyhrman. This project will generate several different forms of data which will be managed as outlined in this document. We will work closely with the Biological-Chemical Oceanography Data Management Office (BCO-DMO: http://www.bco-dmo.org/) to ensure that data used in our analyses and outcomes from our experiments and computational studies are made publicly available and in accordance with NSF guidelines, including formatting and metadata content. Results from this work will be presented at international meetings (e.g. Gordon Research Conferences, ASLO, Ocean Sciences meeting) and will be published in peer-reviewed publications. We have included funds to support open access publications when possible.