Data Management Plan

1. Types of data, samples, physical collections, software, curriculum materials, and other materials to be produced in the course of the project

This project will generate new ecological and genetic data (see table for detailed descriptions), as well as R scripts and bioinformatic pipelines.

Data type	Content of data
Benthic cover	Proportional cover of major benthic components (coral by genus, macroalgae, turf algae, etc.) from 10m x 0.5m transects
Reef rugosity	Chain-and-tape measurements of reef rugosity along 10m transects
Reef fish abundance	Visual estimates of reef fish abundance and size along 50m by 5m (mobile species) or 1m (site-attached species) transects, resolved at least to family, usually to species.
Anemone survey	Species, size (major and minor axis) and location (latitude, longitude) of anemones
Clownfish survey	Species, abundance, and size of clownfish on each anemone
Amphiprion clarkii short-read sequencing	Illumina short-read sequences by individual

2. Standards to be used for data and metadata format and content

We will follow the guidelines for data curation in Hook et al. (2010). These guidelines include 1) define the contents of data files, 2) use consistent data organization, 3) use consistent file structure and stable file formats, 4) assign descriptive file names, 5) perform basic quality assurance, 6) assign descriptive data set titles, and 7) provide documentation. Ecological datasets will be stored in ASCII comma-separated value (CSV) text files to ensure long-term usability. Raw genetic data will be stored as gzipped FASTQ files, while QA/QC'ed genotype data will be stored in CSV files. All file names will include dates or version numbers.

Documentation will include the bioinformatic scripts (a mix of shell, Python, and Perl) used for processing the short-read sequences into single nucleotide polymorphism (SNP) genotypes. R scripts used for statistical analyses will also be included with the data. In addition, we will store metadata with the data files using Ecological Markup Language.

3. Policies for access and sharing including provisions for appropriate protection of privacy, confidentiality, security, intellectual property, or other rights or requirements

All data will be stored on the Amphiprion workstation in the Pinsky lab in the Department of Ecology, Evolution, and Natural Resources at Rutgers. The server is backed up daily to an off-site location. All datasets underlying published papers will shared freely on Dryad (http://www.datadryad.org) and the U.S. National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (http://www.ncbi.nlm.nih.gov/sra). Unpublished data from the project and public presentations will be made freely available to the research community through the Rutgers University Community Repository (RUcore, http://rucore.libraries.rutgers.edu) within a year of the project completion.

4. Policies and provisions for re-use, re-distribution, and the production of derivatives

We will make datasets from published papers freely available for re-use (with appropriate attribution) on Dryad, while unpublished data and presentations will be similarly available through RUcore.

5. Plans for archiving data, samples, and other research products, and for preservation of access

Data, analysis scripts, and outreach materials will be archived and backed up on servers in the Pinsky lab for future use and for responding to requests from other researchers. Published data will be archived on Dryad with a permanent Digitial Object Identifer (DOI). Unpublished data and presentations will be archived on RUcore with permanent URLs.

References

Hook, L. A., S. K. S. Vannan, T. W. Beaty, R. B. Cook, and B. E. Wilson. 2010. Best practices for preparing environmental data sets to share and archive. Oak Ridge National Laboratory, Oak Ridge, TN.