Using Apollo at the i5k Workspace@NAL

NAL USDA-ARS

https://i5k.nal.usda.gov

December 19th, 2017

USDA

Agenda

- Manual annotation general overview
- I5k 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

USDA

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

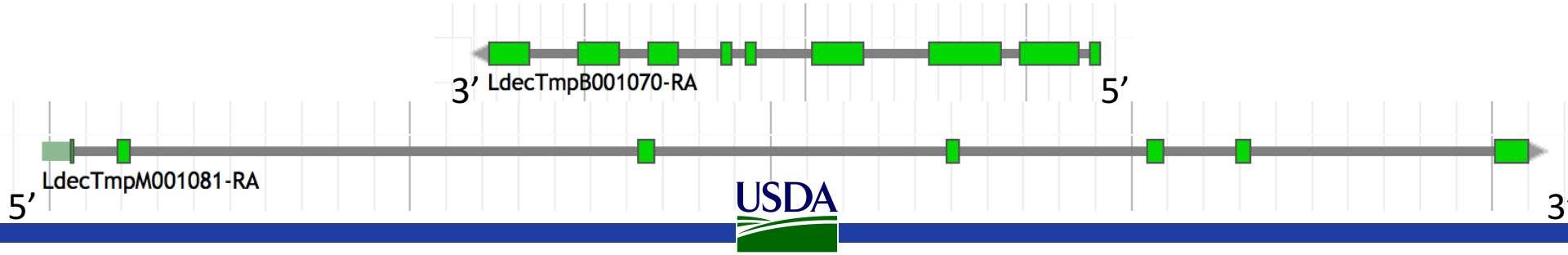
USDA

Manual annotation: i5k Workspace tools

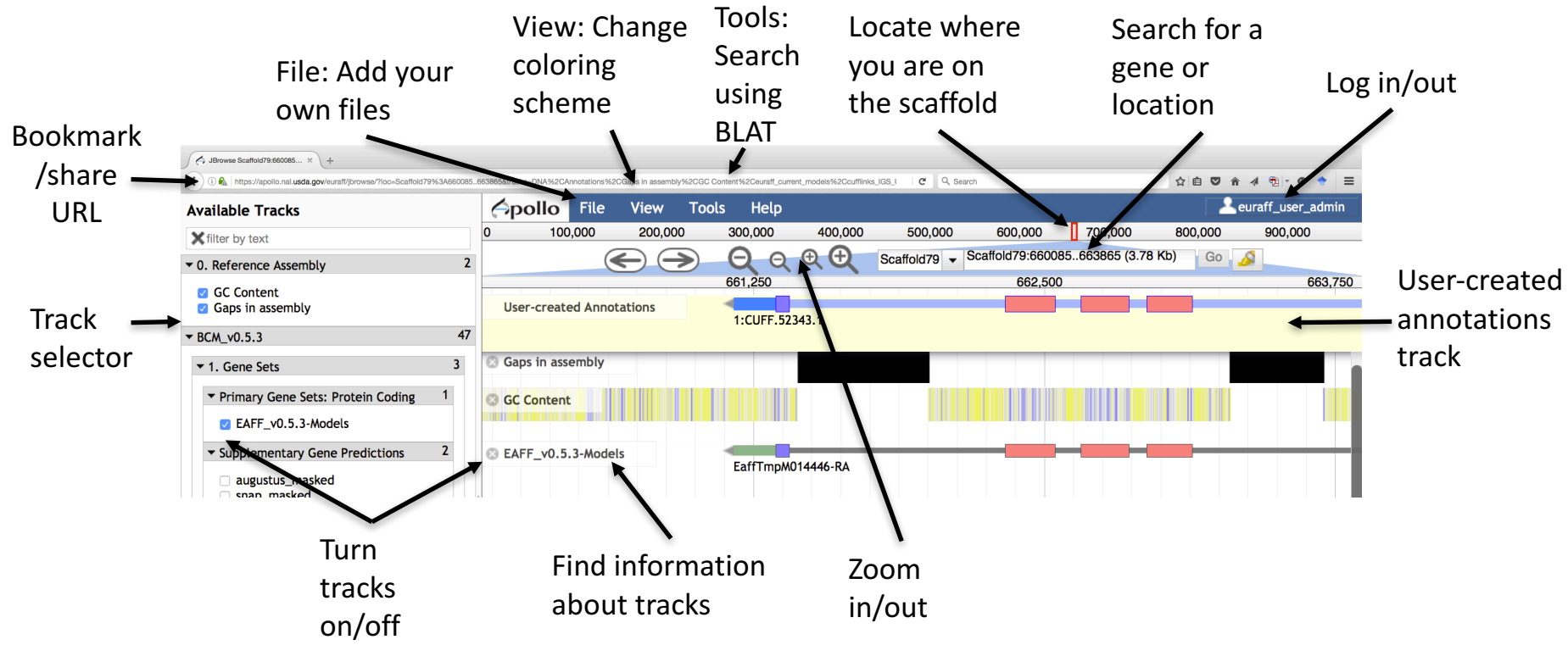
USDA

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



JBrowse is a web- based genome browser

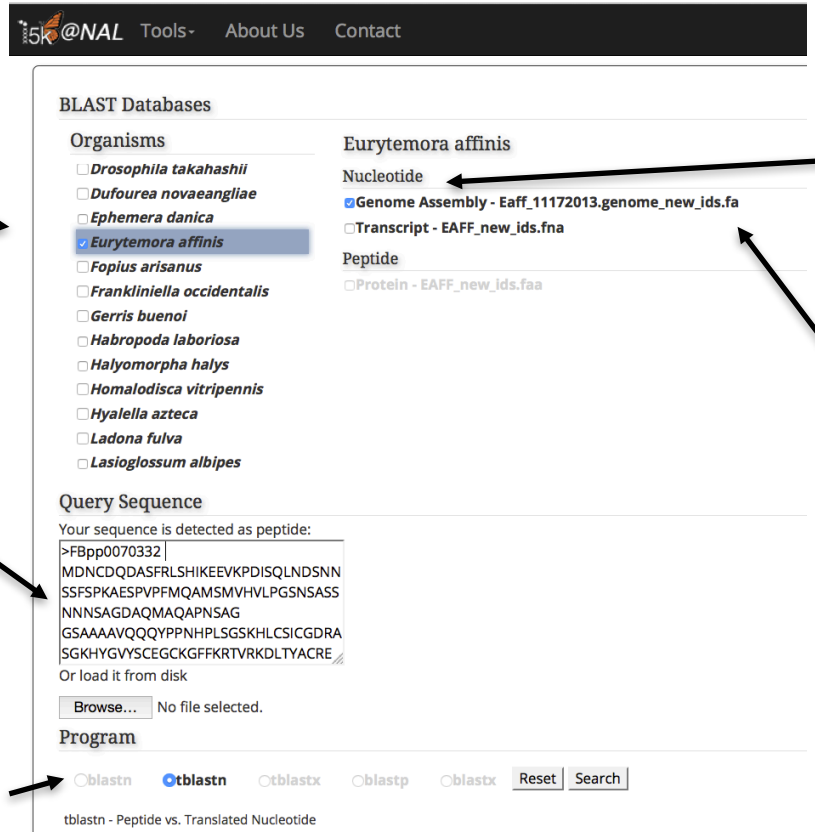
- Visualize features that are mapped to a genome
- These features are displayed as tracks
- Many different types of data may be displayed

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

USDA

i5k Workspace BLAST: one way to access Apollo



Select organism

Select organism-specific database

BLAST against the genome assembly to view HSPs in Jbrowse

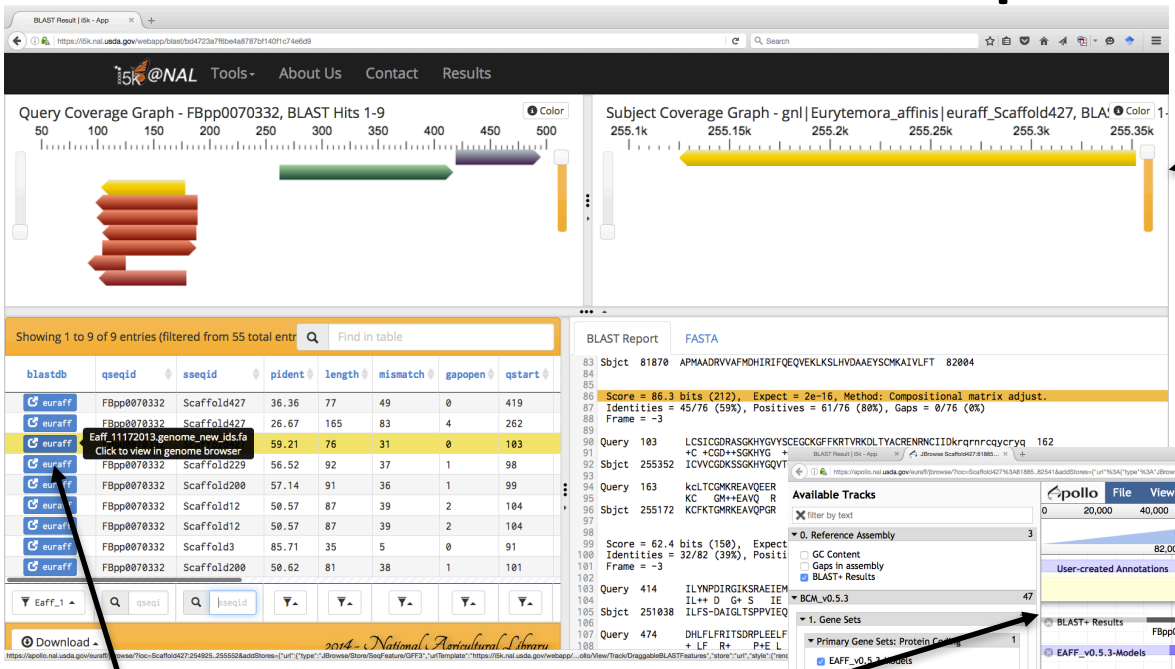
Paste or upload query sequence(s)

Program is automatically selected

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

USDA

i5k Workspace BLAST: one way to access Apollo



BLAST result page with 4 panels

Click on blue blastdb icon next to your favorite HSP

Blast results are displayed in Apollo

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

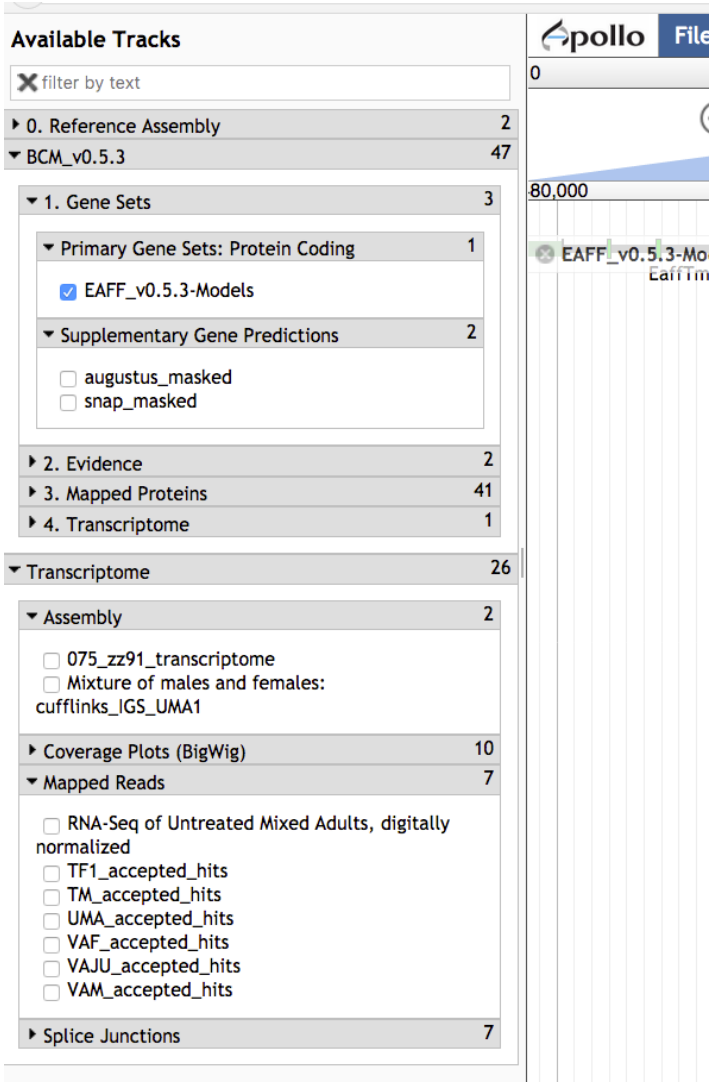
Annotation Example

- 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.

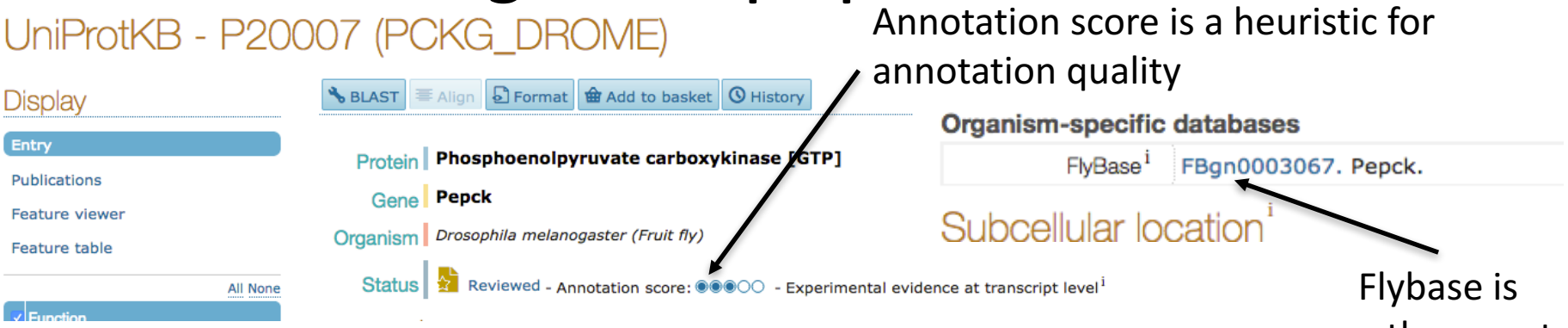
Available tracks for *E. affinis*



- 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

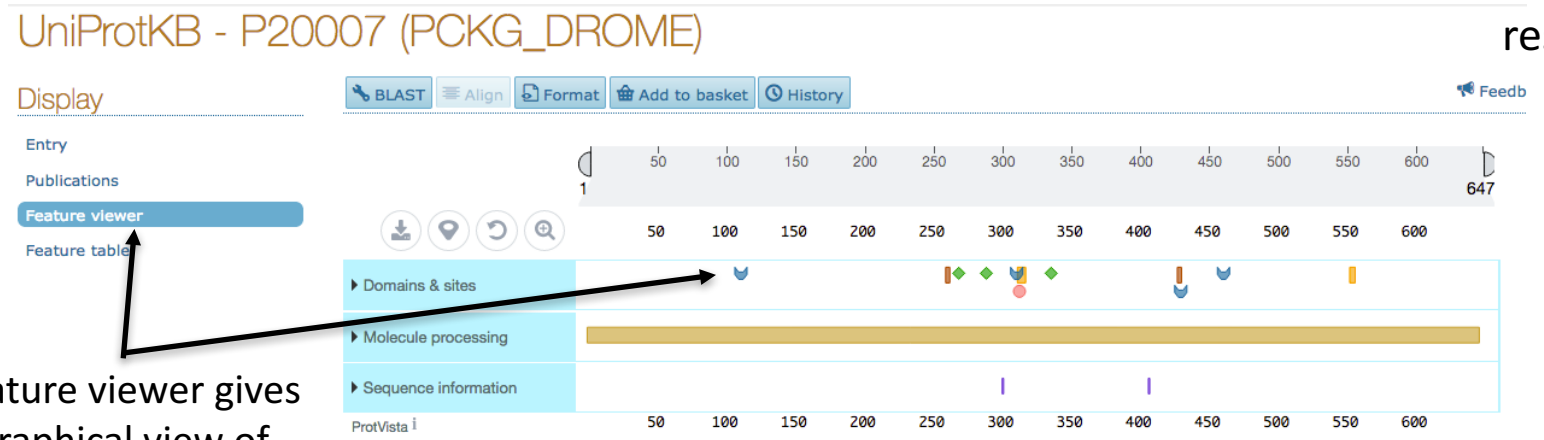
USDA

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



Annotation score is a heuristic for annotation quality

Flybase is another great resource



Feature viewer gives graphical view of domains and sites

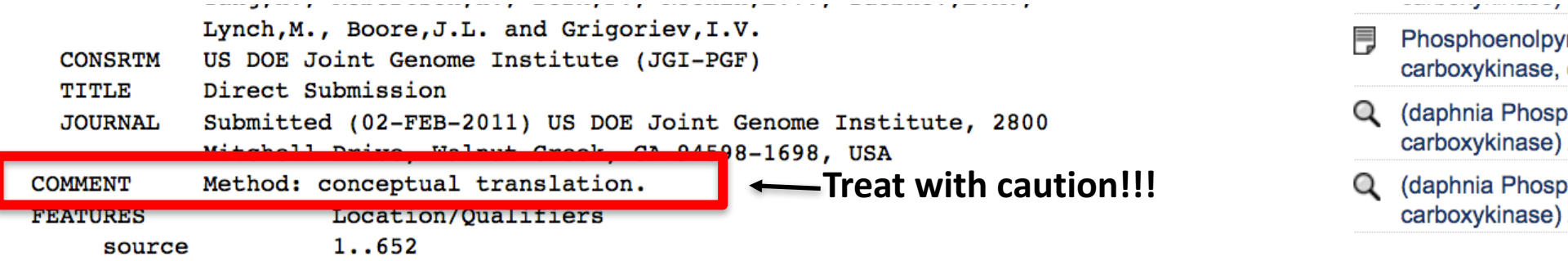
Catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).

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

USDA

Choosing reference proteins: *Daphnia pulex* Pepck

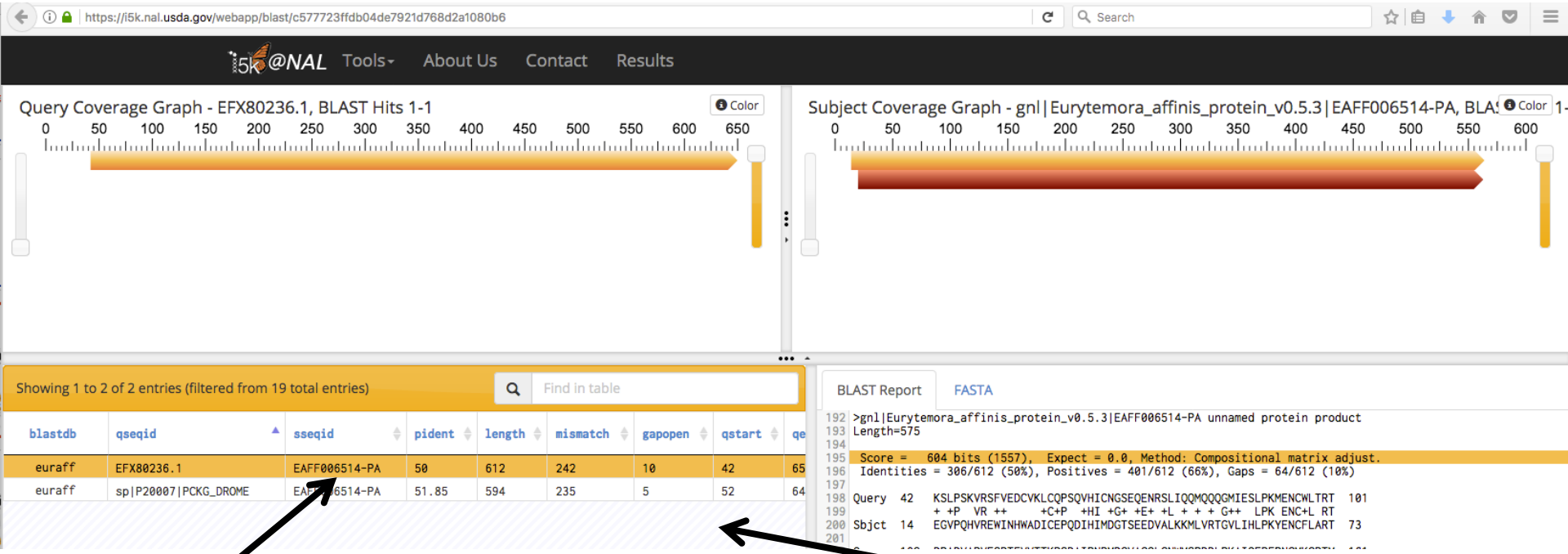
- GenBank record: https://www.ncbi.nlm.nih.gov/protein/EFX80236.1



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/90c0f04749af4d028065c0739068ae10

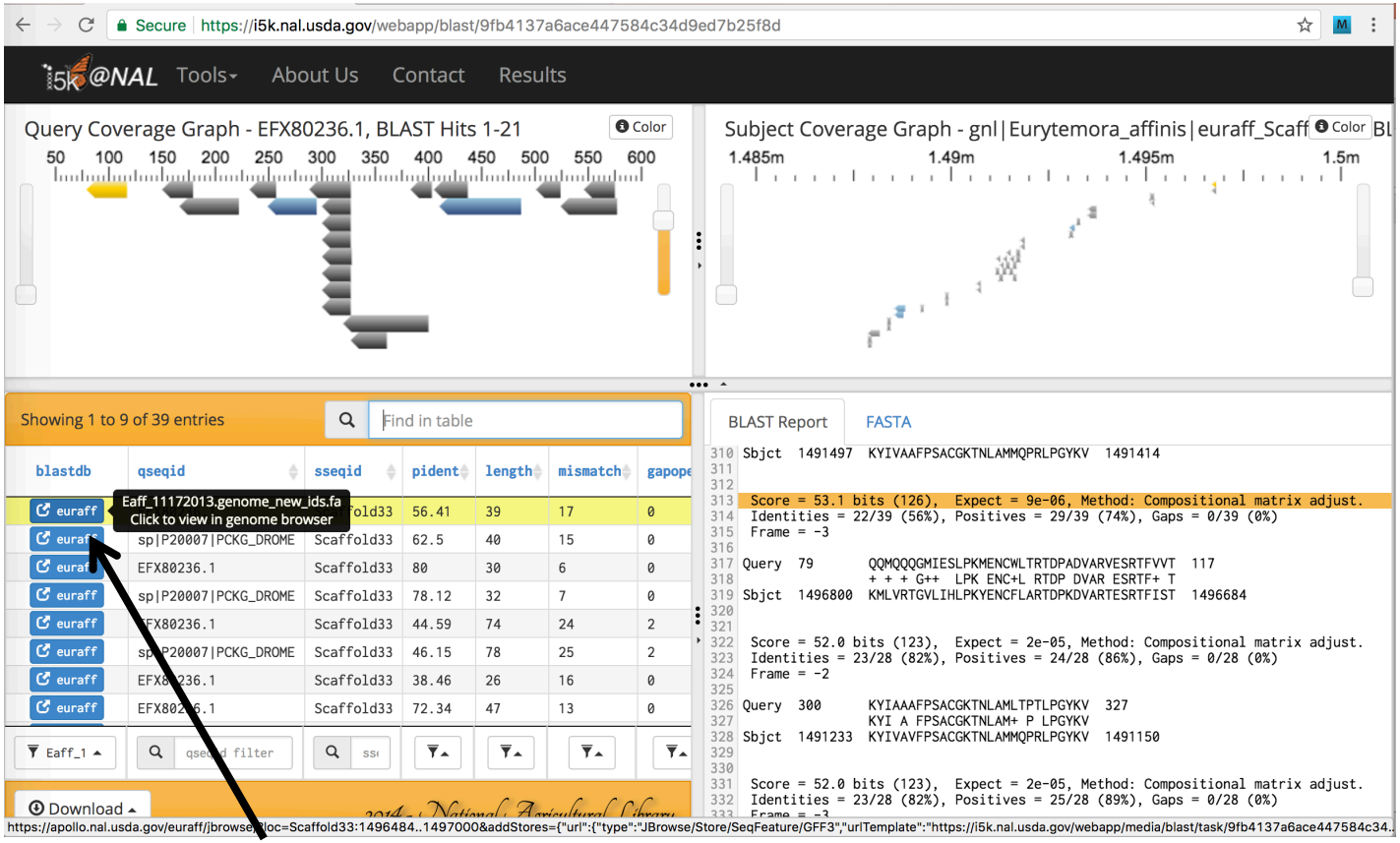
USDA

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

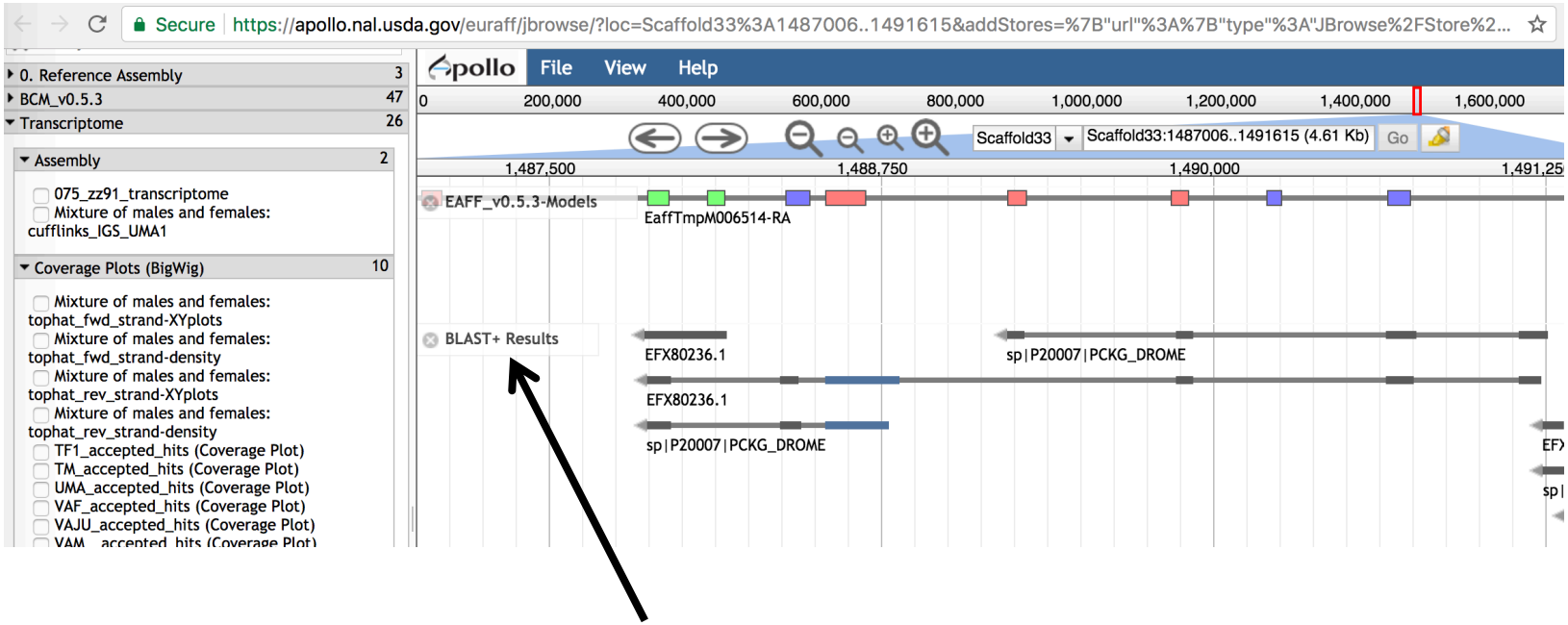
https://i5k.nal.usda.gov/webapp/blast/



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

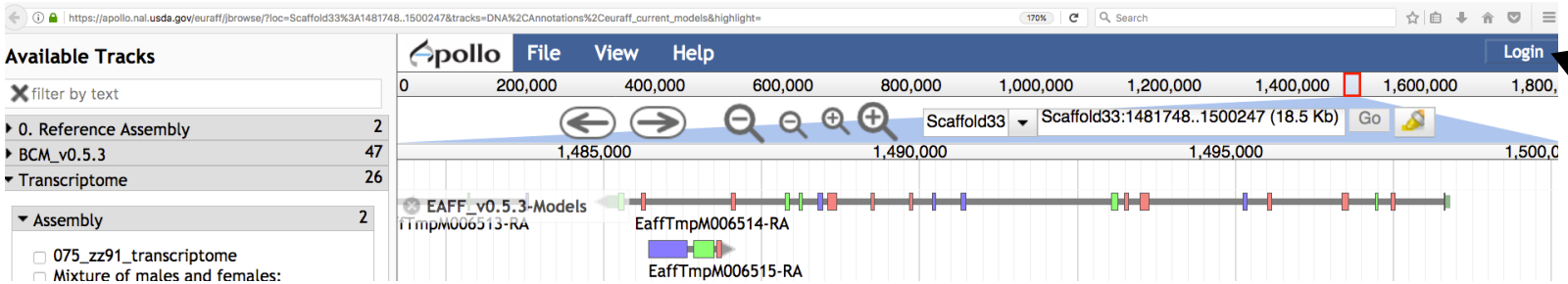


BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

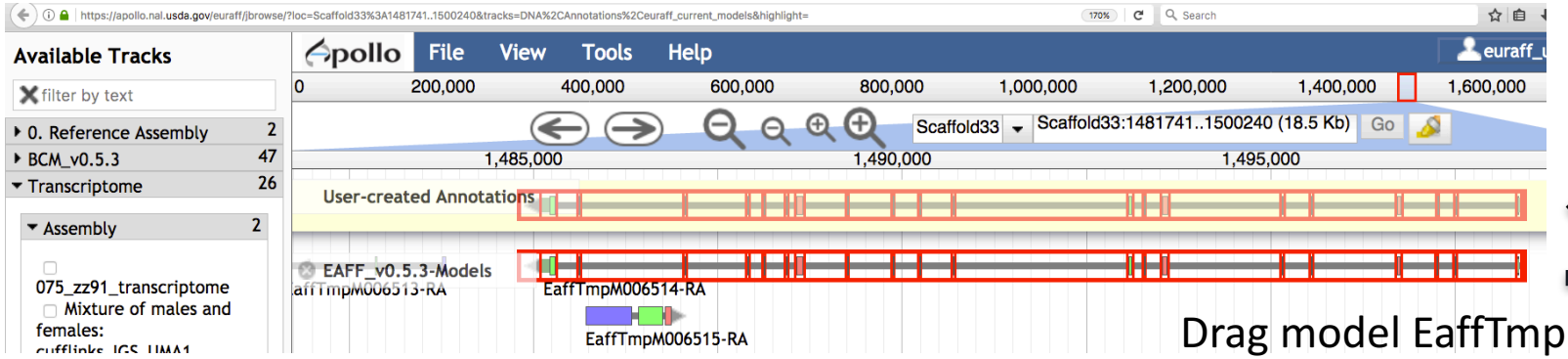
Apollo result URL: https://tinyurl.com/y8fdpzyn

USDA

Create annotation in user-created annotations track



Log in with your Apollo credentials

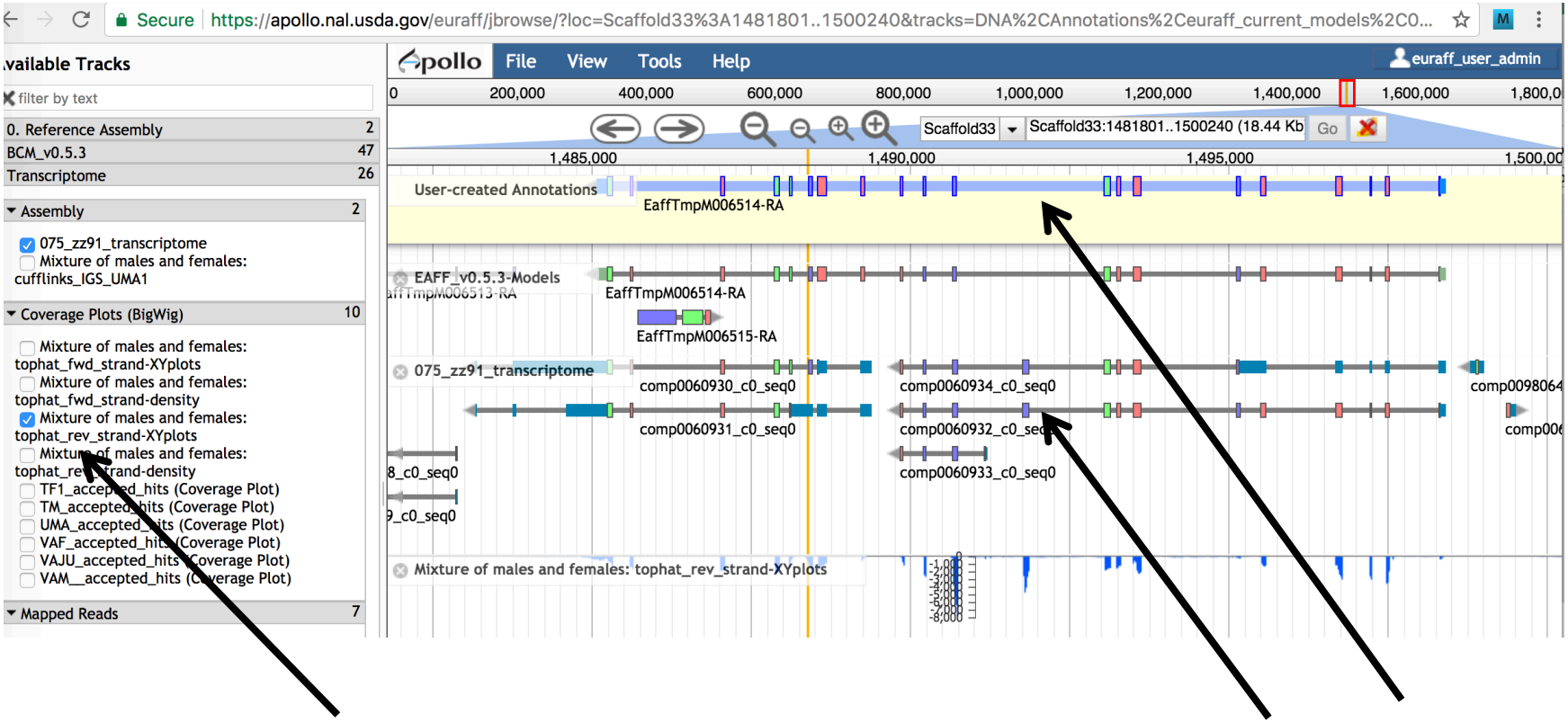


Drag model EaffTmpM006514-RA to User-created Annotations track

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

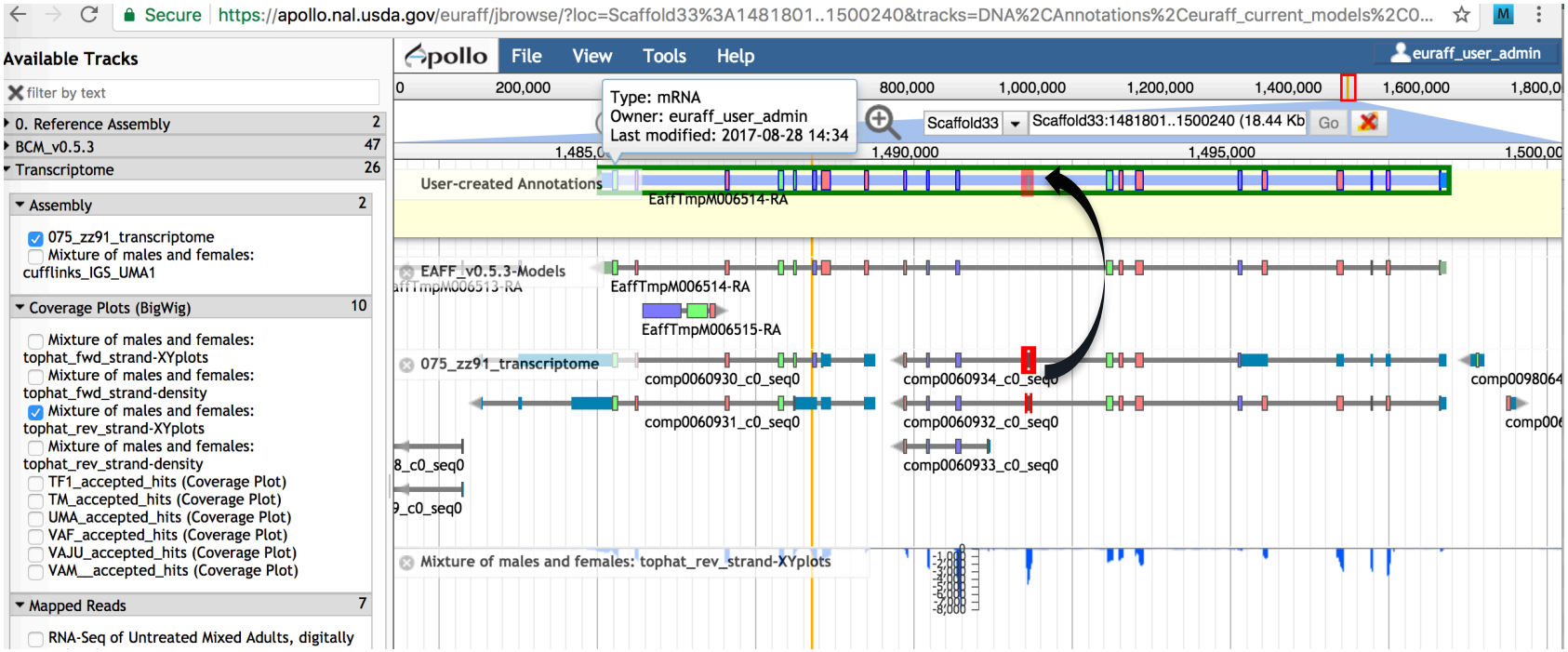


Model is on the reverse strand, so we can take advantage of the stranded RNA-Seq available for this species

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

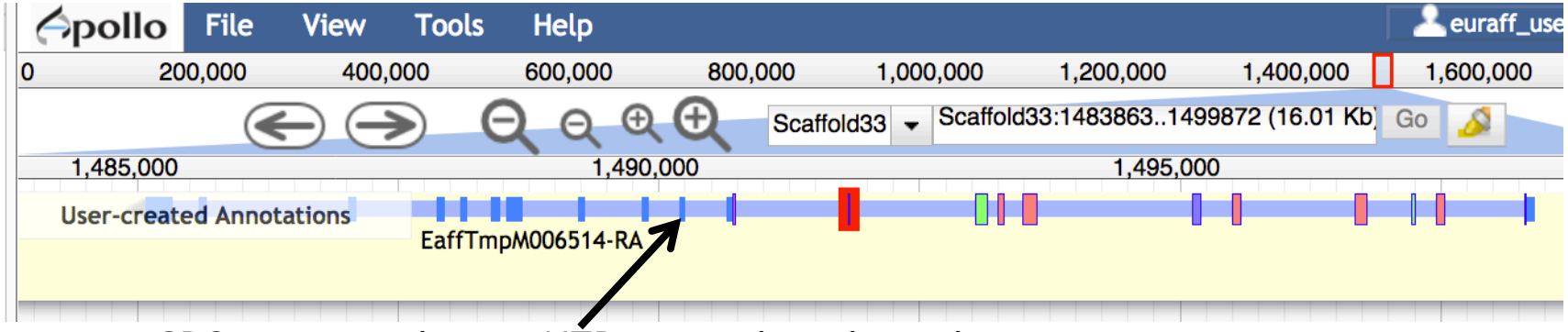
USDA

Add an exon to the model

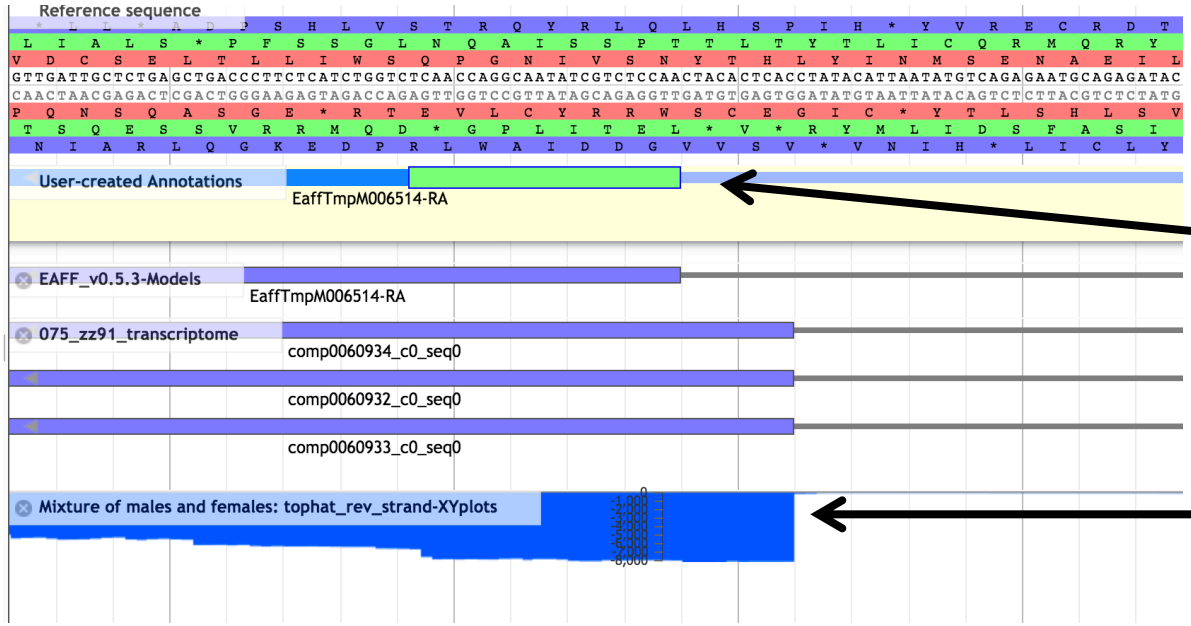


Drag exon from
transcriptome track
into new gene model

Adjust exon boundary



CDS sequence is now UTR –zoom in to investigate

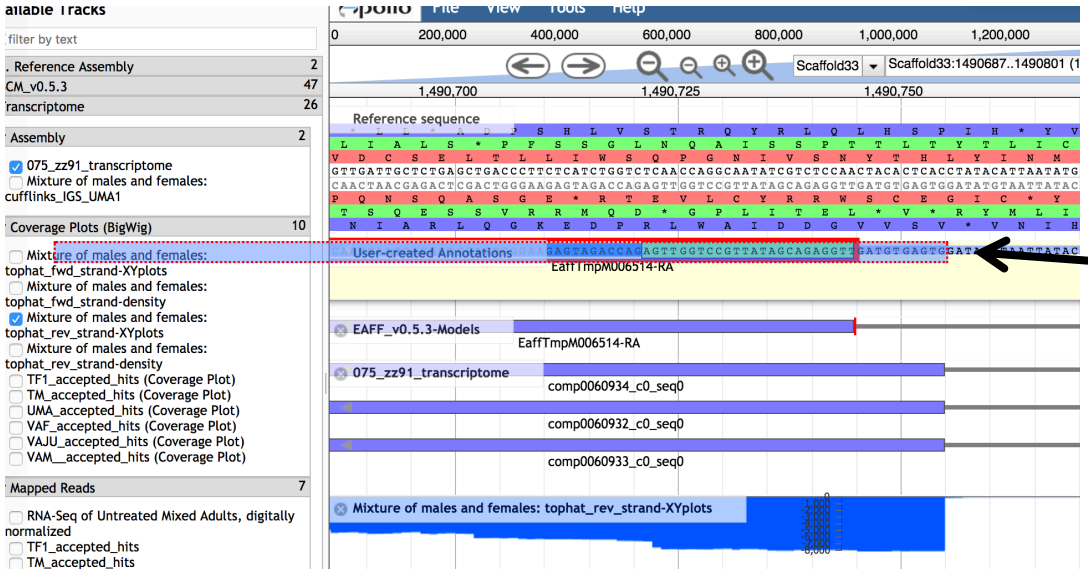


CDS frame has changed from purple to green– we need to fix this

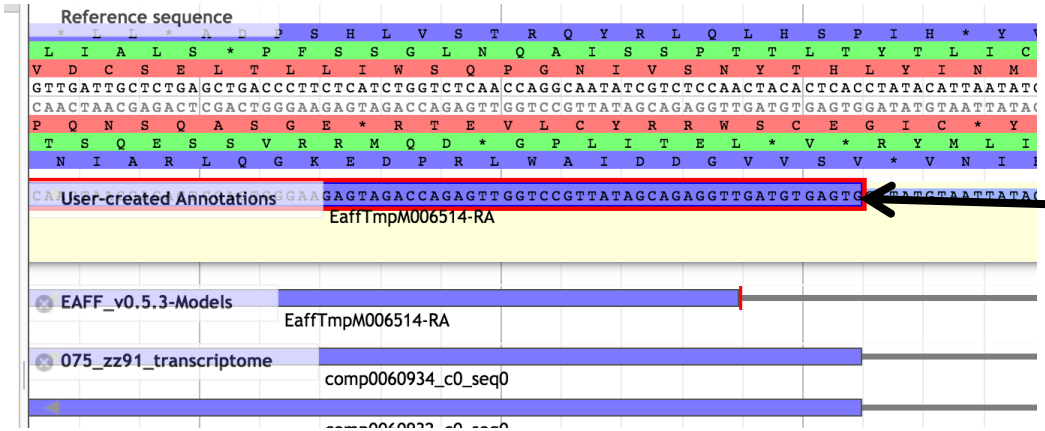
RNA-Seq suggests we need to adjust exon boundary

USDA

Adjust exon boundary



Drag exon boundary to match RNA-Seq and transcriptome tracks



Fixed both reading frame and exon boundary

USDA

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=3FWVAS7301 5 (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

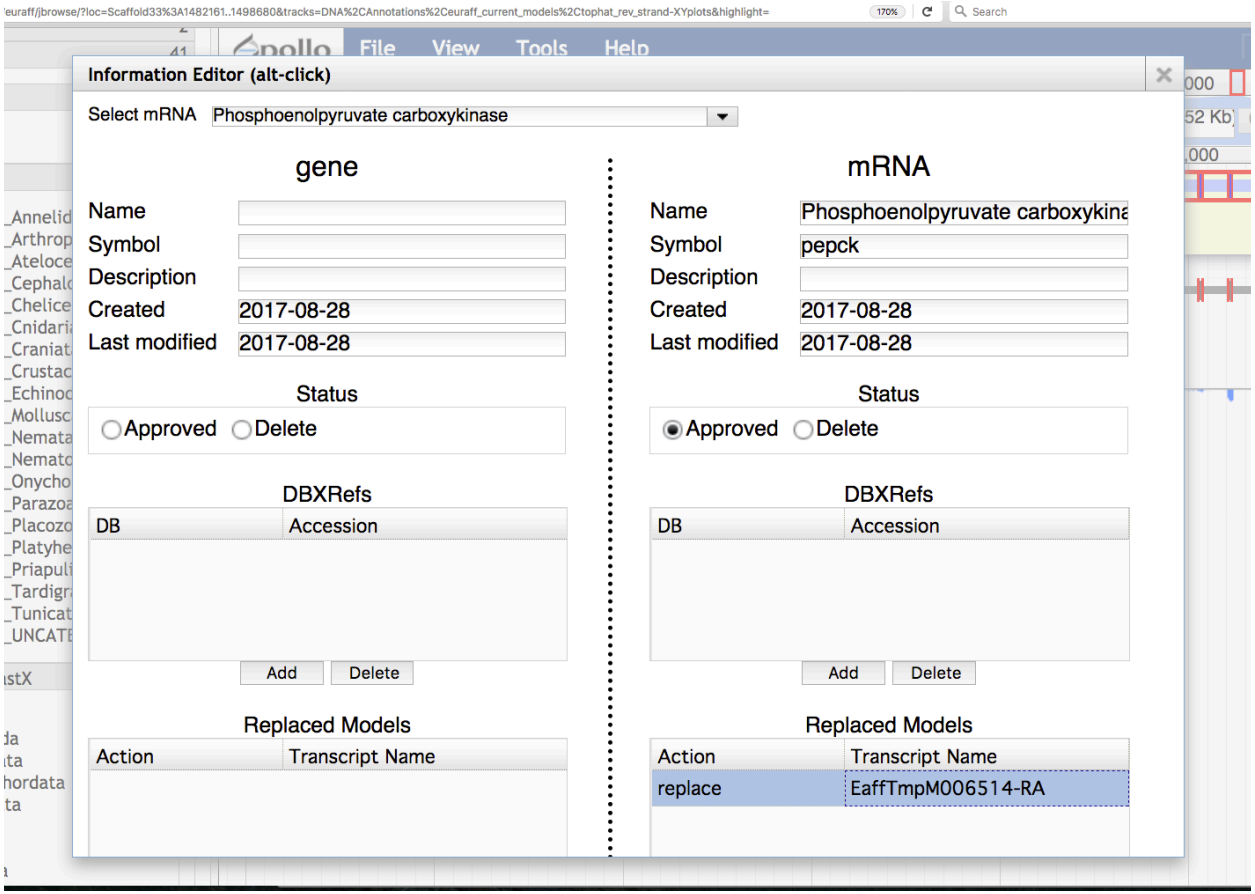
- Clustal result URL: https://i5k.nal.usda.gov/webapp/clustal/03f944fd02f245a8ad12d68cb9681776
- Scroll to bottom of page and click 'colorful' to see color-coded alignment

USDA

Using the Information Editor

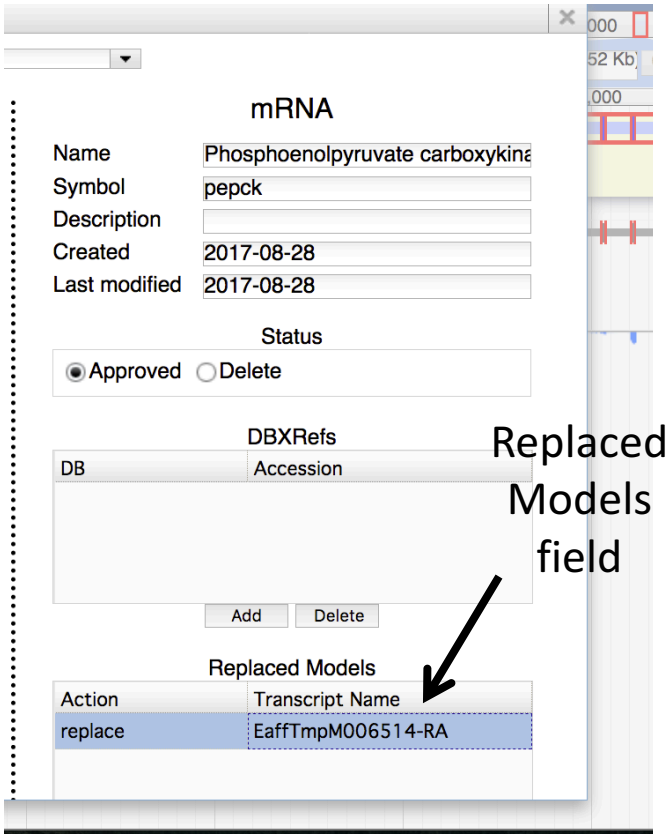
- Select the model in Apollo, then right-click, and select 'Edit Information' from the drop-down menu
  - Use the 'mRNA' section
  - Name: We recommend the UniProt naming guidelines:
    - http://www.uniprot.org/docs/nameprot
    - If a naming convention exists, use it (e.g. for gene families)
    - 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



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!



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

USDA

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

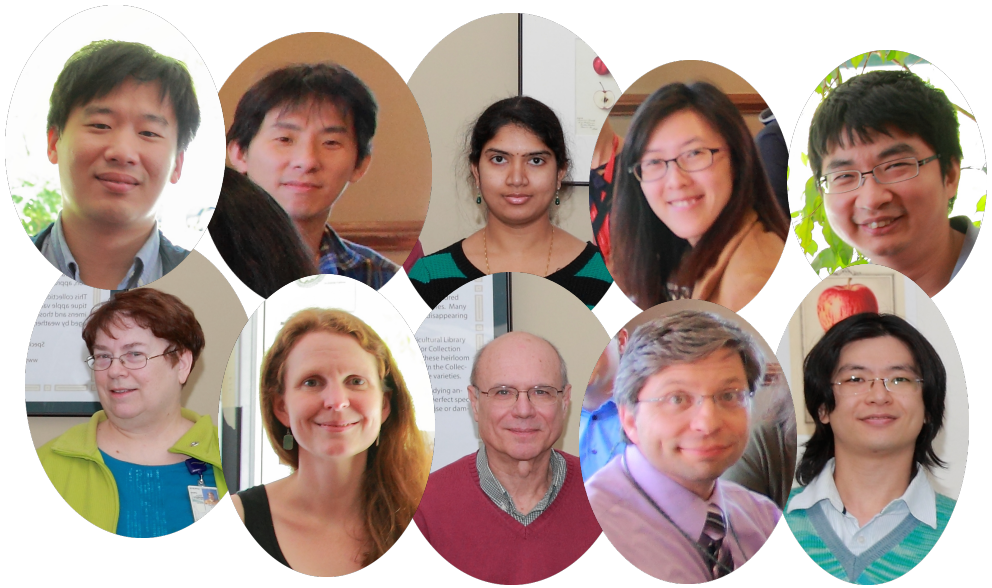
- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Gary Moore
- Susan McCarthy

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!



USDA