

The Workspace@NAL: Introduction, overview and examples

Chris Childers and Monica Poelchau
USDA-ARS, National Agricultural Library
July 14, 2017

Overview

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

The Workspace@NAL

Our focus:

- We support any 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


Overview

1. Background: What is the i5k Workspace?
2. Finding data at the i5k Workspace
 1. General search/Content types
 2. Data downloads
 3. BLAST
 4. Clustal(s)
 5. HMMER
 6. JBrowse/Apollo
3. Improving data at the i5k Workspace via community annotation

Finding Data at the i5k Workspace

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

Organism pages

 i5k Workspace@NAL

Organisms ▾ Data ▾ Tools ▾ Tutorials and Resources ▾ Contact About us

Search


Login

Organisms / Megachile rotundata

Megachile rotundata

[Overview](#)
[Annotation Methods](#)
[Assembly Methods](#)
[NCBI BioProject](#)

Overview



The alfalfa leafcutting bee, *Megachile rotundata* F. (Megachilidae) is a Eurasian solitary bee species that was inadvertently introduced to North America sometime before the 1940s. By the mid 1950s, *M. rotundata* had become established in the farming regions of western United States. With the discovery of *M. rotundata*'s pollination impact on alfalfa seed production, early efforts to increase its populations near alfalfa fields were undertaken a few years later. Currently, *M. rotundata* is the most intensely managed solitary bee species in the world and is surpassed only by the honey bee for its economic impact.

Females are gregarious cavity nesters constructing nests composed of leaf pieces in a linear series of cells in naturally occurring cavities or in artificial nesting boards. In most North American 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

Megachile rotundata data files

Name	Last modified
← Parent Directory	
Current Genome Assembly	2016-05-04 20:17

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	

Gene pages (Official gene sets only)

The screenshot displays the i5k Workspace@NAL website interface. At the top, there is a navigation bar with the USDA logo, the text "i5k Workspace@NAL", and several menu items: "Organisms", "Data", "Tools", "Tutorials and Resources", "Contact", and "About us". A search bar is located on the right side of the navigation bar, and a "Login" link is in the top right corner.

The main content area is titled "Dicer-2, OFAS025276 (gene) Oncopeltus fasciatus". Below the title, there are three tabs: "Overview" (selected), "Sequences", and "Transcripts".

The "Overview" tab contains the following information:


- 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

Below the overview information, there is a section titled "Available Tracks". It includes a "filter by text" input field and two main track categories:

- 0. Reference Assembly** (2 tracks):
 - ☐ GC Content
 - ☐ Gaps in assembly
- 1. Official Gene Set v1.2** (4 tracks):
 - 1. Gene Sets** (4 tracks):
 - ☒ Noncoding (1)
 - ☒ Other features (1)
 - ☐ OGSv1.2 sequence modifications

The main visualization area shows a genomic track for Scaffold23. The track is titled "Scaffold23" and "Scaffold23:319454..445774 (126.32 K)". It displays a genomic map with various features. A specific feature is highlighted: "Dicer-2 -part 1 of 2" with a note: "Note to curator: complete, concatenated CDS:...". The track also shows "OGSv1.2 protein-coding genes" and "D-RA".

Tutorials

 **i5k Workspace@NAL** [Organisms ▾](#) [Data ▾](#) [Tools ▾](#) [Tutorials and Resources ▾](#) [Contact](#) [About us](#) [Login](#)

A place for arthropod genome communities to curate data

- [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](#) ▶

Apollo/JBrowse

[View guidelines](#)


Manually curate a genome with Apollo, or browse a genome and its features with JBrowse.

[REGISTER](#)

[Annotation rules](#)


[Manual curation example](#)


[The Apollo 'Replaced Models' field - explanations and examples](#)


 © 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](#)

<https://i5k.nal.usda.gov/manual-curation-example>

Finding Data at the i5k Workspace

- Website search for metadata (e.g. search term “*Anoplophora glabripennis*”)

The screenshot displays the i5k Workspace@NAL website interface. At the top, the USDA logo and text "United States Department of Agriculture National Agricultural Library" are on the left, and "i5k Workspace@NAL" is on the right. A navigation bar includes links for "Organisms", "Data", "Tools", "Tutorials and Resources", "Contact", and "About us". A search bar with the text "Search" and a magnifying glass icon is present, along with a "My Account" link. Below the navigation bar, a breadcrumb trail reads "Search / Search results / Anoplophora glabripennis / Site".

The main content area is titled "Site" and features a search box with the text "Enter terms" and "Anoplophora glabripenni:" followed by a "Search" button. To the right, a "Filter by content type:" sidebar lists categories: "Feature (22253)", "Analysis (3)", "Page (3)", "iframe (1)", "News (1)", and "Organism (1)".

The "Search results" section is titled "Anoplophora glabripennis" in orange. It contains a paragraph: "The Asian long-horned beetle (**Anoplophora glabripennis**) (ALB) is an invasive pest from Asia that ... beetle (**Anoplophora glabripennis**), a globally significant invasive species, reveals key functional and ... 10.1186/s13059-016-1088-8 **Anoplophora glabripennis** ...".

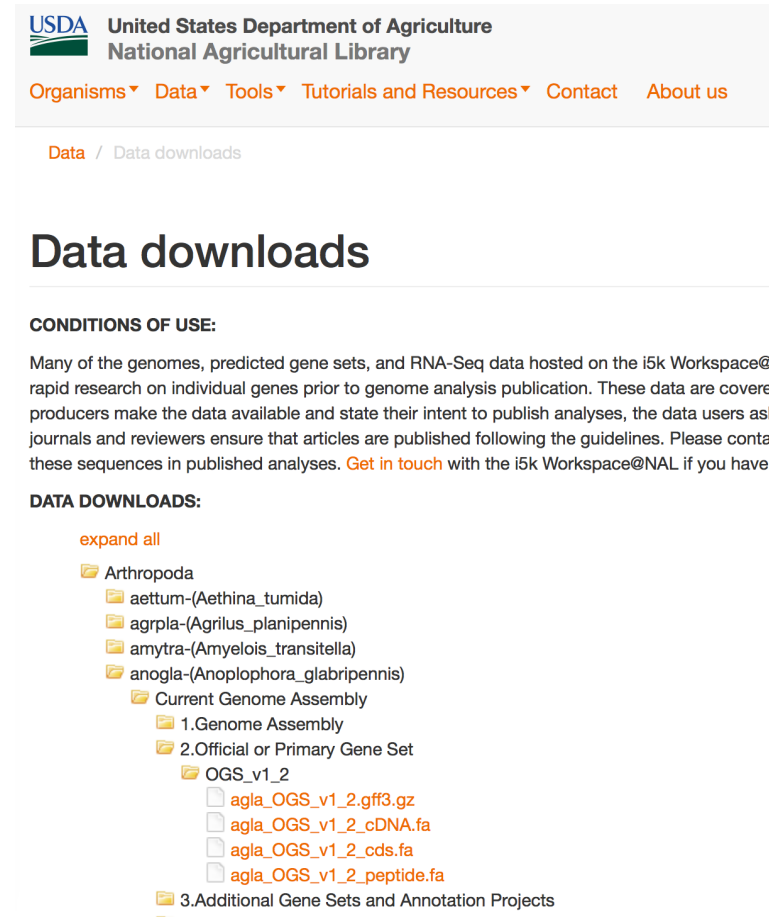
Below this, a section titled "Several new datasets available for viewing on JBrowse/Web Apollo" in orange, followed by a paragraph: "We have a number of new datasets visualized on several genome browsers:- Two new transcriptomes mapped to the *Frankliniella occidentalis* genome- New RNA-Seq data for *Bactrocera dorsalis*- new OGS uploaded for **Anoplophora glabripennis**- Transposable element predictions and miRNA predictions in ...".

Next is a section titled "Available genome browsers" in orange, followed by a paragraph: "(Emerald ash borer) Amyelois transitella (Citrus orange worm) **Anoplophora glabripennis** (Asian long-horned ...".

The final section is titled "Whole genome assembly of Anoplophora glabripennis" in orange.

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/>



The screenshot shows the USDA National Agricultural Library website. The header includes the USDA logo and the text "United States Department of Agriculture National Agricultural Library". Below the header is a navigation bar with links: "Organisms", "Data", "Tools", "Tutorials and Resources", "Contact", and "About us". The "Data" link is highlighted. Below the navigation bar is a breadcrumb trail: "Data / Data downloads". The main heading is "Data downloads". Below this is a section titled "CONDITIONS OF USE:" which states that many genomes, predicted gene sets, and RNA-Seq data are hosted on the i5k Workspace for rapid research. It mentions that data producers make the data available and state their intent to publish analyses, and that data users ask journals and reviewers to ensure articles are published following the guidelines. It also mentions that users can "Get in touch" with the i5k Workspace@NAL if they have these sequences in published analyses. Below this is a section titled "DATA DOWNLOADS:". Under this section is a link "expand all". Below the link is a tree structure of data categories. The first category is "Arthropoda", which includes sub-categories: "aettum-(Aethina_tumida)", "agrpla-(Agrilus_planipennis)", "amytra-(Amyelois_transitella)", and "anogla-(Anoplophora_glabripennis)". The "anogla-(Anoplophora_glabripennis)" category is expanded, showing sub-categories: "Current Genome Assembly", "1.Genome Assembly", "2.Official or Primary Gene Set", and "OGS_v1_2". The "OGS_v1_2" category is expanded, showing sub-categories: "agla_OGS_v1_2_gff3.gz", "agla_OGS_v1_2_cDNA.fa", "agla_OGS_v1_2_cds.fa", and "agla_OGS_v1_2_peptide.fa". The "3.Additional Gene Sets and Annotation Projects" category is also visible.

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 covered by 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 contact 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/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>

The screenshot shows the BLAST+ web interface. At the top is a navigation bar with the i5k@NAL logo and links for Tools, About Us, and Contact. The main section is titled 'BLAST Databases'. Under 'Organisms', a list of species is shown with checkboxes; 'Anoplophora glabripennis' is selected. To the right, under 'Anoplophora glabripennis', there are sections for 'Nucleotide' (with 'Genome Assembly' and 'Transcript' checked) and 'Peptide' (with 'Protein' checked). Below this is the 'Query Sequence' section, which states 'Your sequence is detected as peptide:' and shows a long sequence of amino acids. There is a text area for the sequence and a 'Choose File' button. At the bottom, there is a 'Program' section with radio buttons for 'blastn', 'tblastn' (selected), 'tblastx', 'blastp', and 'blastx', along with 'Reset' and 'Search' buttons.

5k@NAL Tools - About Us Contact

BLAST Databases

Organisms

- ☐ All organisms
- ☐ *Aethina tumida*
- ☐ *Agrilus planipennis*
- ☐ *Amyeloides transitella*
- ☒ *Anoplophora glabripennis*
- ☐ *Athalia rosae*
- ☐ *Bactrocera cucurbitae*
- ☐ *Bactrocera dorsalis*
- ☐ *Bactrocera oleae*
- ☐ *Blattella germanica*
- ☐ *Catajapyx aquilonaris*
- ☐ *Centruroides exilicauda*
- ☐ *Cenobius cinctus*

Anoplophora glabripennis

Nucleotide

- ☒ Genome Assembly - Agla_BtI03082013.gen
- ☒ Transcript - Anoplophora glabripennis cDI
- ☐ Transcript - Anoplophora glabripennis CD

Peptide

- ☐ Protein - Anoplophora glabripennis pepti

Query Sequence

Your sequence is detected as peptide:

```
>FBpp0070037 type=protein;
loc=3L:complement(join(23337875..2333
8046,23334364..23334740,23330698..233
31015,23329886..23330089,23328714..23
329398,23327494..23327645,23326688..2
3326760,23325889..23325956,23325642..
```

Or load it from disk

Choose File no file selected

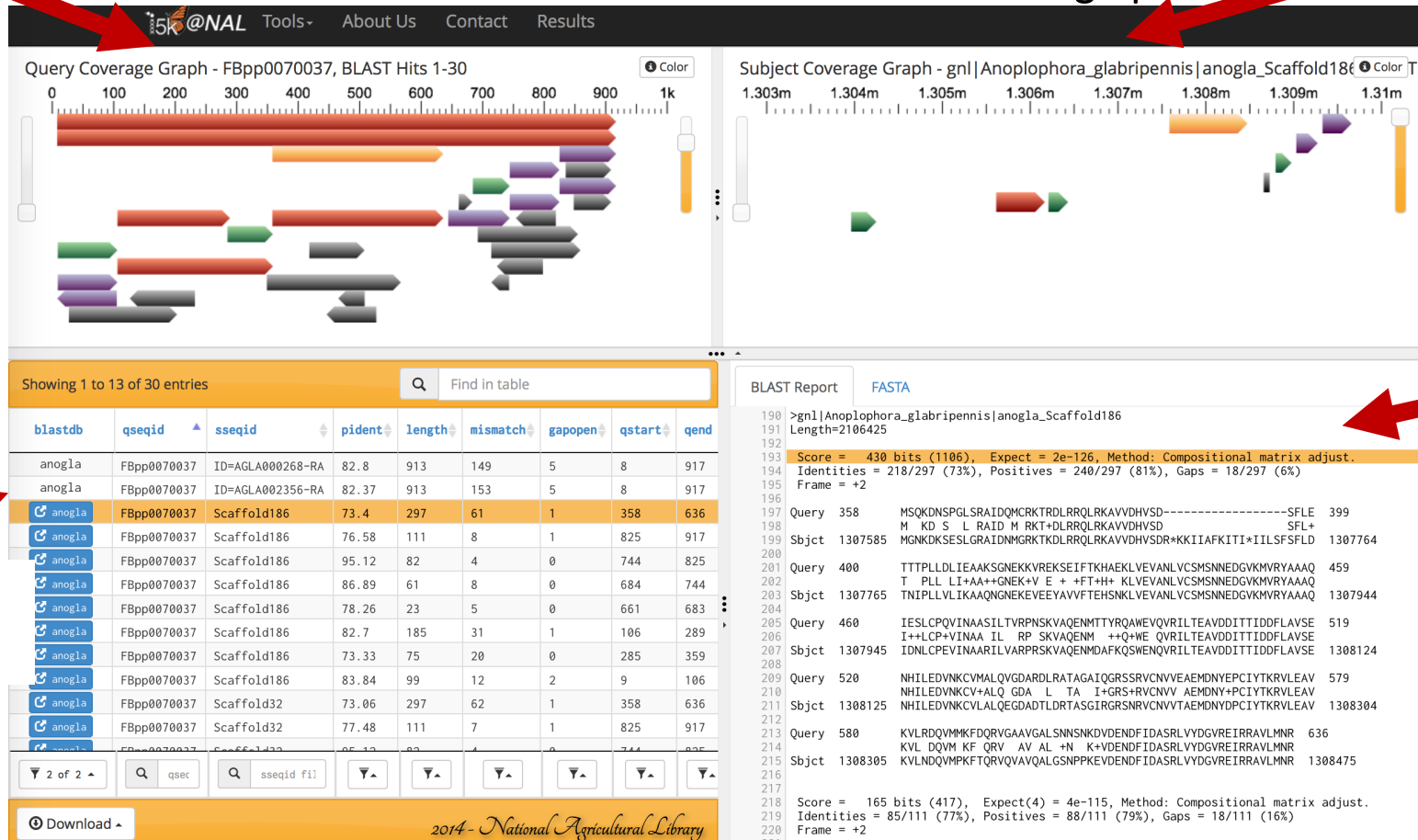
Program

☐ blastn ☒ tblastn ☐ tblastx ☐ blastp ☐ blastx

Sequence Search – BLAST+ Result

Query coverage graph

Subject coverage graph

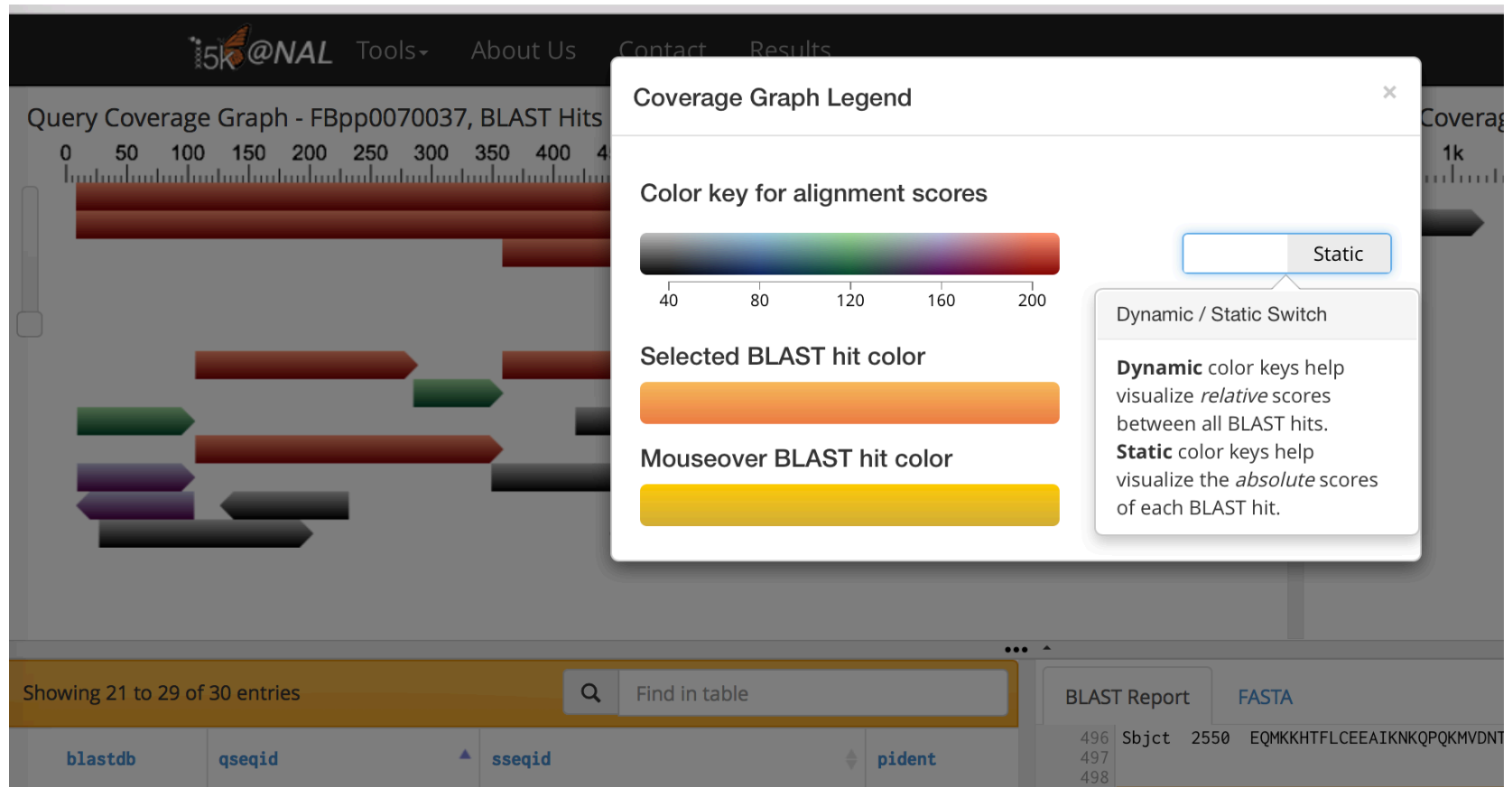


Tabular result

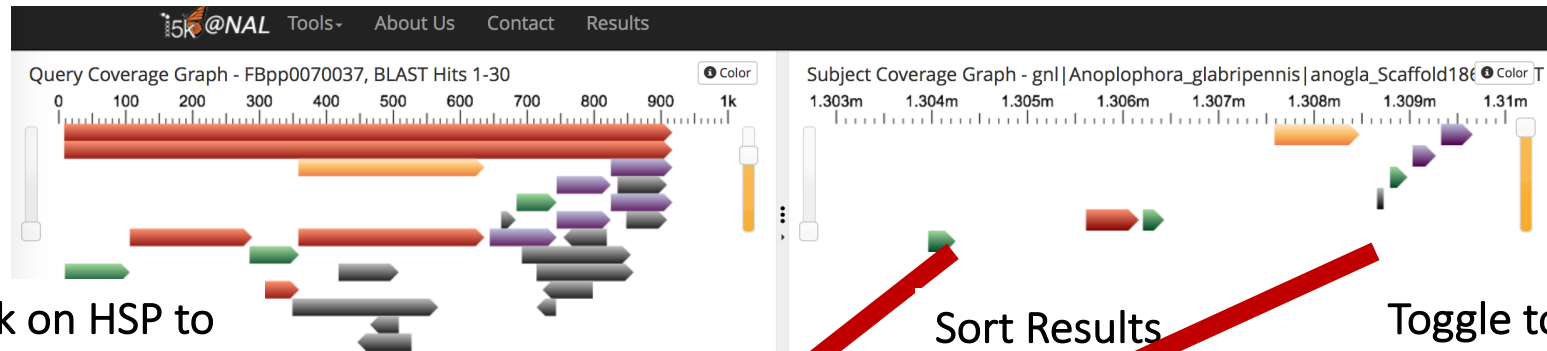
Raw results

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

Sequence Search – BLAST+ Result



Sequence Search – BLAST+ Result



Click on HSP to
'freeze' it

Sort Results

Toggle to
get
sequence

Filter
Results

Download

Showing 1 to 15 of 30 entries

Find in table

blastdb	qseqid	sseqid	pid	length	mismatch	gapopen	qstart	qend
anogla	FBpp0070037	ID=AGLA000268-RA	82.8	913	149	5	8	917
anogla	FBpp0070037	ID=AGLA002356-RA	82.37	913	153	5	8	917
anogla	FBpp0070037	Scaffold186	73.4	297	61	1	358	636
anogla	FBpp0070037	Scaffold186	76.58	111	8	1	825	917
anogla	FBpp0070037	Scaffold186	95.12	82	4	0	744	825
anogla	FBpp0070037	Scaffold186	86.89	61	8	0	684	744
anogla	FBpp0070037	Scaffold186	78.26	23	5	0	661	683
anogla	FBpp0070037	Scaffold186	82.7	185	31	1	106	289
anogla	FBpp0070037	Scaffold186	73.33	75	20	0	285	359
anogla	FBpp0070037	Scaffold186	83.84	99	12	2	9	106
anogla	FBpp0070037	Scaffold32	73.06	297	62	1	358	636
anogla	FBpp0070037	Scaffold32	77.48	111	7	1	825	917

2 of 2

Download

BLAST Report FASTA

>gnl|Anoplophora_glabripennis|anogla_Scaffold186
Length=2106425

Score = 430 bits (1106), Expect = 2e-126, Method: Compositional matrix adjust.
Identities = 218/297 (73%), Positives = 240/297 (81%), Gaps = 18/297 (6%)
Frame = +2

Query 358 MSQKDN SPLSRAIDQMCRTDRLRRQKRAVDHVS-----SFL 399
M KD S L RAID M RKT+DLRRQKRAVDHVS SFL+
Sbjct 1307585 MGNKDKSESLGRAIDNMGKTKDLRRQKRAVDHVS+KKIIAFKITI+IILSFSFLD 1307764

Query 400 TTTPLLDLIEAAKSGNEKKVREKSEIFTKHAEKLVANVCSMSNNEDGVKMYRAAAQ 459
T PLL LI+AA++GNEK+V E + +FT+H+ KLVEANVCSMSNNEDGVKMYRAAAQ
Sbjct 1307765 TNIPLLVLIIAAQNGNEKEVEEYAVVFTEHSNKLVEANVCSMSNNEDGVKMYRAAAQ 1307944

Query 460 IESLCPQVINAASILTVRPNKVAQENMTTYRQAWQVQRILTEAVDDITTDIFLAVSE 519
I++LCP+VINAA IL RP SKVAQENM ++Q+WE QVRILTEAVDDITTDIFLAVSE
Sbjct 1307945 IDNLCPQVINAARILVARPRSKVAQENMDAFKQSWENQVRILTEAVDDITTDIFLAVSE 1308124

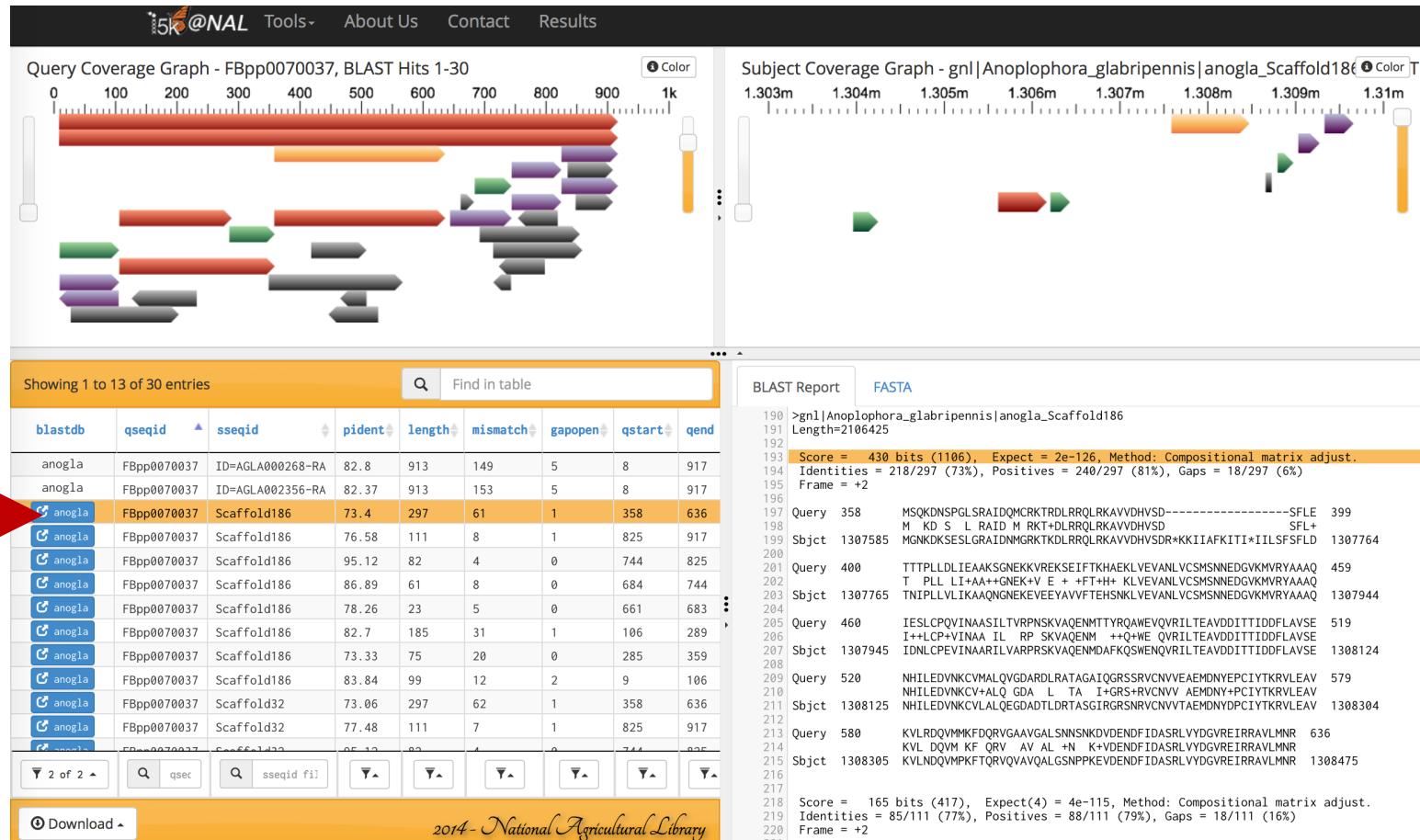
Query 520 NHILEDVNVKCVMALQVGDARDLRATAGAIQGRSSRVNVEAEMDNYPICYTKRVLEAV 579
NHILEDVNVKCV+ALQ GDA L TA I+GRS+RVCNVV AEMDNYPICYTKRVLEAV
Sbjct 1308125 NHILEDVNVKCVLALQEGDADTLDRTAGSIRGRSNRVCNVVTEAEMDNYPICYTKRVLEAV 1308304

Query 580 KVLRDQVMKFDQRVGAAGALSNNSNKVDENDFIDASRLVYDGVREIRRAVLMNR 636
KVL DQVM KF QRV AV AL +N K+VDENDFIDASRLVYDGVREIRRAVLMNR
Sbjct 1308305 KVLNDQVMKFTQRVQVAQALGSNPKPEVDENDFIDASRLVYDGVREIRRAVLMNR 1308475

Score = 165 bits (417), Expect(4) = 4e-115, Method: Compositional matrix adjust.
Identities = 85/111 (77%), Positives = 88/111 (79%), Gaps = 18/111 (16%)
Frame = +2

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

Sequence Search – links to JBrowse



Sequence Search – links to JBrowse

The screenshot displays the JBrowse genome browser interface. The top bar shows the application name 'JBrowse Scaffold186:1307385..1308675' and a 'Login' button. Below the top bar is a navigation menu with 'File', 'View', and 'Help' options. The main display area features a genomic track with a scale from 0 to 2,000,000. A red vertical line indicates the current position at approximately 1,307,500. The track shows 'Scaffold186' and 'Scaffold186:1307385..1308675 (1.29 kb)'. Below the track, a 'BLAST+ Results' track is visible, showing a single hit 'FBpp0070037'. On the left side, the 'Available Tracks' panel is open, listing various tracks. A red arrow points to the 'BLAST+ Results' track, which is checked. The 'Available Tracks' panel also includes a 'filter by text' input field and a list of tracks: '0. Reference Assembly' (3 tracks), '1. Official Gene Set' (2 tracks), '1. Protein Coding Genes' (1 track), and '2. Noncoding Genes' (1 track). The 'BLAST+ Results' track is highlighted in red.

BLAST Result | i5k - App

JBrowse Scaffold186:1307385..1308675

Available Tracks

filter by text

0. Reference Assembly 3

- ☐ GC Content
- ☐ Gaps in assembly
- ☒ BLAST+ Results

1. Official Gene Set 2

- 1. Protein Coding Genes 1
 - ☐ Official Gene Set v1.2 - protein-coding genes
- 2. Noncoding Genes 1
 - ☐ Official Gene Set v1.2 - pseudogenes

JBrowse File View Help Login

0 200,000 400,000 600,000 800,000 1,000,000 1,200,000 1,400,000 1,600,000 1,800,000 2,000,000

Scaffold186 Scaffold186:1307385..1308675 (1.29 kb) Go

1,307,500 1,308,000 1,308,500

BLAST+ Results

FBpp0070037

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=>

The screenshot shows the ClustalW web interface. At the top, there is a navigation bar with the i5k@NAL logo and links for Tools, About Us, and Contact. Below this, the main heading is "CLUSTALO CLUSTALW". The "Query Sequence" section displays a protein sequence identified as Chaperonin Cpn60 from Dana (GF20641). The sequence is shown in a text box with a description: ">7217:002330 ('pub_gene_id':'Dana\GF20641', 'pub_og_id':'EOG091906CT', 'og_name':'Chaperonin Cpn60', 'level':7215, 'description':'Similarity:Belongs to the chaperonin (HSP60) family.').". Below the sequence, there is a "Choose File" button and a "no file selected" message. The "Sequence Input" section has a "Design input sequences" checkbox set to "yes" and a "Search" button. The "Clustering and Iteration" section has three sub-sections: "MBED-Like Clustering Guide-Tree" with a "yes" radio button, "MBED-Like Clustering Iteration" with a "yes" radio button, and "Max HMM Iterations" with a "Default" dropdown. The "Iteration" section has a "Number of Combined Iterations" dropdown set to "Default" and a "Max Guide Tree Iterations" dropdown set to "Default". The "Output" section has a "Format" dropdown set to "Clustal" and an "Out Order" dropdown set to "Aligned".

Sequence alignment – ClustalW and Clustal Omega



Tools- About Us Contact

CLUSTAL Success

Download

Alignment

Submission Details

Report Details

CLUSTAL O(1.2.3) multiple sequence alignment

```
7222:0004a3      MFRSYVRE-SIRSSRAFARAYSKAVTFGAEARARMLHGVDVLADAVAVTLGPKGRSVILE
7230:00256a      MFRSYVRK-AVRSSRAFARAYSKDVAFGADARARMLRGVDLTDAAVAVTLGPKGRSVILE
7244:0019f0      MFRSYVRE-AVRSSRAFARAYSKDVAFGAEARARMLRGVDMLTDAAVAVTMGPKGRSVILE
7260:0029e2      MFRFFARDAAVCTGRNLCRAYSKVRFGEVRLMIRGVDILADAVAVTMGPKGRNVILE
7234:0021c7      MFRHCVRG-ALRGNRNLRLYSKDVRFGEARSMIRGVDLLADAVAVTMGPKGRSVILE
7237:002b4f      MFRHCVRG-VLRGNRNLRLYSKDVRFGEARSMIRGVDLLADAVAVTMGPKGRSVILE
7217:002330      MFRSCVRD-AIRSSRFFTRMYSKEVRFGEVRLMIRGVDVLADAVAVTMGPKGRSVILE
28584:0003f1     MFRSCVSE-AITSSRCFARMYSKEVRFGEVRLMIRGVDVLADAVAVTMGPKGRSVILE
7220:002f9d      MFRSCVPK-AISSSRCFARMYSKEVRFGEVRLMIRGVDVLADAVAVTMGPKGRSVILE
7245:001764      MFRSCVPK-AISSRCFARMYSKDVRFGEVRLMIRGVDVLADAVAVTMGPKGRSVILE
7227:000486      MFRSCVPK-AITSSRCFARMYSKDVRFGEVRLMIRGVDILADAVAVTMGPKGRSVILE
7238:0019ed      MFQSCVPK-AITSSRCFARMYSKDVRFGEVRLMIRGVDVLADAVAVTMGPKGRSVILE
7240:002709      MFRSCVPK-AITSSRCFARMYSKDVRFGEVRLMIRGVDVLADAVAVTMGPKGRSVILE
**      .      :      *      :      *      *      *      *      :      *      :      *      *      *      *      *      *      *
```

Result URL:

<https://i5k.nal.usda.gov/webapp/clustal/ed00819ab40441ca959eacdccb78c0f5>

```
7222:0004a3      -MDEMGAMDGI-----S-KAAEMNDVKSIPGMENVEVHDIDSSQ
7230:00256a      GMDVDGEICNK-----S-KAAEMNEAVQSIPGMEDVTVHDIDSTQ
7244:0019f0      MGMG---MGM-----S-KAEMNQAVQSISAGMEDVTVHDIDSSQ
7260:0029e2      DI-----MDAMGGG-----ASDGGASAKDLNEIVNIPGMEDVTVSDIDSSQ
7234:0021c7      RTR---DMGGD-----GG-----SSTSAEMNEMVKSMPGMENVEVQIDASM
7237:002b4f      GLG---DMGGD-----GG-----SSTSAEMNEMVKSMPGMENVEVQIDASM
7217:002330      GLGGLGDMGDMTGRI-----SASVPKEDDGPTEEMNEMVKAIPGMEQVEVRIDAGL
28584:0003f1     ---GGMGGMGGGFGGMGGGGGMSASKSDGPTAEMNEMVKAIPGMEQVEVRIDSGM
7220:002f9d      GGMGGMAMGGMGGGFGGMGGGGGMSASSSDGPTAEMNEMVKAIPGMEQVEVRIDSGM
7245:001764      GGMGGMAMGGMGAGFGGMGGGGGMSASSSDGPTAEMNEMVKAIPGMEQVEVRIDSGM
7227:000486      ---MGMGGMGGGFGGMGAGGGMSASASNDGPTAEMNEMVKAIPGMEQVEVRIDSGM
7238:0019ed      ---MGMGGMGGGFGGMGAGGGMSASASNEGPTAEMNEMVKAIPGMEQVEVRIDSGM
7240:002709      ---MGMGGMGGGFGGMGAGGGMSASASNEGPTAEMNEMVKAIPGMEQVEVRIDSGM
      . * : : * : * : * : * : * : * : * : * : * : * : * : * : * : *
```

```
7222:0004a3      M-
7230:00256a      L-
7244:0019f0      L-
7260:0029e2      F-
7234:0021c7      MQ
7237:002b4f      MQ
7217:002330      M-
28584:0003f1     M-
7220:002f9d      M-
7245:001764      M-
7227:000486      M-
7238:0019ed      M-
7240:002709      M-
      :
```

uncolorful

To hmmsearch



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=>



i5k@NAL Tools - About Us Contact

Organisms

- ☐ All organisms
- ☐ *Aethina tumida*
- ☐ *Agrilus planipennis*
- ☐ *Amyeloides transitella*
- ☐ *Anoplophora glabripennis*
- ☐ *Athalia rosae*
- ☐ *Bactrocera cucurbitae*
- ☒ *Bactrocera dorsalis*
- ☐ *Bactrocera oleae*
- ☐ *Blattella germanica*
- ☐ *Cataglyphis aquilonaris*
- ☐ *Centruroides exilicauda*
- ☐ *Cenhus cinctus*

Bactrocera dorsalis

Protein

- ☒ Protein - Bactrocera dorsalis BDOR_v1 proteins - MAKER and
- ☒ Protein - Bactrocera dorsalis - NCBI annotation release 100,

Query Sequence / Multiple sequence alignment

Your sequence is detected as fasta:

```
>7217:002330 {"pub_gene_id":"Dana\GF20641", "pub_og_id":"EOG091906CT",
"og_name":"Chaperonin Cpn60", "level":7215, "description":"Similarity:Belongs to
the chaperonin (HSP60) family."}
MFRSCVRDAIRSSRFFTRMYSKEVRFPGPEVRAMIRGVDVLADAVAVTMGPKGRSVIVERPWTSPKITKDGFTVARSLA
LKDQHMLNLAGKLVQDVADNTNQAAGDGTITATVLARAIAKEGFNQITMGANPNEIRRGVMAVDVVKEMLKAMSKSVES
SEEIQQVATISANGDTIGRLIAEATEKVGAKGTITVKDGKRLKDELTTIIQGLRFDITGVSPFFVNSTKGSKEVFSNAL
VLITLKKITALSQIVKGLEQSLRERRPLVIAEDISGEALNALVLNKLRLMGLQCAVKSPSYGEHRKELIGDISAATGA
TIFGDDINYSKIENAKLQDMGQVGEAVITKDSMTLLEKPKPTGQLELRQQIQDELADKQTKPEQKDLRQLRSALTGK
VAVLHIGGISEVEVSEKKDRVVDALHATRAAIEEGIVPGGTAFLRCIPHLEMEVESKDLKKGVEIICNALRMPQTI
AQNAGVDGAMVAVMTKSGDYGIDAMGGQYGRLEKGIIDPTKVVRTAISDAAGVASLLSTTEVVITDTRNEDLLAKL
AGMGGGGMDGLDMNMGGLDELAALAGMGGMGGMGGMGLGGLGGLGMDMTGRISASVPKEDDGPTAEEMNEMV
KAIPGMEQVEVKDIDAGLM
```


Or load it from disk

Choose File no file selected

Program

☒ phmmer ☐ hmsearch

Sequence search – HMMER Result

 [Tools](#) [About Us](#) [Contact](#)

HMMER Success

[Download](#)

[Input file](#)

[Hmmer result](#)

[Submission Details](#)

Report Details


Jump To Dataset BDOR_v1-blast.proteins.fasta ▾

```
# phmmer :: search a protein sequence against a protein database
# HMMER 3.1b2 (February 2015); http://hmmer.org/
# Copyright (C) 2015 Howard Hughes Medical Institute.
# Freely distributed under the GNU General Public License (GPLv3).
# -----
# query sequence file:      db6e627e4fdd46e296226c119e3fcd3.in
# target sequence database: BDOR_v1-blast.proteins.fasta
# output directed to file:  0.out
# sequence reporting threshold: E-value <= 0.01
# domain reporting threshold:  E-value <= 0.03
# domain inclusion threshold:  E-value <= 0.03
# -----

Query:      7217:002330  [L=651]
Description: {"pub_gene_id":"Dana\GF20641", "pub_og_id":"EOG091906CT", "og_name":"Chaperonin Cpn60", "level":7215,
Scores for complete sequences (score includes all domains):
--- full sequence ---    --- best 1 domain ---    -#dom-
E-value  score  bias    E-value  score  bias    exp  N  Sequence                                Description
-----
1.4e-220  733.7  47.0    5.4e-220  731.8  42.5    1.9  1  gnl|Bactrocera_dorsalis_protein_v1|84043
9e-14     50.7   5.6      1.1e-09   37.2   0.3     2.1  2  gnl|Bactrocera_dorsalis_protein_v1|120188
1.7e-09   36.6   0.8      4.5e-05   22.0   0.2     2.1  2  gnl|Bactrocera_dorsalis_protein_v1|207118
1.7e-09   36.6   0.8      4.5e-05   22.0   0.2     2.1  2  gnl|Bactrocera_dorsalis_protein_v1|207162
3e-08     32.5   0.8      3.9e-06   25.5   0.2     2.1  2  gnl|Bactrocera_dorsalis_protein_v1|104162
4.7e-08   31.9   1.6      4.9e-05   21.9   0.2     2.3  2  gnl|Bactrocera_dorsalis_protein_v1|143138
1.6e-07   30.1   4.5      2.5e-07   29.5   0.7     2.0  2  gnl|Bactrocera_dorsalis_protein_v1|33490
```

Genome browser (JBrowse)

- From menu, select
'Tools -> JBrowse/Apollo
-> JBrowse/Apollo
Organisms'
- [https://i5k.nal.usda.gov/
available-genome-
browsers](https://i5k.nal.usda.gov/available-genome-browsers)



United States Department of Agriculture
National Agricultural Library

Organisms ▾ Data ▾ Tools ▾ Tutorials and Resources ▾ Contact About us

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 not require a Java-enabled browser. To log in to Web Apollo from the genome browser, click on the 'login' button at the top right of the screen. If you are having trouble logging in, see the instructions for Web Apollo [here](#).

- Aethina tumida (Small hive beetle)
- Agilus 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 Fly)
- Bactrocera oleae (Olive Fruit Fly)
- Blattella germanica (German cockroach)
- Cataglyphis 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

The screenshot displays the JBrowse genome browser interface. On the left, the 'available Tracks' panel is visible, with a red arrow pointing to the 'Official Gene Set v1.2 - protein-coding genes' track. The main panel shows a genomic track for 'Scaffold1' with a search bar containing 'AGLA000001-RA'. A red arrow points to the search bar. Below the search bar, a list of gene IDs is displayed, including 'AGLA000001-RA' and 'AGLA000002-RA'. A yellow pencil icon is visible next to the search bar. The bottom of the interface shows the 'USDA' logo.

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

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
- 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>

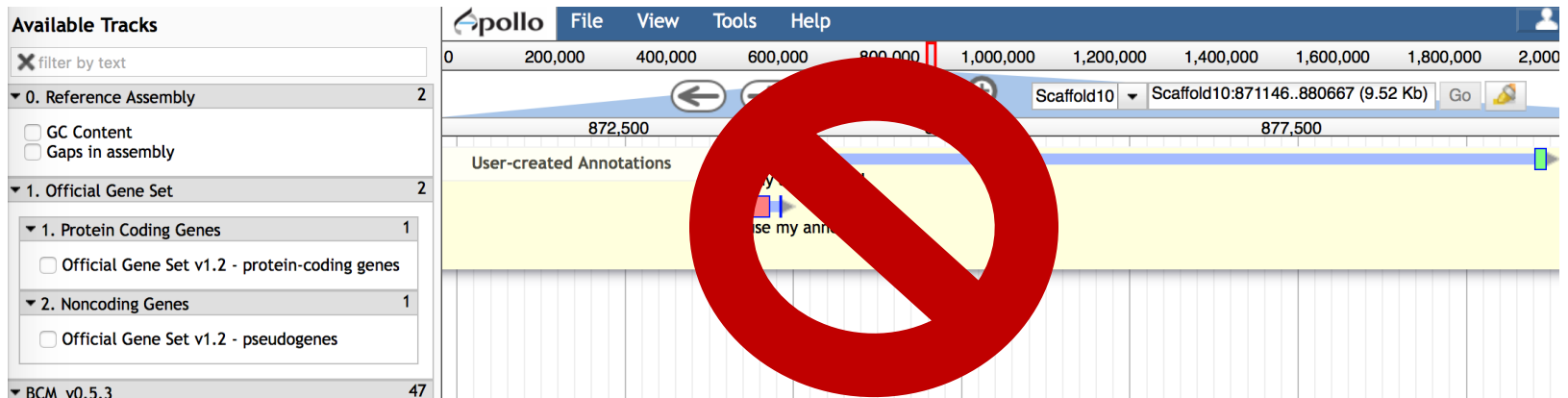
Community curation at the i5k Workspace

Our support for community curation includes:

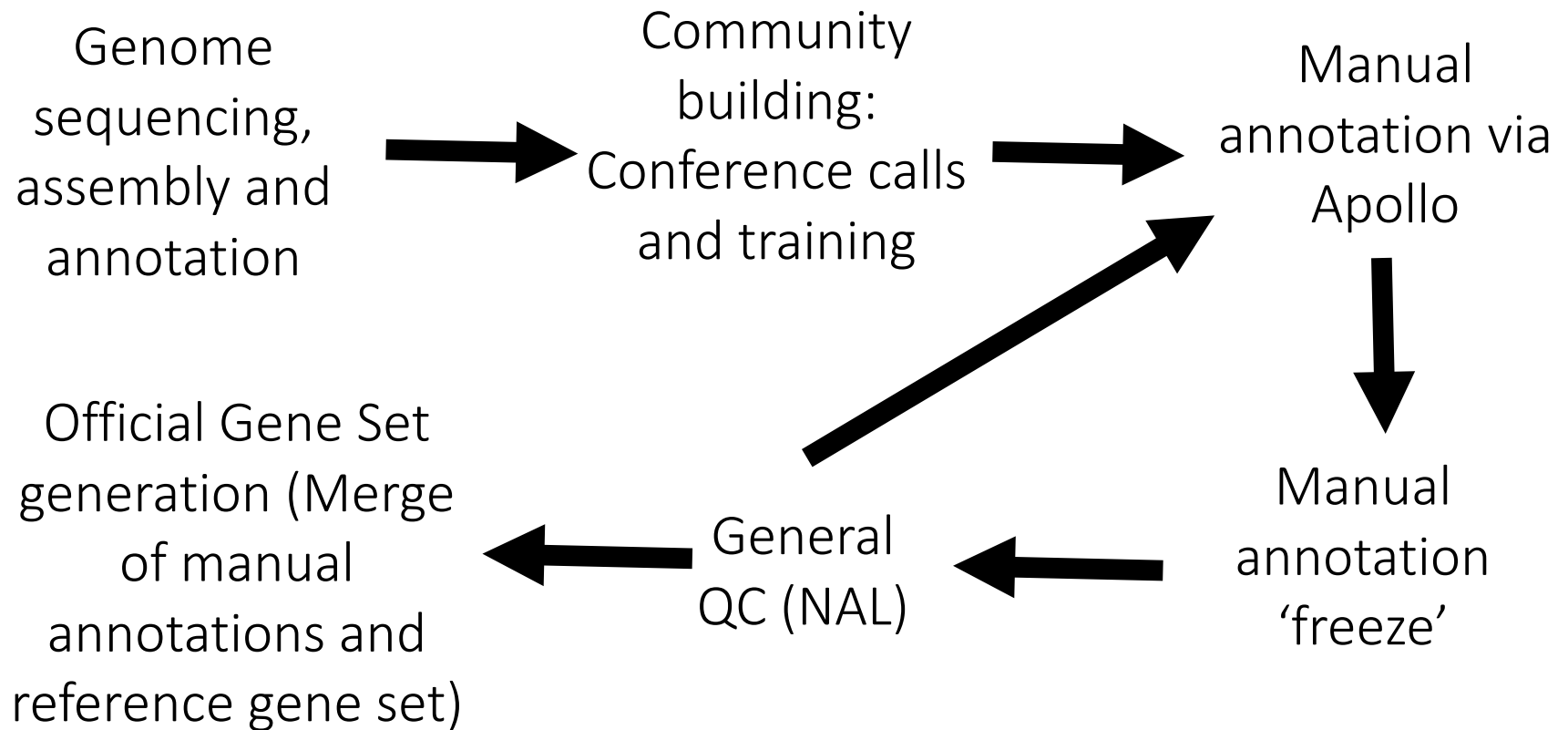
- Access to a large community of curators
- Tutorials, guidelines, webinars
- Registration mechanism for new annotators
- One-on-one support
- Software to evaluate changes between curated and original annotations (Chien-Yueh Lee, <https://github.com/chienyuehlee/gff-cmp-cat>)

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>

i5k@NAL Tools - About Us Contact

BLAST Databases

Organisms

☐ All organisms

☐ *Hyalella azteca*

☒ *Leptinotarsa decemlineata*

☐ *Neodiprion lecontei*

☐ *Oncopeltus fasciatus*

Leptinotarsa decemlineata

Nucleotide

☒ Genome Assembly - Ldec.genome.10062013_new_ids.fa

☐ Transcript - LDEC_new_ids.fna

Peptide

☐ Protein - LDEC_new_ids.faa

Query Sequence

Your sequence is detected as peptide:

>FBpp0070037 type=protein;
loc=3L:complement(join(23337875..23338046,23334364..23334740,23330698..23331015,23329886..23330089,23328714..23329398,23327494..23327645,23326688..23326760,23325889..23325956,23325642,,

Or load it from disk

Choose File no file selected

Program

☐ blastn ☒ tblastn ☐ tblastx ☐ blastp ☐ blastx

tblastn - Peptide vs. Translated Nucleotide

Example workflow: alpha-catenin in the Colorado Potato Beetle

Query Coverage Graph - FBpp00700375889, BLAST Hits 1-8

Subject Coverage Graph - gnl|Leptinotarsa_decemlineata|lepdec

Click on HSP to 'freeze' it

Sort

Filter

qseqid	sseqid	pident	length	mismatch	gapopen	qstart	qend	sstart	send	evalue	bitscore
FBpp00700375889	Scaffold1	73.93	280	51	2	103	360	1970791	1969952	9e-112	388
FBpp00700375889	Scaffold1	78.17	142	31	0	396	537	1966118	1965693	1e-71	234
FBpp00700375889	Scaffold1	78.95	38	8	0	358	395	1966326	1966213	1e-71	61.6
FBpp00700375889	Scaffold1	66.84	193	13	1	684	825	1964288	1963710	7e-66	248
FBpp00700375889	Scaffold1	76.79	112	26	0	530	641	1965356	1965021	1e-40	168
FBpp00700375889	Scaffold1	89.13	92	10	0	15	106	1972073	1971798	2e-32	141
FBpp00700375889	Scaffold1	85.71	42	6	0	822	863	1963456	1963331	1e-10	71.2
FBpp00700375889	Scaffold1	78.26	23	5	0	661	683	1964591	1964523	0.049	42.7

Showing 1 to 8 of 8 entries (filtered from 19 total entries)

BLAST Report FASTA

29 >gnl|Leptinotarsa_decemlineata|lepdec_Scaffold1
30 Length=3793193
31
32 Score = 388 bits (996), Expect = 9e-112, Max
33 Identities = 207/280 (74%), Positives = 232/280
34 Frame = -2
35
36 Query 103 KKTGDAMSIAREFSEDPCCSLKRGNMVRAAI
37 K G AMS+AREFSEDPCCSLKRGNMVRAAI
38 Sbjct 1970791 KIAGTAMSVAREFSEDPCCSLKRGNMVRAAI
39
40 Query 163 EDDLNLKNASSQDELMDNMRFGRNAGELII
41 EDDL KLKNASS EL+DN++ FG+NA EL+
42 EDDLEKLKNASSHGELLDNIAFGQNANELMI
43
44 Query 223 STMLLTASKVYVRHPELDLAKVNRDFILKQVI
45 STMLLTASKVYVRHPEL AK NRD++LKQVI
46 Sbjct 1970431 STMLLTASKVYVRHPELAAKANRDYVLKQVI
47
48 Query 282 AAALDDFD-----EG-
49 AAALDDFD +
50 Sbjct 1970251 AAALDDFDVSHFFIYIS*QIRNIVIVQDHI
51
52 Query 321 LMADADCTDERRRIRVAECNAVRQALQDLL:
53 LMAD+ CTRDERRRIRVAECNAVRQALQDLL:
54 Sbjct 1970071 LMADSSCTDERRRIRVAECNAVRQALQDLL:









2014 - National Agricultural Library

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

Showing 1 to 8 of 8 entries (filtered from 19 total entries)

blastdb	qseqid	sseqid	pident	length
 lepdec	FBpp0070037	Scaffold1	73.93	280
 lepdec	FBpp0070037	Scaffold1	78.17	142
 lepdec	FBpp0070037	Scaffold1	78.95	38
 lepdec	FBpp0070037	Scaffold1	66.84	193
 lepdec	FBpp0070037	Scaffold1	76.79	112
 lepdec	FBpp0070037	Scaffold1	89.13	92
 lepdec	FBpp0070037	Scaffold1	85.71	42
 lepdec	FBpp0070037	Scaffold1	78.26	23

Ldec.genome.10062013_new_ids.fa
Click to view in genome browser

Download

2014 -

[https://apollo.nal.usda.gov/lender_training/browse/?loc=Scaffold1:1966013-1966526&addStores='\"url\"'](https://apollo.nal.usda.gov/lender_training/browse/?loc=Scaffold1:1966013-1966526&addStores='\)

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

Example workflow: alpha-catenin in the Colorado Potato Beetle

The screenshot displays the Apollo genome browser interface. The top navigation bar includes the Apollo logo, menu items (File, View, Help), a search bar, and a 'Login' button. Below the navigation bar is a genomic scale from 0 to 3,500,000. A red vertical line marks a position at approximately 2,000,000. Below the scale are navigation controls (back, forward, zoom in, zoom out) and a dropdown menu for 'Scaffold1'. A search bar shows 'Scaffold1:1966013..1966341 (330 b)' with a 'Go' button. The main display area shows two tracks: 'BLAST+ Results' (red bar) and 'LDEC_v0.5.3-Models' (green bar). The BLAST+ Results track shows a hit for 'FBpp0070037' with a score of 1,966,125. The LDEC_v0.5.3-Models track shows a hit for 'LdecTmpB001070-RA' with a score of 1,966,250. On the left side, the 'Available Tracks' panel is visible, showing a list of tracks with checkboxes. The 'BLAST+ Results' checkbox is checked. The 'LDEC_v0.5.3-Models' checkbox is also checked. A red arrow points to the 'BLAST+ Results' checkbox, and another red arrow points to the 'LDEC_v0.5.3-Models' checkbox.

Available Tracks

filter by text

UC Content

Gaps in assembly

☒ BLAST+ Results

BCM_v0.5.3 47

1. Gene Sets 3

Primary Gene Sets: Protein Coding 1

☒ LDEC_v0.5.3-Models

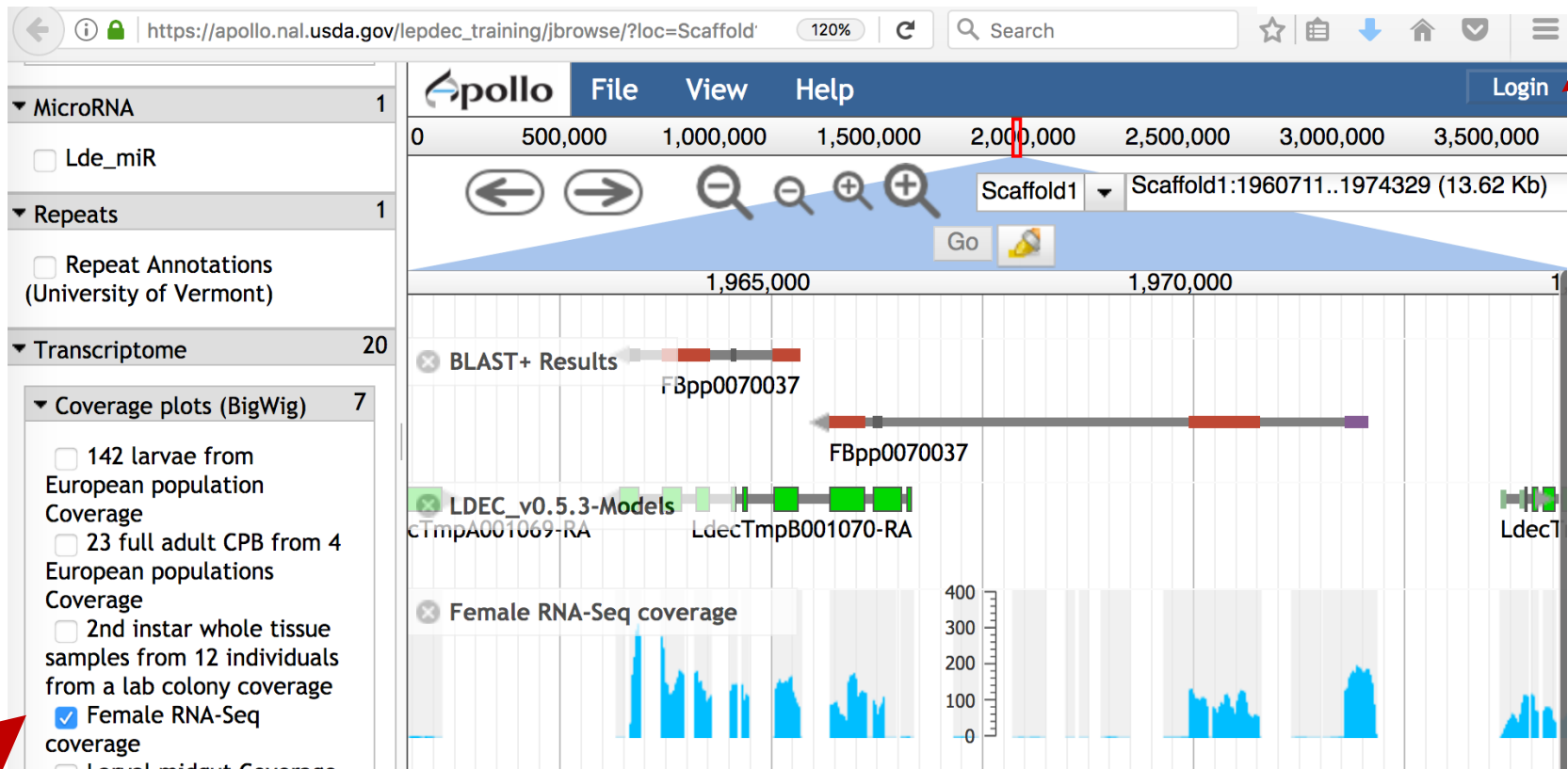
BLAST result track

Gene model track

URL: <https://tinyurl.com/ybf4ehld>

Example workflow: alpha-catenin in the Colorado Potato Beetle

Log in (demo/demo)

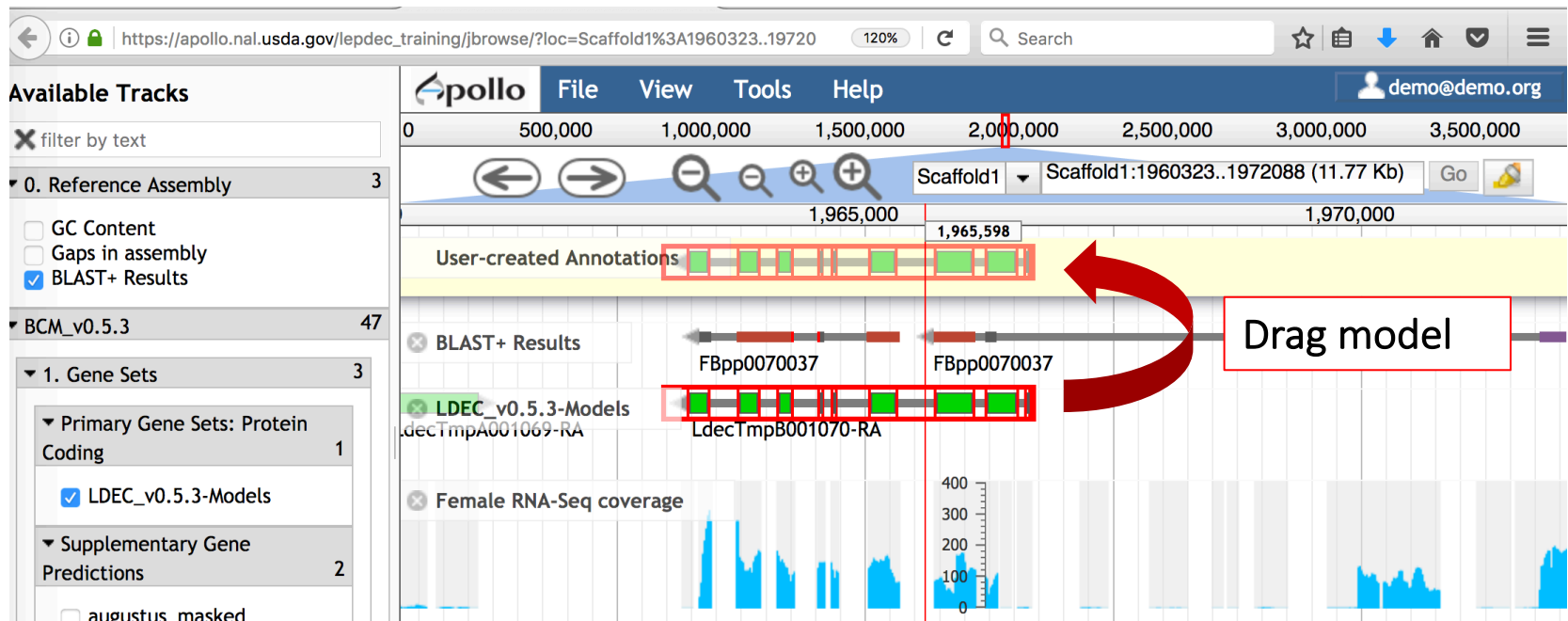


RNA-Seq evidence tracks

URL: <https://tinyurl.com/y8688kgt>

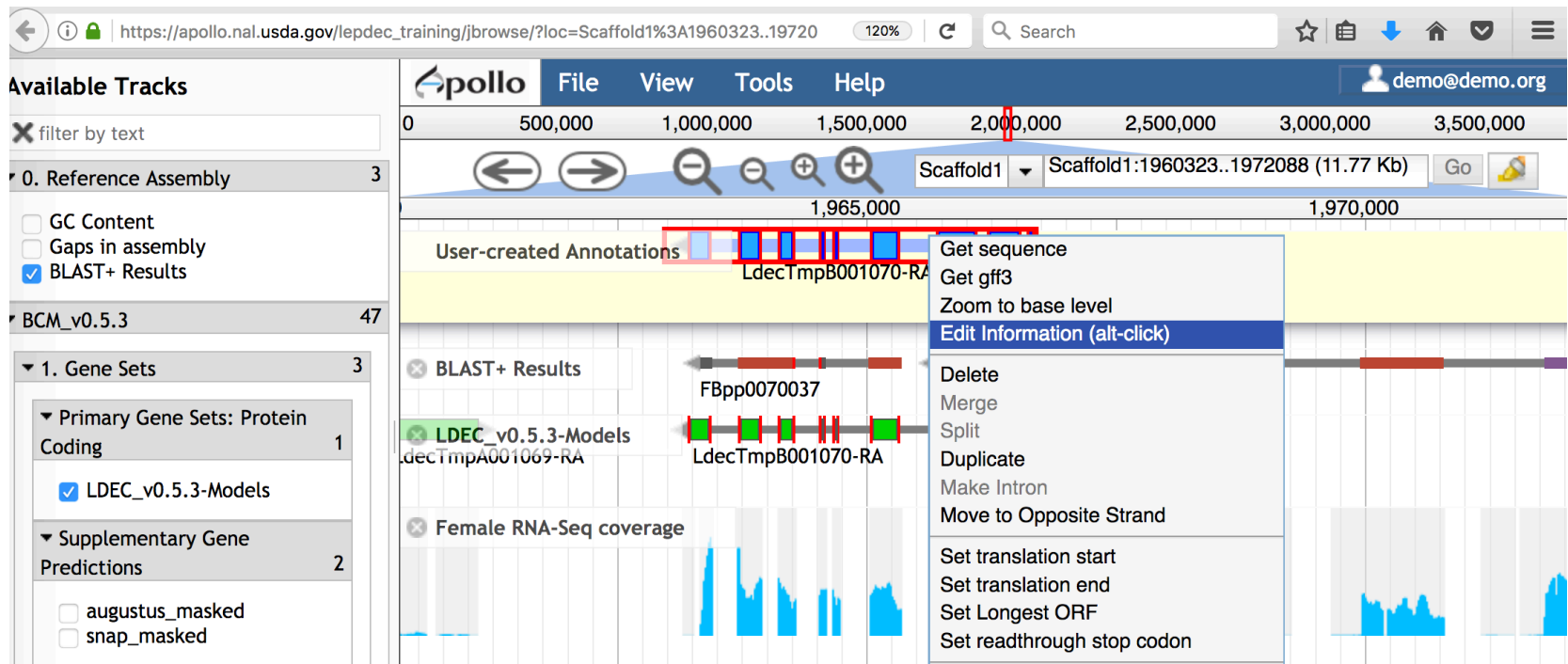


Example workflow: alpha-catenin in the Colorado Potato Beetle



URL: <https://tinyurl.com/y8688kgt>

Example workflow: alpha-catenin in the Colorado Potato Beetle



URL: <https://tinyurl.com/y8688kgt>

Example workflow: alpha-catenin in the Colorado Potato Beetle

Available Tracks

filter by text

0. Reference Assembly

- ☐ GC Content
- ☐ Gaps in assembly
- ☒ BLAST+ Results

BCM_v0.5.3

1. Gene Sets

- Primary Gene Set
- ☒ LDEC_v0.5.3-RA
- Supplementary Gene Set
- ☐ augustus_masked
- ☐ snap_masked

2. Evidence

- Repeats
- ☐ repeatmasker
- ☐ repeatrunner

3. Mapped Proteins

4. Transcriptome

- Assembly
- ☐ est_gff:cufflin

Information Editor (alt-click)

gene

Name

Symbol

Description

Created 2017-05-16

Last modified 2017-05-16

Status

☐ Approved ☐ Delete

DBXRefs

DB	Accession
----	-----------

Add Delete

Replaced Models

Action	Transcript Name
--------	-----------------

Add Delete

mRNA

Name alpha-catenin

Symbol

Description

Created 2017-05-16

Last modified 2017-05-16

Status

☐ Approved ☐ Delete

DBXRefs

DB	Accession
----	-----------

Add Delete

Replaced Models

Action	Transcript Name
replace	LdecTmpB001070-RA

Add Delete

URL: <https://tinyurl.com/y8688kgt>

Post-Annotation QC

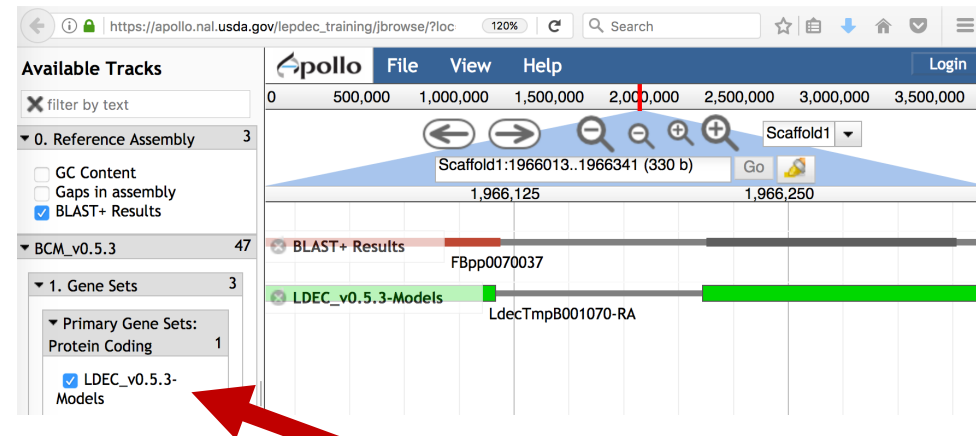
- Manual annotations are run through our Quality Control pipeline
- Some issues need manual intervention
 - Missing required fields
 - Complex splits/merges
 - Incomplete models and those abandoned in process
- Some issues can be automatically corrected
- Iterative process
 - Models requiring inspection 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

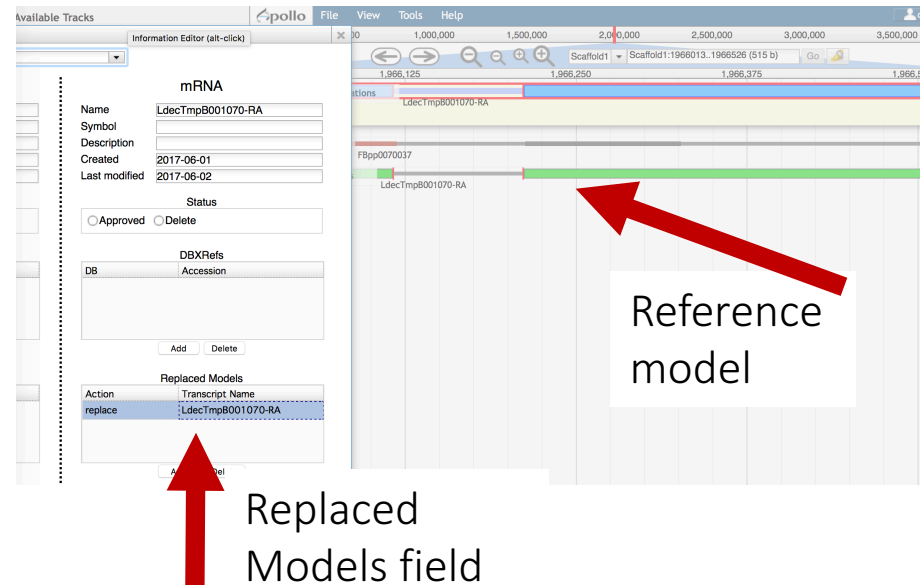
- Requirements:
 - Designate a 'reference gene set' prior to the start of the annotation period
 - Use the 'Replaced Models field' during the manual annotation process



Reference model track

The i5k Workspace 'Replaced Models' field

- Accessible 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>



Need more information?

i5k Workspace@NAL:

- <https://i5k.nal.usda.gov/>

Check us out on GitHub

- <https://github.com/NAL-i5K/>

Questions?

- Email us i5k@ars.usda.gov
- Leave us a note: <https://i5k.nal.usda.gov/contact>

Acknowledgements

The NAL Team

- Christopher Childers*
- Monica Poelchau*
- Gary Moore
- Susan McCarthy
- Chaitanya Gutta
- Yu-yu Lin

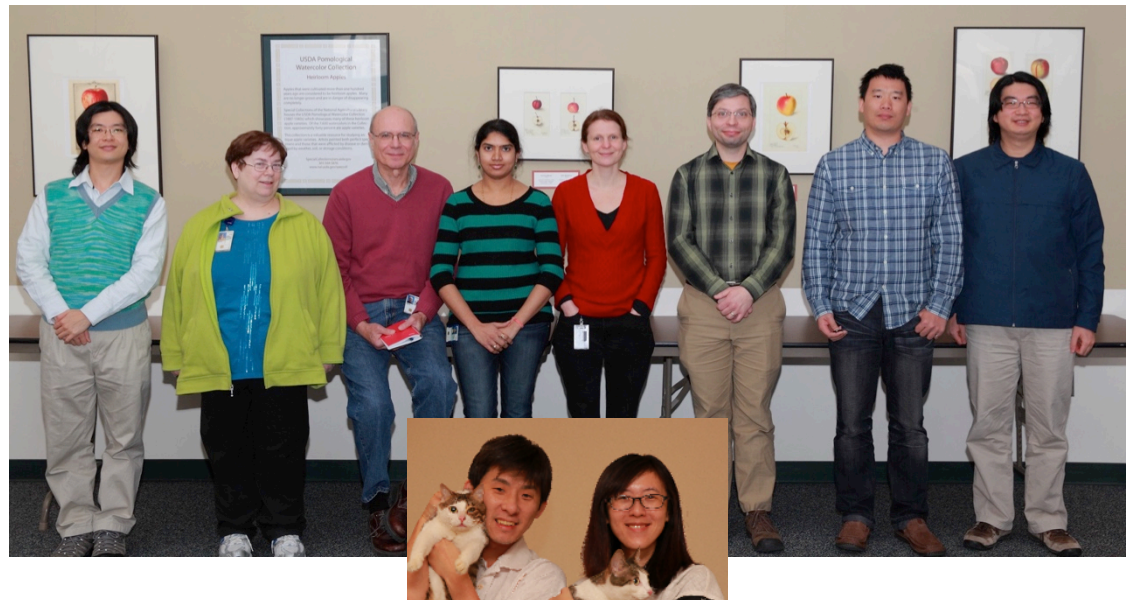
Workspace alumni

- Mei-Ju Chen
- 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
- All of our users and contributors!



Live curation example