# An Introduction to the Dryad Repository

Elena Feinstein
Dryad Librarian and Curator
Metadata Research Center
University of North Carolina at Chapel Hill
curator@datadryad.org


Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences. Dryad enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies.

Dryad is governed by a consortium of journals that collaboratively promote data archiving and ensure the sustainability of the repository.

**http://datadryad.org**

# Brussels Declaration on STM Publishing

"Raw research data should be made freely available to all researchers. Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars"

Signatories include Elsevier, Nature Publishing Group, Springer, Oxford U Press, Wiley-Blackwell. A total of 46 publishers and 13 trade organizations.

http://www.stmassoc.org/public_affairs_brussels_declaration.php

# Defining the need

- Data no longer published within the article
- Data is not shared or only shared selectively
- Data that is not archived is eventually lost
- Specialized repositories (e.g., GenBank) cover only certain data types and do so incompletely
- Supplementary materials not the best long term option
- Funding agencies and journals mandating data archiving

TABLE III[a].

*Measurements of Twenty-eight Adult and Young Females which Perished.*

| | Total Length. | Alar Extent | Weight. | Length of Beak and Head. | Length of Humerus. | Length of Femur. | Length of Tibio-tarsus. | Width of Skull. | Length of Keel of Sternum. |
|---|---|---|---|---|---|---|---|---|---|
| 37 ♀ | 155 | 240 | 26.3 | 31.4 | .709 | .710 | 1.125 | .614 | .815 |
| 38 ♀ | 156 | 240 | 25.8 | 31.5 | .715 | .678 | 1.127 | .597 | .812 |
| 39 ♀ | 160 | 242 | 26. | 32.6 | .740 | .732 | 1.157 | .597 | .854 |
| 40 ♀ | 1521 | 2323 | 23.23 | 30.3 | .6762 | .683 | 1.048 | .590 | .780 |
| 41 ♀ | 160 | 250 | 26.5 | 31.7 | .741 | .731 | 1.187 | .615 | .886 |
| 42 ♀ | 155 | 237 | 24.2 | 31. | .727 | .723 | 1.118 | .610 | .787 |
| 43 ♀ | 157 | 245 | 26.9 | 32.2 | .766 | .751 | 1.2272 | .620 | .841 |
| 44 ♀ | 1653 | 245 | 27.7 | 33.12 | .7801 | .7573 | 1.195 | .633 | .895 |
| 45 ♀ | 1532 | 2312 | 23.9 | 30.1 | .6803 | .6623 | 1.0423 | .592 | .781 |
| 46 ♀ | 162 | 239 | 26.1 | 30.3 | .709 | .685 | 1.092 | .587 | .911 |
| 47 ♀ | 162 | 243 | 24.6 | 31.6 | .741 | .729 | 1.162 | .605 | .840 |
| 48 ♀ | 159 | 245 | 23.6 | 31.8 | .727 | .700 | 1.129 | .610 | .855 |
| 49 ♀ | 159 | 247 | 26. | 30.9 | .711 | .666 | 1.098 | .580 | .7492 |
| 50 ♀ | 155 | 243 | 25. | 30.9 | .730 | .711 | 1.127 | .598 | .839 |
| 51 ♀ | 162 | 252 | 24.8 | 31.9 | .752 | .738 | 1.180 | .615 | .875 |
| 52 ♀ | 1521 | 2301 | 22.82 | 30.4 | .682 | .664 | 1.0423 | .5511 | .7341 |
| 53 ♀ | 159 | 242 | 24.8 | 30.8 | .717 | .667 | 1.090 | .575 | .809 |
| 54 ♀ | 155 | 238 | 24.6 | 31.2 | .706 | .702 | 1.102 | .588 | .7583 |
| 55 ♀ | 163 | 249 | 30.52 | 33.41 | .767 | .7671 | 1.2073 | .6401 | .896 |
| 56 ♀ | 163 | 242 | 24.8 | 31. | .713 | .713 | 1.128 | .607 | .813 |
| 57 ♀ | 156 | 237 | 23.9 | 31.7 | .718 | .716 | 1.090 | .611 | .800 |
| 58 ♀ | 159 | 238 | 24.7 | 31.5 | .726 | .701 | 1.145 | .600 | .800 |
| 59 ♀ | 161 | 245 | 26.9 | 32.1 | .751 | .704 | 1.142 | .607 | .819 |
| 60 ♀ | 155 | 235 | 22.61 | 30.7 | .695 | .692 | 1.119 | .584 | .771 |
| 61 ♀ | 162 | 247 | 26.1 | 31.9 | .751 | .735 | 1.157 | .618 | .802 |
| 62 ♀ | 1532 | 237 | 24.8 | 30.6 | .732 | .718 | 1.172 | .594 | .802 |
| 63 ♀ | 162 | 245 | 26.2 | 32.5 | .728 | .731 | 1.102 | .614 | .832 |
| 64 ♀ | 164 | 248 | 26.1 | 32.3 | .739 | .707 | 1.159 | .592 | .823 |
| Average . . | 158 | 241 | 25.3 | 31.4 | .726 | .709 | 1.131 | .601 | .820 |
| General average for 64 birds . . . | 160 | 245 | 25.8 | 31.5 | .728 | .709 | 1.128 | .601 | .834 |

Bumpus HC (1898) The Elimination of the Unfit as Illustrated by the Introduced Sparrow, *Passer domesticus.* A Fourth Contribution to the Study of Variation. pp. 209-226 in *Biological Lectures from the Marine Biological Laboratory*, Woods Hole, Mass.

# The gap between attitude and practice

- Authors in the British Medical Journal randomly received either a *general* or *specific* request to share their data (n=29)
- Researchers receiving specific requests for data were less likely, and slower, to respond.
- Only one researcher released data. Others requested further information, clarification, or authorship.

"As soon as results of a study are published, authors have a conflict of interest, and are not well placed to judge the suitability of third-party analyses of the data."

Reidpath DD, Allotey P (2001) Data sharing in medical research: an empirical investigation. Bioethics 15, 125-134

# Sharing-on-request is not effective

- Requested data from from 141 articles in American Psychological Association journals.
- "6 months later, after … 400 emails, [sending] detailed descriptions of our study aims, approvals of our ethical committee, signed assurances not to share data with others, and even our full resumes…" only 27% of authors complied

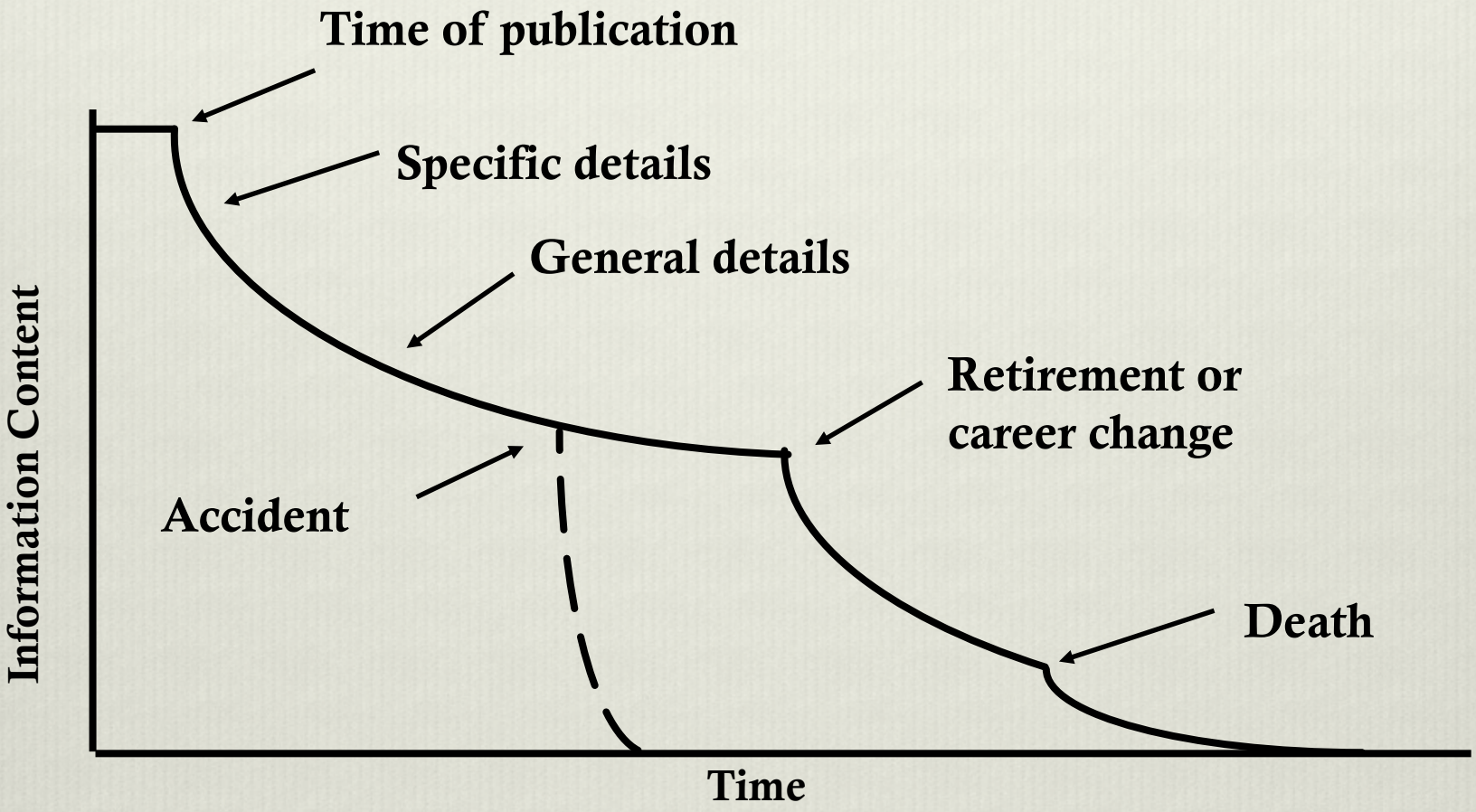
Wicherts, J.M., Borsboom, D., Kats, J., & Molenaar, D. (2006). The poor availability of psychological research data for reanalysis. American Psychologist, 61, 726-728.

# Why do authors withhold data?

- In a survey of 1240 geneticists
  - 47% had been denied at least one request for data or materials in the preceding 3 yrs
  - 28% reported that they had been unable to confirm published research because of data withholding
- The most common reasons cited for withholding:
  - Too much effort to produce the data (80%)
  - Protecting the ability of a junior colleague to publish (64%)
  - Protecting their own ability to publish (57%)

Campbell et al. (2002) Data withholding in academic genetics: Results from a national survey. JAMA 287, 473-480.

# Data entropy




Michener, W. K., J. W. Brunt, J. Helly, T. B. Kirchner, and S. G. Stafford. 1997. Non-geospatial metadata for the ecological sciences. Ecological Applications 7:330-342.

# Potential archiving solutions

**Author-managed websites**

Avoids some of the hazards of informal sharing, but is fragile.

**Specialized databases (e.g. GenBank, PDB)**

Will cover some datatypes well, some not at all; High quality data, but with greater submission burden; Diversity endangers sustainability

**Supplementary materials online**

Publisher provides basic infrastructure, but with low level of service.

**Public repositories**

Institutional or disciplinary

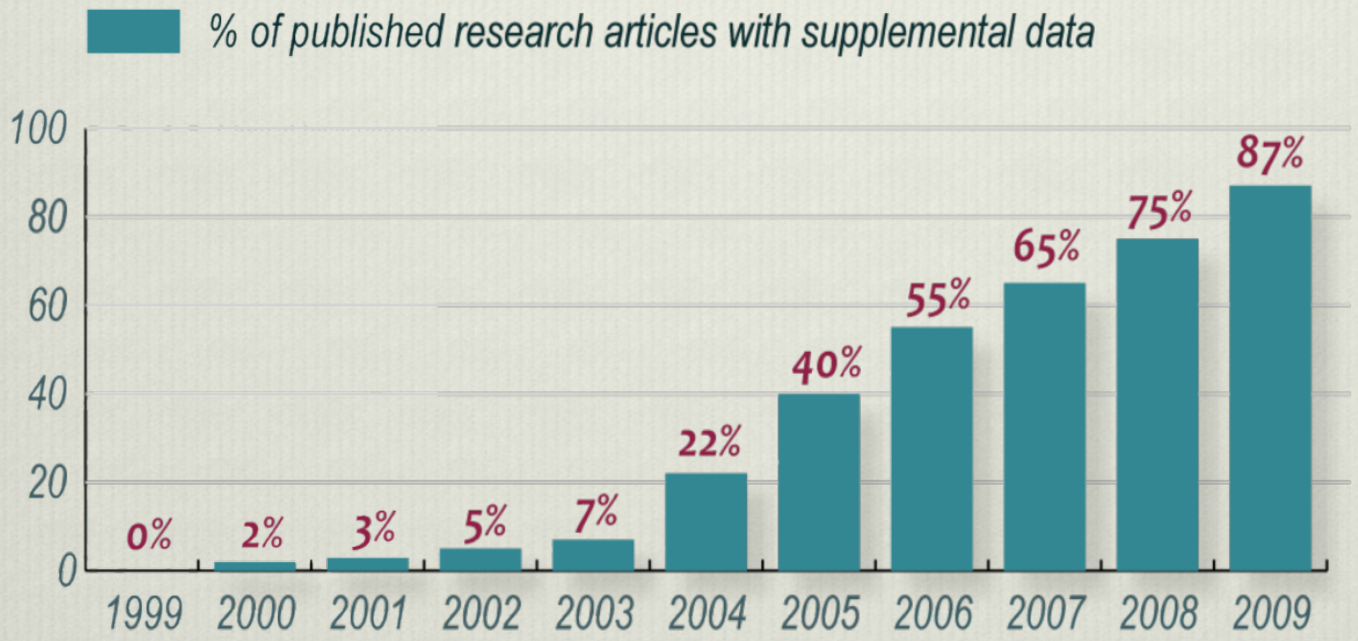
# About 85% of relevant studies submit DNA sequence data to GenBank

| Journal | No. of Studies | Data not submitted to GenBank |
|---|---|---|
| Evolution | 39 | 6 |
| MBE | 109 | 4 |
| Nature | 42 | 3 |
| PLoS Biology | 30 | 3 |
| PNAS | 30 | 1 |
| Science | 30 | 2 |

Noor MAF, Zimmerman KJ, Teeter KC (2006) Data sharing: How much doesn't get submitted to GenBank? PLoS Biol 4(7): e228.

# Rapid growth in supplemental data

JCI The Journal of Clinical Investigation




| Year | %SD |
|---|---|
| 1999 | 0 |
| 2000 | 2 |
| 2001 | 3 |
| 2002 | 5 |
| 2003 | 7 |
| 2004 | 22 |
| 2005 | 40 |
| 2006 | 55 |
| 2007 | 65 |
| 2008 | 75 |
| 2009 | 87 |

*NFAIS/NISO Supplementary Materials Meeting Summary Report*

"Beginning November 1, 2010, **The Journal of Neuroscience will no longer allow authors to include supplemental material** when they submit new manuscripts and will no longer host supplemental material on its web site for those articles. When articles are published, authors will be allowed to include a footnote with a URL that points to supplemental material on a site they support and maintain, together with a brief description of what the supplemental material includes, but that supplemental material will not be reviewed or hosted by The Journal."

The Journal of Neuroscience, August 11, 2010, 30(32):10599-10600

# NSF Data Management Plans

- A new requirement for a one page supplement to all proposals
- To include:
  - The types of data to be produced
  - The standards that would be applied for format, metadata content, etc.
  - Provisions for archiving and preservation
  - Access policies and provisions
  - Plans for eventual transition or termination of the data collection after the NSF funding period

# Joint Data Archiving Policy

Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future.

As a condition for publication, data supporting the results in the article should be deposited in an appropriate public archive.

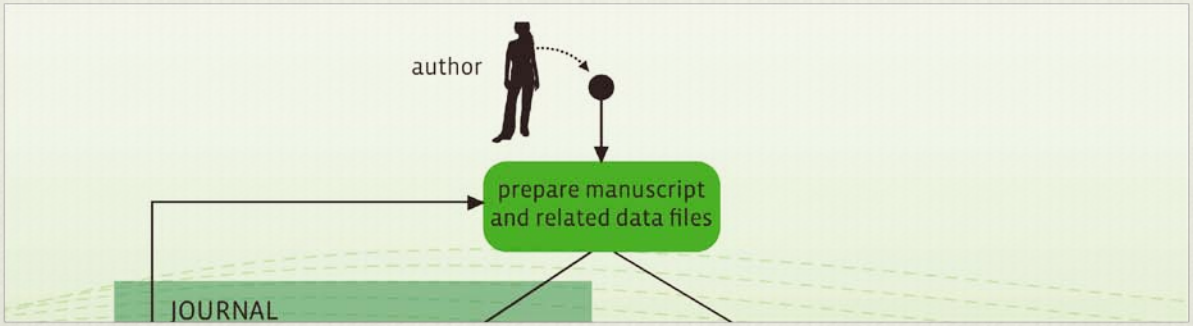
Authors may elect to embargo access to the data for a period up to a year after publication.

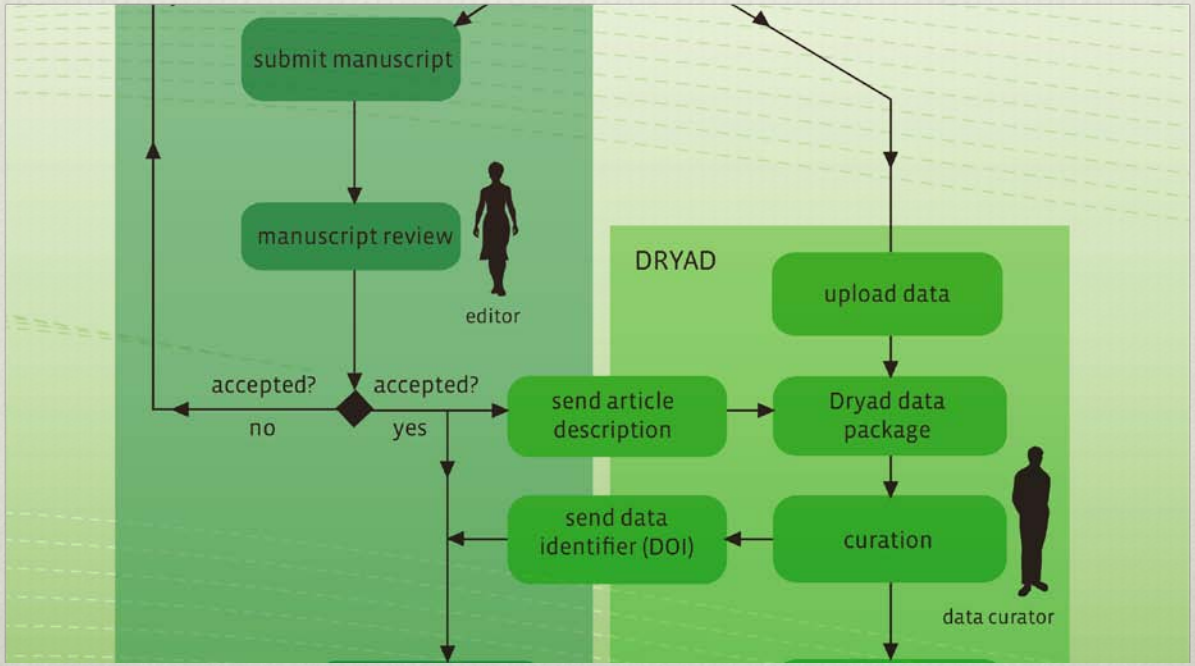
Exceptions may be granted at the discretion of the editor, especially for sensitive information.

Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data Archiving. American Naturalist. 175(2):145-146. doi:10.1086/650340


- **The End**
  - To make data archiving and reuse a standard function of scholarly communication.
- **The Means**
  - Assigning permanent identifiers (DOI) and promoting data citations
  - Publishing data access and download statistics
  - Allowing contents to be updated post-publication
  - Open terms of reuse (Creative Commons Zero), no paywalls
  - Short-term embargoes
  - Searchable across publishers & institutions, by human or machine
  - Metadata are machine harvestable, contents machine-retrievable
  - Preservation services, incl. migration of formats
  - Governed by journals (both publishers and societies)
  - Sustained by the economy of scholarly publishing

# An ingest workflow



# Example Dryad data package

## Data from: Genetic and evolutionary correlates of fine-scale recombination rate variation in Drosophila persimilis

When using this data, please cite the original article:

Stevison LS, Noor MAF (2010) Genetic and evolutionary correlates of fine-scale recombination rate variation in Drosophila persimilis. Journal of Molecular Evolution 71(5-6): 332-345. doi:10.1007/s00239-010-9388-1

Additionally, please cite the Dryad data package:

Stevison LS, Noor MAF (2010) Data from: Genetic and evolutionary correlates of fine-scale recombination rate variation in Drosophila persimilis. Dryad Digital Repository. doi:10.5061/dryad.1877

Cite | Share

| | | |
|---|---|---|
| **Dryad Package Identifier** | doi:10.5061/dryad.1877 | 102 views |
| **Individual Data Files** | SNPMarkers.csv | 38 views 9 downloads |
| | Recombinationintervals.csv | 38 views 11 downloads |
| | chromosome2alignmentannotated.gz | 43 views 10 downloads |
| | Perl scripts for analysis | 33 views 10 downloads |

**Abstract**

Recombination is fundamental to meiosis in many species and generates variation on which natural selection can act, yet fine-scale linkage maps are cumbersome to construct. We generated a fine-scale map of recombination rates across two major chromosomes in Drosophila persimilis using 181 SNP markers spanning two of five major chromosome arms. Using this map, we report significant fine-scale heterogeneity of local recombination rates. However, we also observed "recombinational neighborhoods," where adjacent intervals had similar recombination rates after excluding regions near the centromere and telomere. We further found significant positive associations of fine-scale recombination rate with repetitive element abundance and a 13-bp sequence motif known to associate with human recombination rates. We noted strong crossover interference extending 5–7 Mb from the initial crossover event. Further, we observed that fine-scale recombination rates in D. persimilis are strongly correlated with those obtained from a comparable study of its sister species, D. pseudoobscura. We documented a significant relationship between recombination rates and intron nucleotide sequence diversity within species, but no relationship between recombination rate and intron divergence between species. These results are consistent with selection models (hitchhiking and background selection) rather than mutagenic recombination models for explaining the relationship of recombination with nucleotide diversity within species. Finally, we found significant correlations between recombination rate and GC content, supporting both GC-biased gene conversion (BGC) models and selection-driven codon bias models. Overall, this genome-enabled map of fine-scale recombination rates allowed us to confirm findings of broader-scale studies and identify multiple novel features that merit further investigation.

**Scientific Names**

Drosophila persimilis
Drosophila pseudoobscura
Drosophila miranda

**Keywords**

genome evolution
molecular evolution

**Date Deposited**

2010-08-20T14:58:27Z

# Dryad submission system

Fields marked with an asterisk (*) are required. For more information on expected contents for a field, hold your mouse over the field in question.

**Publication metadata**

**Title*:** Data from: Evolution in extreme environments: replicated phenoty

**Authors*:** Last name, *e.g. Smith* | First name + initial, *e.g. Donald F.* | Add

- Tobler, Michael
- Palacios, Maura
- Chapman, Lauren
- Bierbach, David
- Plath, Martin
- Arias-Rodriguez, Lenin
- Garcia De Leon, Francisco
- Mateos, Mariana

Remove selected

**Journal name*:** Evolution

**Abstract:** We investigated replicated ecological speciation in the livebearing fishes Poecilia mexicana and P. sulphuraria (Poeciliidae), which inhabit freshwater habitats and have also colonized multiple sulfidic springs in southern Mexico. These

**DOI:**

**Journal issue:** Volume | Number | Year

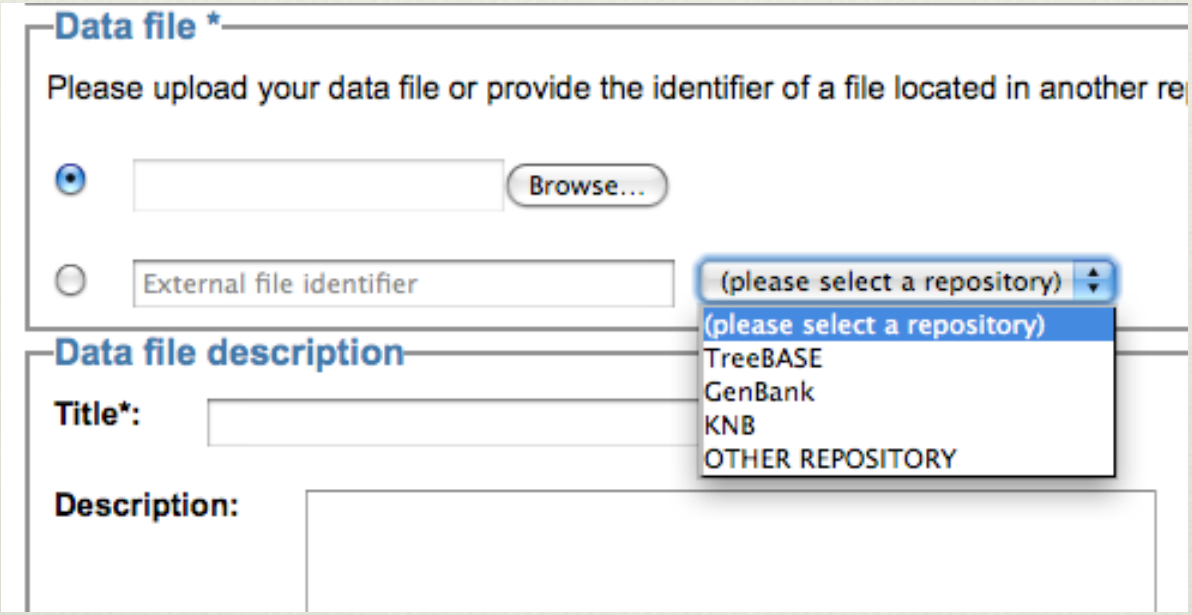
**Primary contact for data associated with this article:** Tobler, Michael

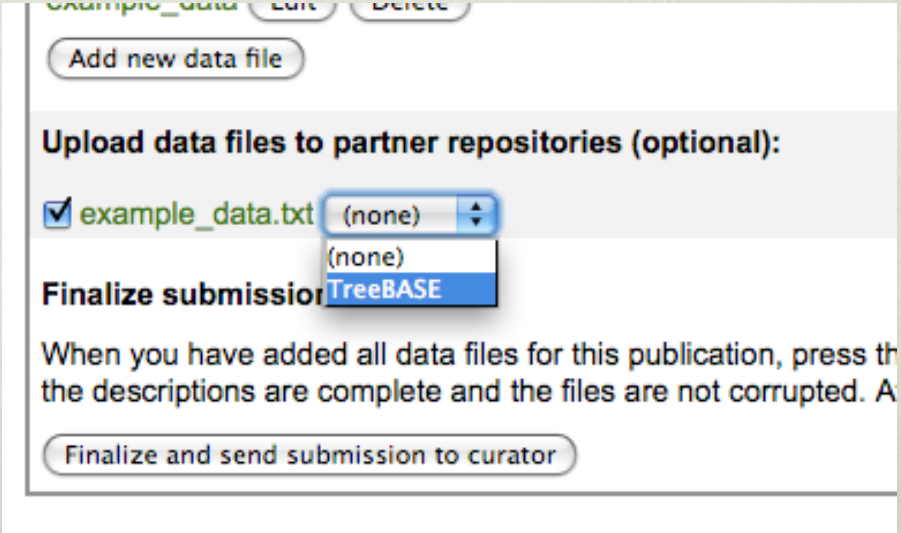
**Subject keywords:** Add

- Ecological speciation
- hydrogen sulfide
- hypoxia
- local adaptation
- morphological differentiation
- Poecilia

Remove selected

# Integrated submission with partner repositories



# Data DOI in a published article

VOL. 177, NO. 4 THE AMERICAN NATURALIST APRIL 2011

## Multiple Benefits Drive Helping Behavior in a Cooperatively Breeding Bird: An Integrated Analysis

**Sjouke A. Kingma,[1,*] Michelle L. Hall,[1,2,3] and Anne Peters[1,4]**

1. Max Planck Institute for Ornithology, Vogelwarte Radolfzell, Schlossallee 2, 78315 Radolfzell, Germany; 2. Mornington Wildlife Sanctuary, Australian Wildlife Conservancy, PMB 925, Derby, Western Australia 6728, Australia; 3. Research School of Biology, Australian National University, Canberra, Australian Capital Territory 0200, Australia; 4. School of Biological Sciences, Monash University, Clayton, Victoria 3800, Australia

*Submitted July 23, 2010; Accepted January 3, 2011; Electronically published March 10, 2011*

*Dryad data:* http://dx.doi.org/10.5061/dryad.8210.

# Another example

**Significant genetic boundaries and spatial dynamics of giant pandas occupying fragmented habitat across southwest China**

LIFENG ZHU[1], SHANNING ZHANG[2], XIAODONG GU[3], FUWEN WEI[1]

Article first published online: 21 JAN 2011

DOI: 10.1111/j.1365-294X.2011.04999.x



Issue


Molecular Ecology

Volume 20, Issue 6, pages 1122–1132, March 2011

Nature Reserve and Wawushan Nature Reserve for help during fieldwork. We thank Dr Wang Shichang for assistance with data analysis.

## Data accessibility

Jump to…

Data deposited at Dryad: doi:10.5061/dryad.8035.

## References

Jump to…

Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. *Molecular Biology and Evolution*, **16**, 37–48.

# Another example, in PLoS ONE

RESEARCH ARTICLE OPEN ACCESS

## Stalking the Fourth Domain in Metagenomic Data: Searching for, Discovering, and Interpreting Novel, Deep Branches in Marker Gene Phylogenetic Trees

Article | Metrics | Related Content | Comments: 13

**Dongying Wu[1], Martin Wu[1,4], Aaron Halpern[2,3], Douglas B. Rusch[2,3], Shibu Yooseph[2,3], Marvin Frazier[2,3], J. Craig Venter[2,3], Jonathan A. Eisen[1*]**

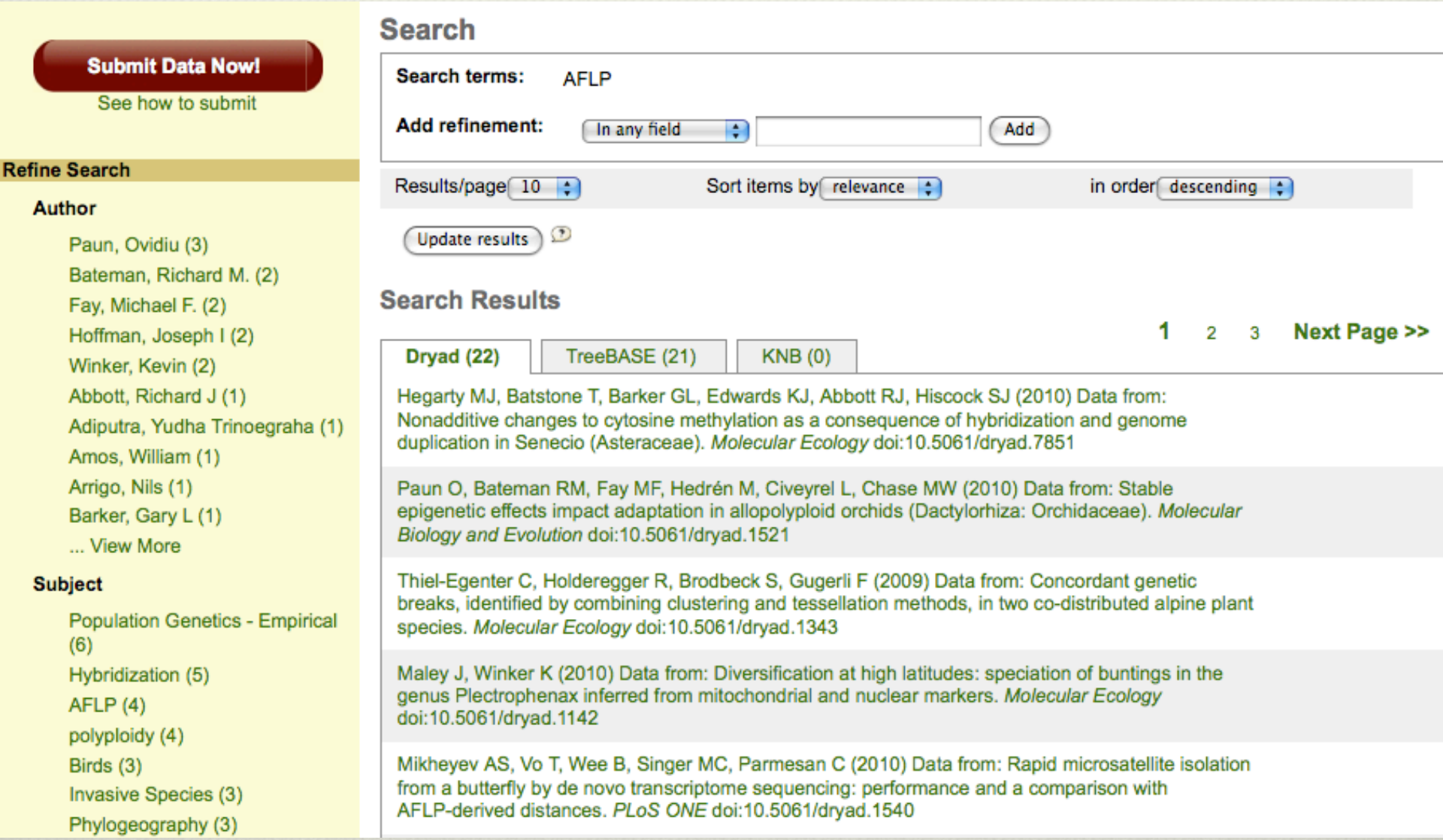
**1** Department of Evolution and Ecology, Department of Medical Microbiology and Immunology, University of California Davis Genome Center, University of California Davis, Davis, California, United States of America, **2** The J. Craig Venter Institute,



### Data and protocol availability

We've made the following data and protocols available for the public: (1) GOS and reference sequences for RecA and RpoB; (2) Subfamilies of RecA and RpoB (Table 1,3); (3) Alignments and Newick format phylogenetic trees of RecA and RpoB (Figure 1,3); (4) Sequences of the genes that share assemblies with the novel *recAs*. (Table 2); (5) GOS ss-rRNA sequence reads; (6) the Lek clustering program. The data and protocols are available at http://bobcat.genomecenter.ucdavis.edu/GOSrecA_DATA/index.html. The data have also been submitted to the Dryad repository http://datadryad.org/ - http://dx.doi.org/10.5061/dryad.8384.

# Searching and browsing



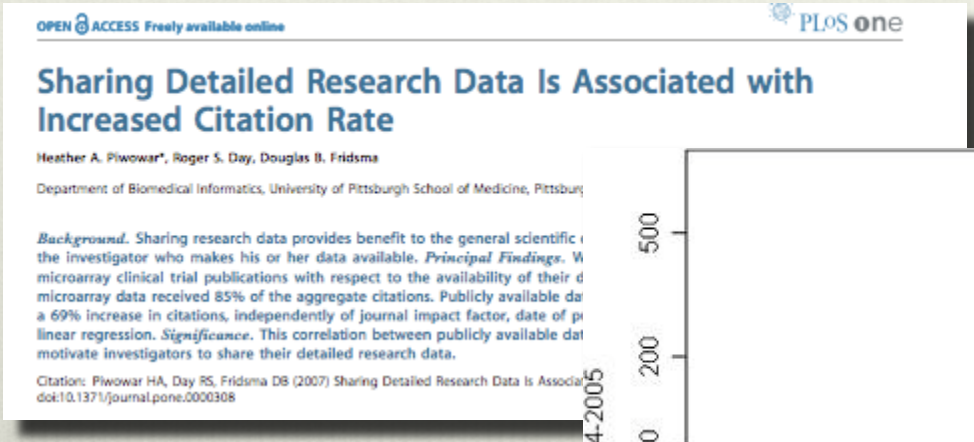
# Dryad metadata

- Emphasis on simplicity and interoperability
- Using Dublin Core plus some additions from PRISM, Darwin Core, and BIBO (planned)
- Exchanging metadata with DataCite, TreeBASE, others
- Working toward making metadata available as RDF and publishing a complete application profile using Singapore Framework guidelines

# Is Dryad meeting its goals?

- Are people using the data?
- Does it improve the efficiency of science?
- Does it improve the quality of science?
- Does it expand the capacity of science?

# Lessons from the Gene Expression Omnibus



OPEN ACCESS Freely available online PLoS one

## Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

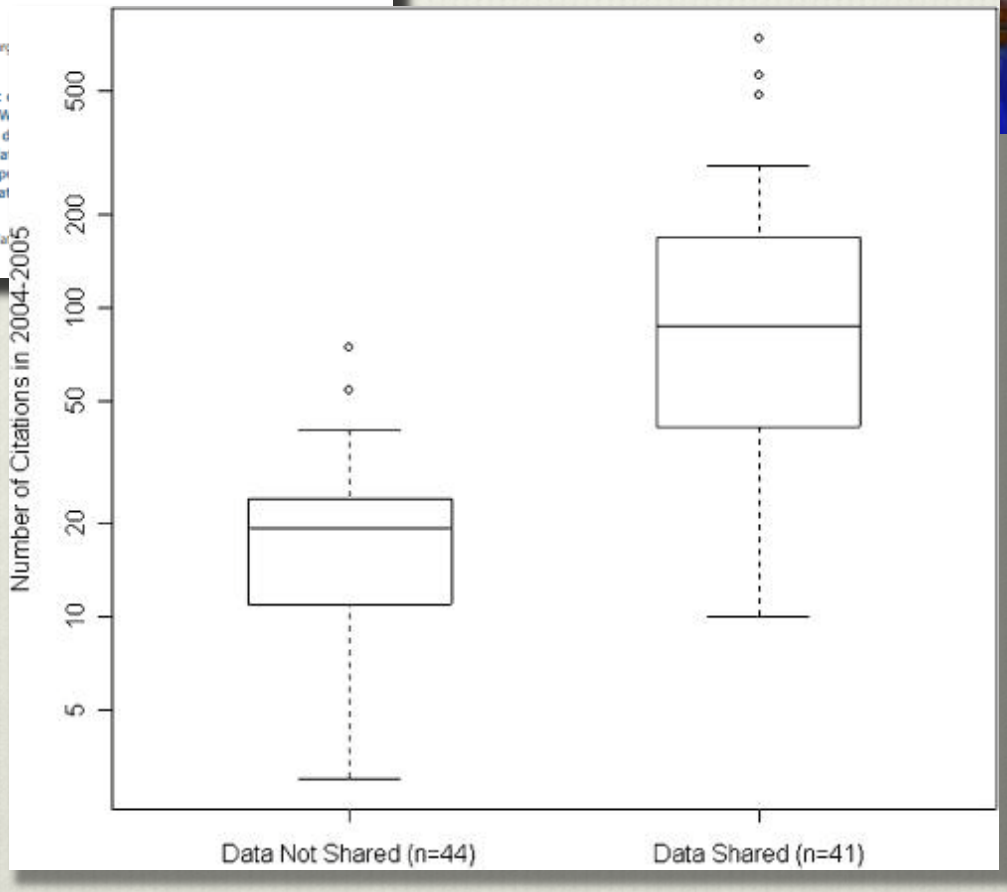
Department of Biomedical Informatics, University of Pittsburgh School of Medicine, Pittsburg

*Background.* Sharing research data provides benefit to the general scientific
the investigator who makes his or her data available. *Principal Findings.* W
microarray clinical trial publications with respect to the availability of their d
microarray data received 85% of the aggregate citations. Publicly available da
a 69% increase in citations, independently of journal impact factor, date of p
linear regression. *Significance.* This correlation between publicly available dat
motivate investigators to share their detailed research data.

Citation: Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associa
doi:10.1371/journal.pone.0000308

Heather Piwowar

Citations were 69% greater for publications that shared microarray data (right) versus those that did not (left), independent of journal impact factor, date of publication, and author country of origin.




Piwowar H, et al. (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308.

# Challenges and next steps for Dryad

- New submission workflows, collaborating with journals to fit into their publication process
- Handshaking with additional partner repositories
- Building tools for more efficient curation
- Linked data / semantic web
- Name authority control – ORCID, other researcher IDs
- Subject metadata and controlled vocabularies – HIVE, another MRC project

# Disciplinary vs. institutional repositories

- The infrastructure for data archiving is playing catch-up with the needs of scientists.
- There is a healthy competition between institutional and disciplinary repositories to meet these needs.
- The ingest bottleneck will drive the solution
  - Disciplinary repositories like Dryad can ingest the long tail of orphan published data.
  - Institutional libraries are best placed to develop the vast array of pre- and post-publication services that data-driven science will require.

# Some of the contributors to Dryad

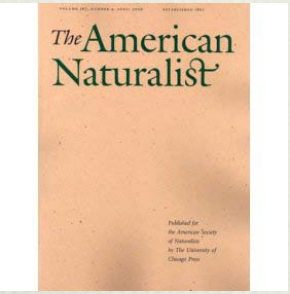
**Dryad Consortium Board, journal partners, and data authors**

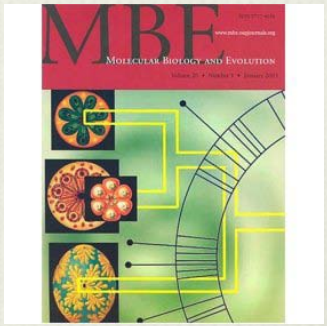
**NESCent:** Kevin Clarke, Hilmar Lapp, Heather Piwowar, Peggy Schaeffer, Ryan Scherle, Todd Vision
**UNC-CH <Metadata Research Center>:** Sarah Carrier, Elena Feinstein, Jane Greenberg, Hollie White
**Duke:** Cliff Cunningham, Mohamed Noor, Kathleen Smith, Marcy Uyenoyama
**U British Columbia:** Michael Whitlock
**NCSU Digital Libraries**: Kristin Antelman
**Yale/TreeBASE:** Youjun Guo, Bill Piel
**UNM/LTER/DataONE:** Bill Michener, Mark Servilla
**Oxford University**: David Shotton
**British Library**: Lee-Ann Coleman, Adam Farquhar

UNC
SCHOOL OF INFORMATION AND LIBRARY SCIENCE
Metadata Research Center <MRC>

NESCent

# Some Partner Journals



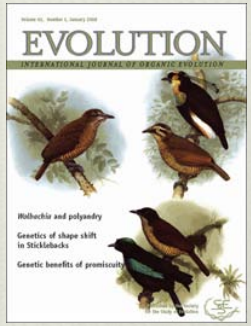




**American Society of Naturalists**

*American Naturalist*

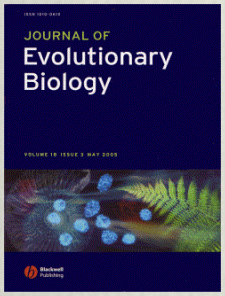

**Ecological Society of America**

*Ecology, Ecological Letters, Ecological Monographs, etc.*




**European Society for Evolutionary Biology**

*Journal of Evolutionary Biology*

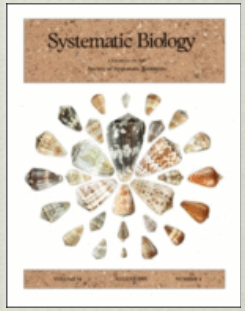



**Society for Integrative and Comparative Biology**

*Integrative and Comparative Biology*

**Society for Molecular Biology and Evolution**

*Molecular Biology and Evolution*




**Society for the Study of Evolution**

*Evolution*

**Society for Systematic Biology**

*Systematic Biology*


**Commercial journals**

*Molecular Ecology*

*Molecular Phylogenetics and Evolution*


# Initial Funding

- National Science Foundation (USA)
- Institute of Museum and Library Services (USA)
- Joint Information Systems Committee (UK)



# Dryad Technology

- DataONE member node
- DSpace repository software (open source)
- Assigning DOIs via California Digital Library
- Integration with specialized repositories and databases
  - Federated searching with TreeBASE and KNB LTER
  - TreeBASE submission (using BagIt and OAI-PMH)
  - GenBank (planned for future)


http://datadryad.org

http://blog.datadryad.org

http://datadryad.org/wiki

http://code.google.com/p/dryad

Facebook & Twitter (#datadryad)