

# Using Apollo at the i5k Workspace@NAL

NAL USDA-ARS

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

April 24<sup>th</sup>, 2018



# 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. Name, symbol, other comments

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

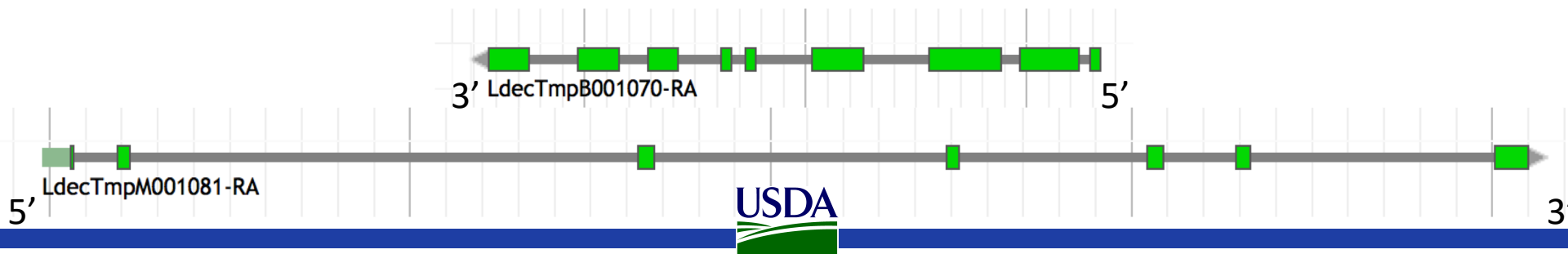
# I5k Workspace ‘Etiquette’

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 image shows a screenshot of the JBrowse and Apollo web interface. The interface is divided into several sections: a left sidebar for track selection, a top menu bar, a main viewing area, and a right sidebar for user actions. Annotations with arrows point to various features:

- Bookmark /share URL**: Points to the address bar showing the URL.
- Track selector**: Points to the 'Available Tracks' sidebar on the left.
- File: Add your own files**: Points to the 'File' menu item.
- View: Change coloring scheme**: Points to the 'View' menu item.
- Tools: Search using BLAT**: Points to the 'Tools' menu item.
- Locate where you are on the scaffold**: Points to the scaffold navigation bar at the top of the main area.
- Search for a gene or location**: Points to the search bar in the top right.
- Log in/out**: Points to the user profile dropdown in the top right.
- User-created annotations track**: Points to the 'User-created Annotations' track in the main area.
- Turn tracks on/off**: Points to the track selection sidebar.
- Find information about tracks**: Points to the track list in the sidebar.
- Zoom in/out**: Points to the zoom controls in the main area.

- 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 web application. The interface includes a header with the i5k@NAL logo and navigation links (Tools-, About Us, Contact). The main content area is divided into sections: BLAST Databases, Query Sequence, and Program. Annotations with arrows point to specific elements:

- Select organism**: Points to the **Organisms** list under **BLAST Databases**, where *Eurytemora affinis* is selected.
- Paste or upload query sequence(s)**: Points to the **Query Sequence** text area, which contains a peptide sequence:   
>FBpp0070332  
MDNCDQDASFRLSHIKEEVKPDISQLNDSNN  
SSFSPKAESPVPFMQAMSMVHVLPGSNSASS  
NNNSAGDAQMAQAPNSAG  
GSAAAQVQQYPPNHPLSGSKHLCSICGDRA  
SGKHVGVYSCGCKGFFKRTVRKDLTYACRE
- Program is automatically selected**: Points to the **Program** section, where **tblastn** is selected.
- Select organism-specific database**: Points to the **Nucleotide** section, where **Genome Assembly - Eaff\_11172013.genome\_new\_ids.fa** is selected.
- BLAST against the genome assembly to view HSPs in Jbrowse**: Points to the **Genome Assembly - Eaff\_11172013.genome\_new\_ids.fa** selection.

The **Query Sequence** section also includes a **Browse...** button and the text "No file selected." The **Program** section includes radio buttons for **tblastn**, **tblastx**, **tblast**, and **tblastx**, along with **Reset** and **Search** buttons.

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



# i5k Workspace BLAST: one way to access Apollo

BLAST Result | i5k - App

Query Coverage Graph - FBpp0070332, BLAST Hits 1-9

Subject Coverage Graph - gnl|Eurytemora\_affinis|euraff\_Scaffold427, BLAST Hits 1-9

Showing 1 to 9 of 9 entries (filtered from 55 total entries)

blastdb	qsseqid	sseqid	pid	length	mismatch	gapopen	qstart
euraff	FBpp0070332	Scaffold427	36.36	77	49	0	419
euraff	FBpp0070332	Scaffold427	26.67	165	83	4	262
euraff	Eaff_11172013.genome_new_ids.fa		59.21	76	31	0	103
euraff	FBpp0070332	Scaffold4229	56.52	92	37	1	98
euraff	FBpp0070332	Scaffold200	57.14	91	36	1	99
euraff	FBpp0070332	Scaffold12	58.57	87	39	2	104
euraff	FBpp0070332	Scaffold12	58.57	87	39	2	104
euraff	FBpp0070332	Scaffold13	85.71	35	5	0	91
euraff	FBpp0070332	Scaffold200	58.62	81	38	1	101

BLAST Report

Score = 86.3 bits (212), Expect = 2e-16, Method: Compositional matrix adjust.

Identities = 45/76 (59%), Positives = 61/76 (80%), Gaps = 0/76 (0%)

Frame = -3

Query 103 LCSCGDRASGKHVGVSCGCKGFKRTVRKDLTYACRENRCIIIDKgrnrcqcyry 162

Subject 255352 ICVCGDKSSGKHVGVQVT

Query 163 kclTCGKREAVQEER KC GH+EAVO R KCFKTGHRKEAVQPR

Subject 255172

Score = 62.4 bits (150), Expect = 3e-12, Method: Compositional matrix adjust.

Identities = 32/62 (51%), Positives = 41/62 (66%), Gaps = 0/62 (0%)

Frame = -3

Query 414 ILYNPDIRGKSKRAEIM IL++ D G+ S IE

Subject 251038 ILFS-DAIGLTPPPVIEQ

Query 474 DHLFLFRITSDRPLELF + LF R+ P+E L

Subject 251038

Available Tracks

- 0. Reference Assembly
- 1. Gene Sets
- 2. Evidence
- 3. Mapped Proteins
- Other
- Protein2genome

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

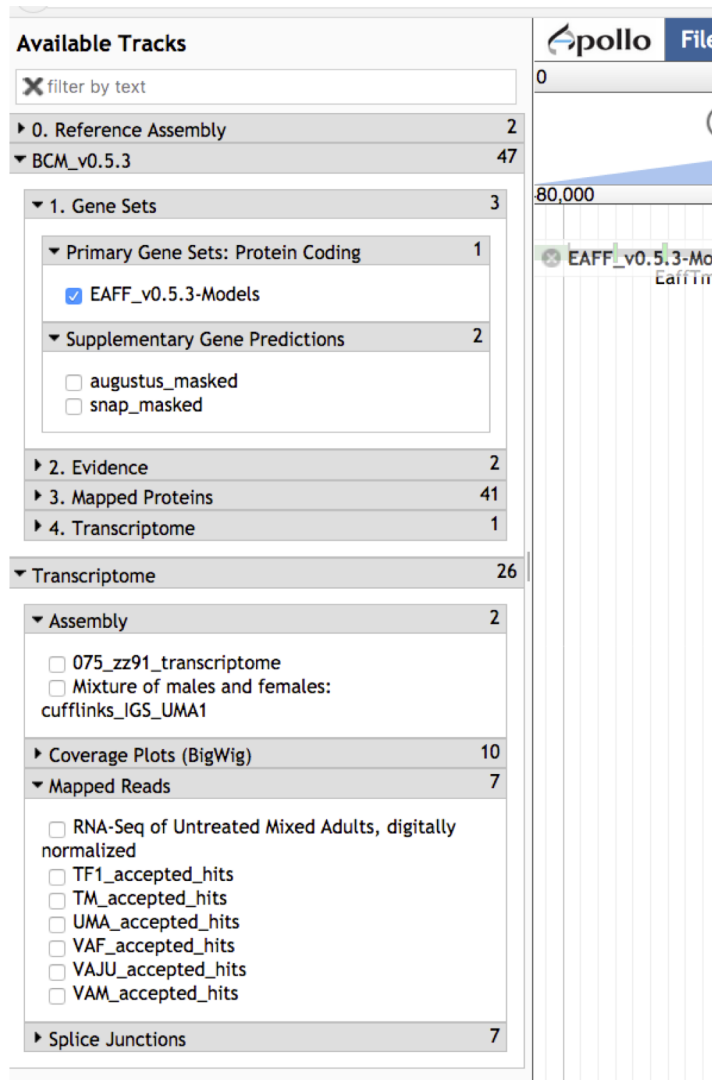
# Annotation Example

- 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](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 categories include Reference Assembly (2), BCM\_v0.5.3 (47), Gene Sets (3), Evidence (2), Mapped Proteins (41), Transcriptome (26), and Splice Junctions (7). The "Gene Sets" category is expanded, showing "Primary Gene Sets: Protein Coding" (1) and "Supplementary Gene Predictions" (2). Under "Primary Gene Sets: Protein Coding", the track "EAFF\_v0.5.3-Models" is selected with a blue checkmark. Under "Supplementary Gene Predictions", the tracks "augustus\_masked" and "snap\_masked" are listed but not selected. The "Transcriptome" category is also expanded, showing "Assembly" (2), "Coverage Plots (BigWig)" (10), and "Mapped Reads" (7). Under "Assembly", the tracks "075\_zz91\_transcriptome", "Mixture of males and females: cufflinks\_IGS\_UMA1", and "cufflinks\_IGS\_UMA1" are listed. Under "Coverage Plots (BigWig)", the tracks "RNA-Seq of Untreated Mixed Adults, digitally normalized", "TF1\_accepted\_hits", "TM\_accepted\_hits", "UMA\_accepted\_hits", "VAF\_accepted\_hits", "VAJU\_accepted\_hits", and "VAM\_accepted\_hits" are listed. Under "Mapped Reads", the tracks "RNA-Seq of Untreated Mixed Adults, digitally normalized", "TF1\_accepted\_hits", "TM\_accepted\_hits", "UMA\_accepted\_hits", "VAF\_accepted\_hits", "VAJU\_accepted\_hits", and "VAM\_accepted\_hits" are listed. On the right, the main panel shows a genomic track view with a blue bar representing the "EAFF\_v0.5.3-Models" track. The track is labeled "EAFF\_v0.5.3-Models" and "EaffTm". The track is positioned at a scale of 80,000. The Apollo logo and a "File" button are visible at the top of the main panel.

Category	Count
0. Reference Assembly	2
BCM_v0.5.3	47
1. Gene Sets	3
Primary Gene Sets: Protein Coding	1
Supplementary Gene Predictions	2
2. Evidence	2
3. Mapped Proteins	41
4. Transcriptome	1
Transcriptome	26
Assembly	2
Coverage Plots (BigWig)	10
Mapped Reads	7
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 level<sup>i</sup>

Annotation score is a heuristic for annotation quality

Organism-specific databases

FlyBase<sup>i</sup> FBgn0003067. Pepck.

Subcellular location<sup>i</sup>

Flybase is another great resource

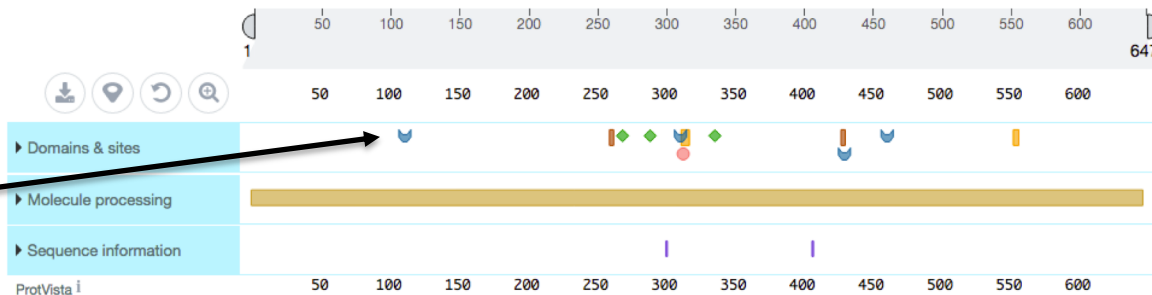
UniProtKB - P20007 (PCKG\_DROME)

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,

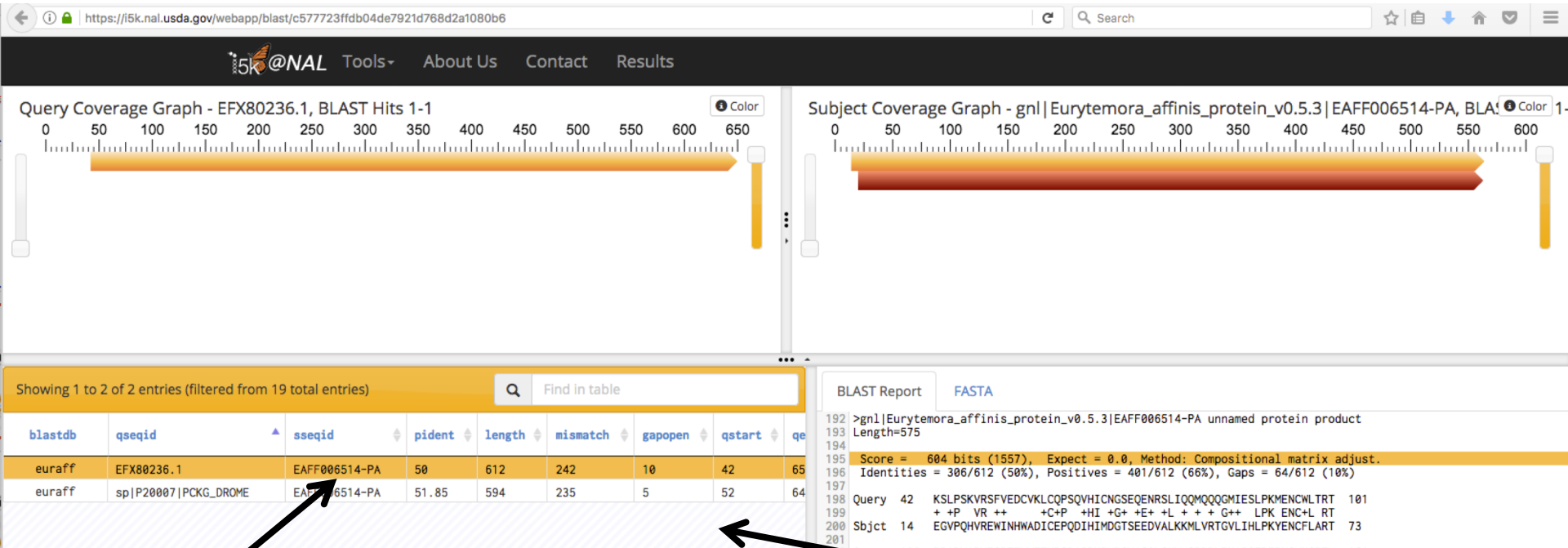
(daphnia Phosp  
carboxykinase)

(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'  
EAFF006514 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/68b677fb267d4cfe93b0570dd87449f7>



# 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 tooltip is visible over the 'blastdb' column for the first row, showing 'Eaff\_11172013.genome\_new\_ids.fa' and a link to 'Click to view in genome browser'. A black arrow points from the text 'Click on blue blastdb button next to your favorite HSP to view it in JBrowse' to the 'blastdb' column header. Below the table, there are filters and a 'Download' button. On the right side, there is a 'BLAST Report' section with 'FASTA' format, showing sequence alignments and scores.

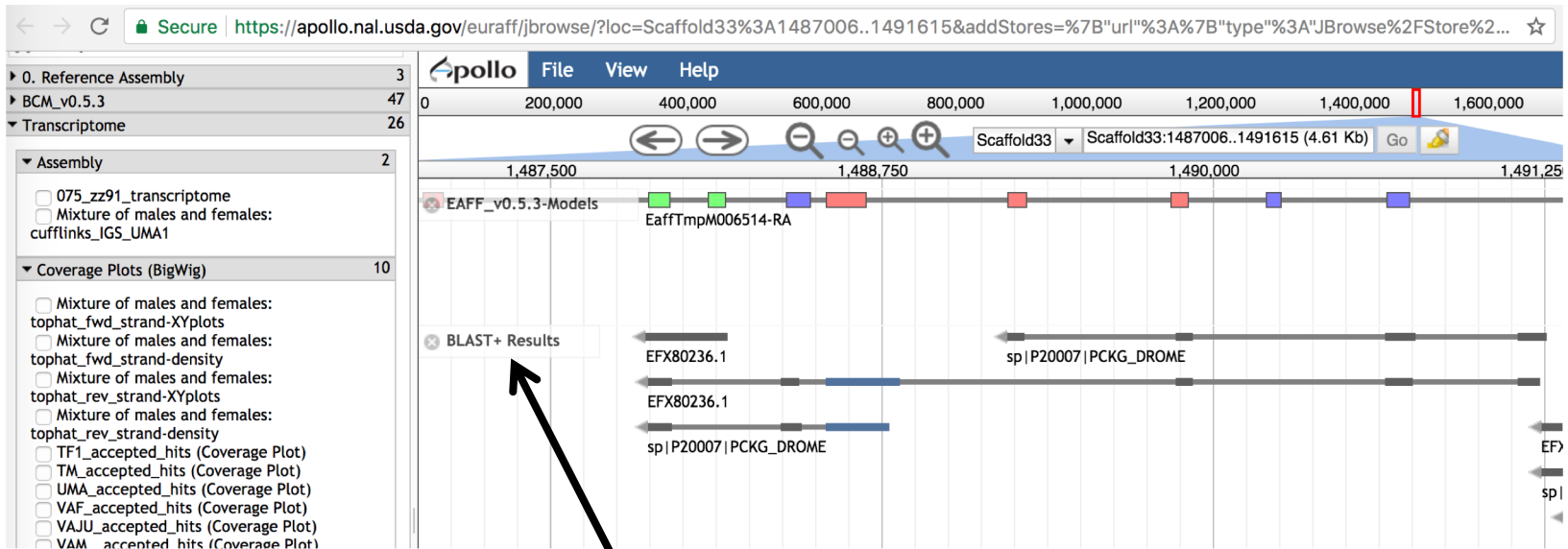
blastdb	qseqid	sseqid	pident	length	mismatch	gapope
euraff	Eaff_11172013.genome_new_ids.fa	Scaffold133	56.41	39	17	0
euraff	sp P20007 PCKG_DROME	Scaffold133	62.5	40	15	0
euraff	EFX80236.1	Scaffold133	80	30	6	0
euraff	sp P20007 PCKG_DROME	Scaffold133	78.12	32	7	0
euraff	EFX80236.1	Scaffold133	44.59	74	24	2
euraff	sp P20007 PCKG_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

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/1dd580d46260410da7473f974da76a54>



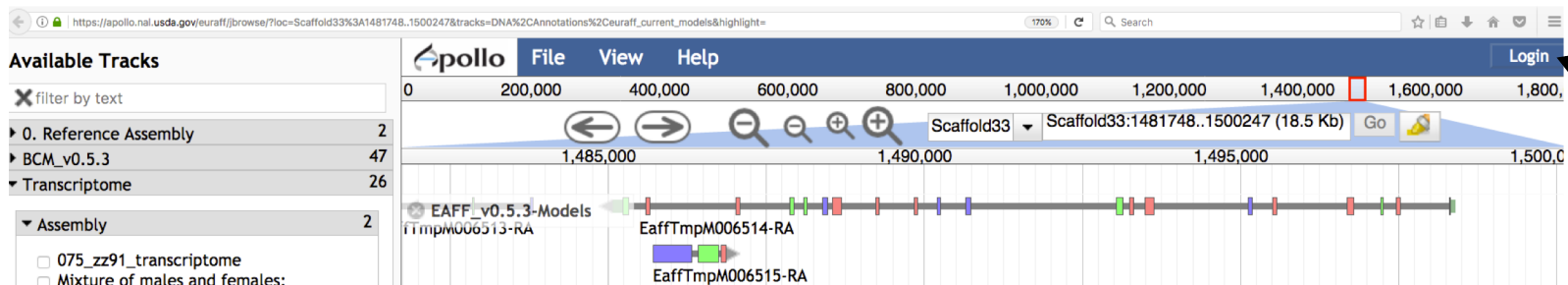
# Another approach: BLAST against the genome



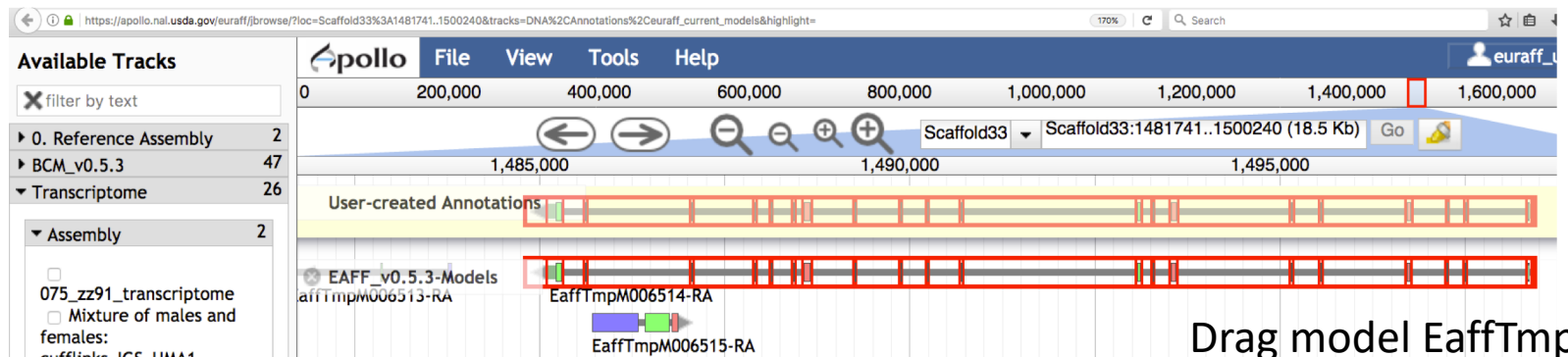
BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

Apollo result URL: <http://tiny.cc/lwuzsy>

# Create annotation in user-created annotations track



Log in with  
your  
Apollo  
credentials



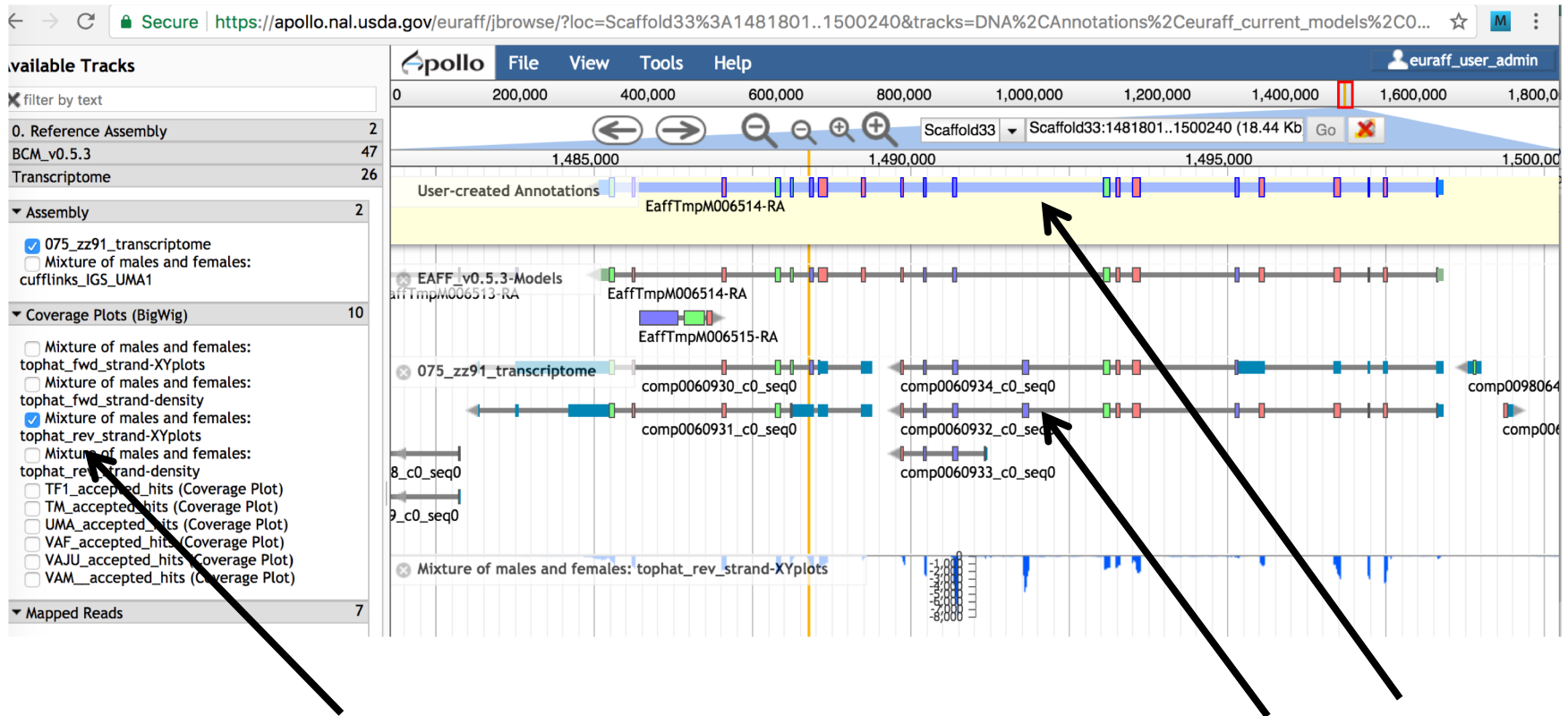
Drag model EaffTmpM006514-  
RA to User-created Annotations  
track

# 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?



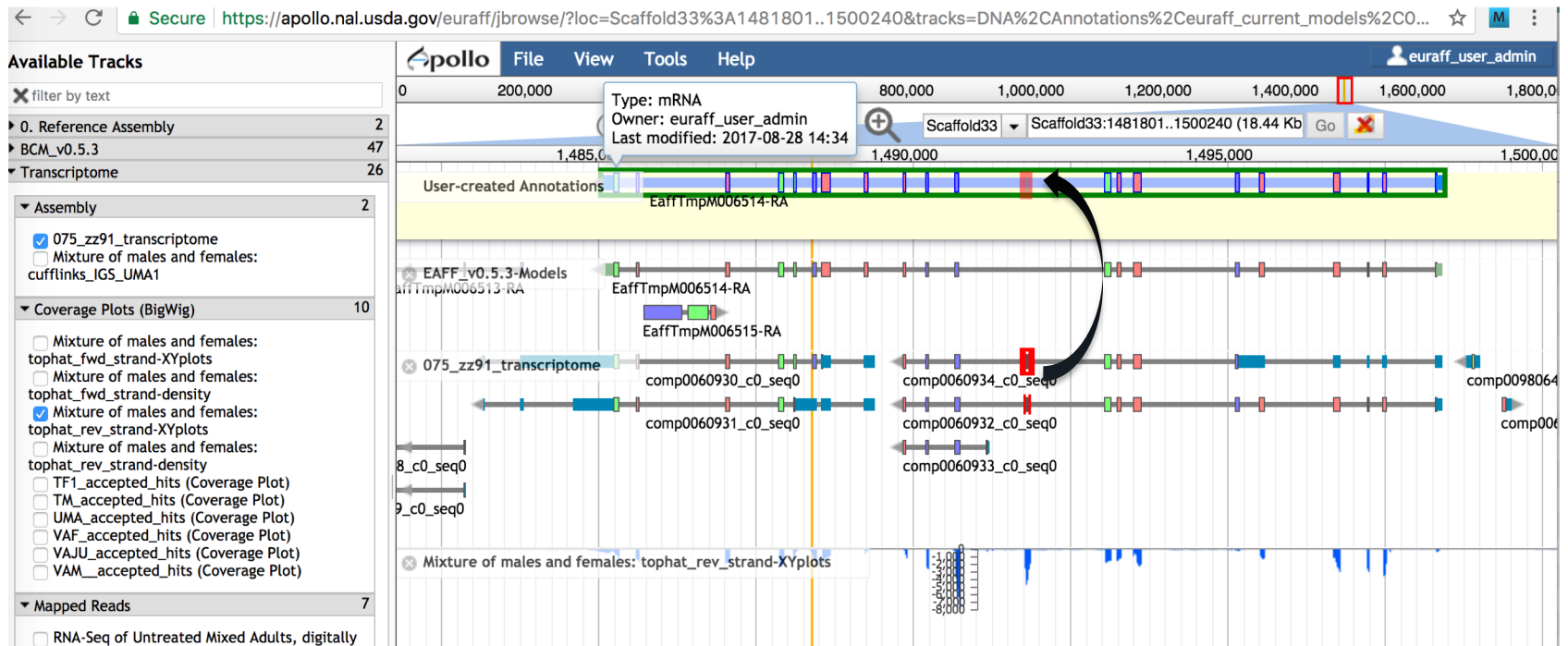
# View available evidence



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

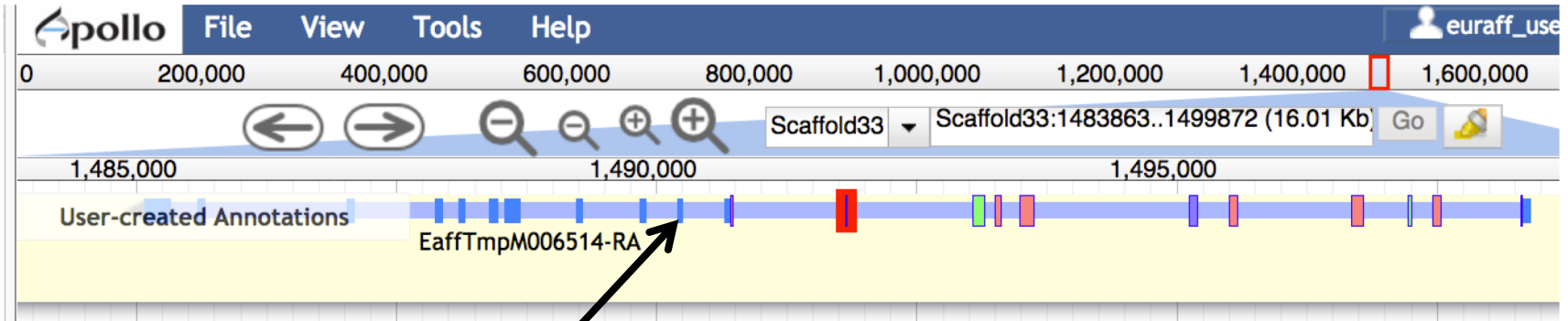
RNA-Seq and transcriptome tracks suggest that one exon is missing

# Add an exon to the model

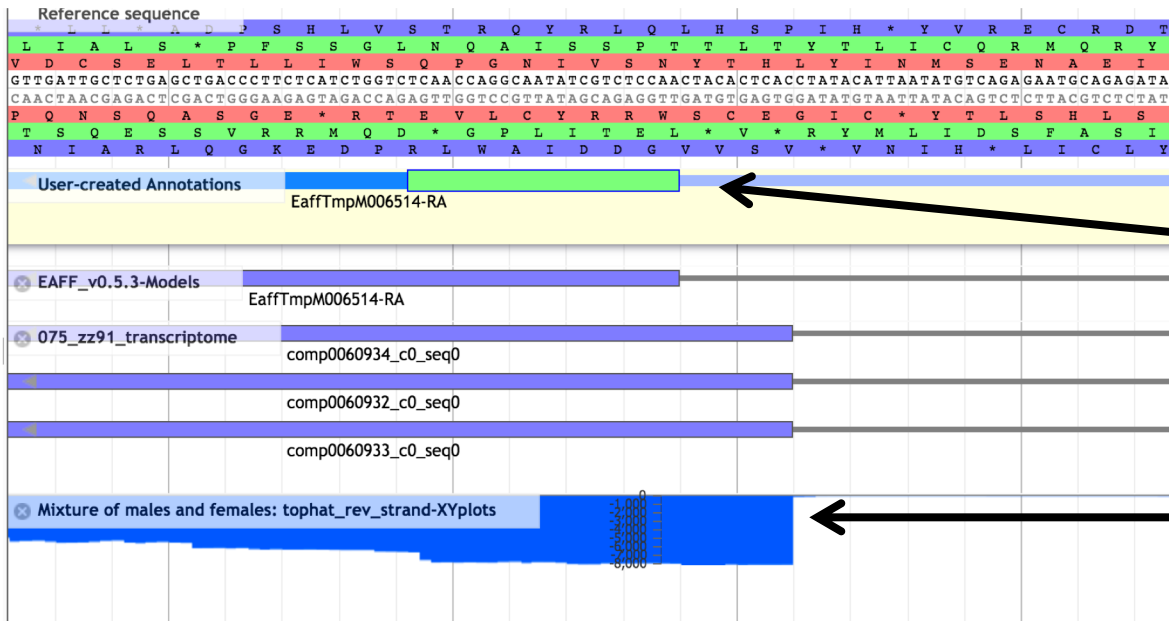


Drag exon from  
transcriptome track  
into new gene model

# Adjust exon boundary



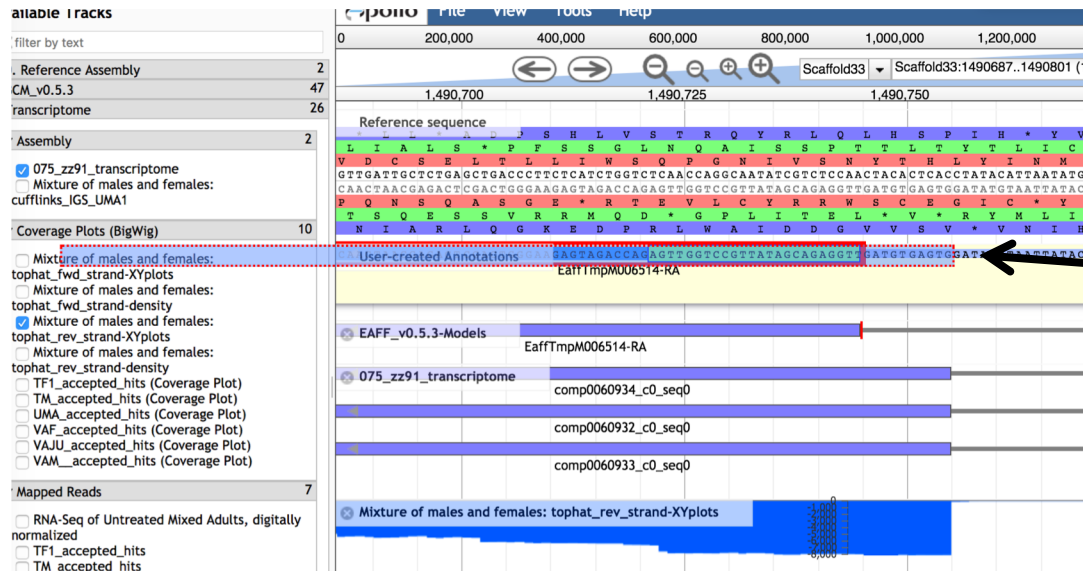
CDS sequence is now UTR –zoom in to investigate



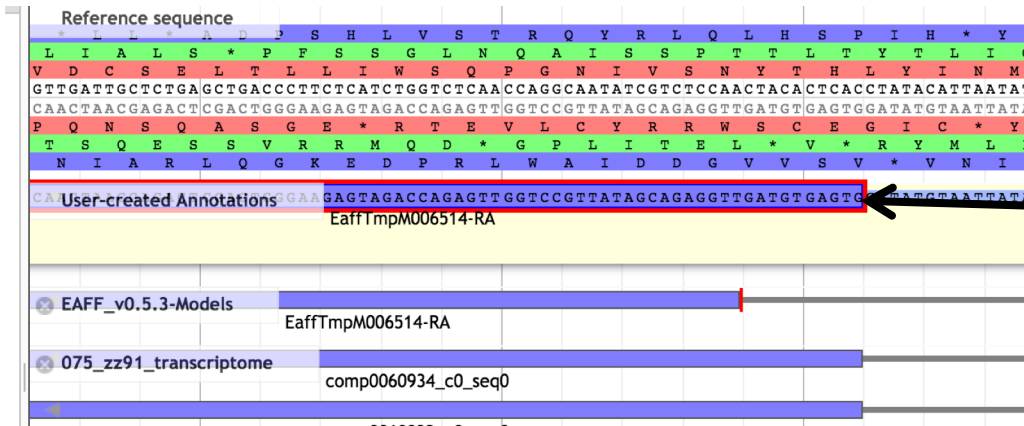
CDS frame has changed from purple to green—we need to fix this

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

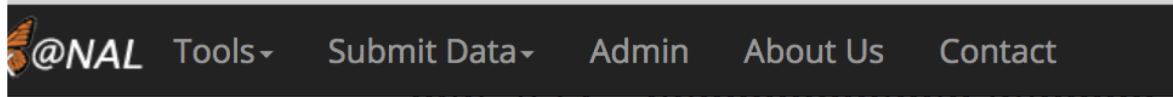
- Blast modified EAFF006514-PA sequence to NCBI's nr database
  - Make sure it doesn't match a potential contaminant
  - Get an idea whether you have the right sequence
  - Blastp home:
    - [https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE\\_TYPE=BlastSearch&LINK\\_LOC=blasthome](https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome)
  - Result URL:
    - <https://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Get&RID=DY9HRCRA015> (expires end of day 4/25)
- 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-----VGDDIAWLRPDEKGQLRAI
ISGITNSQGEKKYIVAAFPSCGKTNLAMMQPRLP-----VGDDIAWLRPDEKGQLRAI
ILGITDPKGEKKYITAAFPSCGKTNLAMNPSLANYKVECVGDDIAWMKFD SQVLRAI
ILGITNPQGQKKYIAAAPPSCGKTNLAMLTPTLPGYKVECVGDDIAWMHFDKEGRLRAI
*                               *****: :*.:* ****
```

New exon added

euraff\_old\_pepck  
euraff\_new\_pepck  
sp|P20007|PCKG\_DROME  
EFX80236.1

```
NPENGFFGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMGDKPK
NPENGFFGVAPGTSYTSNPVA-----MQSIFKDTIFSNVAMTDDGGVWVEGMGDKPK
NPENGFFGVAPGTSMETNPPIA-----MNTVFKNTIFTNVASTSDGGVFWEGMESSLA
NPENGFFGVAPGTNYATPNACYNFFLYAMLTIQKNTIFTNVAKTSDDGGVFWEGLEKEV-
*****: :* * * : : * : : : : * : : : : : : : : : : . .
```

euraff\_old\_pepck  
euraff\_new\_pepck  
sp|P20007|PCKG\_DROME  
EFX80236.1

```
ERSSCIDWK GK-PWRPTSSNPAHPNSRFCTPLLNC PVLDESAEDPAGVPIAAILFGGRR
ERSSCIDWK GK-PWRPTSSNPAHPNSRFCTPLLNC PVLDESAEDPAGVPIAAILFGGRR
PNVQITDWLGK-PWTKDSGKPAHPNSRFCTPAAQCPIIDEAWEDPAGVPIAAILFGGRR
TGVDITSWLGDANWTKSSGKPAHPNSRFCTPAAQCPIIDEAWEDPAGVPIAAILFGGRR
. . * * * * : : : : : : : : : : : : : : : : * * * * *
```

euraff\_old\_pepck  
euraff\_new\_pepck  
sp|P20007|PCKG\_DROME  
EFX80236.1

```
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFK GKQIMHDPF SMRPFFG-----HW
PSGVPLVYQAISWEHGVFMGACVKSEATAAAEFK GKQIMHDPF SMRPFFG-----HW
PAGVPLIYEARDWTHGVFI GAAMRSEATAAAEHK GKVIMHDPFAMRPFFGYNFGDYVAHW
PRGVPLVYEALNWKHG VFVGASVSEATAAAEHK GRSIMHDPFAMRPFFGYNAGNYLGHW
* ****: :* . * ****: :* : : ****: :* : : ****: :* : : ****
```

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

- Clustal result URL:  
<https://i5k.nal.usda.gov/webapp/clustal/49a4d63c24fd4ed3b3a67cf71a0369df>
- 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
  - **Please review our naming guidelines:**
    - <https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines>
    - 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



# Using the Information Editor

eucaff/browse/?loc=Scaffold33%3A1482161..1498680&tracks=DNA%2CAnnotations%2CEuraff\_current\_models%2Ctphat\_rev\_strand-XYplots&highlight=

170% Search

File View Tools Help

Information Editor (alt-click)

Select mRNA Phosphoenolpyruvate carboxykinase

gene	
Name	
Symbol	
Description	
Created	2017-08-28
Last modified	2017-08-28
Status	
<input type="radio"/> Approved <input type="radio"/> Delete	
DBXRefs	
DB	Accession
<input type="button" value="Add"/> <input type="button" value="Delete"/>	

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

# 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
    - 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.
  - You can also try out the process yourself: <https://github.com/NAL-i5K/GFF3toolkit/>
  - 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>

# 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!

