The i5k Workspace@NAL

Chris Childers and Monica Poelchau
USDA-ARS, National Agricultural Library
Arthropod Bioinformatics Workshop 2017
University of Notre Dame

USDA

Overview

1. Background: What is the i5k Workspace?
2. Submitting data
3. Finding data at the i5k Workspace
    1. General search/Content types
    2. Data downloads
    3. BLAST
    4. Clustal(s)
    5. HMMER
    6. Jbrowse/Apollo
4. Improving data at the i5k Workspace via community annotation
    1. See Monica Munoz Torres' workshop for full use of Apollo
5. Other data management resources at the NAL – the Ag Data Commons

The i5k Workspace@NAL

Our focus:

- We support any ‘orphaned’ arthropod genome project:
  - Genome assembly needs to be in GenBank/ENA/DDBJ
  - Data should be open access (no private repositories)
- We enable and support community curation.
- We enable content search and retrieval

Submitting data to the i5k Workspace

- **All** of our data is user-submitted.
- The i5k Workspace centers data around ***projects.***
  - I5k Workspace project: A collection of data centered around the genome assembly of an arthropod
    - Genome assembly – **must be accessioned by the INSDC**
    - Gene predictions
    - Any other data that is mapped to the assembly
- Each i5k Workspace project has a ***project coordinator.***
  - Serves as the point of contact for questions about the project
  - Main responsibility: approve or reject new Apollo users

Criteria for starting an i5k Workspace project

- You need to have an **arthropod** genome assembly, **accessioned by NCBI** (or another INSDC member)
  - Using GenBank's accession numbers avoids confusion about assembly version
  - The GenBank contamination screen improves the assembly quality
  - Using a stable assembly is beneficial for the labor-intensive community annotation process
- All manual annotation efforts need to be at one database

What do we do with your data?

- We set up:
  - An organism page;
  - The JBrowse genome browser, including the Apollo manual annotation tool;
  - The BLAST+ sequence search tool;
  - HMMER, Clustal Omega and ClustalW applications;
  - Bulk data downloads.
- Unofficial help:
  - We can help with setting up an Official Gene Set for your assembly.
- Future plans:
  - Moving all Official Gene Sets into GenBank
  - Updating Official Gene Sets and manual annotations to new genome assemblies

What don’t we do with your data?

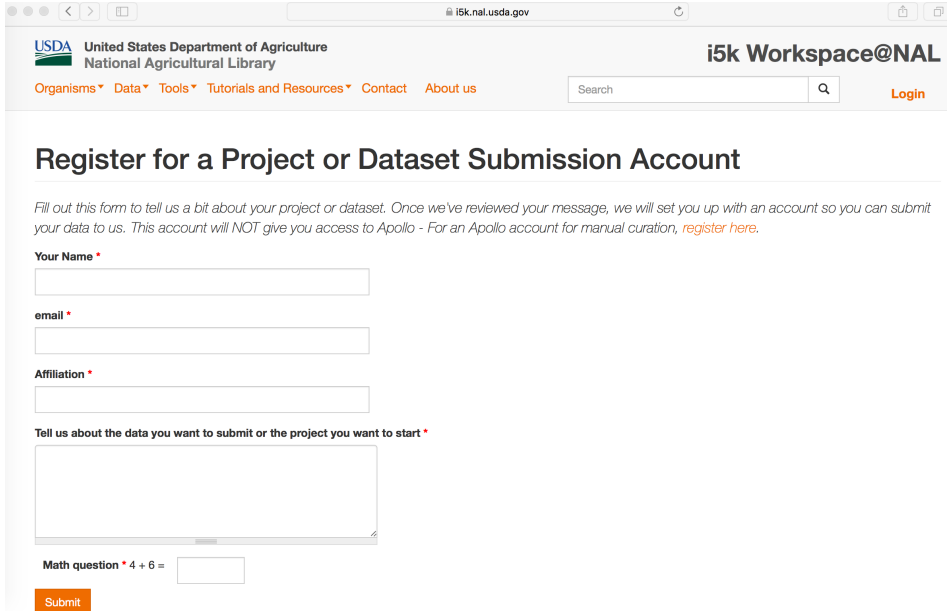
- Long-term archive.
  - That’s what NCBI is for.
  - We may also be able to help you submit your data to the Ag Data Commons repository.
- Our data management policy: https://i5k.nal.usda.gov/data-management-policy
- Our long-term project management policy: https://i5k.nal.usda.gov/long-term-i5k-workspace-project-management

Other things to consider before submitting

- ***All data submitted to the i5k Workspace is public.***
  - However, we do state whether Ft. Lauderdale/Toronto agreements of data sharing should apply
- Is your genome an 'orphan', or is there another suitable database?
  - We can host genomes that are already hosted elsewhere, but we keep in touch with other known database providers

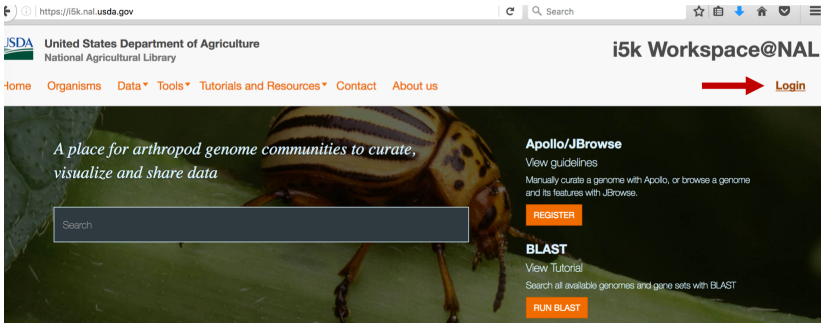
Getting an account

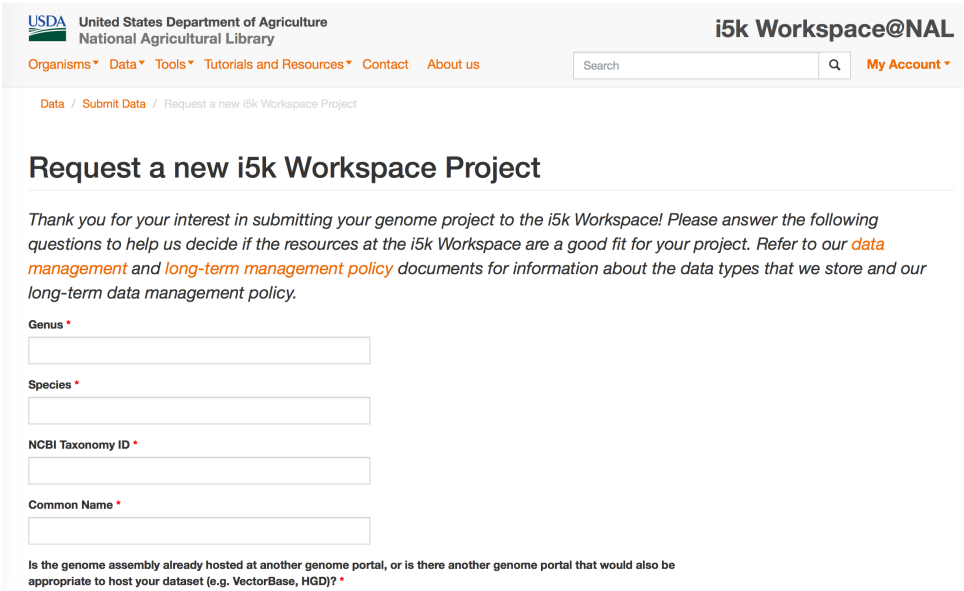
- Apply for a dataset submission account: https://i5k.nal.usda.gov/register/project-dataset/account
- For this workshop, you can use
  - Username: demo
  - Password: demo123



Start an i5k Workspace Project

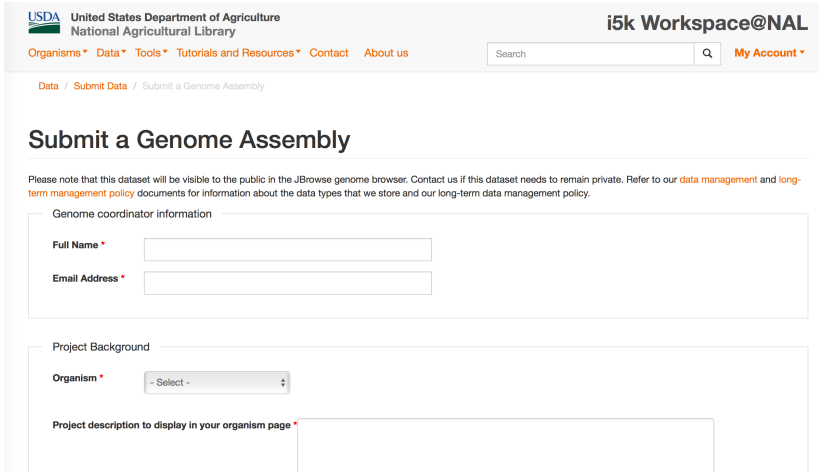
- Log in (demo/demo123)
  - https://i5k.nal.usda.gov/user
- From menu, select 'Data -> Submit data -> Request a new i5k Workspace Project'
  - https://i5k.nal.usda.gov/datasets/request-project
- We'll review your submission and will get in touch with you





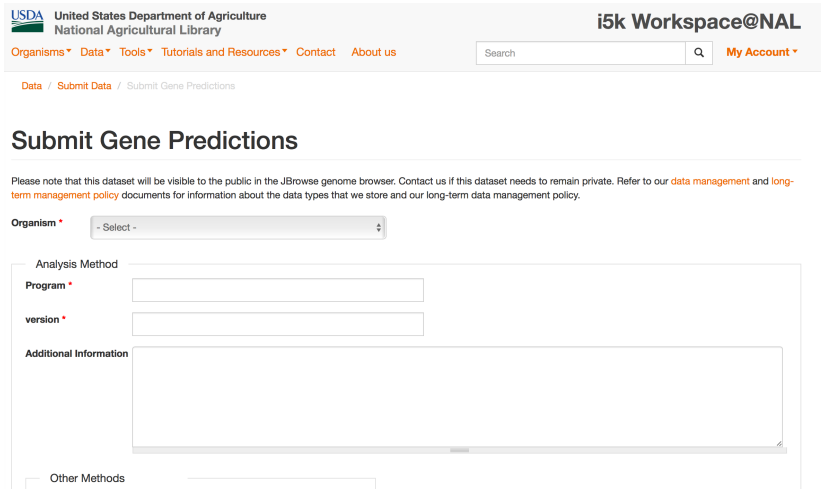
Submit your genome assembly

- *All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file md5sum)*
- From menu, select ‘Data -> Submit data -> Submit a genome assembly’
  - https://i5k.nal.usda.gov/datasets/assembly-data



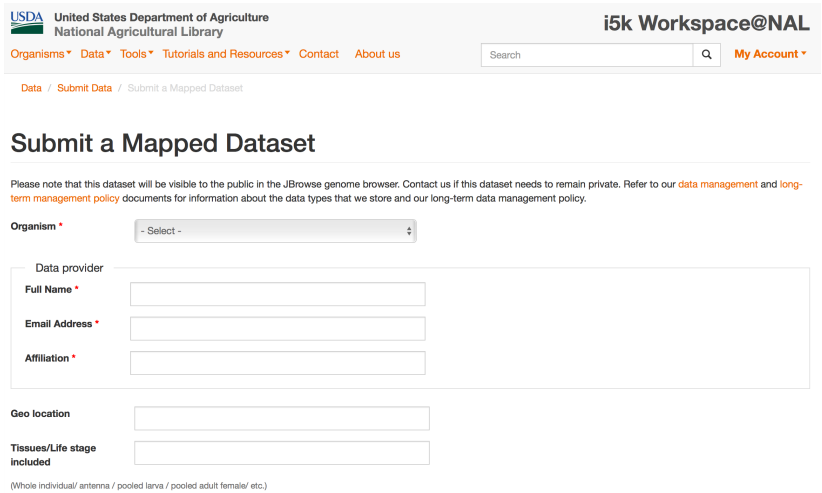
Submit gene predictions

- *All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file md5sum)*
- Under menu bar, select ‘Data -> Submit data -> Submit Gene Predictions’
  - https://i5k.nal.usda.gov/datasets/gene-prediction



Submit mapped datasets

- *All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file md5sum)*
- Under menu bar, select ‘Data -> Submit data -> Submit a Mapped Dataset’
  - https://i5k.nal.usda.gov/datasets/mapped



Send us your files

- Forms are only for metadata
- General guidelines for sharing files with us:
  - https://i5k.nal.usda.gov/content/sharing-files-us
- We'll get in touch with you about the best way to get your data to us
- File uploads via the submission forms: coming soon

Overview

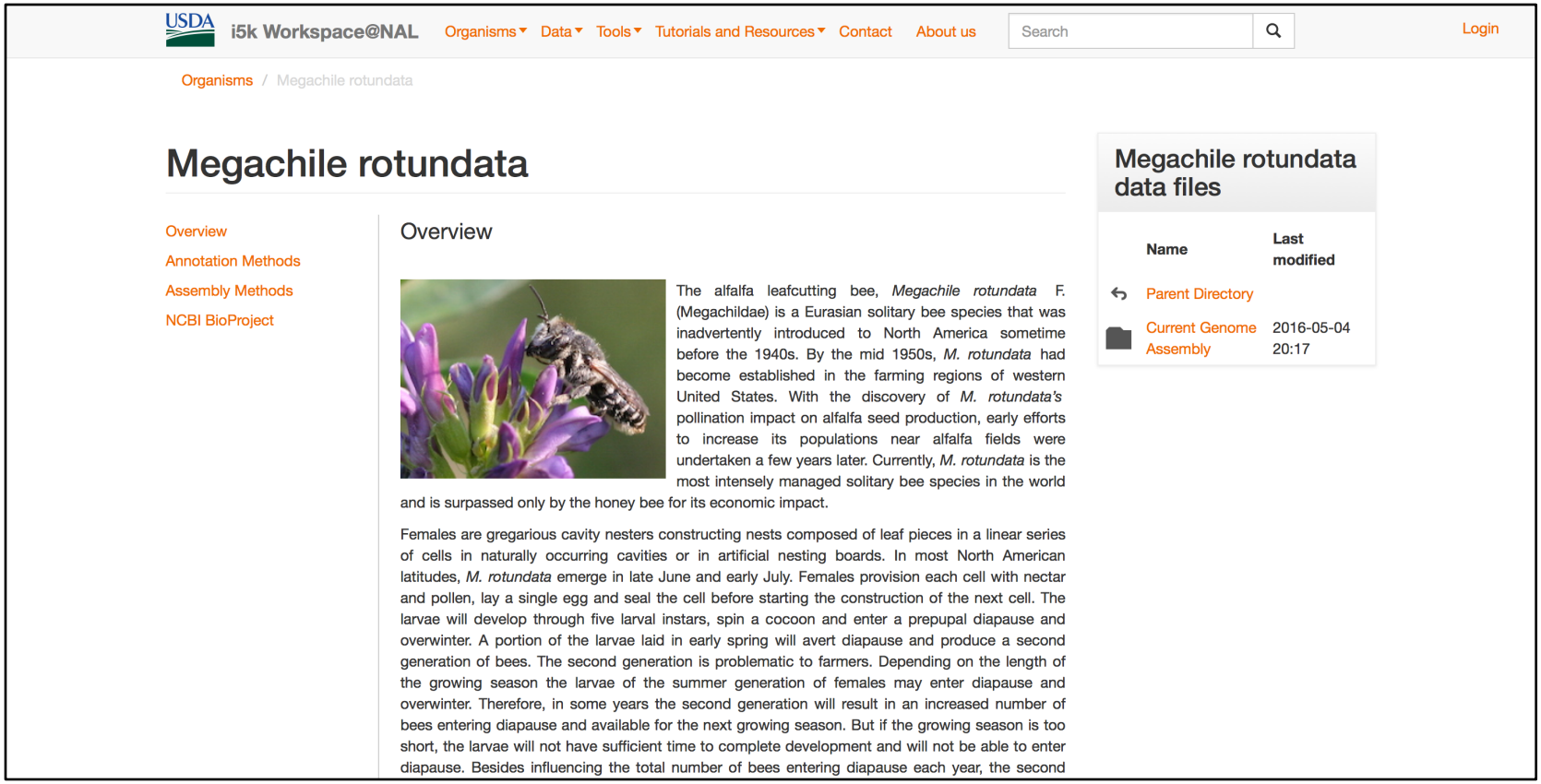
1. Background: What is the i5k Workspace?
2. Submitting data
3. Finding data at the i5k Workspace
    1. General search/Content types
    2. Data downloads
    3. BLAST
    4. Clustal(s)
    5. HMMER
    6. Jbrowse/Apollo
4. Improving data at the i5k Workspace via community annotation
    1. See Monica Munoz-Torres' workshop for full use of Apollo
5. Other data management resources at the NAL – the Ag Data Commons

USDA

Finding Data at the i5k Workspace

- We have different kinds of information to search for:
  - Information about each i5k Workspace project (project metadata, available in our organism pages)
  - For Official Gene Sets: Gene names, gene metadata (available in our gene pages)
  - Sequence data
  - Flat files (bulk data downloads)

Organism pages



Organism pages

latitudes, *M. rotundata* emerge in late June and early July. Females provision each cell with nectar and pollen, lay a single egg and seal the cell before starting the construction of the next cell. The larvae will develop through five larval instars, spin a cocoon and enter a prepupal diapause and overwinter. A portion of the larvae laid in early spring will avert diapause and produce a second generation of bees. The second generation is problematic to farmers. Depending on the length of the growing season the larvae of the summer generation of females may enter diapause and overwinter. Therefore, in some years the second generation will result in an increased number of bees entering diapause and available for the next growing season. But if the growing season is too short, the larvae will not have sufficient time to complete development and will not be able to enter diapause. Besides influencing the total number of bees entering diapause each year, the second generation has been implicated as a major factor in the spread of chalk brood, the primary disease of *M. rotundata*. The development of a *M. rotundata* genome database is an important advancement for understanding basic physiology and disease management of this important pollinator.

Community contact: George Yocum Karen Kapheim Hailin Pan
Image Credit: Theresa Pitts-Singer, U.S. Department of Agriculture. Public domain.

Assembly Information

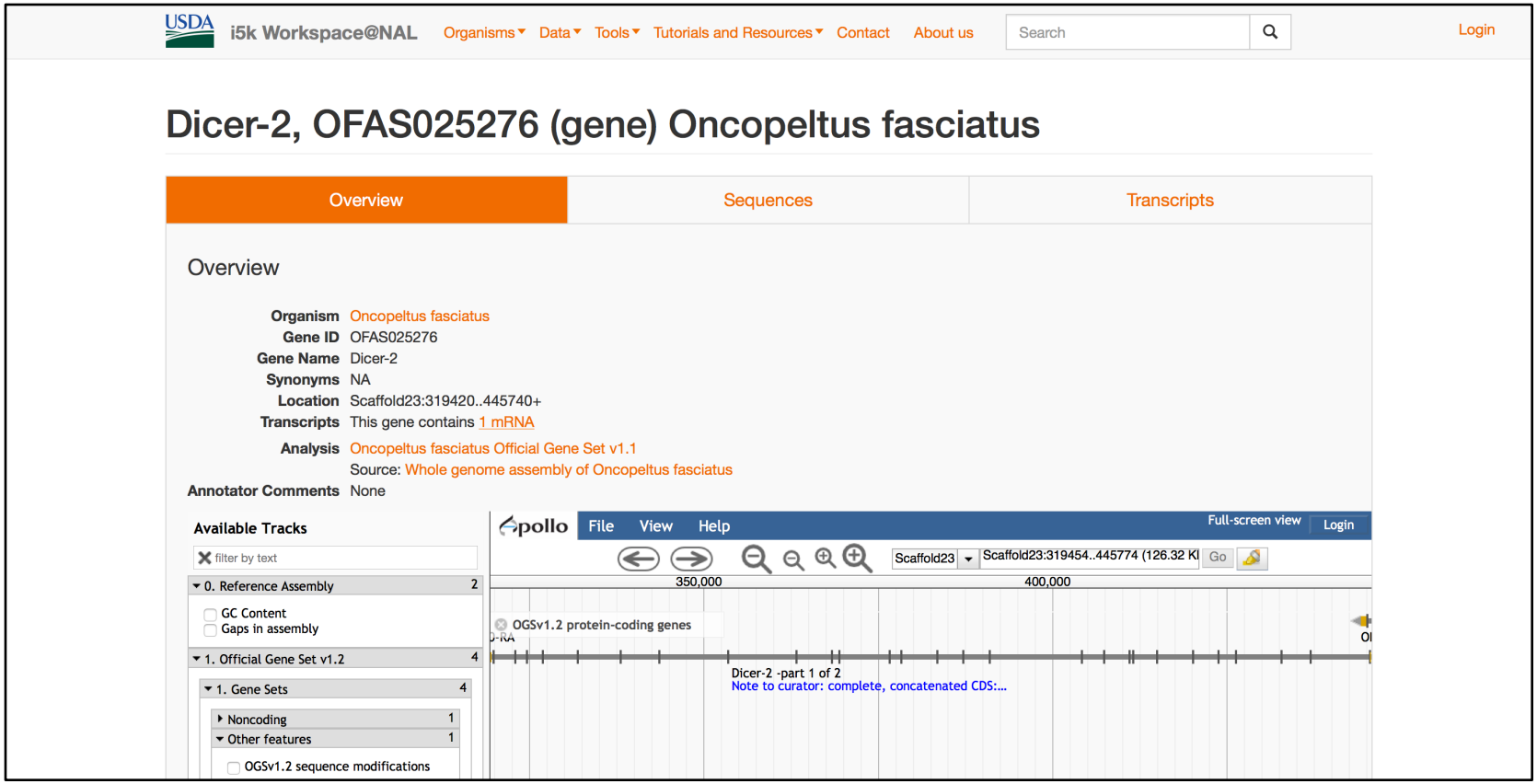
Analysis Name	Megachile rotundata genome assembly MROT_1.0 (GCF_000220905.1)
Software	SOAPdenovo Assembler (1.05)
Source	BioProject PRJNA66515
Date performed	2016-03-23
Materials & Methods	

Statistics

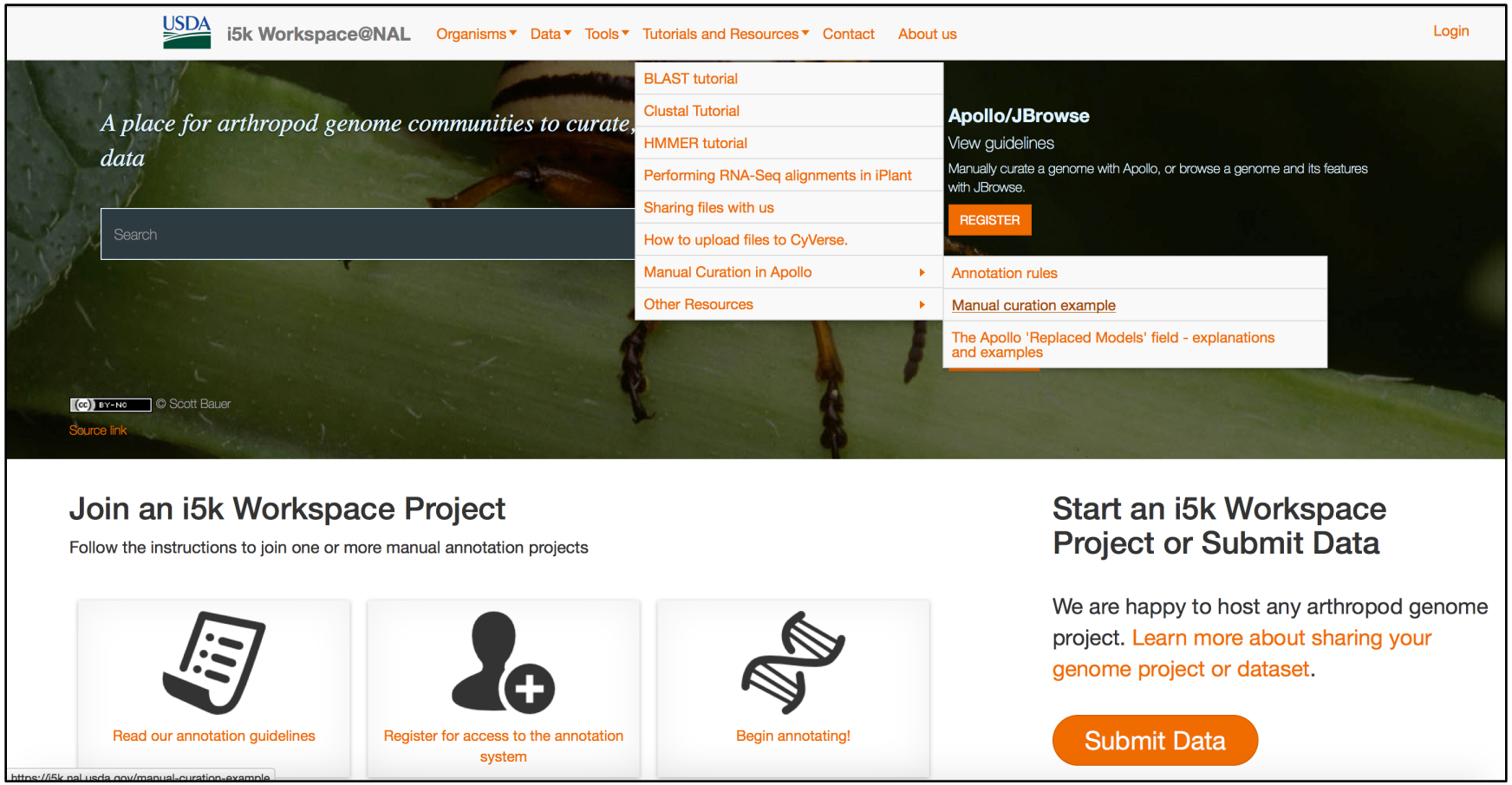
Assembly Metrics	
Contig N50	NA
Scaffold N50	1699680
GC Content	40.54
Manual Annotations	

USDA

Gene pages (Official gene sets only)

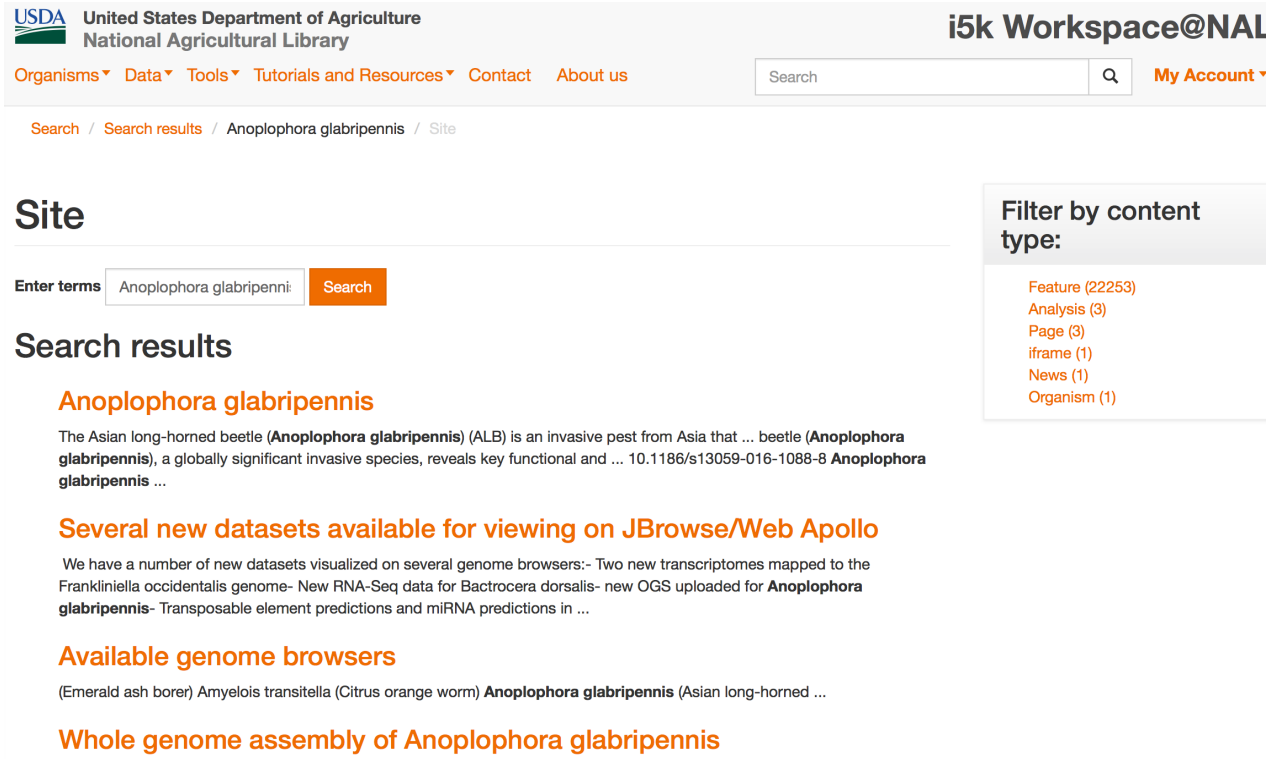


Tutorials


USDA
i5k Workspace@NAL
Organisms
Data
Tools
Tutorials and Resources
Contact
About us
Login
A place for arthropod genome communities to curate, data
Search
BLAST tutorial
Clustal Tutorial
HMMER tutorial
Performing RNA-Seq alignments in iPlant
Sharing files with us
How to upload files to CyVerse.
Manual Curation in Apollo
Other Resources
Annotation rules
Manual curation example
The Apollo 'Replaced Models' field - explanations and examples
Apollo/JBrowse
View guidelines
Manually curate a genome with Apollo, or browse a genome and its features with JBrowse.
REGISTER
© Scott Bauer
Source link
Join an i5k Workspace Project
Follow the instructions to join one or more manual annotation projects
Read our annotation guidelines
Register for access to the annotation system
Begin annotating!
Start an i5k Workspace Project or Submit Data
We are happy to host any arthropod genome project. Learn more about sharing your genome project or dataset.
Submit Data

Finding Data at the i5k Workspace

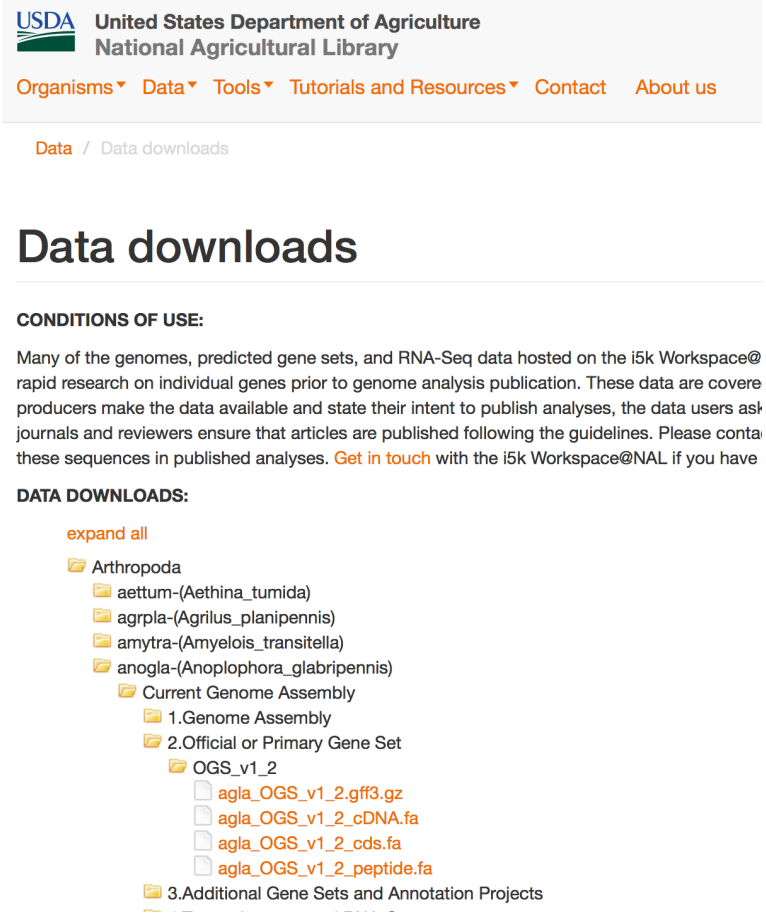
- Website search for metadata (e.g. search term "Anoplophora glabripennis")



USDA

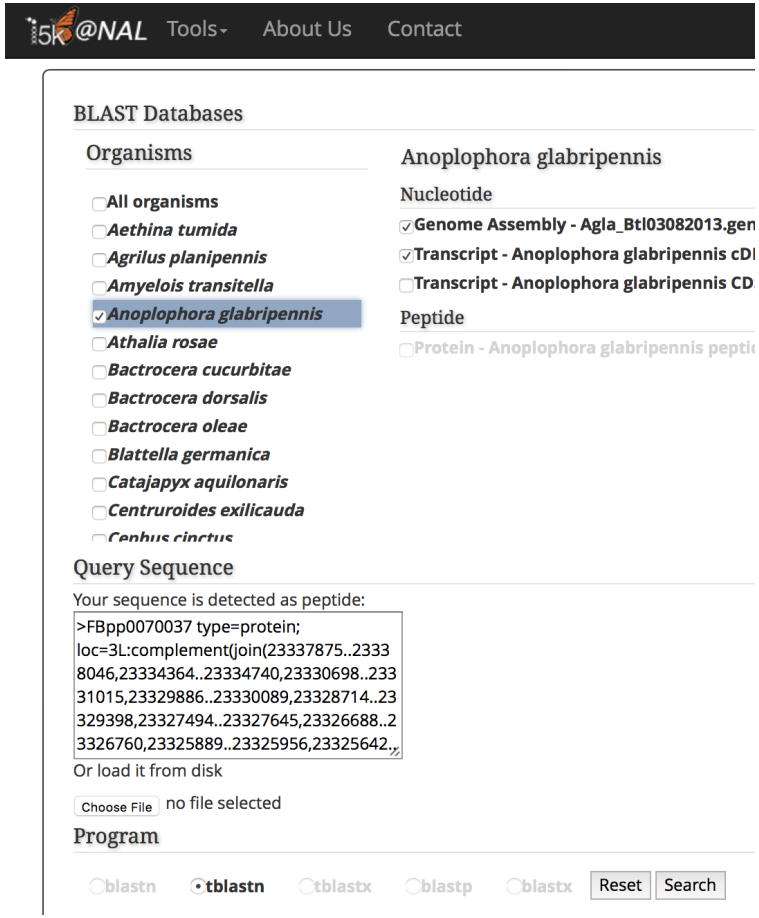
Bulk data downloads for full files

- From menu, select ‘Data -> Data Downloads’
  - https://i5k.nal.usda.gov/content/data-downloads, or
  - https://i5k.nal.usda.gov/data/



Sequence Search – BLAST+

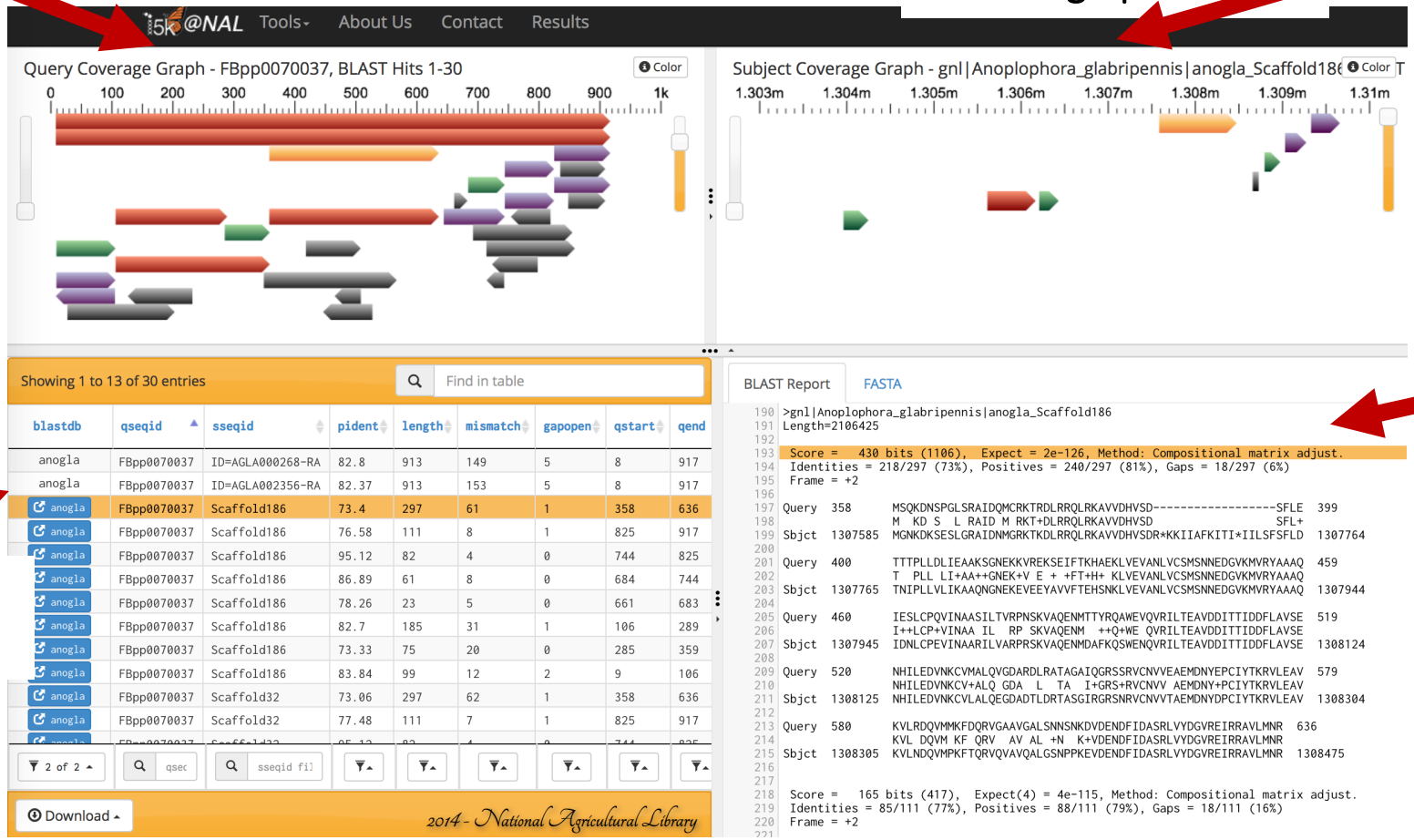
- From menu, select ‘Tools -> BLAST’
  - https://i5k.nal.usda.gov/webapp/blast/
- Tutorial:
  - https://i5k.nal.usda.gov/content/blast-tutorial
- Example query:
  - http://flybase.org/cgi-bin/getseq.html?source=dmel&id=FBpp0070037&chr=3L&dump=PrecompiledFasta&targetset=translation



Sequence Search – BLAST+ Result

Query coverage graph

Subject coverage graph

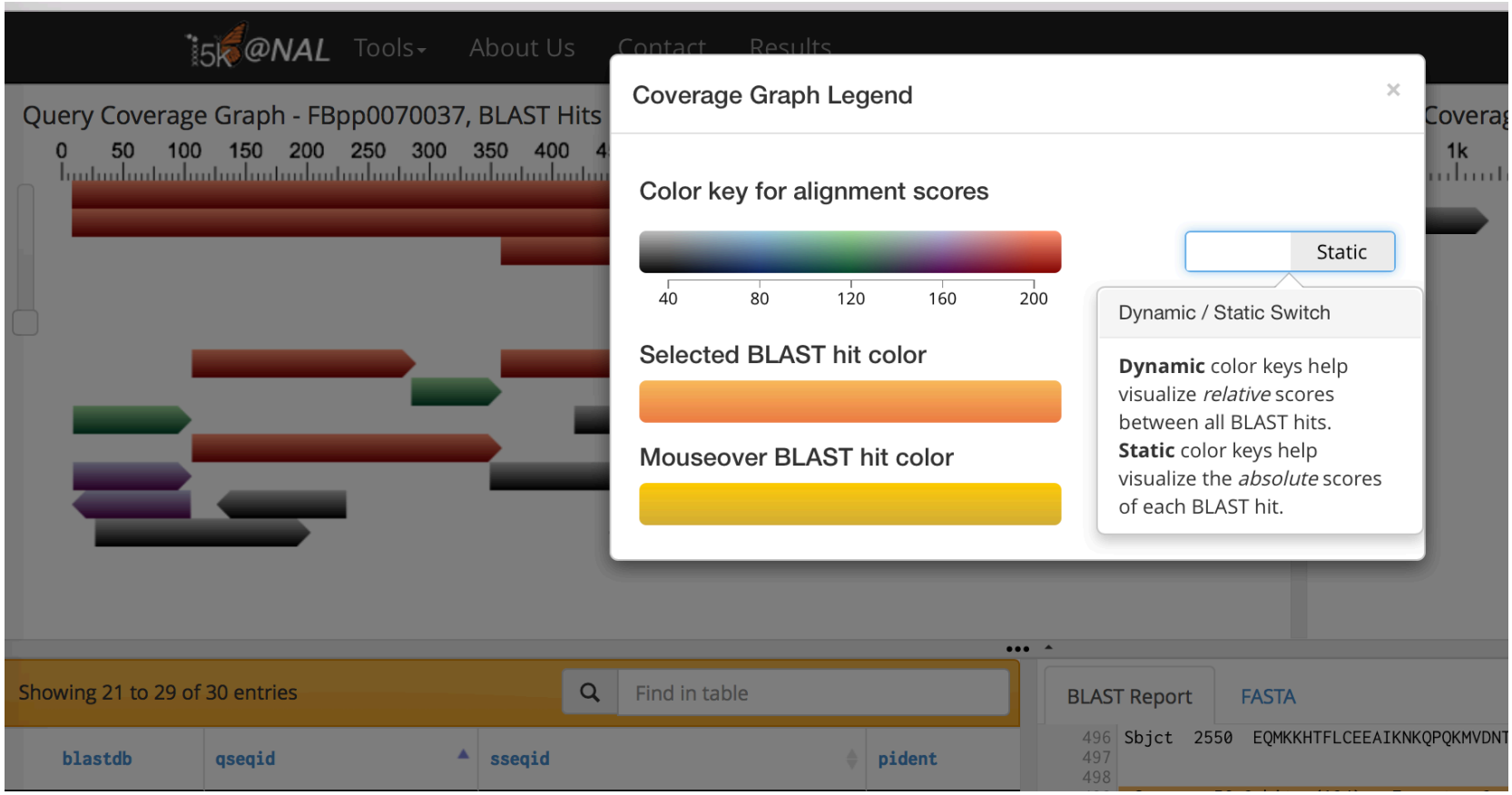


Tabular result

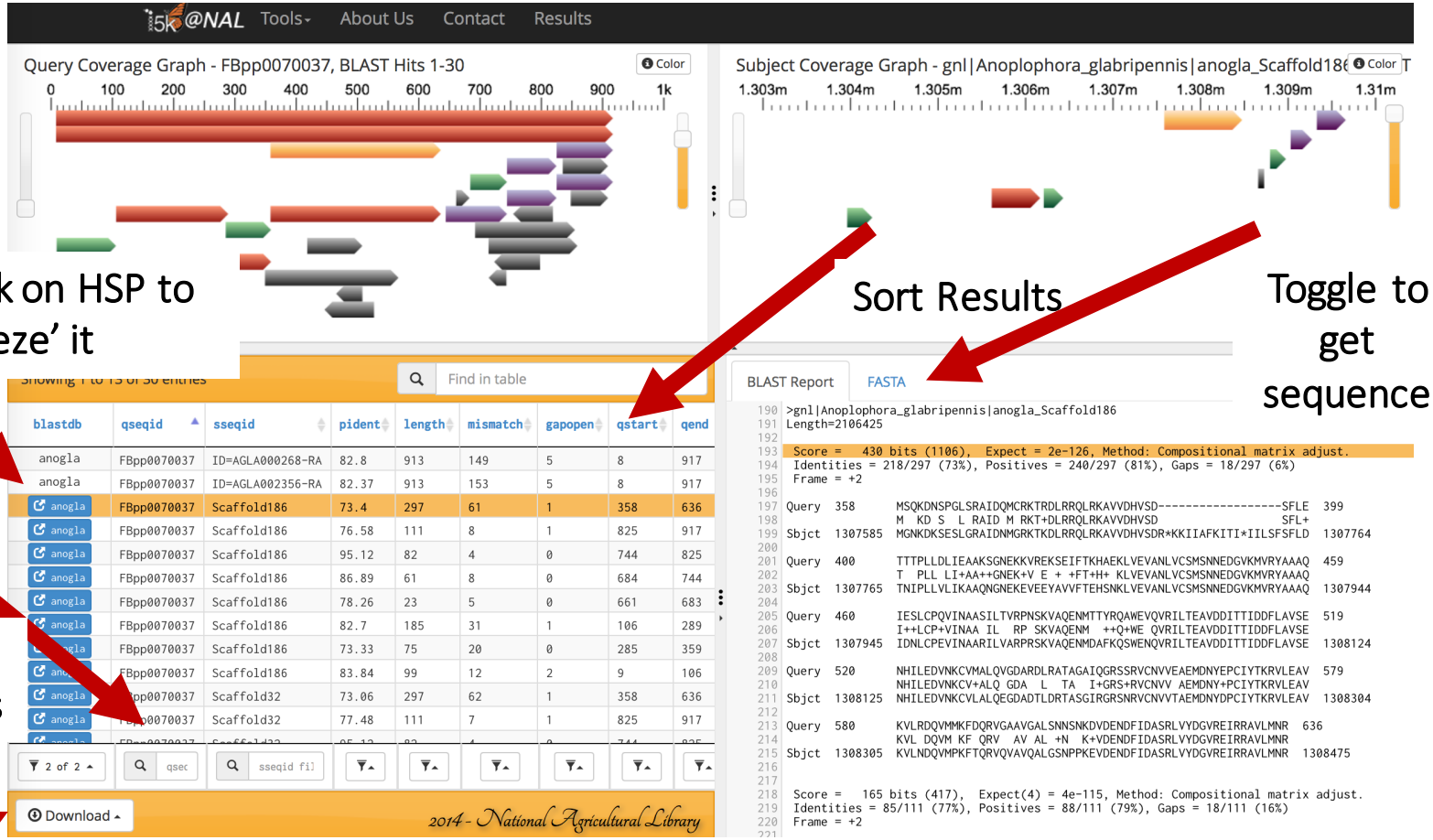
Raw results

Result URL: https://i5k.nal.usda.gov/webapp/blast/0b0f5690f06b446f86005bcc9f1bd3d7

Sequence Search – BLAST+ Result



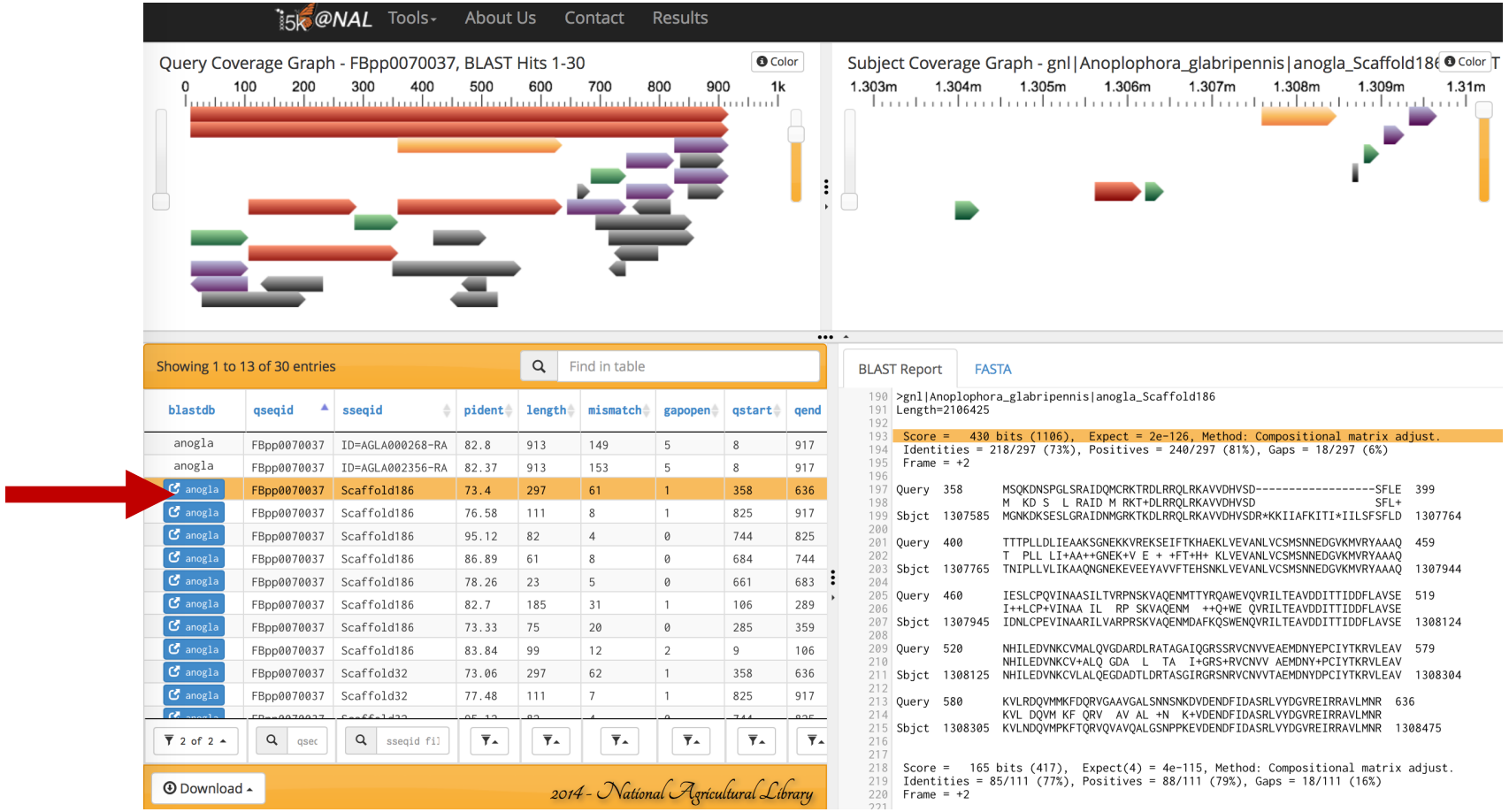
Sequence Search – BLAST+ Result



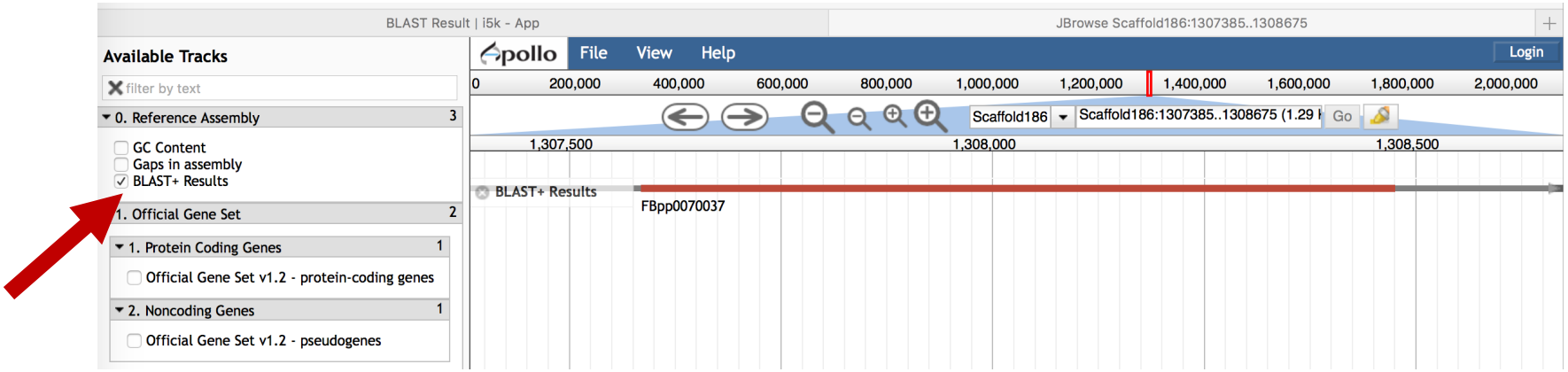
Result URL: https://i5k.nal.usda.gov/webapp/blast/0b0f5690f06b446f86005bcc9f1bd3d7

USDA

Sequence Search – links to JBrowse

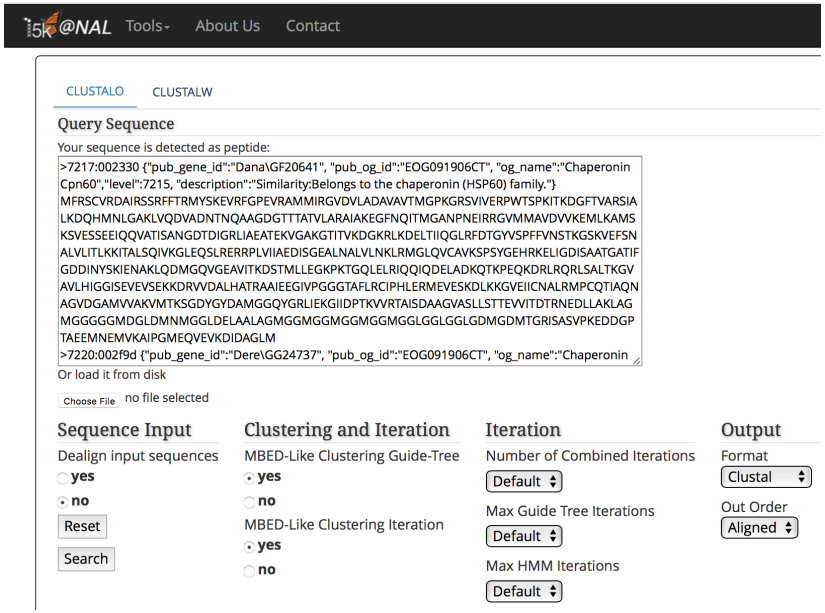


Sequence Search – links to JBrowse

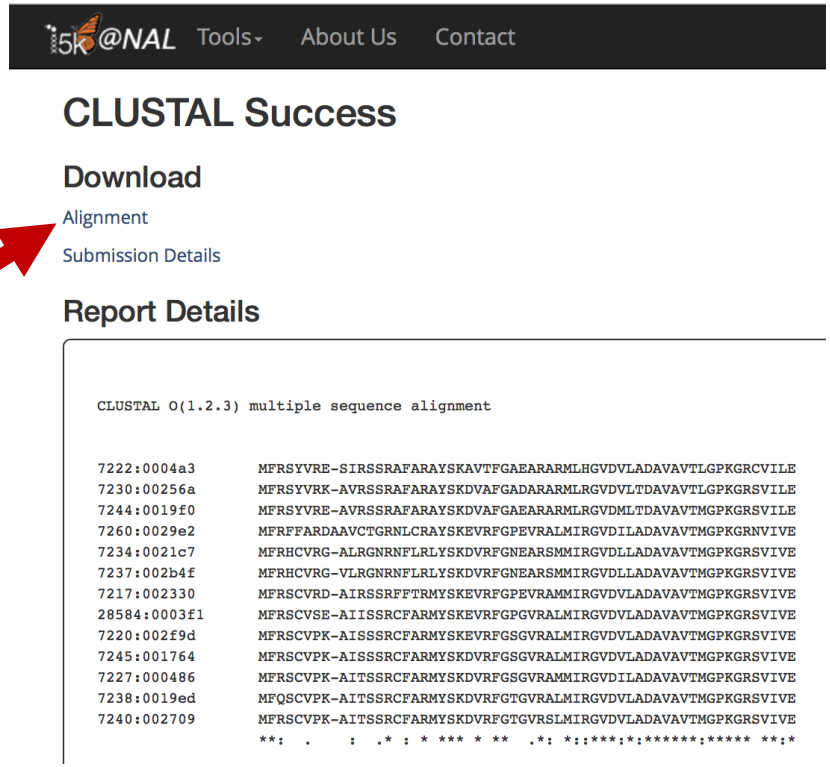


Sequence alignment – ClustalW and Clustal Omega

- From menu, select 'Tools -> Clustal (beta)'
  - https://i5k.nal.usda.gov/webapp/clustal/
- Example query sequences:
  - http://www.orthodb.org/fasta?query=EOG091906CT&level=&species=&universal=&singlecopy=



Sequence alignment – ClustalW and Clustal Omega

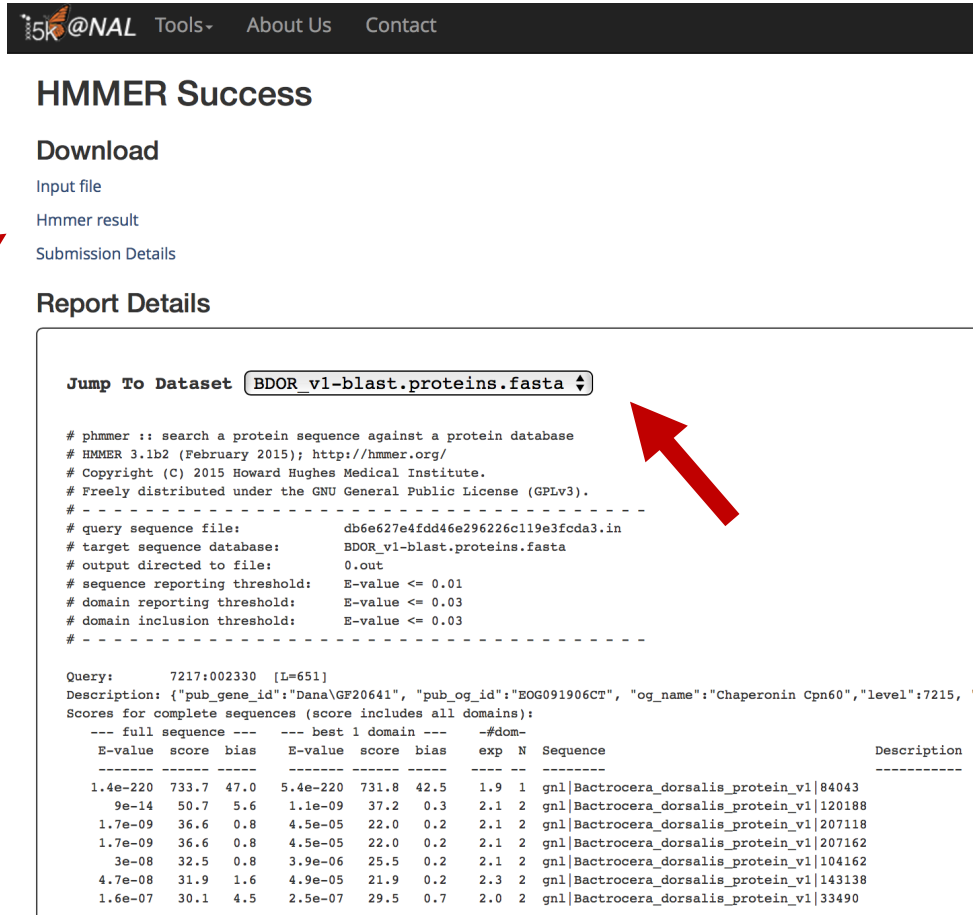


Result URL:
https://i5k.nal.usda.gov/webapp/clustal/ed00819ab40441ca959eacdccb78c0f5

Sequence search – HMMER

- From menu, select 'Tools -> HMMER (beta)'
  - https://i5k.nal.usda.gov/webapp/hmmer/
- Example query sequences (fasta, restrict to <10):
  - http://www.orthodb.org/fasta?query=EOG091906CT&level=&species=&universal=&singlecopy=

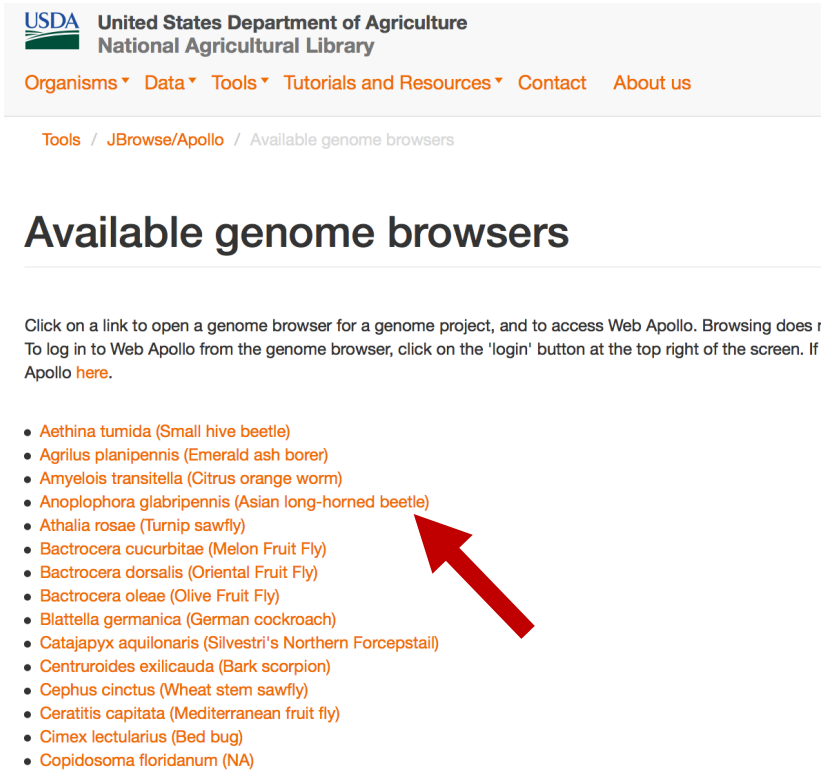
Sequence search – HMMER Result



Result URL: https://i5k.nal.usda.gov/webapp/hmmer/1aafe206cc7749ce8fbeab582d999432

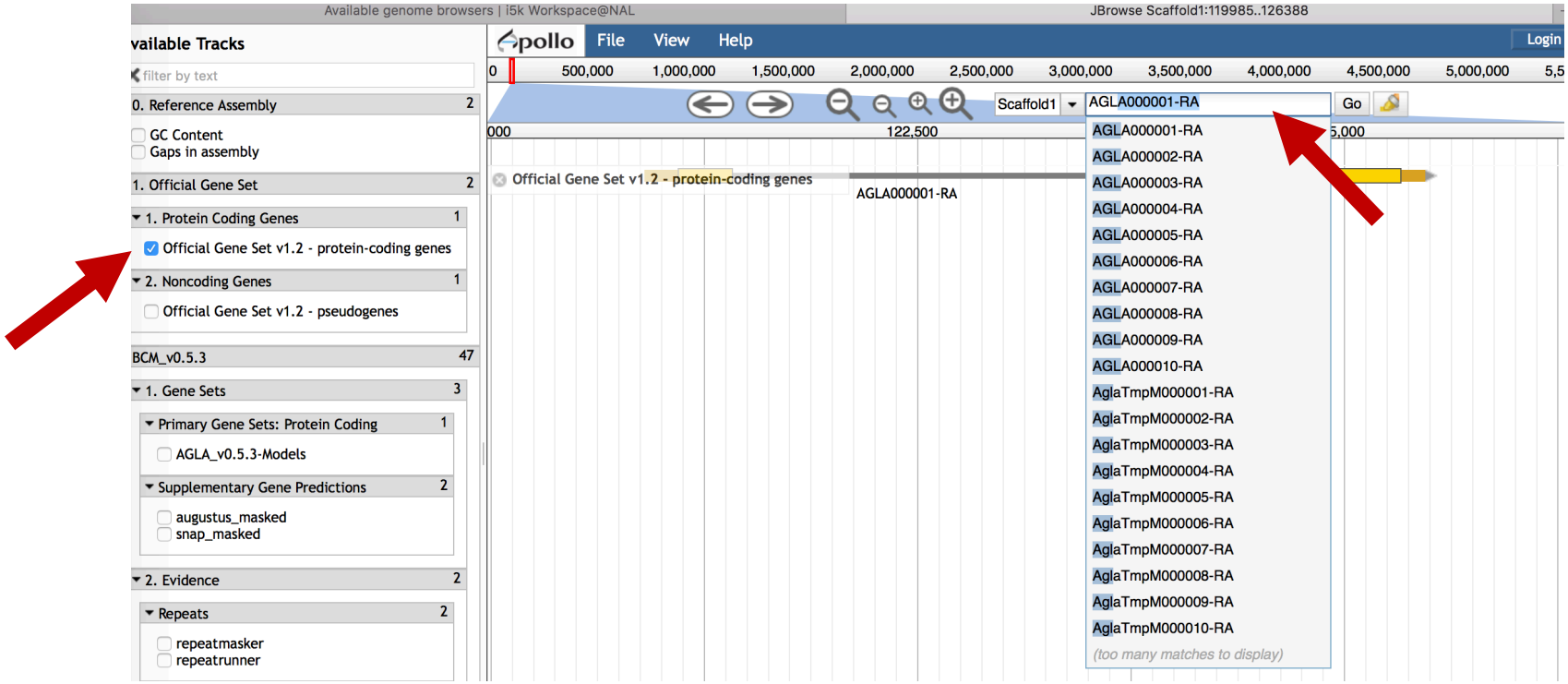
Genome browser (JBrowse)

- From menu, select ‘Tools -> JBrowse/Apollo -> JBrowse/Apollo Organisms’
  - https://i5k.nal.usda.gov/available-genome-browsers



Genome browser (JBrowse)

- If you know the gene ID of your gene of interest, you can paste it into the JBrowse 'Search' bar



Overview

1. Background: What is the i5k Workspace?
2. Submitting data
3. Finding data at the i5k Workspace
    1. General search
    2. Data downloads
    3. BLAST
    4. Clustal(s)
    5. HMMER
    6. Jbrowse/Apollo
4. Improving data at the i5k Workspace via community annotation
    1. See Monica Munoz-Torres' workshop for full use of Apollo
5. Other data management resources at the NAL – the Ag Data Commons

USDA

Improving Data at the i5k Workspace via Manual Annotation

- What is manual annotation?
  - Verify or improve the biological validity of computationally predicted gene models
  - Assign function to gene models via comparative analysis
- Why manually annotate?
  - Automated gene predictions often contain errors
  - Improve gene models for specific analyses
- For more background, visit Moni Munoz-Torres' workshop
- Apollo documentation:
  - http://genomearchitect.github.io/users-guide/
  - https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project
  - https://i5k.nal.usda.gov/manual-curation-example

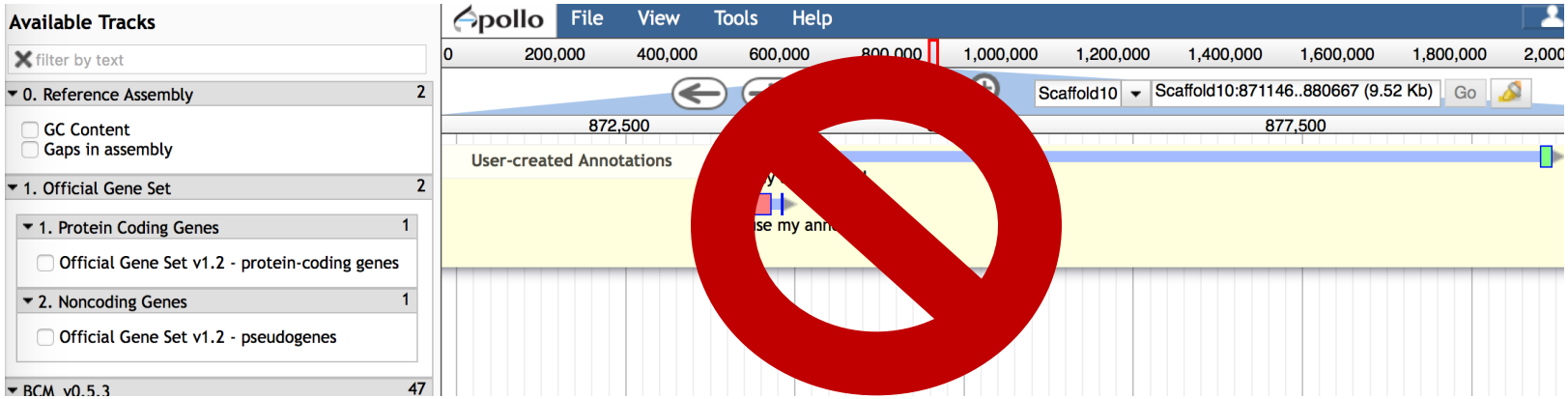
Community curation at the i5k Workspace

Our support for community curation includes:

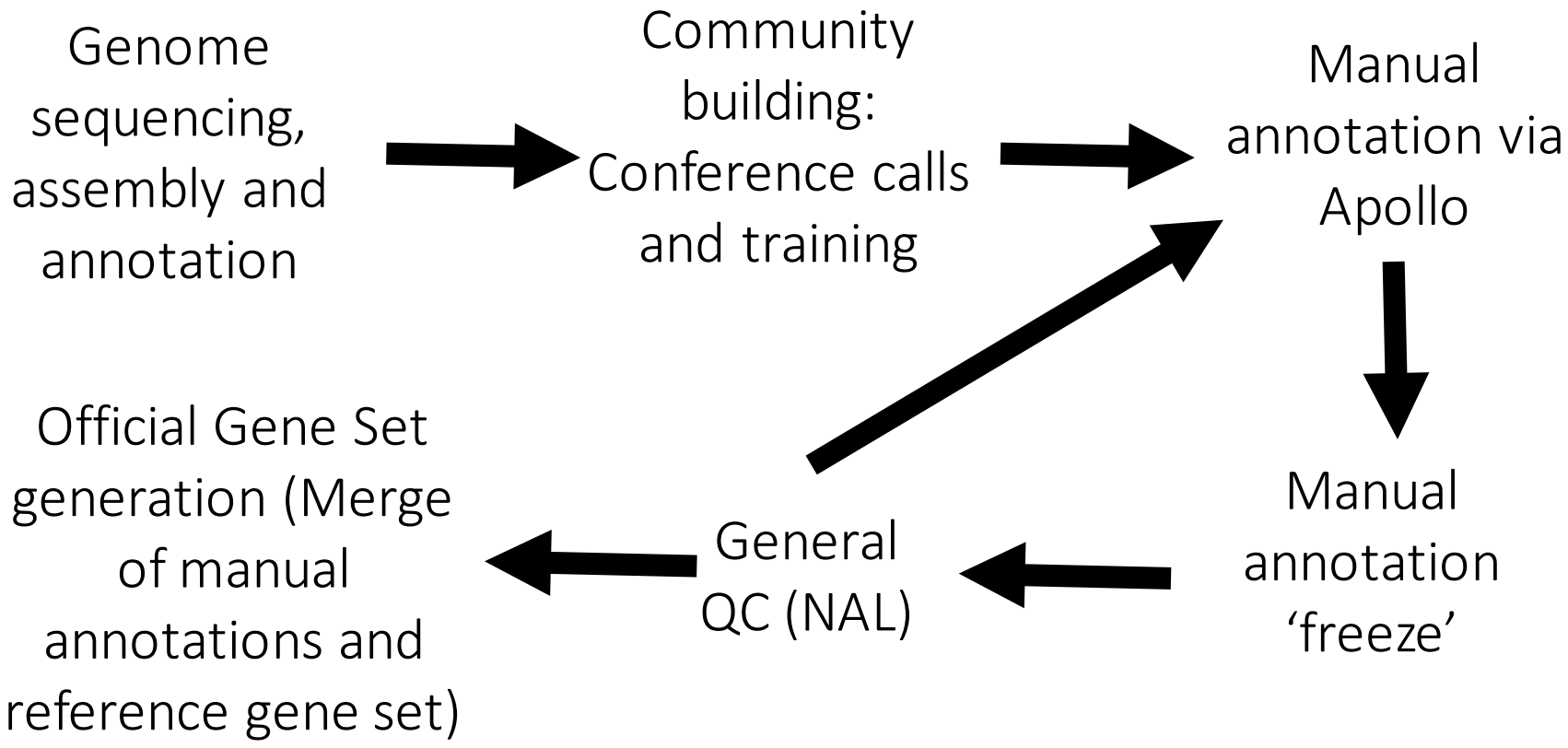
- Access to a large community of curators
- Tutorials, guidelines, webinars
- Registration mechanism for new annotators
- One-on-one support
- Software to evaluate changes between curated and original annotations (Chien-Yueh Lee, https://github.com/chienyuehlee/gff-cmp-cat)

Principles of community annotation

- Collaborative effort across many individuals, often in different time zones and countries
- We encourage annotators to work together to find the best solution
- We work with each project coordinator to facilitate communication and collaboration whenever possible.



Manual annotation life cycle (end goal: OGS)

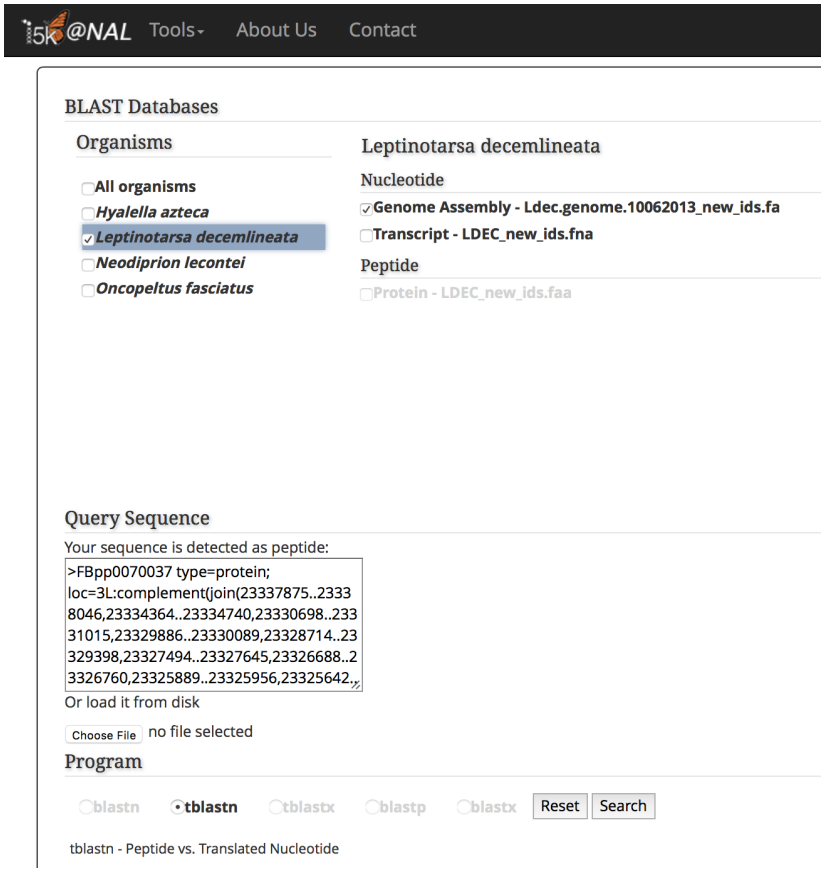


Some Apollo notes

- We're still using Apollo1 – Apollo2 has a slightly different interface
- Here, we'll use our 'Training' applications
- Apollo credentials for training applications:
  - Username: demo
  - Password: demo
- To annotate on an actual project, you'll need to register first:
  - From menu, select 'Tools -> JBrowse/Apollo -> Apollo registration form'
  - https://i5k.nal.usda.gov/web-apollo-registration
  - Registration is only for the organisms that you select

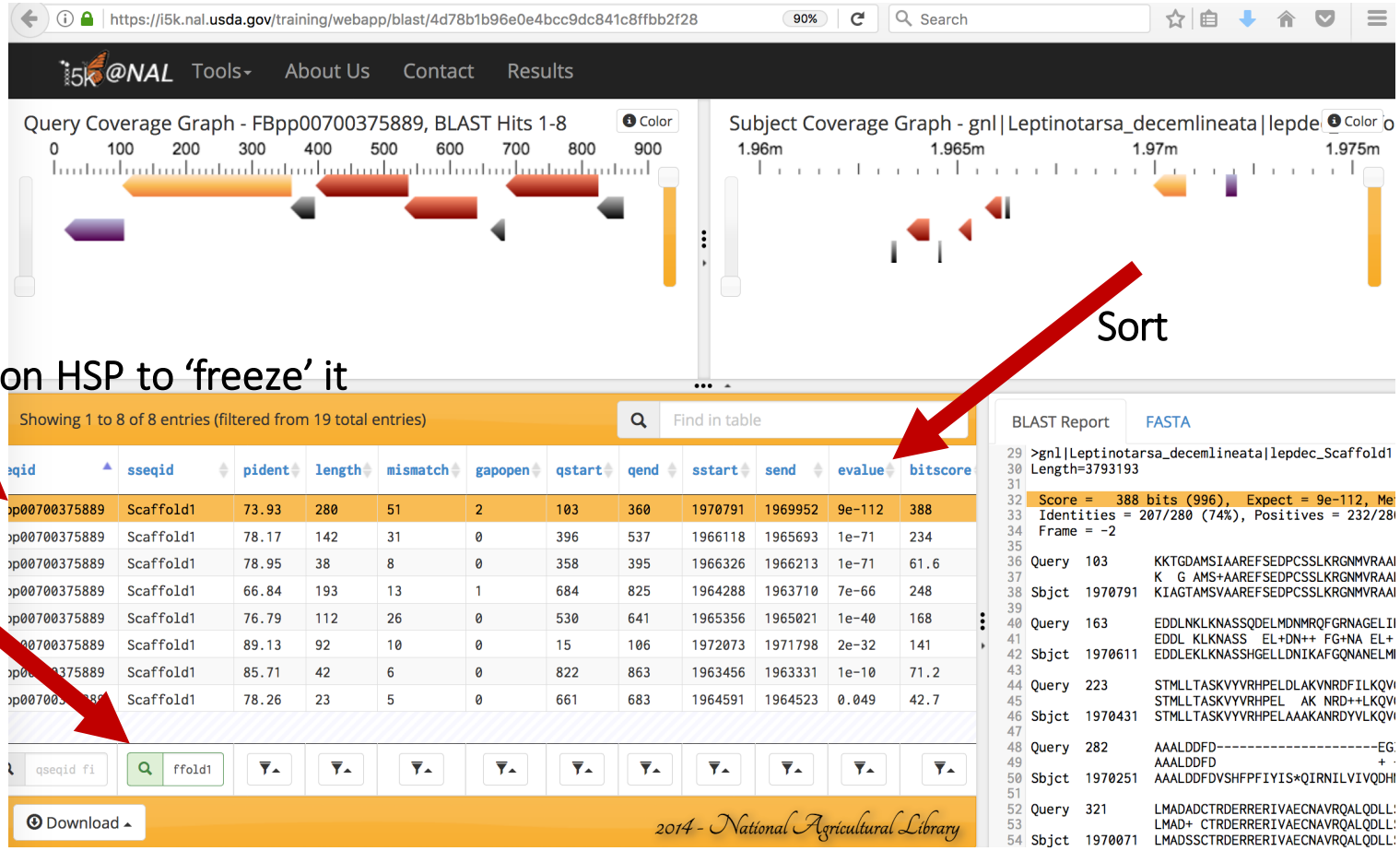
Example workflow: alpha-catenin in the Colorado Potato Beetle

- From menu, select 'Tools -> Training tools -> Training BLAST'
  - https://i5k.nal.usda.gov/training/webapp/blast
- Query sequence:
  - http://flybase.org/cgi-bin/getseq.html?source=dmel&id=FBpp0070037&chr=3L&dump=PrecompiledFasta&targetset=translation



USDA

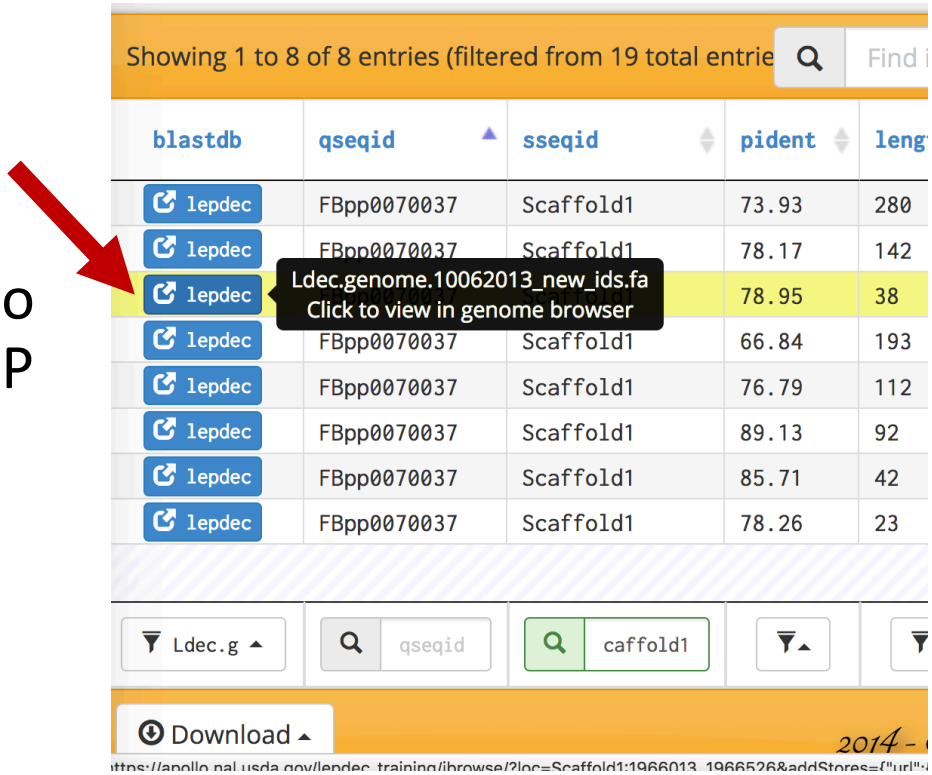
Example workflow: alpha-catenin in the Colorado Potato Beetle



Result URL: https://i5k.nal.usda.gov/training/webapp/blast/613bde5948cb450da1b1d2d891995c9d

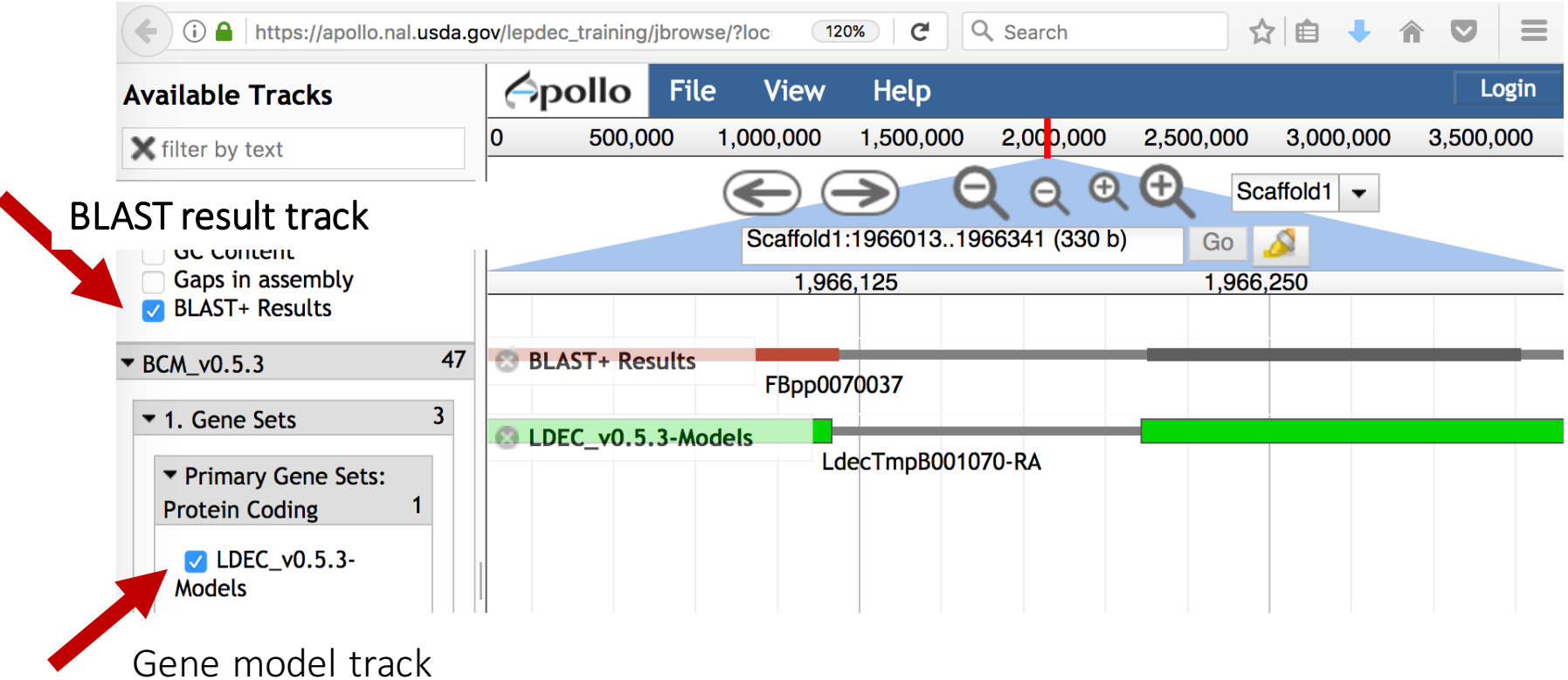
Example workflow: alpha-catenin in the Colorado Potato Beetle

- To view HSP in the genome browser:
  - Go to result table on bottom left
  - click on the blue box to the left of the best HSP result in the 'blastdb' column



Result URL: https://i5k.nal.usda.gov/training/webapp/blast/613bde5948cb450da1b1d2d891995c9d

Example workflow: alpha-catenin in the Colorado Potato Beetle

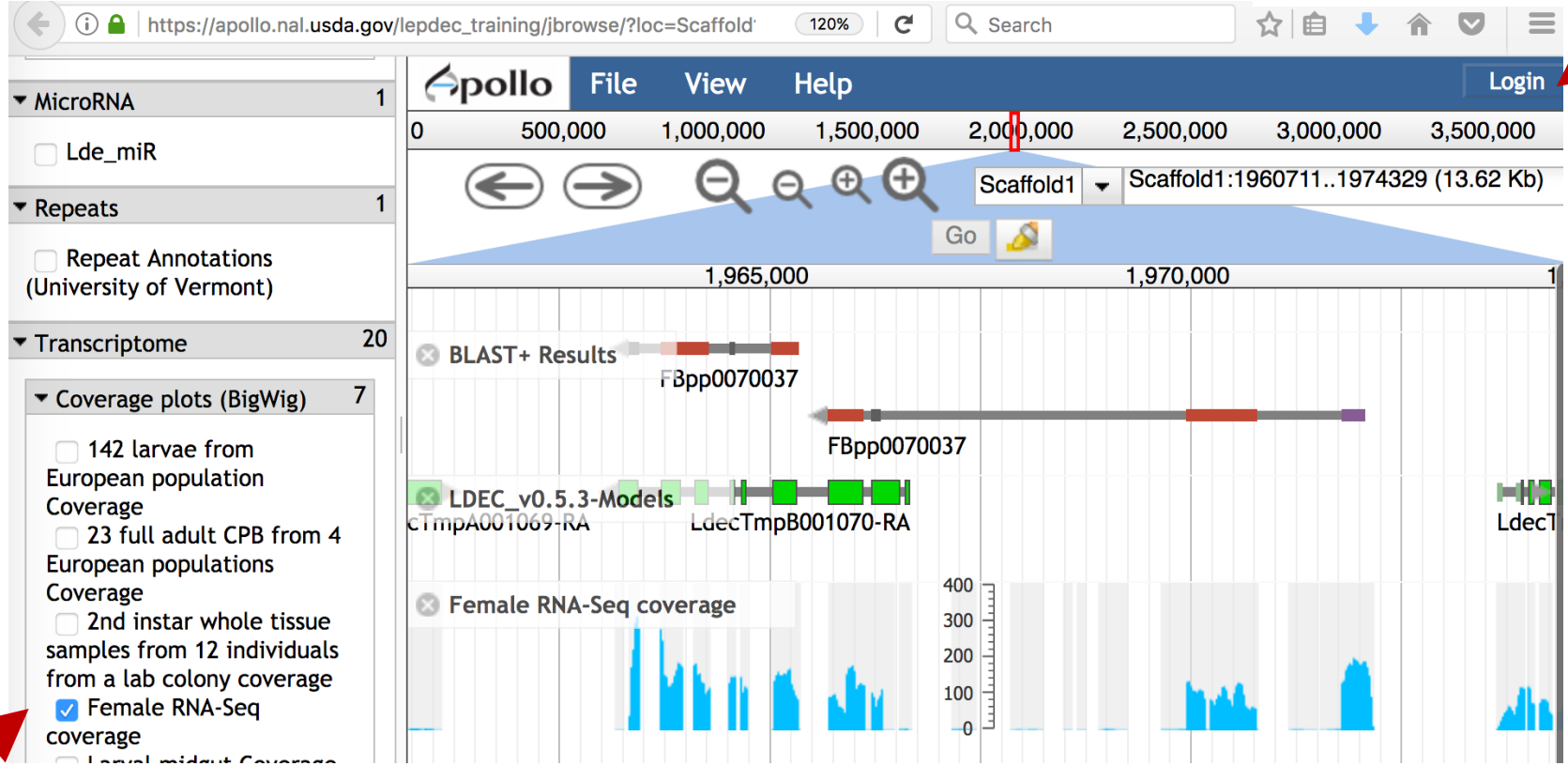


URL: https://tinyurl.com/ybf4ehld

USDA

Example workflow: alpha-catenin in the Colorado Potato Beetle

Log in (demo/demo)

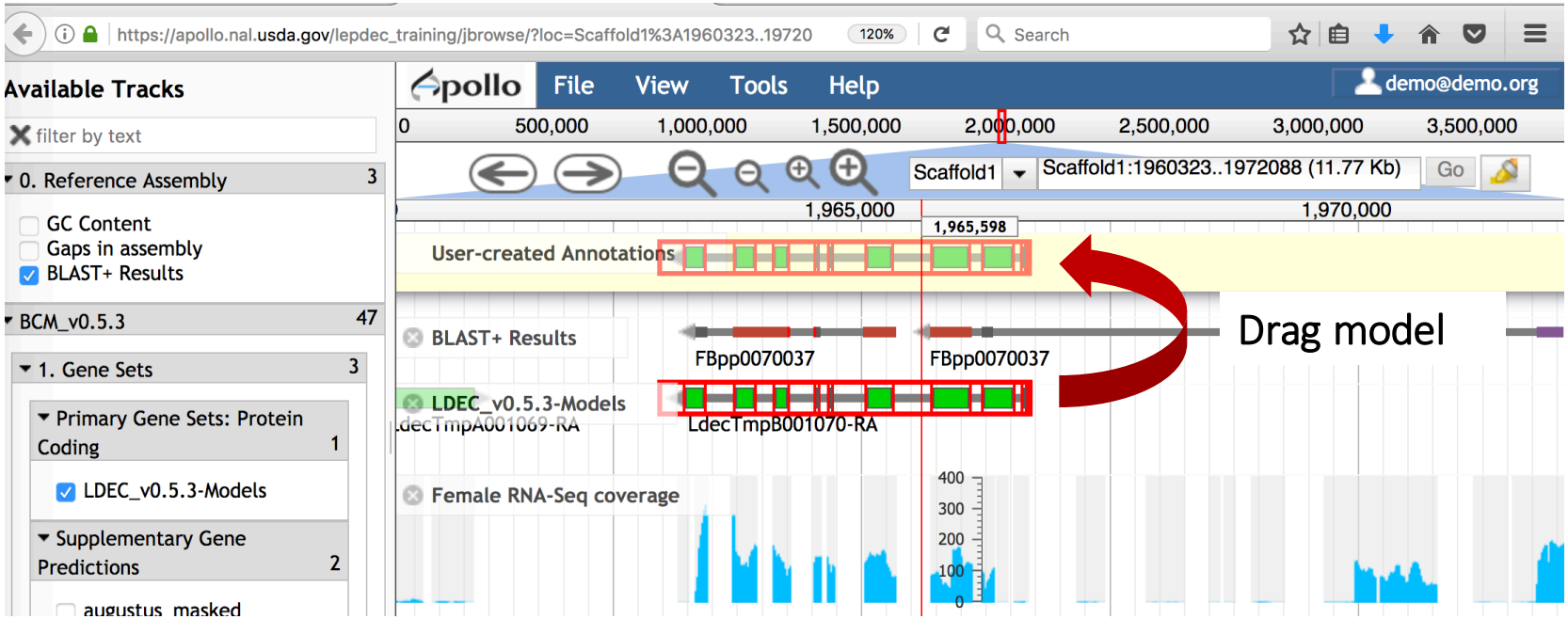


RNA-Seq evidence tracks

URL: https://tinyurl.com/y8688kgt

USDA

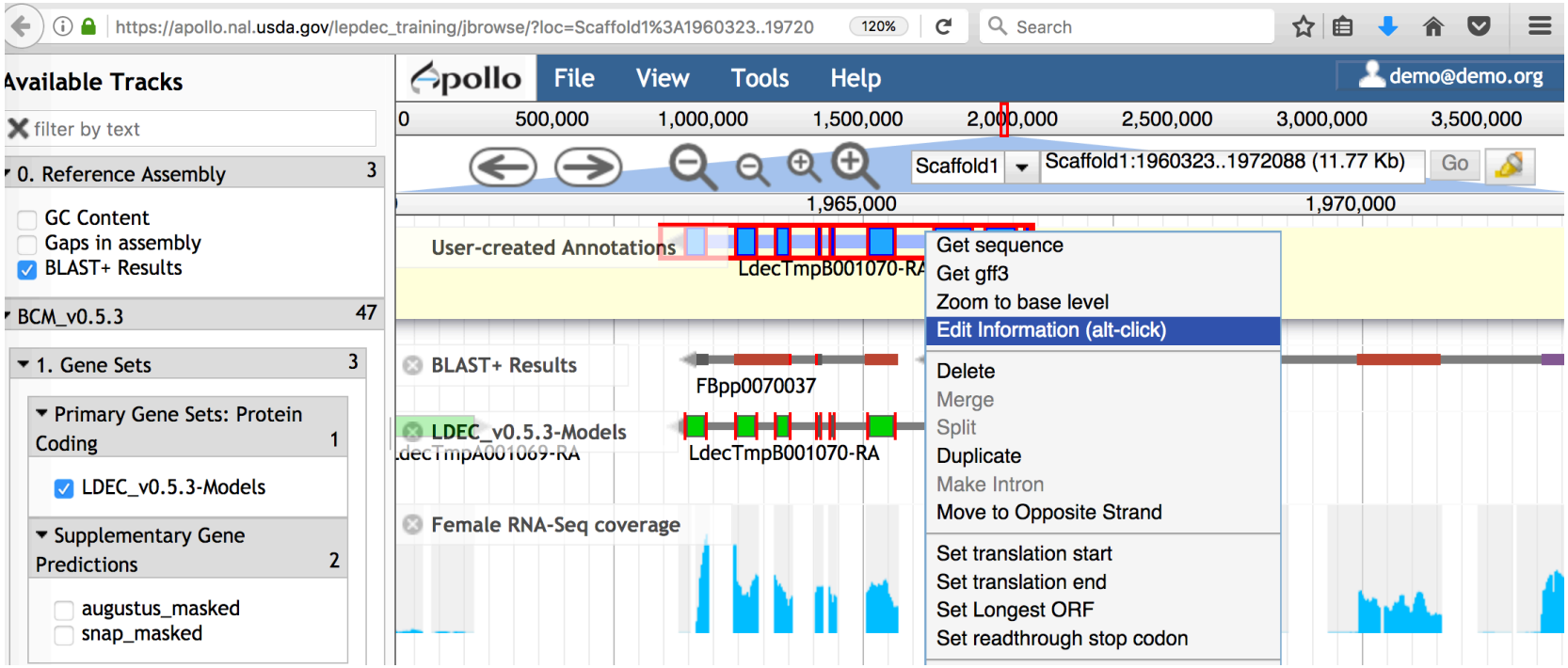
Example workflow: alpha-catenin in the Colorado Potato Beetle



URL: https://tinyurl.com/y8688kgt

USDA

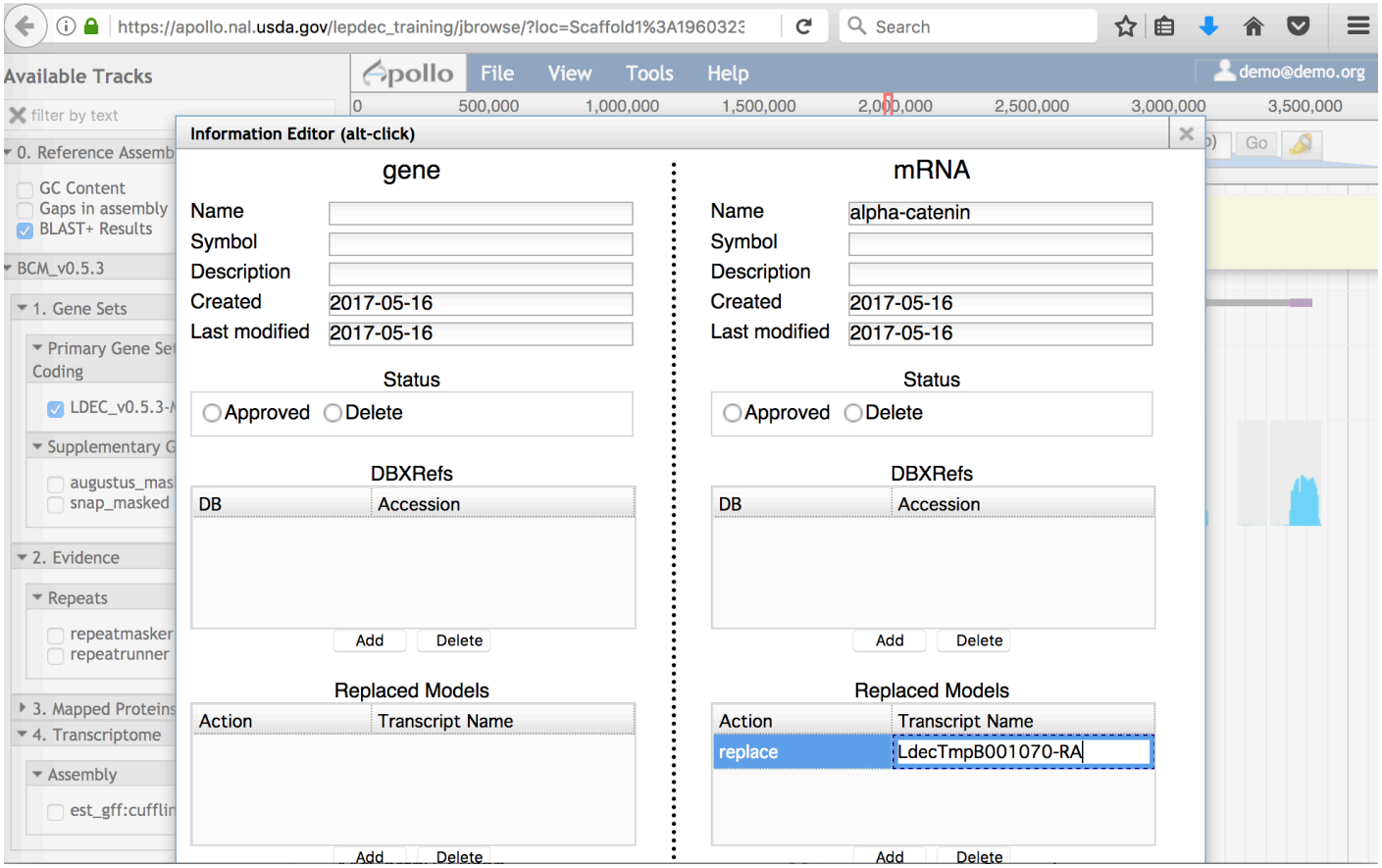
Example workflow: alpha-catenin in the Colorado Potato Beetle



URL: https://tinyurl.com/y8688kgt

USDA

Example workflow: alpha-catenin in the Colorado Potato Beetle



URL: https://tinyurl.com/y8688kgt

USDA

Post-Annotation QC

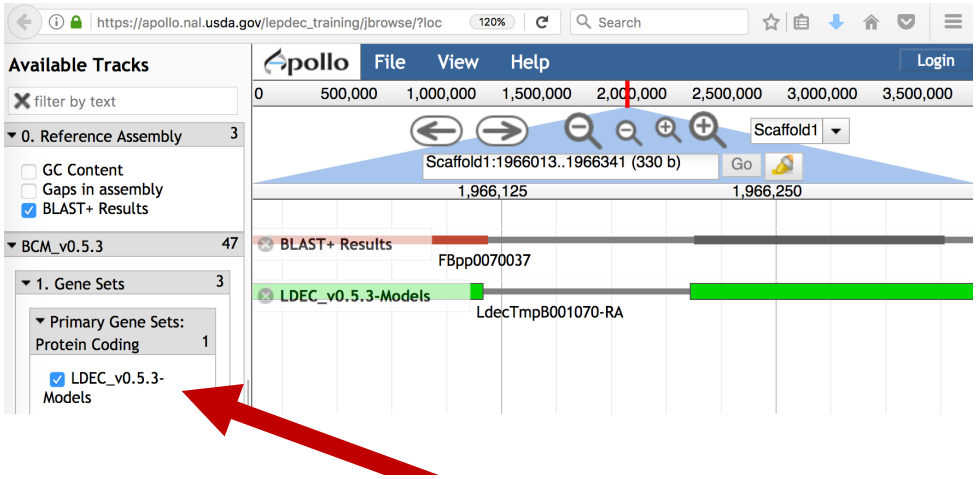
- Manual annotations are run through our Quality Control pipeline
- Some issues need manual intervention
  - Missing required fields
  - Complex splits/merges
  - Incomplete models and those abandoned in process
- Some issues can be automatically corrected
- Iterative process
  - Models requiring inspection are referred back to curators
  - After resolution models are screened again to screen for additional issues

OGS (Official Gene Set) Generation

- An Official Gene Set is the gene set chosen by the community to be the representative set of gene models for that organism
- Our system takes a single existing gene set and incorporates the validated manual annotations
- The gene set may be a previous OGS or other gene set (e.g. Maker models)
- Manual curations are used to
  - Update models
  - Flag models for removal from the final set
- The resulting set is then tested for errors and once approved, disseminated to the community

OGS (Official Gene Set) Generation

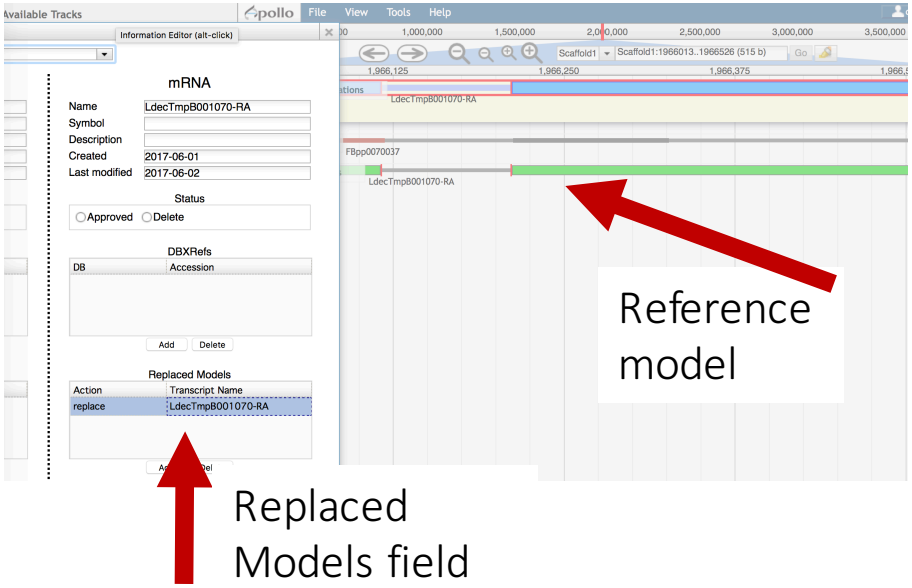
- Requirements:
  - Designate a ‘reference gene set’ prior to the start of the annotation period
  - Use the ‘Replaced Models field’ during the manual annotation process



Reference model track

The i5k Workspace ‘Replaced Models’ field

- Accesible via the Information Editor
- Enter the name or ID of the ***reference*** gene model that your manually curated model replaces.
- Information is used to merge your annotation with reference gene set to make an OGS (Official Gene Set)
- More information:
  - https://i5k.nal.usda.gov/apollo-replaced-models-field-explanations-and-examples



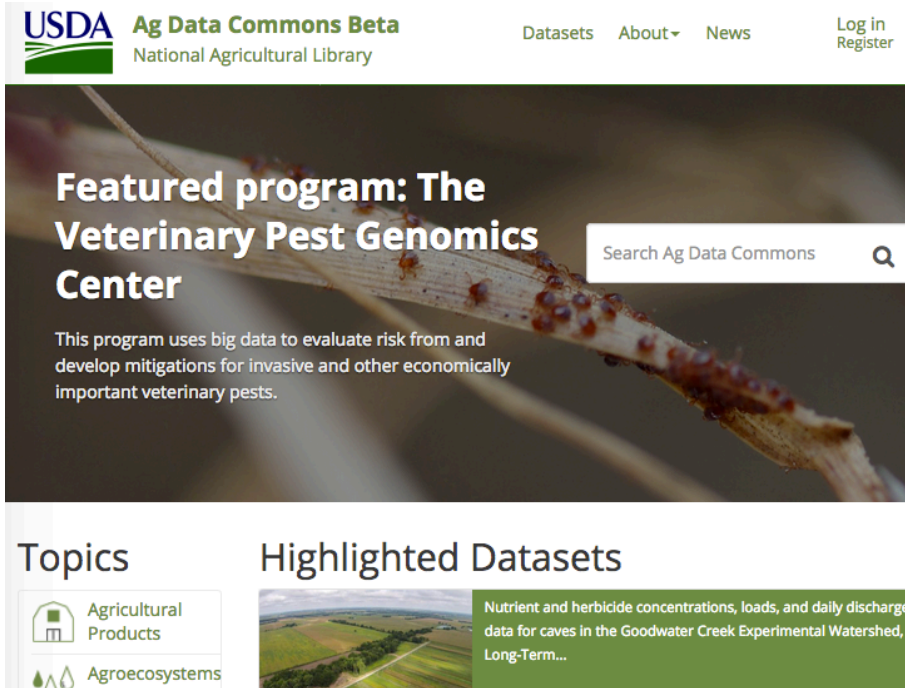
Overview

1. Background: What is the i5k Workspace?
2. Submitting data
3. Finding data at the i5k Workspace
    1. General search
    2. Data downloads
    3. BLAST
    4. Clustal(s)
    5. HMMER
    6. Jbrowse/Apollo
4. Improving data at the i5k Workspace via community annotation
    1. See Monica Munoz-Torres' workshop for full use of Apollo
5. Other data management resources at the NAL – the Ag Data Commons

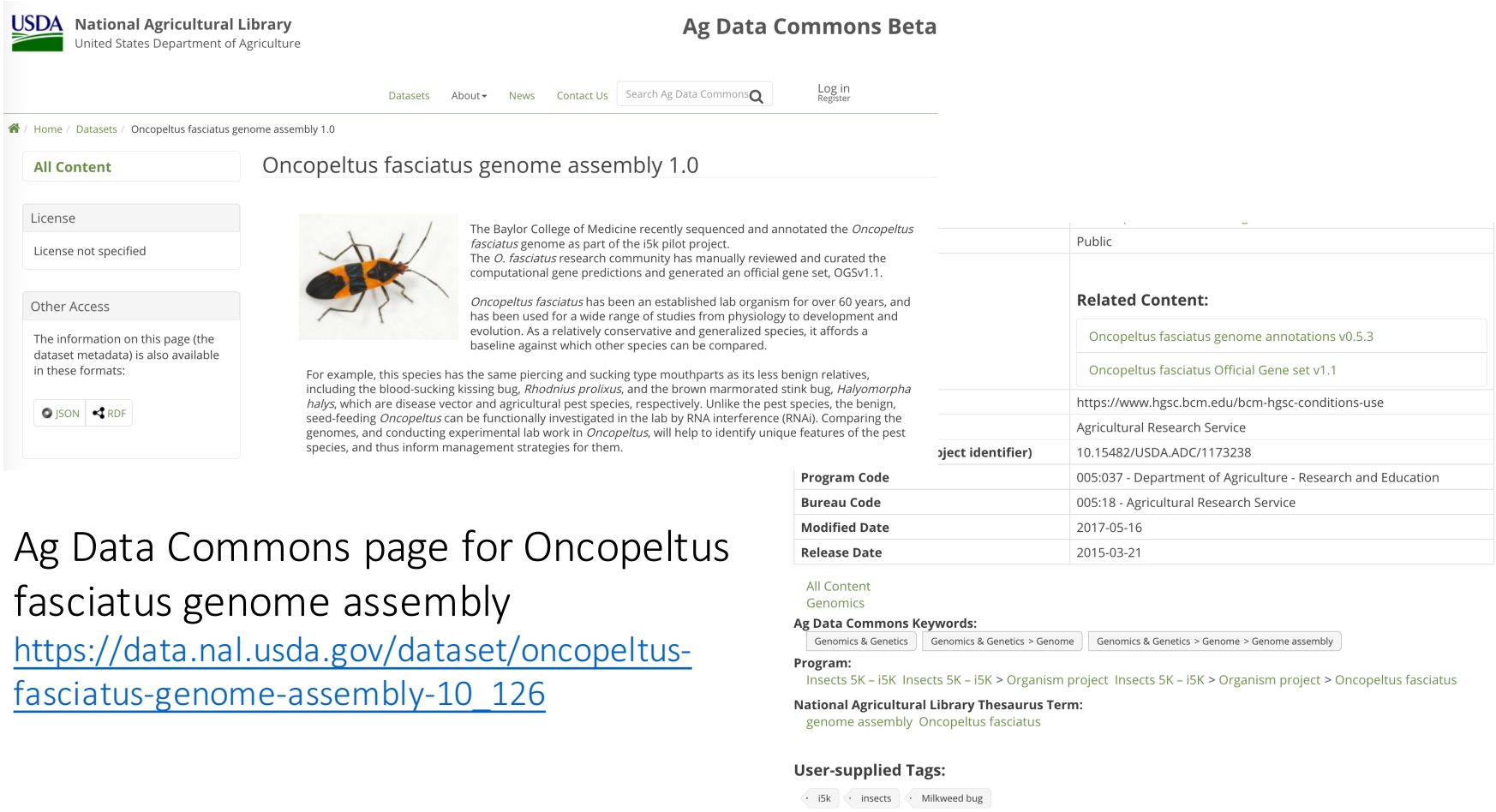
USDA

Other resources at the NAL: the Ag Data Commons

- Hosts any dataset funded by the USDA
- Landing page
- Citable DOI
- https://data.nal.usda.gov/
- 9 i5k datasets already available



Other resources at the NAL: the Ag Data Commons



Ag Data Commons page for Oncopeltus fasciatus genome assembly https://data.nal.usda.gov/dataset/oncopeltus-fasciatus-genome-assembly-10_126

Feedback?

- We LOVE feedback.
- If you have comments, suggestions, critiques, get in touch:
  - https://i5k.nal.usda.gov/contact
  - Monica.poelchau@ars.usda.gov
  - Christopher.childers@ars.usda.gov
  - We'll be available at tables outside the poster sessions and throughout the conference

Need more information?

i5k Workspace@NAL:

- https://i5k.nal.usda.gov/
- https://github.com/NAL-i5K/
- Poster during the Friday session

The i5k initiative:

- New website: http://i5k.github.io/

Ag Data Commons:

- https://data.nal.usda.gov/

USDA