Data Management Plan

1. Data Policy Compliance

Data generated from this project will comply with the NSF Division of Ocean Sciences Data and Sample Policy as directed in document NSF 17-037. In addition, we will adhere to Bigelow Laboratory's institutional data management policy. We will coordinate data sharing and archiving with the Biological and Chemical Oceanography Data Management Office (BCO-DMO). *We will comply with the Policy requirement that all project data be archived at BCO-DMO within 2 years of collection.*

2. Overview

We propose a 3-year project involving field observations, sample collections, laboratory analyses, and statistical models that will reveal niche partitioning among herbivores on coral reefs and how patterns of niche partitioning shift with changing reef conditions. Four expeditions will take place to the Dominican Republic over the first two years (two per year). Field observations will include surveys of fishes, algae, and corals found in the environment as well as timed observations of fish behavior. Sample collection will include herbivorous fishes for gut content and stable isotope analysis and algal samples for augmenting DNA barcode repositories and developing isotope mixing models. Lab-generated data will include raw DNA metabarcode sequence data and DNA barcodes from algal species found in the intestinal tract of each herbivore species, as well as isotopic profiles of herbivorous fish tissues and isotopic end-members (algae) in the environment. A summary of these data can be found on **Table 1**.

Project metadata will be provided to BCO-DMO when the project commences, and all project data will be archived with BCO-DMO within 2 years of being generated. Methods of data collection will be included with associated metadata when deposited at BCO-DMO. In addition, raw Illumina NovaSeq FASTA files, final amplicon sequence variant tables, and DNA barcodes will be archived in public repositories (i.e., NCBI Sequence Read Archive; Barcode of Life Data Systems; see below).

Table 1. For each type of data generated (a row), we list the Personnel who will manage the curating and archiving of the data, a description of the data and data type, and the file format of the data.

Curator	Archiver	Data Description	Type	File
Rasher	Rasher	Fish density (#/200 m^2) and biomass (grams/200 m^2) at each site	Obs.	.csv
Rasher	Rasher	Abundance (% cover) of benthic taxa (corals, algae) at each site	Obs.	.jpeg/.csv
Adam	Adam	Herbivorous fish feeding rates (bites/min) at each site	Obs.	.csv
Adam	Adam	Herbivorous fish space use (inferred from GPS) at each site	Obs.	.csv
Leray	Leray	Herbivorous fish alimentary tract morphometrics	Obs.	.jpeg/.csv
Leray	Leray	Relative abundance (pt. count) of algae in herbivore gut contents	Obs.	.jpeg/.csv
Postdoc	Casey	Raw DNA sequences/barcodes of herbivore gut contents, environmental samples, and mock communities	Obs.	.fastq
Postdoc	Casey	Curated DNA sequences/barcodes of herbivore gut contents, environmental samples, and mock communities	Obs.	.sra
Casey	Casey	End-member isotopic profiles of algae at each site	Obs.	.csv
Casey	Casey	Herbivorous fish isotopic profiles at each site	Obs.	.csv
Leray	Leray	Algal biomass consumed by fishes (grams/hr), relative to controls	Exp.	.csv
Postdoc	Casey	Raw DNA sequences of herbivore gut contents post-trail	Exp.	.fastq
Postdoc	Casey	Curated DNA sequences of herbivore gut contents post-trial	Exp.	.sra

3. Data Handling

As field observations are collected, they will be recorded on underwater paper or in field notebooks and then entered in Microsoft Excel spreadsheets (.csv format) at the end of each day. Data contained in Microsoft Excel files, as well as digital images from lab or fieldwork, will be backed-up on external hard disk drives while in the field (and on the cloud when internet is locally available). Prior to return travel, digital files and photos will be consolidated and distributed among the team (as a back-up measure while traveling). Upon return from the field, digitized data and images will be stored in a shared project Google Drive folder, as well as on redundant servers (section 4). Additionally, physical data sheets and miscellaneous field notes will be scanned (in .pdf format), shared, and stored after each expedition.

DNA extracts of herbivore gut contents – generated immediately in the field upon dissection – will be transported to the University of Texas; there, an aliquot will be sent to the University of New Hampshire for sequencing. These laboratory-generated data (raw, unprocessed Illumina sequence data in .fastq format) will be shared and backed-up as described above. At regular intervals during the project and no later than 2 years after collection, the raw .fastq data as well as SRA archive formatted (.sra) data will be deposited at BCO-DMO, NCBI's Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra), the European Nucleotide Archive (http://www.ebi.ac.uk/ena), the DNA Data Bank of Japan (http://www.ddbj.nig.ac.jp/), and the Barcode of Life Data System archive (http://www.boldsystems.org). All data deposits will contain sample metadata corresponding to the MIMarks Data Packages. Specifically, .sra files will be generated for each individual sample/primer set and deposited under a common project ID containing the following metadata for each sample: fish traits [growth phase, weight (g), total length (cm), total gut length (cm)], latitude and longitude of collection site, unique barcode sequence, forward and reverse primer sequences, date, and any other information pertinent to collection. Raw and archived sequence data will also be backed-up on the data warehouse component of Bigelow's High-Performance Compute Cluster. The remaining DNA extracts will be preserved and archived at -80 degrees Celsius for the life of the project and 5 years thereafter, in the event that we or someone else wishes to revisit the samples that underlie the data. Compound-specific stable isotope samples from fish white muscle, algae, and sediment will be taken in the field and immediately frozen. Samples will be kept on ice and brought back to the University of Texas, where they will be freeze dried, lipid extracted, homogenized, and analyzed. All data will be generated as .csv files and stored on Google Drive to ensure accessibility among all project researchers.

With this approach to handling field and laboratory data, data back-ups will be redundant, all project researchers will have access to all the data, and if there is a premature departure of any personnel, the data remains with the PI for archiving and publication. Such an approach ensures that the data will be archived at BCO-DMO within 2 years of collection.

Outputs from all project analyses will be managed through figshare (https://figshare.com/), which will provide links to the underlying raw data stored at the national repositories including BCO-DMO. Within figshare, we can upload data and analytical outputs, share results with collaborators, and make results available to the broader community. A free figshare account offers 20Gb of private space and unlimited space for open-access data. Figshare can also generate a Digital Object Identifier (DOI) for public data so that our open-access data can be found and cited. We will make all code (pipelines, workflows, and algorithms) used to process and analyze our data available through figshare and as supplemental information in published manuscripts. For all primer sets (16S rRNA, rbcL, tufA, and 23S rRNA), we will generate single data packages using *dada2* and *phyloseq* in R and provide the workflow in R markdown (http://rmarkdown.rstudio.com/). The data package and markdown file can be downloaded and run through R, which will facilitate analysis and discussion among the members of this collaborative research project. This will make our work reproducible and accessible to anyone.

4. Data Back-up at Bigelow Laboratory for Ocean Sciences

Bigelow Laboratory has a robust, multi-tiered data backup and retention strategy. Local data is stored on a scalable enterprise NetApp system that utilizes redundant power from uninterruptable power supplies with generator backup, dual high availability (HA) controllers, and redundant networking to ensure uptime. RAID DP allows for two disk failures per raid group without losing data, providing time to replace the disks. Hourly read-only snapshots provide the ability to rapidly recover from a ransomware attack or accidentally deleted files. A second similar system is co-located at an off-site data center for full redundancy in case of a physical disaster affecting the local system. Volumes are mirrored to the backup location daily. Lastly, volumes that are deemed to be ready for permanent archive are encrypted and stored on Amazon Glacier. Data submitted to BCO-DMO are maintained in perpetuity.