Data Management Plan

Co-PIs R. Toonen, B. Bowen and Postdoctoral researcher K. Selkoe will be responsible for creating and maintaining our data storage and sharing system. There is no standard format for storage of genetic data. We will draw from the field of community ecology to find database structures appropriate for analysis of large arrays of species and environmental data in a multivariate framework. All data will be accessible at all times for all project team members and dataset owners using secure website repositories.

Data Storage and Sharing

Genetic data: Given the large amount of raw data and results associated with this project, careful data management is key. All genetic datasets contributed to this project will be put into a standardized format (tailored to sequence and allele based genotypes) with a standardized meta-data table outlining data owner(s), associated publications, number of samples, and details about sampling sites, etc. Two other data tables will be linked to each dataset: a table of the genetic summary statistics produced from the raw data, and a table of life history traits gathered for the species. As the input file formats for each of the analyses are generated for each species, they will also be linked to each raw dataset. The series of multi-page databases for each species will be kept as Excel files labeled with species names and uploaded to a secure website such as the new EcoData Retriever associated with the Ecological Data Wiki (http://ecologicaldata.org/) that enables sharing data among project team members. There is no need for a more tailored database storage system for our purposes.

Environmental Data: All environmental data used for the multivariate analyses must be turned into "driver" datasets and processed into a format compatible with the genetic data. This means tabulating it into arrays of values for each genetic sampling site and matrices of pairwise values for each pair of genetic sampling sites. These tables of driver datasets and their meta-data will be uploaded to our data sharing folder.

Results Storage and Sharing: We will be generating many types of new data with our suite of analysis approaches. These data will be tabulated in two ways: for each species, all the data produced from analyses will go into a single multi-sheet Excel file. Because most data generated is multivariate, it will be stored in matrix formats compatible with R, the main software program used for our statistical analyses. Ancillary data that are of tabular format (spreadsheets, CSV, etc.) will be documented in EML, Ecological Metadata Language (http://knb.ecoinformatics.org/software/eml) in preparation for later data archiving. Images such as graph theoretic networks and heat maps will be linked to each associated output dataset.

In the end of the first year of the project, when we complete our analysis of multi-species clustering of genetic connectivity, we will create a public Project Website presenting an atlas of all network images with an explanation of the results and significance for science and management, geared toward a lay audience. When we have created community-level maps of genetic, oceanographic, ecological and geographic versions of connectivity, we will add these to the website in an analogous fashion. This website will be maintained at least through the publication of all papers from the funding. We also expect to contribute our findings and images to other websites geared toward science and education outreach for the Hawaiian Islands, such as the Papahānaumokuākea Marine National Monument (http://www.papahanaumokuakea.gov/), Hawaii Institute of Marine Biology (http://www.hawaii.edu/himb/), Raising Islands Blog (http://raisingislands.blogspot.com/) and Big Ocean Managers (http://www.bigoceanmanagers.org/) websites. Before concluding maintenance of the website, all data on the website will be archived in existing data repositories, enumerated below.

Product Sharing and Data Archiving

The primary data generated as part of this proposal will be published in a timely manner in refereed journals. All software and code produced by this project will be downloadable from our Project Website as they become available, and be deposited in two searchable open access databases: The Integrated Earth Data Applications (IEDA) repository, and the National Center for Ecological Analysis and Synthesis Data Repository. Any R code will be made available to the public on a CRAN site.

Below is an example of some of our products as they would be archived in the IEDA repository.

Expected data product #1
Data type: Analytical
Responsible investigator: Toonen, Bowen & Selkoe
Product description: Graph theoretic and genetic heat maps of population connectivity across the Hawaiian Archipelago both as overall and as species subsets of interest for resource management application.
Preservation plan: Web access of images in addition to direct submission to the resource managers and relevant policy-makers
Timeline for data release: Upon completion of analysis

Expected data product #2
Data type: Model Output
Responsible investigator: Toonen, Bowen & Selkoe
Product description: Derived oceanographic dispersal distances from simulated larval dispersal in physical oceanographic simulations using the flow field generated from the HYCOM ocean circulation model.
Preservation plan: Summary output will be incorporated into publications, and made available to resource managers, but no long-term storage of the data product is intended.
Timeline for data release: Upon completion of analysis

Expected data product #3
Data type: Model Output
Responsible investigator: Toonen, Bowen & Chan
Product description: Hierarchical Approximate Bayesian Computational (HABC) model testing of the role of historical factors in determining shared patterns of population genetic structure across species.
Preservation plan: Summary output will be incorporated into publications, and made available to resource managers, but no long-term storage of the data product is intended.
Timeline for data release: Upon completion of analysis