Data management plan

We are committed to broadly disseminating new data and datasets publicly in a timely fashion (i.e. within 2 years of collection) during project analyses and manuscript preparation. We adopt University of California guidelines for data management https://dmptool.org/guidance?public_guidance_page=2, and all team members will participate in and are committed to data management, with overall responsibility maintained by Dawson, consistent with NSF requirements (e.g. <https://www.nsf.gov/pubs/2017/nsf17037/nsf17037.jsp>). We will make extensive use of the PISCO data repository <http://www.piscoweb.org/data> and ensure consistency with NSF requirements (below) . There are no ethical nor privacy issues with the proposed data and no human research subjects are included as part of this study (i.e. no IRB protocol).

Data management during project: collection & storage

All collections are made under valid research permits, in accordance with landowner, State, State Parks, and National Parks regulations. Collections are curated with unique identifiers in a fully searchable databased archive in Dawson's lab. The tissue samples are stored in liquid preservative (ethanol, RNAlater) in -20°C freezers that are supplied with emergency power back-up systems in the laboratory. Tissue samples, trackable using the unique identifier, will be shared among partner labs during the project, and with non-project labs via the database (see #7, below) and formal requests.

All field data (e.g. geolocated records, photographs of specimens, environmental datasets, etc.) are archived in original file formats and, if appropriate, exported as tab-delimited text files; the original files are never modified, with all analyses being completed on working copies. These and all other data on personal computers will be backed up daily using Apple Time Machine to an onsite external hard drive, and weekly to an offsite hard drive. In addition, we use the San Diego Super Computer's enterprise class Cloud Service <https://cloud.sdsc.edu/hp/index.php> for all original project data; subsets of data that are to be used for analyses will be duplicated and transferred to a PBworks collaborative research site which also has access management and wiki options for sharing data with the research community and the general public. We maintain all sequence data in at least three locations, including one lab computer, one personal computer, and the SDSC 'cloud'-based server. All three 'localities' are themselves backed up on a daily-to-weekly basis.

To ensure data availability for public use and potential secondary uses, there will be no restrictions on sharing, using, or re-using our data. We will …

[1] share sequence and other data with collaborators pre-publication to facilitate their own independent or collaborative efforts to fully describe project research results.

[2] upload complete NextGeneration sequence data to the Sequence Read Archive (SRA) of the National Center for Biotechnology Information (NCBI), and link to the project BCO-DMO site.

[3] upload genetic datasets used in analyses to GitHub, along with code and conditions for generating the analyzed dataset from the original SRA (see #2), and link to the project BCO-DMO site.

[4] deposit UCSC survey datasets in the PISCO data archive <http://www.piscoweb.org/data> (see also #5). The PISCO archive facilitates integration of this project's dataset with comparable datasets from other non-NSF funded research. Data will be linked to the project BCO-DMO site.

[5] deposit published SNP trees in TreeBase <http://www.treebase.org/>, with link to BCO-DMO.

[6] register the project with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) and will curate all ecological data, project metadata, and links out to datasets in appropriate

genetic archives (e.g. see #s 2, 3, & 4 above) . We have an established history of depositing data annually with BCO-DMO for other BIO-OCE funded projects and will continue this practice. When data may overlap with existing initiatives (e.g. #4), data will be thoroughly annotated to describe any replication and to enable recovery of the exact same dataset from either source. This approach is taken to facilitate cross-project integration (e.g. #4) with availability of publication-ready datasets (e.g. #3, #5).

[7] make relevant entries from Dawson's laboratory sample database publicly available on GitHub as a tab delimited text file amenable to regular expression queries, also linked to the BCO-DMO site.

[8] make graduate students' Ph.D. theses available electronically within one year of filing via the University of California Library.

[9] publish in open access formats to the maximum extent possible, using UC Merced's reduced rates negotiated between the University of California and select journals (see <http://osc.universityofcalifornia.edu/alternatives/submit_work.html> including Nature and PNAS [25% discount]).

Dawson works closely with the Library at University of California, Merced, on a variety of Open Access initiatives, and this will continue during this project, per the University of California's recommendations for data management. If services above are found to be lacking, then we will use the University of California & California Digital Library's UC3Merritt curation services <http://www.cdlib.org/services/uc3/dmp/index.html>.

PISCO Data Sharing Policy:

General Policies

1. Datasets will be uploaded to the data catalog for availability within PISCO within one year of collection.
2. Full documentation (metadata) will be developed for each dataset and will be publicly available within one year of collection.
3. All data will be available to the public via the data catalog within two years of collection, or at the time of publication of the main findings of the project, whichever comes first.

Exceptions and Use Policies

There will be instances where sensitive information cannot be shared prior to publication or where data summaries rather than raw datasets are more appropriate for sharing. PISCO has a formal internal process for reviewing such exceptions. In such instances, consistent with the Division of Ocean Sciences requirements, this project's datasets will be deposited with BCO-DMO and an embargo requested. Use of the data and how to appropriately acknowledge PISCO is specified in the Usage Rights section of the metadata for each dataset.