BII: Evolving Meta Ecosystems (EVOME)

Data Management Plan
Update Feb 10, 2025

This Data Management Plan is designed to ensure permanency, standardization, and broad accessibility to all types of data and products and incorporates best practices developed by publicly-funded efforts including DataONE, the Consortium of Universities Allied for Hydrologic Science, Inc. (CUAHSI), the NSF Arctic Data Center, the NSF Long Term Ecological Research Program and the Environmental Data Initiative (EDI). We use FAIR (Findable, Accessible, Interoperable, and Reusable) and TRUST principles of data management.

1 & 2. Data to be Collected, Formats and Standards, Storage and Preservation of Data
The EvoME conceptual framework will be used as a mechanism for structuring and organizing data, developing databases, producing metadata, sharing data, and archiving data. The combination of data type and source will dictate how data are handled, processed, and archived. Data will be in commonly used formats, and may include CSV and Excel files, Shape and other files compatible with ArcGIS, genomics sequences, mp3 recordings, pdfs, and jpegs. Metadata will be highly specific, given the diversity of data types. Data will enter into a multi-institutional storage and backup process that will be secured, and password protected. Data will be preserved, archived for use by others, and published in scientific journals. This project will produce six data types:
1) Sequence data (raw reads). Immediately after they are generated, raw sequencing reads for this project will be deposited in a secure location accessible to all project collaborators. UConn provides permanent, redundant storage on its bioinformatics cluster, Xanadu, for this purpose. Upon publication, the raw sequence data (both short- and long-reads) will be deposited in the NCBI Sequence Read Archive Database (www.ncbi.nlm.nih.gov/sra). The deposited data will include an overview of the project and associated metadata. The metadata will describe the individuals and information about the sequencing experiment (library preparation and sequencing platform).
2) Assemblies, annotations, transcripts, and variants. DNA sequences assembled from high throughput sequencing technologies will be submitted to the Whole Genome Shotgun (WGS) collection of GenBank (www.ncbi.nlm.nih.gov/genbank) upon publication. These data will be submitted in FASTA format and will be associated with gene-level annotations in the form of a GenBank Genome record. In addition, we will submit final genome assemblies to the Earth Biogenome Project hub. https://goat.genomehubs.org/projects. Public GitLab will host Binary formats (BAM) to record alignments of short reads to the assemblies and text formats, such as VCF, will be used to store Single Nucleotide Polymorphism (SNP) data derived from GBS. Annotated genome assemblies will be made available in GFF3 file format (submitted to Ensembl). Variant information along with associated metadata, cannot be directly submitted to GenBank and will be shared instead with EBI in association with study metadata. EBI has a new agreement with CartograPlant to exchange information for georeferenced plant populations that will be implemented here.

3) Environmental data. We will produce a diverse set of environmental data from field observations, and experiments, and laboratory analyses that will be stored in formats appropriate for the data types. Many forms of environmental data will be initially collected via spreadsheet files (templates associated with appropriate metadata), and linked to existing ontological frameworks. Physical data will include time-stamped records of air temperature, water temperature, and water-level. We will compile georeferenced data on the abundance of key taxa, including communities of aquatic and terrestrial insects, willows, white-crowned Sparrows, and Arctic grayling and their responses to the ecosystem manipulation experiment. Other products that will be stored in a digital format include georeferenced digital imagery and images from phenology cameras. These sources will be integrated into the Evome framework (based on CartograPlant).
4) Phenotypic data. Trait values for body mass, growth, function, and phenology will be stored in a georeferenced sample-by-trait matrix and distributed as a comma separated value formatted file. We will associate the phenotypes with existing ontologies defined through existing standards. As an example, the Plant and Trait Ontology will be used for the willow and the NSF LTER standards for environmental and observation data. Term assignment to relevant biological ontologies and metadata will be completed prior to loading the traits into the database. All traits, their descriptions, and their values with units will be associated with georeferenced individuals/populations through EBI/CartograPlant.
5) Genome assembly, annotation, variant detection, and association analysis pipelines. Software packages and associated in-house scripts will be documented and made available in GitLab (http://about.gitlab.com) or Protocols.io (https://www.protocols.io/). The DOI associated with the Protocols.io entries will be used in publications reporting results associated with the genomes and population studies. Intermediate processes leading up to the final assemblies, annotations, and variant sites will also be maintained in GitLab for all project members. The final accepted assembly and annotation will represent the final WGS and Genome entry for that species in GenBank or Ensembl.
6) Bioinformatics and analysis scripts. Custom scripts developed for processing and analyzing all data will be made available through the GitLab repository described above. Separate sub-projects will be maintained under the master project to encompass different analytic goals, such as exome capture, genome annotation, and gene family analysis. To ensure that all critical steps are sufficiently documented and repeatable, the final accepted pipelines and protocols will be described at Protocols.io and associated with a DOI for publication.
3. Personnel Responsibilities, Including Contingency Plans
The PI (Deegan) will have overall responsibility for data management and will monitor compliance with the Data Management Plan. Compilation and use of these data will be the joint responsibility of all researchers who will ensure that the data collected under their supervision are made available according to the plan, continuing through the timeframe for release of all data, including unpublished data, and commit to publishing data after the grant ends. The Data Manager will coordinate data organization, data sharing among PIs and laboratories, and submission to data banks and archives, and work closely with all researchers to identify data management issues, define data curation and use responsibilities. Lead authors of each publication will be responsible for following the Data Management Plan for all associated data

and analyses. Because data is shared between PIs and laboratories, in the unlikely event that one of the PIs departs from the project, another PI would assume these responsibilities.

4. Dissemination Methods to Make Data and Metadata Available

A core principle underlying our data management and access plan is transparency. We will strive to make data organized by the project available as quickly as possible following data organization and quality assurance. Once data is ready for distribution, it will be posted to an open access platform along with sufficient metadata to describe the location and methods underlying the data. Publications will appear in journals that comply with NSF's publication access policy. Each paper will include references to all associated vouchers, datasets, results, and analysis methods and identify where those are accessible. Data and associated metadata will be deposited in a repository for each data type. Postaward all PIs commit to making sure the data is deposited in appropriate online repositories and made freely available following these procedures.

5 & 6. Policies for Archiving, Data Sharing and Public Access

Data generated from this grant will be freely distributed according to the NSF Policy on Dissemination and Sharing of Research Results. The data do not include any protected, personally identifiable information or information protected by confidentiality agreements. Data deposited in the repositories we selected are publicly and freely available for reuse. All scripts will be designated as open source and thus will be modifiable or reuseable. Duplicated museum vouchers for the individuals sequenced will be deposited at the University of Connecticut repository. To meet a 10-year minimum data preservation standard, data be archived in the NSF Arctic Data Center or other locations as appropriate for the data (e.g., Genbank).