# The i5k Workspace@NAL

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



### Overview

- 1. Background: What is the i5k Workspace?
- 2. Submitting data
- 3. Finding data at the i5k Workspace
  - General search/Content types
  - Data downloads
  - 3. BLAST
  - 4. Clustal(s)
  - 5. HMMER
  - 6. Jbrowse/Apollo
- 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 §5

### 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.
  - 15k 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 laborintensive 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: <u>https://i5k.nal.usda.gov/long-term-i5k-workspace-</u> project-management



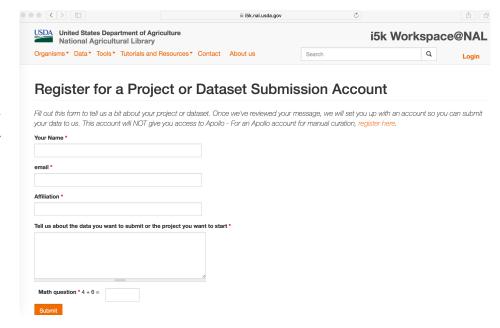
# Other things to consider before submitting

- 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

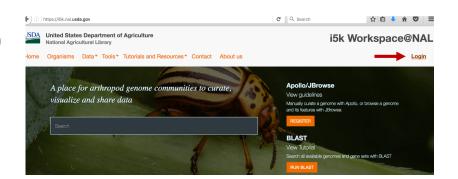
- Apply for a dataset submission account: <a href="https://i5k.nal.usda.gov/register/project-dataset/account">https://i5k.nal.usda.gov/register/project-dataset/account</a>
- For this workshop, you can use
  - Username: demo
  - Password: demo123





## 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/re quest-project
- We'll review your submission and will get in touch with you



National Agricultural Library		iok workspace@ivAL	
Organisms * Data * Tools * Tutorials and Resources * Contact About us	Search	Q My Account	
Data / Submit Data / Request a new i5k Workspace Project			
Request a new i5k Workspace Proje	ct		
Thank you for your interest in submitting your genome project to questions to help us decide if the resources at the i5k Workspace management and long-term management policy documents for is long-term data management policy.	e are a good fit for your pr	oject. Refer to our <mark>data</mark>	
Genus *			
Species •			
NCBI Taxonomy ID *			
Common Name •			
Is the genome assembly already hosted at another genome portal, or is there another gen	ome portal that would also be		
appropriate to host your dataset (e.g. VectorBase, HGD)? *	one portar that would also be		

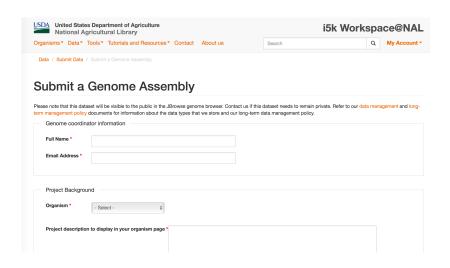
iEk Workensoo@NAI



USDA United States Department of Agricultur

# Submit your genome assembly

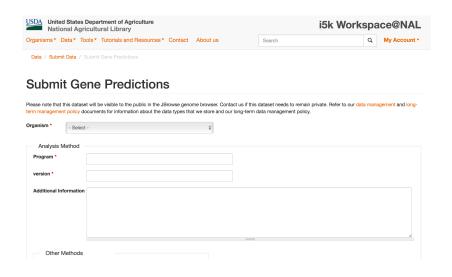
- All information submitted through this form will be reformatted for display at the i5k Workspace (except for email address and file md5sum)
- - https://i5k.nal.usda.gov/ datasets/assembly-data





# Submit gene predictions

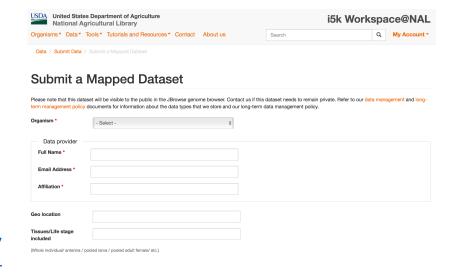
- All information submitted through this form will be re-formatted for display at the i5k Workspace (except for email address and file md5sum)
- Under menu bar, select 'Data -> Submit data -> Submit Gene Predictions'
  - https://i5k.nal.usda.gov/ datasets/geneprediction





## 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
- Other data management resources at the NAL the Ag Data Commons

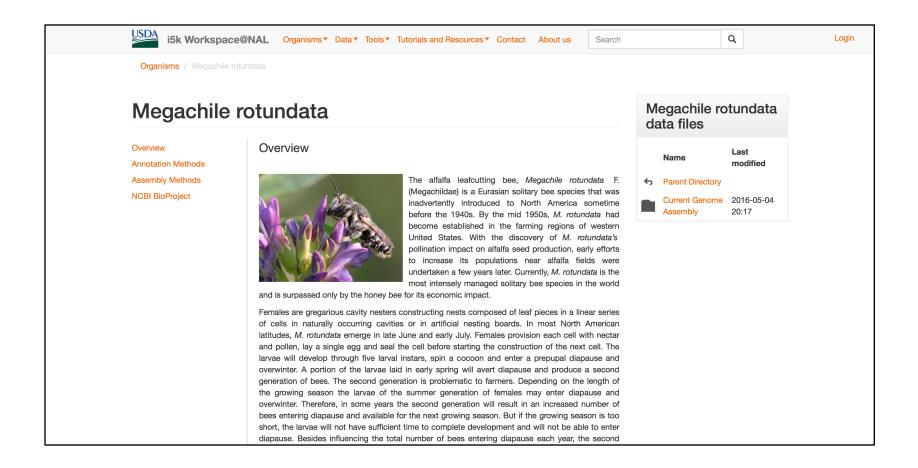


# Finding Data at the i5k Workspace

- 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		

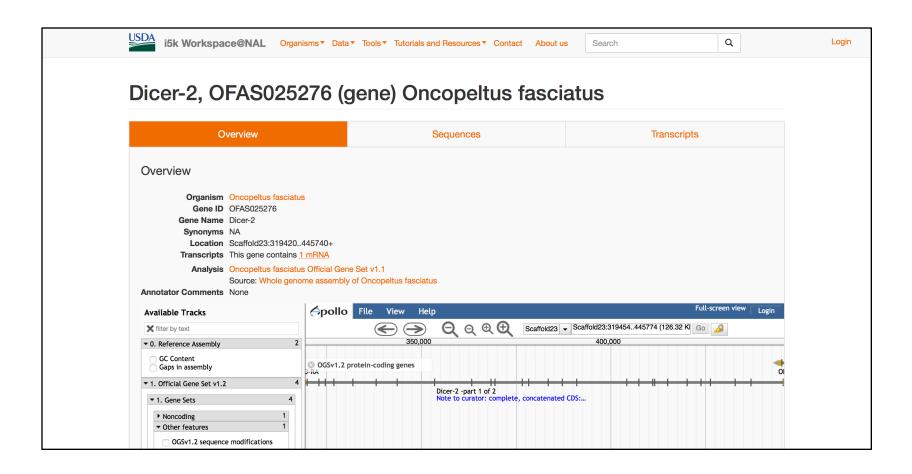
#### **Statistics**

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

NAL Home | USDA.gov | Agricultural Research Service | Plain Language | FOIA | Accessibility Statement | Information Quality | Privacy Policy | Non-Discrimination Statement | USA.gov | White House Please cite the use of our resources: doi: 10.1093/nar/gku983

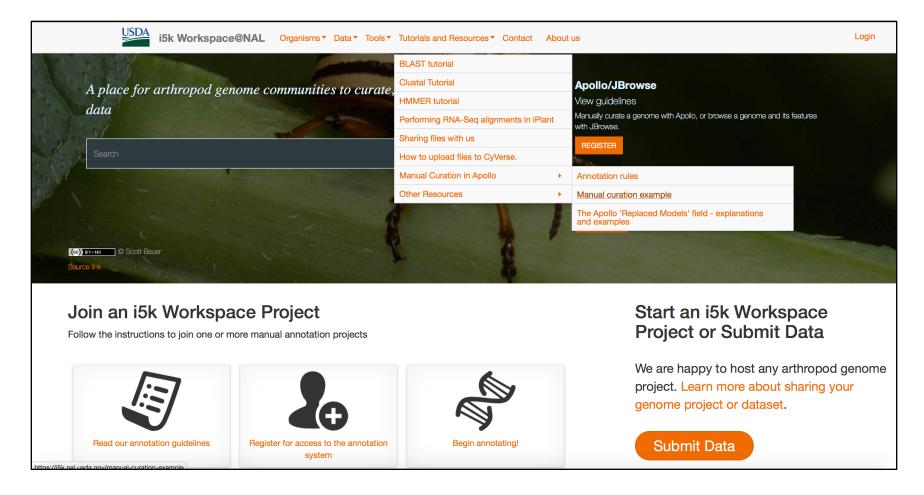


# Gene pages (Official gene sets only)





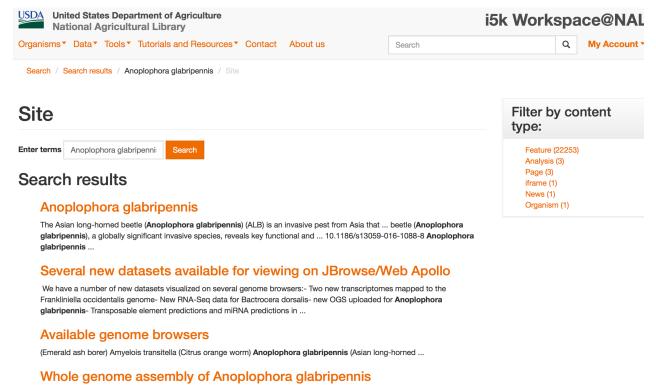
### **Tutorials**





# Finding Data at the i5k Workspace

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





## Bulk data downloads for full files

- From menu, select
   'Data -> Data
   Downloads'
  - https://i5k.nal.usda.gov/ content/datadownloads, or
  - https://i5k.nal.usda.gov/ data/



### Data downloads

#### **CONDITIONS OF USE:**

Many of the genomes, predicted gene sets, and RNA-Seq data hosted on the i5k Workspace@ rapid research on individual genes prior to genome analysis publication. These data are covere producers make the data available and state their intent to publish analyses, the data users ask journals and reviewers ensure that articles are published following the guidelines. Please contain these sequences in published analyses. Get in touch with the i5k Workspace@NAL if you have

#### **DATA DOWNLOADS:**

```
expand all

Arthropoda

aettum-(Aethina_tumida)

agrpla-(Agrilus_planipennis)

amytra-(Amyelois_transitella)

anogla-(Anoplophora_glabripennis)

Current Genome Assembly

1.Genome Assembly

2.Official or Primary Gene Set

OGS_v1_2

agla_OGS_v1_2.gff3.gz

agla_OGS_v1_2.cDNA.fa

agla_OGS_v1_2.cds.fa

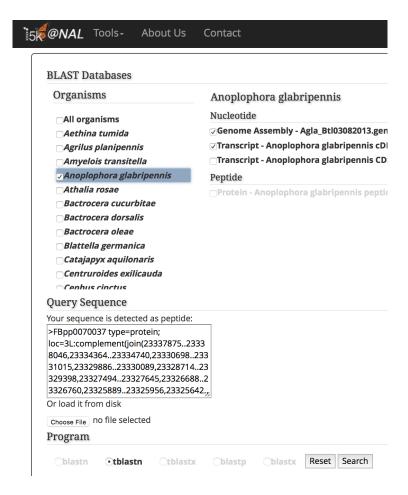
agla_OGS_v1_2.peptide.fa

3.Additional Gene Sets and Annotation Projects
```



## Sequence Search — BLAST+

- From menu, select 'Tools
   -> BLAST'
  - https://i5k.nal.usda.gov/w ebapp/blast/
- Tutorial:
  - https://i5k.nal.usda.gov/co ntent/blast-tutorial
- Example query:
  - http://flybase.org/cgibin/getseq.html?source=d mel&id=FBpp0070037&ch r=3L&dump=PrecompiledF asta&targetset=translation





## Sequence Search – BLAST+ Result



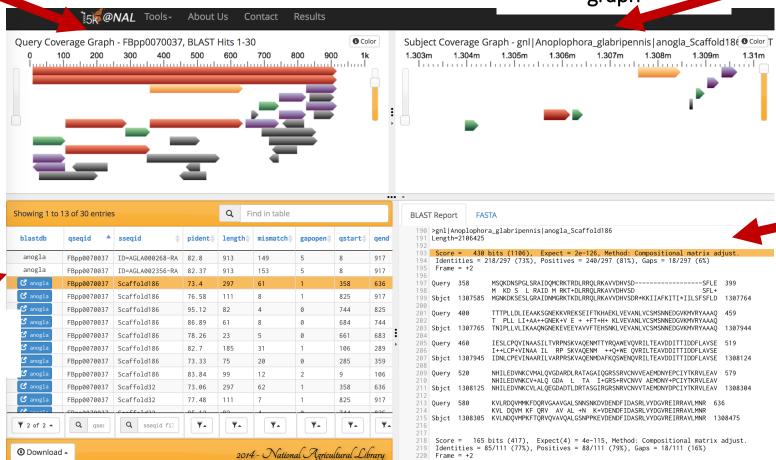
Tabular

result

Subject coverage graph

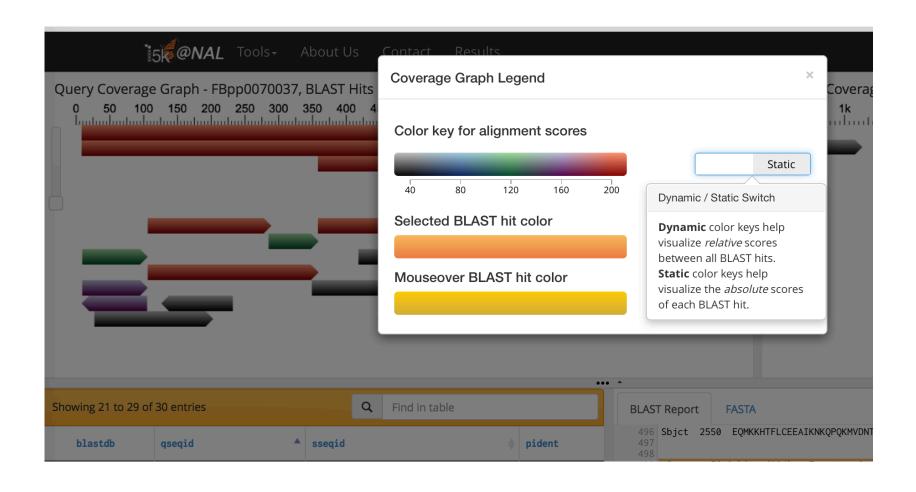
Raw

results

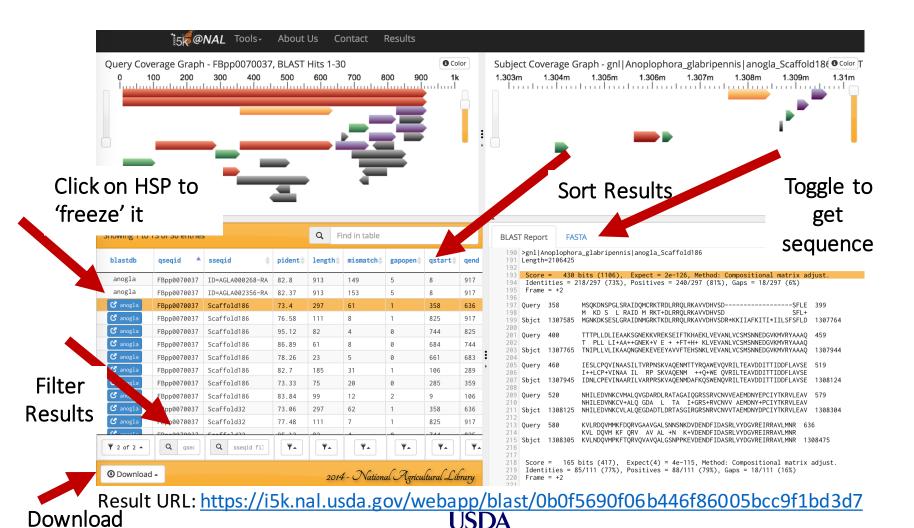


Result URL: https://i5k.nal.usda.gov/webapp/blast/0b0f5690f06b446f86005bcc9f1bd3d7

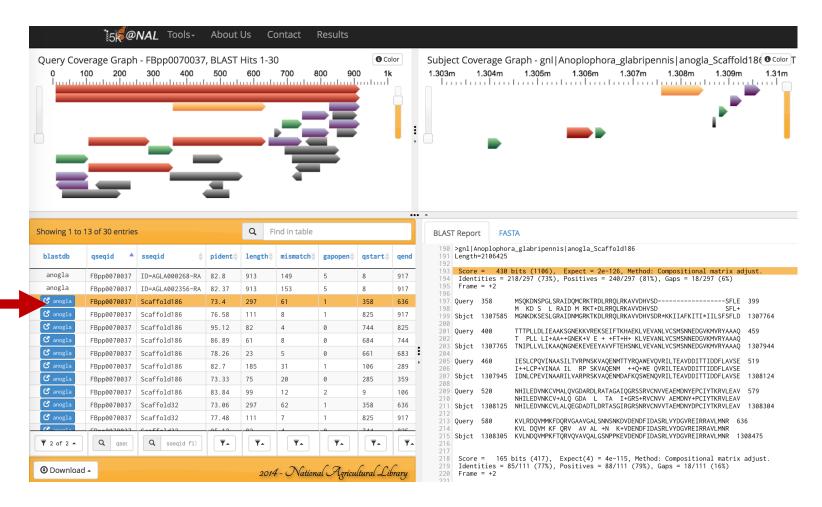
## Sequence Search — BLAST+ Result



# Sequence Search — BLAST+ Result

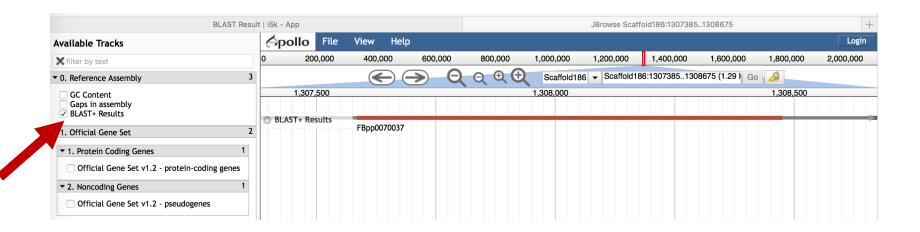


# Sequence Search — links to JBrowse





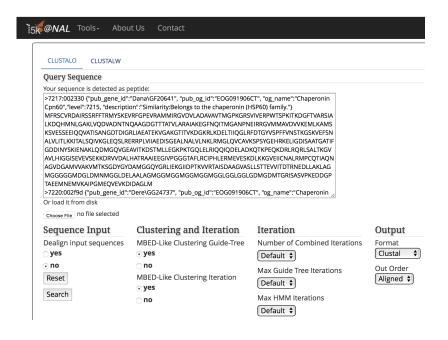
# Sequence Search — links to JBrowse





# 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=EOG09190 6CT&level=&species=&u niversal=&singlecopy=





# Sequence alignment – ClustalW and Clustal Omega



#### **CLUSTAL Success**

#### Download

Alignment

Submission Details

#### **Report Details**

CLUSTAL O(1.2.3) multiple sequence alignment 7222:0004a3 MFRSYVRE-SIRSSRAFARAYSKAVTFGAEARARMLHGVDVLADAVAVTLGPKGRCVILE 7230:00256a MFRSYVRK-AVRSSRAFARAYSKDVAFGADARARMLRGVDVLTDAVAVTLGPKGRSVILE 7244:0019f0 MFRSYVRE-AVRSSRAFARAYSKDVAFGAEARARMLRGVDMLTDAVAVTMGPKGRSVILE 7260:0029e2 MFRFFARDAAVCTGRNLCRAYSKEVRFGPEVRALMIRGVDILADAVAVTMGPKGRNVIVE 7234:0021c7 MFRHCVRG-ALRGNRNFLRLYSKDVRFGNEARSMMIRGVDLLADAVAVTMGPKGRSVIVE 7237:002b4f MFRHCVRG-VT.RGNRNFT.RT.YSKDVRFGNEARSMMTRGVDT.T.ADAVAVTMGPKGRSVTVF 7217:002330 MFRSCVRD-ATRSSRFFTRMYSKEVRFGPEVRAMMTRGVDVTADAVAVTMGPKGRSVTVE 28584:0003f1 MFRSCVSE-AIISSRCFARMYSKEVRFGPGVRALMIRGVDVLADAVAVTMGPKGRSVIVE 7220:002f9d MFRSCVPK-ATSSSRCFARMYSKEVRFGSGVRATMTRGVDVTADAVAVTMGPKGRSVTVE 7245:001764 MFRSCVPK-AISSSRCFARMYSKDVRFGSGVRALMIRGVDVLADAVAVTMGPKGRSVIVE 7227:000486 MFRSCVPK-AITSSRCFARMYSKDVRFGSGVRAMMIRGVDILADAVAVTMGPKGRSVIVE 7238:0019ed MFQSCVPK-AITSSRCFARMYSKDVRFGTGVRALMIRGVDVLADAVAVTMGPKGRSVIVE 7240:002709 MFRSCVPK-ATTSSRCFARMYSKDVRFGTGVRST.MTRGVDVT.ADAVAVTMGPKGRSVTVE 

#### Result URI:

https://i5k.nal.usda.gov/webapp/clustal/ed00819ab 40441ca959eacdccb78c0f5

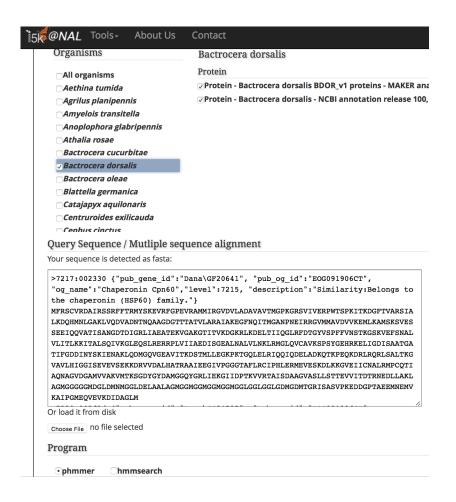
```
7222:0004a3
                 -MDEMGAMDGI-----SKAAEMNDAVKSIPGMENVEVHDIDSSQ
7230:00256a
                 GMDDVGEICNK-----SKAAEMNEAVOSIPGMEDVTVHDIDSTO
7244:0019f0
7260:0029e2
7234:0021c7
7237:002b4f
7217:002330
                 GLGGLGDMGDMTGRI-----SASVPKEDDGPTAEEMNEMVKAIPGMEOVEVKDIDAGI
28584:0003f1
                 ----GGMGGMGGGFGGMGGGGGGMSASKSDGPTAEEMNEMVKAIPGMEOVEVRDIDSGM
7220:002f9d
                 GMGGMGAMGGMGGGFGGMGGGGGMSASSSSDGPTAEEMNEMVKAIPGMEOVEVRDIDSGM
7245:001764
                 CMCCMCAMCCMCACFCCMCCCCCMSASSSSDCPSAFEMNEMVKATPCMEOVEVRDTDSC
7227:000486
                 ----MGGMGGMGGGFGGMGAGGGMSASASNDGPTAEEMNEMVKAIPGMEOVEVRDIDSGN
7238:0019ed
                 ---MGGMGGMGGGFGGMGAGGGMSASASNEGPTAEEMNEMVKAIPGMEOVEVRDIDSGN
7240:002709
                 ----MGGMGGMGGGFGGMGAGGGMSASASNEGPTAEEMNEMVKAIPGMEQVEVRDIDSGM
7222:0004a3
7230:00256a
7244:0019f0
7260:0029e2
7234:0021c7
7237:002b4f
                 MQ
7217:002330
28584:0003f1
7220:002f9d
7245:001764
7227:000486
7238:0019ed
7240:002709
```





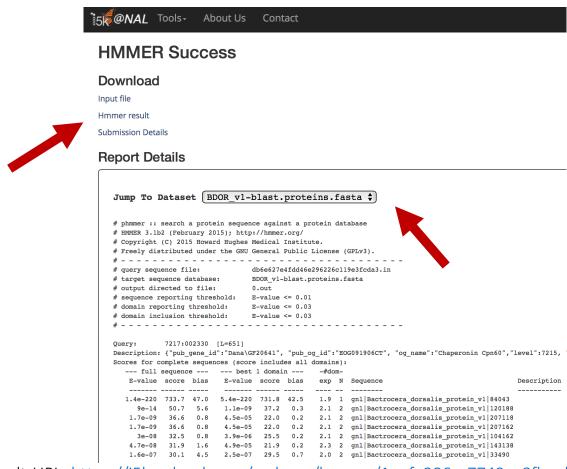
## Sequence search – HMMER

- From menu, select
   'Tools -> HMMER (beta)'
  - https://i5k.nal.usda.gov/ webapp/hmmer/
- Example query sequences (fasta, restrict to <10):</li>
  - http://www.orthodb.org /fasta?query=EOG09190 6CT&level=&species=&u niversal=&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-genomebrowsers



Tools / JBrowse/Apollo / Available genome browsers

### Available genome browsers

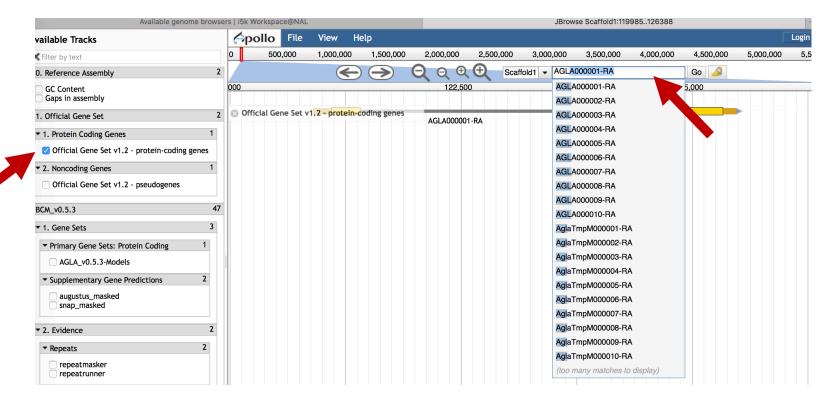
Click on a link to open a genome browser for a genome project, and to access Web Apollo. Browsing does read to log in to Web Apollo from the genome browser, click on the 'login' button at the top right of the screen. If Apollo here.

- · Aethina tumida (Small hive beetle)
- Agrilus planipennis (Emerald ash borer)
- Amyelois transitella (Citrus orange worm)
- Anoplophora glabripennis (Asian long-horned beetle)
- Athalia rosae (Turnip sawfly)
- Bactrocera cucurbitae (Melon Fruit Fly)
- Bactrocera dorsalis (Oriental Fruit Flv)
- Bactrocera oleae (Olive Fruit Fly)
- Blattella germanica (German cockroach)
- Catajapyx aquilonaris (Silvestri's Northern Forcepstail)
- Centruroides exilicauda (Bark scorpion)
- Cephus cinctus (Wheat stem sawfly)
- · Ceratitis capitata (Mediterranean fruit fly)
- Cimex lectularius (Bed bug)
- Copidosoma floridanum (NA)



# 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



# 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/
  - <a href="https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project">https://i5k.nal.usda.gov/content/rules-web-apollo-annotation-i5k-pilot-project</a>
  - https://i5k.nal.usda.gov/manual-curation-example



# Community curation at the i5k Workspace

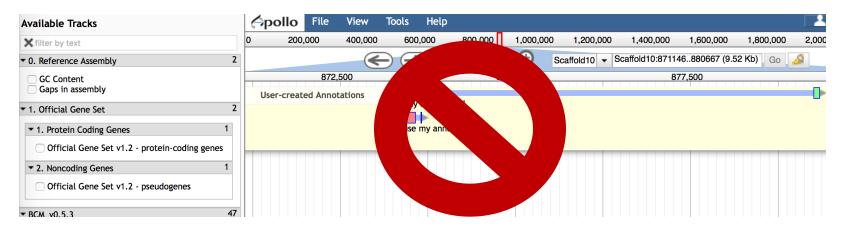
Our support for community curation includes:

- 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)

Community Genome Manual building: sequencing, annotation via Conference calls assembly and Apollo and training annotation Official Gene Set Manual generation (Merge General annotation of manual QC (NAL) 'freeze' annotations and reference gene set)

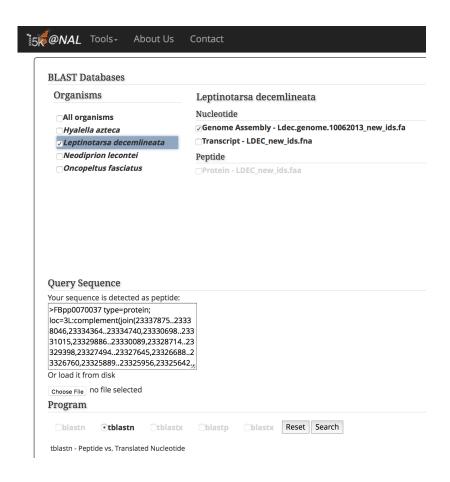


### Some Apollo notes

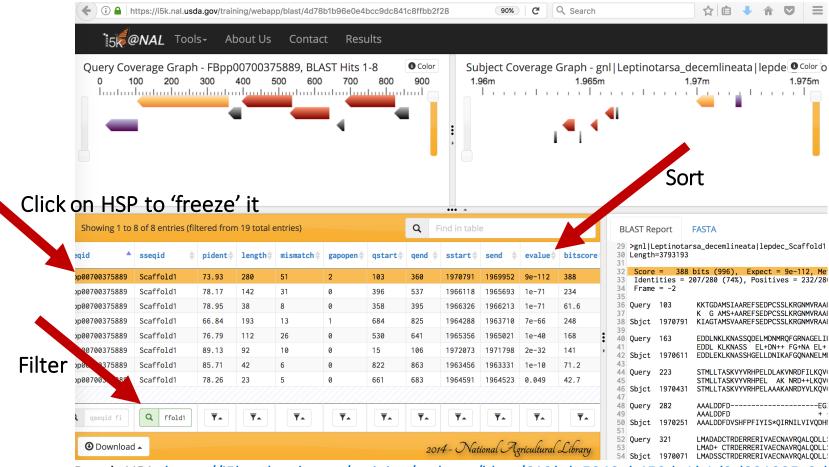
- We're still using Apollo1 Apollo2 has a slightly different interface
- Here, we'll use our 'Training' applications
- Apollo credentials for training applications:
  - Username: demo
  - Password: demo
- To annotate on an actual project, you'll need to register first:
  - From menu, select 'Tools ->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



- From menu, select
   'Tools -> Training tools > Training BLAST'
  - https://i5k.nal.usda.gov/ training/webapp/blast
- Query sequence:
  - http://flybase.org/cgibin/getseq.html?source =dmel&id=FBpp007003 7&chr=3L&dump=Preco mpiledFasta&targetset= translation







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

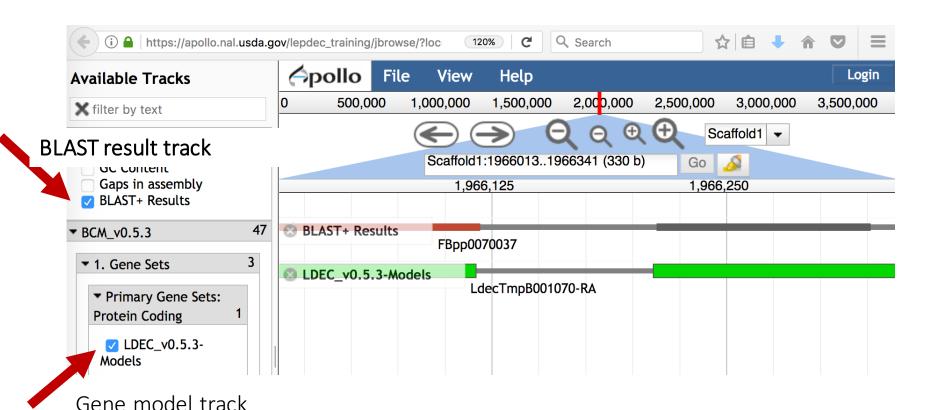


- To view HSP in the genome browser:
  - Go to result table on bottom left
  - click on the blue box to the left of the best HSP result in the 'blastdb' column





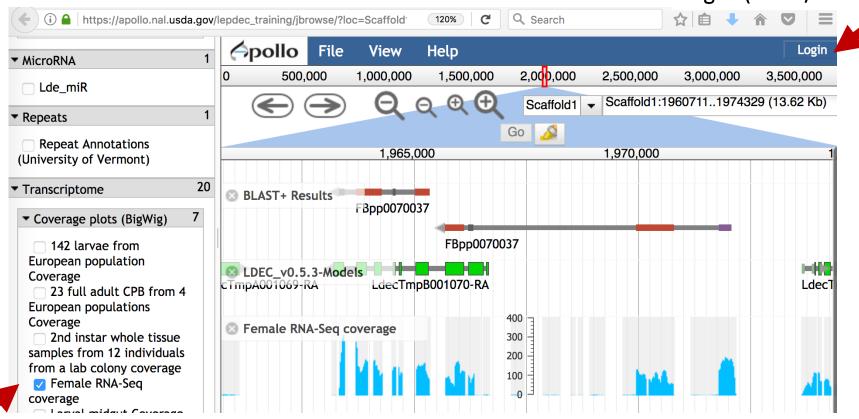




URL: <a href="https://tinyurl.com/ybf4ehld">https://tinyurl.com/ybf4ehld</a>



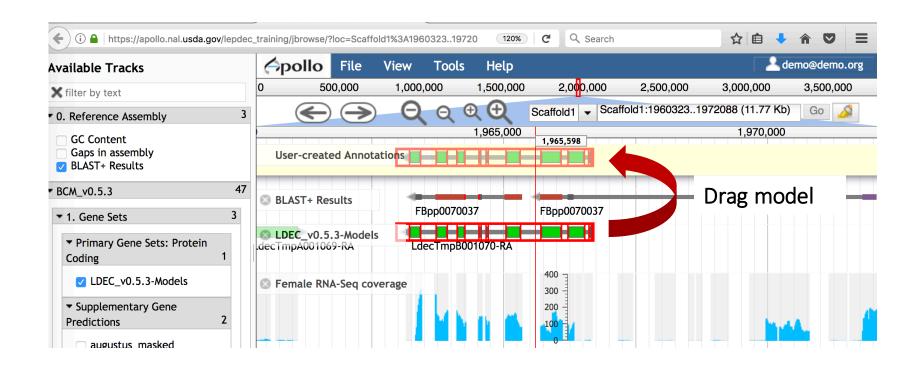
Log in (demo/demo)



RNA-Seq evidence tracks

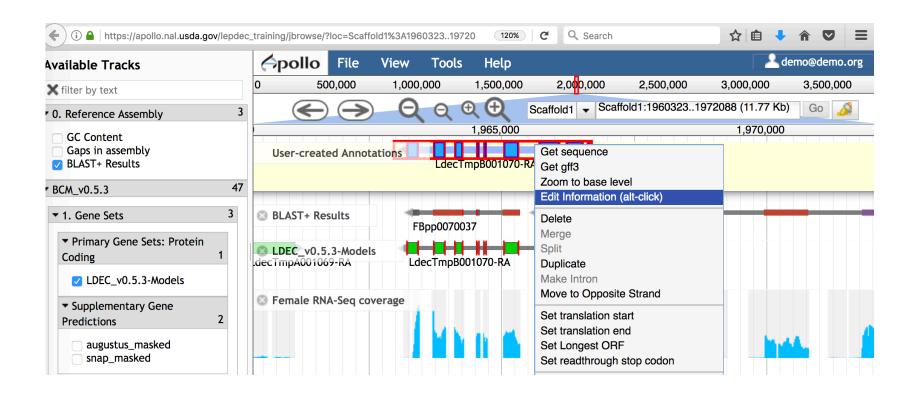
URL: <a href="https://tinyurl.com/y8688kgt">https://tinyurl.com/y8688kgt</a>





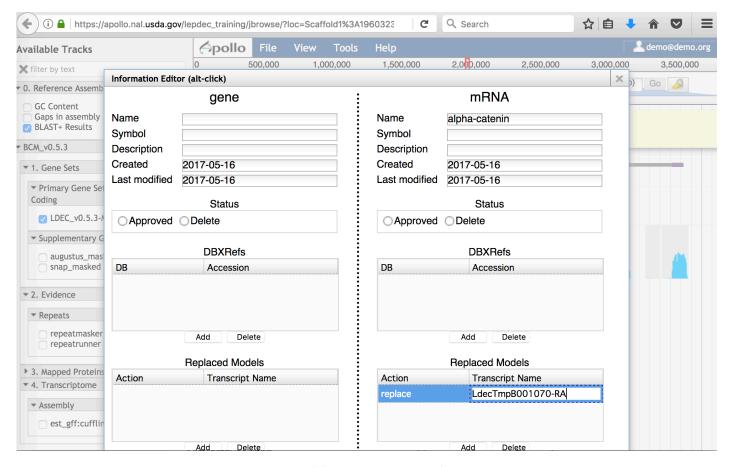
URL: https://tinyurl.com/y8688kgt





URL: https://tinyurl.com/y8688kgt





URL: <a href="https://tinyurl.com/y8688kgt">https://tinyurl.com/y8688kgt</a>



### 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



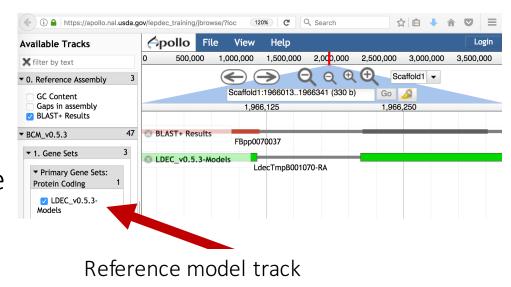
### OGS (Official Gene Set) Generation

- 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

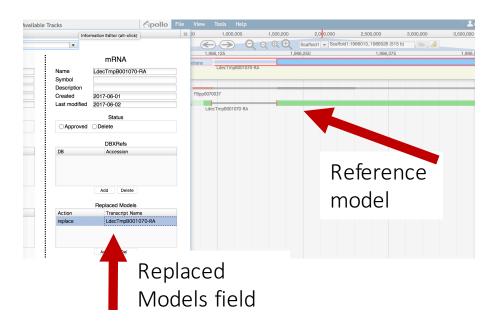
- Requirements:
  - Designate a 'reference gene set' prior to the start of the annotation period
  - Use the 'Replaced Models field' during the manual annotation process





## 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/apoll o-replaced-models-fieldexplanations-and-examples





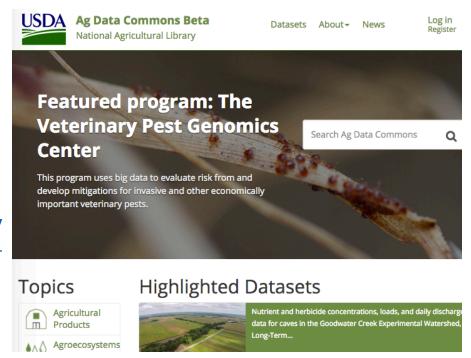
### Overview

- 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



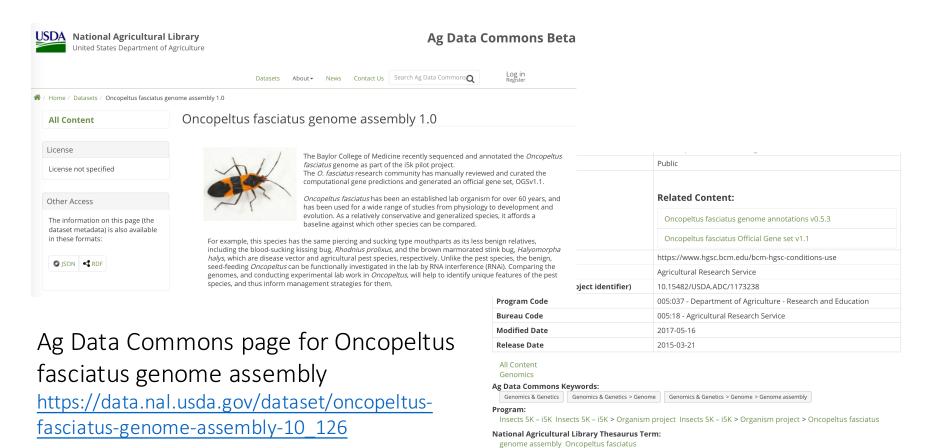
# Other resources at the NAL: the Ag Data Commons

- 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





**User-supplied Tags:** 

### 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: <a href="http://i5k.github.io/">http://i5k.github.io/</a>

#### Ag Data Commons:

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