USING POPULATION GENETIC MODELS TO RESOLVE AND PREDICT DISPERSAL KERNELS OF MARINE LARVAE

DATA POLICY COMPLIANCE

The proposal investigators will comply with the general data management and dissemination policies described in the NSF Award and Administration Guide (PAPPG, Chapter II.D.2.i.ii), and specific policy of the Division of Ocean Sciences (NSF 17-037). Upon initiation of any award, the PIs will register their project with The Biological and Chemical Oceanography Data Management Office (BCO-DMO).

ROLES AND RESPONSIBILITIES

PIs Gaither and Crandall are founding steering committee members of the Genomic Observatories Metadatabase (GEOME) and recognize the importance of findable, accessible, interoperable and reusable data (FAIR; see Deck et al. 2017, Riginos et al. 2020, Toczydlowski et al. 2021, Crandall et al. 2023). PI Beger is similarly committed to FAIR data. All project personnel will receive training in open science practices and reproducible research during the project workshops in Years 1 (virtual) and 3 (in-person). PI Crandall will take the lead in ensuring open and reproducible data and code are archived appropriately.

DESCRIPTION OF DATA TYPES

The research proposed here will produce many observational, molecular, and modeled datasets, as well as analytical code and lesson materials. The observational data will be collected in Fiji, Vanuatu and New Caledonia during Years 1 and 2, while modeled data will be developed during all three years of the project. Data will be archived at and made accessible from three principal repositories: (1) the Biological and Chemical Data Management Office (BCO-DMO; primary archival location), (2) the Genomic Observatories Metadatabase (GEOME), and (3) the International Nucleotide Sequence Database Collaboration Sequence Read Archive (INSDC SRA).

For each data type below, we give data formats and metadata standards (where appropriate) that we will use to store each type of data, together with plans for long-term archival and open access.

Observational Samples and Data

Tissue samples: Tissues collected as part of this proposal will be stored in 95% ethanol as is standard for genetic samples. For all samples, duplicate tissues will be collected with one stored at UCF and the other at the University of the South Pacific (USP) Marine Collection. Additionally, at least one intact voucher specimen for each species from each locality will be stored at USP.

DNA extractions: DNA will be extracted from tissues. The extractions will be archived for long-term storage at -20°C at either UCF or Penn State and will remain available for future advancements in genomic technologies. Information about DNA extractions will be tracked through the Geneious (Biomatters, New Zealand) Laboratory Information Management System, which interfaces with GEOME.

Fish census data: Raw data from fish censuses at each locality will be made available in CSV text format in the BCO-DMO repository.

Ocean current observations: Data from acoustic Doppler current profilers will be made available in CSV or NetCDF format in the BCO-DMO repository.

Molecular Datasets

Illumina sequence data: All raw DNA sequence data from Illumina sequencing runs will be uploaded to the INSDC SRA, which in the U.S. is maintained by the National Center for Biotechnology Information (NCBI). NCBI is part of the United States National Library of Medicine, a branch of the National Institutes of Health. INSDC SRA is the standard repository for large batches of second and third generation sequence data.

Model Output
Oceanographic models: Model output for each seascape and year, in the form of netCDF files, will be shared in the BCO-DMO repository.

Biophysical models: Model output for all iterations over all species-seascape combinations will be shared as dispersal matrices in CSV or RData format in the BCO-DMO repository.

Reproducible Analysis Code
Graduate students and postdocs will be instructed by PI Crandall in the use of Quarto notebooks to store annotated R and Python code and bash command line inputs in such a way as to allow a relatively naïve investigator to reproduce their analyses from raw data to final figures and tables. These notebooks will be developed and served from Github, and final notebook versions will be stored in the BCO-DMO repository upon publication of associated papers.

Capacity Development Workshop Materials
Lesson plans and materials from the capacity development workshops in Years 1 and 3 will be shared on the Diversity of the Indo-Pacific Network website and archived in the BCO-DMO repository.

DATA AND METADATA FORMATS AND STANDARDS
Observational Samples and Data
Sample and sampling event metadata will be crucial to interpreting the sequence data both for project personnel and future users. These metadata include standard sampling event data (such as date, GPS coordinates, and habitat where specimen was collected) as well as specimen metadata (including taxonomic information and preservative). Metadata associated with each sample will be stored in the Genomic Observatories MetaDatabase (GEOME), developed as part of the NSF RCN Diversity of the Indo-Pacific initiative. GEOME follows the Darwin Core (Wieczorek et al. 2012) and Minimum Information for Any Sequence (MIxS; Field et al. 2008) data standards, and is listed on the Registry of Research Data Repositories. The metadata will also be shared into the Global Biodiversity Information Facility (GBIF) as occurrence records, as is customary for all GEOME metadatasets. Output from ADCPs will be in NetCDF format, and conform to National Centers for Environmental Information metadata standards.

Molecular Datasets
Sequence data in FASTQ format will be submitted to the SRA through the GEOME portal, ensuring that metadata will be permanently linked to the sequence data. Metadata specifically relating to the sequence data (e.g. about library preparation protocol) will added to each SRA BioSample record. FASTQ format includes information about the Phred quality score of each nucleotide base. Derived SNP genotype and haplotype datasets will be stored in VCF format (which also includes quality scores for SNP calling) in the BCO-DMO repository with their URL linked to each relevant record in GEOME.

Model Output
Oceanographic models: Output from hydrodynamic models will be in NetCDF format, and will conform to EPA metadata standards: ISO 19115 for geospatial data.

Biophysical models: Model output for all iterations over all species-seascape combinations will be shared as dispersal matrices in CSV or Rdata format in the BCO-DMO repository.

DATA STORAGE AND ACCESS DURING THE PROJECT
During the project, investigators will store project data on personal and laboratory computers that will be mirrored on at institutional locations. At the University of Central Florida (UCF), all data will be stored on the UCF Advanced Research Computing Center's high-performance computing system Stokes which

is backed up weekly on alternating USB hard drives that are stored remotely. At the Pennsylvania State University (Penn State), data will be mirrored on a network-attached storage device with 53 TB of RAID5-redundant storage space. Data will be shared among project members using the Open Science Framework, whose portal allows connection of a wide variety of cloud storage providers, including Microsoft OneDrive (used by PSU, UCF & Leeds). As papers are published, associated datasets will be publicly archived as described above.

MECHANISMS AND POLICIES FOR ACCESS, SHARING, RE-USE, AND RE-DISTRIBUTION

Data and code will be made available at the three databases indicated above upon publication of relevant manuscripts. We will ask permission from local villages (as well as obtaining appropriate permits from national governments) to collect the samples and data, but do not anticipate that there will be any restrictions placed on the data. In the absence of restrictions from permit-issuing authorities, data and metadata will published under a CC0 license, waiving any intellectual property rights.

Because we are proposing to create the first data-assimilated biophysical models of larval dispersal, we expect quite a bit of interest within the marine connectivity community in our model outputs. We also expect that the molecular data can be re-used for purposes such as monitoring genetic diversity of marine populations or for more synthetic macrogenetic analyses.

PLANS FOR ARCHIVING

All data products will be archived at BCO-DMO as the primary repository for all datasets, with raw genomic reads stored at INSDC SRA and sample metadata stored at GEOME. PI Crandall will work with BCO-DMO to ensure the long-term stability of these data products at the end of the project. All data will be made freely and openly available within two years of collection.