Genome project management resources at the National Agricultural Library

Chris Childers and Monica Poelchau

USDA-ARS, National Agricultural Library

USDA

So you have a genome project. Where will you store your data?

- Make your data available through NCBI when applicable (or other INSDC organizations).
- To make your data even more useful for your community, consider also making it available in a taxon-specific repository.
- Advantages for you:
  - Greater visibility for your dataset
  - 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-

So you have a genome project. Where will you store your data?

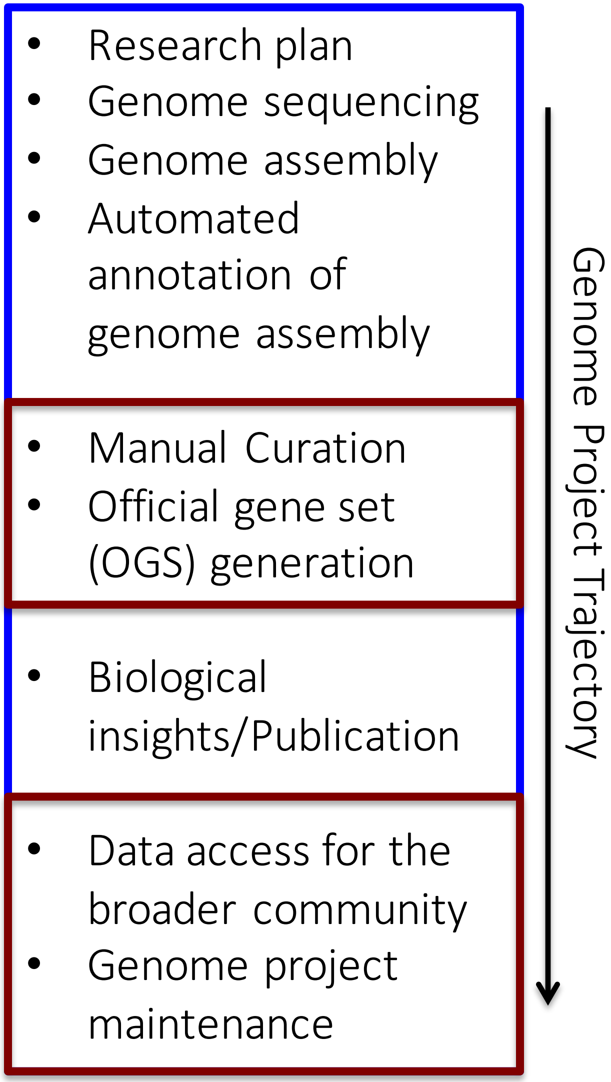
- Advantages for the scientific community:
  - Helps facilitate knowledge discovery for humans (and sometimes machines);
  - Easier to find data for comparative analyses;
  - Promotes reproducible research;
  - General repositories (e.g. GenBank) may not meet the needs for storing all data types, in particular for non-standard data types (e.g. phenotypic data).

Genome data management resources for arthropods – how to choose

- What species is the data from?
  - Many taxon-specific genome databases are here at this workshop
- What kind data do you have?
  - Raw data, genome assemblies, transcriptome assemblies, gene annotations, can and should all be stored at NCBI (or other INSDC organization)
  - Some or all of these data types can also be made accessible at genome databases (just ask)
  - Generic repositories (e.g. Dryad, Ag Data Commons) can be used for data types that don't fit the mold

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



- Research plan
- Genome sequencing
- Genome assembly
- Automated annotation of genome assembly

- Manual Curation
- Official gene set (OGS) generation

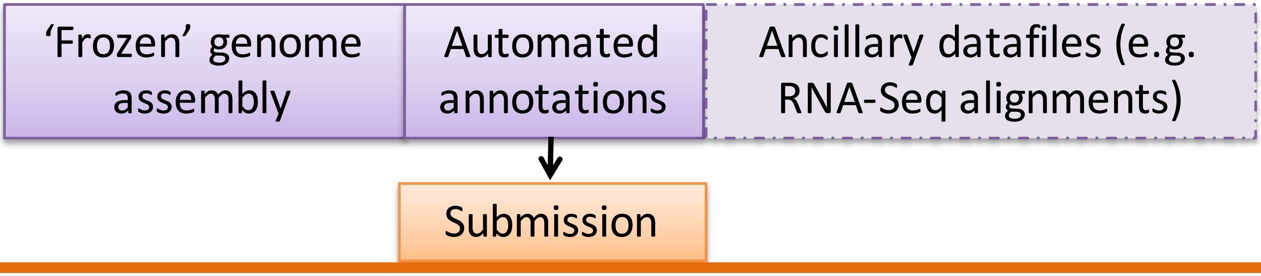
- Biological insights/Publication

- Data access for the broader community
- Genome project maintenance

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
- Brief introduction to i5k at the beginning of the i5k session on Thursday



'Frozen' genome assembly

Automated annotations

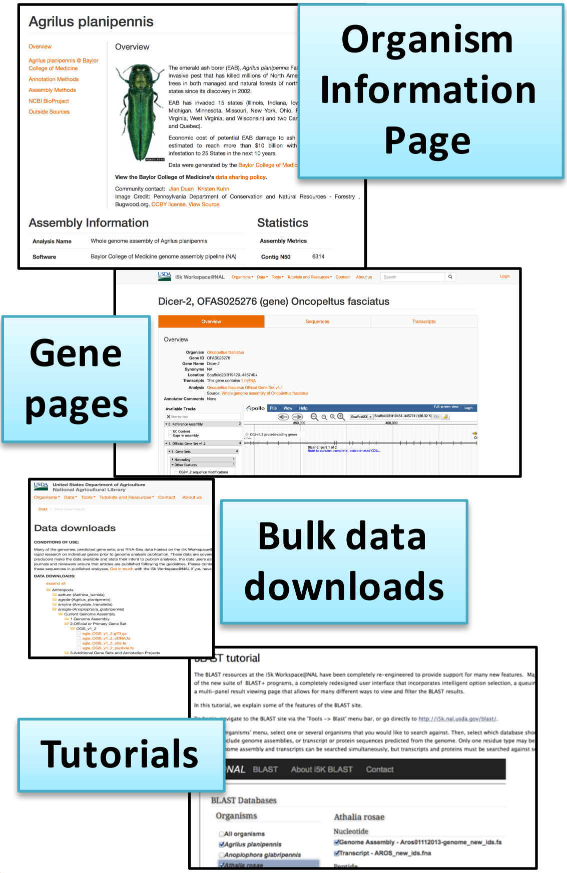
Ancillary datafiles (e.g. RNA-Seq alignments)

Submission

i5K Workspace@NAL

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

Resources



Agrilus planipennis

Overview

Agrilus planipennis @ Baylor College of Medicine

Annotation Methods

Assembly Methods

NCBI BioProject

Outside Sources

Overview

The emerald ash borer (EAB), *Agrilus planipennis* Fa… invasive pest that has killed millions of North Ame… trees in both managed and natural forests of north… states since its discovery in 2002.

EAB has invaded 15 states (Illinois, Indiana, Io… Michigan, Minnesota, Missouri, New York, Ohio, P… Virginia, West Virginia, and Wisconsin) and two Ca… and Quebec).

Economic cost of potential EAB damage to ash … estimated to reach more than $10 billion with… infestation to 25 States in the next 10 years.

Data were generated by the Baylor College of Medic…

View the Baylor College of Medicine's data sharing policy.

Community contact: Jian Duan Kristen Kuhn

Image Credit: Pennsylvania Department of Conservation and Natural Resources - Forestry , Bugwood.org. CCBY license. View Source.

Assembly Information

Analysis Name	Whole genome assembly of Agrilus planipennis
Software	Baylor College of Medicine genome assembly pipeline (NA)

Statistics

Assembly Metrics	
Contig N50	6314

USDA i5k Workspace@NAL Organisms Data Tools Tutorials and Resources Contact About us Search Login

Dicer-2, OFAS025276 (gene) Oncopeltus fasciatus

Overview | Sequences | Transcripts

Overview

Organism	Oncopeltus fasciatus
Gene ID	OFAS025276
Gene Name	Dicer-2
Synonyms	NA
Location	Scaffold23:319420..445740+
Transcripts	This gene contains 1 mRNA
Analysis	Oncopeltus fasciatus Official Gene Set v1.1 Source: Whole genome assembly of Oncopeltus fasciatus
Annotator Comments	None

USDA United States Department of Agriculture National Agricultural Library

Organisms Data Tools Tutorials and Resources Contact About us

Data / Data downloads

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 cover… producers make the data available and state their intent to publish analyses, the data users as… journals and reviewers ensure that articles are published following the guidelines. Please conta… 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

…ST tutorial

The BLAST resources at the i5k Workspace@NAL have been completely re-engineered to provide support for many new features. Ma… of the new suite of BLAST+ programs, a completely redesigned user interface that incorporates intelligent option selection, a queui… a multi-panel result viewing page that allows for many different ways to view and filter the BLAST results.

In this tutorial, we explain some of the features of the BLAST site.

…vigate to the BLAST site via the 'Tools -> Blast' menu bar, or go directly to http://i5k.nal.usda.gov/blast/.

…rganisms' menu, select one or several organisms that you would like to search against. Then, select which database sho… …clude genome assemblies, or transcript or protein sequences predicted from the genome. Only one residue type may be… …ome assembly and transcripts can be searched simultaneously, but transcripts and proteins must be searched against s…

☐ All organisms
☑ *Agrilus planipennis*
☐ *Anoplophora glabripennis*
☑ *Athalia rosae*

☑ Genome Assembly - Aros01112013-genome_new_ids.fa
☑ Transcript - AROS_new_ids.fna

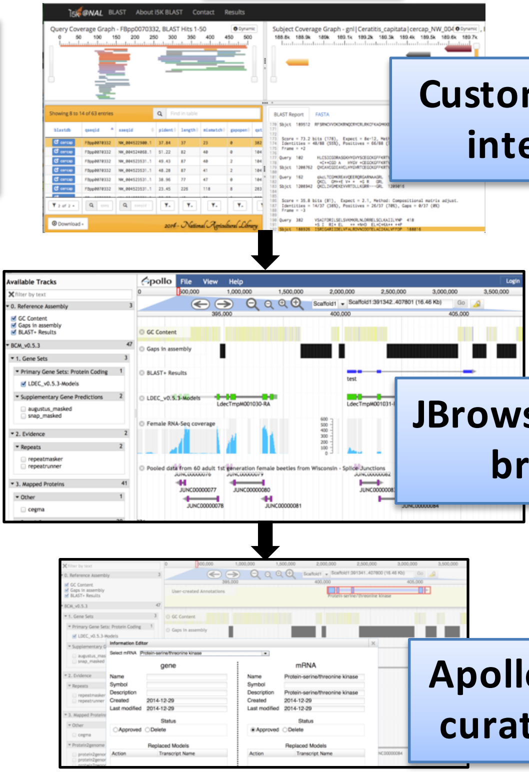
Organism Information Page

Gene pages

Bulk data downloads

Tutorials

Tools



gene		mRNA	
Name		Name	Protein-serine/threonine kinase
Symbol		Symbol	
Description		Description	Protein-serine/threonine kinase
Created	2014-12-29	Created	2014-12-29
Last modified	2014-12-29	Last modified	2014-12-29
Status	○ Approved ○ Delete	Status	● Approved ○ Delete

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)

i5k Workspace content – 57 species and counting

Order	Quantity	Order	Quantity
Amphipoda	1	Hemiptera	7
Araneae	3	Hymenoptera	14
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.)

Community annotation at the i5k Workspace

- What is community annotation?
  - Scientists collectively examine and improve gene models (usually computationally predicted)
- Why annotate?
  - Verify quality of automated gene predictions
  - Improve gene models for specific analyses
  - Link gene models to existing literature and ontologies
- Our community: Over 400 registered annotators have curated over 10,000 gene models using the Apollo software

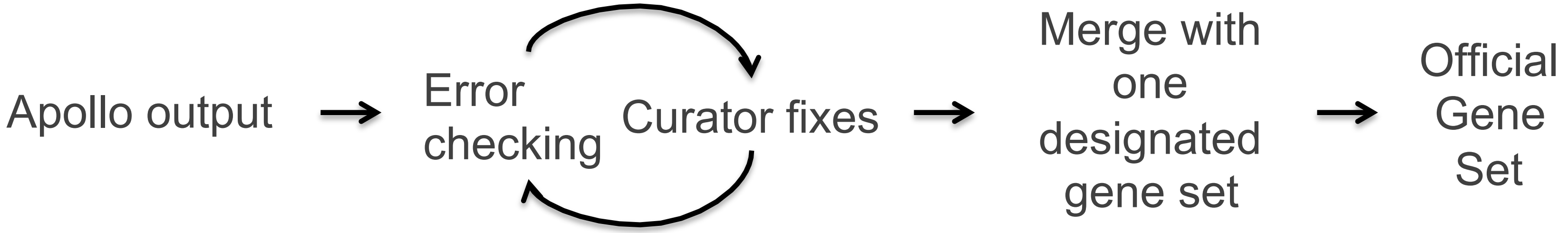
Community annotation at the i5k Workspace

Our support for community annotation 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)

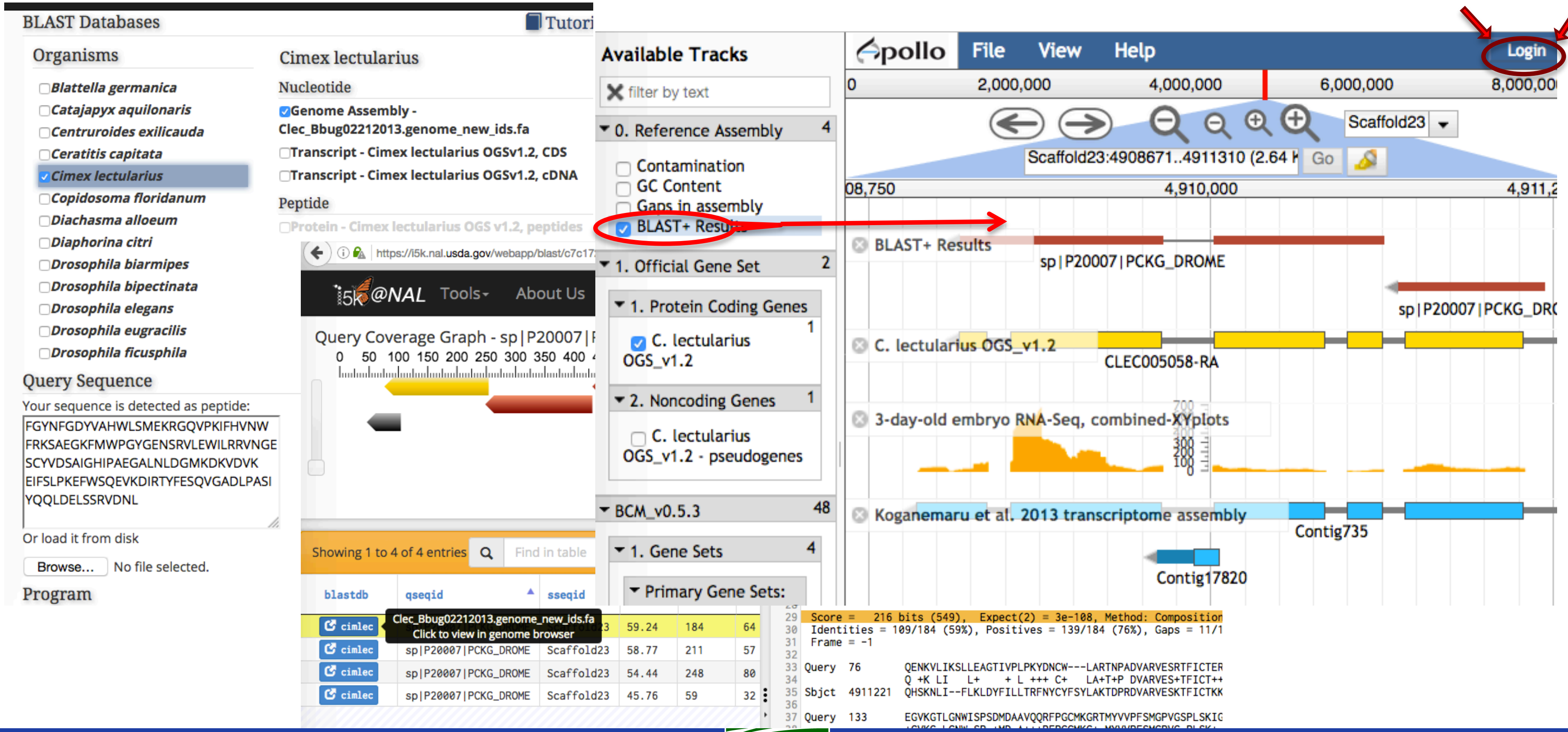
QC and OGS pipeline

- QC program corrects common formatting errors from the curation process
- OGS generation program merges curated models with one designated gene set using curator-supplied information
- Still in development, already 6 OGS's produced (Mei-Ju Chen)



Genome already hosted elsewhere?

- You can also use our tools to query the datasets that we host.



□ *Blattella germanica*
□ *Catajapyx aquilonaris*
□ *Centruroides exilicauda*
□ *Ceratitis capitata*
☑ *Cimex lectularius*
□ *Copidosoma floridanum*
□ *Diachasma alloeum*
□ *Diaphorina citri*
□ *Drosophila biarmipes*
□ *Drosophila bipectinata*
□ *Drosophila elegans*
□ *Drosophila eugracilis*
□ *Drosophila ficusphila*

☑ Genome Assembly - Clec_Bbug02212013.genome_new_ids.fa
□ Transcript - Cimex lectularius OGSv1.2, CDS
□ Transcript - Cimex lectularius OGSv1.2, cDNA
□ Protein - Cimex lectularius OGS v1.2, peptides

Your sequence is detected as peptide:

```
FGYNFGDYVAHWLSMEKRGQVPKIFHVNW
FRKSAEGKFMWPGYGENSRVLEWILRRVNGE
SCYVDSAIGHIPAEGALNLDGMKDKVDVK
EIFSLPKEFWSQEVKDIRTYFESQVGADLPASI
YQQLDELSSRVDNL
```

Or load it from disk

Browse... No file selected.

https://i5k.nal.usda.gov/webapp/blast/c7c17

▾ 0. Reference Assembly 4
□ Contamination
□ GC Content
□ Gaps in assembly
☑ BLAST+ Results
▾ 1. Official Gene Set 2
▾ 1. Protein Coding Genes 1
☑ C. lectularius OGS_v1.2
▾ 2. Noncoding Genes 1
□ C. lectularius OGS_v1.2 - pseudogenes
▾ BCM_v0.5.3 48
▾ 1. Gene Sets 4
▾ Primary Gene Sets:

Scaffold23:4908671..4911310 (2.64 k

Showing 1 to 4 of 4 entries

Find in table

blastdb	qseqid	sseqid			
cimlec	sp|P20007|PCKG_DROME	Scaffold23	59.24	184	64
cimlec	sp|P20007|PCKG_DROME	Scaffold23	58.77	211	57
cimlec	sp|P20007|PCKG_DROME	Scaffold23	54.44	248	80
cimlec	sp|P20007|PCKG_DROME	Scaffold23	45.76	59	32

Clec_Bbug02212013.genome_new_ids.fa
Click to view in genome browser

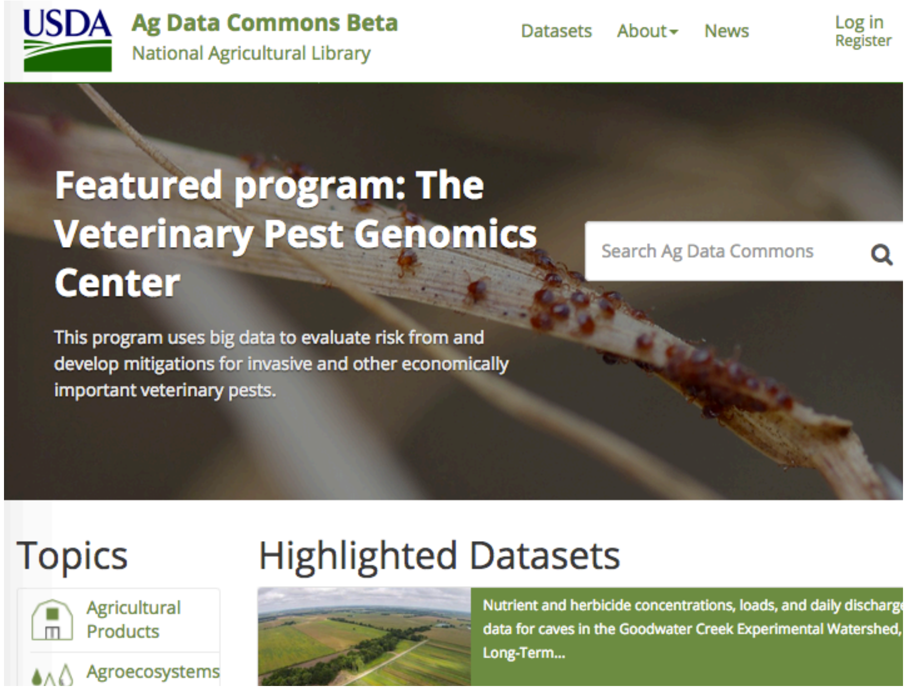
```
 Score =   216 bits (549),  Expect(2) = 3e-108, Method: Composition
 Identities = 109/184 (59%), Positives = 139/184 (76%), Gaps = 11/1
 Frame = -1

Query  76       QENKVLIKSLLEAGTIVPLPKYDNCW---LARTNPADVARVESRTFICTER
                Q +K LI   L+     + L +++ C+    LA+T+P DVARVES+TFICT++
Sbjct  4911221  QHSKNLI--FLKLDYFILLTRFNYCYFSYLAKTDPRDVARVESKTFICTKK

Query  133      EGVKGTLGNWISPSDMDAAVQQRFPGCMKGRTMYVVPFSMGPVGSPLSKIG
```

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/
- Nine i5k datasets already available



Ag Data Commons Beta
National Agricultural Library

Featured program: The Veterinary Pest Genomics Center

This program uses big data to evaluate risk from and develop mitigations for invasive and other economically important veterinary pests.

Topics

Agricultural Products

Agroecosystems

Highlighted Datasets

Nutrient and herbicide concentrations, loads, and daily discharg data for caves in the Goodwater Creek Experimental Watershed, Long-Term...

What we'll talk about tomorrow

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

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/

Acknowledgements

The NAL Team

- Gary Moore
- Susan McCarthy
- Yu-yu Lin
- Mei-Ju Chen

Workspace alumni

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli

i5k Workspace@NAL advisory committee

- Jay Evans
- Kevin Hackett
- Simon Liu
- Ursula Pieper

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