Genome project management resources at the National Agricultural Library

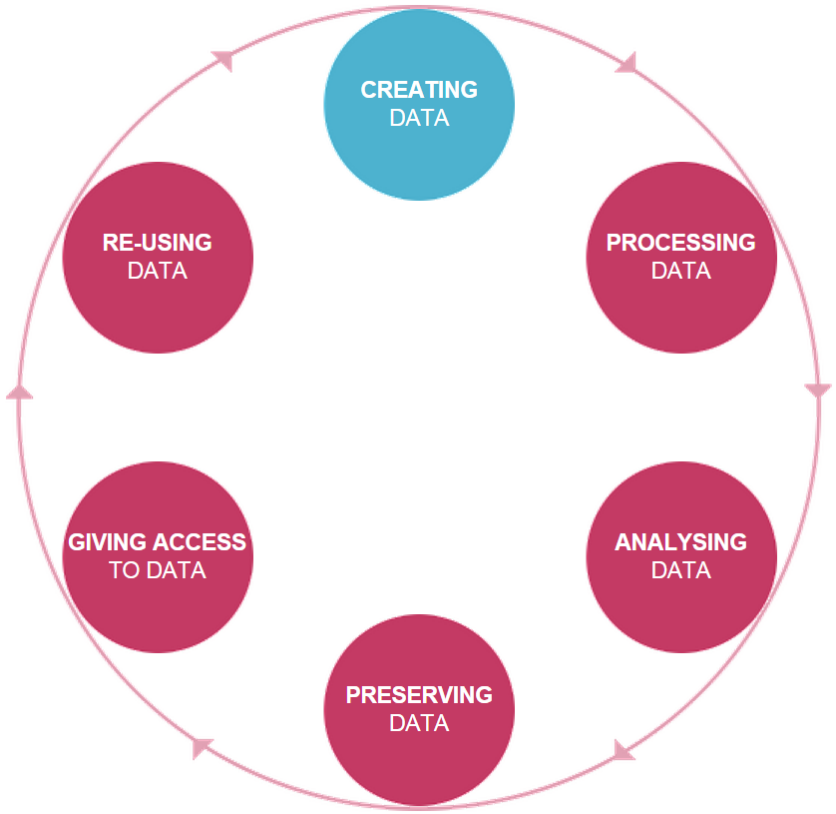
Monica Poelchau and Chris Childers
USDA-ARS, National Agricultural Library
Entomological Society of America Meeting 2017
November 8th, Denver, CO

USDA

Outline

1. Data management activities
2. Repository suggestions
    1. Genome warehouses
    2. Generic repositories
    3. ‘Boutique’ genome databases
3. The i5k Workspace@NAL

1. Data management activities

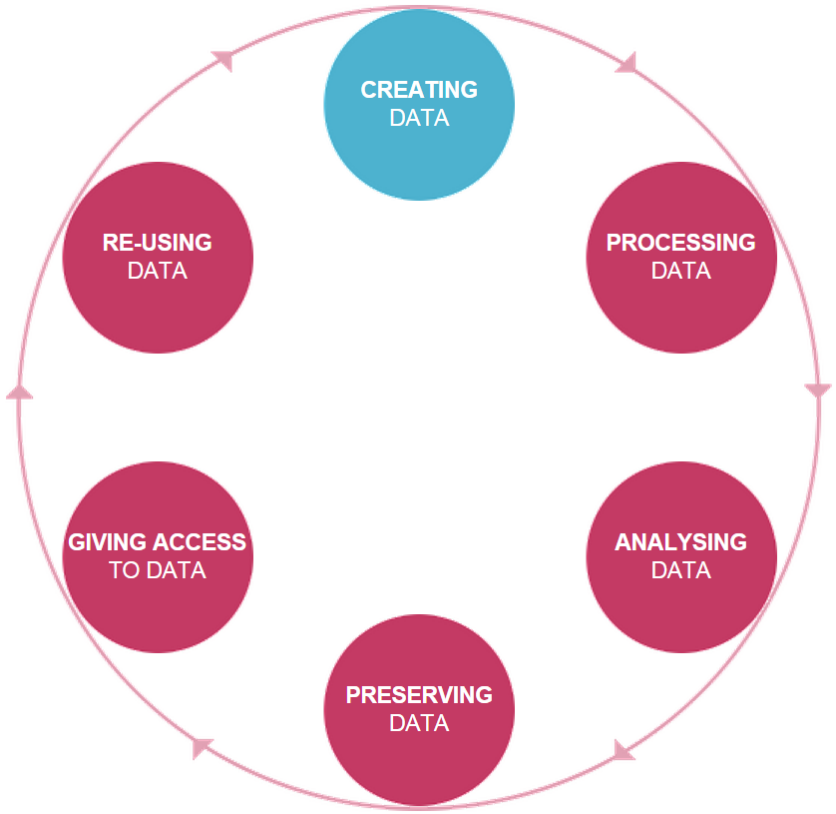


- Data has a life cycle – the data you generate can outlive the project you generated it for
- Proper data management lets others build on existing research

https://library.leeds.ac.uk/images/research-data-lifecycle-850x850.png

USDA

1. Data management activities



Proper data management activities can include:

- Organizing: Write a data management plan prior to beginning your research
- Documenting:
  - Record metadata
  - Properly record analysis methods
- *Preserving and sharing: Submit data to appropriate repositories

https://library.leeds.ac.uk/images/research-data-lifecycle-850x850.png

1. Data management activities

- Write a data management plan prior to beginning your research
  - Requirements will often depend on your funding agency
  - Some tools are available (e.g. https://dmptool.org/)
  - Your institution (e.g. library) might have someone to help out

1. Data management activities

- Record metadata
  - Metadata: Data about your data
    - Has different levels. Could be your name, your institution, the machine you sequenced your DNA on, or the accession number and name of the gene you cloned
  - 'Data without metadata is garbage' or 'Metadata is a love note to the future' (https://twitter.com/textfiles/status/119403173436850176)
    - Basically, releasing your data without context makes it unusable for others
    - Proper metadata will make that context machine-readable, so it can be better integrated by future generations to gain new scientific insights
  - For you, this generally means submitting your experimental details to a database so they can render them in a format that is machine-readable
  - Often, you will officially record the metadata when you submit your data to a repository (e.g. NCBI)
  - There are some 'helper' tools available, e.g. CyVerse: https://learning.cyverse.org/projects/sra_submission_quickstart/en/latest/

1. Data management activities

- Properly record analysis methods
  - Can use GitHub (for version control)
  - Zenodo (for long-term preservation, DOI)

1. Data management activities

- Submit data to appropriate repositories
- Advantages to submitting:
  - Greater visibility for your dataset
    - "…studies that made data available in a public repository received 9% more citations than similar studies for which the data was not made available". https://doi.org/10.7717/peerj.175
  - Value-added tools for searching and browsing, analysis
  - Curation tools to improve annotation quality
  - Help with data management
  - Increasing mandate from journals and funding bodies to make research data fully accessible post-publication[1, 2]

[1]http://www.nature.com/authors/policies/data/data-availability-statements-data-citations.pdf
[2]https://obamawhitehouse.archives.gov/the-press-office/2013/05/09/executive-order-making-open-and-machine-readable-new-default-government-

2. Repositories

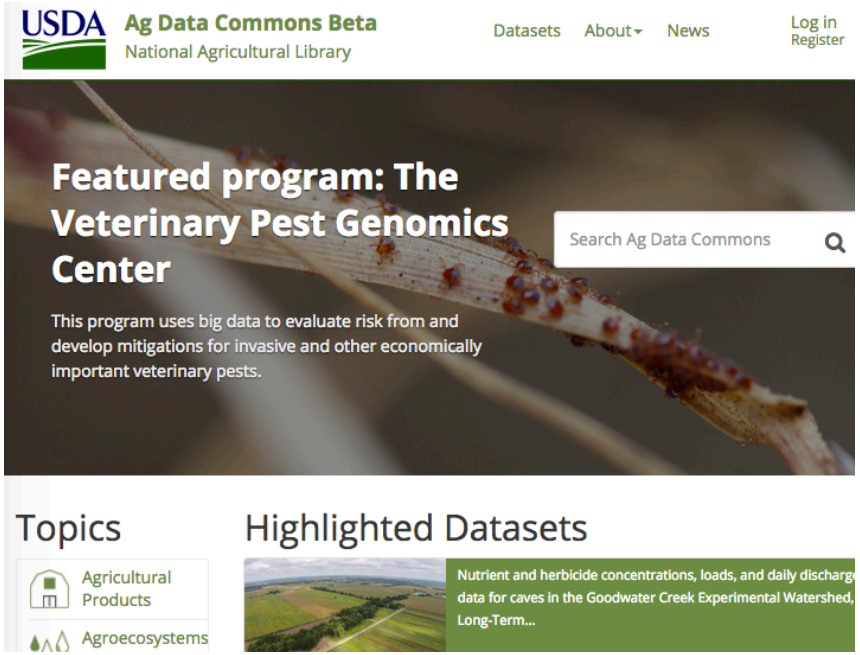
- Genome warehouses
  - E.g. NCBI, ENA, DDBJ
  - Specific to genomic data types
  - Long-term preservation/Archiving
- Generic repositories
  - E.g. Dryad, FigShare, *Ag Data Commons
  - Are good for data types that aren't usually submitted to NCBI or genome databases
- 'Boutique' genome databases (don't always guarantee long-term preservation)
- Can't find a good fit for your data? Try https://fairsharing.org/

2. Repositories – Warehouses

- In general, once you have data in a format that can be deposited to your repository of choice, do it!
  - If you delay, you will forget or displace the appropriate metadata.
  - You will have to do it anyway.
  - You can set an embargo.
- Raw reads
  - Deposit in NCBI's SRA repository as soon as possible
  - CyVerse has tools to help with SRA submission
- Assembled genome
  - Submit to NCBI's GenBank **BEFORE** performing major downstream analyses
  - Your assembly will change after submission, due to GenBank's QA/QC process
  - Downstream analyses should be performed on a dataset that is already accessioned somewhere, so it can be reproduced.

2. Generic Repositories – The Ag Data Commons

- A platform to gather agricultural data and transform it into agricultural knowledge so the farmer can translate knowledge into action.
- Hosts any dataset funded by the USDA
- Landing page
- Citable DOI
- https://data.nal.usda.gov/
- 17 i5k datasets already available
  - https://data.nal.usda.gov/i5k

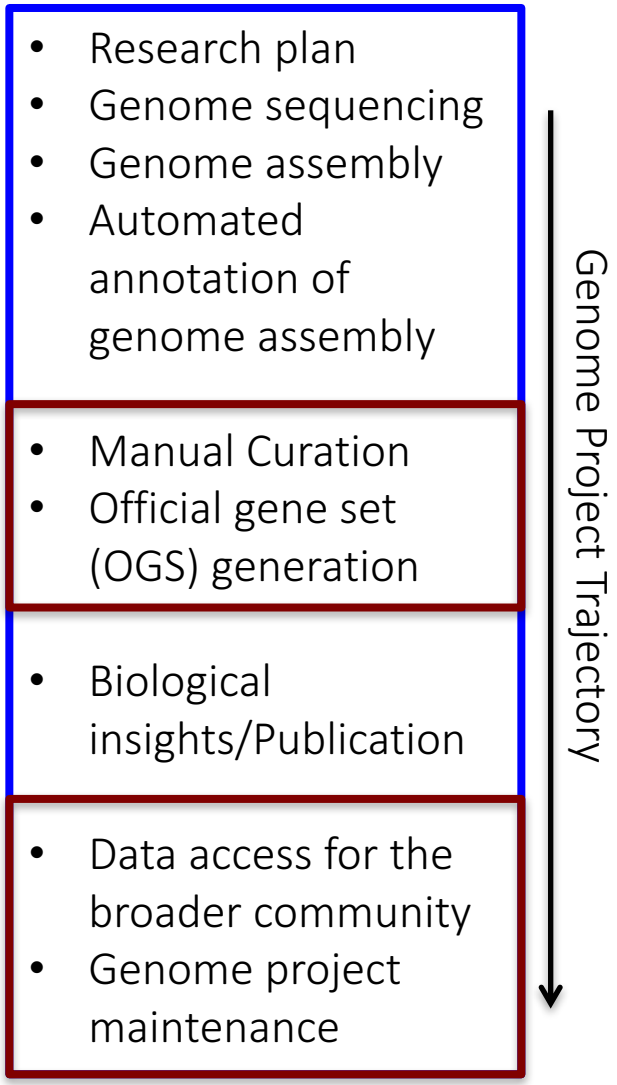


2. Repositories: ‘Boutique’ genome databases

- A non-exhaustive list: http://i5k.github.io/share
  - *Any arthropod: i5k Workspace@NAL
  - Hymenoptera: Hymenoptera Genome Database
  - Ants: Fourmidable
  - Insect vectors of disease: VectorBase
  - Aphids: AphidBase
  - Lepidoptera: LepBase
- Often provide value-added curation services and tools to make clade-specific data easier to find and use;
- ‘Warehouse’ repositories (e.g. GenBank) may not store all data types (e.g. phenotypic data).

3. The i5k Workspace@NAL

- We support any 'orphaned' arthropod genome project.
  - Connect researchers to the data
  - Create standardized tools for accessing the data in useful ways
  - Provide resources to facilitate manual curation projects
- Supported data types:
  - Genome assembly
  - Anything that you can map to or predict from the genome assembly
- Main requirements:
  - Genome assembly needs to be in GenBank/ENA/DDBJ
  - Data should be public (no private repositories)
  - Manual annotation only occurs at one genome database at a time



USDA

3. The i5k Workspace@NAL

Our background:

- Originally set up to support genomes sequenced as part of the i5k initiative
- I5k: International effort to prioritize insect genomes for sequencing; provide guidelines for genome sequencing and curation; and seek funding
- I5k Goal: coordinate the sequencing and assembly of 5000 insect or related arthropod genomes

‘Frozen’ genome assembly | Automated annotations | Ancillary datafiles (e.g. RNA-Seq alignments)

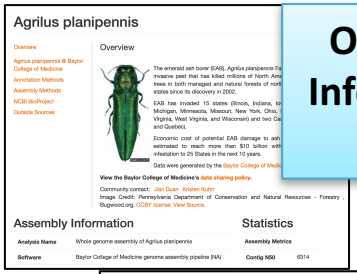
↓

Submission

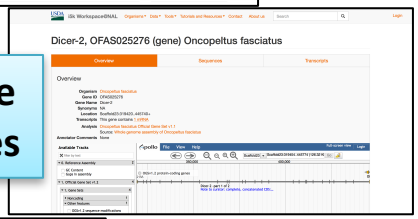
i5k Workspace@NAL

https://i5k.nal.usda.gov/

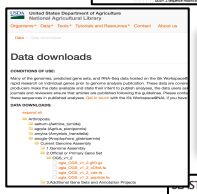
Resources



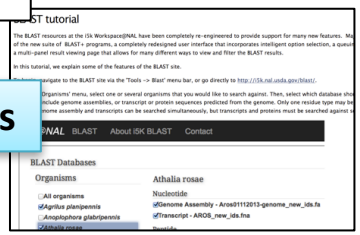
Organism Information Page



Gene pages

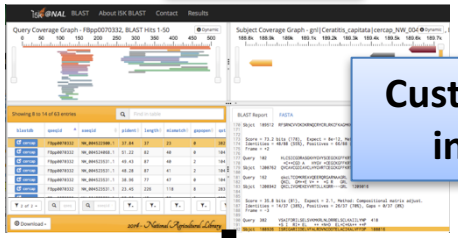


Bulk data downloads

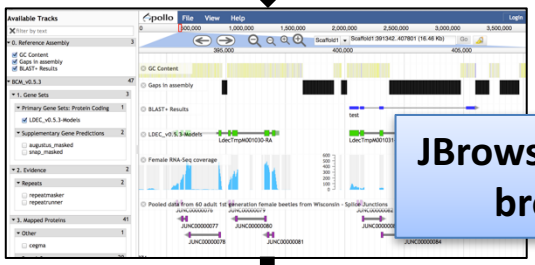


Tutorials

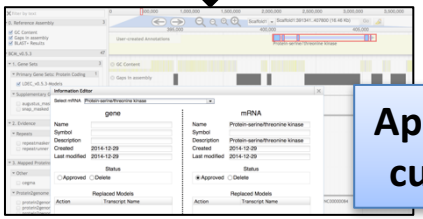
Tools



Custom BLAST interface



JBrowse genome browser



Apollo manual curation tool

HMMer

Clustal

Services

Manual annotation quality control

Official gene set generation

Challenges

Non-standard data formatting

Failure to submit all metadata (ex: sample origin; analysis methods)

3. i5k Workspace content – 59 species and counting

Order	Quantity	Order	Quantity
Amphipoda	1	Hemiptera	8
Araneae	3	Hymenoptera	15
Blattodea	1	Lepidoptera	2
Calanoida	1	Odonata	1
Coleoptera	7	Orthoptera	1
Diplura	1	Scorpiones	1
Diptera	13	Thysanoptera	1
Ephemeroptera	1	Trichoptera	1
Harpacticoida	1		

- Many other datasets mapped to, or predicted from each genome assembly (gene predictions, transcriptomes, RNA-Seq, etc.)

Need more information?

i5k Workspace@NAL:

- https://i5k.nal.usda.gov/
- https://github.com/NAL-i5K/

The i5k initiative:

- New website: http://i5k.github.io/
  Ag Data Commons:
- https://data.nal.usda.gov/

Thank you!

The NAL Team

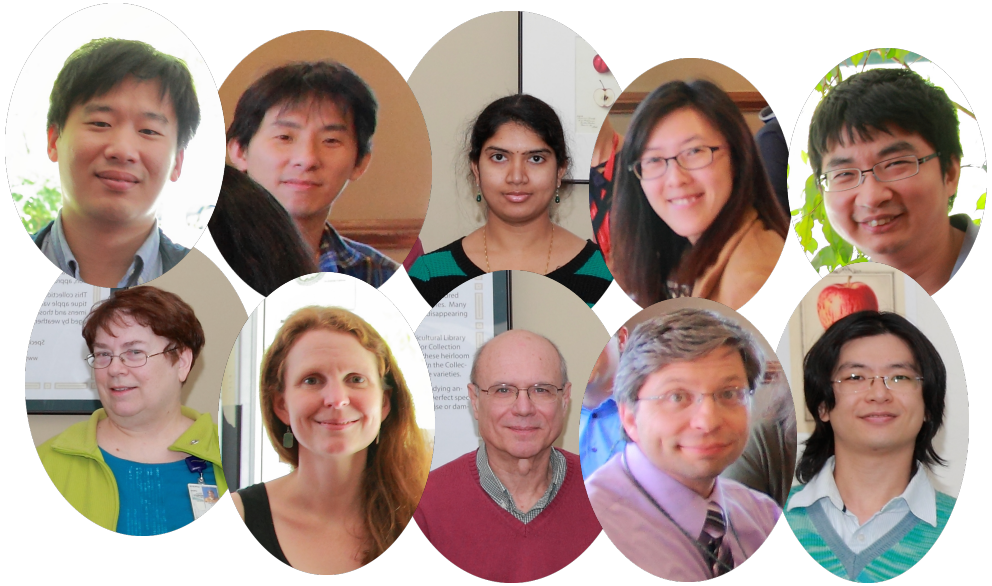
- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

I5k Workspace alumni

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

i5k Workspace@NAL advisory committee

- i5k Coordinating Committee
- i5k Pilot Project
- Apollo & JBrowse Development Teams
- GMOD/Tripal community
- All of our users and contributors!



USDA