Annotation Resources at the National Agricultural Library

NAL USDA-ARS

https://i5k.nal.usda.gov

September 19th, 2016

USDA

Agenda

- Initial conventions
- Apollo Access/Overview
- Introduction to the Apollo interface
  - Viewing evidence tracks
  - Logging in/out
- BLAST
- Annotation
  - Process overview
  - Finalizing annotations
- Additional tips and tricks
- Example annotation

Questions for you.

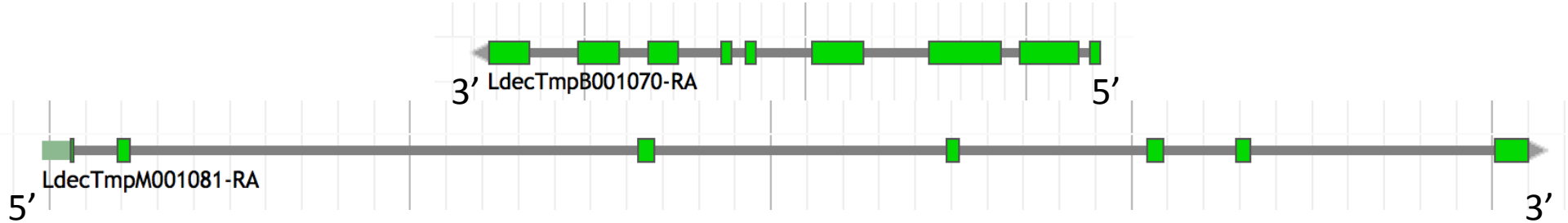
- Do you have any experience using Apollo already?
- How familiar are you with the terms exon, intron, UTR, 3’ end, 5’ end, splice site? Isoform?
- Do you know what the tracks in the *E. affinis* genome browser are/where they came from?
- Do you know what genes you want to annotate in the *E. affinis* genome?

Other resources.

- Moni Munoz-Torres' tutorial to your group from last year is pretty comprehensive:
  - http://www.slideshare.net/MonicaMunozTorres/introduction-to-apollo-i5k-e-affinis
- The official Apollo annotation guide:
  - http://genomearchitect.org/users-guide/
- Another manual curation tutorial: https://i5k.nal.usda.gov/manual-curation-example
- For more information about the *E. affinis* genome project, including data downloads:
  - https://i5k.nal.usda.gov/Eurytemora_affinis

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
  - 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



Apollo/JBrowse

- 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

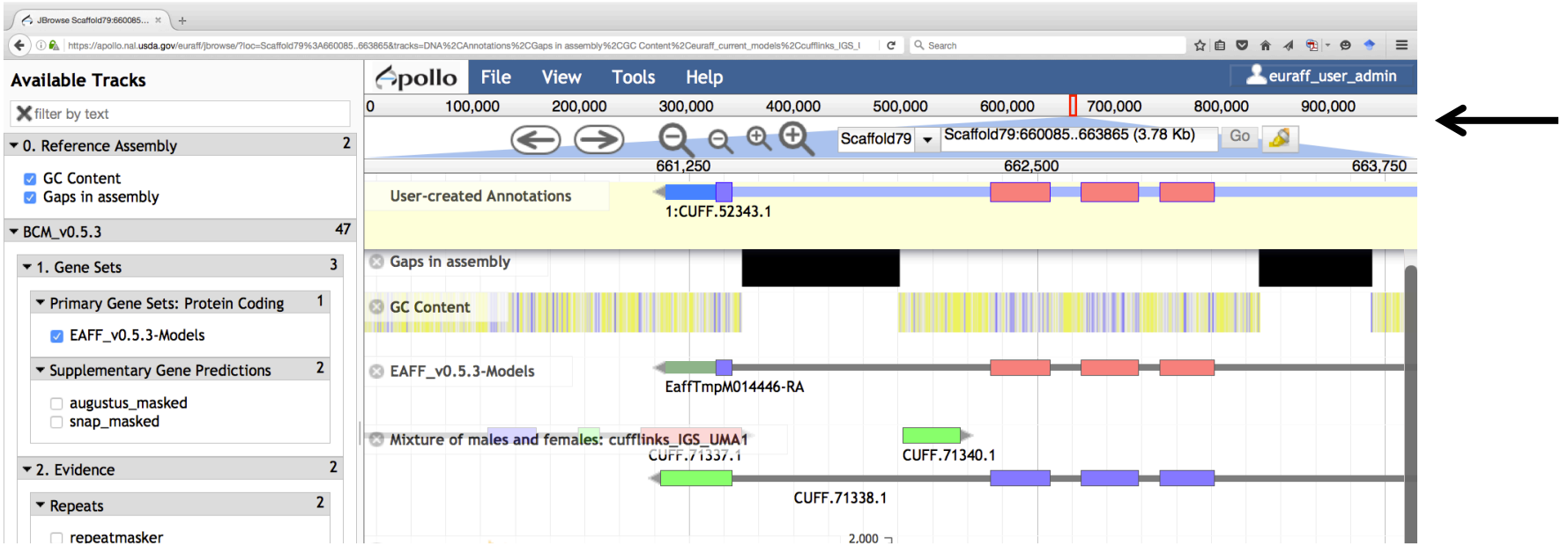
Viewing evidence

The “Available Tracks” panel lists all tracks (grouped hierarchically)

Viewing evidence

The Navigation panel describes the viewed region

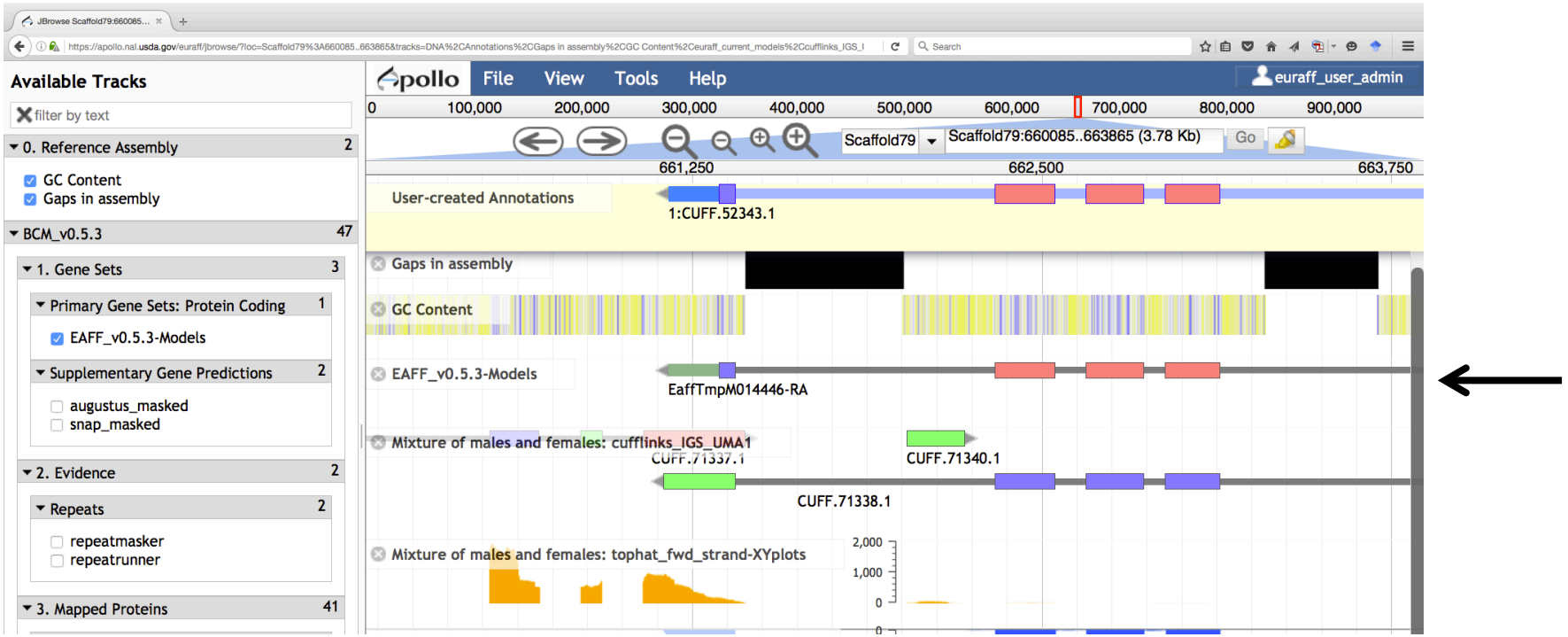
- Coordinate ranges for the entire sequence on top
- Navigation controls including pan, zoom and search in the middle
- Coordinate ranges for the current view along the bottom



Viewing evidence

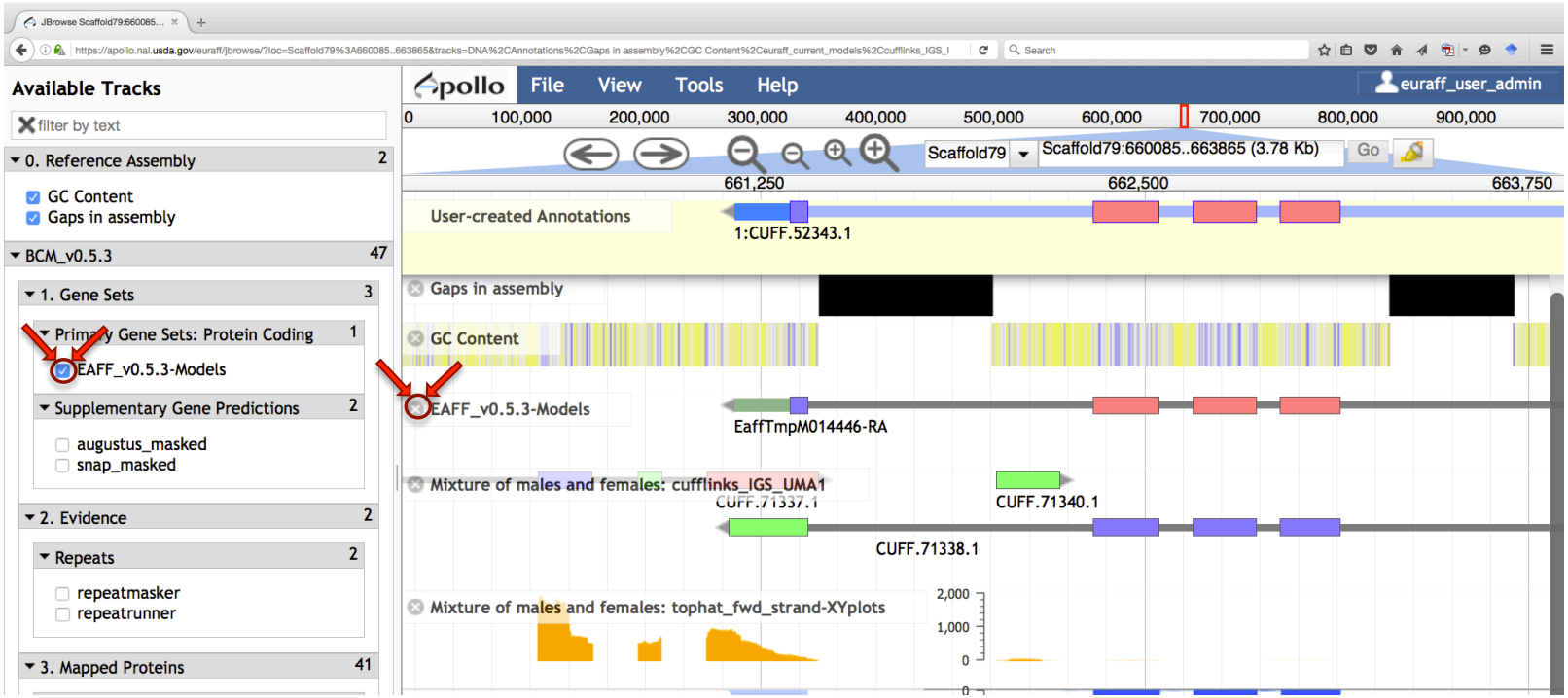
The Evidence panel shows the features in genomic context

- Each type of data is a separate track



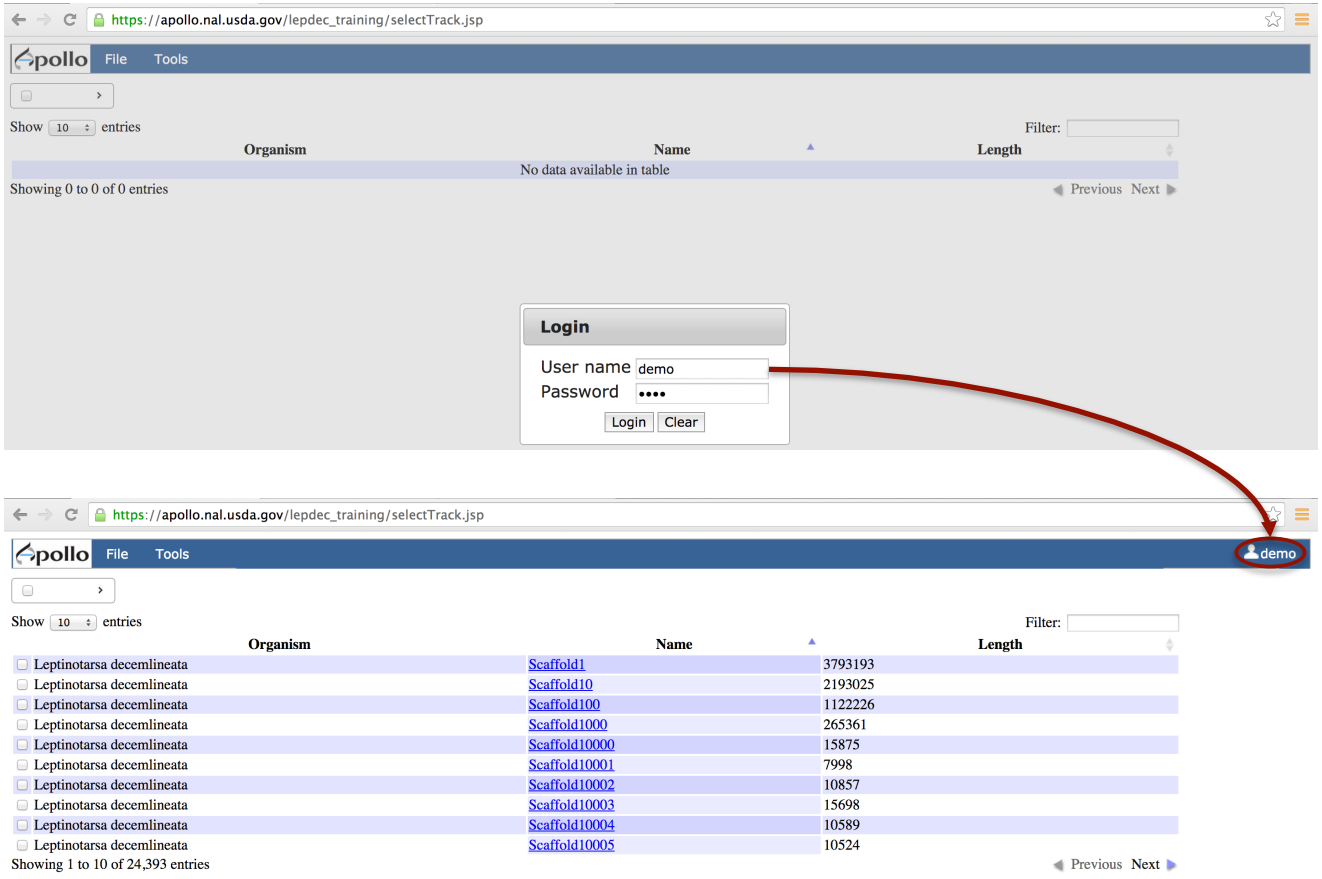
Viewing evidence

- Turning tracks on or off
  - Click a tracks checkbox to turn it on or off
  - Click ‘X’ by the track label also turns the track off



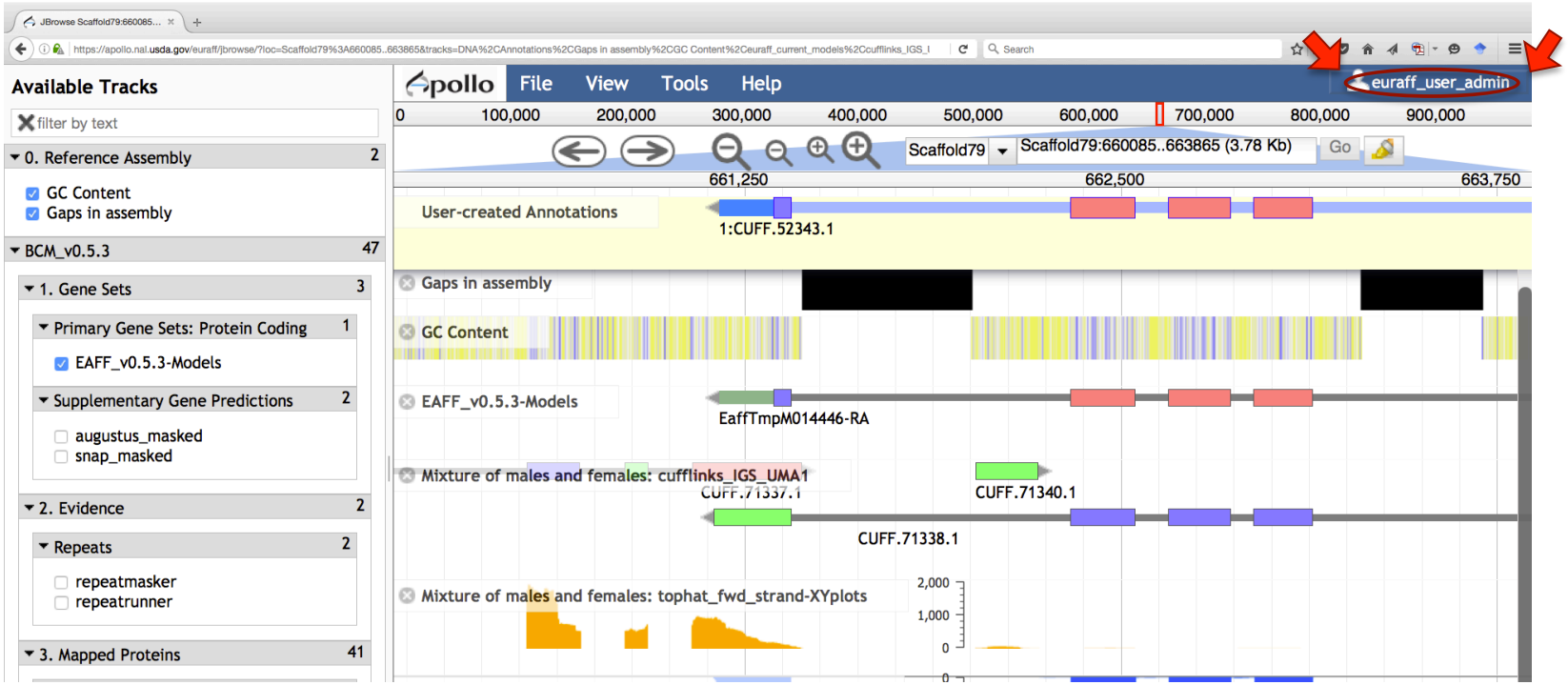
Login

- For scaffold selection view:



Login

- Directly from JBrowse genome browser

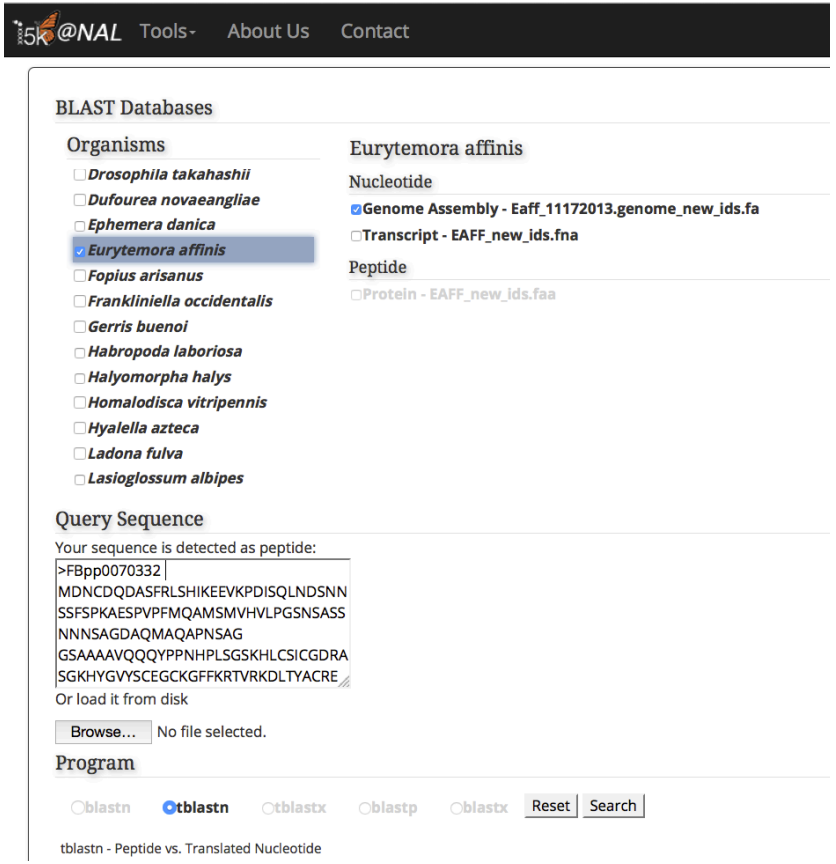


The user-created annotations track

BLAST at the i5k Workspace

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

- Interface updates options automatically
  - Based on the combination of query and database
  - Chooses correct options (blastn, blastx, tblastn, etc)

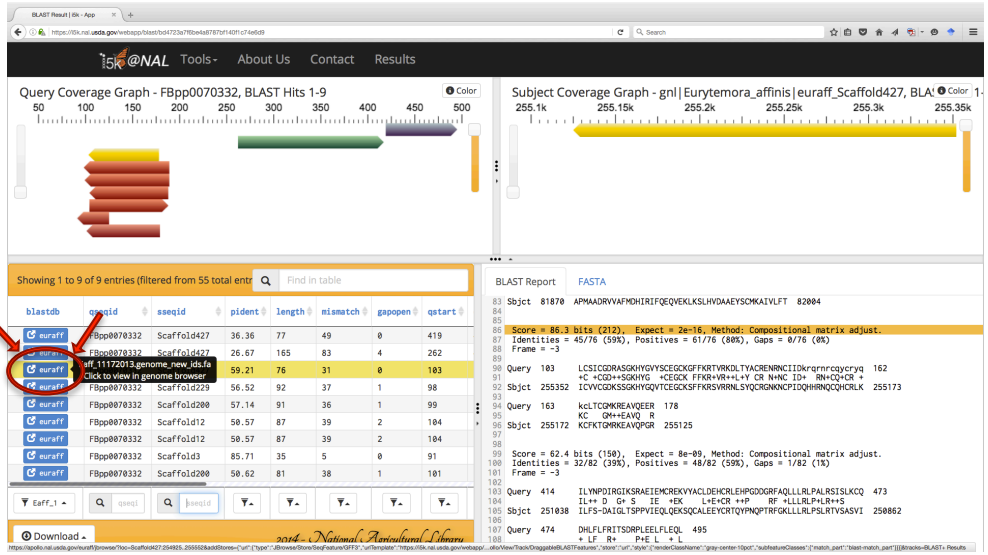


USDA

BLAST at the i5k Workspace

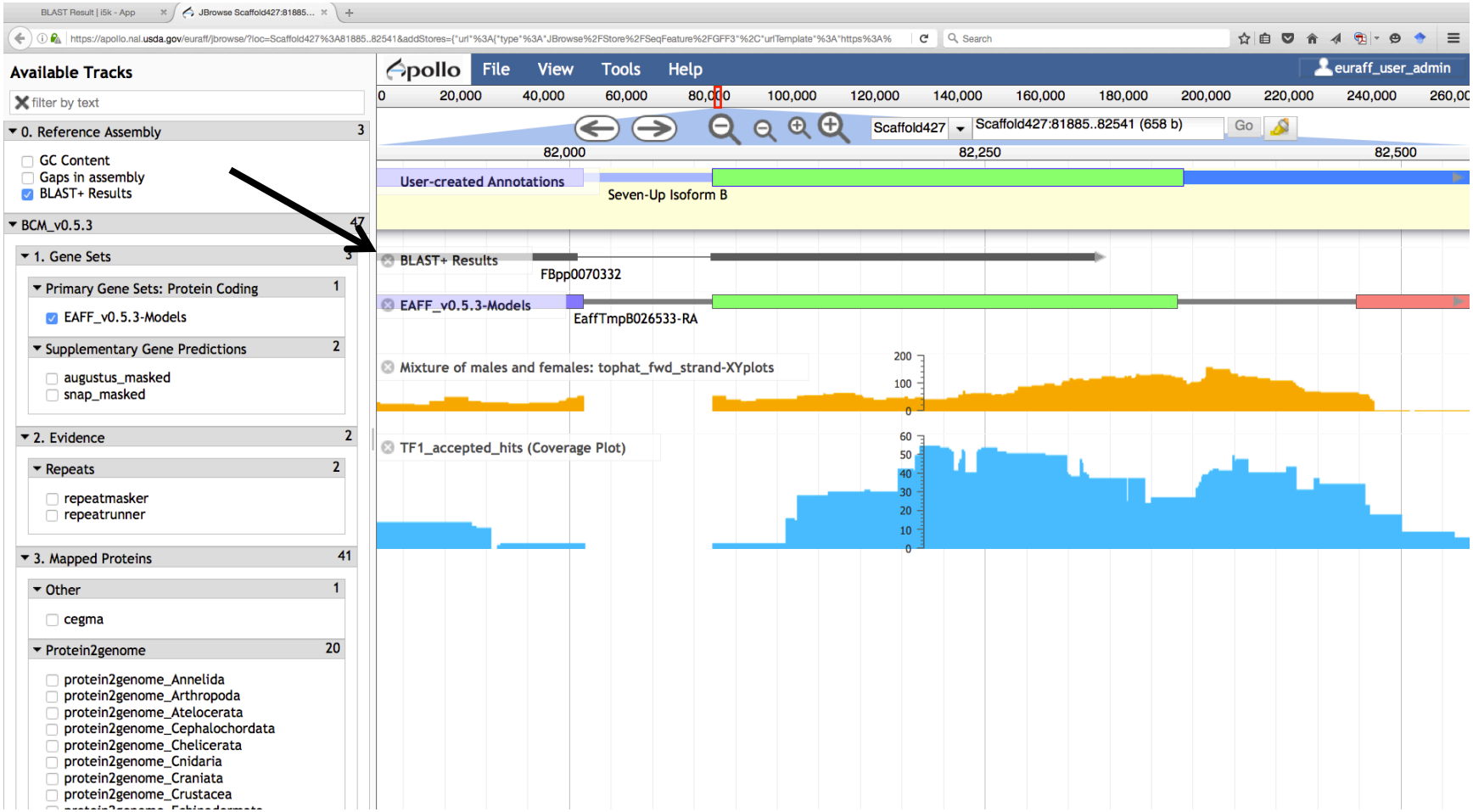
https://i5k.nal.usda.gov/webapp/blast/

- Results persist for one week
  - The results URL can be saved or shared
  - USE THE FILTERS!
- HSPs are viewable in the genome browser
  - HSPs are conservatively chained together
  - Visible as connected blocks



Result URL:
https://i5k.nal.usda.gov/webapp/blast/bd4723a7f6be4a8787bf140f1c74e6d9

BLAST results in Apollo



Annotation Guidelines

cf. http://www.slideshare.net/MonicaMunozTorres/introduction-to-apollo-i5k-e-affinis

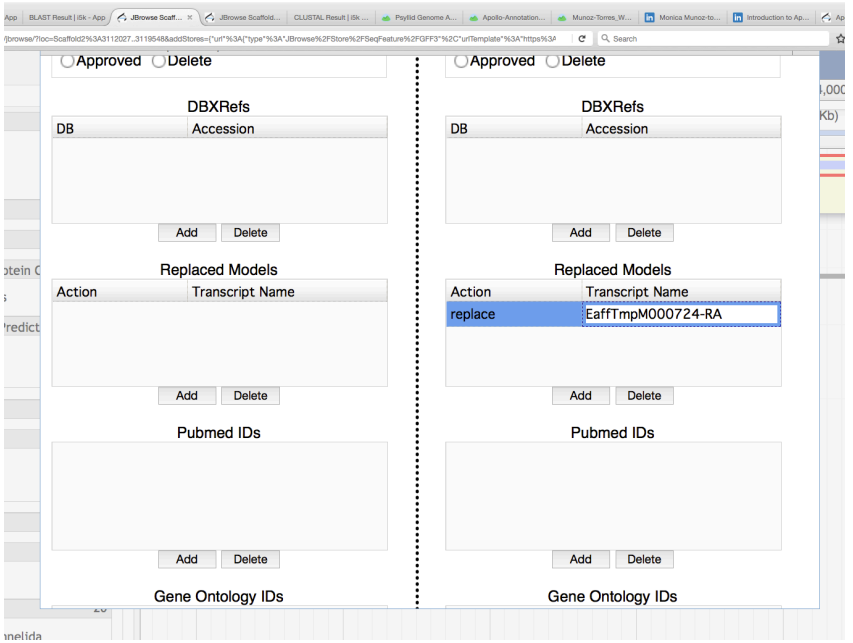
1. Find sequence of interest from one or several other species (e.g. *Drosophila melanogaster* and *Daphnia pulex*)
    1. If you're annotating a protein, use the protein sequence, and know where the protein domains are
2. Align sequence to genome or protein set in order to locate potential gene region
3. Examine available evidence (tracks) to begin model curation
4. Adjust model based on available evidence
5. Compare sequence of annotated model to related sequences
6. Repeat steps 1 through 5 as needed to refine model
7. Add annotation details in the "Information Editor"
    - Replaced model, name, symbol, other comments

When finalizing annotations

- Check start, stop and exon boundaries (splice sites)
- 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
  - Add comments that describe
    - your evidence for the annotation
    - Modifications that you made to the gene model

The Replaced Models field

- Used to generate a merged, non-redundant gene set from the manually curated models in Apollo and the Maker models (EAFF_v0.5.3)



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

USDA

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 }
- You can pin tracks to the top
- If you know the name or ID of the MAKER gene that you'd like to annotate, you can paste it into the search box to navigate to it

Notes on *E. affinis* genome/browser

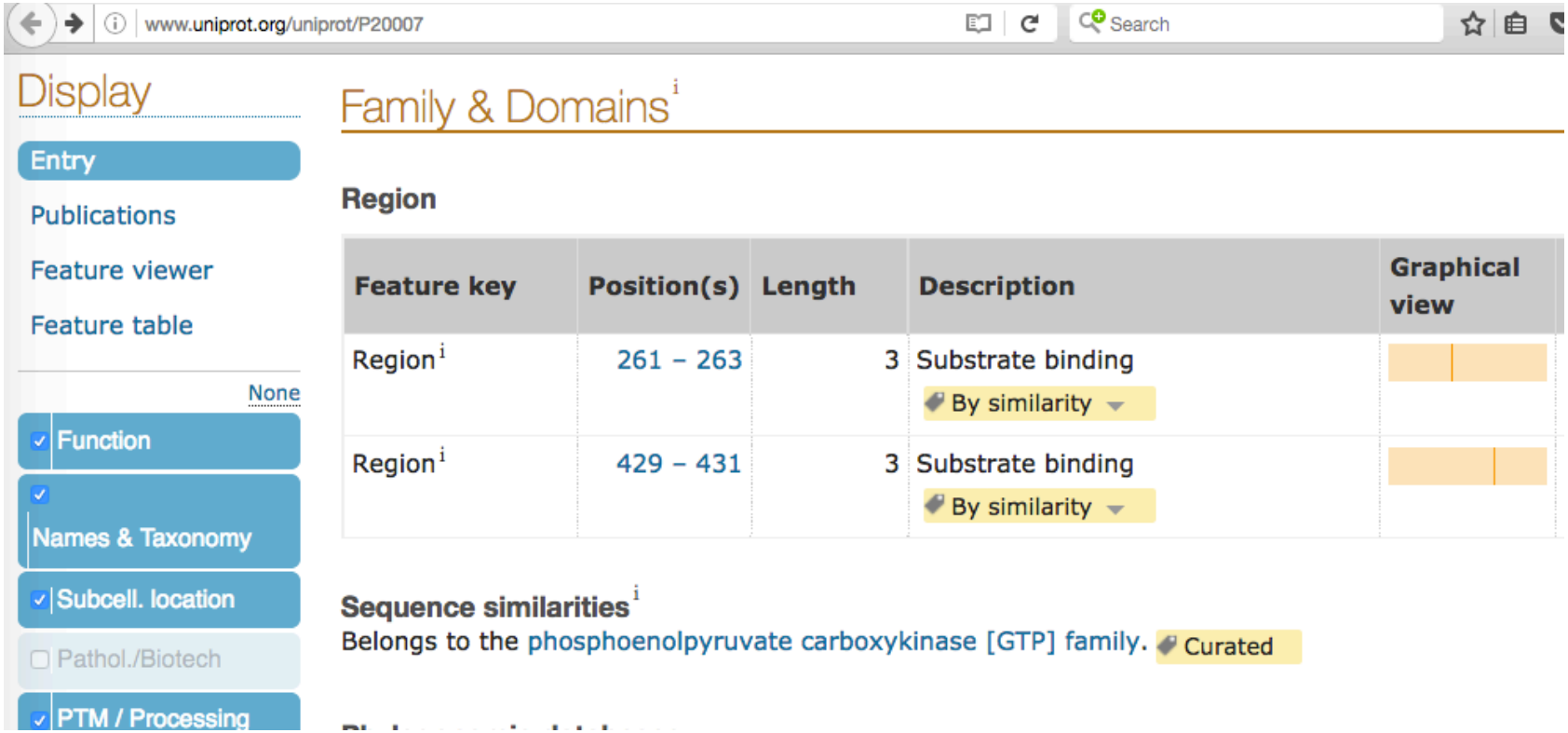
- Big advantage for annotation: you have lots of RNA-Seq and transcriptome data to use as contributing evidence for your gene models
  - Includes strand-specific RNA-Seq
- Disadvantage: No close reference genomes, so it may be harder to find homologs for your genes of interest to inform your annotations.

Live Annotation Example

- Phosphoenolpyruvate carboxykinase (pepck) in *Eurytemora affinis*
- What reference protein do I choose to find my protein in the *E. affinis* genome?
  - Learn how to use and understand UniprotKB and NCBI databases
- How do I know that I found the right *E. affinis* protein?
- Once I am fairly confident of the protein, how do I correct the sequence? Do I need to?
- What do I enter into the Information Editor?

Choosing reference proteins: *D. melanogaster* pepck

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

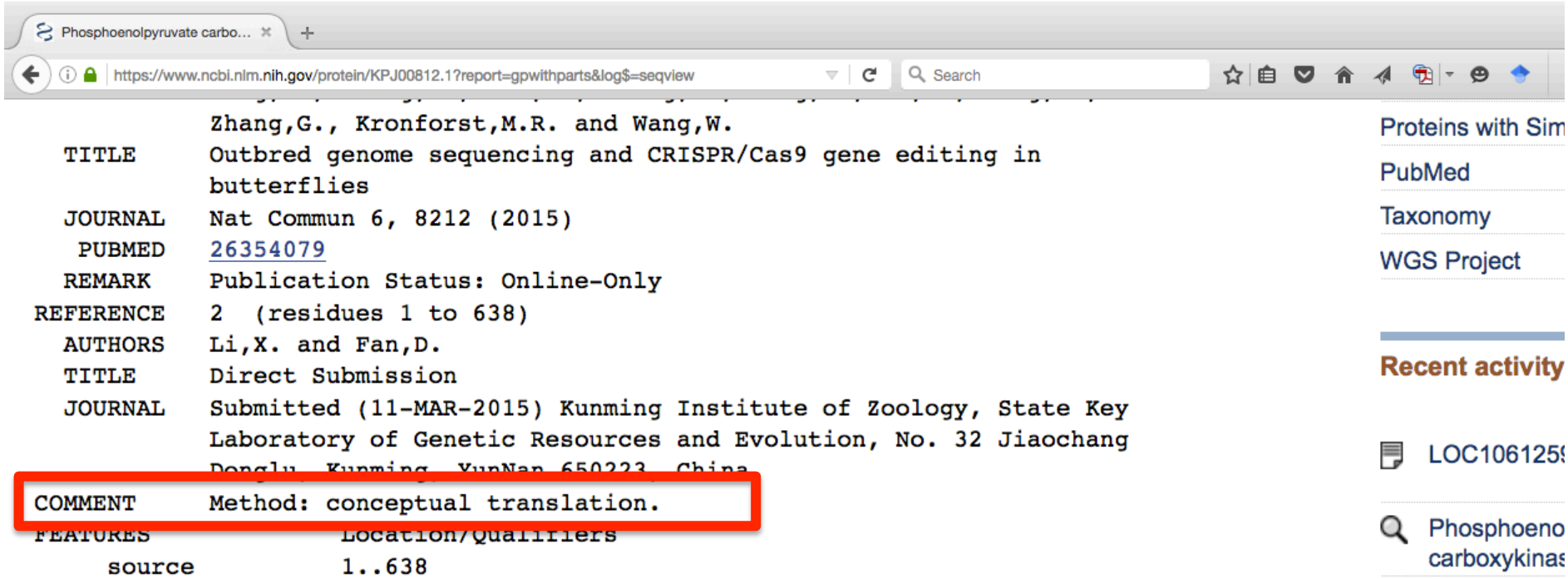


Source: http://www.uniprot.org/uniprot/P20007

USDA

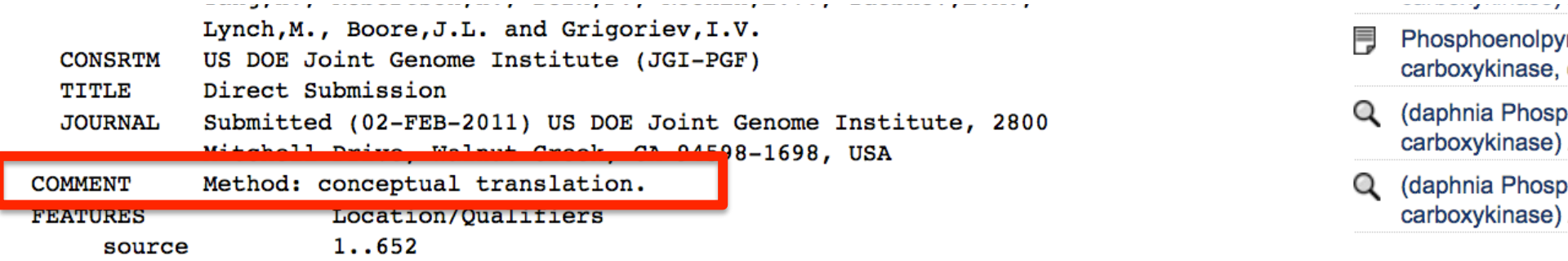
Choosing reference proteins: *Papilio xuthus* Pepck

- GenBank record: https://www.ncbi.nlm.nih.gov/protein/KPJ00812.1?report=gpwithparts



Choosing reference proteins: *Daphnia pulex* Pepck

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

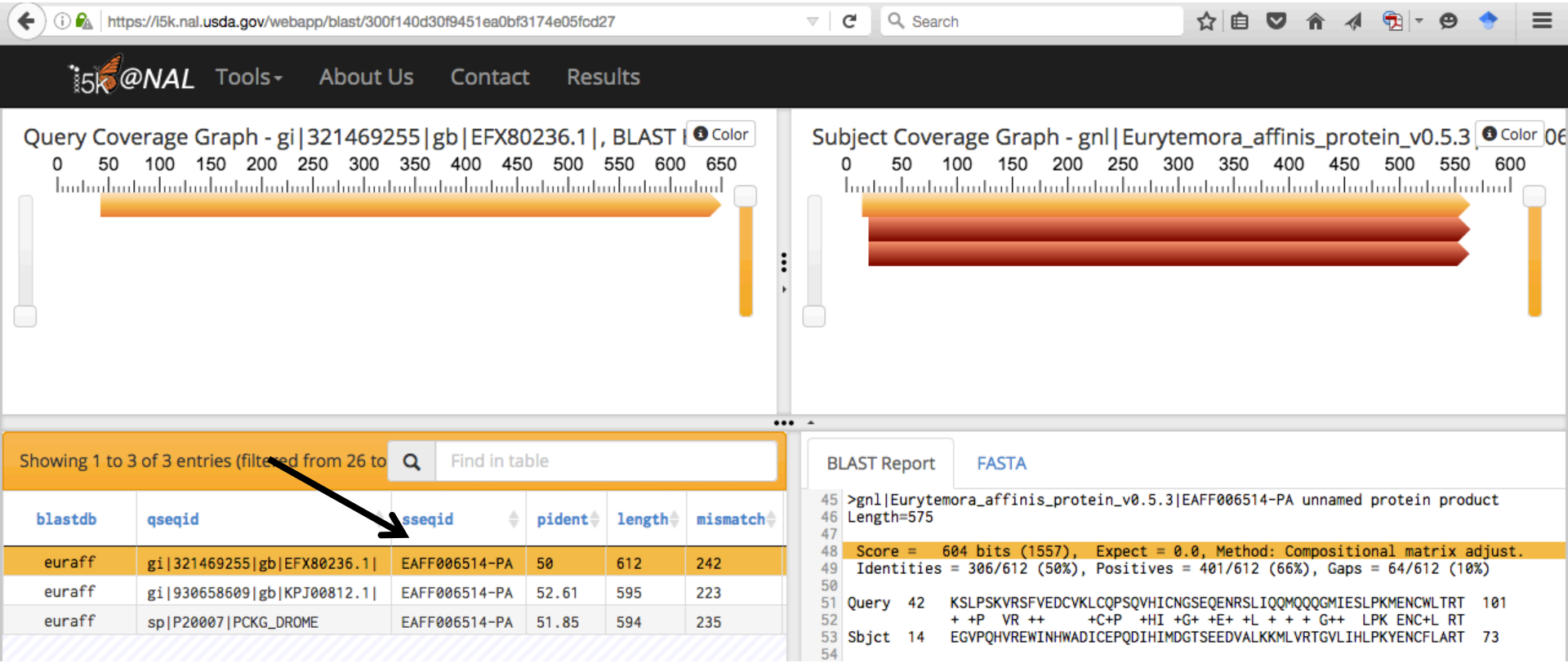


```
            Lynch,M., Boore,J.L. and Grigoriev,I.V.
  CONSRTM   US DOE Joint Genome Institute (JGI-PGF)
  TITLE     Direct Submission
  JOURNAL   Submitted (02-FEB-2011) US DOE Joint Genome Institute, 2800
                                                    98-1698, USA
  COMMENT   Method: conceptual translation.
FEATURES             Location/Qualifiers
     source          1..652
```

BLAST dmel, dpul, pxut proteins against *E. affinis* proteins and genome

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

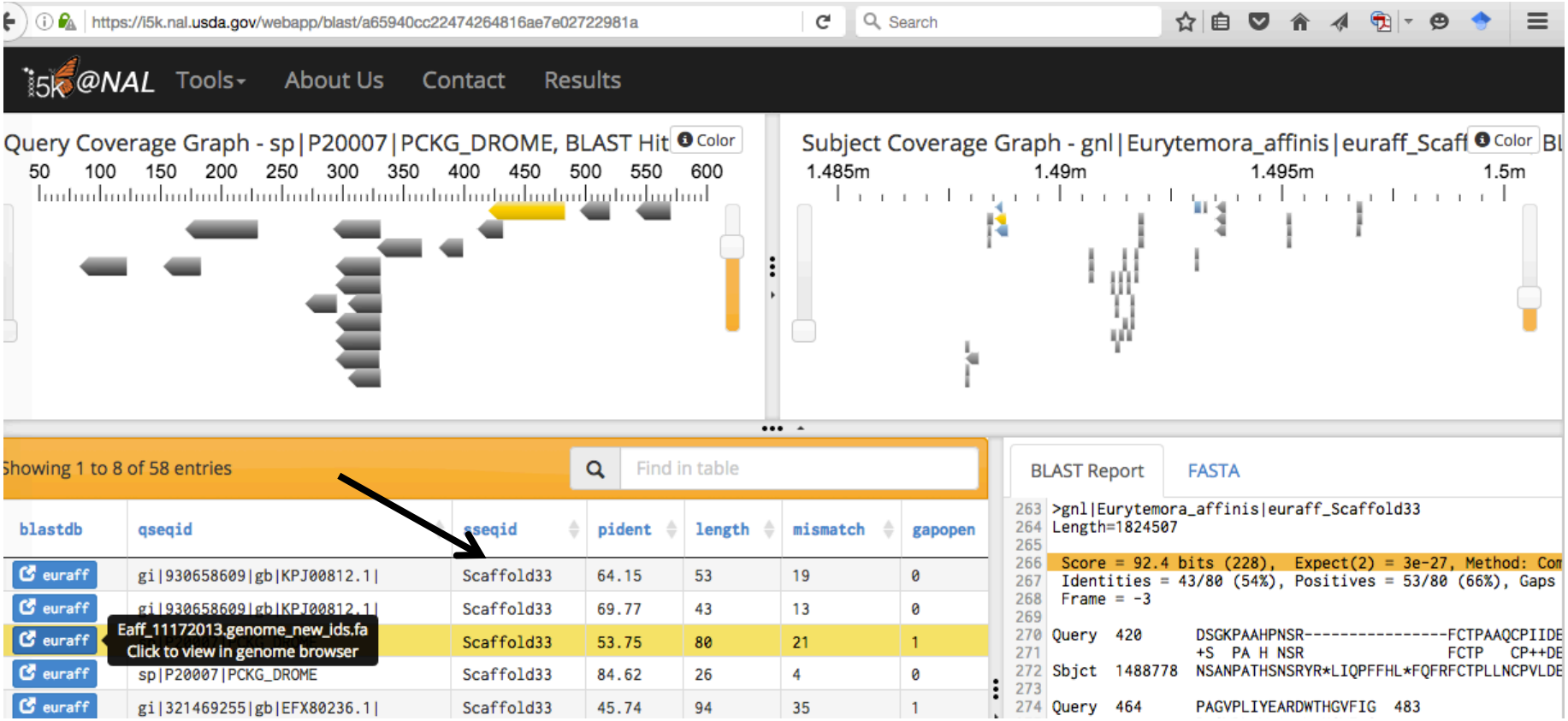
- First, protein results (top hit is on Scaffold33):



Result URL: https://i5k.nal.usda.gov/webapp/blast/300f140d30f9451ea0bf3174e05fcd27

BLAST dmel, dmag, pxut proteins against *E. affinis* proteins and genome

- Genome results:



Result URL: https://i5k.nal.usda.gov/webapp/blast/a65940cc22474264816ae7e02722981a

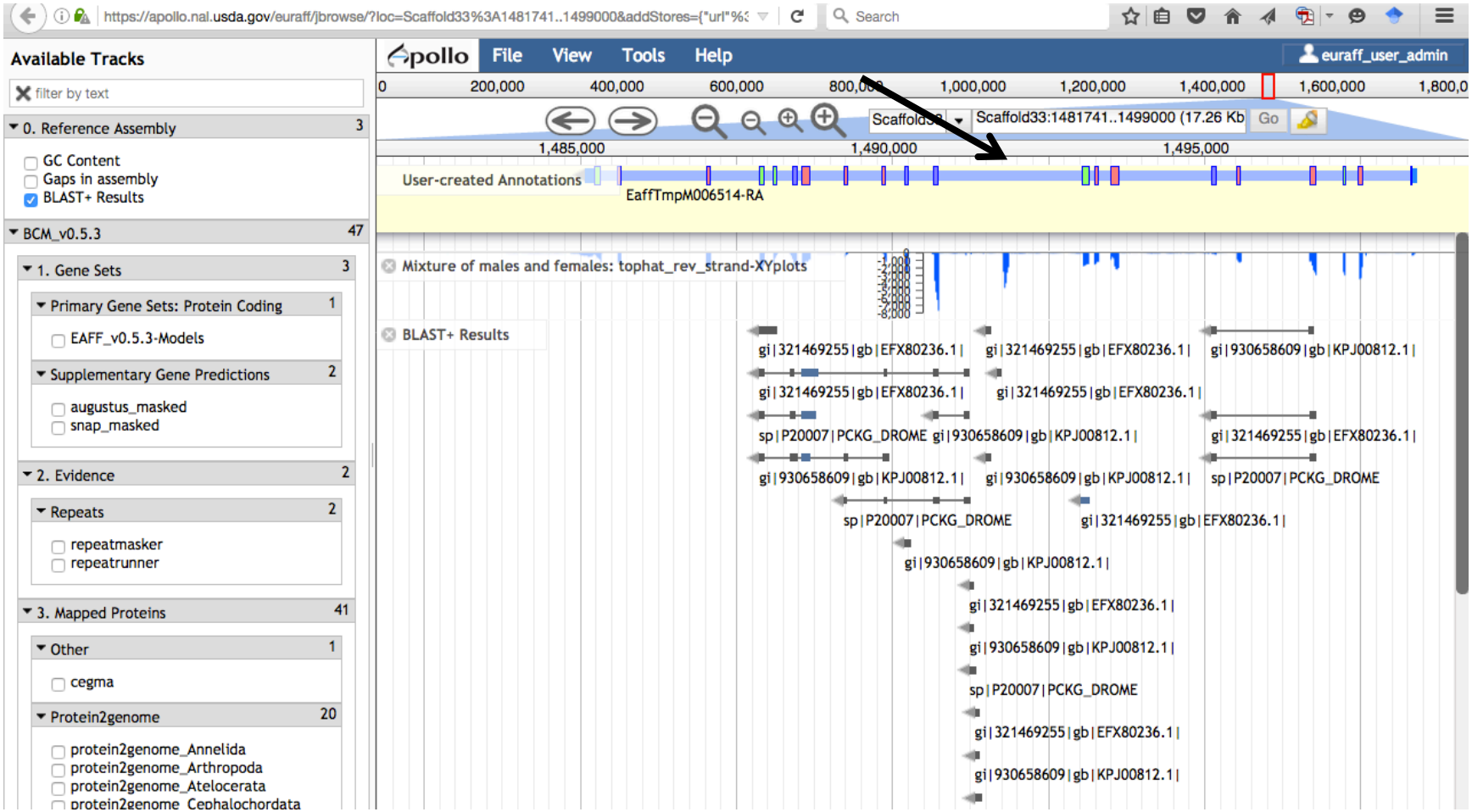
Modify *E. affinis* model sequence in Apollo

- Click on blue 'euraff' button next to HSP of choice in tabular view to view Blast HSPs in JBrowse
- Alternatively, go to Apollo URL: https://apollo.nal.usda.gov/euraff/jbrowse/
  - find 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

Modify *E. affinis* model sequence in Apollo

- Questions:
  - What evidence do you choose?
  - 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



Evaluating protein sequence

- Add missing exon to model
- Align EAFF006514-PA sequence to dmel, dmag, ptux proteins
  - Clustal (https://i5k.nal.usda.gov/webapp/clustal/)
  - Can also use NCBI Blast: http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins&PROGRAM=blastp&BLAST_PROGRAMS=blastp&PAGE_TYPE=BlastSearch&BLAST_SPEC=blast2seq&DATABASE=n/a&QUERY=&SUBJECTS=
- Check alignment extent, %ID

Clustal result URL:
https://i5k.nal.usda.gov/webapp/clustal/68361de413724675a836e4fcf8a21361

Evaluating protein sequence

- Blast modified EAFF006514-PA sequence to nr
  - Make sure it doesn't match a potential contaminant
  - Get an idea whether you have the right sequence
- Once contamination is ruled out, it's better to blast your sequence against a smaller db of high-quality proteins
- If you notice that parts of the protein are missing, check the 'Gaps in assembly' track in the browser

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 – if you’re not sure, leave it blank
  - Replaced Models Field – the Maker model (EAFF_v0.5.3) that your new model will replace in the OGS
  - 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

Closing thoughts

- Your annotation work is a community effort.
  - Thank you for contributing!
  - Because annotating via Apollo is a community effort, you might notice that someone else might already be working on your model of choice. In that case, get in touch with them and collaborate!
  - Keep in mind that your work will be used by the scientific community once you're done.