Data Management Plan

1 Products of the research

The data will consist of organismal specimen collections (including arthropods, marine invertebrates, herpetofauna, and plants), genetic and genomic sequence data, ecological measurements, geospatial layers, and subsequent analyses. The effective transmission of data from field to laboratory and analyses will be handled through a data management pipeline involving online tools and platforms that we will build as part of our research objectives.

Physical specimens: For the sequencing of voucher specimens we expect to collect on the order of 100 specimens in each of the groups of arthropods, marine invertebrates, herpetofauna, and plants. Collection and processing of specimens and tissue samples will follow standard techniques, and be conducted with other uses of the specimens in mind. For plants, silica gel samples will be made of every collection. In vertebrates, skin swabs and blood samples will be collected whenever appropriate and possible. Co-PIs Carnaval, Robinson, Gillespie, Dawson, and collaborators Michelangeli, Lyra, Rodrigues, and Freitas have or will obtain necessary collecting permits and will ensure that all US participants obtain authorization to work and collect before traveling abroad. Any collected vouchers will be deposited in local collections, while sub-samples of tissues or DNA extractions may be temporarily transported to the Co-PIs laboratories, for sequencing. Material Transfer Agreements will be established when needed. For each project involving vertebrates, animal specimen handling will follow respective Institutional Animal Care and Use Committees.

Digitized specimen data: All specimen data will be digitized and initially stored locally in our database and curated permanently in open repositories e.g. iDigBio and GBIF.

Genetic data: We have developed next generation, Illumina sequencing-based tools for rapid, cost efficient and large scale analysis of communities. We will generate sequence information for voucher specimens of arthropods, marine invertebrates, herpetofauna, and plants. Markers used will vary depending on taxon, but will be consistent within taxa across the projects.

Abundance data: For herpetofauna, bird, and plant surveys of abundance, no physical specimens will be taken. All abundance data will be recorded as simple text files with or without physical specimens.

Environmental data: Environmental data will be compiled as part of this project, maintained in their native format (`.las`, `.tiff`, `.shp`, etc.), and linked in our database. These data will be utilized as predictor variables in our joint model of species abundance, phylo- and population genetic structure, and trait distributions.

Geographical data: Candidate sites for new data collection will be selected in Geographic Information System (GIS) using layers for geological, environmental, land use and ownership, and remotely sensed physical and biological attributes. In the field, precise geographical coordinates of all collections will be determined using GPS devices. These sampling areas will serve as geospatial scope for species occurrences and abundances.

Software: The proposed `R` packages will be fully documented according to published standards and give users access to informatics and modeling tools developed during the course of this project.

2 Standards to be used, metadata format, and content

Specimen data will be digitized to conform to the Darwin Core (DwC; Wieczorek et al. 2012) metadata standard. Metadata associated with nucleic acid sequences will conform to the MIxS standards for metagenomes and marker genes, which extends the typical INSDC format used by sequence repositories. We plan to use common standards for phylogenetic trees, e.g, Newick and NeXML.

We will deposit new trait data into TraitBase, utilizing trait ontologies. We will store trait data assembled from other resources (e.g. TRY for plant functional traits). Humboldt Core will serve as standard for reporting inventory-based abundance data. Metadata concerning the environments sampled will conform to the Environmental Ontology (EnvO) standards.

Raw collections data, abundances, and derived, analyzed datasets, with their metadata, will be available to all project co-PIs on an ongoing basis and made publicly available upon publication. We will register and archive ecological data as simple text files with Humboldt Core and Ecological Metadata Language (EML) as established by the Knowledge Network for Biocomplexity.

3 Data access, dissemination, and preservation:

All data produced during this research will be freely available to the public and released to repositories as quickly as data curation practices allow; we anticipate no sensitive or confidential data. Our informatics platform is a key means for disseminating standardized data that can be joined as needed. The platform will be accessible via an R package and Shiny app.

Voucher specimens will be archived at the Bishop Museum in Honolulu, the Essig Museum at UC Berkeley, the California Academy of Sciences, and the Smithsonian Institution. These physical specimens will be matched with genetic barcode accession numbers to facilitate future work.

Ecological data: Prior to publication, metadata documenting data collections or archives will be posted publicly within one year of collection, regardless of eventual disposition of the data themselves. All metadata will minimally contain information on citation, access, data holder contact information, methods of discovery, and data structure. There are multiple options for promoting re-use of these data, including Map Of Life, on which co-PI Guralnick is a collaborative PI.

Genetic and genomic data: All sequence data will be deposited in the NCBI Genbank, with raw sequence reads deposited in the NCBI Sequence Read Archive. Copies of raw sequence data will also be stored in NERSC.

Software: All software and other code will be publicly available through GitHub and distributed under a GNU General Public License (GPL) immediately upon completion.

4 Re-use policies and PI responsibilities

We will respect all licenses on products used in modeling and will not use data products that are not conformant to community standards and practices. We will always license data to assure broadest re-use, with aspirations to CC-0 licenses but often with CC-BY as well. The lead PI Rominger will decided issues related to data use and policy and consult especially with team members who are familiar with the social, legal, and technical challenges to assure data conform to FAIR principles.

5 Tracking and annotation

Globally unique identifiers (GUIDs) will be assigned to specimens, field observations, and associated data objects to enable reliable tracking throughout dispersal among different institutions and data domains. This allows tracking metadata associated with curated specimens and their physical and electronic derivatives. Version control will be handled by Google Sheets, Amazon RDS, and GitHub.