Data Management Plan: **Hybridization: the key to threatened coral species survival or the harbinger of extinction?**

1. Products of Research

At all stages, data inventories and lab notebooks will be carefully generated and stored, and protocols for the lab, field, and analysis work will be generated and shared among all partners of the project using email, FTP servers, dropbox, and teleconferencing.
Project 1: High quality sperm samples of *Acropora palmata* are in hand for sequencing of the genomes and plan to collect sperm samples from *Acropora cervicornis* in Aug 2015. The samples stem from colonies in culture at Ken Nedimyer's coral nursery in the Florida Keys and thus are accessible for future research with lower permitting requirements. DNA samples of adults will be taken from PI Fogarty's and Baums' extensive collection of *Acroporid* corals from across the Caribbean Sea as well as from new collections for phenotypic analyses. We will focus on three hybrid zones: USVI, Curacao and Belize. Additional samples from other Caribbean sites where the hybrid is rare will supplement these samples. 20 DNA samples per parental will be sequenced at 5x with Illumina HiSeq. All 850 existing DNA samples will be screened with 16 SNP markers to select complex hybrids. Hybrids will be sequenced at 5x coverage with Illumina Hi-seq. The sequencing data will be used to determine rate and direction of introgression across the genome and across the geographic range. Hybrid skeletal samples and 3-D images will be used for phenotypic trait analyses. Fertilization assays to determine the extent of later generation hybridization will be conducted in St. Thomas, USVI.
Educational Activities: Surveys will be given to the University Bound and South Broward High School students to quantify what they have learned and how we can improve their experience in the following years. We will share these surveys with University Bound director Ms. Rohan-Rhymer.
Project 2: Tissue samples from the thermal tolerance experiment will be collected and preserved for histology and gene expression analysis. After decalcification, histoslides of these samples will be generated at NSUOC. Esther Peters will train Fogarty's students to quantify coral tissue and cell degradation and conduct zooxanthellae counts. After analysis, Esther Peters will conduct quality control checks on data. Data on the oxidative stress and lipid peroxidation will be generated by Cliff Ross. After analysis, a report summarizing the data and interpretation will be sent to PI Fogarty at NSUOC. Gene expression analysis will yield count data on differentially expressed genes and gene networks as well as associated metadata. Fogarty will manage environmental and ecological data.
Project 3: Tissue samples from the disease experiments will be preserved for histology and future gene expression analysis. Tissue samples for gene expression will be stored in the Baums' laboratory at PSU. After decalcification, histoslides of these samples will be generated at NSUOC. Esther Peters will train Fogarty's students to quantify coral tissue and cell degradation and conduct zooxanthellae counts. Fogarty will manage environmental and ecological data.

2. Data Storage and Preservation

Project 1,2&3: DNA extractions from all samples are archived at -20C in 96-well-plate format, and when available, additional tissue is in long term storage at -80C. Metadata regarding each sample is currently stored in a custom Filemaker database available on the Baums' Lab internal server and backed up regularly by the Penn State University ITS backup service. For the samples used in this study, metadata will be associated with genetic sequence data archived on publicly accessible NCBI servers and presented as supplemental information in resulting publications. The sequences themselves will be stored in raw read format along with relevant barcoding and adapter sequence information on the Baums' lab servers. Trimmed Illumina reads for all portions of the study will be submitted to the NCBI short read archive. Assembled genome contigs will be available from the Galaxy free public web server (galaxyproject.org), as will SNP calls from the whole genome sequences. Data on Galaxy will be associated with additional information such as the number of reads per allele and per individual where applicable. SNP data will also be submitted to dbSNP to allow for retrieval from cross database searches via NCBI Entrez. DNA samples from project 3 will be stored at -80C until future analysis.
Project 2: Histoslides will be stored in Fogarty Lab at NSUOC. Data generated by Esther Peters and Cliff Ross will be stored in Fogarty Lab at NSUOC and will be backed up daily using Carbonite software and archived via DRYAD. All raw Illumina sequencing reads will be archived in the NCBI Sequence read archive (SRA) (http://www.ncbi.nlm.nih.gov/Traces/sra/) and on Galaxy (see Project 1 above). Counts of differentially expressed genes will be submitted to the GEO database (http://www.ncbi.nlm.nih.gov/geo/).

Gene networks of co-expressed genes will be submitted to DRYAD and/or the PSU ScholarSphere. Project 3: Histoslides will be stored in Fogarty Lab at NSUOC. Data generated by Fogarty's students will be backed up daily using Carbonite software and archived via DRYAD.

3. Data Formats and Metadata

Project 1: Each DNA sample in the Baums' Lab database has a unique identifier which is recorded along with relevant metadata regarding collection time, date, depth, GPS coordinates, etc. This data is contained in a Filemaker database for ease of access to lab members. Data tables can be exported in other formats such as tab delimited text files as needed. DNA sequence data will be stored as both raw and quality/adapter trimmed reads in Fastq format files. Assembled genome data will be stored in Fasta format files and SNP call data will be stored in tab delimited text files in accordance with expectations for the SNPASSAY and SNPPOPUSE formats required by NCBI dbSNP. Additional details regarding SNP calls will also be available in tab delimited text format on the Galaxy web server.
Project 1,2&3: Metadata concerning phenotype and environmental data will be recorded following the recommendations of Michner et al. (1997) for ecological datasets in the ecological metadata language (EML).
Project 2: Data and metadata files will be formatted in compliance with the minimum information about any (x) sequence (MIxS) specifications as defined by the genomic standards consortium (Yilmaz et al. 2011). All gene expression data will be stored in accordance with expectations for the GEO database formats required by NCBI.

4. Data Dissemination & Policies for Data Sharing and Public Access

Project 1&2: Because of the limited nature of the DNA extractions, access to the samples used in this study will be by request to PI Baums. SNP data will be made publically available at the time of publication. This will ensure that the data has undergone all required quality control steps and readers are able to test the hypotheses presented in the papers. Genome data will be made available immediately after quality control and annotation on Galaxy.
Project 2&3: Access to skeletal and histoslides, will be provided by request to PI Fogarty. All data will be made available to the community through peer-reviewed publications or 18 months after completion of the study through free-of charge online databases (such as DRYAD data archive), whichever comes first. RNAseq data will be available through the NCBI SRA web server (http://www.ncbi.nlm.nih.gov/Traces/sra/) and via Galaxy. All data files will be made publicly available via PSU Scholarsphere website, DRYAD (http://datadryad.org/). DRYAD and ScholarSphere require users to agree to a Creative Commons license. ScholarSphere's default Creative Commons license is Attribution-Non-Commercial-No-Derivs 3.0, or CC-BY-NC-ND. With this license the data provider (the PI's) share work with others and allow them to download it, provided they attribute the providers as the creators; they must also refrain from changing the content in any way and from using it for commercial means. DRYAD uses the CCO 1.0 Universal version. All ecological data will be deposited in the Biological and Chemical Oceanograph Data Management Office (BC-DM). The providers (the PI's) dedicate the work to the public domain by waiving all of their rights to the work worldwide under copyright law, including all related and neighboring rights, to the extent allowed by law. Users can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission.
Educational activities: Other Education and Outreach materials will be disseminated to the general public and scientific community as appropriate. Facebook posts on the coral research website will be made available to the public. Social media posts will be used to educate the public on our coral research activities.

5. Roles and Responsibilities

The data management plan will be implemented by the PIs. PI Baums will be responsible for submitting the raw and assembled genome sequence data and the gene expression data to the NCBI SRA and Galaxy web servers as it is available. PI Miller will be responsible for submitting the SNP calls resulting from the genome wide SNP study to the Galaxy web server and the NCBI dbSNP. PI Fogarty will be in charge of depositing environmental, ecological, and phenotypic data.