DATA MANAGEMENT PLAN

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 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.

Data Generated

The proposed research will produce several types of data, including genomic/transcriptomic sequence data, bioinformatic pipelines, non-genomic data (including lab, field, and modeling records), and model derivations and data.

Data formats and storage

Genomic and transcriptomic data (including raw read files from Illumina sequencing platforms) will be stored as compressed FASTA/FASTQ files stored on redundant RAID storage devices. This will include raw data (read files) from DNA amplicon and RNAseq data collection proposed in the Project Description. Genomic sequence data will be maintained in direct association with paired metadata providing full details of experiments that gave rise to these data. These metadata for all samples for which genetic sequence data is generated for in the grant will accompany public data depositions (NCBI).

Genomic and transcriptomic sequence data comprised mostly of raw read files from the Illumina sequencing platform will be stored as compressed FASTA/FASTQ files stored on redundant RAID storage devices. These RAID devices will include redundant drive striping and a reciprocal backup in house, and an offsite backup. In addition to raw data, we will also maintain copies of intermediate datasets generated during analyses (e.g., transcriptome assemblies) in FASTA form, or other appropriate standard filetype. To facilitate archiving, files will be highly compressed (e.g. bzip2) for mid-long term storage. Ultimately all raw files will be permanently archived in the NCBI Gene Expression Omnibus and Sequence Read Archive.

Bioinformatic Pipelines and Strategies for analyzing high-throughput sequencing data are described in the specific objectives of this proposal. The labs using genomic data will use published and standardized pipelines wherever possible for the projects conducted under this proposal. If no standardized pipelines are available, novel Nextflow-integrated pipelines will be generated and 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.

This proposal will also yield **non-genomic data**, including field data and images, experimental disease transmission data, immune proteins, and modeling results. Field notes will be collected on underwater paper *in situ,* collated into notebooks upon return to the lab and scanned into PDF files for storage in electronic format. Databases will be maintained by the PIs at their respective institutions (University of the Virgin Islands, Woods Hole Oceanographic Institution, Rice University, University of Texas, Arlington and Louisiana State University) and shared with the other PIs. Digital photographs taken in the field will be downloaded, labeled appropriately, and saved on external disc drives and on shared cloud storage providers (i.e., Google Drive and UVI's Microsoft OneDrive). Images will be stored in JPEG and/or TIFF formats. From these images, the image processing programs CPCe and ImageJ will be used

to calculate tissue loss rates. In compliance with NSF guidelines, hard copy versions of field and laboratory data will be retained for at least three years following the award period.

All datasets will be annotated with meta-data. The same procedure will be utilized for the data generated by our **modeling approaches**. Biophysical modelling data, including connectivity matrices and particle trajectories, will be stored in netcdf files for access and archival purposes. Data collected from each survey on the abundance and health condition [healthy, diseased (type of disease), injured] for each major reef-building coral species will be entered in Microsoft access or excel spreadsheets, organized, edited for errors and stored for later analyses. A Metadata file will be produced to have all relevant information on the methods and data collected readily available. The short-term data storage plan for the **non-genomic data** generated by the experiments and surveys will saved as metadata files and Excel spreadsheets (saved as .csv files) to an external drive, a Microsoft OneDrive data storage site maintained by UVI, and a UTA server that is backed up nightly. All modeling data will be saved to external drives daily. In addition, appropriate field and experimental data (with links to genomic data) will also be deposited in The Biological and Chemical Oceanography Data Management Office (BCO-DMO) housed at Woods Hole, MA (http://bco-dmo.org).

All samples will be collected and manipulative experiments performed under an appropriate permit from the Department of Planning and Natural Resources, Division of Fish and Wildlife, US Virgin Islands.

Data access and sharing

Data access and sharing will comply with the guidelines of the NSF OCE Data and Sample Policy. Data from this study will be made available through the Biological and Chemical Oceanography Data Management Office [BCO-DMO] for broad dissemination immediately following publication (http://www.bco-dmo.org). Utilization of this service ensures that the raw data sets will be available in useful formats in perpetuity. Databases that include raw and annotated sequences (as described above) will be available for download from the PI lab's webpages and further promotion of the databases will occur though sites with open access coral genomic and transcriptomic data, such as the Coral List Serve (http://coral.aoml.noaa.gov/mailman/listinfo/coral-list/), Matz lab data (https://matzlab.weebly.com/data--code.html), the Coral Microbiome Portal (https://vamps2.mbl.edu/portals/CMP) and the Coral Trait Database (https://opentraits.org/datasets/coral-traits.html). 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 (through the labs various Github pages).

For data that are not immediately published, data will be embargoed through BCO-DMO for no more than two years after collection, processing, and generation of data. Resources at the PIs' universities will be used to ensure redundant long-term retention of the complete datasets. There will be no charge for the data. Intellectual Property ownership will be determined in accordance with Title 35 of the United States Code for inventions and Tile 17 of the United States Code for works of authorship. Authorship will be determined through discussions among the PI and Co-Pis. IP disputes will first be discussed with the PI and Co-PIs and escalated to the Sponsored Program Offices if disputes are unable to be resolved.

User groups

We anticipate that the data generated from this project will be most useful in published formats, with broad applicability to research scientists and marine resource managers interested in the effects of disease on coral reef communities. However, we recognize that some individuals may be interested in the raw data files for their own meta-analyses and/or comparative studies. The aforementioned data files will allow those analyses.