Companion planting
Generalist and specialist repositories working together to promote agricultural data sharing and reuse

Sarah Lippincott, Head of Community Engagement, Dryad AgBioData Webinar Series July 2023
About me

- Head of Community Engagement at Dryad

- A trained librarian with an interest in advancing open and equitable scholarly communication
About Dryad

An open data publishing platform and community committed to the open availability and routine re-use of all research data

50,00+ data publications
200,000+ researchers
70,000+ institutions
1,270+ academic journals

Image CC0 https://doi.org/10.5061/dryad.xd2547dd5
Data and metadata are:

➔ Curated by our expert team to ensure suitability for publishing and to facilitate discovery and reuse

➔ Permanently stored in our CoreTrustSeal repository

➔ Published under a Creative Commons Public Domain (CC0) license

➔ Accessible via our open API
What to expect

- A brief introduction to Dryad
- A perspective on the value of generalist and specialist data publishers
- A case study of data sharing in multiple platforms
- Collaborative strategizing on enhancing connections
Discoveries do not become knowledge until the findings are shared with the scientific community, to be vetted, challenged, and expanded on.
Data sharing policies

“Whenever possible, researchers should deposit data in discipline-specific or data-type specific repositories.

Otherwise, they should use a trusted generalist repository.”

What happens when a project has both?
Generalist repositories

Accommodate heterogeneous data

Provide a home for data that might not have a place elsewhere

Facilitate broad and serendipitous discovery

Expands potential audience for data

May support custom metadata and linking to related content managed elsewhere

May be free to use (within limits)

Specialist repositories

Optimized for disciplinary needs

Enhance discovery and reuse by using discipline-specific metadata

Support niche file formats

May support embedded visualization or analysis

Serve as “community hubs”
Data sharing in multiple platforms

- **When producing heterogeneous data types or formats**
  - Are there subsets of data that belong in an accepted specialist repository due to their file type or subject matter?
  - Are there funder requirements governing where data should be deposited?

- **When there are security and confidentiality considerations**
  - If raw data is too sensitive for broad sharing, are there subsets or processed datasets that could be openly distributed?
Focus on reuse

Reproduce, replicate, repurpose, build-upon, perform meta-analysis, seed machine learning algorithms
“Too often open data are uploaded piecemeal, with no accompanying metadata or missing context on processing that happened before data deposition.

As a result, potential for their reuse in either replicated studies, or in metanalyses ... or their use in generating novel results ... is reduced.”
Make and sustain connections

Publishers, researchers, institutions, and data repositories globally are collaborating to advance open science
As a researcher, my study uses genomic data in combination with landscape, dispersal, and occupancy data, to inform [conservation unit] CU delineation in Nevada populations of the Great Basin Distinct Population Segment of the Columbia spotted frog (*Rana luteiventris*).

Forester, Brenna et al. (2022), Genomics-informed delineation of conservation units in a desert amphibian, Dryad, Dataset, [https://doi.org/10.5061/dryad.w6m905qqn](https://doi.org/10.5061/dryad.w6m905qqn)
Raw sequencing data

Filtered data
(Variant Call Format [VCF]) and metadata (text, TSV)

NIH National Library of Medicine
National Center for Biotechnology Information

BioProject
A BioProject is a collection of biological data related to a single initiative, originating from a single organization or from a consortium. A BioProject record provides users a single place to find links to the diverse data types generated for that project.

DRYAD
Not-for-profit and online since 2008. Learn more: datadryad.org
Build connections and provide context

- Indicate use of multiple repositories in DMSP
- Add persistent identifiers (PIDs) for related datasets (and other outputs)
- Describe how and why data has been divided between repositories
- Strategically apply metadata
Indicate use of multiple repositories in DMSP

Element 4: Data Preservation, Access, and Associated Timelines

A. Repository where scientific data and metadata will be archived:

Raw, demultiplexed sequencing data will be made available on the NCBI Sequence Read Archive.

Filtered VCFs and metadata will be made available via Dryad.
Add persistent identifiers (PIDs)

Genomics-informed delineation of conservation units in a desert amphibian

Forester, Brenna, Colorado State University, https://orcid.org/0000-0002-1608-1904
Murphy, Melanie, University of Wyoming
Mellison, Chad, United States Fish and Wildlife Service
Petersen, Jeffrey, Nevada Department of Wildlife
Pilliod, David, United States Geological Survey
Van Horne, Rachel, US Forest Service
Harvey, Jim, US Forest Service
Funk, W. Chris, Colorado State University, https://orcid.org/0000-0002-6466-3618
brenna.forester@colostate.edu, chris.funk@colostate.edu
Publication date: August 25, 2022
Publisher: Dryad
https://doi.org/10.5061/dryad.w6m905qqn

Data files
Download dataset
August 25, 2022
*changes not displayed to the public

Related works
Article
https://doi.org/10.1111/mec.16660
Dataset
DATA AVAILABILITY STATEMENT

Raw, demultiplexed sequencing data are available on the NCBI Sequence Read Archive under BioProject PRJNA869693: [https://www.ncbi.nlm.nih.gov/bioproject/PRJNA869693](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA869693). Filtered VCFs and metadata are available on Dryad (Forester et al., 2022): [https://doi.org/10.5061/dryad.w6m905qqn](https://doi.org/10.5061/dryad.w6m905qqn).
Strategically apply metadata

- Biological sciences
- Adaptive differentiation
- conservation genomics
- conservation units
- double digest RADseq (ddRADseq)
- evolutionarily significant units
- genetic rescue
- management units
- Rana luteiventris

Dryad

- adaptive differentiation
- conservation genomics
- evolutionarily significant units
- genetic rescue
- management units
- Rana luteiventris

Molecular Ecology

NCBI

<table>
<thead>
<tr>
<th>Organism</th>
<th>Rana luteiventris [Taxonomy ID: 58176]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Eukaryota</td>
<td>Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Amphibia; Batrachia; Anura; Neobatrachia; Ranoidea; Ranidae; Rana; Rana; Rana luteiventris</td>
</tr>
</tbody>
</table>
ASHRAE global database of thermal comfort field measurements

Parkinson, Thomas, University of Sydney, ORCiD: https://orcid.org/0000-0002-0088-8754
Tartarini, Federico, Berkeley Education Alliance for Research in Singapore

Works Referencing This Dataset


Citation

Parkinson, Thomas et al. (2022), ASHRAE global database of thermal comfort field measurements, Dryad, Dataset, https://doi.org/10.6078/D1F671
Methods

Sequencing-based miRNA expression profiles from 378 clinical samples, comprising 239 neuroendocrine neoplasm (NEN) cases and 139 site-matched non-NEN controls, were used in this study. Expression profiles were either compiled from published studies (n=149) or generated through small RNA sequencing (n=229). Prior to sequencing, total RNA was isolated from formalin-fixed paraffin-embedded (FFPE) tissue blocks or fresh-frozen (FF) tissue samples. Small RNA cDNA libraries were sequenced on HiSeq 2500 Illumina platforms using an established small RNA sequencing (Hafner et al., 2012 Methods) and sequence annotation pipeline (Brown et al., 2013 Front Genet) to generate miRNA expression profiles. Scaling our existing approach to miRNA-based NEN classification (Panarelli et al., 2019 Endocr Relat Cancer; Ren et al., 2017 Oncotarget), we constructed and cross-validated a multi-layer classifier for discriminating NEN pathological types based on selected miRNAs.

Usage Notes

Diagnostic histopathology and small RNA cDNA library preparation information for all samples are presented in Table S1 of the associated manuscript.

Funding

Academic Health Sciences Center Alternative Funding Plan Innovation Fund
Canada Foundation for Innovation John R Evans Leaders Fund
Carcinoid and Neuroendocrine Tumor Society Canada
Ontario Research Fund-Research Infrastructure
Southeastern Ontario Academic Medical Organization
Robertson Therapeutic Development*
Rockefeller University, Award: U11TR001866
Ontario Institute for Cancer Research
Robertson Therapeutic Development
Room to grow

- Support discipline-specific metadata
- Strengthen connections with discipline-specific repositories
- Support author “wayfinding”
- What else?
Thank you!

hello@datadryad.org
sarah@datadryad.org
www.datadryad.org