Genome project management resources at the National Agricultural Library

Monica Poelchau/Chris Childers, USDA-NAL
Monica Muñoz-Torres, BBOP/LBNL

Aphid Genomics Symposium. September 25th, 2016

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

- Make your data available through NCBI (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:
  - Greater visibility for your dataset
  - Easier to find data for comparative analyses
  - Value-added tools for searching and browsing, analysis
  - Curation tools to improve annotation quality

The i5k Workspace@NAL

Our focus:

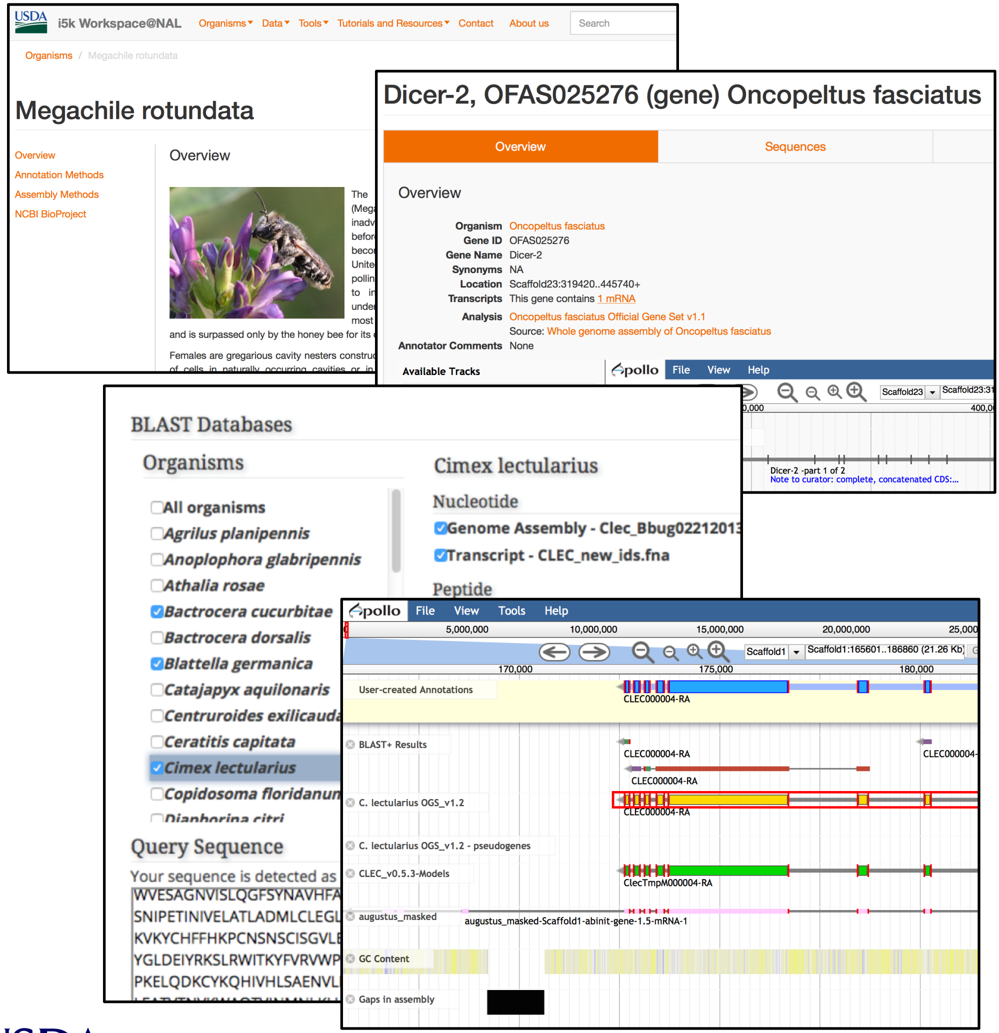
- We support **any** ‘orphaned’ arthropod genome project:
  - Genome assembly needs to be in GenBank/ENA/DDBJ
  - Data should be open access (no private repositories)
- We enable and support **community curation.**

Our background:

- Originally set up to support genomes sequenced as part of the i5k initiative
- To learn more about the i5k initiative, visit **Booth #320**

The i5k Workspace@NAL

- **Resources:**
  - Organism landing page (Tripal software)
  - Gene pages for official gene sets
  - Tutorials
- **Tools:**
  - BLAST, HMMER, Clustal
  - JBrowse genome browser
  - Apollo curation software
- **Support:**
  - Semi-automated QC of manual annotations
  - OGS generation pipeline



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

Organisms / Megachile rotundata

Megachile rotundata

Overview
Annotation Methods
Assembly Methods
NCBI BioProject

Overview

The (Meg inadv befor becor Unite pollin to in under most and is surpassed only by the honey bee for its

Females are gregarious cavity nesters construc of cells in naturally occurring cavities or in

Dicer-2, OFAS025276 (gene) Oncopeltus fasciatus

Overview | Sequences

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

Available Tracks

BLAST Databases

Organisms

□All organisms
□Agrilus planipennis
□Anoplophora glabripennis
□Athalia rosae
☑Bactrocera cucurbitae
□Bactrocera dorsalis
☑Blattella germanica
□Catajapyx aquilonaris
□Centruroides exilicauda
□Ceratitis capitata
☑Cimex lectularius
□Copidosoma floridanum
□Diaphorina citri

Cimex lectularius

Nucleotide

☑Genome Assembly - Clec_Bbug02212013

☑Transcript - CLEC_new_ids.fna

Peptide

Query Sequence

Your sequence is detected as

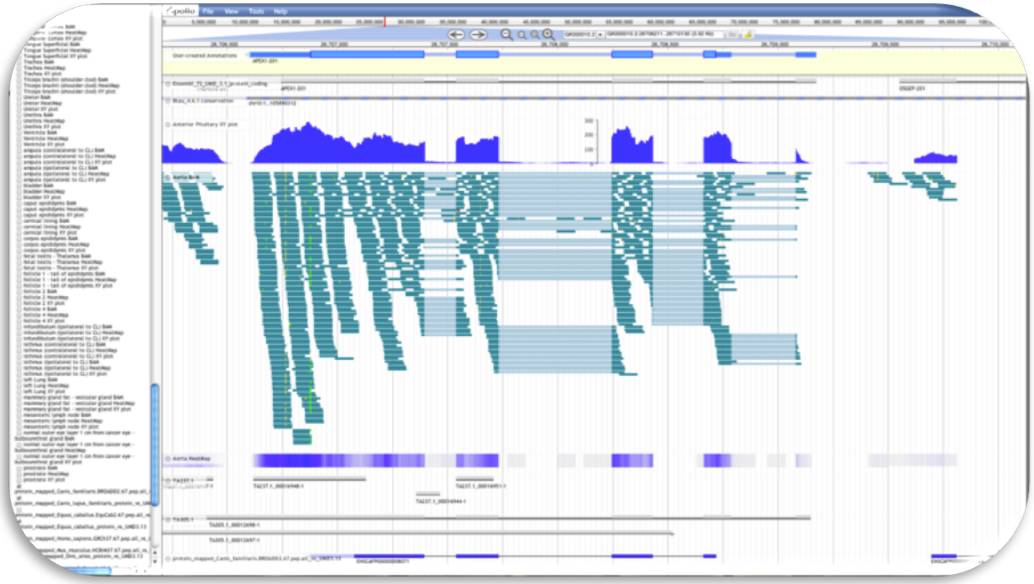
```
WVESAGNVISLQGFSYNAVHFA
SNIPETINIVELATLADMLCLEGL
KVKYCHFFHKPCNSNSCISGVLE
YGLDEIYRKSLRWITKYFVRVWP
PKELQDKCYKQHIVHLSAENVL
```

i5k Workspace data – 53 species and counting

Order	Quantity	Order	Quantity
Amphipoda	1	Harpacticoida	1
Araneae	3	Hemiptera	7
Blattodea	1	Hymenoptera	13
Calanoida	1	Lepidoptera	2
Coleoptera	5	Odonata	1
Diplura	1	Scorpiones	1
Diptera	13	Thysanoptera	1
Ephemeroptera	1	Trichoptera	1

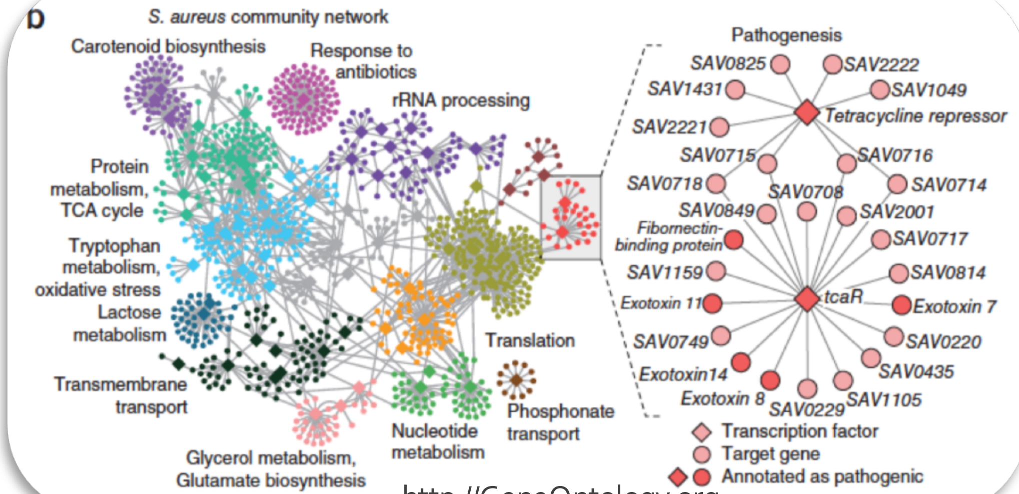
- Many other datasets mapped to, or predicted from each genome assembly (gene predictions, transcriptomes, RNA-Seq, etc.)

Curation



Identifies elements that best represent the underlying biology & **eliminates** elements that reflect systemic errors of automated analyses.

Assigns function through comparative analysis of similar genome elements from closely related species using literature, databases, and experimental data.



http://GeneOntology.org

Community curation at the i5k Workspace

- Why curate?
  - Verify quality of automated gene predictions
  - Improve gene models for specific analyses

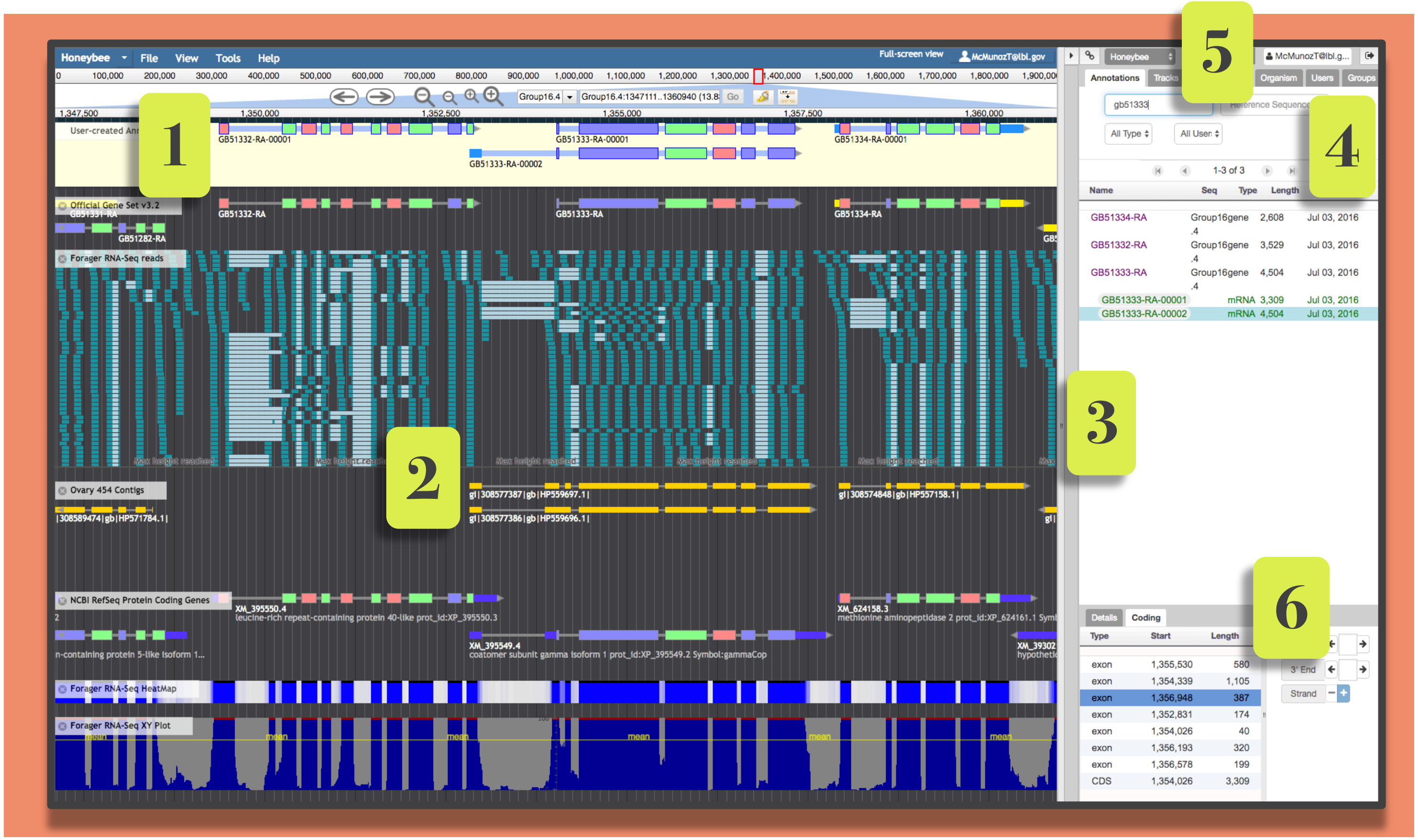
- **Our community:** Over 400 registered annotators have curated more than10,000 gene models using the Apollo genome annotation editor.

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)

Apollo: Collaborative, instantaneous, web-based



Name	Seq	Type	Length	
GB51334-RA	Group16gene.4		2,608	Jul 03, 2016
GB51332-RA	Group16gene.4		3,529	Jul 03, 2016
GB51333-RA	Group16gene.4		4,504	Jul 03, 2016
GB51333-RA-00001		mRNA	3,309	Jul 03, 2016
GB51333-RA-00002		mRNA	4,504	Jul 03, 2016

Type	Start	Length
exon	1,355,530	580
exon	1,354,339	1,105
exon	1,356,948	387
exon	1,352,831	174
exon	1,354,026	40
exon	1,356,193	320
exon	1,356,578	199
CDS	1,354,026	3,309

1. User-created Annotations
2. Evidence Tracks: Experimental data, alignments
3. Annotator Panel: Removable dock
4. Tabs for searching, editing, and exporting data
5. Switch between organisms and sequences
6. Visualize and edit annotation details

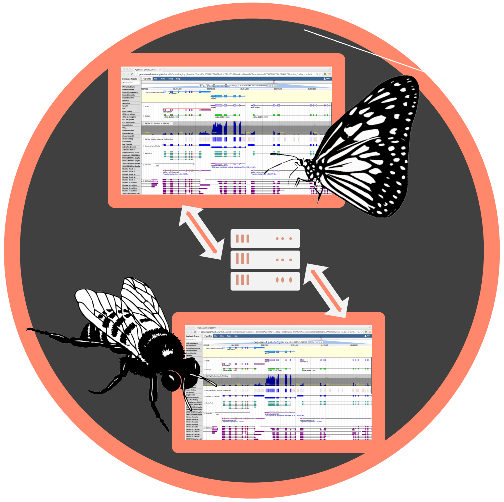
Apollo

Powered by JBrowse

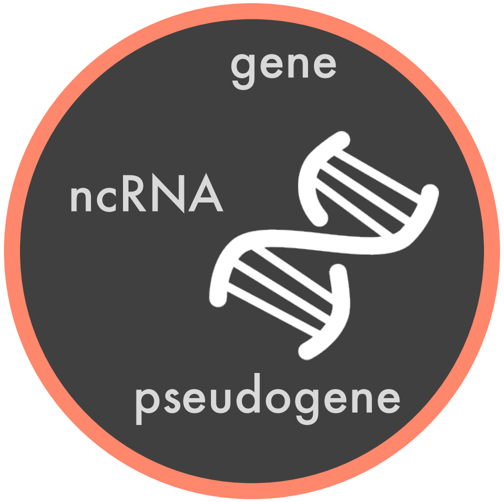
BERKELEY LAB
Lawrence Berkeley National Laboratory

Apollo: Latest Features

Annotate multiple organisms per server



Change type of genomic element to annotate



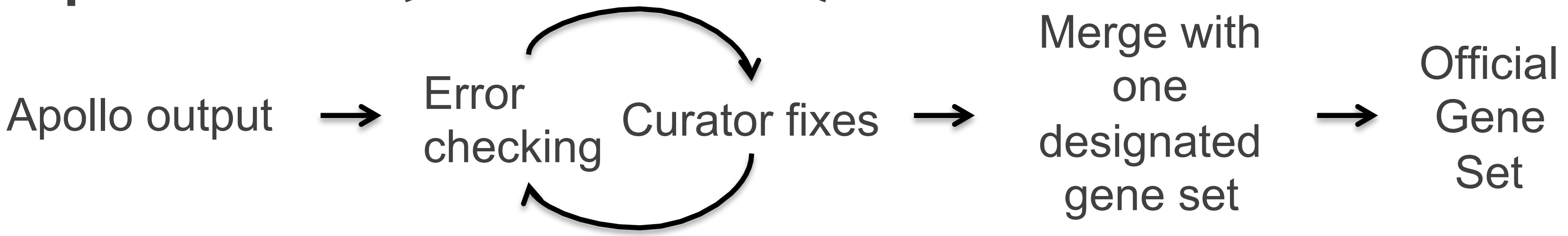
Galaxy / Apollo Integration

Export and update a Chado database

BERKELEY LAB
Lawrence Berkeley National Laboratory

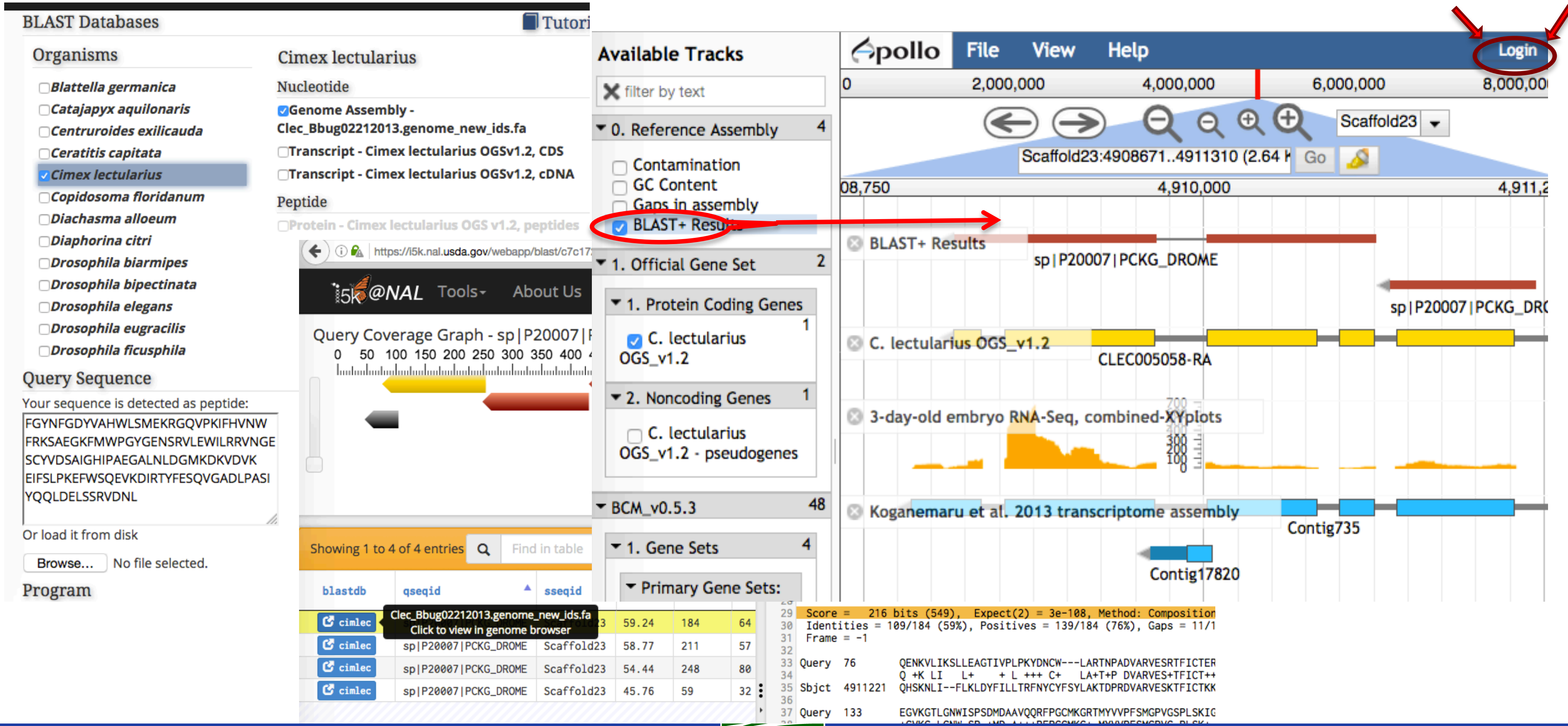
QC and OGS pipeline

- 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 4 OGS's produced (Mei-Ju Chen)


Apollo output → Error checking ⟲ Curator fixes → Merge with one designated gene set → Official Gene Set

Genome already hosted elsewhere?

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



BLAST Databases

Organisms

- ☐ *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*

Cimex lectularius

Nucleotide

- ☑ Genome Assembly - Clec_Bbug02212013.genome_new_ids.fa
- ☐ Transcript - Cimex lectularius OGSv1.2, CDS
- ☐ Transcript - Cimex lectularius OGSv1.2, cDNA

Peptide

- ☐ Protein - Cimex lectularius OGS v1.2, peptides

Query Sequence

Your sequence is detected as peptide:

```
FGYNFGDYVAHWLSMEKRGQVPKIFHVNW
FRKSAEGKFMWPGYGENSRVLEWILRRVNGE
SCYVDSAIGHIPAEGALNLDGMKDKVDVK
EIFSLPKEFWSQEVKDIRTYFESQVGADLPASI
YQQLDELSSRVDNL
```

Or load it from disk

Browse... No file selected.

Program

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

Showing 1 to 4 of 4 entries

Find in table

blastdb	qseqid	sseqid			
cimlec			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

Available Tracks

filter by text

- 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:

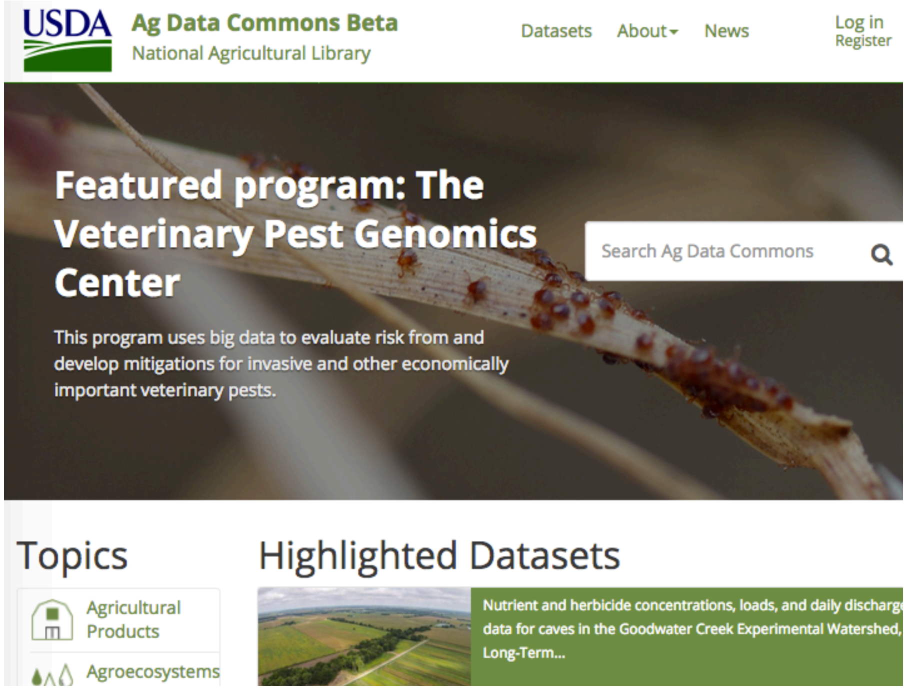
```
 Score =   216 bits (549),  Expect(2) = 3e-108, Method: Composition
 Identities = 109/184 (59%), Positives = 139/184 (76%), Gaps = 11/1
 Frame = -1

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



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

Highlighted Datasets

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

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/
- New webinar series coming soon!

Apollo:

- http://GenomeArchitect.org/

Ag Data Commons:

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

Learn more about
i5k Workspace@NAL at
Poster D3385 on Tuesday

Visit us this week at
Booth 320; 1-5 PM

Acknowledgements

The NAL Team

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

Workspace alumni

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

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
- **All of our users and contributors!**