Data Management Plan

To support the reproducibility of the science conducted under this proposal, this project aims to manage data in an open access and transparent manner. This project will compile and store a large number of specimens (e.g., fecal, coral tissue slurry, pelleted symbionts), collected metadata, digital images, and high-throughput sequencing data. Laboratory notebooks will be archived and secured in the PI's laboratory and will be available for review by all interested scientific entities. Specimens will be available to the research community upon request.

NSF OCE Data Policy Compliance: All field and experimental data will be submitted to the Biological and Chemical Oceanography Data Management Office (BCO-DMO) for archiving and public dissemination. The data will be contributed to BCO-DMO within 2 years of their production to comply with NSF OCE data dissemination and archiving policy.

Access and Sharing of Ecological Data: This proposal will generate a large amount of data on the distribution and abundance of Symbiodiniaceae in fish feces, and in corals and nearby seawater. To properly store this information, database software (Microsoft Access & SQL) will be used to manage the data and facilitate statistical analyses. All species abundance, distribution, and diversity data for Symbiodiniaceae, fishes, and corals will be submitted to the Ocean Biogeographic Information System (OBIS) (http://www.iobis.org). We will also store the data in a freely accessible folder through Rice University's Shared Computing Service Center.

***Access and Sharing of Next-Generation Sequence Data*:** Sequence data will be deposited in the appropriate major databases, such as the federal National Center for Biotechnology Information's (NCBI) GenBank/EMBL/SRA databases. We will share nucleic acid sequences with wider research communities through deposition in the publicly available the Meta Genome Rapid Annotation using Subsystem Technology (MG-RAST) (http://metagenomics.nmpdr.org/), and the Earth Microbiome Project's global environmental sample database (http://www.earthmicrobiome.org/global-environmental-sample-database/). We will also contribute data to the SymPortal (https://symportal.org) database. Doing so will allow for cross-comparisons with hundreds of studies and facilitate raw data dissemination, promoting the reuse of data generated under this proposal. We will store all generated raw sequence read libraries in a freely accessible folder through Rice University's Shared Computing Service Center.

***Data Entry/Management**:* One of the keys to a successful project is data centralization and standardization through the use of commonly available tools for data entry and sharing. In addition to relying on systems like GitHub for sharing data and code, the lab will also employ easy-to-use, widely available tools like DropBox and Google Docs for managing files and generating products associated with this project. We use standardized sheets for data collection that are accessible to the lab group through Google Drive. The PI will oversee procedures related to data quality control and data collection standardization efforts (e.g., technical implementation, documentation, training, etc.), as well as the general implementation of the data management plan. Postdocs, graduate and undergraduate students in the lab will be trained to use these data management tools.

***Analysis of Sequence Data**:* Strategies for analyzing high-throughput sequencing data are described in the specific objectives of this proposal. The Correa Lab will follow the latest directives from the Genomic Standards Consortium (GSC) for the development of the minimal information checklists for any marker-gene amplicon datasets we generate. These datasets, "Minimal Information about a Marker Sequence" (MIMARKS), provide a curated standard format layer for the acquisition and display of information associated with sample acquisition, processing, handling, sequencing, and analysis. These are community standards, agreed upon using consensus and updated where necessary by annual meetings of the GSC (www.gensc.org). These standards are recognized by the International Nucleotide Sequence Database Collaboration and reported by a keyword (GSC) for compliant sequences. We will adhere to both standards for sequencing data generated using this proposal. All of the data generated in this study will be made publicly available upon publication in a peer-reviewed journal or within 2 years of the completion of the project. To generate results, the Correa Lab will use published and standardized pipelines wherever possible for the projects conducted under this proposal. If no standardized pipelines are available, the Correa Lab will generate novel, Nextflow-integrated pipelines with proper documentation that will be submitted for peer-review when possible. Nextflow is a bioinformatics pipeline manager that enables scalability and reproducibility of scientific workflows using software containers like Docker or Singularity (https://www.nextflow.io/). In either case, all bash code for sequence data processing and any subsequent analyses will be recorded in a text file (.txt) in an organized format with substantial commentary. R-based statistical analysis workflows will be stored as Rmarkdown files with extensive additional comments explaining the results and reasoning of each step and links to the raw data. The project-specific text and RMarkdown files, as well as any novel workflows, will be shared in an organized format in a public repository on GitHub (https://github.com/CorreaLab).

Data Backups: In addition to user backups, the Correa Lab uses storage space on both Rice University's computer cluster and on the commercial Dropbox software. All researchers also individually back up to hard-drives approximately every two weeks.

Data Publication and Presentation: We aim to publish data collected under our Research Plan in peer-reviewed international scientific journals in a timely manner following the proposed timeframe in the project description, and to use the data in teaching undergraduate courses; these practices are already routinely performed by the PI. We have budgeted funds to make some of publications resulting from this project 'open access' to allow for a broader community of researchers and the public to acquire these manuscripts.

Procedures to Manage and Maintain the Confidentiality of Personally Identifiable Information (PII) of Human Subjects Data: All human subjects data (collected to evaluate and assess the results of programs in the Education Plan of this proposal) will be in compliance with the policies of the Rice University Institutional Review Board (IRB). PII management occurs primarily through the management structure of the project. All data collection from human subjects will be exclusively conducted in coordination with the Office of Assessment and Evaluation of STEM Programs (see letter of collaboration from Dr. Chris Barr) and not shared outside of the evaluation team (i.e., project research assistants working with Dr. Barr). Following data collection, but prior to all data analysis, the evaluation team will conduct a data anonymization procedure. In this procedure, human subjects are assigned a randomly generated study ID. The evaluation team members remove all PII from the analysis dataset, and maintain a linking file that contains the randomly generated study ID, as well as the PII. This linking file is necessary because the outcome evaluation involves longitudinal data, and it is imperative to link pre- and post- data, as well as any delayed longitudinal data for students of RET teachers or participants in the RESP program. As stated above, all analytic datasets used by the evaluation team to investigate evaluation questions, will be stripped of all PII. With regard to the linking file that contains the randomly generated study ID and PII, it will be stored exclusively on Dr. Barr's University-provided Rice Box account. Rice Box is a cloud-based file storage system, using real-time backup, behind an HTTPS firewall, and only accessible on the Rice University IP block. At the conclusion of the study, when all longitudinal data have been linked, this PII study ID linking file will be deleted.