Data Management Plan

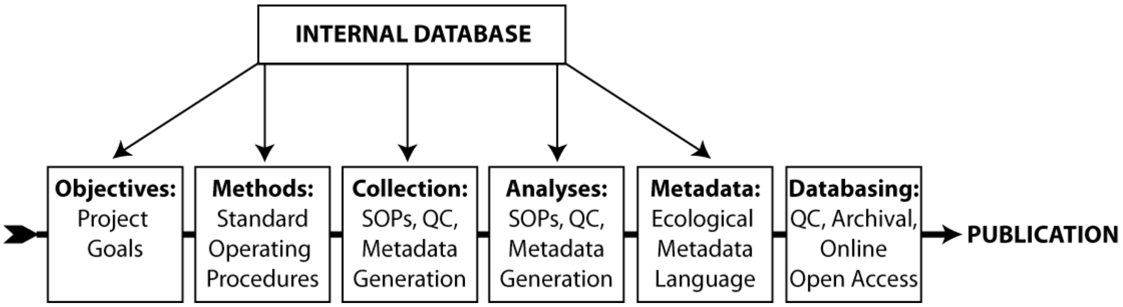
INTERNAL DATABASE
Objectives: Project Goals
Methods: Standard Operating Procedures
Collection: SOPs, QC, Metadata Generation
Analyses: SOPs, QC, Metadata Generation
Metadata: Ecological Metadata Language
Databasing: QC, Archival, Online Open Access
PUBLICATION

Figure 1. *Data Management represents an iterative process that ensures the objectives of the project proceed through a series of steps leading from project goals to publication and archival in an organized, logical, rigorous, and well-documented fashion. This ensures high data quality, experimental repeatability, reproducible analyses, and long-term, organized data archival, and open source accessibility.*

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

- Illumina amplicon, methylome, and transcriptome 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, pH, total alkalinity, water flow, and physiological measures and metadata) stored as jpg, csv, and R scripts.

***Sample and Active Data Storage*:** Genomic DNA and RNA (cDNA) samples will be retained and archived in the Putnam and Bhattacharya laboratories, stored at -80° C for long-term storage, when not fully consumed in analyses. Upon acquisition from the sequencing instruments and facilities data will be quality controlled and added to the project server (Rutgers >50 TB server), will also be submitted to the appropriate NCBI repository, and will be platformed on the project website for secure storage with backup. All data on the project server (housed at Rutgers) 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 two desktop personal computers and/or local servers, whereas an additional copy will be used to share research data with the global scholarly community and will be used for public access. Data archiving services are available through the Rutgers University RUresearch portal of RUcore, the Libraries' Institutional Repository. RUcore offers data preservation designed to make research data broadly and permanently available, and meet or exceed all federal requirements for data sharing and protection capabilities for an indefinite project term. RUcore is managed by data management and information technology specialists, and ensures the preservation of data via enterprise-level security policies and nightly, weekly and offsite (tape) backup support. Where practical, common data file formats will be migrated to new standards as they evolve to ensure availability to a wide variety of research platforms for the foreseeable future. All institutional partners will make their data findings available by the end of the project term. These data will be ingested into RUcore. In the event that some data generated with partners is considered "Limited Rights Data" that may not be made publicly shareable, the RUcore repository can provide a dark archive that will preserve the data results while limited access to a defined set of users, as required. Finally, Putnam and Bhattacharya will annually review the practices with team members to ensure compliance with this Data Management Plan, and will work closely with RUresearch to properly preserve the data in the RUcore repository.

Data Archival: Data will be archived permanently in the original data format and also in more common, non-proprietary formats (e.g., tiff, csv, txt, fasta, etc...) to facilitate future data usage. Data generated by the research and related metadata will be deposited in and be accessible through NCBI (raw data via the

NCBI Sequence Read Archive [SRA], assembled transcriptome contigs via the NCBI Transcriptome Shotgun Assembly archive [TSA]), as well as to the iMicrobe project and through a link on the project web site hosted at Rutgers. Physical and physiological data will be archived at BCO-DMO. The project website will centralize all resources and will also provide access to all downstream analyses such as gene annotations (e.g., Blast2GO, NCBI) and output of network modeling.

***Documentation and Metadata*:** Documentation for this project will include the formation of written methodologies for sample collection and processing, made openly available on the project website (designed by a project-supported web developer), 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, 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.

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. 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 NSF as the original funding source.
- Users have the final responsibility for any errors in their external and secondary analyses, while the PI and project participants will conduct quality control on the primary data and ensure accuracy of the primary data to the best of their abilities.
- The PI and project participants will not release any private or confidential information to the public, and in-house databases will be password protected.
- The PI and project participants 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. All significant findings from the proposed research will be promptly prepared and submitted for publication with authorship that accurately reflects the contributions of those involved. Data and products from this project will be used in courses at URI and Rutgers and course syllabi will be posted on the PIs websites as well as on the project website. The URL for the project website, NCBI BioProject ID, and BCO-DMO doi for each sample / study will be provided in all publications generated by the proposed work.

Participant Roles: The PIs are responsible for supervising all data management in cooperation with the project participants. All participants are responsible for data collection, quality control, internal database management/curation, and data publication as applicable to their research responsibilities within the project. The graduate and undergraduate students will be trained in and involved in data collection and quality control across the process from collection to publication (Figure 1).