Data Management Plan:

Project: Impact of freshwater runoff from Hurricane Harvey on coral reef benthic organisms and associated microbial communities

Solicitation Info: NSF Unit: 1. OCE – Hurricane Harvey 2017

Our data management plan is based on guidelines established by the National Science Board and the National Science Foundation and covers dissemination and sharing of materials and data that are expected to be collected as part of the research detailed in the project description. We intend to make our data from the 2016 East Bank Mortality Event as well as the 2017/18 Hurricane Harvey-related cruises as open access as possible in the shortest amount of time that is needed for securing publications. For this collaborative project, the data management plan is agreed upon by all institutions, and may be considered common to all. This project will compile and store coral and sponge specimens and their associated microbial communities, collected metadata, photography, and high-throughput sequence data. Water column microbes and chemistry will also be collected. We will sub-sample DNA and RNA samples as appropriate for long-term storage and access at each of the laboratories. Physical laboratory notebooks (and digital ones as described below) will be archived in PI or co-PI laboratories and will be available for review by all interested scientific entities. These notebooks will be available to the research community upon any valid request.

Access and Sharing of Oceanographic/Ecological Data*:* This proposal will generate a large amount of oceanographic metadata, species abundance/distribution data, and sequence data on the microbiomes of the samples, as well as on host gene expression. To properly store this information, database software (Microsoft Access & SQL) will be used to manage the data and facilitate statistical analyses. All oceanographic metadata (e.g., seawater salinity, carbonate chemistry, nutrient chemistry) will be submitted to the National Oceanographic Data Center (NDOC) (http://www.nodc.noaa.gov/) and to the Biological and Chemical Oceanography Data Management (BCO-DMO) office (http://www.bco-dmo.org) in compliance with the guidelines for the Division of Ocean Sciences Data and Sample Policy. Further, all species abundance, distribution, and diversity data for sponges and corals will be submitted to the Ocean Biogeographic Information System (OBIS) (http://www.iobis.org). Image data will also be stored on CyVerse (http://www.cyverse.org/) for any user to access.

***Access and Sharing of Next-Generation Sequence Data*:** Sequence data not appropriate for the data management sites above (e.g., BCO-DMO) will be deposited in the appropriate major databases, such as the federal National Center for Biotechnology Information's (NCBI) GenBank/EMBL/SRA database. We will share nucleic acid sequences with wider research communities through deposition in the publically available Meta Genome Rapid Annotation using Subsystem Technology (MG-RAST) (http://metagenomics.anl.gov), CyVerse, and the Earth Microbiome Project and its global environmental sample database (http://www.earthmicrobiome.org/global-environmental-sample-database/). We will also contribute data to the QIIME database (http://www.microbio.me/qiime/index.psp), which will enable comparisons with hundreds of studies and raw data download.

Data Entry/Management*:* One of the keys to a successful collaborative project is a process of data centralization that uses commonly available tools for data entry and sharing. In addition to relying on systems like GitHub for sharing data and program development, the labs will also maintain easy to use, widely available tools like DropBox and Google Docs. The Correa lab will oversee the collection, documentation, and quality control of general data (e.g., GPS coordinates) collected as part of this project, as well as the general implementation of the data management plan. The graduate students, technicians, and post-docs in all labs will be trained to use these data management tools as part of their education.

Analysis of All Sequence Data*:* Our strategies for analyzing high-throughput sequencing data are described in the project description of this proposal. We will follow the latest directives from the

Genomic Standards Consortium (GSC) for the development of the minimal information checklists for any genomes, metagenomes, and 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 using consensus and updated where necessary by annual meetings of the GSC (www.gensc.org). In addition, these standards are recognized by the International Nucleotide Sequence Database Collaboration (INSDC) and reported by a keyword (GSC) for compliant sequences. Analysis of all tag-seq transcriptomic data will follow the previously established pipeline publically available on github (https://github.com/z0on/tag-based_RNAseq). Downstream analyses will be conducted in the R statistical environment and all code and associated counts files will be made publically available upon publication. We will adhere to both standards for sequencing data generated using this proposal. All data will be made publicly available as soon as modeling and quality controls are completed.

To ensure that analyses involving numerous steps on the command line are replicable, we will use a system of "runnable lab notebooks". For each analytical product, we generate a 'procedure' text file to document the exact steps of the analysis, starting from the shared raw data present in the sequencing center repository. However, rather than being static text, the file will be a BASH script (with extensive additional comments explaining the results and reasoning of each step) *to allow large portions of the analysis to be regenerated in a single command-line step*. This simple system provides an easy way to share procedures with collaborators or lab-members.

Data Backups: In addition to user backups, we use both an on-site cluster with storage at Rice University's Shared Computing Cluster (SCC) and the commercial Dropbox software. TAMU and Rice faculty use unlimited backup on Google Drive for all raw data. All researchers will also individually back up hard drives nightly. Fastq files from transcriptome sequencing will be stored on Boston University's SCC, which has over a petabyte of data storage and all data are backed up for long-term data storage using a service called *STASH*. These layers of redundancy will be sufficient for internal analyses, paper manuscripts, etc. However, they are not sufficient for data sharing with the broader community or permanent data storage. We will archive sequences in appropriate online repositories (see above).

Coding Practices: All software will be implemented as an open-source software package in the Python programming language (wrapping ecological modules from R, etc. as needed). This package will be developed openly through GitHub, which allows for good versioning practices and community input. It is our policy that all scientific software be accompanied by test code. Test code acts in a similar fashion to control experiments in molecular biology, and helps to ensure that code changes (to optimize speed, etc.) do not introduce biological errors. All code will follow a consistent, documented coding style (http://pycogent.org/coding_guidelines.html) and include substantial commentary.

Data Publication and Presentation:
We aim to publish our data 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. When possible, we will make publications 'open access' to allow for a broader community of researchers and the public to acquire manuscripts easily.

Roles and Responsibilities:
Lead PI Correa will ensure compliance with this Data Management Plan. She will be responsible for deposition of all genetic data generated at Rice University. BU PI Davies will be responsible for metazoan transcriptomic data, UH PI Santiago will be responsible for sponge microbiome data, TAMU PI Sylvan will be responsible for deposition of water column microbiome data and nutrient chemistry and TAMU PI Shamberger will be responsible for deposition of water column carbonate chemistry data.