Collaborative Research: RUI: Combined spatial and temporal analyses of population connectivity during a northern range expansion

DATA MANAGEMENT PLAN

Data Products

Specific data products from this project will include: 1) an annotated transcriptome of Kellet's whelk (*Kelletia kelletii*); 2) RAD-seq data for the population at each location; 3) targeted amplicon sequencing loci from each of these two sources for genotyping thousands of individuals; and 4) the population genetic allelic data from genotyping those individuals using the targeted amplicon loci. Additionally, this project will produce a curated collection of 8 years of spatially age-structured (recruits & adults) individuals from across the historical and recently expanded species range of *K. kelletii* and archived common-garden reciprocal crosses of each as tissue samples for future work. Ancillary data such as environmental, oceanographic and life history data from field collections will also be recorded and linked through open-access data repositories as appropriate.

Data Types and Formats

Specimens and DNA extractions:

Tissues will be archived in the White lab at Cal Poly, where all DNA extractions will be performed and archived for future use. Metadata associated with each tissue sample will be stored in GeOMe (Deck *et al.* 2017), a standards compliant georeferenced database developed as part of the DIPnet RCN initiative (Toonen et al. NSF-DEB 14-57848).

Sequence Data:

Raw sequence data will be deposited in NCBI's Sequence Read Archives. Through GeOMe, raw sequence data are permanently linked to the metadata through persistent identifiers generated by EZID (California Digital Library). In compliance with university and national standards, all digital data will be backed up at local (redundant personal cloud storage, see UH, Purdue and Cal Poly Facilities: Computing & Data Storage Capacity) and institutional locations and maintained in identifiable and retrievable format for a minimum of 5 years post publication. Individual data sets (individual genotypes with full metadata, including accurate georeferencing, habitat notes, etc.) will be deposited either as data publications (Such as Figshare, DataONE, re3data, or DASH), as openly accessible files at the Dryad Digital Repository (an open access repository for ecological and evolutionary data), or open access publications that include the data as part of the peer-reviewed product. All genetic data and metadata will be made available online through GeOMe within 2 years of collection and served to both GBIF and OBIS as is customary with GeOMe submissions.

***Metadata Standards, Formats and Content*:**

Metadata are crucial to repeatable science and interpreting data but often undervalued in Data Management strategies. Metadata associated with sequencing including information on library preparation protocols and details of sequencing events will be carefully formatted according to the Genomic Standards Consortium recommendations (MIMS, MIMARKS, and MINSEQE) and long-term archiving and curation of genomic metadata will occur through GeOMe. Although not able to archive large volumes of sequence data such as this, we will collaborate with the Biological and Chemical Oceanography Data Management Office (BCO-DMO) for any datasets they wish to archive, and all datasets resulting in publications will also be archived in raw input form through DataDryad.

Data Access Approach

Short-term Data Storage and Organization: Data will be collected and added to the lab databases of each PI and mirrored on personal computers of the individuals who are working with specific datasets. Additional backups on alternating USB hard drives will be stored remotely for individual workflows. As data are generated, they will be collated and added to shared institutional data repositories that will be redundant and accessible by each PI on the project. Data will be archived in the original data format as well as non-proprietary formats in accordance with GBIF/TDWG standards (following protocols and formats used for the GBIF Integrated Publishing Toolkit; IPT).

Data Archiving:

All data associated with this project will be aggregated and maintained on both the Toonen lab personal cloud storage and the Purdue University Research Repository (PURR). At Cal Poly, data also will be backed up locally (24TB external G-RAID), and raw sequencing data will be stored in our cloud Illumina BaseSpace account. Combined with the short-term data storage, this strategy ensures redundant off-site storage of all data sets and minimizes risk of data corruption or loss.

Access To Data, Tissue Samples and DNAs: All of the raw data generated in this study will be made publicly available upon publication in a peer-reviewed journal or within 2 years of collection, whichever comes first. Non-electronic data products or samples accumulated through this project will be made publicly available at no more than the cost of reproduction or shipping & handling upon completion of the project.

Data Milestones

Data will be collected nearly continuously throughout the project, and added to the redundant data archives at each institution quarterly. QA/QC will be completed on genetic sequences as part of the analysis process and QCed data will be deposited to SRA annually with reporting throughout the project. Derived data products will be completed by the end of the project, and all data will be made publicly available following publication or within two years of collection, whichever comes first. Any data desired by BCO-DMO will be delivered (or to other appropriate data facilities) over the course of the project, and all will be made publicly available within two years of collection.

Prior Experience in Data Management

Data management will be overseen by Co-P.I. Toonen at HIMB, who has extensive experience with data management and curation, data archiving and sharing (through DIPnet).

References

Deck, J., Gaither, M.R., Ewing, R., Bird, C.E., Davies, N., Meyer, C. *et al.* (2017). The Genomic Observatories Metadatabase (GeOMe): A new repository for field and sampling event metadata associated with genetic samples. *Plos Biol*, 15.