DATA MANAGEMENT PLAN

Data Policy Compliance: This project will comply with the NSF OCE Data and Sample Policy, with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) as the primary data repository. The DNA sequence information will be archived with NCBI and linked with our data at BCO-DMO, as well as with our open access code repositories and notebooks.

***Nature of Data and Collections*:** Multiple data types will be generated in this project including:

- Amplicon and as needed genomic sequence data, stored as raw data (long-term storage), as processed quality controlled data, and as assembled contigs for analysis.
- Experimental physical and biological data (e.g., temperature, light, salinity, pH, and physiological measures and metadata) stored as jpg, csv, and linked to R scripts.
- Analyses of the above datasets will require statistical code (e.g., using the programs R and associated packages). Code will be made available either as a supplement, or by deposition on the first author's GitHub page within a OpenScienceFramework repository with a DOI and archived in BCO-DMO.

***Sample and Active Data Storage*:** All samples will be collected under appropriate permits and protocols through the government of French Polynesia and UC Gump Station. Nucleic acids and raw samples will be retained and archived on site as a backup where possible, transported to and stored in the Putnam lab at -80° C for long-term storage, when not fully consumed in analyses. Upon acquisition from the sequencing or analytical instruments and facilities, data will be quality controlled and added to project servers at URI and immediately following QC, will be submitted to the appropriate NCBI repositories for public access and backup. All data on the project will be accessible by all persons involved in the project. Data will be backed up, specifically the original and one copy will be stored on the hard drive of local computers and servers as write only, whereas an additional copy will be used to share research data with the global scholarly community and will be available for public access (e.g., GitHub and OpenScienceFramework).

Data Archival: Data will be archived permanently in the original data format and also in common, non-proprietary formats (e.g., tiff, csv, txt, fasta, etc...) to facilitate future data usage. Physical and physiological data will be archived at BCO-DMO. Sequence data will be uploaded to NCBI and linked to BCO-DMO. For each publication, raw data and reproducible pipelines (including scripts used to create publication figures) will be archived permanently at BCO-DMO. Within two years of data collection, the primary data will be made publically available on BCO-DMO following the Division of Ocean Sciences Sample and Data Policy.

***Documentation and Metadata*:** Documentation for this project will include the formation of written methodologies for sample collection and processing, made openly available on the PI's GitHub page and lab webpage, and published as peer-reviewed or online repository methodologies where applicable (e.g., Molecular Ecology Resources, Protocols.io, GitHub). Quality control will be conducted at each stage of the data acquisition, processing and analyses, including the development of metadata forms detailing the outline of the project, instrumentation

used, the format of data, QA/QC standards and controls, and funding source amongst other details. Metadata forms will be utilized to create and organize the project database for local use, and upon publication for increased ease of data dissemination (see "Publication and Presentation" section below). Analyses will be scripted to facilitate reproducible science. Metadata associated with this proposed research, including information on sites, experiments, and data collected (e.g., date, time, location, experimental treatments and maintenance, and environmental variables measured) will be documented for all data.

Policies for Data sharing and Public Access: Policies for access and sharing will include provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements. All of the raw data and processed data generated in this study will be made publicly available upon quality control. Raw data from secondary procedures including mapping and genome feature analysis will be available in real-time on via online lab notebooks and code repositories. Data will be in non-proprietary formats. Physical samples will be made available upon request where not consumed by analyses.

To ensure accuracy and data tracking, the project will have a specific data use policy including:

- User requests require current and valid contact information that will be used by the PI for tracking and documenting data usage.
- Users are required to cite the project publications and acknowledge the funding source, NSF.
- Users have the final responsibility for any errors in their external and secondary analyses, while the PI and project team will conduct quality control on the primary data and ensure the accuracy of the primary data to the best of their abilities.
- The PI and project team will not release any private or confidential information to the public, and in-house team will be password protected.
- The PI and project team will retain intellectual property rights, except where explicitly released for publication and documentation.

***Publication and Presentation*:** Our results will be disseminated in presentations at scientific meetings and peer-reviewed journal articles along with popular press and social media. All significant findings from the proposed research will be promptly prepared and submitted for publication with inclusive authorship that accurately reflects the contributions of those involved. NCBI Accession numbers, and code and data repository DOIs for each sample / study will be provided in all publications generated by the proposed work.

Roles: The PI is responsible for supervising all data management in cooperation with the project team. All team members are responsible for data collection, quality control, internal database management/curation, and data publication as applicable to their research responsibilities within the project. All trainees associated with the project will be trained in and involved in data collection and quality control across the process from collection, to publication, to archival.