

Data Management Plan

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

- DNA methylome and RNA transcriptome data stored as raw data (fastq, long-term storage), as processed quality controlled data, and as assembled contigs for analysis.
- Physical and biological data (e.g., temperature 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 lab stored at -80° C for long-term storage, when not fully consumed in analyses. Homogenized tissues or intact fragments will be archived in the Barott lab stored at -20 or -80° C for long-term storage, when not fully consumed in analyses. Upon acquisition, data will be quality controlled and added to the project server (URI compute and storage server) and github repositories. Sequence data will also be submitted to the appropriate NCBI repository for secure storage with backup. All data on the project server (URI Putnam compute node) will be accessible by the PIs and personnel. 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 (NCBI SRA and Open Science Framework (osf.io)).

Data Archival: Data will be archived in the original data format and also in more common, non-proprietary formats (e.g., tiff, csv, txt, fasta, etc...) where possible to facilitate future data usage (e.g., osf.io). Sequence 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]). Physical and physiological data will be archived at BCO-DMO. Github repositories will be used to centralize all data and code, which will be archived on Zenodo.

Documentation and Metadata: Documentation for this project will include the formation of written methodologies for sample collection and processing, made openly available (GitHub and lab websites) and published as peer-reviewed or online repository methodologies where applicable (e.g., Molecular Ecology Resources, Protocols.io,). 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 included to organize the project database for local use and public use for increased ease of data dissemination (see “Publication and Presentation” section below). Analyses will be scripted and code will be provided openly (e.g., Github) 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 PIs 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 PIs 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 PIs and project participants will not release any private or confidential information to the public, and in-house databases will be password protected.
- The PIs 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 UPenn. The 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.