

Using Apollo at the i5k Workspace@NAL

NAL USDA-ARS

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

December 19th, 2017



Agenda

- Manual annotation general overview
- 15k Workspace tools for manual annotation
 - BLAST, Clustal, HMMER
 - Apollo
- Manual annotation example: preparation
- Manual annotation live example

Other resources

- Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
 - <https://www.slideshare.net/MonicaMunozTorres/presentations>
 - I recommend these slides if you need more background:
 - <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>
 - Note - there are two versions of Apollo. The i5k Workspace still uses the older version with a slightly different interface
 - If you are new to Apollo, or need a refresher, we **highly recommend** that you review one of her presentations
- The official Apollo annotation guide:
 - <http://genomearchitect.org/users-guide/>
- Other manual curation tutorials:
 - <https://i5k.nal.usda.gov/manual-curation-example>
 - <http://genomecuration.github.io/genometrain/d-feature-curation-crossing/>

Manual annotation general overview

What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Often, but not always: drawing on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model
 - Structural annotation – defining the gene structure (e.g. exon boundaries)
 - Functional annotation – describing the gene function (e.g its name)

Why manually annotate?

- “Incorrect annotations poison every experiment that makes use of them”
- “Worse still, the poison spreads because incorrect annotations from one organism are often unknowingly used by other projects to help annotate their own genomes.”
 - Yandell and Ence 2012, doi:10.1038/nrg3174

General process of manual annotation

1. Select a chromosomal region of interest (e.g. scaffold)
 1. E.g. find sequence of interest from one or several other species, and align against proteins or genome sequence from your species
2. Select appropriate evidence (tracks in Apollo, or your own files)
3. Determine whether a feature in your evidence provides a reasonable starting gene model
 1. If yes: select and drag the feature to the 'user-created annotations' area, creating an initial gene model. If necessary use editing functions to adjust the model.
 2. If not – get in touch with us!
4. Edit model if necessary
5. Check your edited gene model for integrity and accuracy by comparing it with available homologs
 1. Verify that the gene model is the best representation of the underlying biology
6. Repeat steps 1 through 5 as needed to refine model
7. Add annotation details in the “Information Editor”
 1. Replaced model, name, symbol, other comments

Adapted from <https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015>

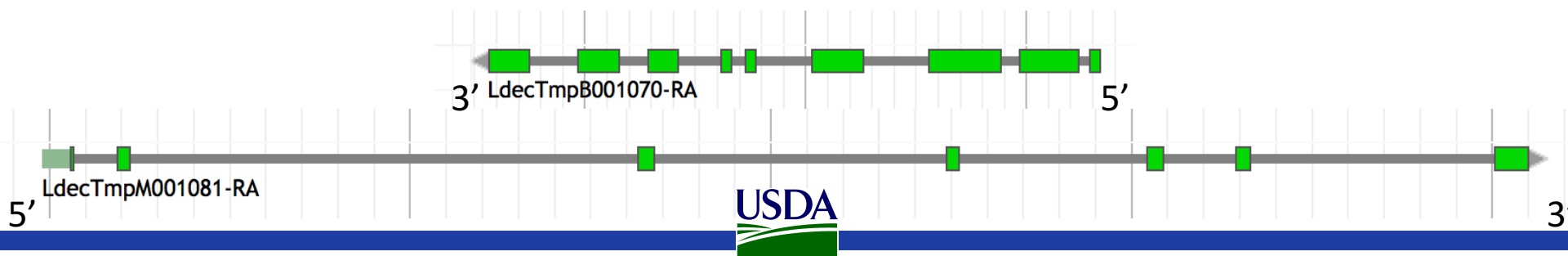
I5k Workspace ‘Etiquette’

1. Use Apollo to improve a gene model in an i5k Workspace assembly.
 1. If you just want to practice – use one of our training instances.
 1. <https://i5k.nal.usda.gov/jbrowseapollo-training>
 2. If you just want to view the data – you probably can get what you want without using Apollo. All of the data that we host is public.
2. Your annotation work is a community effort.
 1. If you notice that someone else is working on your model of choice, get in touch with them (or us) and collaborate – don’t make a 2nd model or delete the other model.
 2. Keep in mind that your work will be used by the scientific community once you’re done.
3. If you publish any of your work generated in the i5k workspace:
 1. Get in touch with the genome contact first (you can find the contact info on the organism page; <https://i5k.nal.usda.gov/species>);
 2. Please cite the i5k Workspace paper! This helps us continue to exist.
 1. <https://doi.org/10.1093/nar/gku983>

Manual annotation: i5k Workspace tools

First, some conventions

- HSP – High scoring pair in BLAST/BLAT alignments
 - The ‘Hits’ in an alignment result set
 - A subsection of a pair of sequences with sufficient score
 - HSPs can change based on the alignment parameters
- Five prime end and three prime end
 - Based on direction of transcription
 - Initiation site is at the five prime end
 - Stop codon is at the three prime end
- In the genome browser, arrowheads indicate direction



JBrowse and Apollo

The screenshot shows the JBrowse web interface with the Apollo extension. The interface includes a top menu bar with 'File', 'View', 'Tools', and 'Help'. A left sidebar contains an 'Available Tracks' panel with a search filter and a track selector. The main area displays a genomic track for Scaffold79, showing various data layers like GC Content, Gaps in assembly, and EAF v0.5.3-Models. Annotations are visible as colored blocks on the tracks. Arrows point from text labels to specific interface elements: 'Bookmark /share URL' points to the address bar; 'File: Add your own files' points to the 'File' menu; 'View: Change coloring scheme' points to the 'View' menu; 'Tools: Search using BLAT' points to the 'Tools' menu; 'Locate where you are on the scaffold' points to the scaffold selection dropdown; 'Search for a gene or location' points to the search bar; 'Log in/out' points to the user profile; 'Track selector' points to the track selector in the sidebar; 'Turn tracks on/off' points to the track checkboxes; 'Find information about tracks' points to the track names in the sidebar; 'Zoom in/out' points to the zoom controls; and 'User-created annotations track' points to the 'User-created Annotations' track.

Annotations:

- Bookmark /share URL
- File: Add your own files
- View: Change coloring scheme
- Tools: Search using BLAT
- Locate where you are on the scaffold
- Search for a gene or location
- Log in/out
- Track selector
- Turn tracks on/off
- Find information about tracks
- Zoom in/out
- User-created annotations track

JBrowse is a web- based genome browser

- Visualize features that are mapped to a genome
- These features are displayed as tracks
- Many different types of data may be displayed

Apollo adds editing functions to JBrowse

- Manual gene curation
- Changes automatically saved back to server
- Edits are visible to other annotators in real-time
- Editing history is tracked

i5k Workspace BLAST: one way to access Apollo

The screenshot shows the i5k Workspace BLAST interface. The top navigation bar includes the i5k@NAL logo and links for Tools, About Us, and Contact. The main content area is titled "BLAST Databases" and is divided into three sections: "Organisms", "Nucleotide", and "Peptide".

- Organisms:** A list of organisms with checkboxes. *Eurytemora affinis* is selected, indicated by a blue highlight and a checkmark.
- Nucleotide:** Two options are listed: "Genome Assembly - Eaff_11172013.genome_new_ids.faa" (selected with a blue checkmark) and "Transcript - EAFF_new_ids.fna".
- Peptide:** One option is listed: "Protein - EAFF_new_ids.faa".

Below the database selection is the "Query Sequence" section. It contains a text area with a peptide sequence: >FBpp0070332 MDNCDQDASFRLSHIKEEVKPDISQLNDSNN SSFSPKAESPVPFMQAMSMVHVLPGSNSASS NNNAGDAQMAQAPNSAG GSAAAAVQQYPPNHPLSGSKHLCSICGDRA SGKHYGVVSCGCKGFFKRTVRKDLTYACRE. Below the text area is a "Browse..." button and the text "No file selected.".

At the bottom is the "Program" section, which includes radio buttons for "tblastn" (selected), "otblastn", "tblastx", "blastp", and "blastx", along with "Reset" and "Search" buttons.

Annotations with arrows point to the following elements:

- "Select organism" points to the *Eurytemora affinis* selection in the Organisms list.
- "Paste or upload query sequence(s)" points to the query sequence text area.
- "Program is automatically selected" points to the selected "tblastn" radio button.
- "Select organism-specific database" points to the "Genome Assembly - Eaff_11172013.genome_new_ids.faa" selection in the Nucleotide section.
- "BLAST against the genome assembly to view HSPs in Jbrowse" points to the "Genome Assembly" selection in the Nucleotide section.

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

i5k Workspace BLAST: one way to access Apollo

The screenshot displays the i5k Workspace BLAST interface, which is divided into four main panels:

- Query Coverage Graph:** A bar chart showing coverage for query hits 1-9. The x-axis represents position (50 to 500) and the y-axis represents coverage (0 to 500). The bars are colored in a gradient from red to yellow.
- Subject Coverage Graph:** A bar chart showing coverage for subject hits. The x-axis represents position (255.1k to 255.35k) and the y-axis represents coverage (0 to 1). The bars are colored in a gradient from red to yellow.
- BLAST Results Table:** A table showing BLAST results for query 103. The table has columns for blastdb, qseqid, sseqid, pident, length, mismatch, gapopen, and qstart. The table shows 9 entries (filtered from 55 total entries). The first entry is highlighted in yellow.
- Apollo Genome Browser:** A view of the genome browser showing the Scaffold427. The browser displays various tracks including User-created Annotations, Seven-Up Isoform B, and TF1_accepted_hits (Coverage Plot). The browser also shows a list of available tracks on the left side.

Annotations:

- Click on blue blastdb icon next to your favorite HSP
- Blast results are displayed in Apollo
- BLAST result page with 4 panels

HMMER and Clustal

- Use HMMER to detect remote protein homologs
- <https://i5k.nal.usda.gov/webapp/hmmer/>
- Use Clustal to perform multiple sequence alignments
- <https://i5k.nal.usda.gov/webapp/clustal/>

Tips and Tricks

- The i5k Workspace BLAST results persist for one week
 - You can bookmark and share searches
 - BLAST HSPs are ‘draggable’ and can be used in annotations
- Jbrowse/Apollo URLs can be shared
 - Allow you to share the exact view (including active tracks) with others
 - Great for troubleshooting with collaborators
- In Apollo “walk” feature boundaries
 - Square brackets walk exon boundaries: [and]
 - Curly brackets walk gene boundaries: { and }
- In Apollo, you can pin tracks to the top
- If you know the name or ID of the gene that you’d like to annotate, you can paste it into the search box in Apollo to navigate to it

Manual annotation example: preparation

Annotation Example

- Phosphoenolpyruvate carboxykinase (pepck) in the copepod *Eurytemora affinis*
- Pepck catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).
- More information about the copepod:
https://i5k.nal.usda.gov/Eurytemora_affinis
- Apollo URL:
<https://apollo.nal.usda.gov/euraff/jbrowse/>
 - Note: There are no demo accounts for this species

Notes on *E. affinis* genome/browser

- Big advantage for annotation: lots of RNA-Seq and transcriptome data are available to use as contributing evidence for your gene models
 - Includes strand-specific RNA-Seq
- Disadvantage: No close reference genomes, so it may be harder to find homologs for your genes of interest to inform your annotations.

Available tracks for *E. affinis*

The screenshot displays the Apollo genome browser interface. On the left, a sidebar titled 'Available Tracks' lists various genomic data categories and their counts. The 'Primary Gene Sets: Protein Coding' section is expanded, showing 'EAFF_v0.5.3-Models' selected. The 'Transcriptome' section is also expanded, showing 'Assembly' and 'Coverage Plots (BigWig)' options. The main panel on the right shows a genomic track with a blue line representing the 'EAFF_v0.5.3-Models' gene set. The track is labeled 'EAFF_v0.5.3-Models' and 'Eaff1m'.

Track Category	Count
0. Reference Assembly	2
BCM_v0.5.3	47
1. Gene Sets	3
Primary Gene Sets: Protein Coding	1
EAFF_v0.5.3-Models	1
Supplementary Gene Predictions	2
2. Evidence	2
3. Mapped Proteins	41
4. Transcriptome	1
Transcriptome	26
Assembly	2
075_zz91_transcriptome	1
Mixture of males and females: cufflinks_IGS_UMA1	1
Coverage Plots (BigWig)	10
Mapped Reads	7
RNA-Seq of Untreated Mixed Adults, digitally normalized	1
TF1_accepted_hits	1
TM_accepted_hits	1
UMA_accepted_hits	1
VAF_accepted_hits	1
VAJU_accepted_hits	1
VAM_accepted_hits	1
Splice Junctions	7

- Baylor Maker annotations:
 - Primary Gene Set:
 - EAFF_v0.5.3-Models
 - Other tracks that were used to generate the primary gene set
- Transcriptome/RNA-Seq
 - Transcriptome assemblies
 - Coverage plots, Mapped RNA-Seq data, Splice junctions
 - Some of the RNA-Seq libraries are stranded

Choosing reference proteins: *D. melanogaster* pepck in UniProt

UniProtKB - P20007 (PCKG_DROME)

Display

- Entry
- Publications
- Feature viewer
- Feature table
- All None
- Function

BLAST Align Format Add to basket History

Protein | Phosphoenolpyruvate carboxykinase [GTP]
Gene | Pepck
Organism | *Drosophila melanogaster* (Fruit fly)
Status | Reviewed - Annotation score: ●●●○○○ - Experimental evidence at transcript levelⁱ

Annotation score is a heuristic for annotation quality

Organism-specific databases

FlyBaseⁱ FBgn0003067. Pepck.

Subcellular locationⁱ

Flybase is another great resource

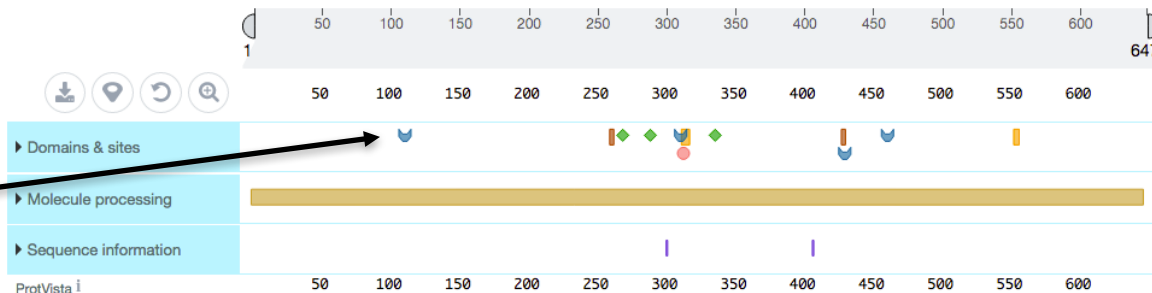
UniProtKB - P20007 (PCKG_DROME)

Display

- Entry
- Publications
- Feature viewer
- Feature table

BLAST Align Format Add to basket History

Feeds



Feature viewer gives graphical view of domains and sites

Catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).

Source: <http://www.uniprot.org/uniprot/P20007>

Choosing reference proteins: *Daphnia pulex* Pepck

- GenBank record:

<https://www.ncbi.nlm.nih.gov/protein/EFX80236.1>

```
.....
Lynch,M., Boore,J.L. and Grigoriev,I.V.
CONSRTM  US DOE Joint Genome Institute (JGI-PGF)
TITLE     Direct Submission
JOURNAL   Submitted (02-FEB-2011) US DOE Joint Genome Institute, 2800
          Mitchell Drive, Walnut Creek, CA 94598-1698, USA
COMMENT   Method: conceptual translation.
FEATURES             Location/Qualifiers
     source            1..652
```

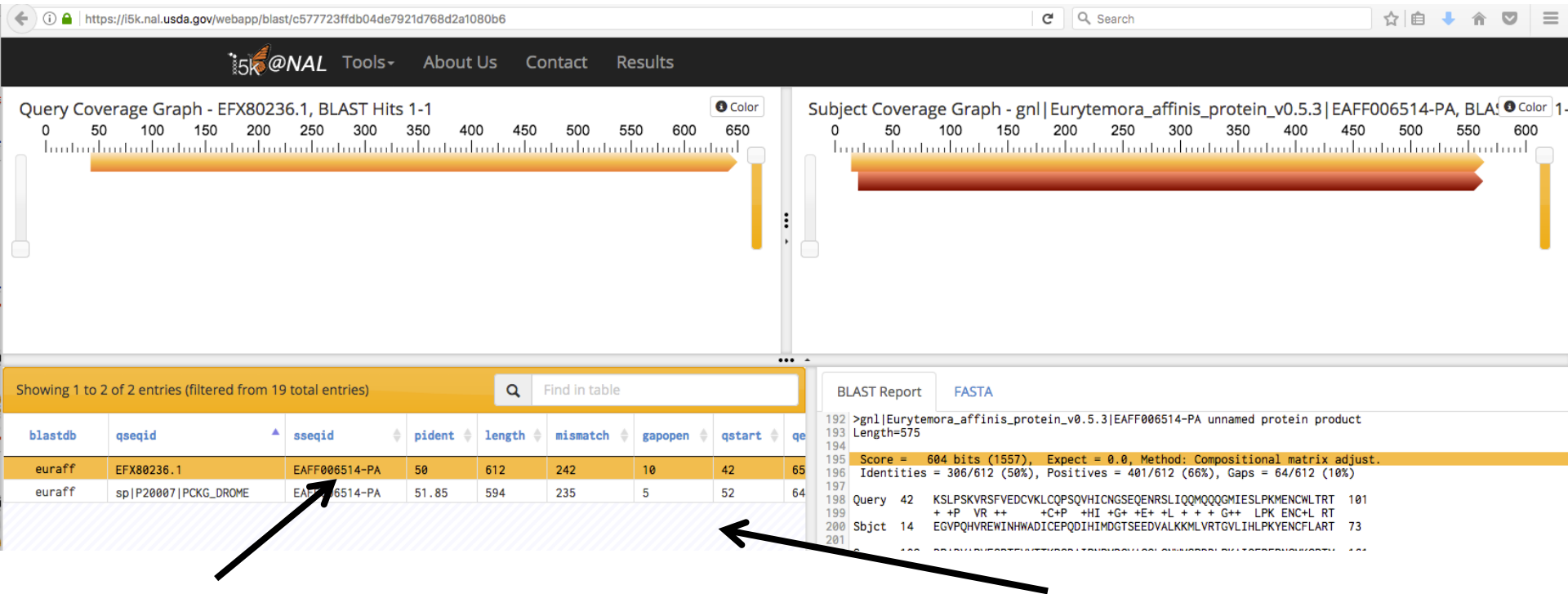
← Treat with caution!!!

Phosphoenolpy
carboxykinase,
(daphnia Phosp
carboxykinase)
(daphnia Phosp
carboxykinase)

Manual annotation live example

BLAST dmel, dpul proteins against *E. affinis* proteins

<https://i5k.nal.usda.gov/webapp/blast/>



Copy the protein 'base name'
EAF006514 for searching in Apollo

Results are filtered by e-value; only
one protein in the *E. affinis* dataset has
a significant match

Result URL: <https://i5k.nal.usda.gov/webapp/blast/90c0f04749af4d028065c0739068ae10>

Modify *E. affinis* model sequence in Apollo

- Go to Apollo URL:
<https://apollo.nal.usda.gov/euraff/jbrowse/>
 - Find mRNA of EAFF006514-PA in genome browser by pasting EAFF006514 into search box, selecting EAFF006514-RA
- Log in to Apollo
- Drag EAFF006514-RA into the yellow annotation track
- Check available evidence for model

Another approach: BLAST against the genome

<https://i5k.nal.usda.gov/webapp/blast/>

The screenshot displays the i5k@NAL BLAST web interface. At the top, there are navigation links: Tools, About Us, Contact, and Results. Below the navigation bar, there are two coverage graphs: 'Query Coverage Graph - EFX80236.1, BLAST Hits 1-21' and 'Subject Coverage Graph - gnl| Eurytemora_affinis| euraff_Scaff'. The main content area shows a table of BLAST hits. The table has columns: blastdb, qseqid, sseqid, pident, length, mismatch, and gapope. The first row is highlighted in yellow. A blue 'blastdb' button is next to the first row. A tooltip points to this button, stating: 'Eaff_11172013.genome_new_ids.fa Click to view in genome browser'. Below the table, there are filters and a 'Download' button. On the right side, there is a 'BLAST Report' section with 'FASTA' format. It shows three hits with their respective scores, identities, and positives. The first hit is highlighted in yellow.

blastdb	qseqid	sseqid	pident	length	mismatch	gapope
euraff	Eaff_11172013.genome_new_ids.fa	Scaffold133	56.41	39	17	0
euraff	sp P20007 PKG_DROME	Scaffold133	62.5	40	15	0
euraff	EFX80236.1	Scaffold133	80	30	6	0
euraff	sp P20007 PKG_DROME	Scaffold133	78.12	32	7	0
euraff	EFX80236.1	Scaffold133	44.59	74	24	2
euraff	sp P20007 PKG_DROME	Scaffold133	46.15	78	25	2
euraff	EFX80236.1	Scaffold133	38.46	26	16	0
euraff	EFX80236.1	Scaffold133	72.34	47	13	0

Showing 1 to 9 of 39 entries

Download

BLAST Report FASTA

310 Sbjct 1491497 KYIVAAFPSACGKTNLMMQRLPGYKV 1491414

311

312

313 Score = 53.1 bits (126), Expect = 9e-06, Method: Compositional matrix adjust.

314 Identities = 22/39 (56%), Positives = 29/39 (74%), Gaps = 0/39 (0%)

315 Frame = -3

316

317 Query 79 QMQQQGMIESLPKMENCLTRTPADVARVESRTFVVT 117

318 +++ G++ LPK ENC+L RTDP DVAR ESRTF+ T

319 Sbjct 1496800 KMLVRTGVLHLPKYENCFLARTDPKDVARTESRTFIST 1496684

320

321

322 Score = 52.0 bits (123), Expect = 2e-05, Method: Compositional matrix adjust.

323 Identities = 23/28 (82%), Positives = 24/28 (86%), Gaps = 0/28 (0%)

324 Frame = -2

325

326 Query 300 KYIAAFPSACGKTNLMLTPTLPYKV 327

327 KYI A FPSACGKTNLAM+ P LPGYKV

328 Sbjct 1491233 KYIVAVFPSACGKTNLMMQRLPGYKV 1491150

329

330

331 Score = 52.0 bits (123), Expect = 2e-05, Method: Compositional matrix adjust.

332 Identities = 23/28 (82%), Positives = 25/28 (89%), Gaps = 0/28 (0%)

333 Frame = -2

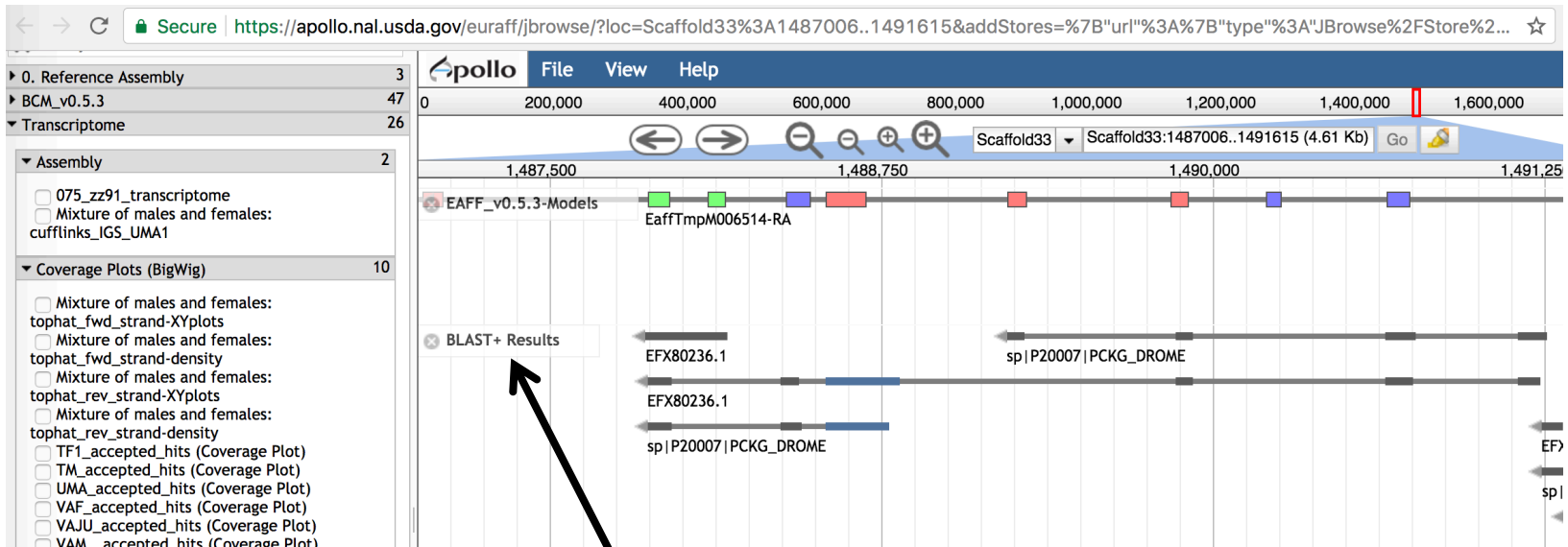
334

Click on blue blastdb button next to your favorite HSP to view it in JBrowse

BLAST result URL: <https://i5k.nal.usda.gov/webapp/blast/cb2ec8536f8d400595842f710fe8f2c2>



Another approach: BLAST against the genome



BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

Apollo result URL: <https://tinyurl.com/y8fdpzyn>

Create annotation in user-created annotations track

The screenshot shows the Apollo genome browser interface. On the left, the 'Available Tracks' panel lists '0. Reference Assembly' (2), 'BCM_v0.5.3' (47), 'Transcriptome' (26), and 'Assembly' (2). The 'Assembly' track is expanded, showing '075_zz91_transcriptome' and 'Mixture of males and females:'. The main view displays a genomic track for 'Scaffold33' with a zoomed-in region from 1,485,000 to 1,500,000. The track shows 'Eaff_v0.5.3-Models' with features like 'EaffTmpM006513-RA', 'EaffTmpM006514-RA', and 'EaffTmpM006515-RA'. A 'Login' button is visible in the top right corner of the Apollo header.

Log in with
your
Apollo
credentials

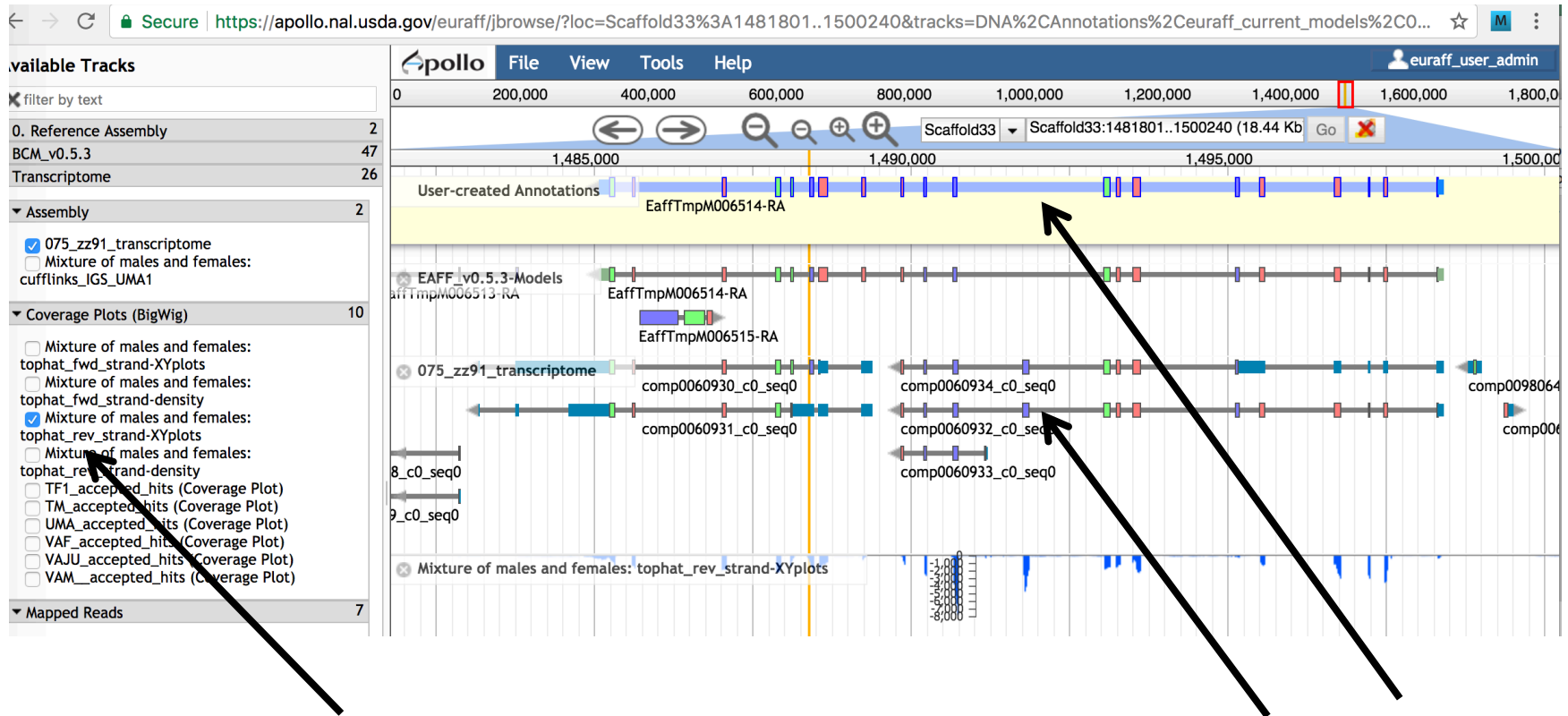
The screenshot shows the Apollo genome browser interface after logging in. The 'Available Tracks' panel is the same. The main view shows the same genomic track for 'Scaffold33' with the zoomed-in region from 1,485,000 to 1,495,000. A new track, 'User-created Annotations', has been added and is highlighted in yellow. This track contains a red box around the 'EaffTmpM006514-RA' feature, indicating it has been dragged into the user-created annotations track. The 'Eaff_v0.5.3-Models' track is still visible below it.

Drag model EaffTmpM006514-
RA to User-created Annotations
track

Modify *E. affinis* model sequence in Apollo

- Questions:
 - What evidence do you choose to check the integrity of the model?
 - Do you need additional evidence?
 - How do you evaluate whether the protein sequence is as complete as it can be?
 - Should you add/modify UTRs?

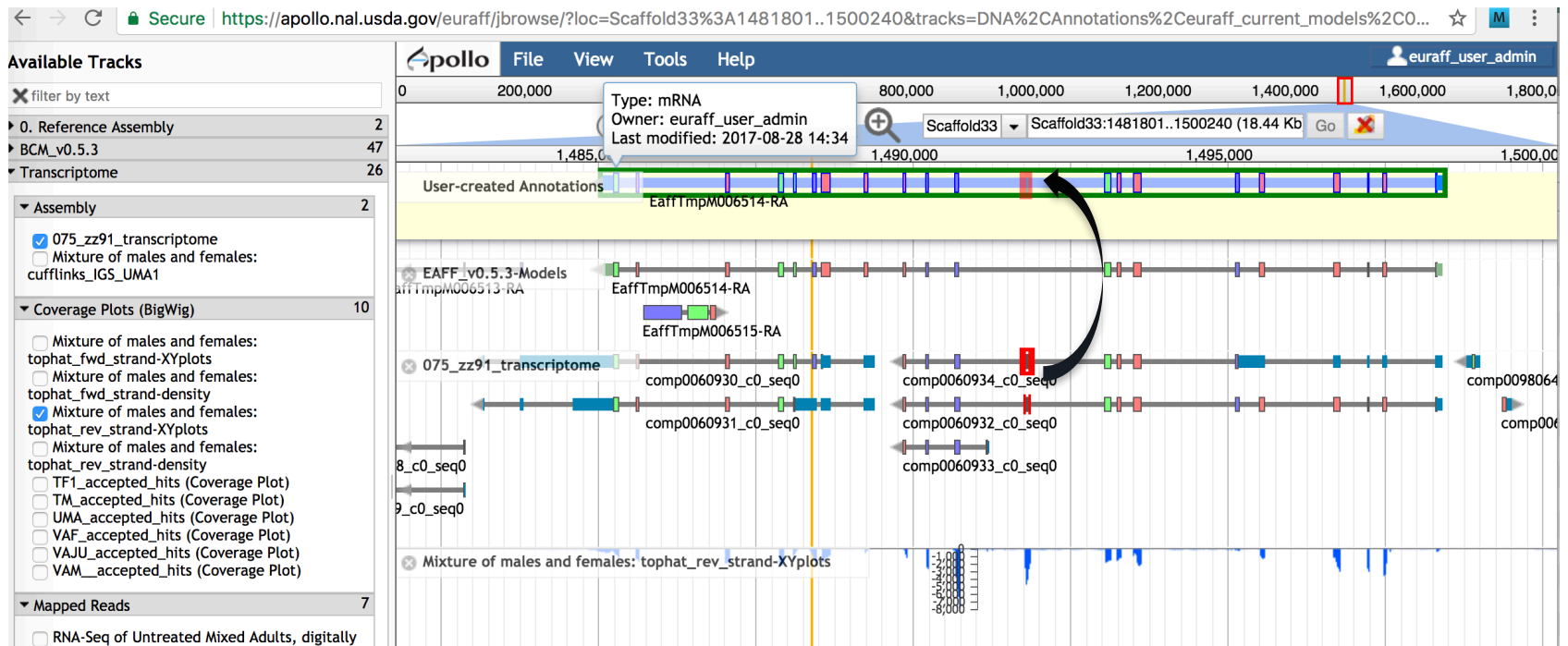
View available evidence



Model is on the reverse strand, so we can take advantage of the stranded RNA-Seq available for this species

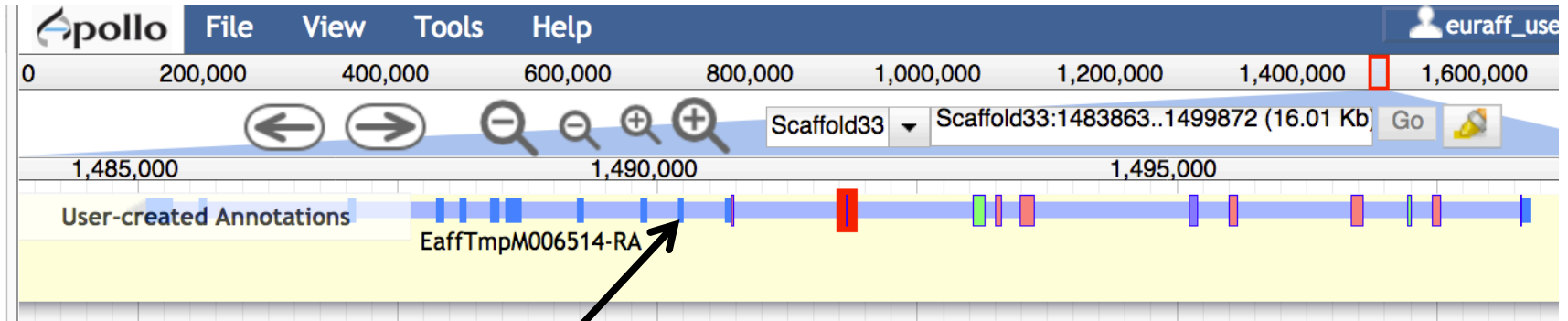
RNA-Seq and transcriptome tracks suggest that one exon is missing

Add an exon to the model



Drag exon from
transcriptome track
into new gene model

Adjust exon boundary



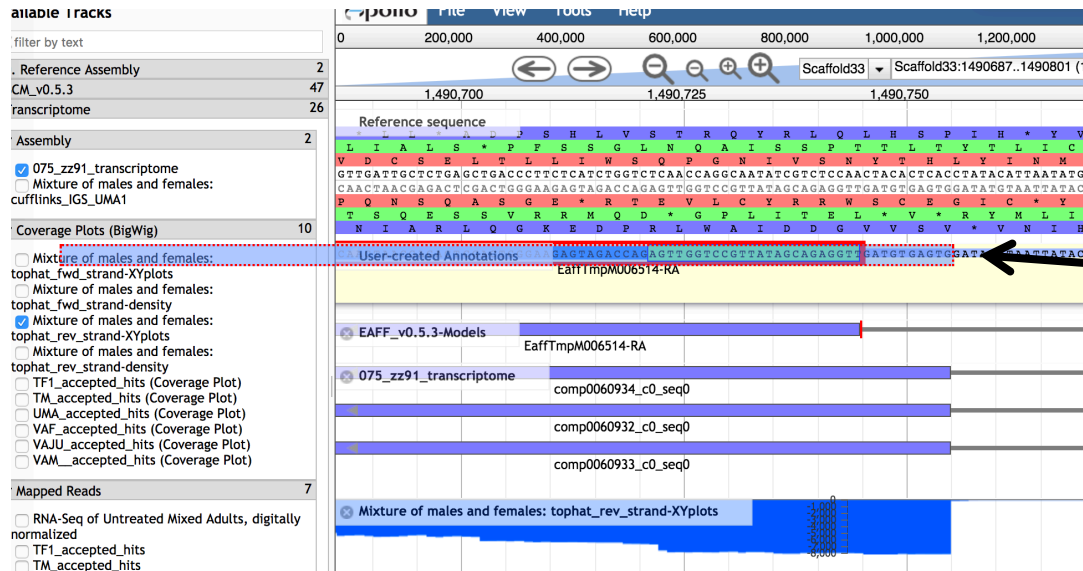
CDS sequence is now UTR –zoom in to investigate



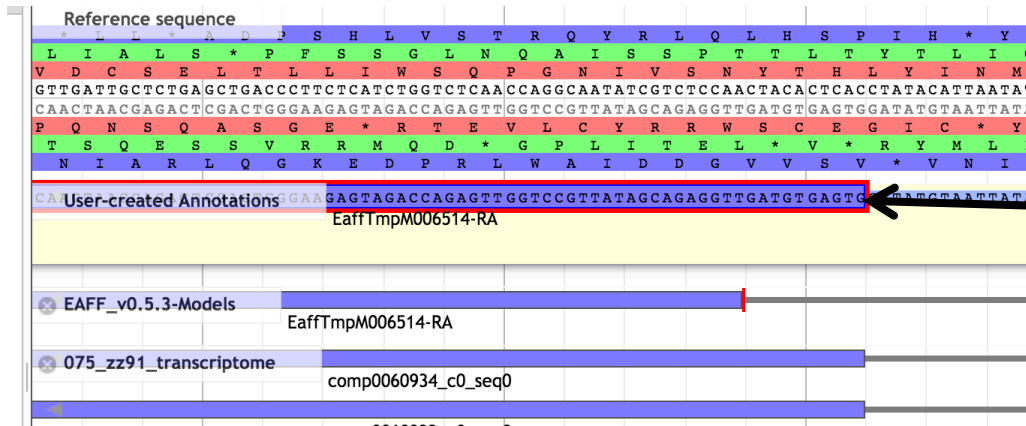
CDS frame has changed from purple to green—we need to fix this

RNA-Seq suggests we need to adjust exon boundary

Adjust exon boundary



Drag exon boundary to match RNA-Seq and transcriptome tracks



Fixed both reading frame and exon boundary

Evaluate new protein sequence

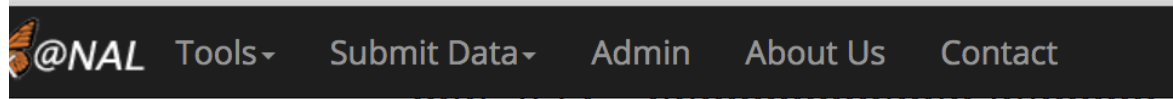
- Blast modified EAFF006514-PA sequence to NCBI's nr database
 - Make sure it doesn't match a potential contaminant
 - Get an idea whether you have the right sequence
 - Blastp home:
 - https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
 - Result URL:
 - <https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=3FWVAS73015> (expires end of day 12/19)
- Once contamination is ruled out, it's better to align your sequence against a smaller set of high-quality proteins
- If you notice that parts of the protein are missing, check the 'Gaps in assembly' track in the browser

Evaluate new protein sequence

- Get *E. affinis* pepck protein sequence from old model and new model
- Align new and old sequence to dmel and dmag protein sequences
 - Clustal (<https://i5k.nal.usda.gov/webapp/clustal/>)
 - Can also use NCBI Blast
- Check alignment extent, %ID

Clustal Results

:/i5k.nal.usda.gov/webapp/clustal/105850a3594e4234a21b07d93cbbd71



euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
IS-----VGDDIAWLRPDEKQQLRAI
ISGITNSQGEKKYIVAAFPSCGKTNLAMMQRLP-----VSVVGDDIAWLRPDEKQQLRAI
ILGITDPKGEKKYITAAFPSCGKTNLAMLNPSLANYKVECVGDDIAWMKFD SQVLRAI
ILGITNPQGQKKYIAAAPPSCGKTNLAMLTPTLPGYKVECVGDDIAWMHFDKEGRLRAI
*                               *****:*.:* ****
```

New exon added

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
NPENGFFGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMDKPK
NPENGFFGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMDKPK
NPENGFFGVAPGTSMETNPVIA-----MNTVFKNTIFTNVASTSDGGVFWEGMESSLA
NPENGFFGVAPGTNYATPNACYNFFLYAMLTIQKNTIFTNVAKTSDGGVFWEGLEKEV-
*****. : ** * : : * : * : * : * : * : * : * : * : * : *
```

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
ERSSCIDWK GK-PWRPTSSNPAHPNSRFCTPLLNC PVLDESAEDPAGVPIAAILFGGRR
ERSSCIDWK GK-PWRPTSSNPAHPNSRFCTPLLNC PVLDESAEDPAGVPIAAILFGGRR
PNVQITDWLGK-PWTKDSGKPAHPNSRFCTPAAQCPIIDEAWEDPAGVPI SAMLFGGRR
TGV DITSWLGDANWTKSSGKPAHPNSRF CAPASQCPIIDPLWESPEGVPI DAILFGGRR
. . * * . * : * : * : * : * : * : * : * : * : *
```

euraff_old_pepck
euraff_new_pepck
sp|P20007|PCKG_DROME
EFX80236.1

```
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFK GKQIMHDPFSMRPFFG-----HW
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFK GKQIMHDPFSMRPFFG-----HW
PAGVPLIYEARDWTHGVFI GAAMRSEATAAAEHK GKVIMHDPFAMRPFFGYNFGDYVAHW
PRGVPLVYEA LNWKHG VFVGASVSSEATAAAEHK GRSIMHDPFAMRPFFGYNAGNYLGHW
* ****:* * . * ****:* * : *****.* : *****:***** **
```

Another exon might be missing (we're not going to handle this today)

- Clustal result URL:
<https://i5k.nal.usda.gov/webapp/clustal/03f944fd02f245a8ad12d68cb9681776>
- Scroll to bottom of page and click 'colorful' to see color-coded alignment



Using the Information Editor

- Select the model in Apollo, then right-click, and select 'Edit Information' from the drop-down menu
 - Use the 'mRNA' section
 - Name: We recommend the UniProt naming guidelines:
 - <http://www.uniprot.org/docs/nameprot>
 - If a naming convention exists, use it (e.g. for gene families)
 - Use name from an orthologous protein if you are sure that your gene model is orthologous.
 - Document your justification for the name in the Comments field (e.g. "88% sequence similarity via blastp to D. melanogaster pepck P20007")
 - If you create a new name, it should be unique and attributed to all orthologs (as far as possible)
 - Comments – Document what changes you performed, and your justification for the name. These notes will be visible in the OGS, so make sure that others understand them
 - Replaced Models Field – the Maker model (EAFF_v0.5.3) that your new model will replace in the OGS

Using the Information Editor

Information Editor (alt-click)

Select mRNA: Phosphoenolpyruvate carboxykinase

gene		mRNA	
Name		Name	Phosphoenolpyruvate carboxykinase
Symbol		Symbol	pepck
Description		Description	
Created	2017-08-28	Created	2017-08-28
Last modified	2017-08-28	Last modified	2017-08-28
Status		Status	
<input type="radio"/> Approved <input type="radio"/> Delete		<input checked="" type="radio"/> Approved <input type="radio"/> Delete	
DBXRefs		DBXRefs	
DB	Accession	DB	Accession
<input type="button" value="Add"/> <input type="button" value="Delete"/>		<input type="button" value="Add"/> <input type="button" value="Delete"/>	
Replaced Models		Replaced Models	
Action	Transcript Name	Action	Transcript Name
		replace	EaffTmpM006514-RA

The Replaced Models field

- We use the information in this field to generate a merged, non-redundant gene set from the manually curated models and the official or primary gene set
- Your official or primary gene set is listed in the category field of the track selector
- If you don't know what your project's gene set is, contact us!

mRNA

Name Phosphoenolpyruvate carboxykinase
Symbol pepck
Description
Created 2017-08-28
Last modified 2017-08-28

Status
☒ Approved ☐ Delete

DBXRefs

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

Add Delete

Replaced Models

Action	Transcript Name
replace	EaffTmPM006514-RA

Replaced Models field

<https://i5k.nal.usda.gov/apollo-replaced-models-field-explanations-and-examples>

Checklist for accuracy and integrity

- Check start, stop and exon boundaries (splice sites)
 - Try to fix non-canonical splice sites if possible
- Check if you can annotate UTRs (e.g. using RNA-Seq data)
- Check for gaps in the genome
- If you change the genome sequence, add a justification comment to the corresponding gene model
- Use BLAST or a multiple sequence aligner
 - To look at completeness of model
 - To verify the appropriateness of the gene name
- In the Information editor **mRNA** field
 - Fill in the Replaced Model for the **Maker** gene (EAFF_v0.5.3-Models)
 - Update the Name if appropriate
 - Add comments that describe
 - your evidence for the annotation
 - Modifications that you made to the gene model

cf. <https://www.slideshare.net/MonicaMunozTorres/editing-functionality-apollo-workshop>

What happens to my annotation when I'm done?

- This depends on the genome project that you're working on.
- If the genome coordinator has asked us to generate an OGS (Official Gene Set), we will do so
 - We are still working on this process, so if you ask us to do this, 1) it will take some time, and 2) we will probably ask you for co-authorship if you publish a paper on the OGS.
 - We are working on a pipeline to submit Official Gene Sets to GenBank, where they will be archived/accessioned
- Otherwise, don't assume that your annotation will be archived.
 - If you need it to be, get in touch with us and we'll figure out what to do.
- Get in touch with us and the genome project coordinator if you're not sure about the status of a genome project.
- <https://i5k.nal.usda.gov/data-management-policy>

Upcoming webinars (tentative schedule)

- February: i5k Workspace roadmap and Q&A
- April: Orientation and resources for project coordinators
- June: Overview of i5k Workspace resources

Thank you!

The NAL Team

- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

I5k Workspace alumni

- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

i5k Workspace@NAL advisory
committee

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

