Data Management Plan

1 Products of the research

The data will consist of organismal specimen collections (including arthropods, plants, and microbes), genetic and genomic sequence data, ecological measurements, geospatial layers, and subsequent models, analyses, and archives of all data. The effective transmission of data from field, to laboratory, and analyses, will be handled through a data management pipeline involving online tools and portals to upload, share, and collaborate between participants.

1.1 Physical specimens

For the macroecological (quantitative community) component of the work we expect to collect microbes in excess of several million specimens; arthropods in excess of several hundred thousand specimens of 300–400 arthropod species in $\geq$ 15 orders (18 sites, with 3 replicate plots each, and 3 collection methods: timed beating, litter Berlese funnel, and soil Berlese funnel).

1.2 Digital images and morphological data/metadata

Size measurements (to the nearest mm) will be taken for quantitative community analyses, using taxon specific regressions to estimate body mass. Images will be taken of all macro-organisms to be added to the reference collection, including those that will be processed for metagenomics.

1.3 Genetic data

We have developed next generation sequencing-based tools for rapid, cost efficient and large scale analysis of arthropod communities, applying Illumina sequencing to: (1) multi-locus phylogenetic and taxonomic analyses based on a comprehensive barcode reference library for Hawaiian arthropod taxa; and (2) metabarcoding for qualitative and quantitative analyses of the species composition in mixed arthropod samples. We will generate sequence information for mixed arthropod community samples, collected across elevation, and precipitation gradients on the Hawaiian Archipelago. The samples will be roughly pre-sorted taxonomically and grouped into different body size classes to allow abundance estimates using a combination of amplicon sequencing and PCR-free approaches, we will estimate species richness, species turnover and species abundances across environmental gradients. Mitochondrial COI and nuclear ribosomal 28s or 18s rDNA amplicons will be used. If these markers provide insufficient resolution for differentiating taxa, in addition we will design novel biomarkers. The previously generated reference database and community phylogeny will assist in taxon identification. Synoptic vouchers will be retained while building the reference library.

1.4 Environmental data

Environmental data will also be collected as part of project field efforts. Much of these data collections will follow and replicate NEON site initiation protocols from the Core Terrestrial site at PuâĂŹu MakaâĂŹala Natural Area Reserve. Core environmental metrics would include (protocols at `http://data.neonscience.org/data-product-catalog`): precipitation, relative humidity, soil temperature, soil physical and chemical properties, litter chemical properties, and soil microbe biomass and abundances, community, and functional composition, and metagenome sequences. These data will be maintained and stored in spreadsheets, linked, and documented in the main project database and provided to NEON staff. These data will be utilized as predictor variables for modeling deviations from the maximum entropy theory of ecology, niche occupancy, network structure, and phylogenetic diversity. Analyses will include generalized linear models and Bayesian hierarchical modeling.

1.5 Geographical data

Candidate sites will be selected in Geographic Information System (GIS) using layers for geological, environmental, land use and ownership, and remotely sensed physical and biological attributes. Meta-data will be compiled and archived with geographic shapefiles. Within collecting locales, we will record precise geographical data. The NEON Airborne Observation Platform (AOP) will generate additional datasets following NEON protocols, such as: canopy nitrogen, leaf area index, total aboveground biomass, ecosystem structure, elevation, slope, and aspect.

1.6 Software

The proposed `R` package will be fully documented according to published standards and give users access to the hierarchical models 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) metadata standard. This standard, ratified in 2009 by Biodiversity Information Standards Taxonomic Databases Working Group (http://www.tdwg.org/), has been internationally adopted and extended to specialized areas including gene bank data and georeferencing best practices.

Meta-data associated with nucleic acid sequences will conform to the MIMS or MIMARKS standards for metagenomes and marker genes respectively. Meta-data concerning the environments sampled will confirm to the Environmental Ontology (EnvO) standards.

3 Data access, sharing, and preservation

All data produced during this research will be freely available to the public and released to NEON; we anticipate no sensitive or confidential data. Data will be embargoed for up to 1 year while undergoing quality assurance and quality control.

3.1 Specimen-level data

Specimen digitization will build on a Moore-funded database platform built by the Berkeley Science Technology group (BSCIT, `http://bscit.berkeley.edu`) for the Moorea Biocode initiative (UC Berkeley). The database uses open-source software systems developed using LAMP (Linux OS, Apache web server, MySQL database, Perl/PHP). This provides a web-based system usable on any platform without the need to install special applications, and is easily shared among institutions. The BSCIT system includes: (i) Flexible querying and browsing (web-based database queries on a collectionâĂŹs full metadata or a restricted subset); (ii) Flexible display/ delivery of results (displayed initially with selected fields, links to full records, full results downloadable; points can be displayed on a map); (iii) Real-time database record creation and correction. (iv) Automatic verification and data enhancement during upload process; (v) DiGIR/TAPIR compliance and accessibility (providing access to all specimen collection data via the DiGIR/TAPIR protocol, mapped to the Darwin Core metadata standard); (vi) Mapping software, e.g., BerkeleyMapper (`http://berkeleymapper.berkeley.edu`)—the user can zoom in, click on a point to view the database record, image (via CalPhotos, `http://calphotos.berkeley.edu`).

We will use existing geospatial tools to analyze, model, and predict species distributions based on locality data and underlying environmental variables. In a GIS context, georeferenced specimen data provided by this project will be combined with remotely sensed land-cover and climate data to test hypotheses of how biodiversity interacts with the dynamic Hawaiian landscape. Once specimens are logged in the database and georeferenced, existing tools, including the DiGIR/Tapir protocol (see `http://www.tdwg.org/activities/tapir`) and the GBIF data exchange protocol (see `http://www.gbif.org`) will be used to query aggregated data across multiple collections - to create a distribution map or to use in a GIS framework. In both cases, a core set of data is made available using a predefined exchange format, based on an established standard (Darwin Core). The delivery mechanism

used for systems that support distributed queries like DiGIR and GBIF typically involve either live distributed queries to multiple sites, or a query to a single cache that contains data from multiple sites. We propose to create a single cache for this project using Darwin Core to map fields from the existing databases of participating collections to those in the cache. We propose using this cache, rather than live distributed queries, because of latency issues that have been observed with distributed queries in previous research projects. Cache creation will involve identification of the core set of fields needed for mapping environmental change (habitat use and climate), including taxonomy, location and date, and the relevant fields for each collection. Data from many of the focal taxonomic groups are already available online and searchable across different taxa and region via GBIF. Order, family, genus, and species-level pages will be linked to specimen queries thus creating dynamic species lists.

Voucher specimens will be archived at the Bishop Museum in Honolulu, the Essig Museum at UC Berkeley, and at the Smithsonian Institution, though held at collaborating institutions prior to deposition. These physical specimens will be matched with genetic barcode accession numbers to facilitate future work.

3.2 Ecological data

Raw collections data 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 EML (Ecological Markup Language), as established by the Knowledge Network for Biocomplexity (`http://knb.ecoinformatics.org/index.jsp`) through KNB or DataONE (`http://www.dataone.org`). 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.

3.3 Genetic and genomic data

All sequence data will be deposited in the NCBI Genbank (`http://www.ncbi.nlm.nih.gov/genbank`), with raw sequence reads deposited in the NCBI Sequence Read Archive (`http://www.ncbi.nlm.nih.gov/sra`). Copies of raw sequence data will also be stored in NERSC (`https://www.nersc.gov`).

3.4 Software

All software and other code will be publically availible through `github.com` and distributed under a GPL (GNU General Public Library) license immediately upon completion.

4 Tracking and annotation

Critical to interoperability will be the use of globally unique identifiers (GUIDs), which will be assigned to specimens 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. The software and database system described above has been developed to track collected material, tissue samples, DNA extractions, and sequencing steps. The main components are the Field Information Management System (FIMS) and the Laboratory Information Management System (LIMS), both open-source and developed through the Moorea project (`http://biocode.berkeley.edu`). FIMS captures all collecting event, specimen, tissues, and photo records for each sample. LIMS tracks tissue samples through DNA extraction and sequencing. A plug-in to the Geneious software developed by Biomatters (`http://www.biomatters.com`) LIMS offers the ability to track all steps of the laboratory process and integrate with BOLD, Genbank, and the Biocode FIMS system. Existing database software will build on the BSCIT system. Collections and specimens will be assigned GUIDs to allow dynamic access and assembly of derived datasets; filters will constrain data entry to acceptable value ranges. A subset of the data will be double entered to

check for repeatability and the need for further quality control diagnostics. Remote, automated backup engines will create data backup files and record changes to these data over time.

Field metadata and photos will be uploaded to a FIMS, assigned GUIDs, and shared with partner institution databases. Voucher specimens will be distributed to 3 holding institutions (Berkeley, Cornell, Maryland), where the FIMS data will be transformed to databased catalog entries. DNA extractions will be generated onsite and sent to Berkeley, where they are registered, sequenced, tracked and recorded in their LIMS. Resulting sequence data and GUID identifiers from both FIMS and LIMS data will be uploaded to Genbank. Given the expertise of our team, we expect to be able to identify most taxa to species, with undescribed taxa being the focus of taxonomic studies. Using GUIDs, BiSciCol Tracker (`http://biscicol.blogspot.com`) we will track and distribute annotations among participating institutions.