Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL

April 21st, 2020

USDA

Agenda

- Manual annotation general overview
- I5k Workspace tools for manual annotation
  - BLAST, Clustal, HMMER
  - Apollo2
- Manual annotation example: preparation
- Manual annotation live example
- Isoform annotation example

Other resources

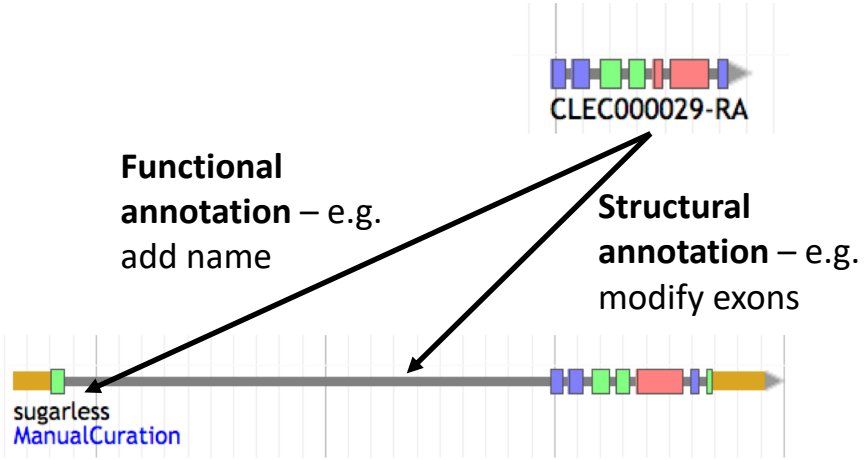
- Monica Munoz-Torres from the Apollo group has a number of comprehensive tutorials:
  - https://www.slideshare.net/MonicaMunozTorres/presentations
    - I recommend these slides if you need more background:
      - https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015
  - If you are new to Apollo, or need a refresher, I **highly recommend** that you review one of her presentations
- The official Apollo annotation guide:
  - http://genomearchitect.org/users-guide/
- I5k Workspace manual annotation landing page: https://i5k.nal.usda.gov/manual-annotation-and-apollo
- Other manual curation tutorials: http://genomecuration.github.io/genometrain/d-feature-curation-crossing/
- VEuPathDB will be holding an Apollo training webinar on May 21st: https://eupathdb.org/eupathdb/webinars.jsp

MANUAL ANNOTATION GENERAL OVERVIEW

USDA

What is manual annotation?

- Manual review and improvement of an existing gene prediction
- Draw on external evidence (e.g. RNA-Seq, cDNA, genes from other species) to improve a computationally predicted gene model



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
- Link gene models to existing literature and ontologies, providing richer data

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

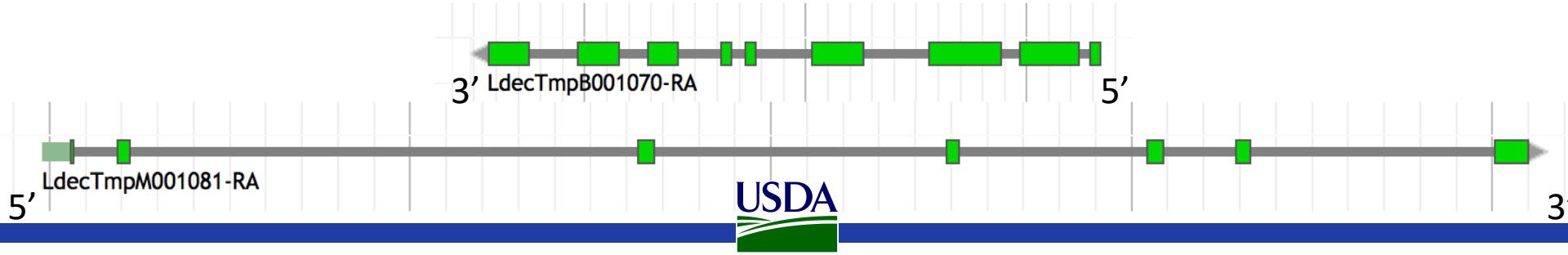
Adapted from https://www.slideshare.net/MonicaMunozTorres/apollo-workshop-at-ksu-2015

MANUAL ANNOTATION: I5K WORKSPACE TOOLS

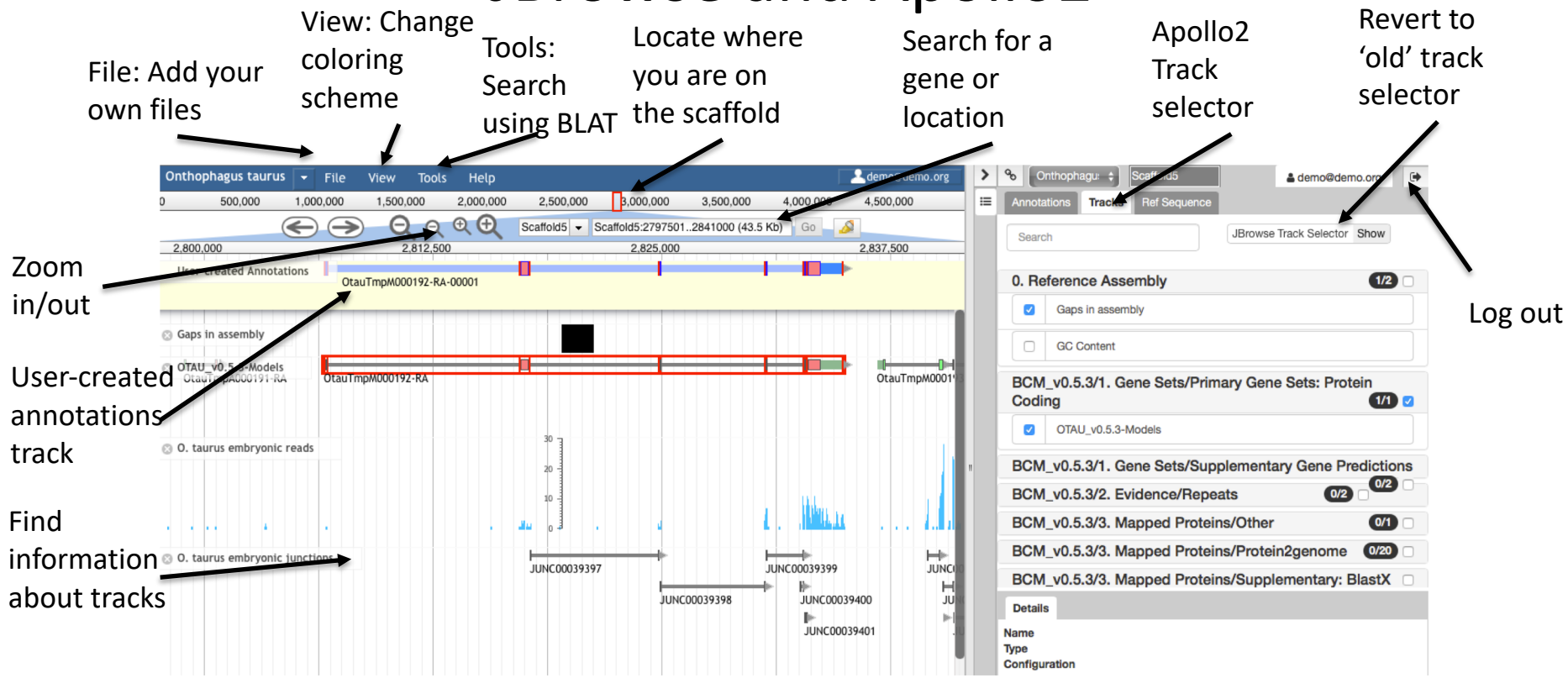
USDA

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 Apollo2



JBrowse is a web- based genome browser

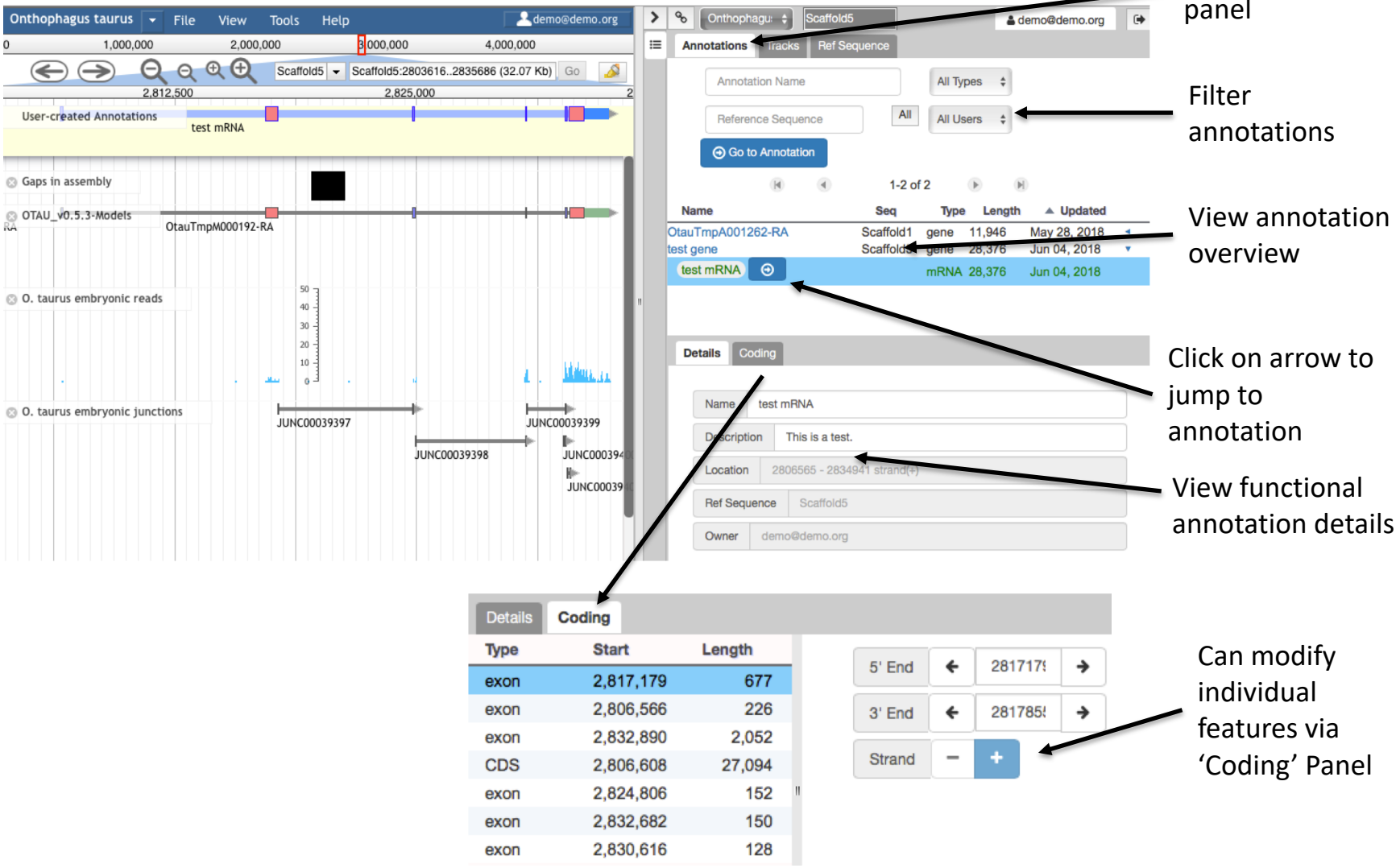
- Visualize features that are mapped to a genome
- These features are displayed as tracks
- Many different types of data may be displayed

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

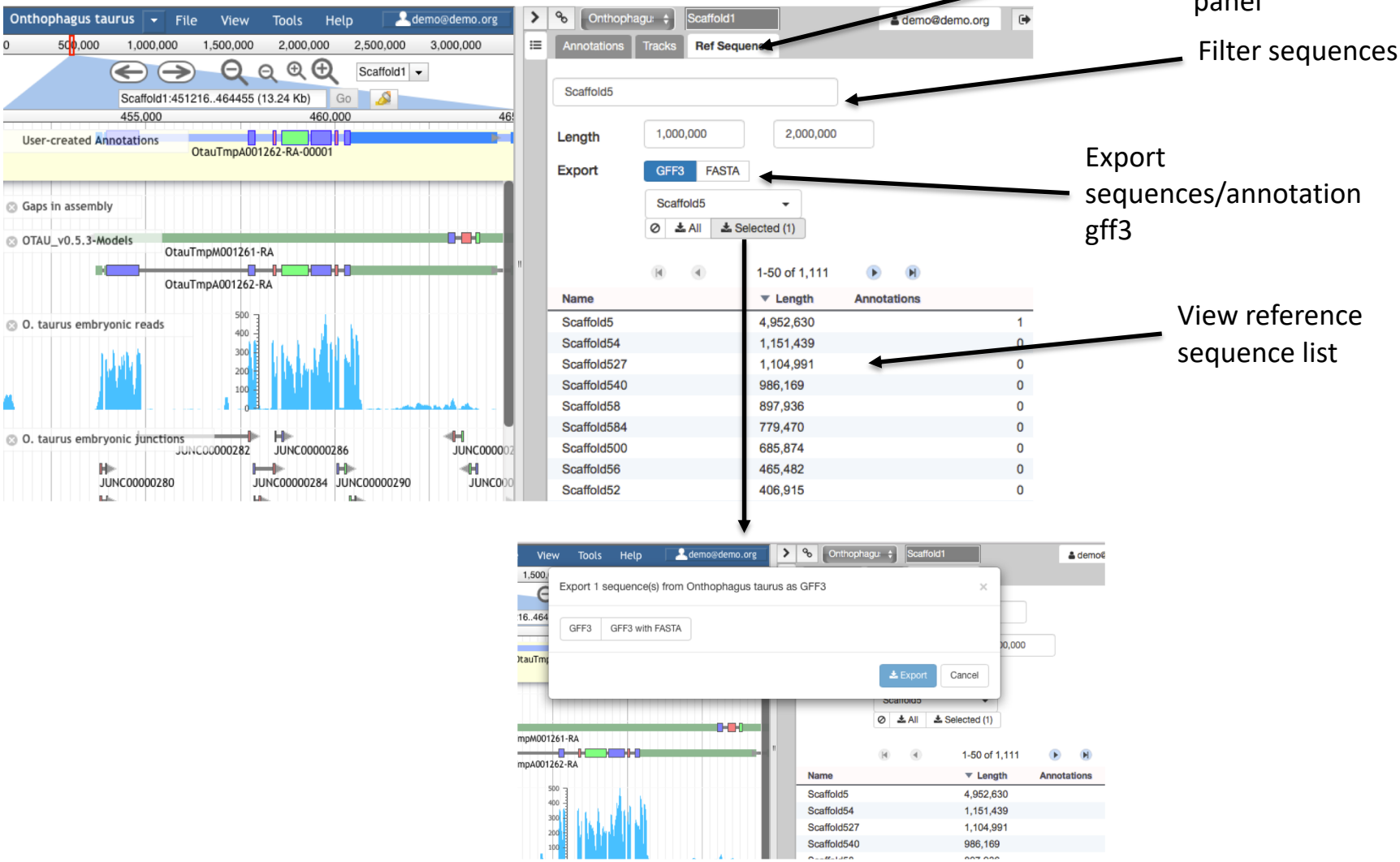
USDA

Apollo2 – Annotations Panel



USDA

Apollo2 – Ref Sequence Panel

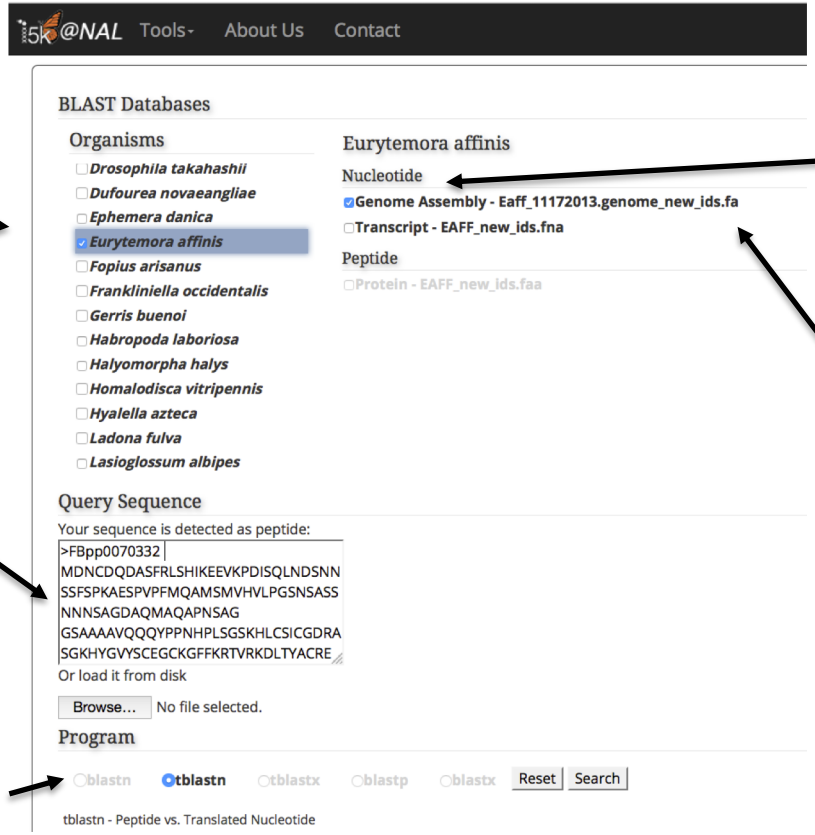


i5k Workspace BLAST: one way to access Apollo

Select organism

Paste or upload query sequence(s)

Program is automatically selected



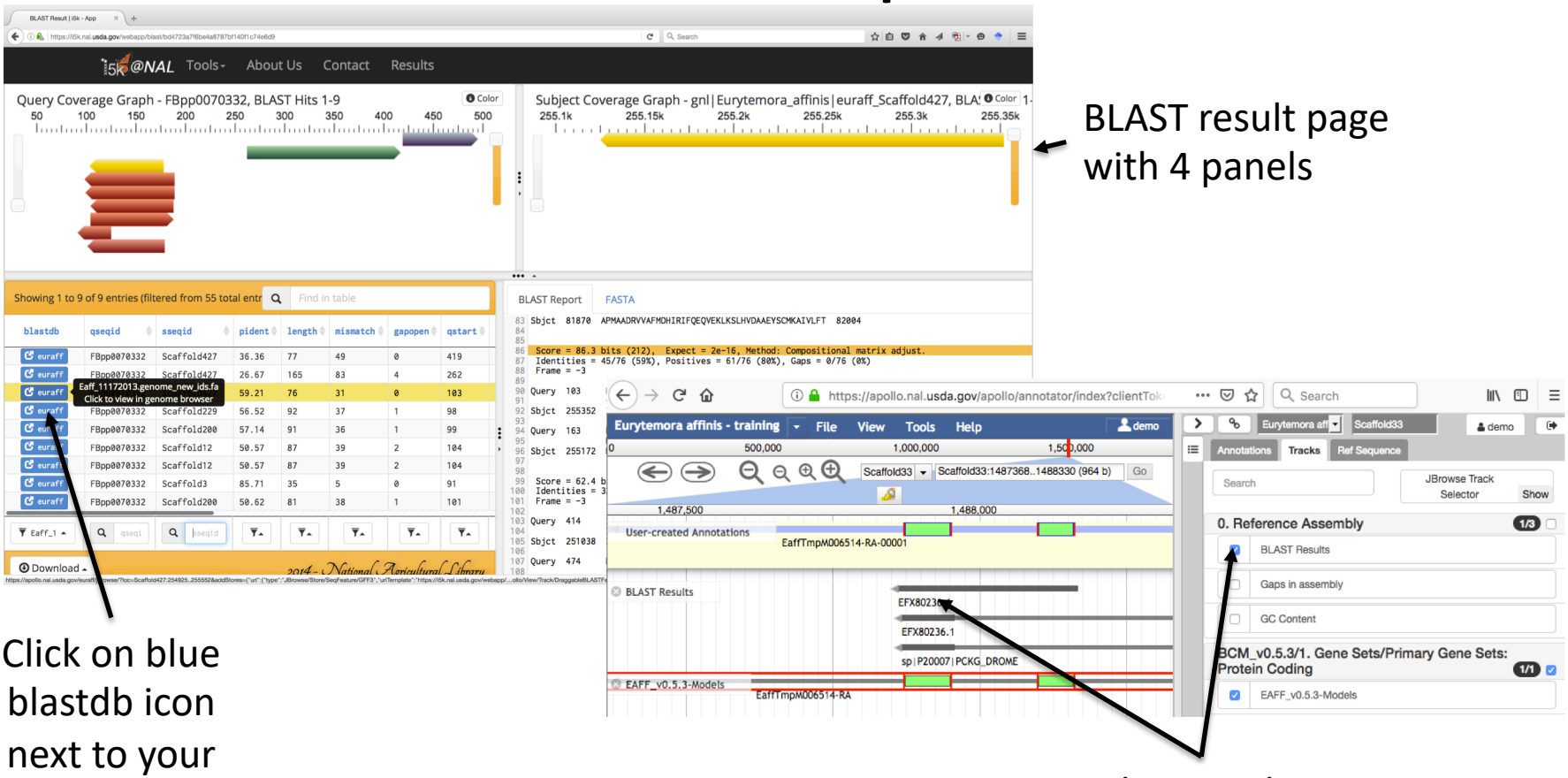
Select organism-specific database

BLAST against the genome assembly to view HSPs in Jbrowse

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

USDA

i5k Workspace BLAST: one way to access 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

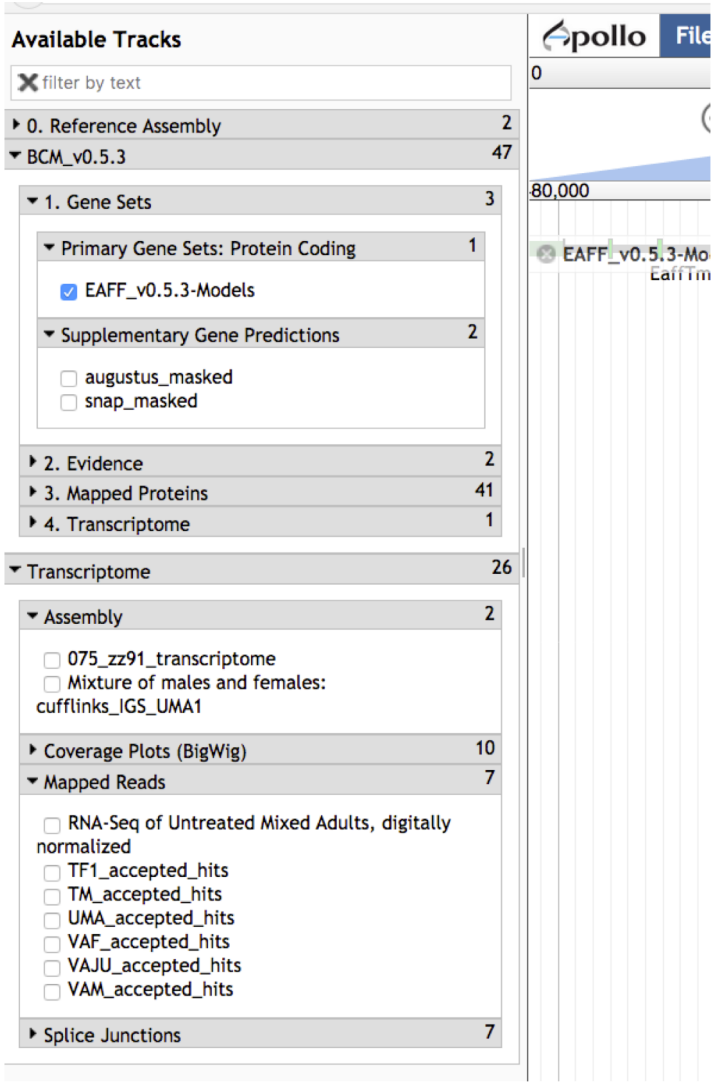
Annotation Example

- 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 (for training only): https://apollo.nal.usda.gov/apollo/3068161/jbrowse/index.html
  - Login credentials: demo/demo

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.

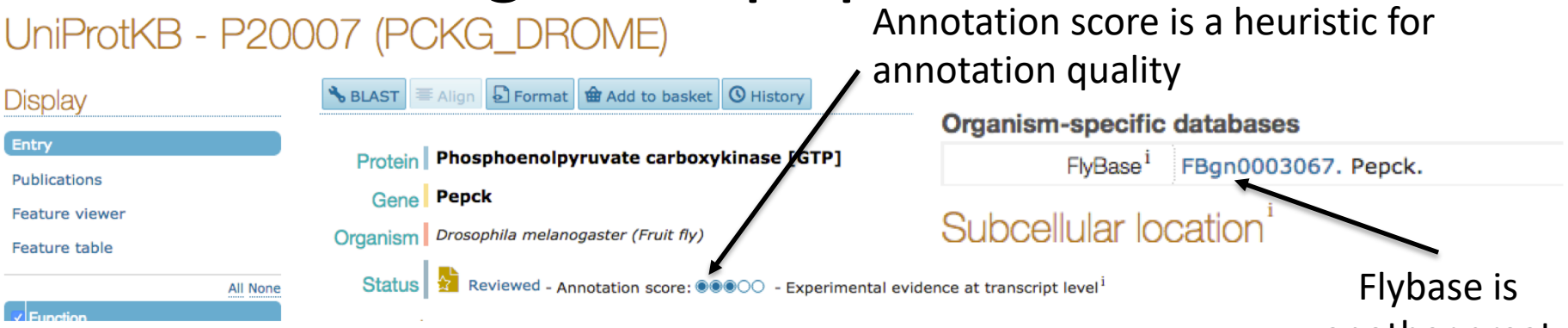
Available tracks for *E. affinis*



- 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

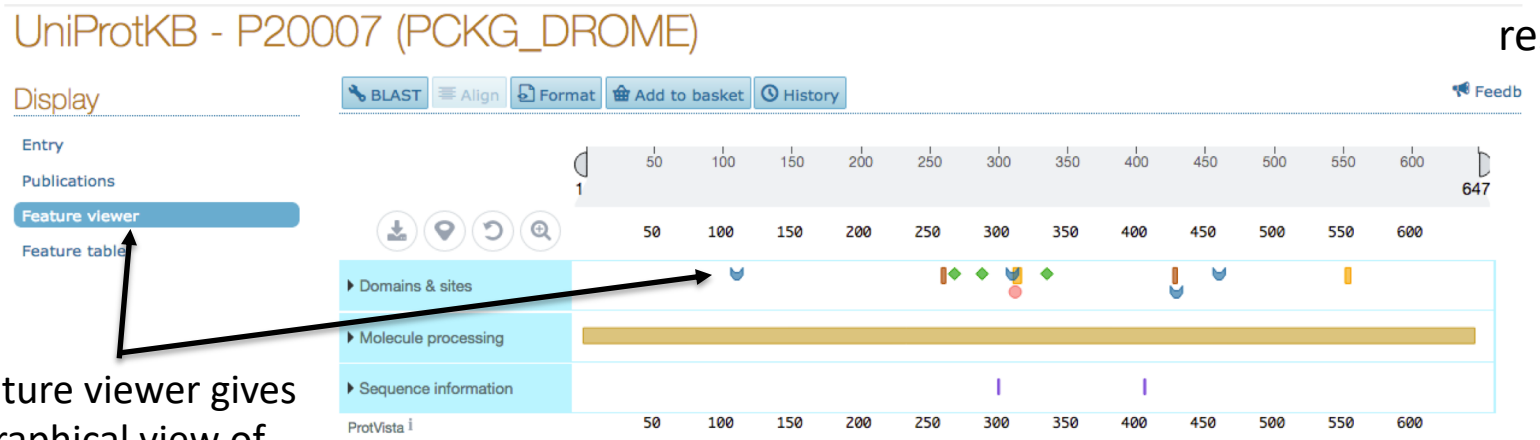
USDA

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

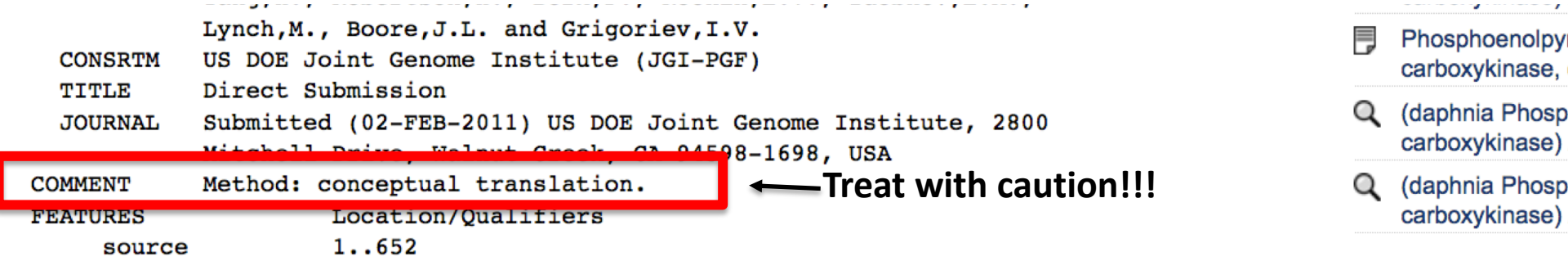
Catalyzes the conversion of oxaloacetate (OAA) to phosphoenolpyruvate (PEP).

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



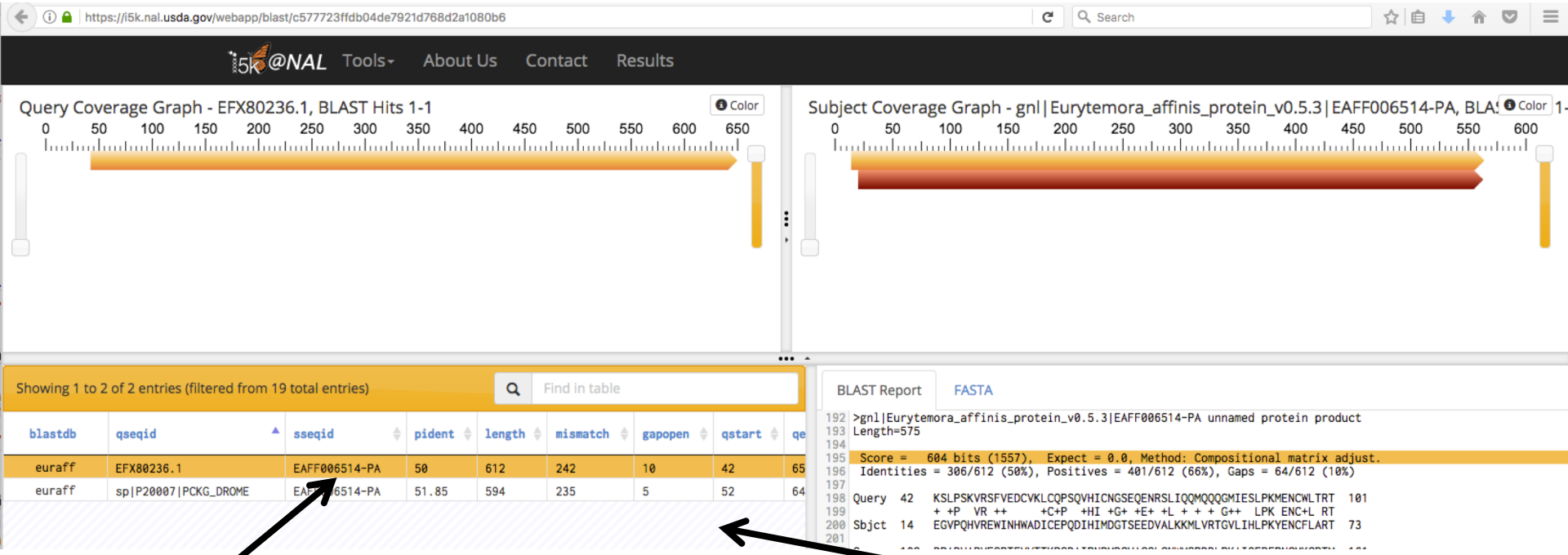
Resources for learning about insect gene/protein structure and function

- UniProt: https://www.uniprot.org/
- OrthoDB: https://www.orthodb.org/
- FlyBase: http://flybase.org/
- VectorBase: https://www.vectorbase.org/
- Hymenoptera Genome Database: http://hymenopteragenome.org/
- AphidBase/BIPAA: https://bipaa.genouest.org/is/

MANUAL ANNOTATION LIVE EXAMPLE

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

https://i5k.nal.usda.gov/training/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

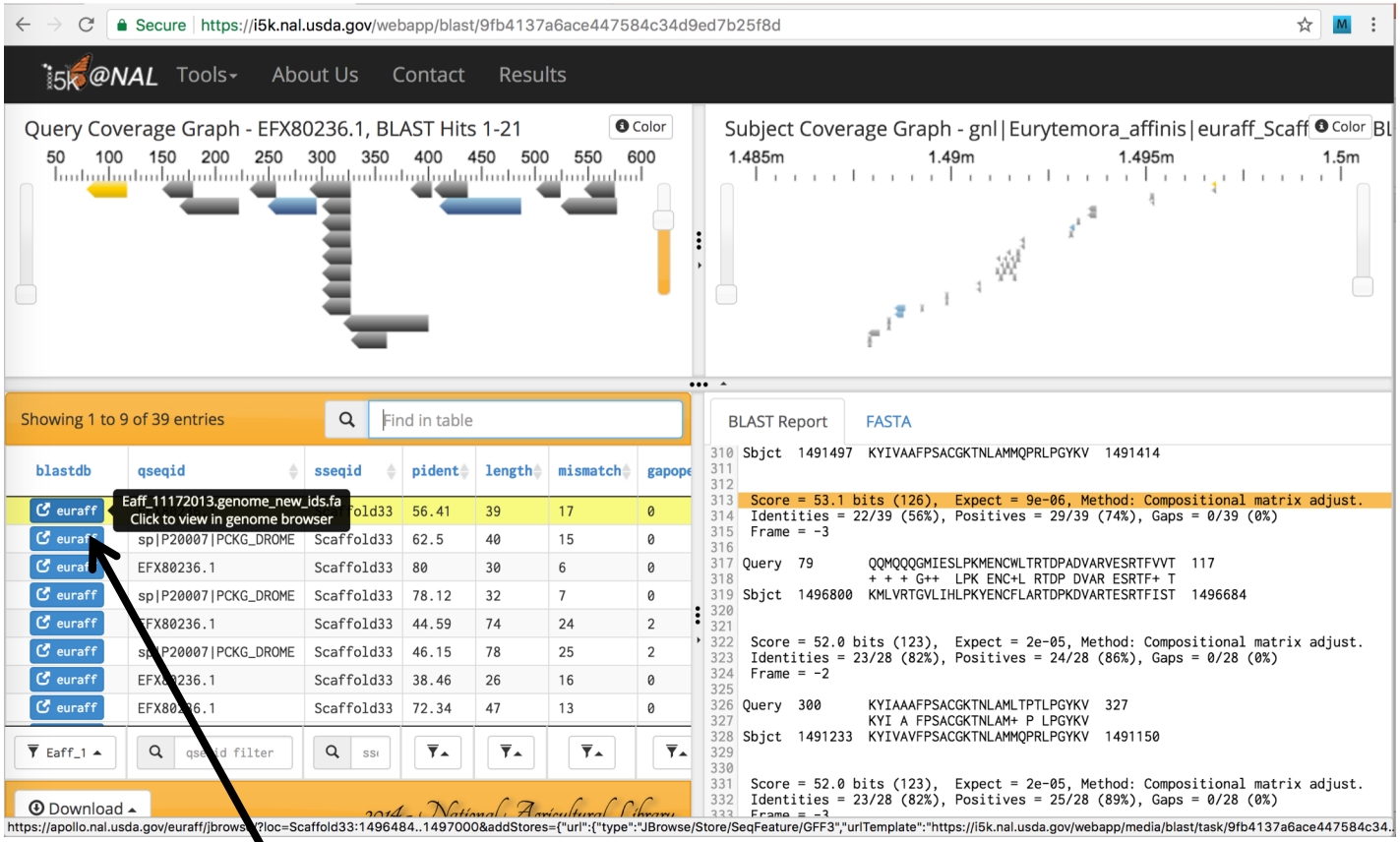
USDA

Modify *E. affinis* model sequence in Apollo

- Go to Apollo URL: https://apollo.nal.usda.gov/apollo/3068161/jbrowse/index.html
  - 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

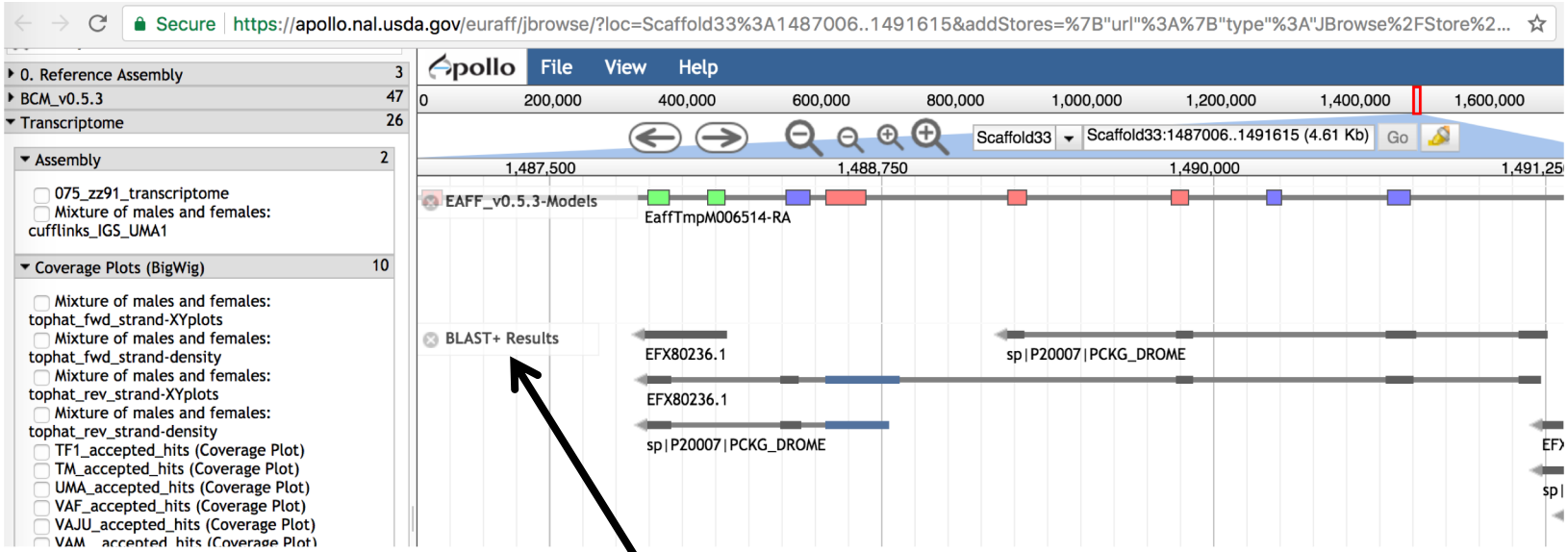
Another approach: BLAST against the genome

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



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

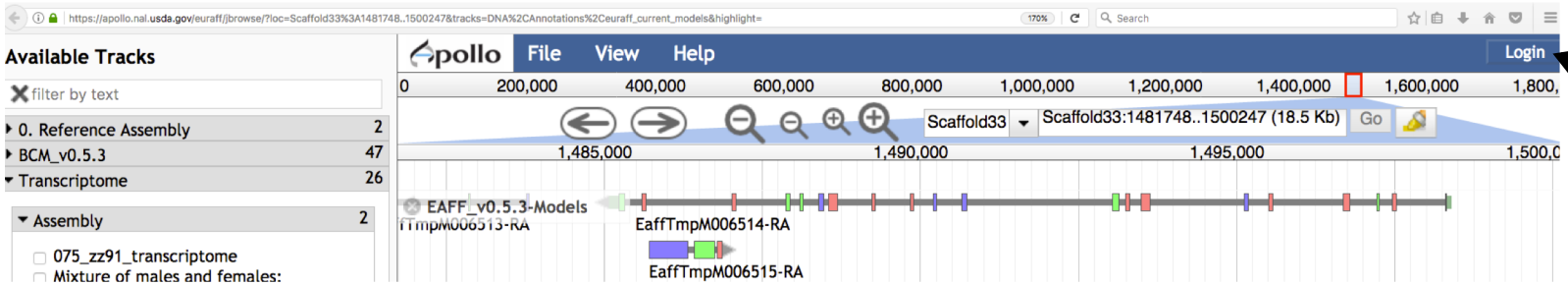
Another approach: BLAST against the genome



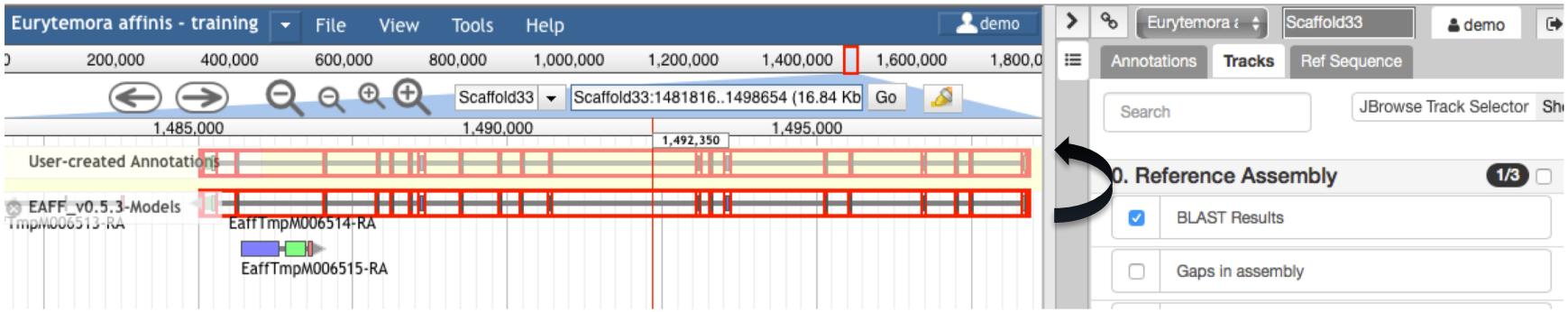
BLAST results are displayed as glyphs in browser; can be used as annotation starting points if the alignment is high quality

USDA

Create annotation in user-created annotations track



Log in with your Apollo credentials



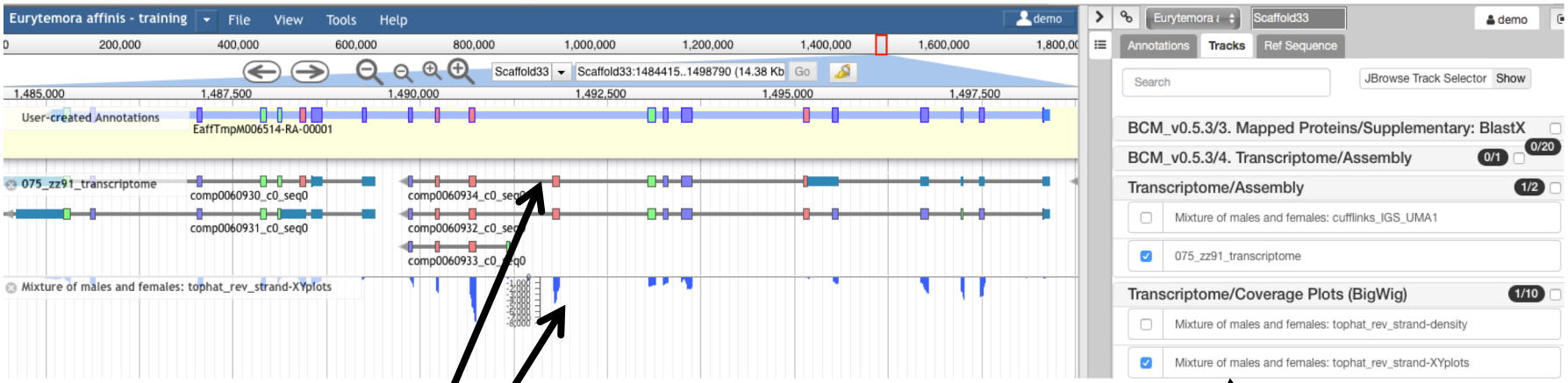
Drag model EaffTmpM006514-RA to User-created Annotations track

USDA

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

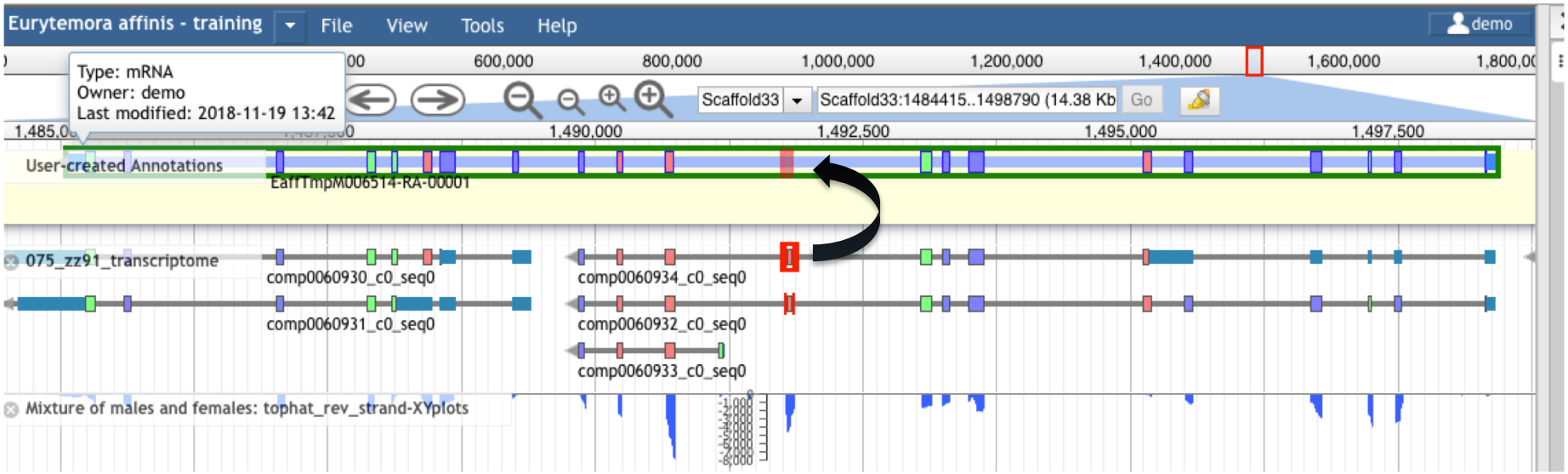


RNA-Seq and transcriptome tracks suggest that one exon is missing

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

USDA

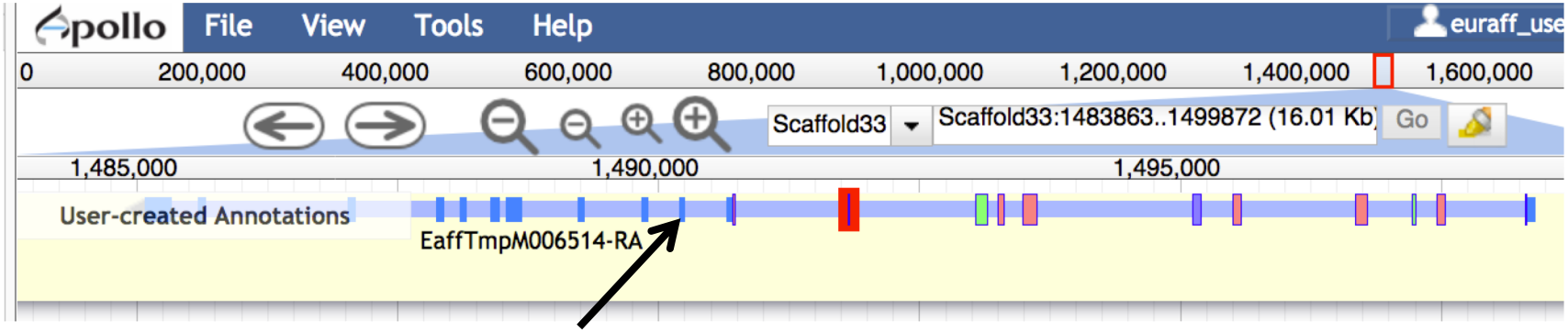
Add an exon to the model



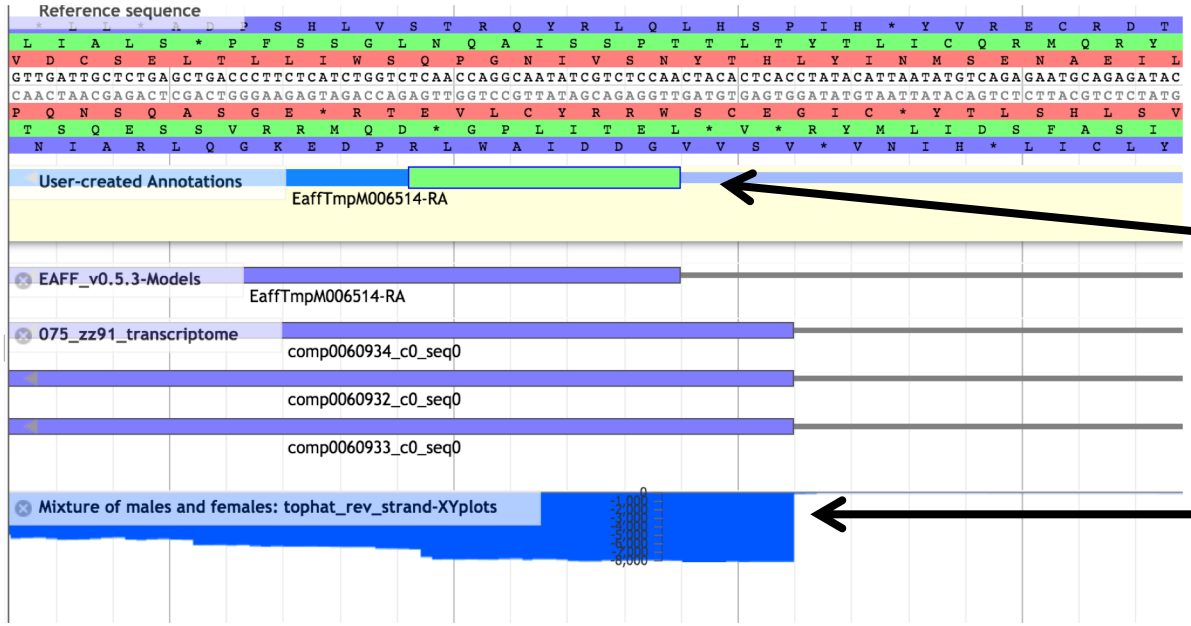
Drag exon from
transcriptome track
into new gene model

USDA

Adjust exon boundary



CDS sequence is now UTR –zoom in to investigate

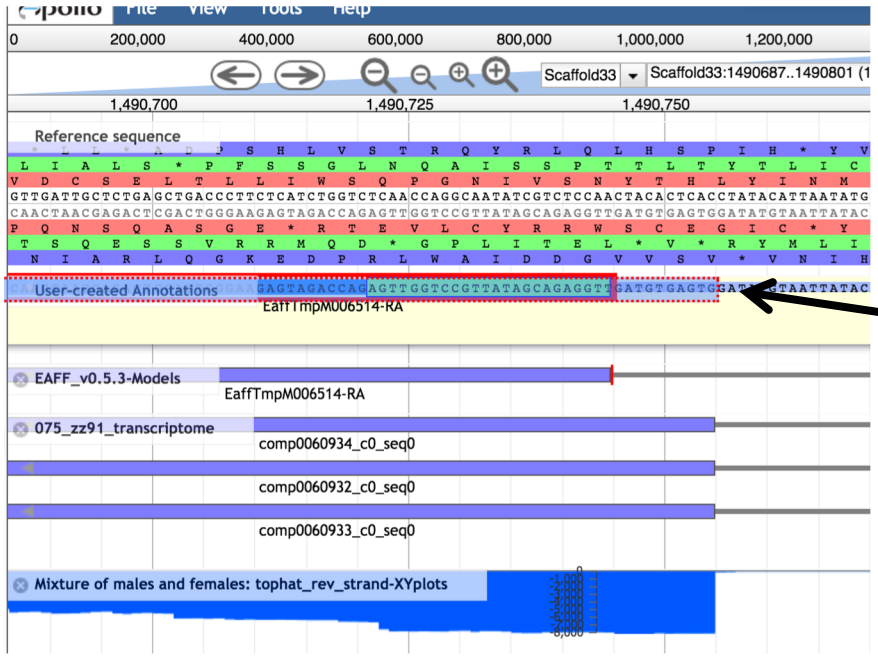


CDS frame has changed from purple to green– we need to fix this

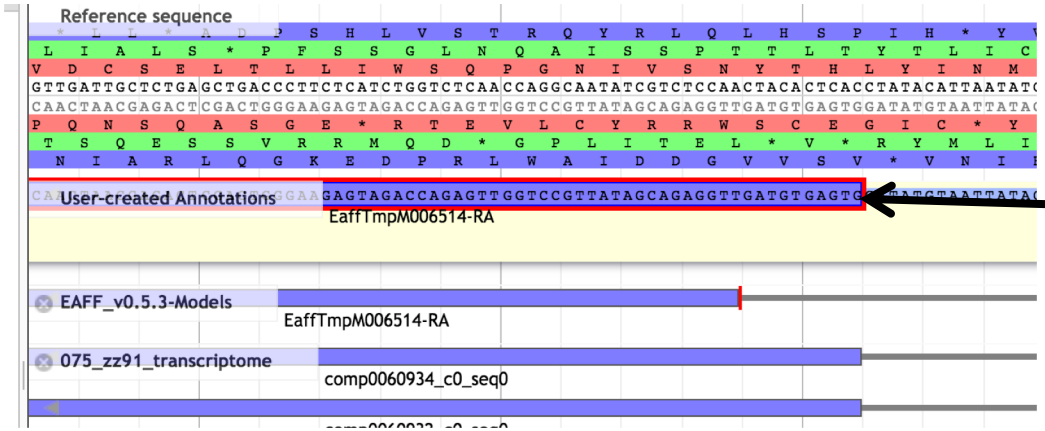
RNA-Seq suggests we need to adjust exon boundary

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

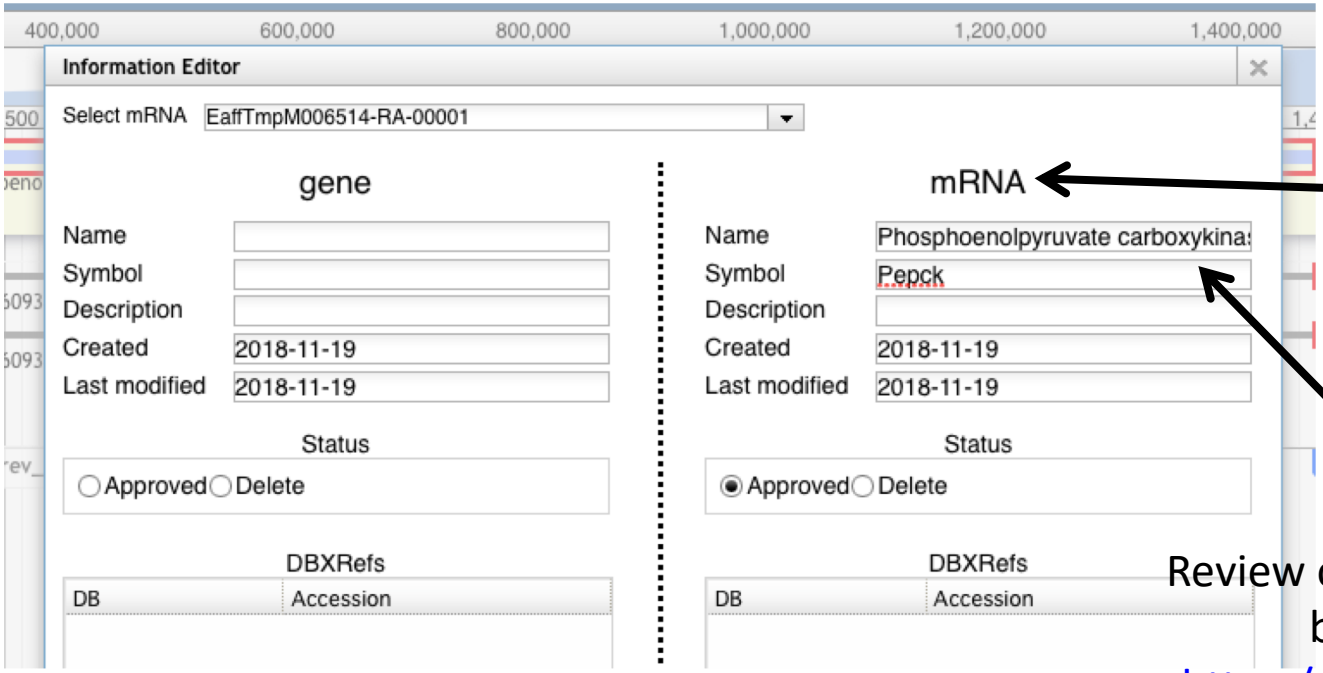
- Blast modified EAFF006514-PA sequence to NCBI's nr database
  - Make sure it doesn't match a potential contaminant
  - Get an idea whether you have the right sequence
  - Blastp home:
    - https://blast.ncbi.nlm.nih.gov/Blast.cgi?PROGRAM=blastp&PAGE_TYPE=BlastSearch&LINK_LOC=blasthome
- Once contamination 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

Using the Information Editor



Use the mRNA/transcript side of the IE

Review our naming guidelines before naming: https://i5k.nal.usda.gov/i5k-workspace-gene-and-protein-naming-guidelines

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 INSDC naming guidelines:
    - http://www.uniprot.org/docs/nameprot
    - If a naming convention exists, use it (e.g. for gene families)
    - Name should be unique and attributed to all orthologs (as far as possible)
    - Use name from an orthologous protein if you are sure that your gene model is an ortholog.
    - Document your justification for the name in the Comments field (e.g. "88% sequence similarity via blastp to D. melanogaster pepck P20007")
  - 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

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.
  - 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

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 may 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

Additional training if you have time:

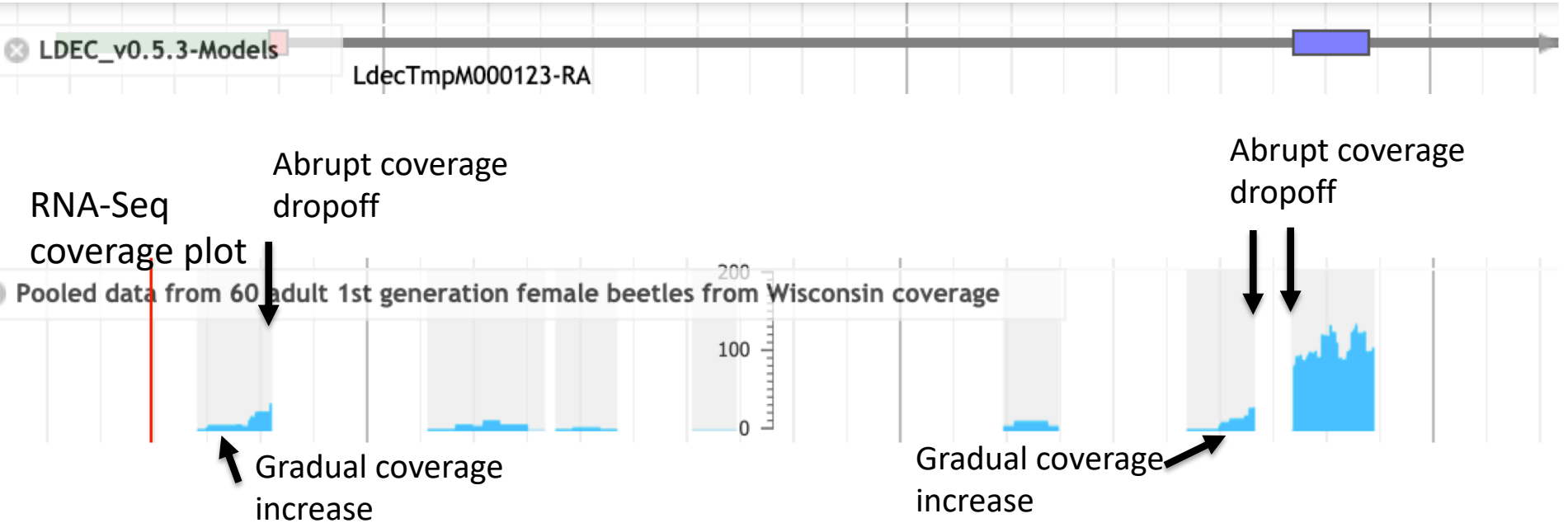
ISOFORM ANNOTATION EXAMPLE

Isoform annotation example

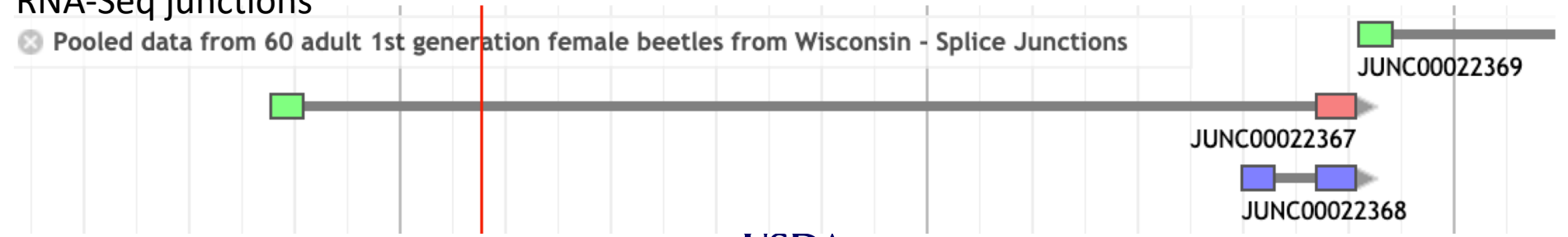
- Here, I'll show you an example annotation of tyrosine protein kinase isoforms in the Colorado Potato Beetle, *Leptinotarsa decemlineata.*
- URL: https://apollo.nal.usda.gov/lepdec_training/jbrowse/?loc=Scaffold2%3A4014701..4033650
- Login: demo/demo

Isoform annotation example

5’ end of MAKER tyrosine
protein kinase gene prediction

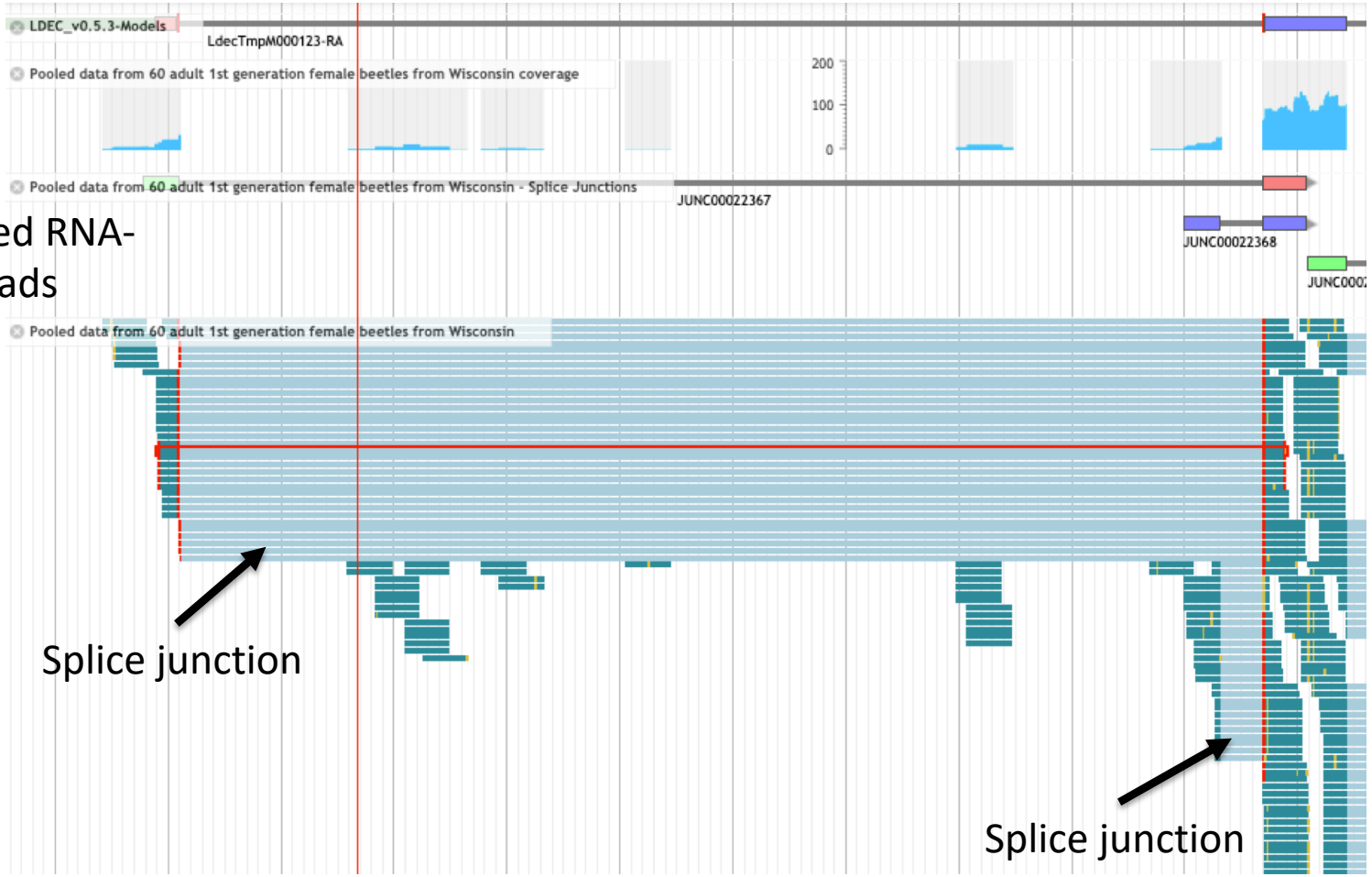


RNA-Seq junctions



Isoform annotation example

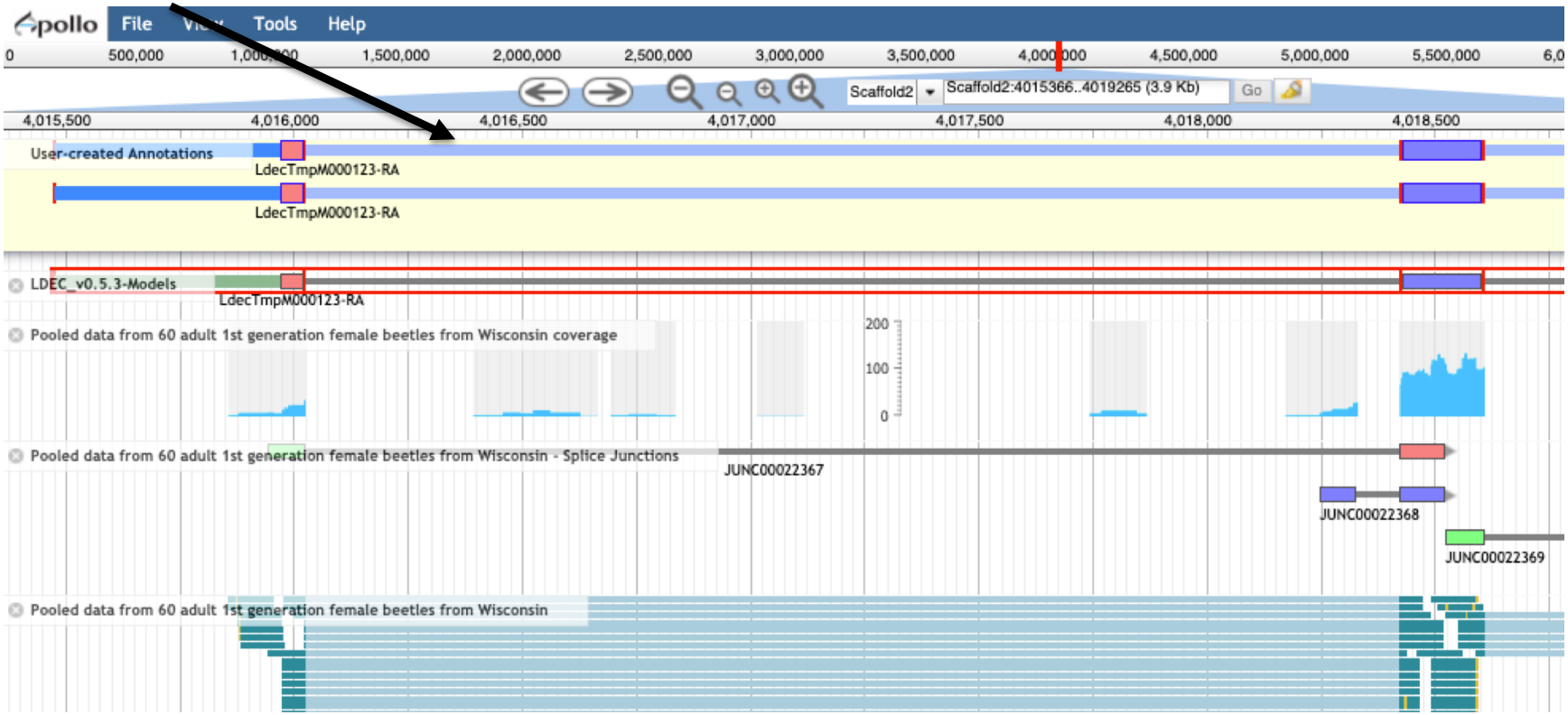
Mapped RNA-Seq reads



USDA

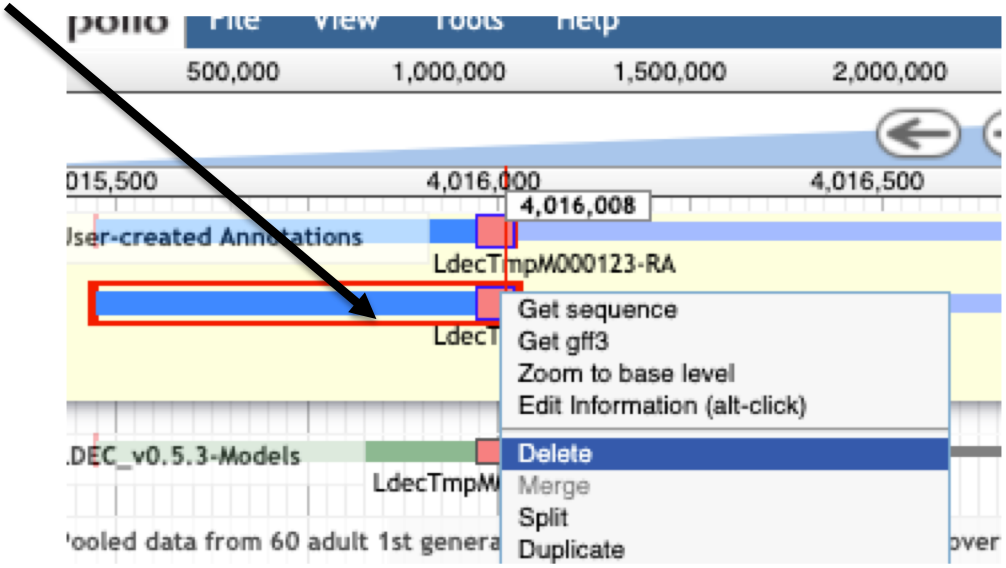
Isoform annotation example

Create 2 isoforms
from Maker model



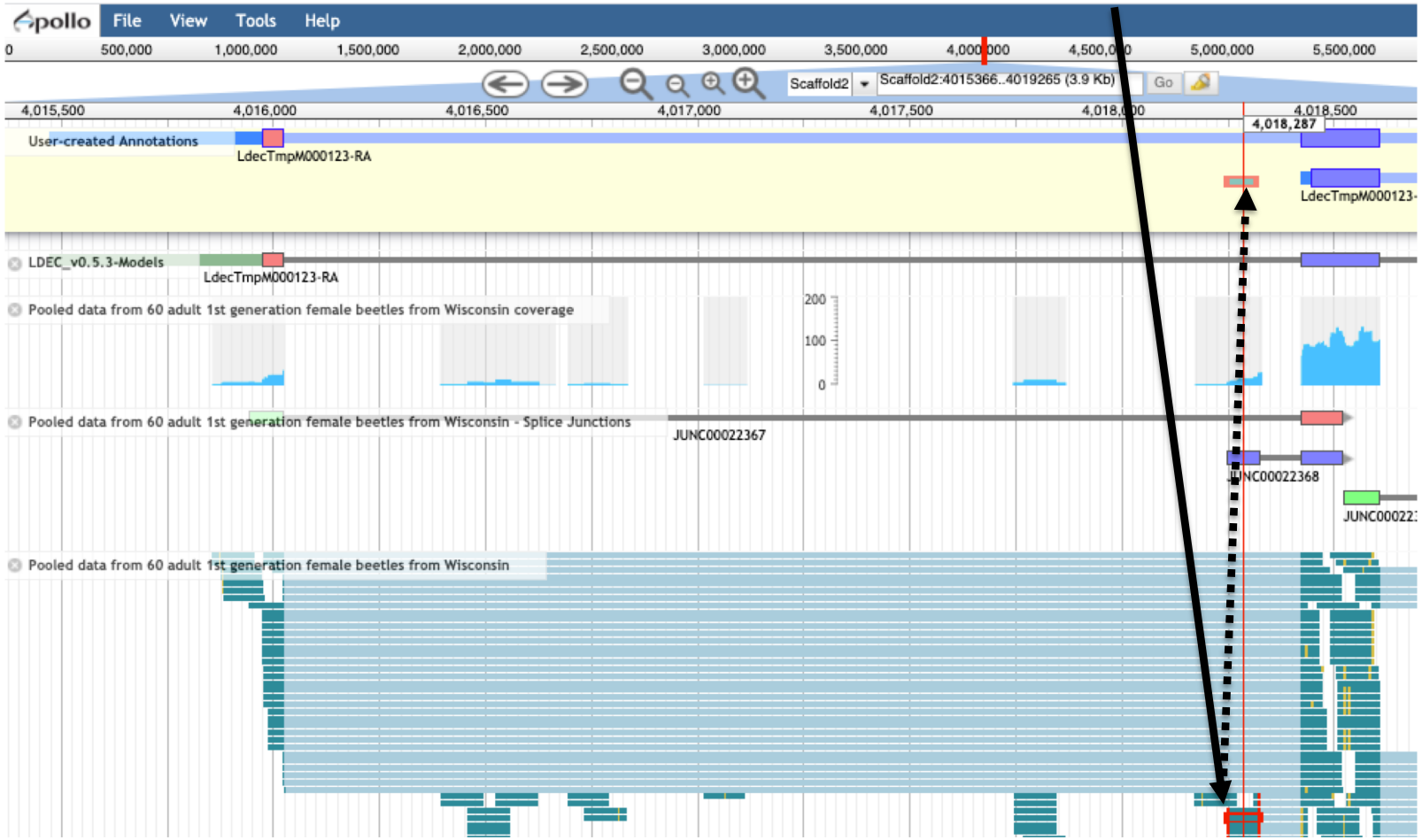
Isoform annotation example

Select and delete 5’ exon from one of the isoforms



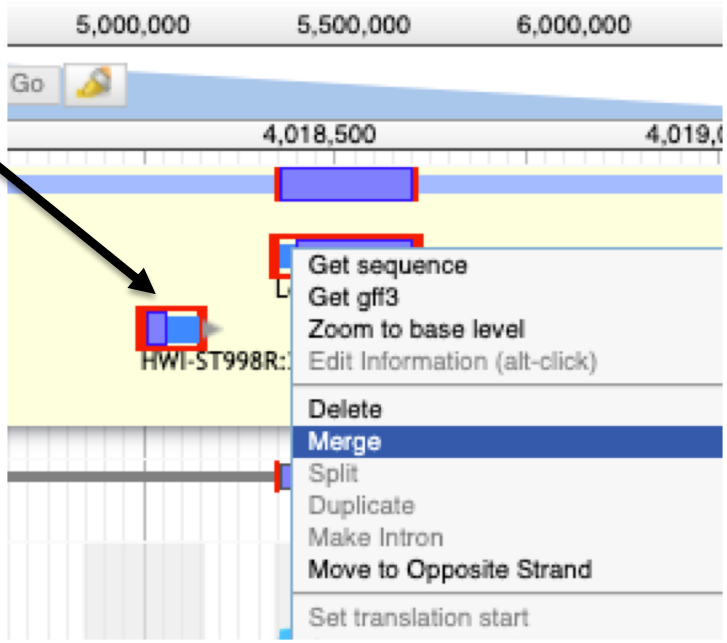
Isoform annotation example

Add a new 5’ exon from mapped RNA-Seq evidence



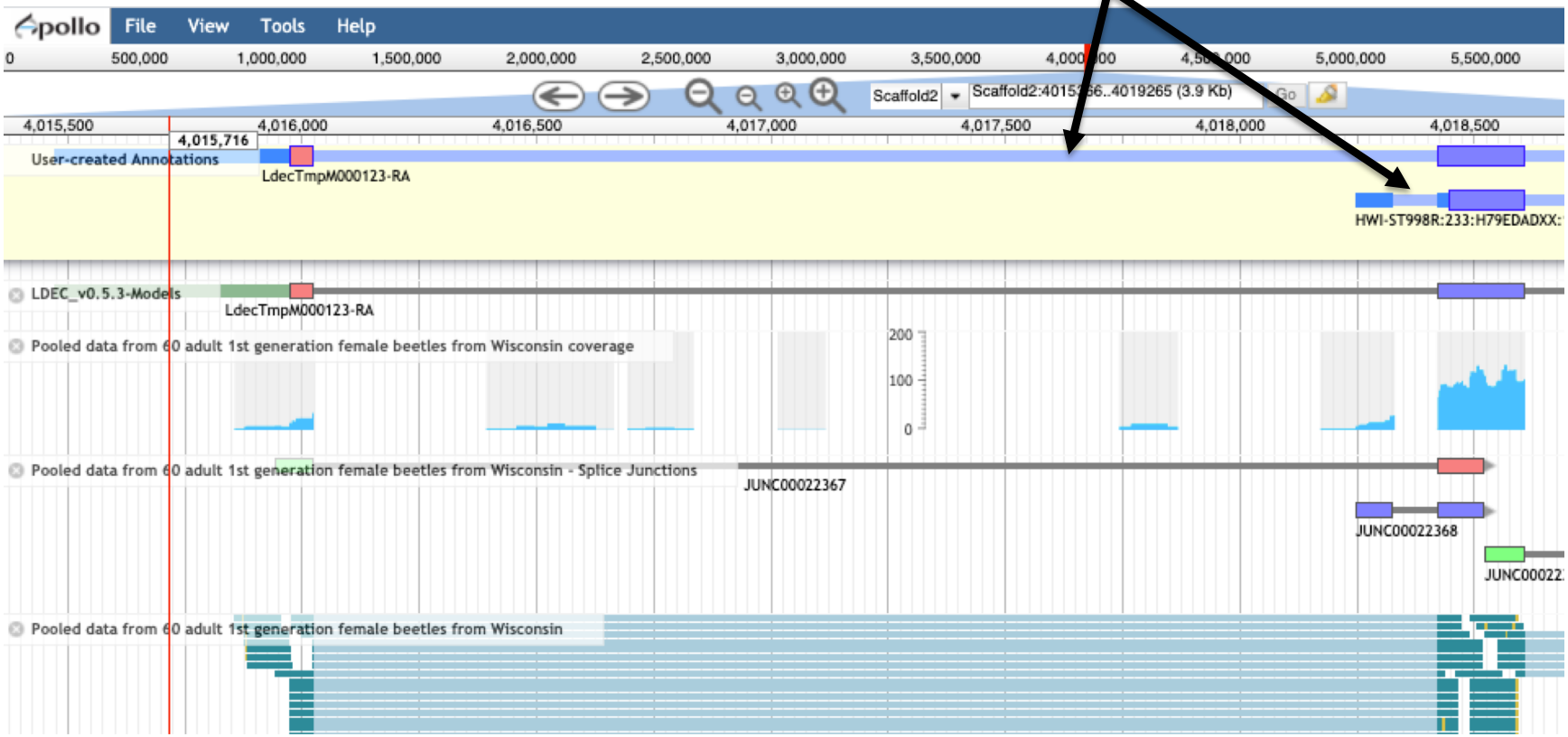
Isoform annotation example

Merge the new 5’ exon with the rest of the model



Isoform annotation example

2 isoforms supported by RNA-Seq evidence

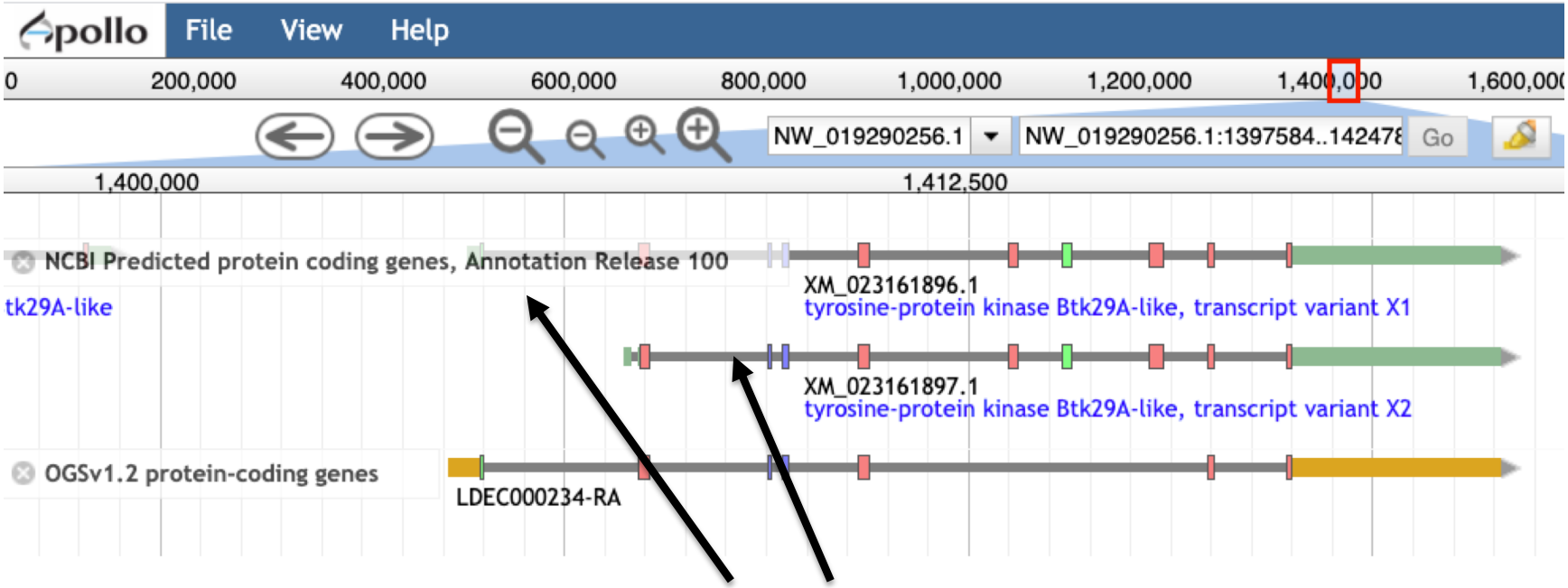


USDA

Isoform annotation example

- In our experience, lots of mapped RNA-Seq reads are critical for good manual isoform annotation
- Before evaluating RNA-Seq for isoforms, it helps to understand how to interpret gradual and abrupt drops in coverage
  - Gradual – usually means 5’ start or 3’ end of expression
  - Abrupt – usually means splice junction
- 5’ splice variants – in RNA-Seq, check for sequential mapped blocks with gradual coverage build-up and abrupt drop – usually means multiple 5’ exons
- Checking junction reads (if available) helps as well
- NCBI’s RefSeq annotation tends to predict lots of isoforms

Isoform annotation example



NCBI predicted these isoforms
in an updated assembly

https://apollo.nal.usda.gov/apollo/2494545/jbrowse/index.html?loc=NW_019290256.1%3A1402381..1422071&tracks=DNA%2CAnnotations%2CNCBI_Annotation_Release_100_Protein_Coding%2Clepdec_current_models&highlight=

USDA

Thank you!

The NAL Team

- Chris Childers
- Gary Moore
- Susan McCarthy
- Min-Chen Hsu
- Chun-Hung Lin
- Chia-Tung Wu

I5k Workspace alumni

- Yu-yu Lin
- Chaitanya Gutta
- Li-Mei Chiang
- Yi Hsiao
- Chien-Yueh Lee
- Han Lin
- Jun-Wei Lin
- Vijaya Tsavatapalli
- Mei-Ju Chen
- Chao-I Tuan

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

Contact us:

- https://i5k.nal.usda.gov/contact
- i5k@ars.usda.gov
- Monica.Poelchau@usda.gov
- Christopher.Childers@usda.gov

USDA