Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL

September 20th, 2022

USDA

Agenda

- Basic RNA-Seq evaluation
- Basic structural changes- splitting and merging a model, adding and removing exons
- UTRs –when and how to add and adjust
- Changing translation start and stop sites, and open reading frames
- Non-canonical splice sites
- Annotating isoforms

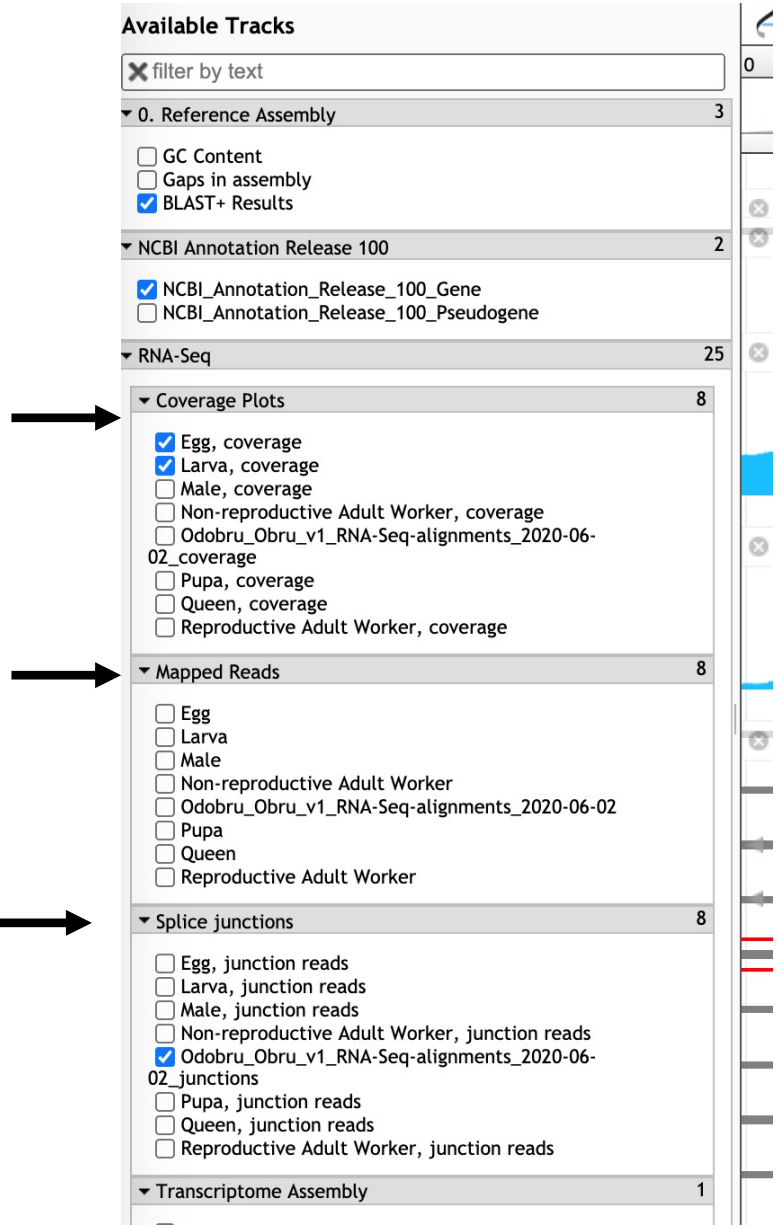
Other resources

- I5k Workspace manual annotation landing page: https://i5k.nal.usda.gov/manual-annotation-and-apollo
- An additional Apollo webinar with a worked example: https://www.youtube.com/watch?v=dol99KExLgY&feature=youtu.be
- 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:
  - https://genomearchitect.readthedocs.io/en/latest/UsersGuide.html
- Other manual curation tutorials: http://genomecuration.github.io/genometrain/d-feature-curation-crossing/
- VEuPathDB Apollo training webinar: https://veupathdb.org/veupathdb/app/static-content/webinars.html#apollo

Basic RNA-Seq evaluation

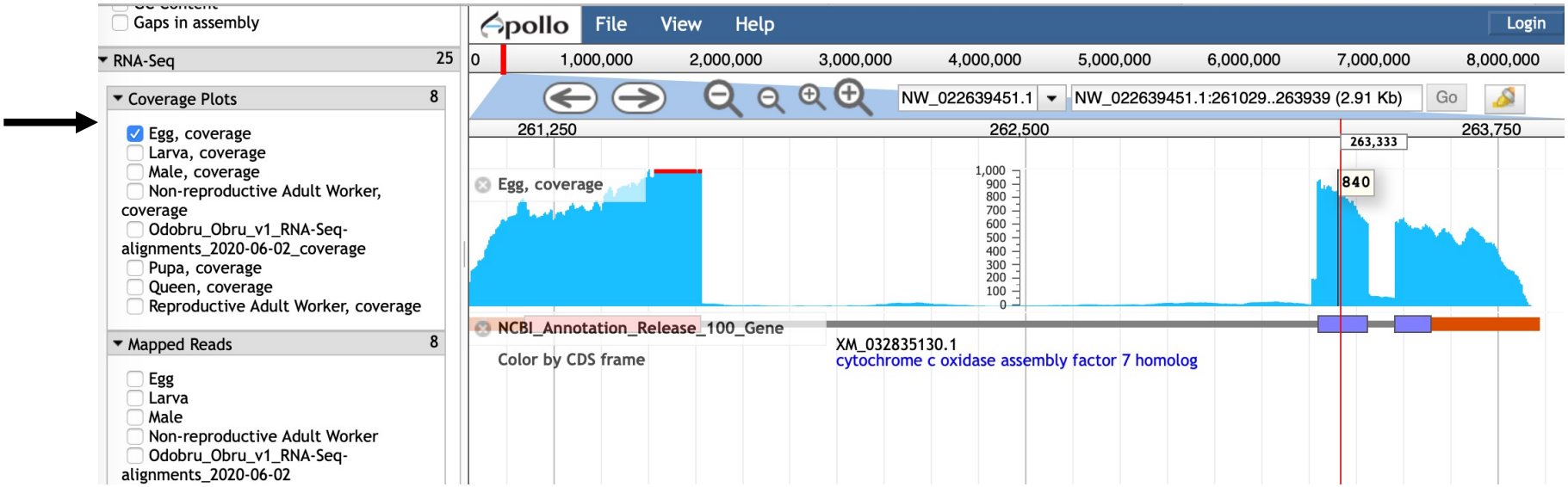
RNA-Seq tracks

- **Coverage plots:** Histogram of the number of mappings at each nucleotide
- **Mapped reads:** Individual glyphs of each mapped read.
- **Junction reads:** Show where mapped reads are spliced.



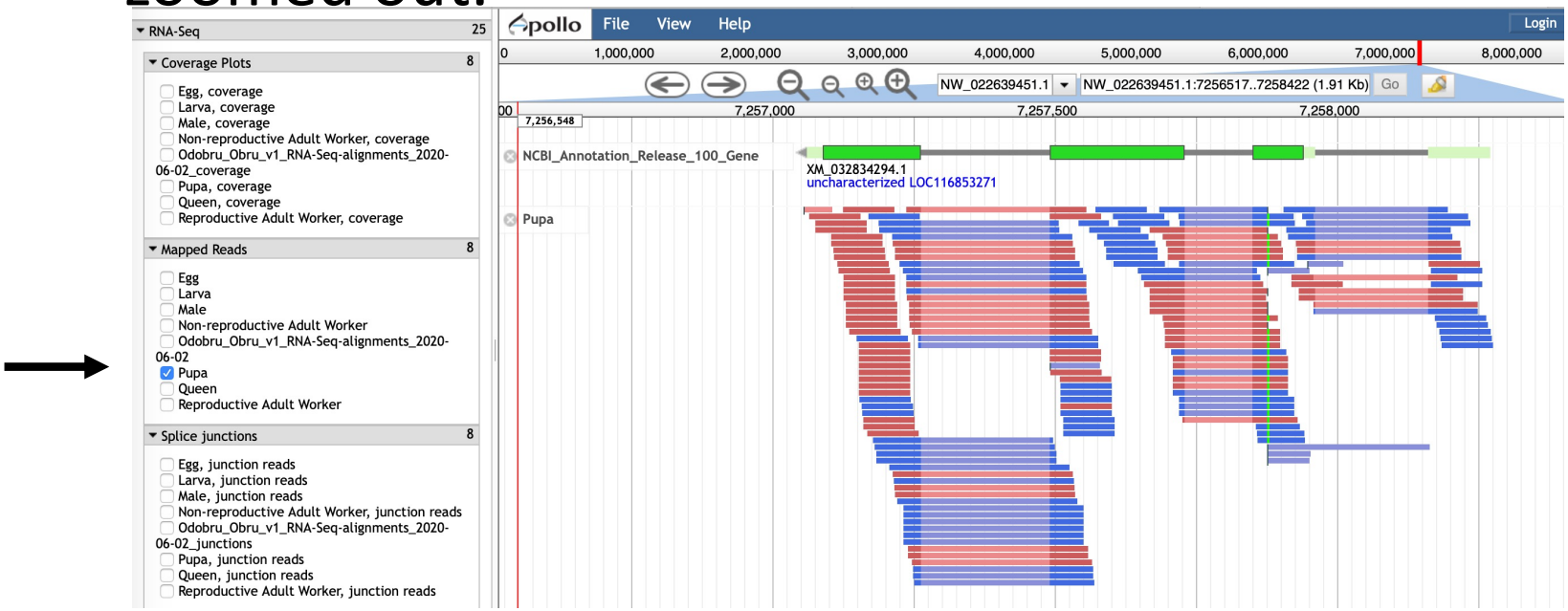
RNA-Seq tracks

- **Coverage plots:** Histogram of the number of mappings at each nucleotide; hover over the blue area to see the value



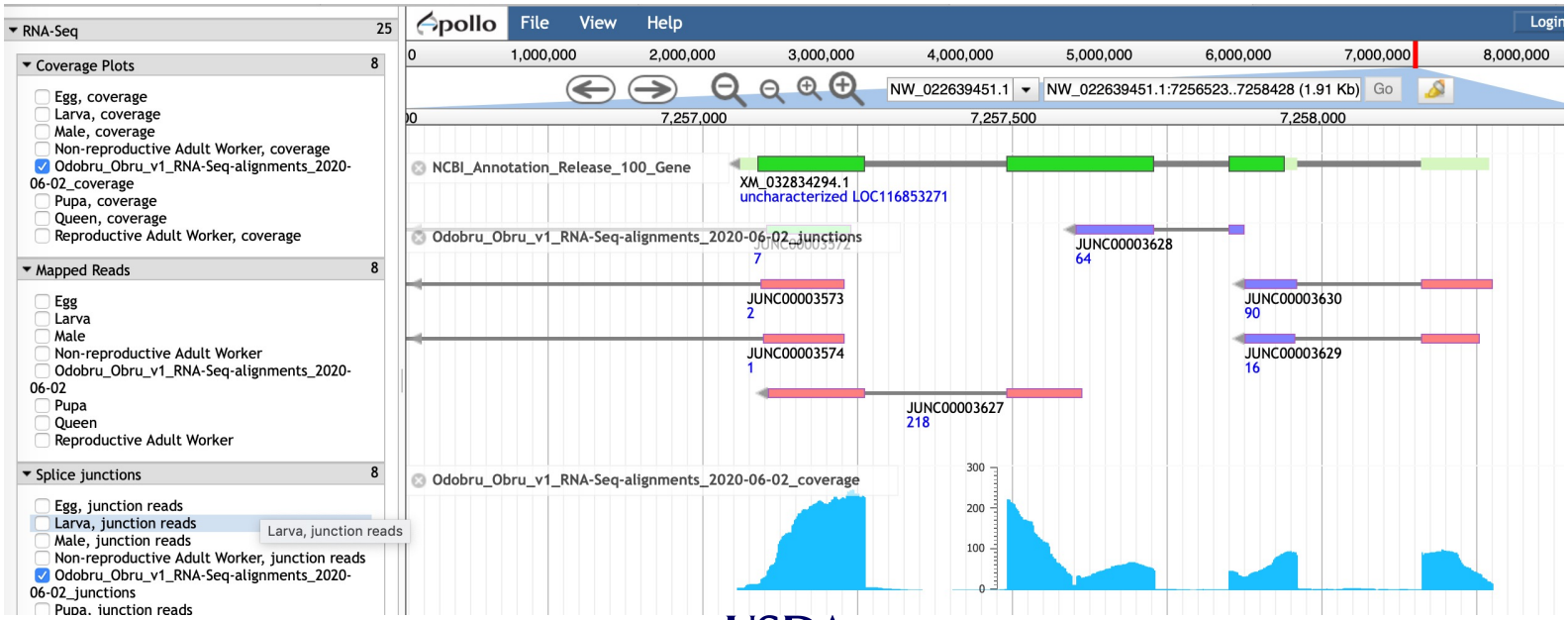
RNA-Seq tracks

- **Mapped reads:** Individual glyphs of each mapped read. Show mapped and spliced areas, and SNPs/indels. Informative, but hard to work with when zoomed out.

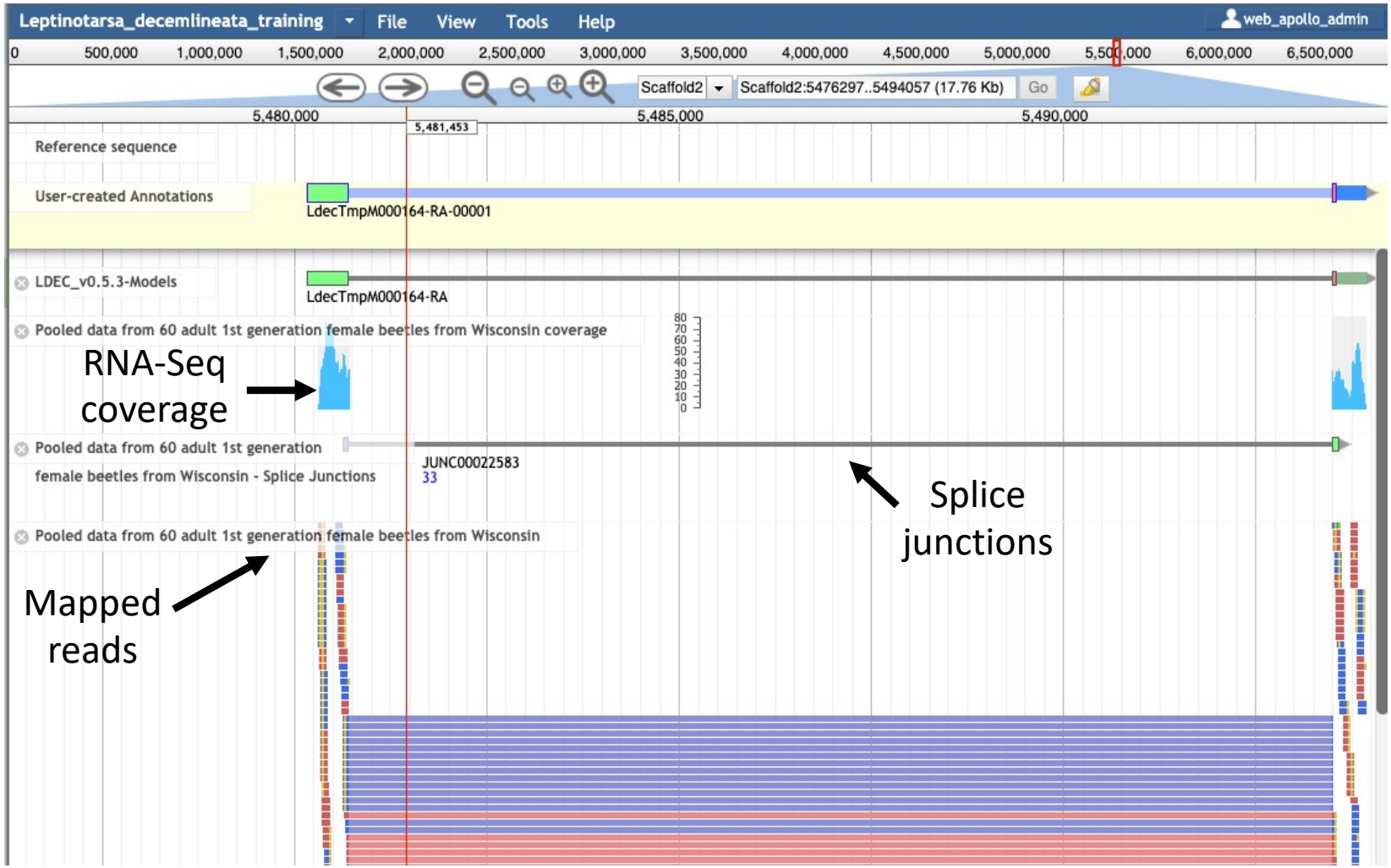


RNA-Seq tracks

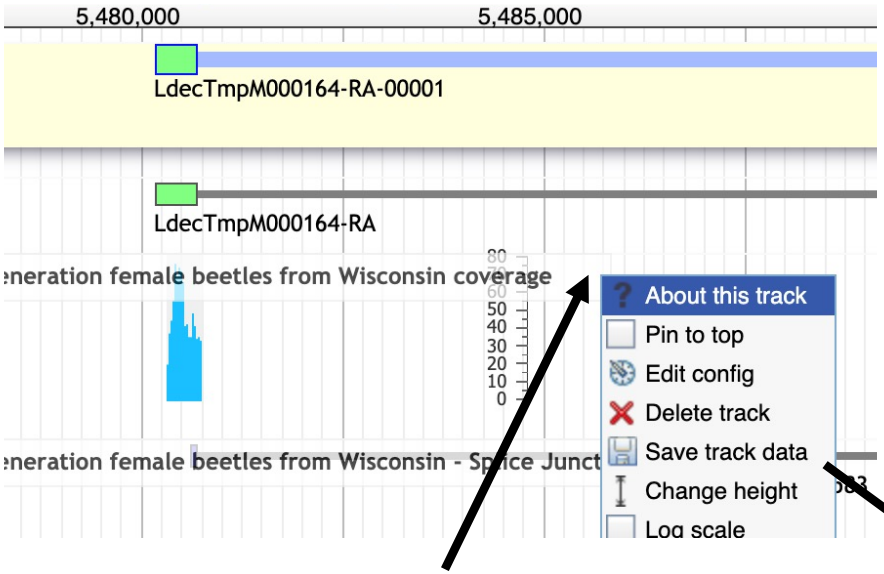
- **Junction reads:** Useful combined with coverage plots; show where mapped reads are spliced. The blue numbers show the ‘score’ – the number of mappings that support the splice junction.



A simple case

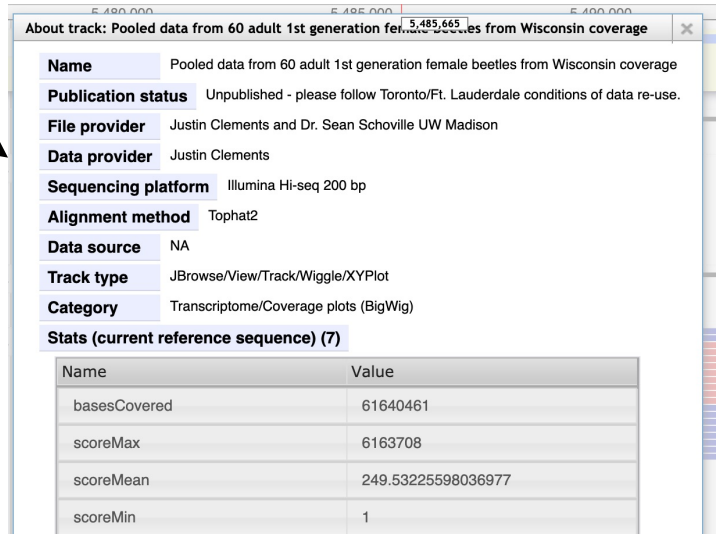


A simple case

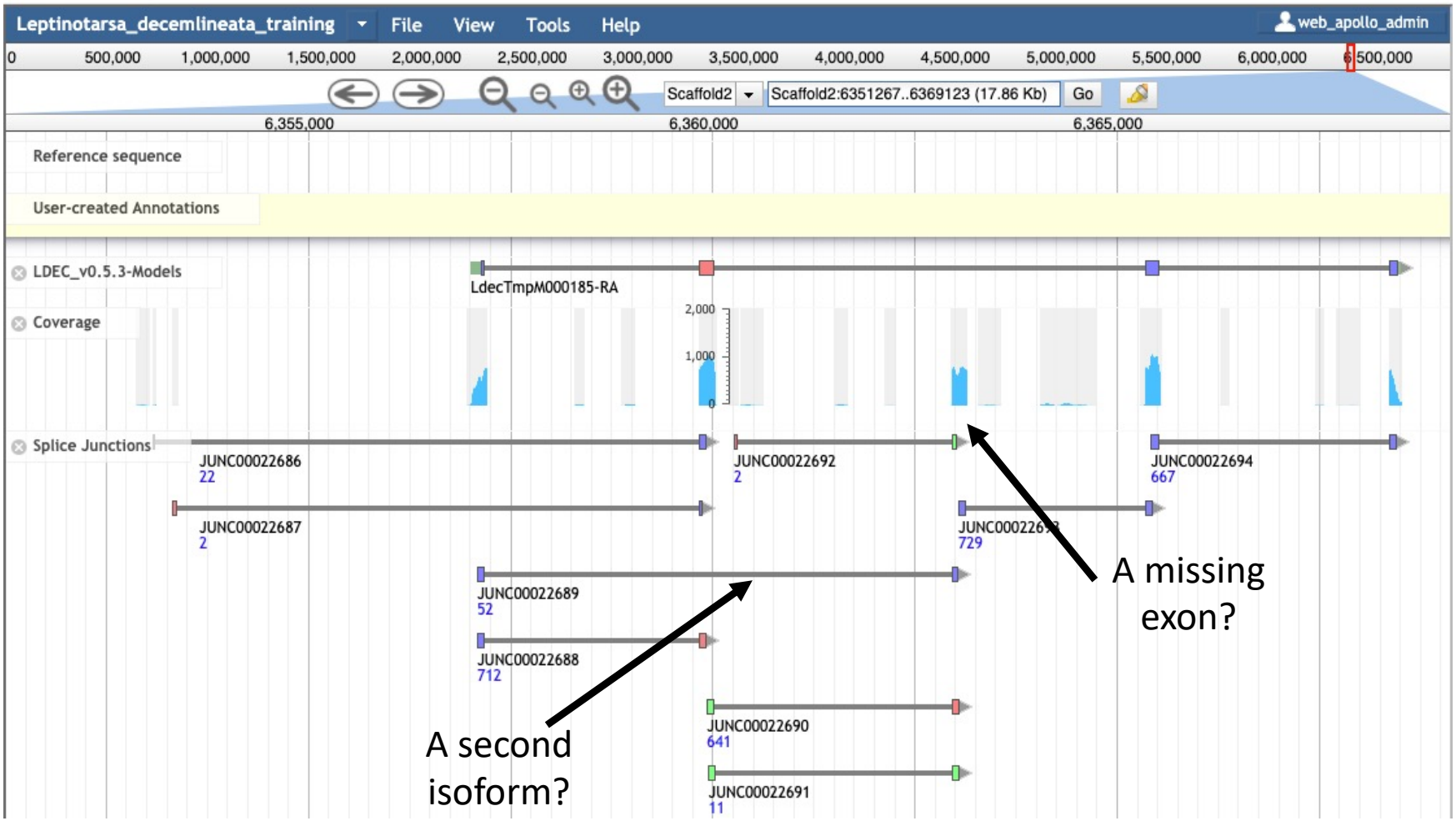


Select ‘About this track’ from drop-down menu

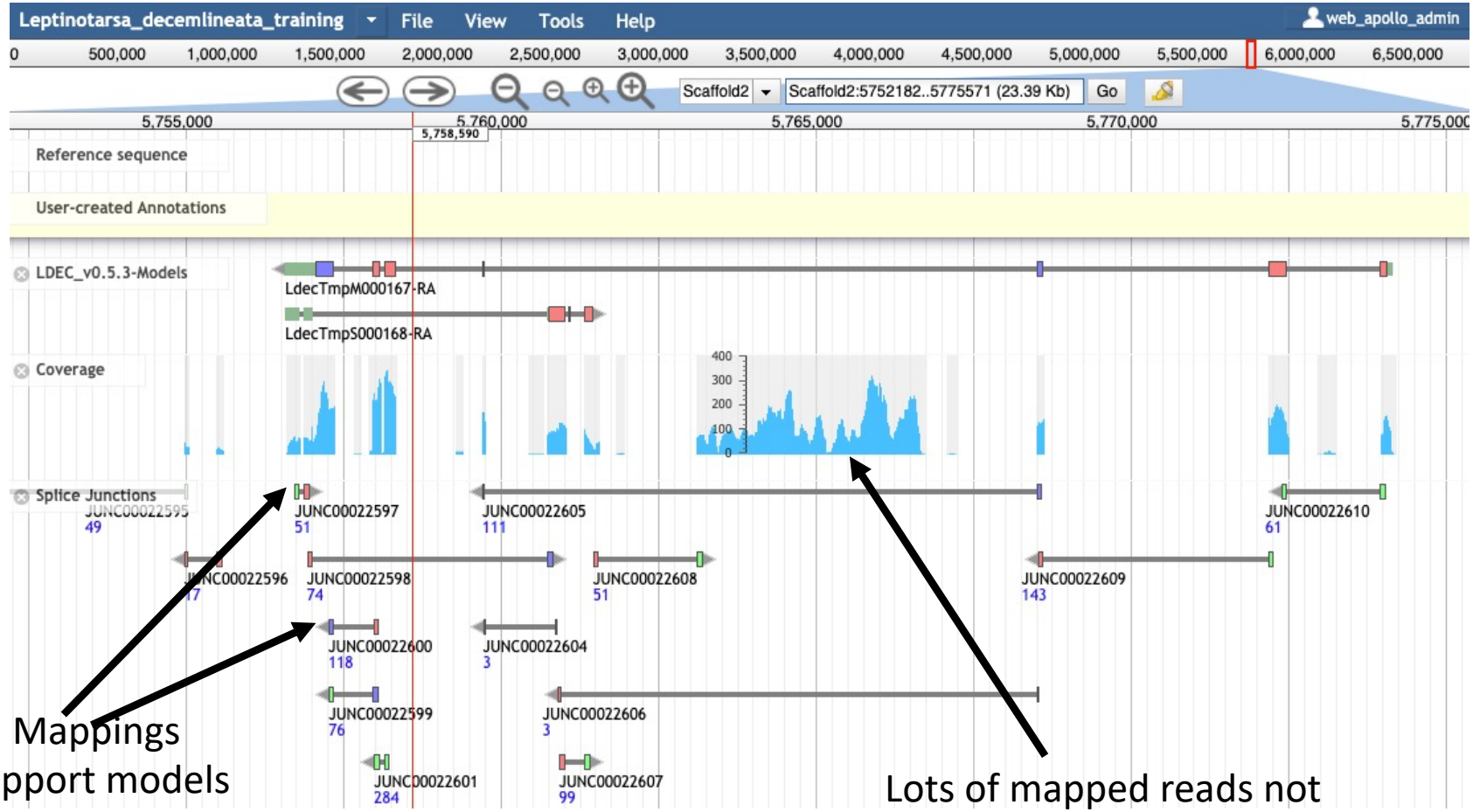
Information about methods



A more complex case



A really messy case

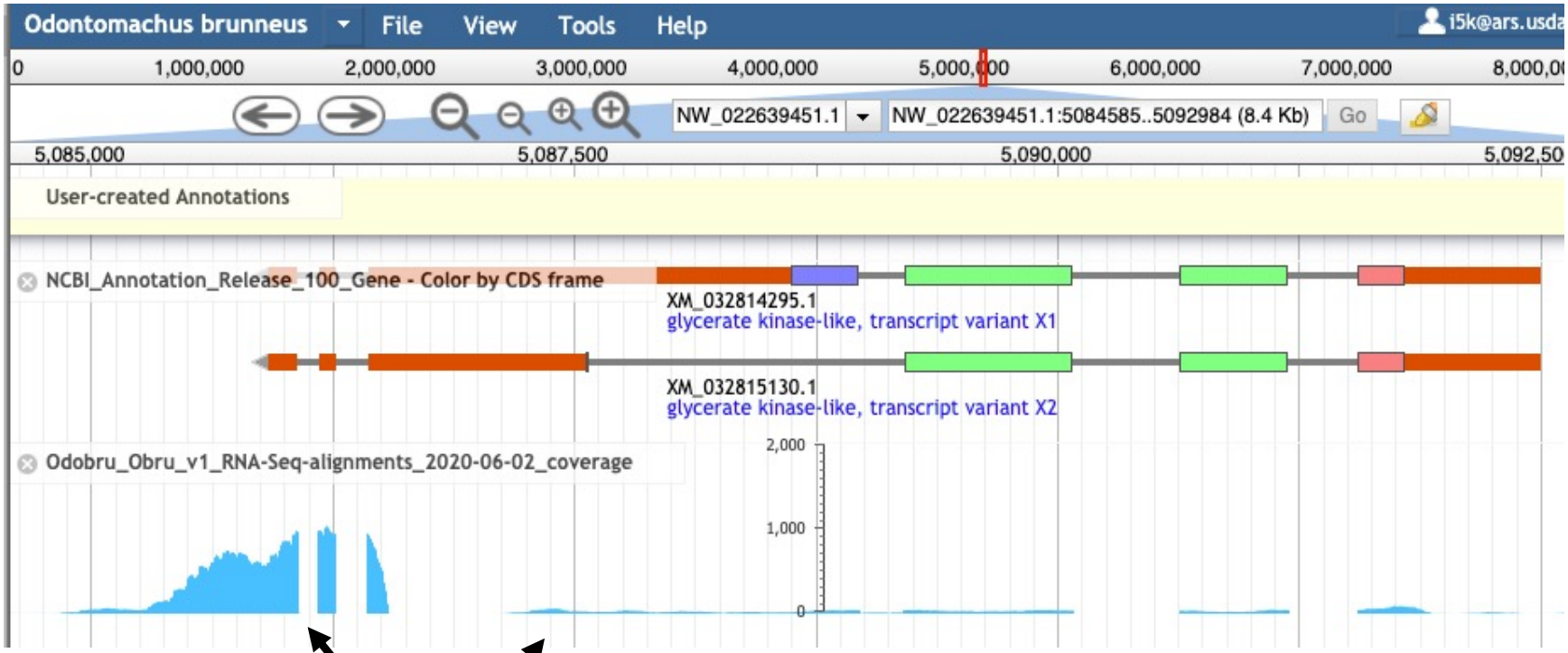


Mappings support models on forward and reverse strands

Lots of mapped reads not reflected in gene model

Basic structural changes – splitting and merging a model, adding and removing exons

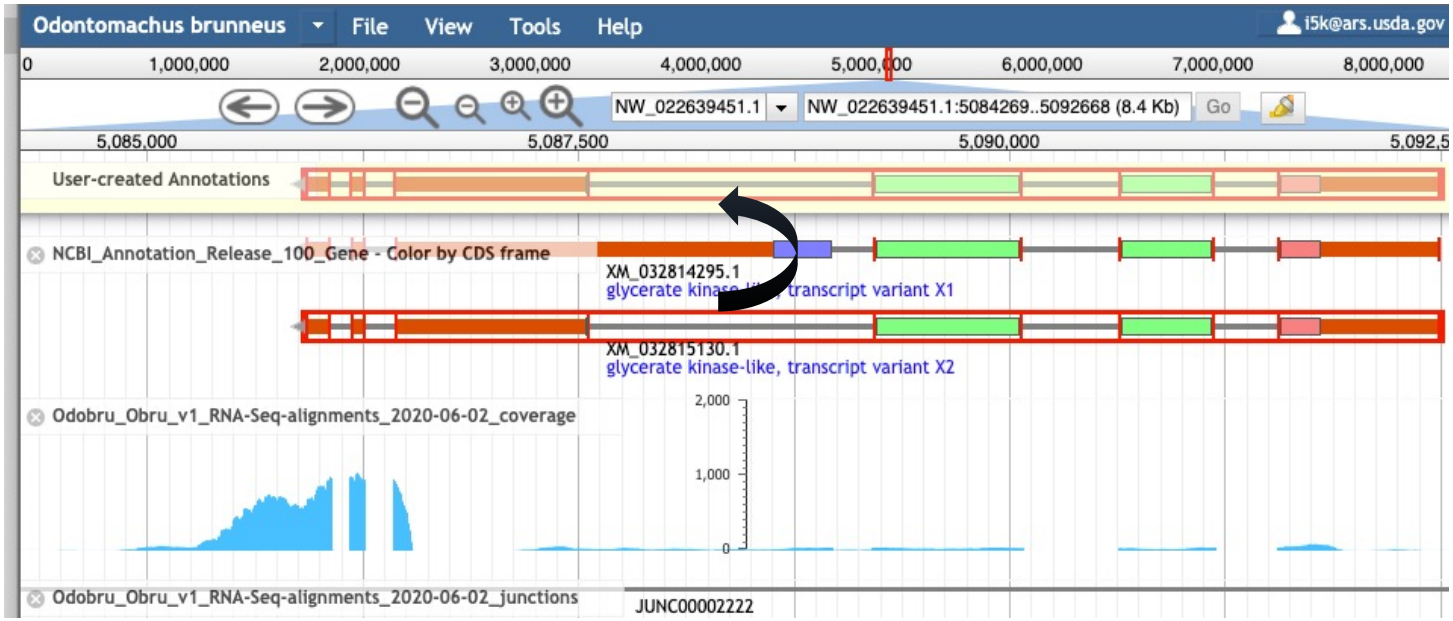
RNA-Seq evaluation



- Very different coverage between UTR and CDS
- No RNA-Seq coverage between high and low expression areas
  - 2 separate models?

Create new model in user-created annotations track

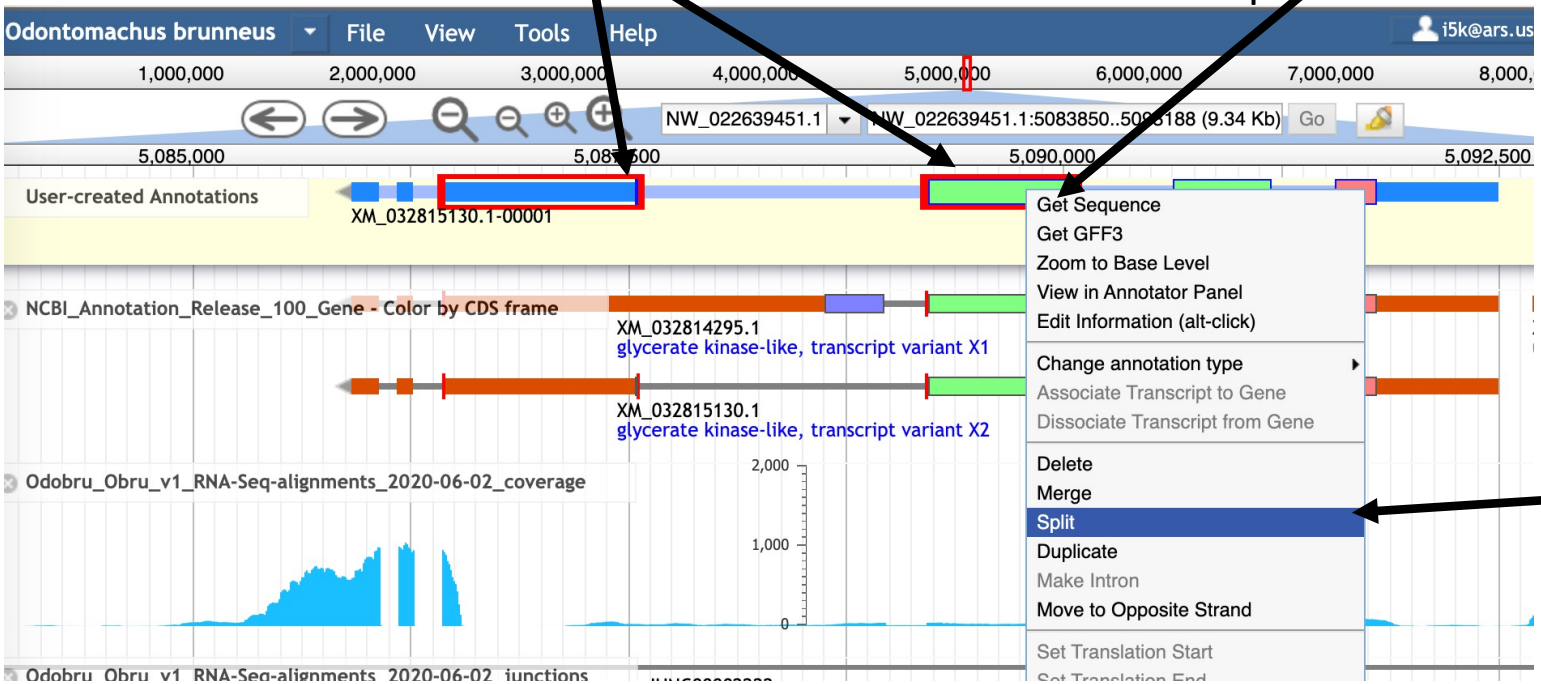
Drag evidence to UcA track (or right-click and select “create annotation”)



Split model

Select exons on which to split the model using the ‘shift’ key

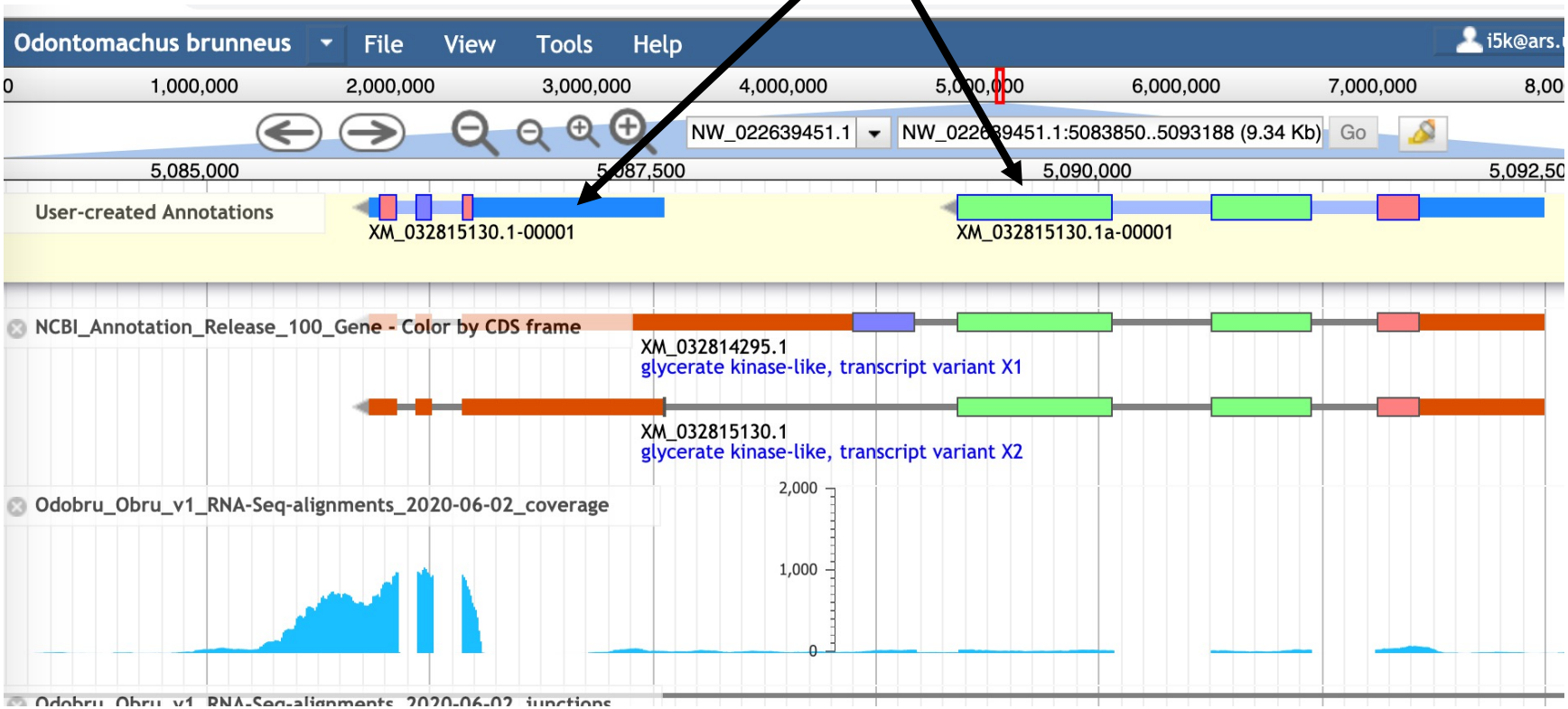
Right-click on the model while continuing to hold shift to get the drop-down menu



Select ‘split’

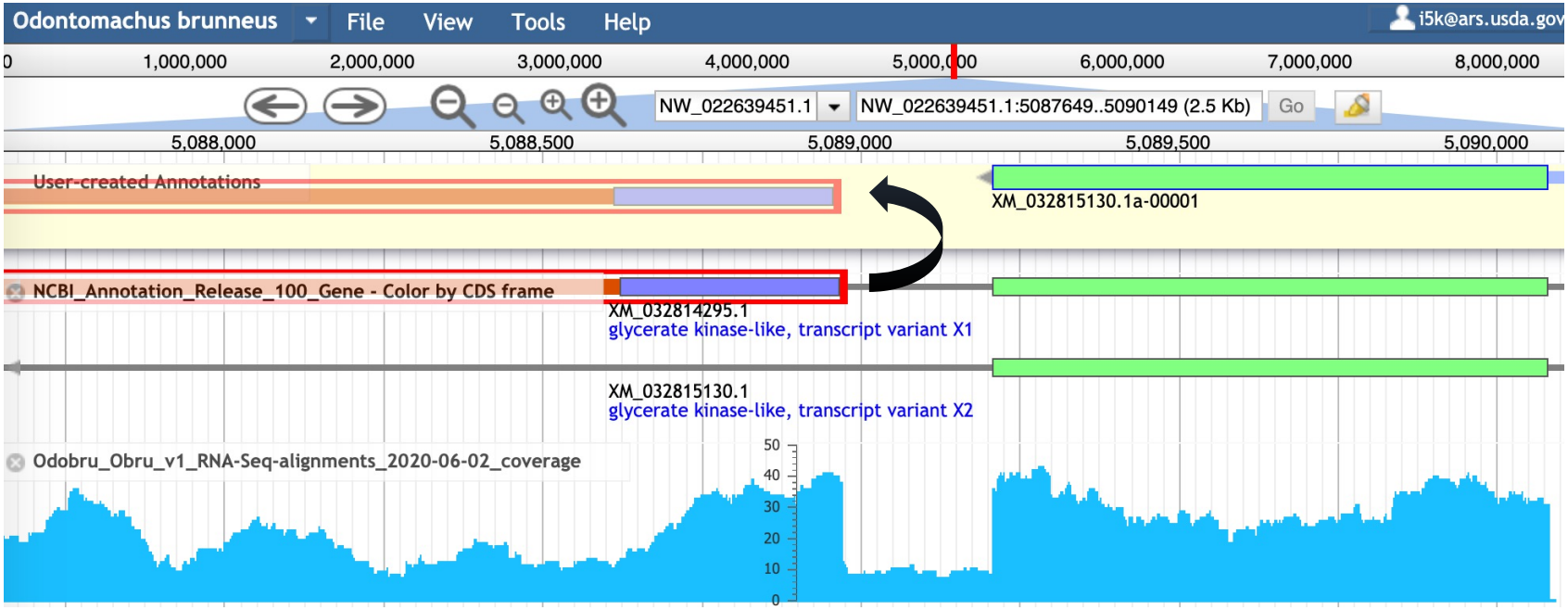
Split model

You now have 2 models! Let's start fixing the model on the right – it needs a 3' exon.



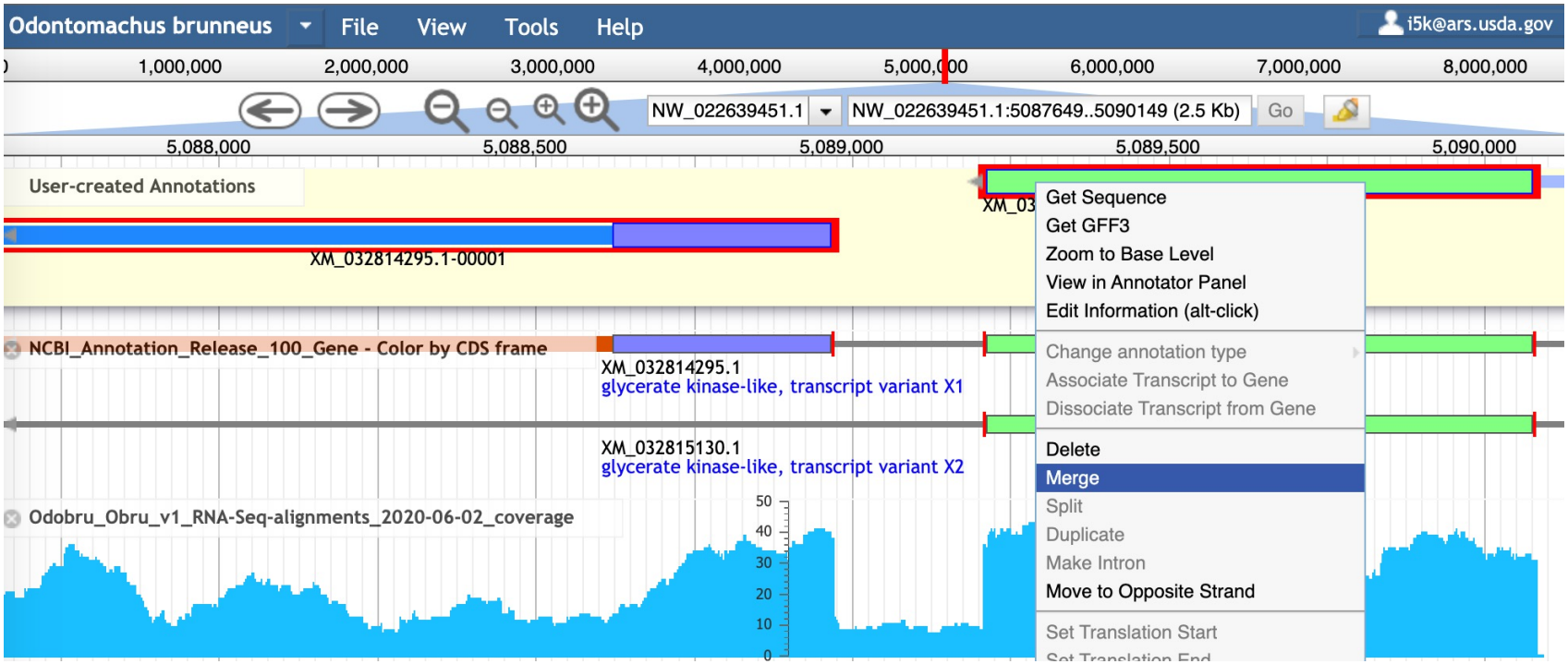
Add an exon

Zoom in, select the missing exon,
drag up to UcA track



Merge exons

Shift-select both exons, shift-right click, then select ‘merge’ from the dropdown menu



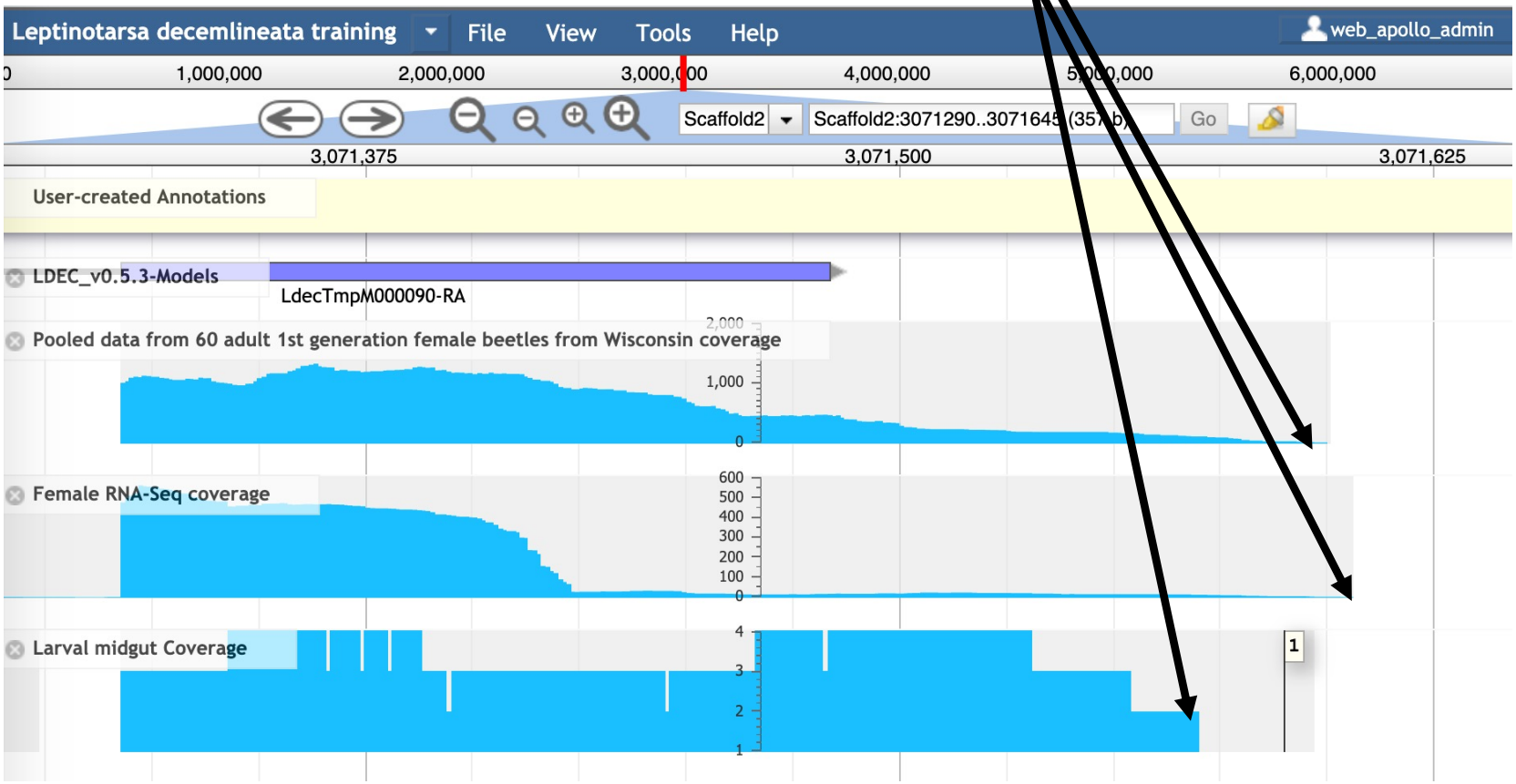
UTRs – how and when to add or adjust

Adding or adjusting UTR boundaries

- When should you add or change UTRs?
  - Only if you have RNA-Seq evidence with sufficient coverage relative to the rest of the model
  - Adding or changing UTRs is helpful, but not necessary if you're only interested in the protein sequence
  - Deciding where the UTR ends is usually a judgement call
- Apollo tools for gene boundary changes:
  - Manual edge-matching to available evidence
  - Automated edge-matching to available evidence

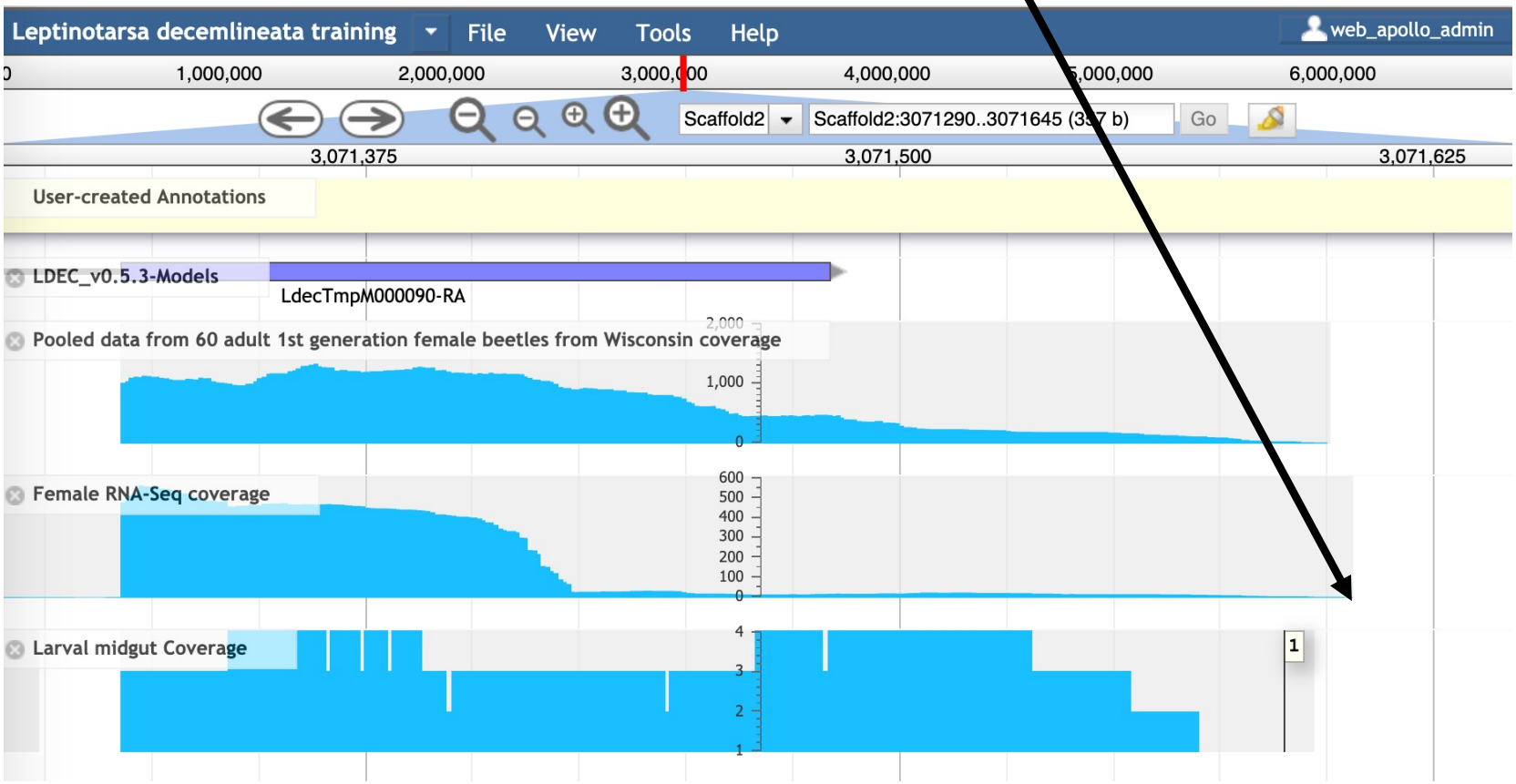
Adjusting gene boundaries

RNA-Seq evidence ends in different places for each track – how do you decide?

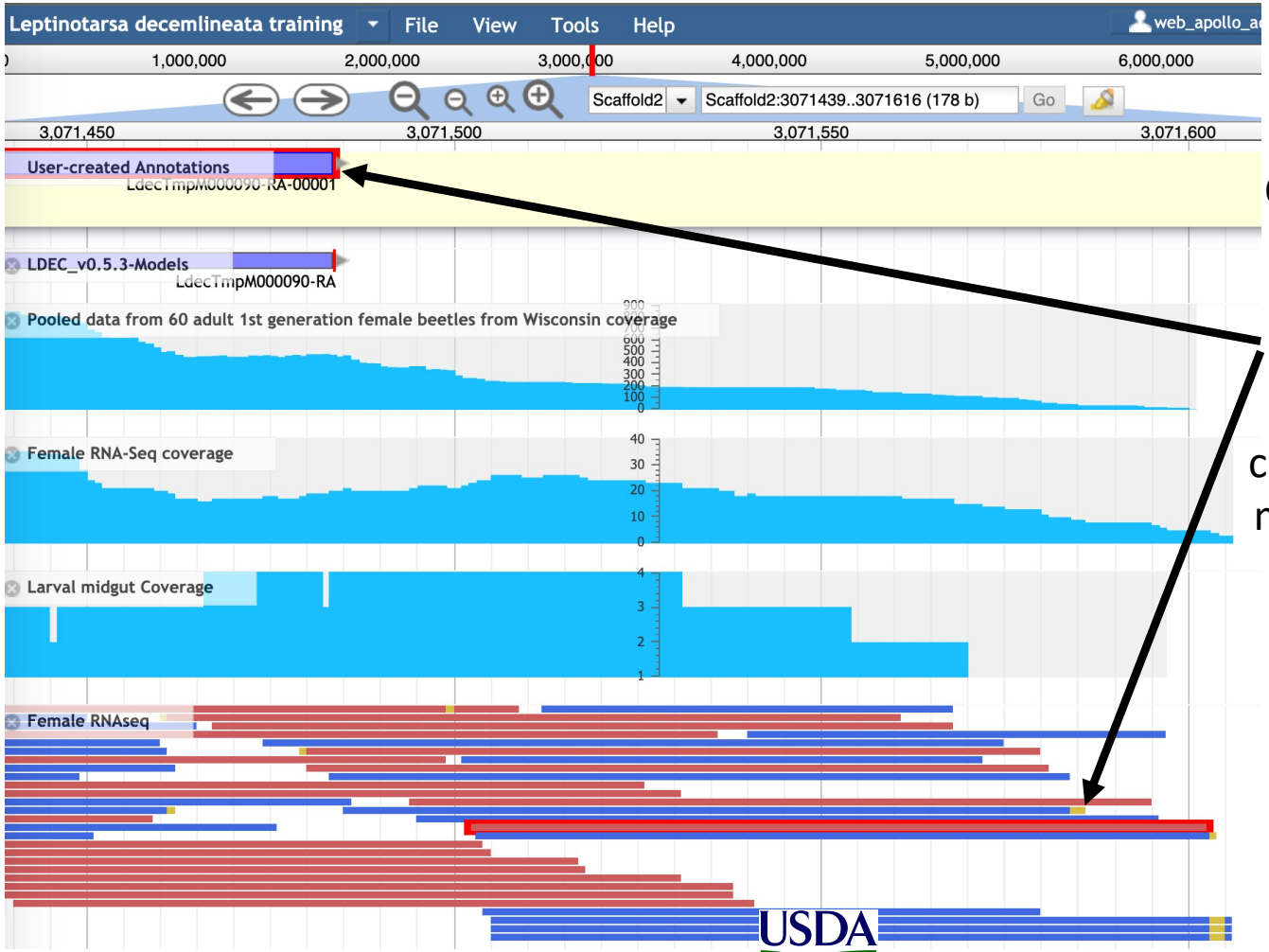


Adjusting gene boundaries

Pick the longest boundary available, and note which track you used in the ‘Comments’ section

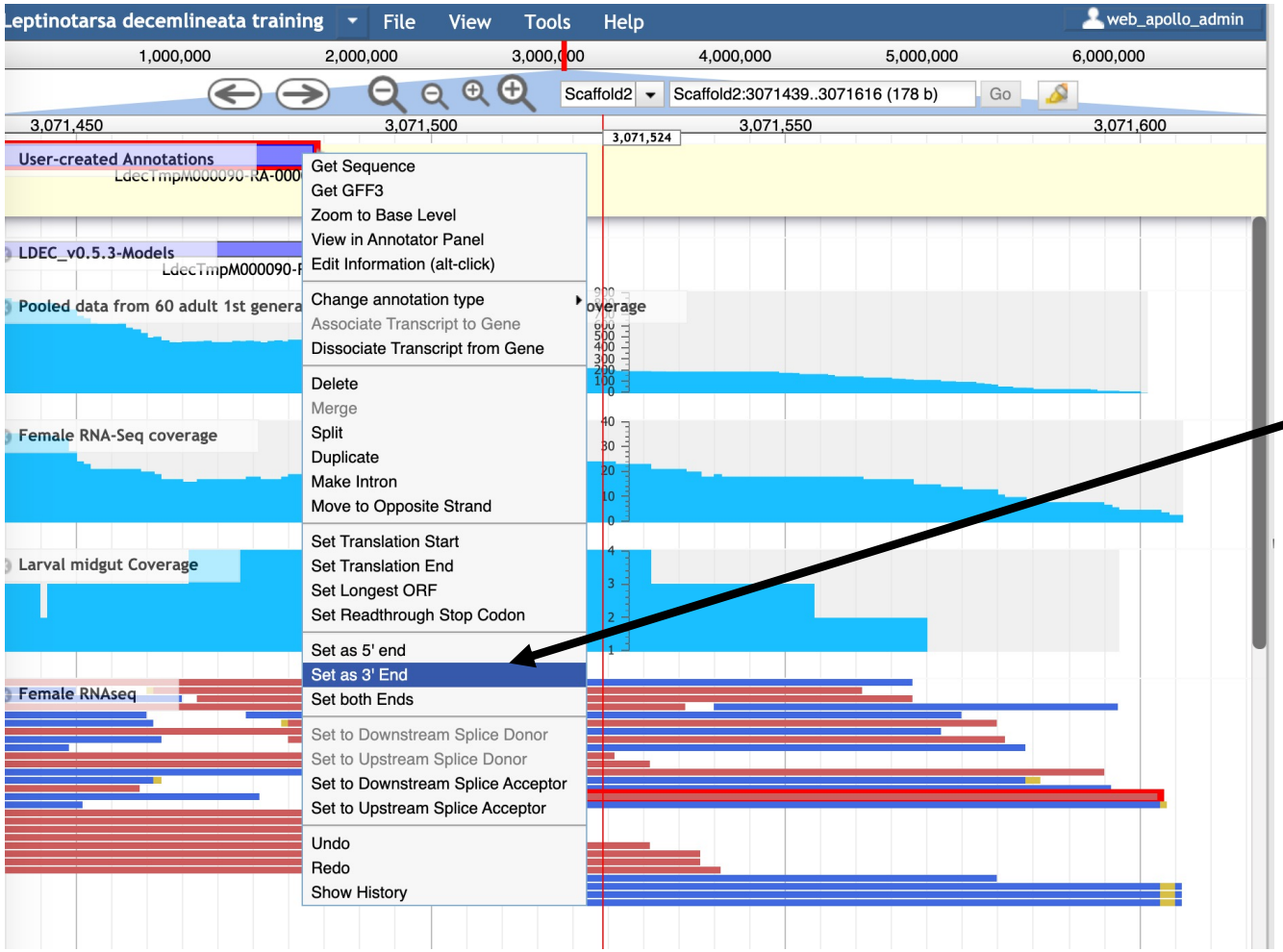


Adjusting gene boundaries



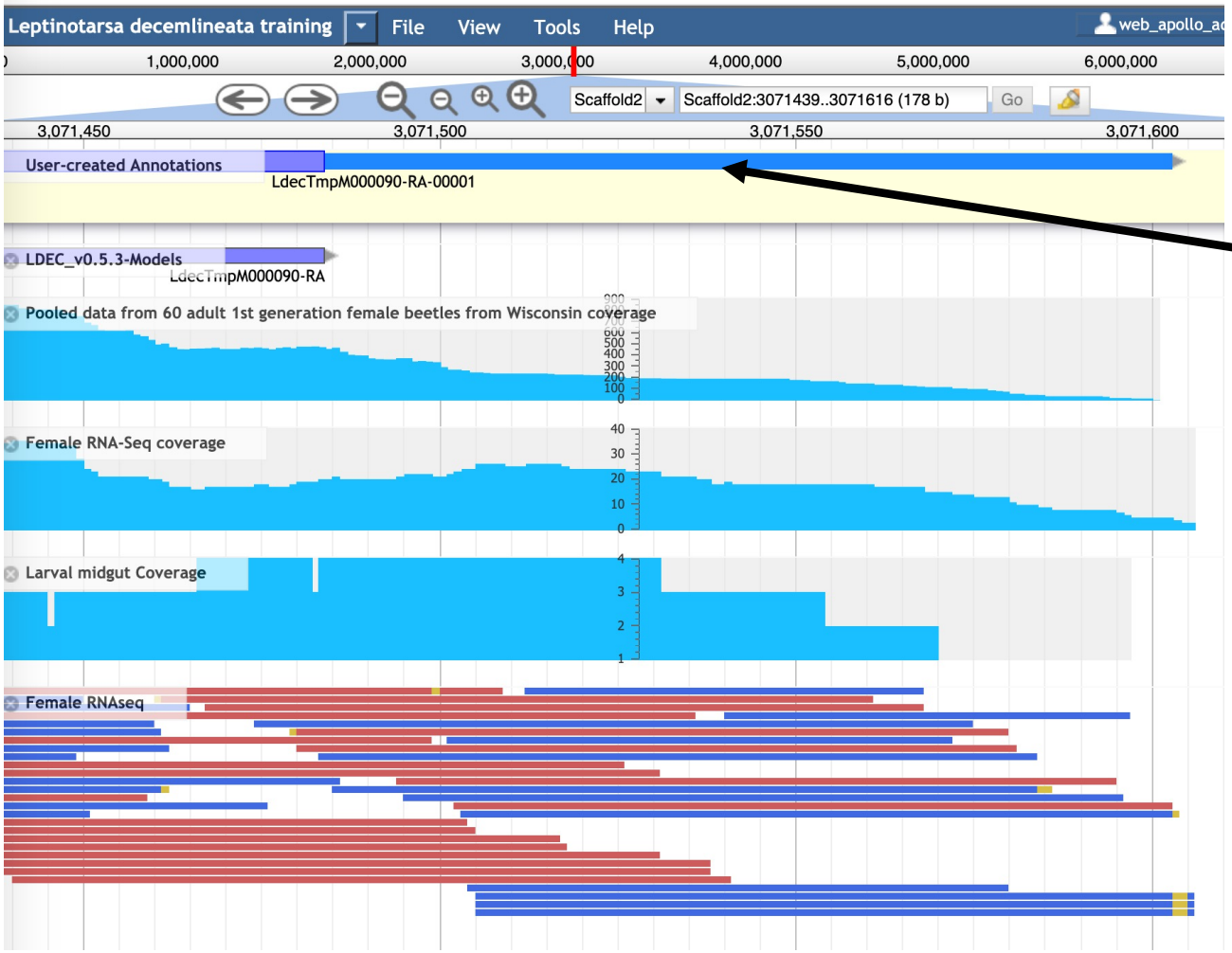
One way to change the boundary: find a mapped read on the same strand as the model; hold shift and click on the read and the model to highlight them both

Adjusting gene boundaries



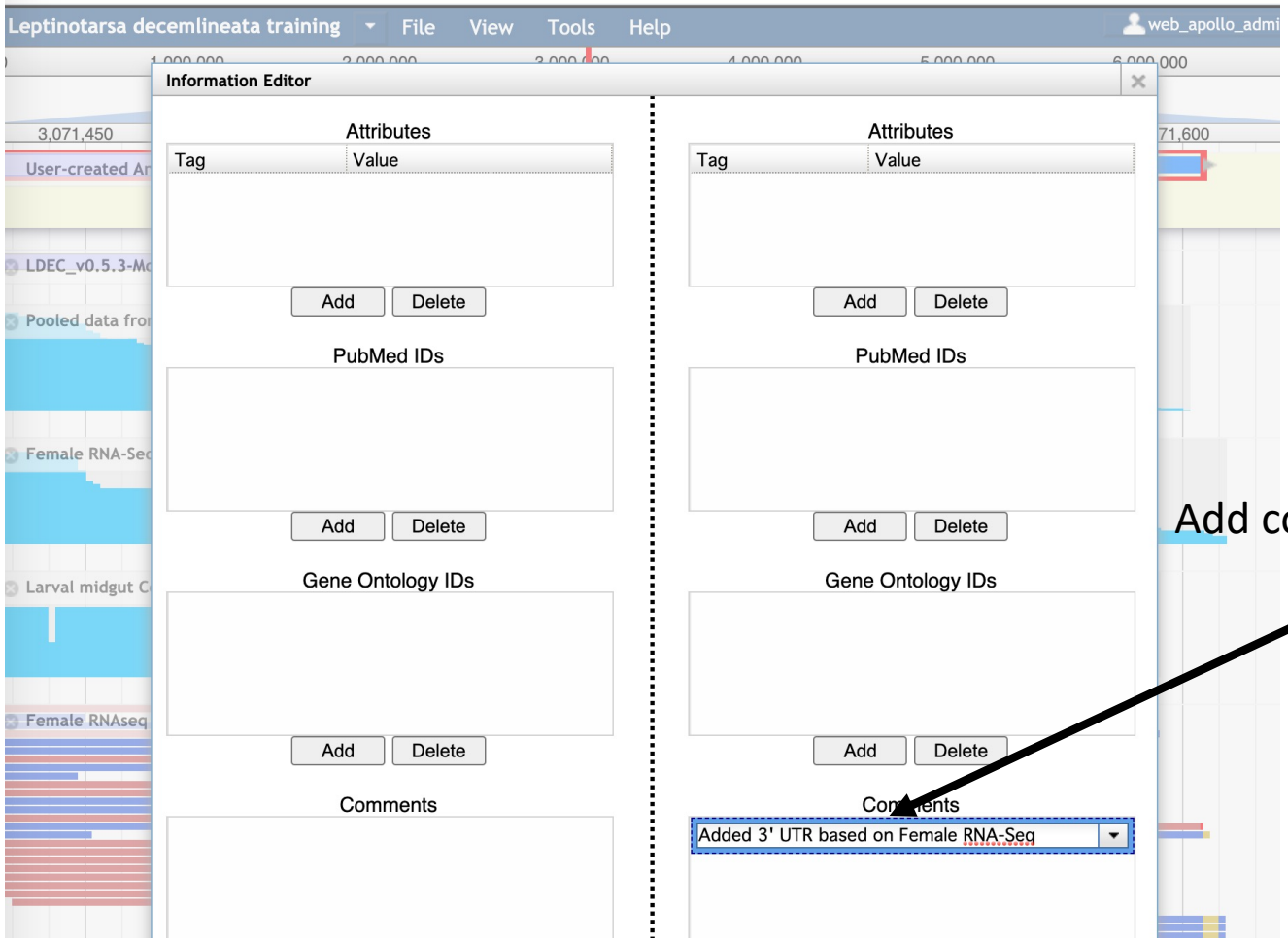
Right-click on model in user-created annotations track, and select ‘Set as 3’ end’ from the drop-down menu

Adjusting gene boundaries



New UTR is there!

Adjusting gene boundaries

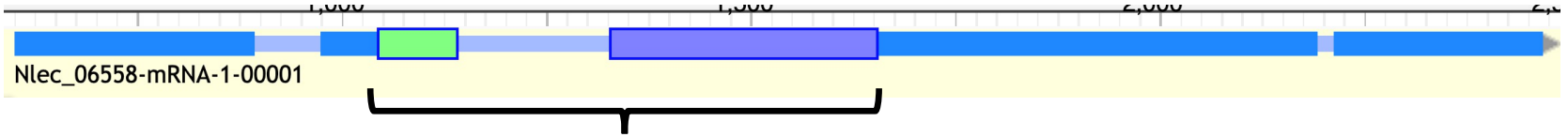


Starts, stops, open reading frames

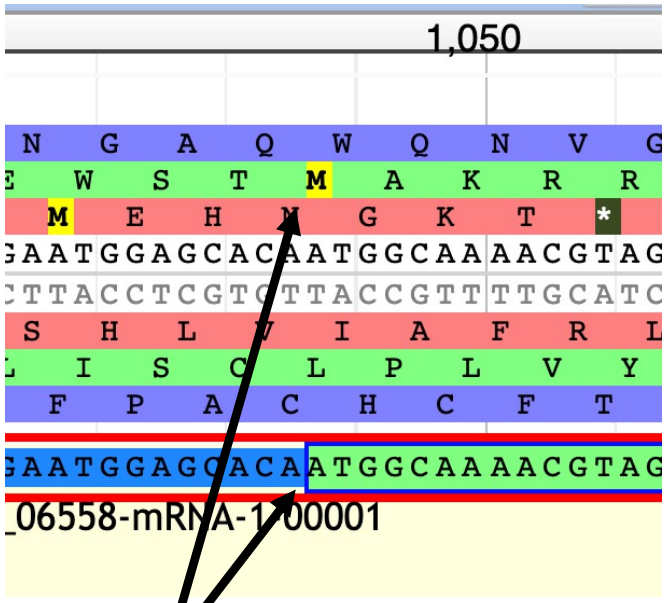
Setting the sequence start, stop, and open reading frame (ORF)

- Apollo will automatically calculate the longest possible ORF that includes canonical ‘Start’ and ‘Stop’ signals (https://genomearchitect.readthedocs.io/en/latest/UsersGuide.html#start-and-stop-sites)
- However, in some fringe cases, you will need to double-check
- You can change a model’s start and stop sites if needed

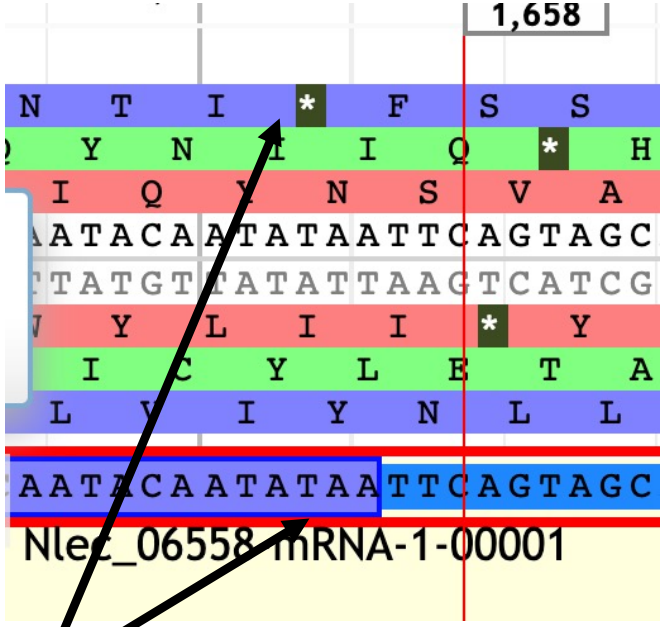
Starts, stops, ORFs



Open reading frame (ORF): translated region

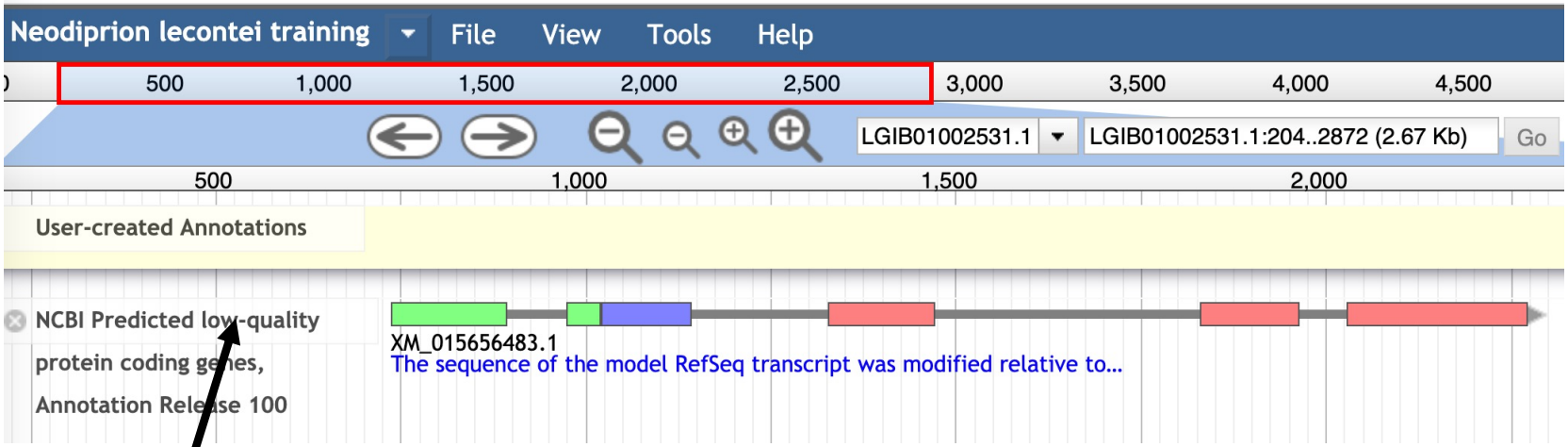


Translation start at Methionine



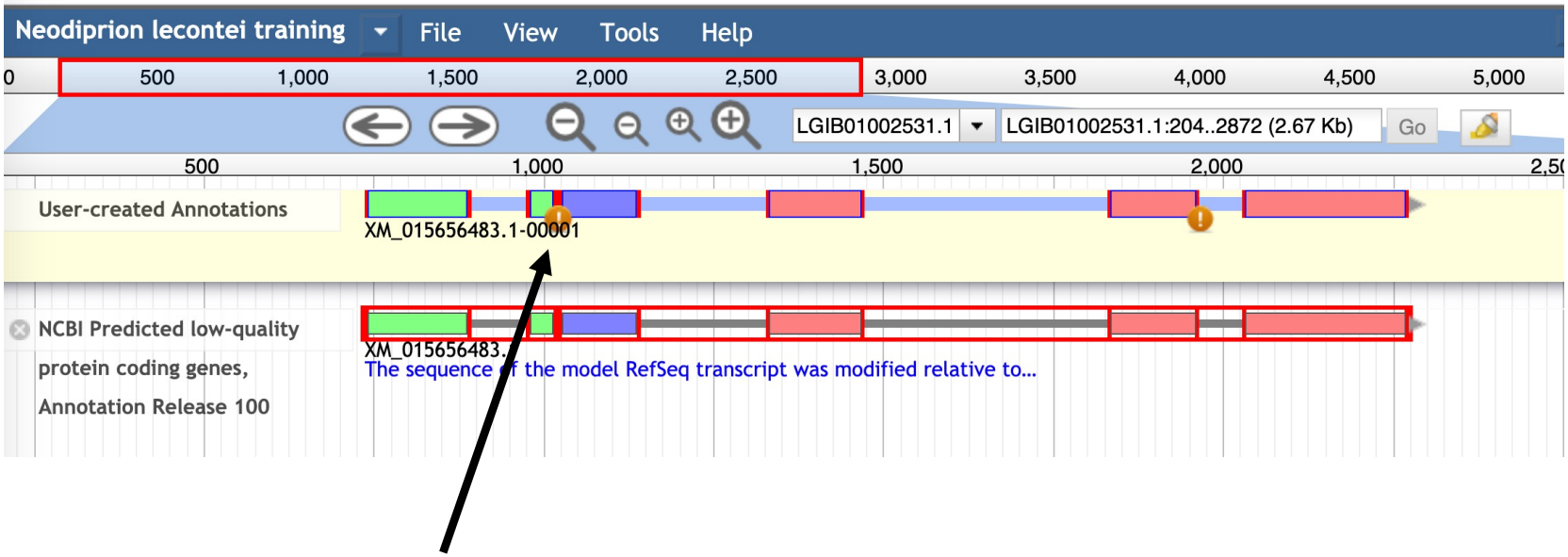
Translation stop at stop codon

Starts, stops, ORFs



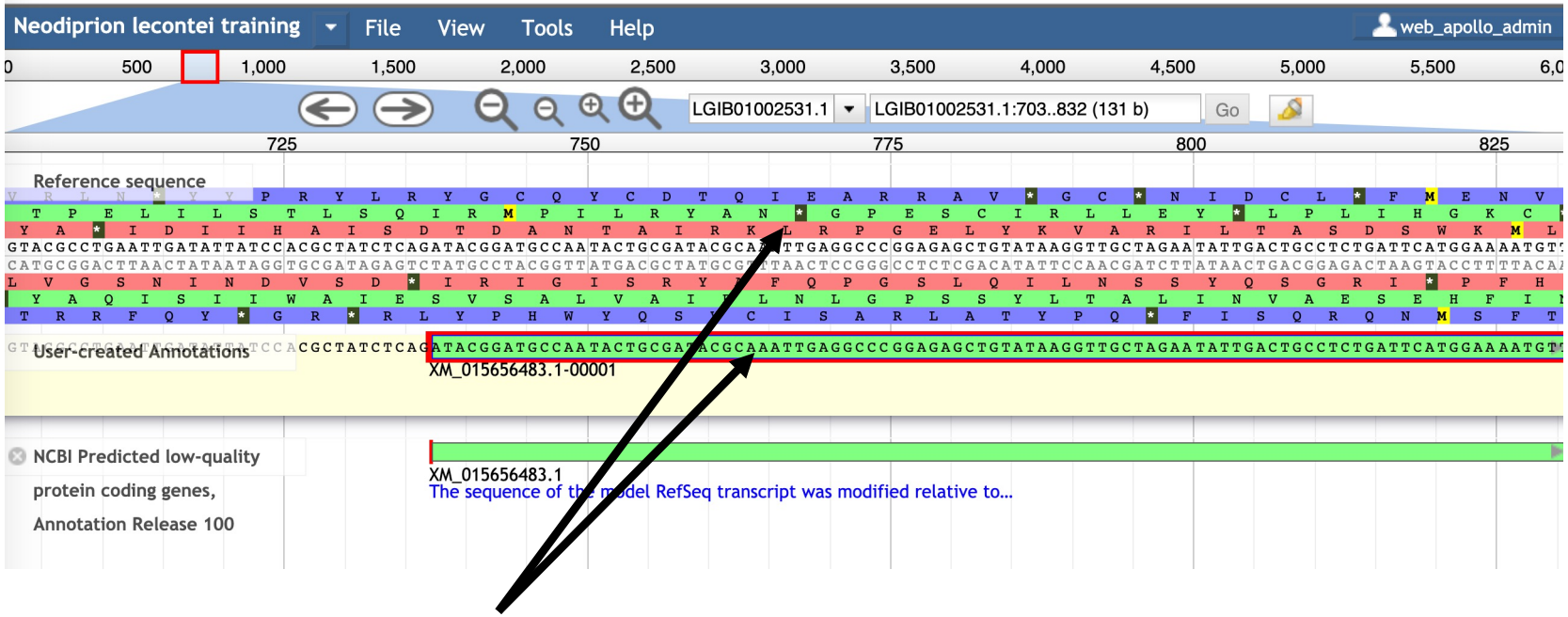
This is a ‘low quality’ protein coding gene from NCBI – it will likely show some problems in Apollo

Starts, stops, ORFs



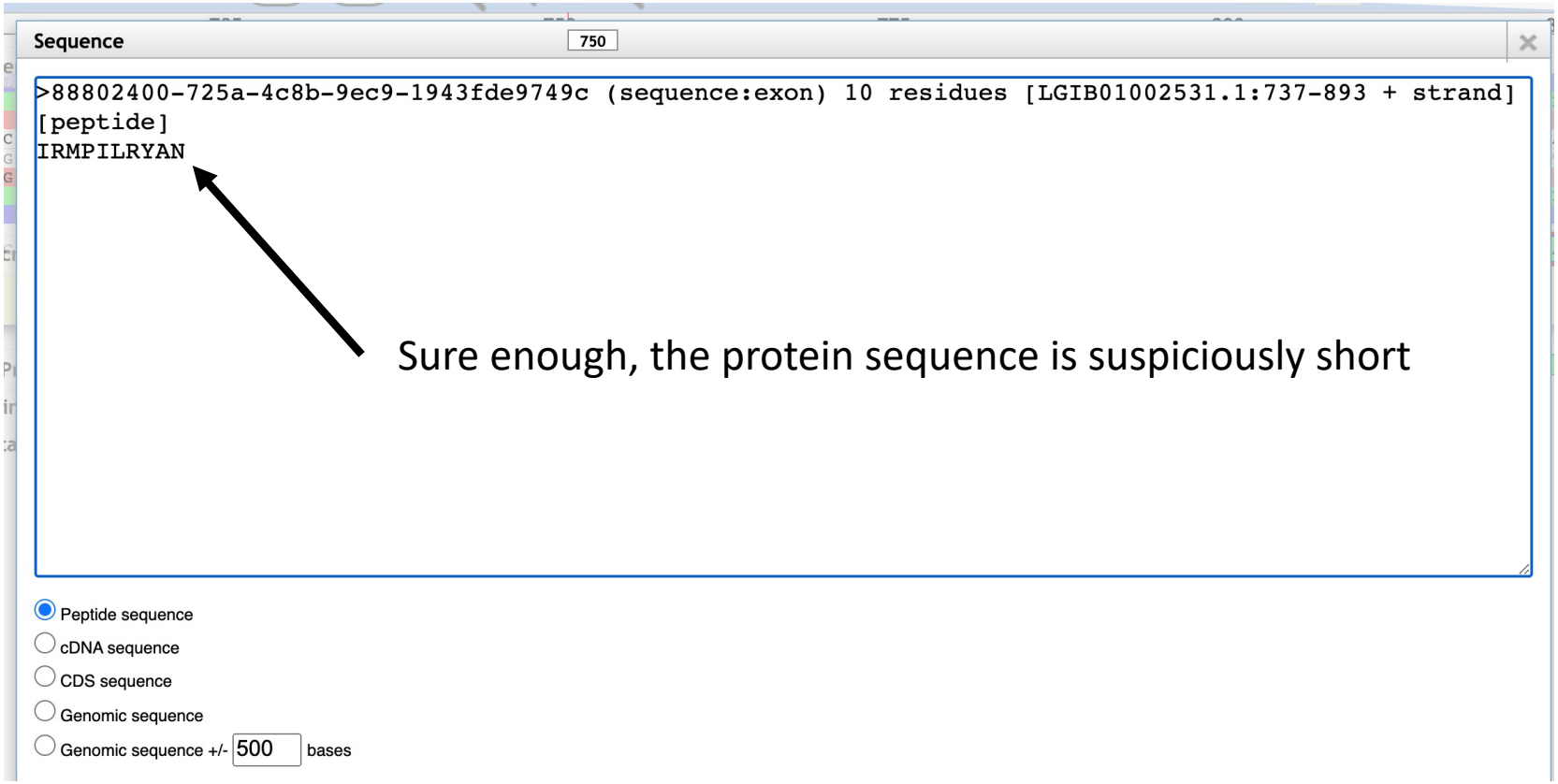
We can see a non-canonical splice site in the UcA (more on that later). Let's zoom to the start of the model.

Starts, stops, ORFs

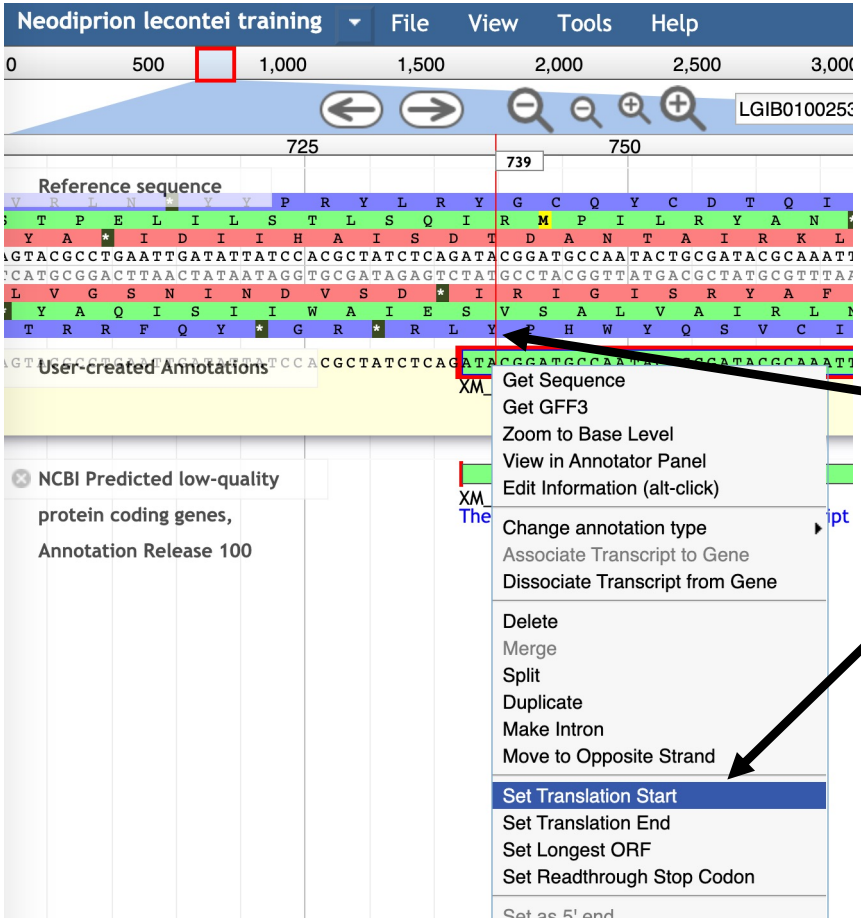


Apollo shows this model in the green reading frame – however, we can see a stop pretty early on in the genome sequence - but that's not reflected in the Apollo model! It looks like the pink reading frame doesn't have stops.

Starts, stops, ORFs

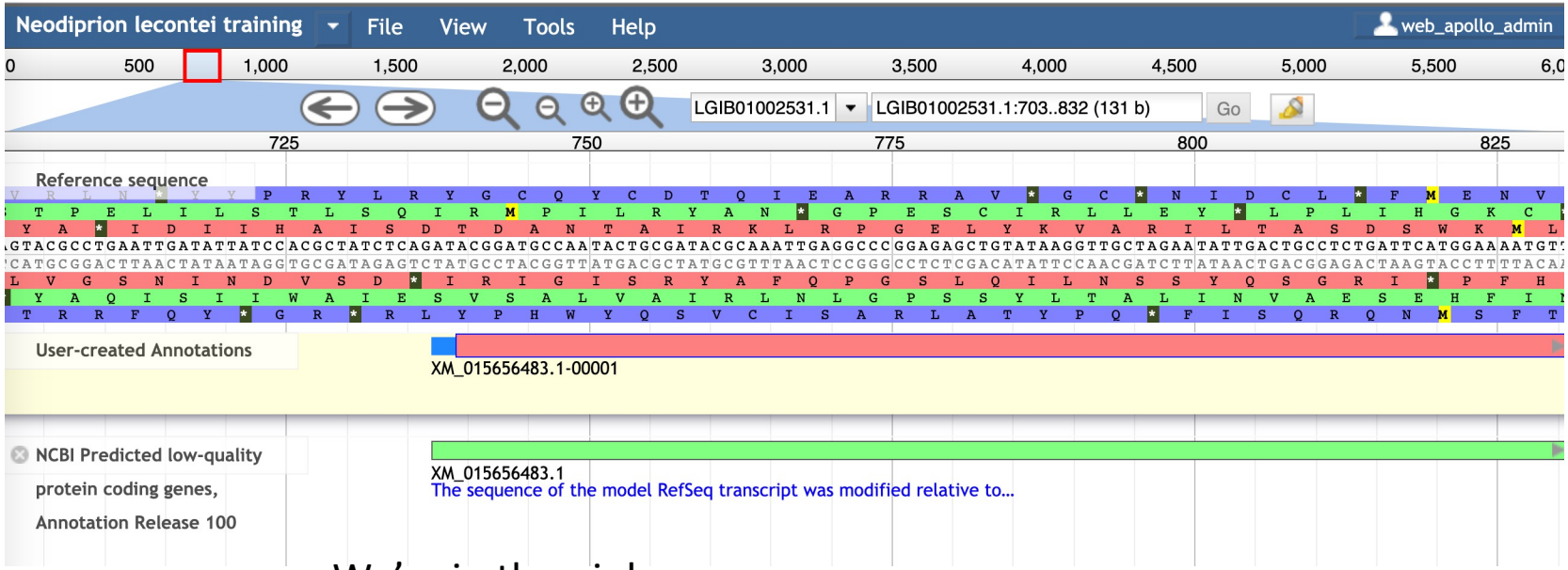


Starts, stops, ORFs



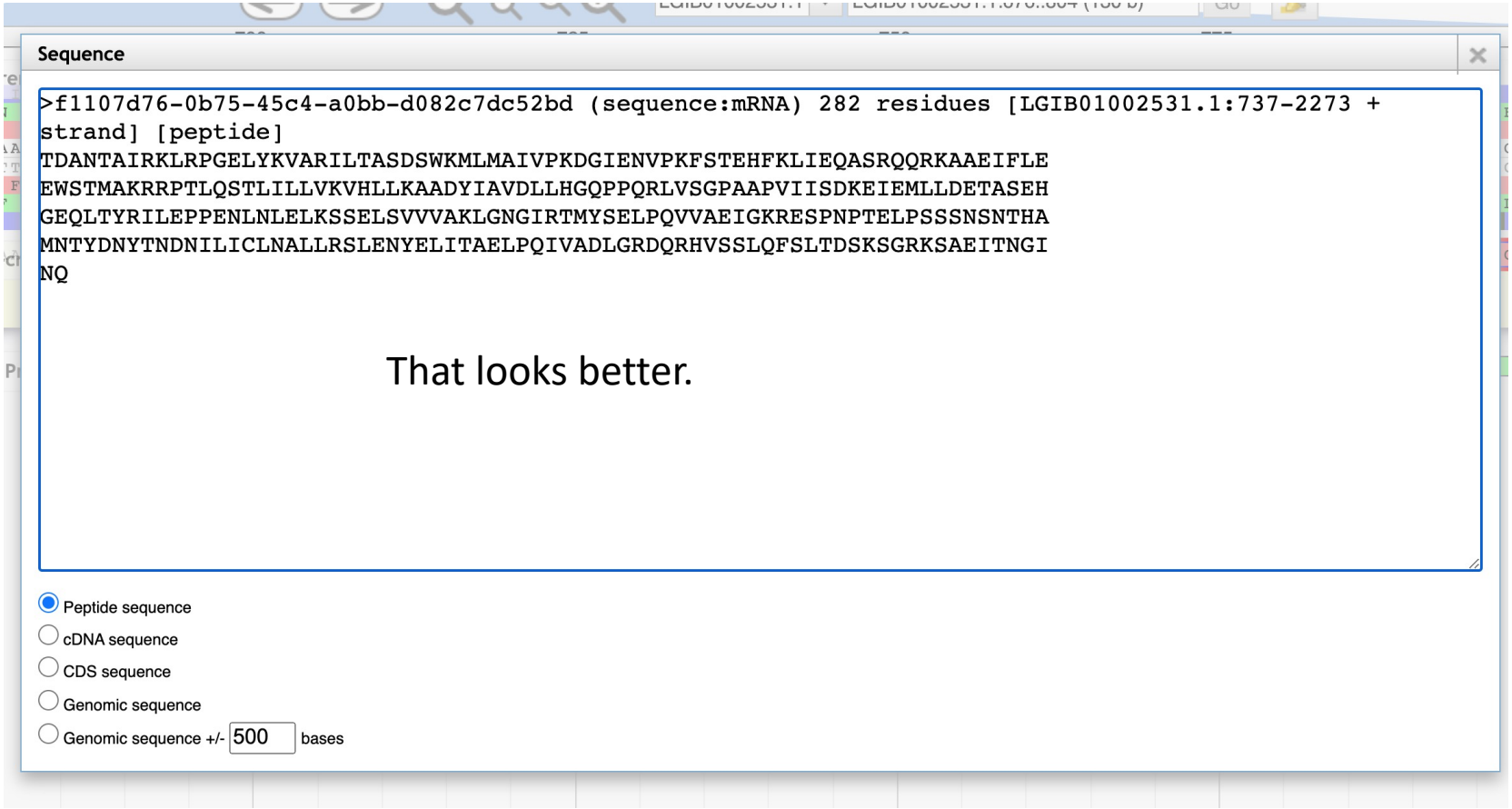
Let's set the translation start in the pink reading frame – click on the 3rd nucleotide in the UcA, right-click, and select 'Set Translation Start'

Starts, stops, ORFs

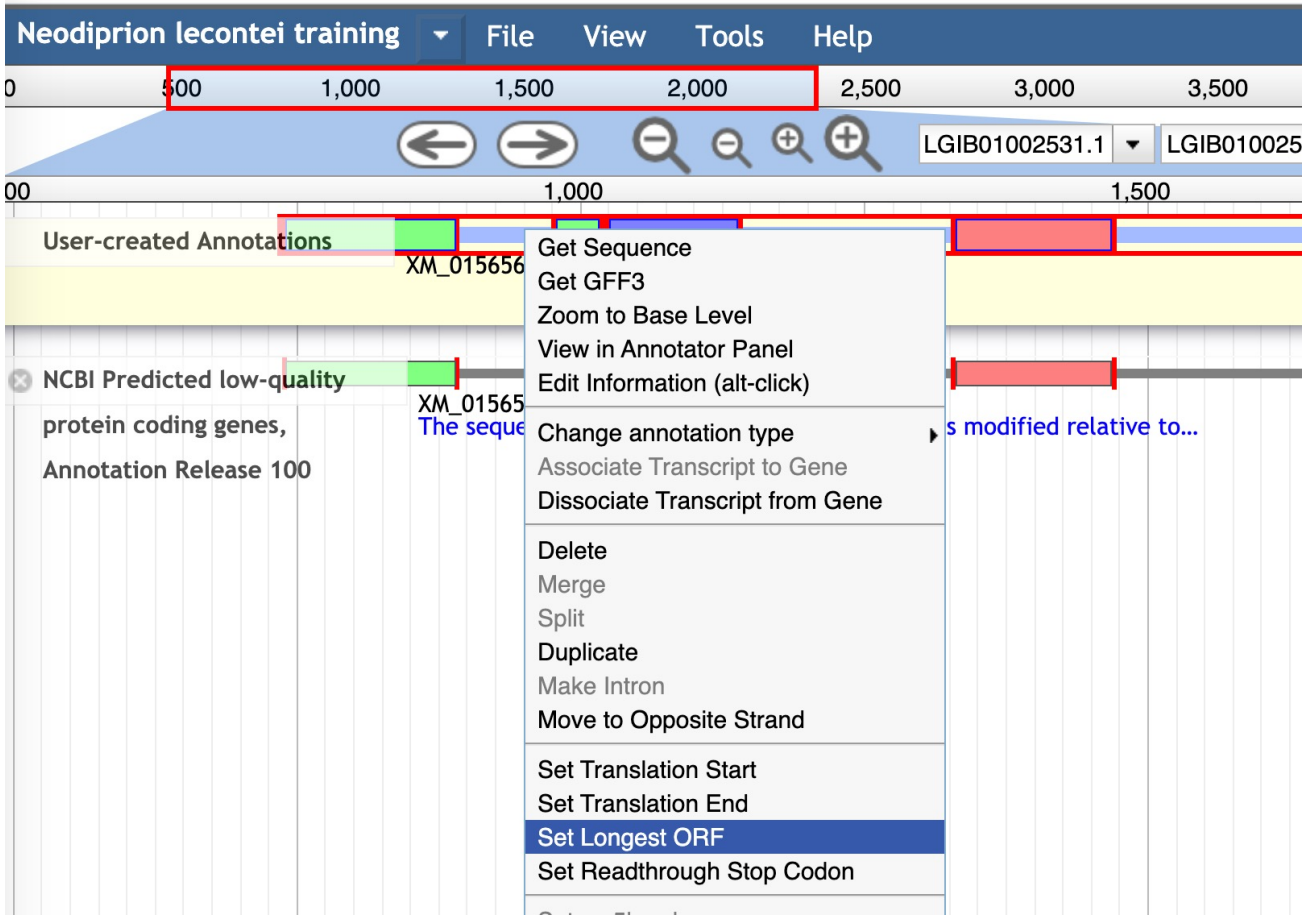


We're in the pink reading frame now – let's check the protein sequence

Starts, stops, ORFs

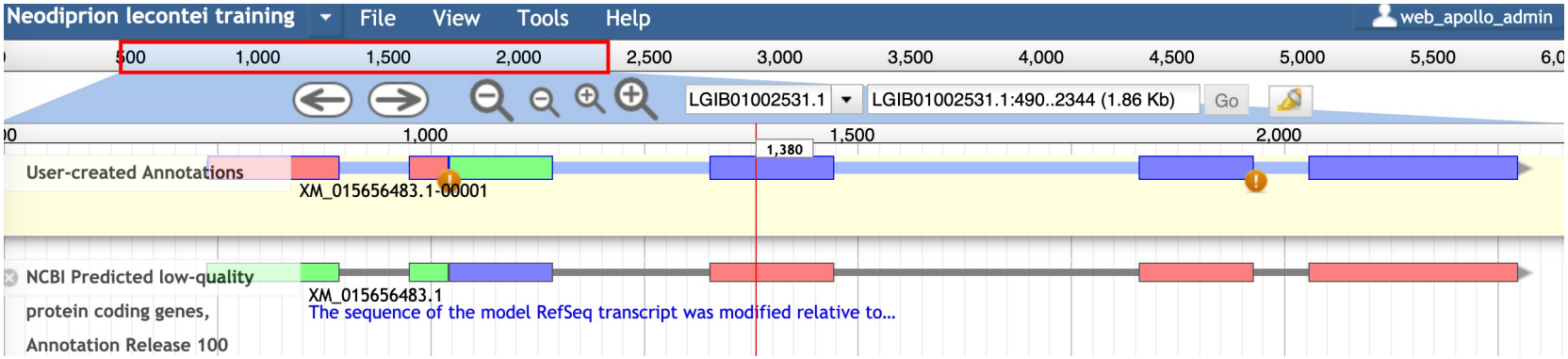


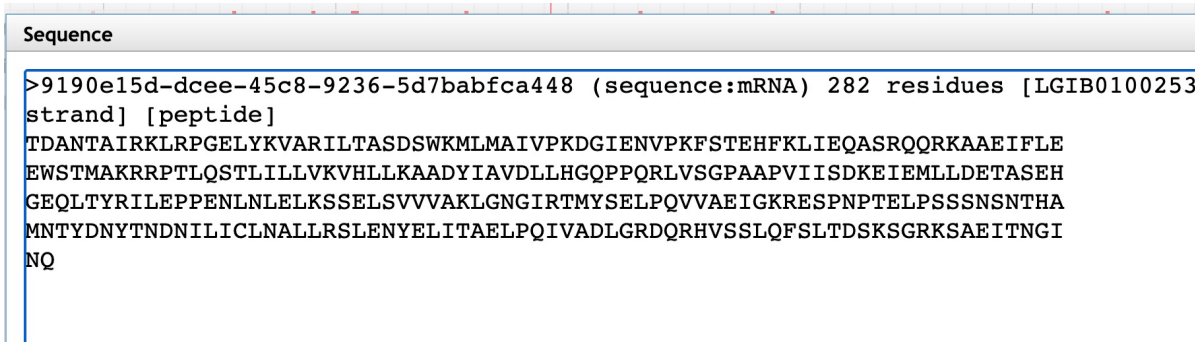
Starts, stops, ORFs



Sometimes it can be hard to tell what the protein sequence should be – in that case you can right-click and select 'Set Longest ORF'

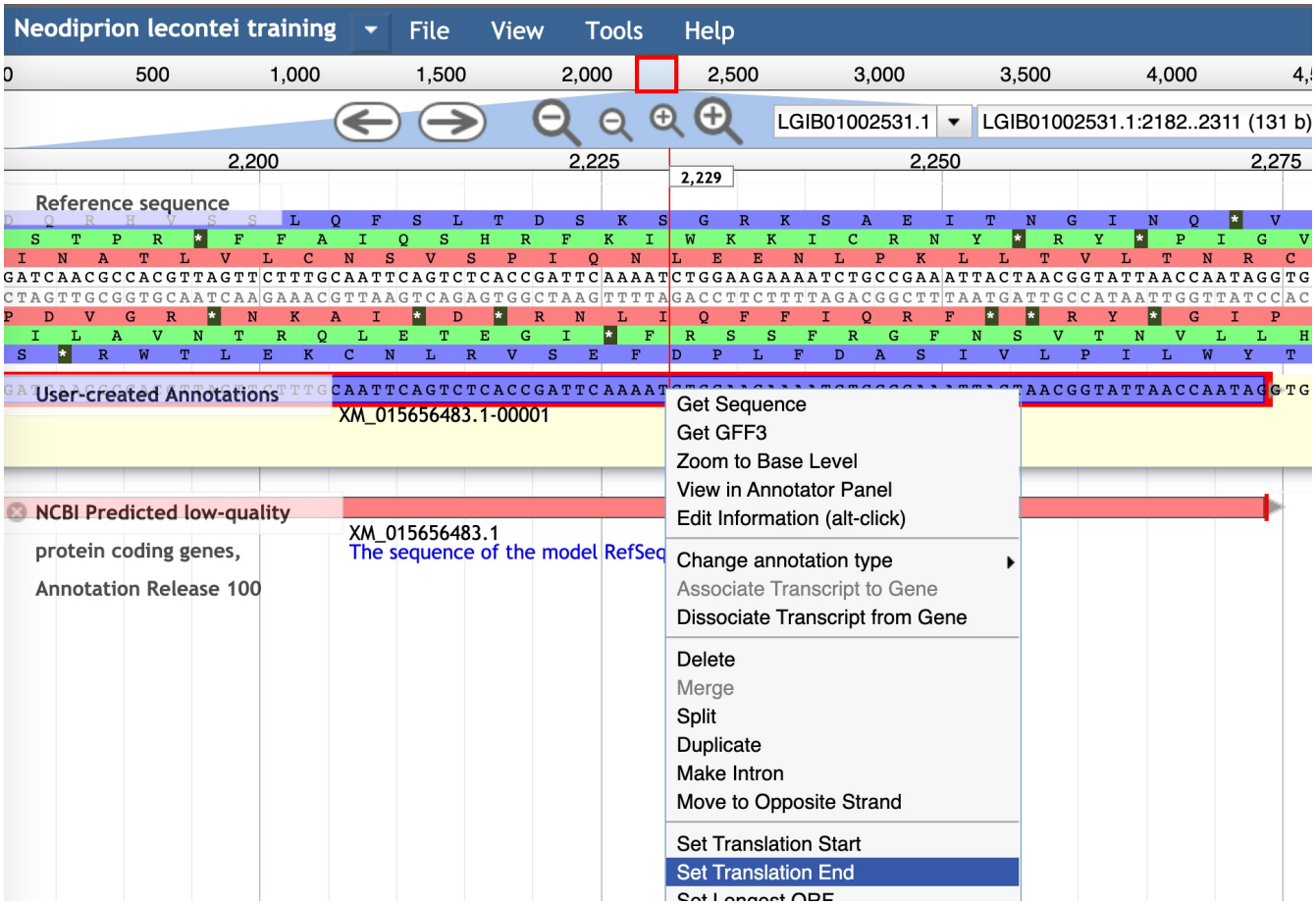
Starts, stops, ORFs





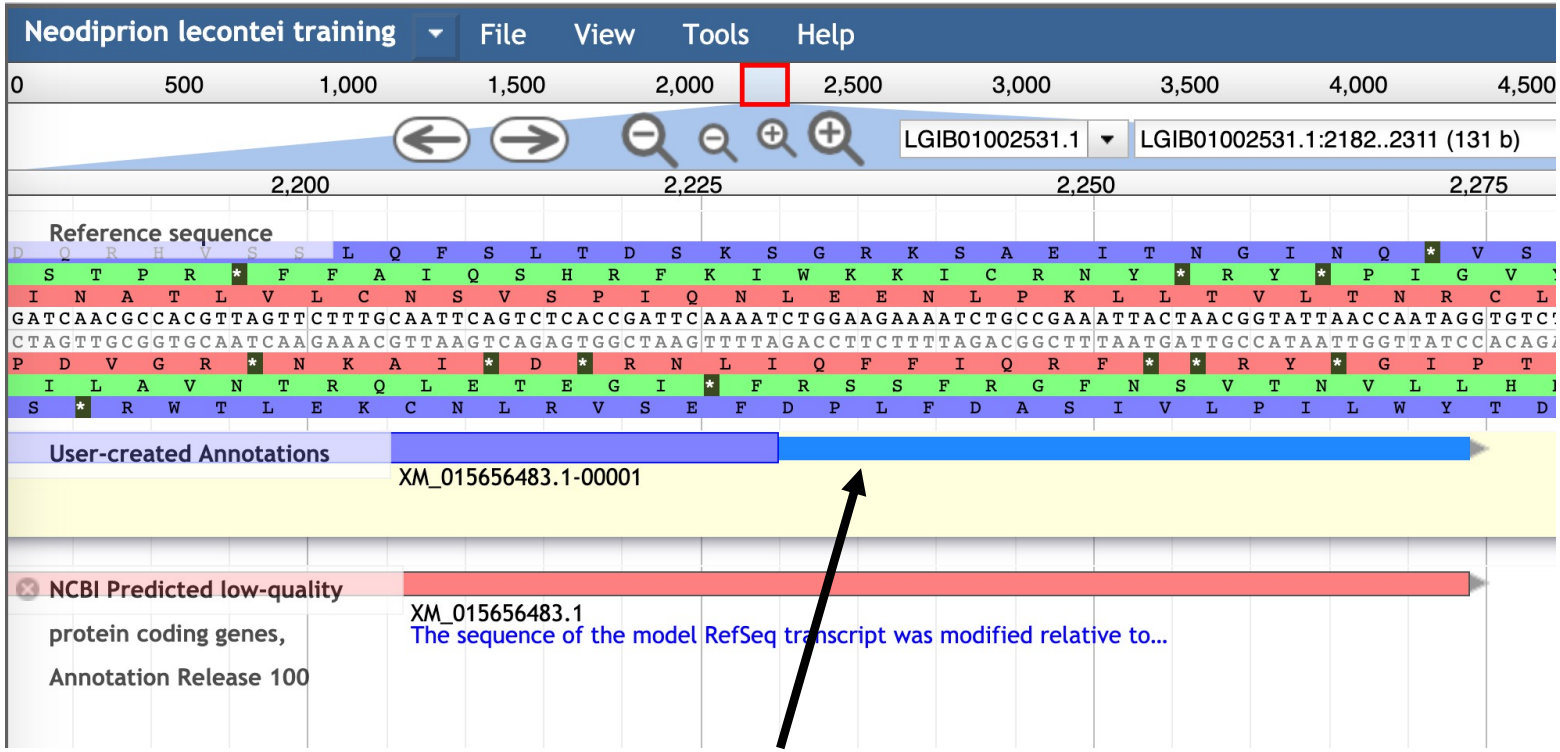
This also fixed the reading frame.

Starts, stops, ORFs



Similarly, if you have evidence to change the translation end, you can click on the corresponding nucleotide, right-click, and select ‘Set Translation End’

Starts, stops, ORFs



Now the sequence after the translation end is 3’ UTR.

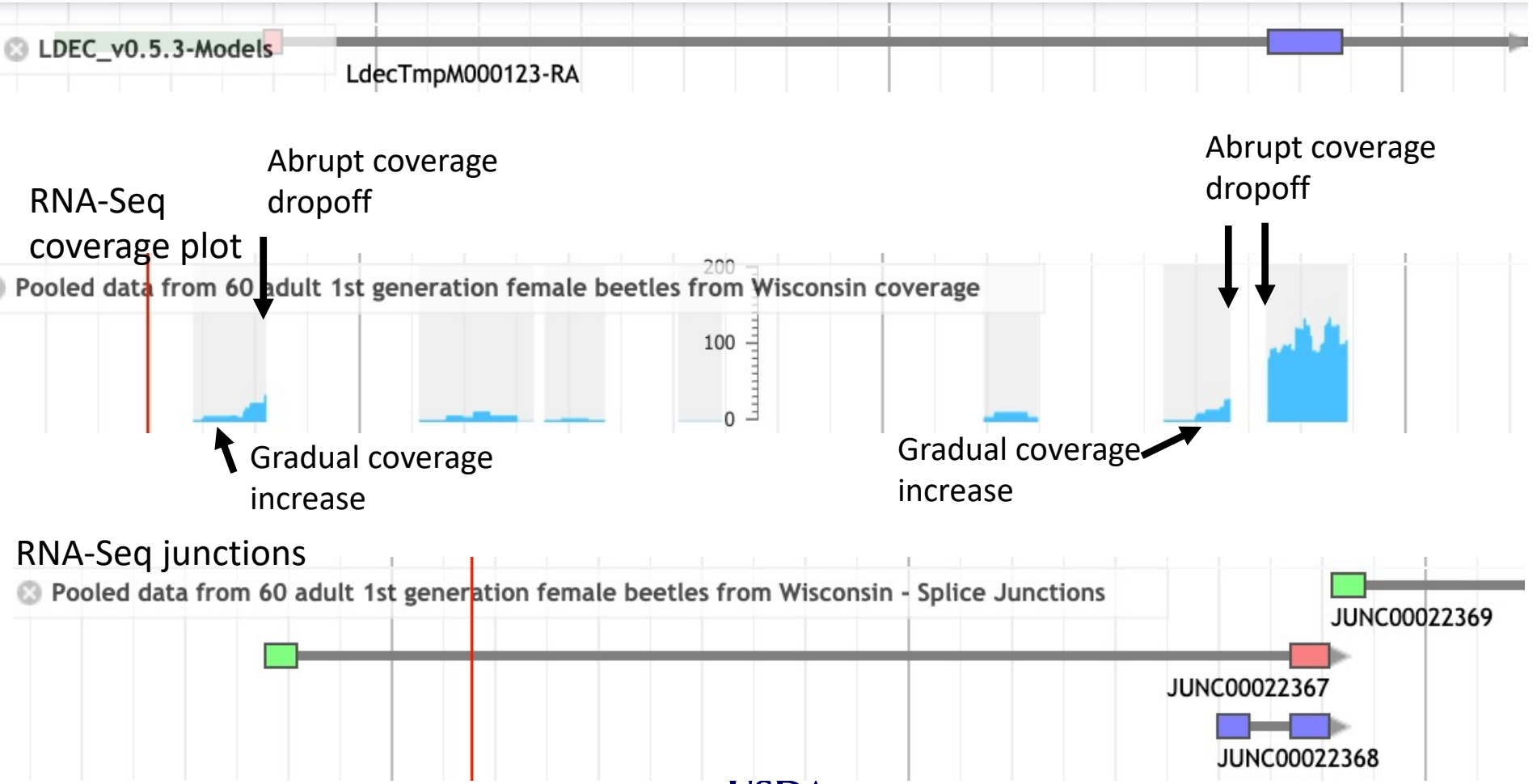
Annotating isoforms

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
- Checking junction reads (if available) is incredibly useful

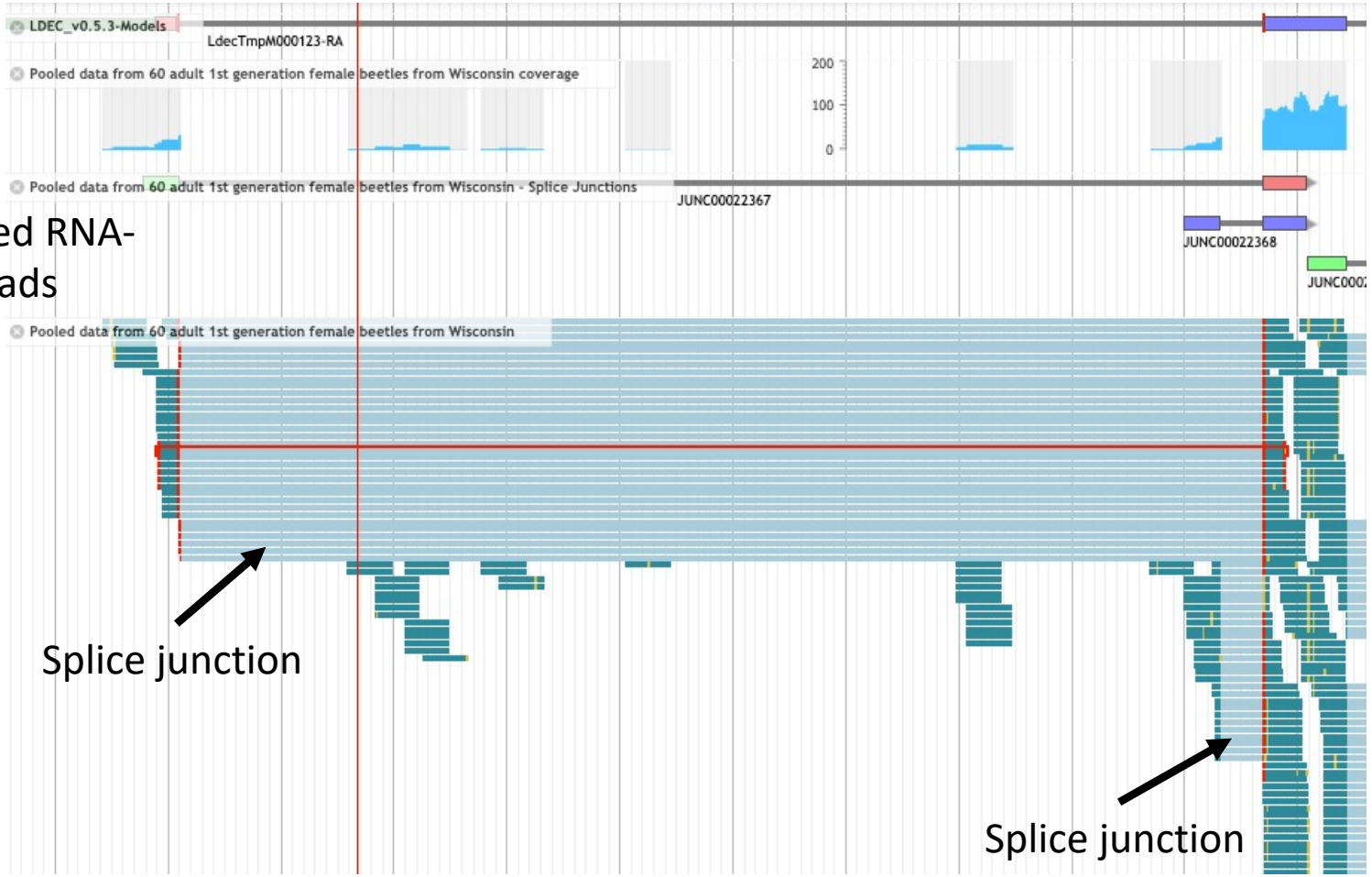
Isoform annotation example

5’ end of MAKER tyrosine protein kinase gene prediction



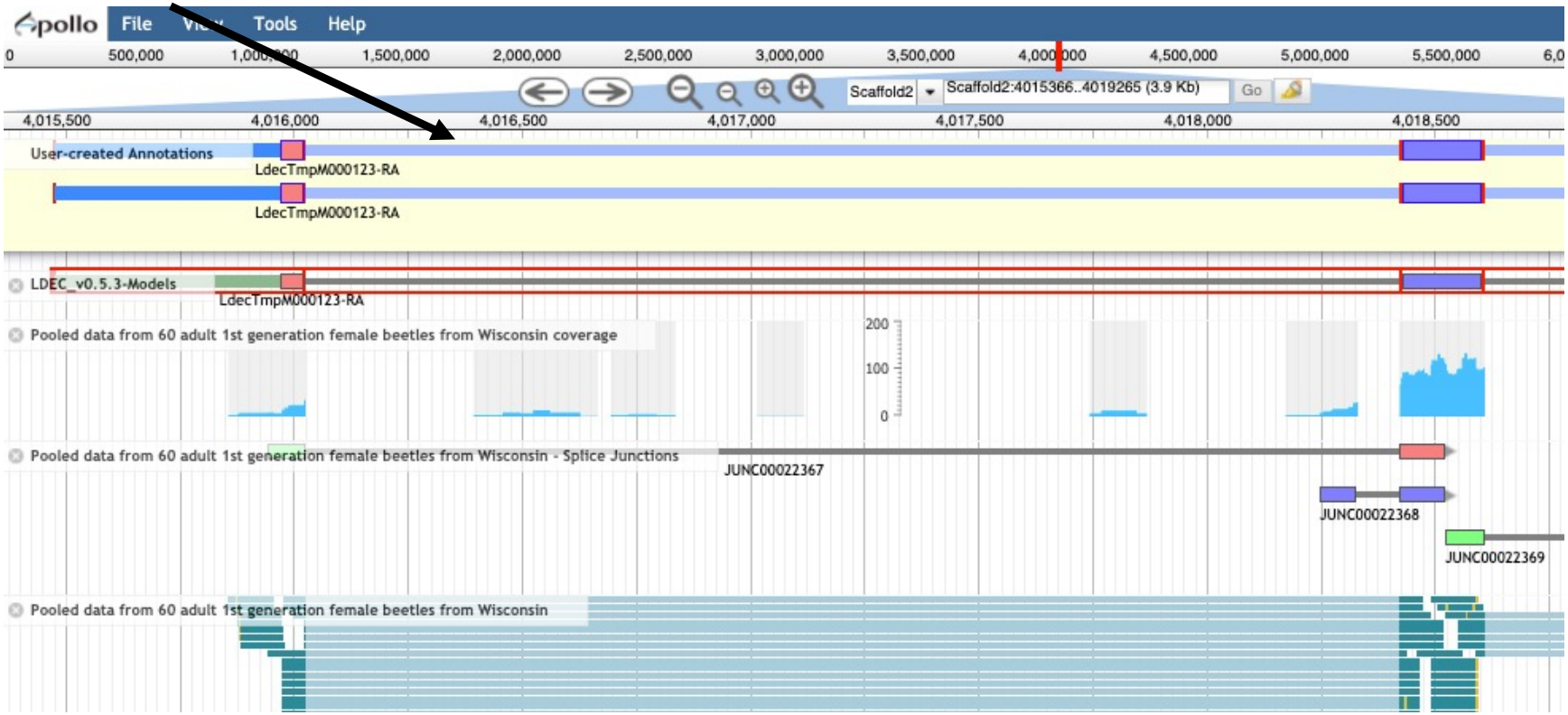
Isoform annotation example

Mapped RNA-Seq reads



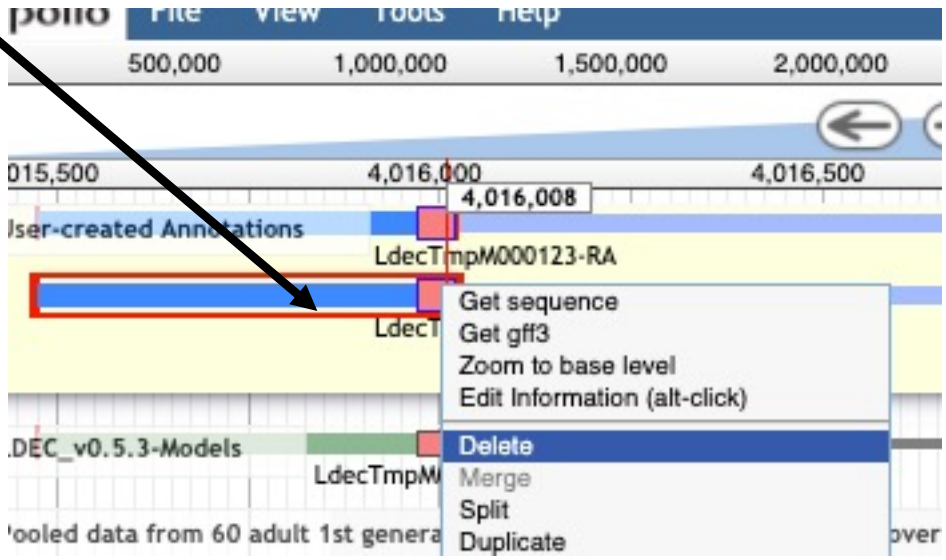
Isoform annotation example

Create 2 isoforms
from Maker model



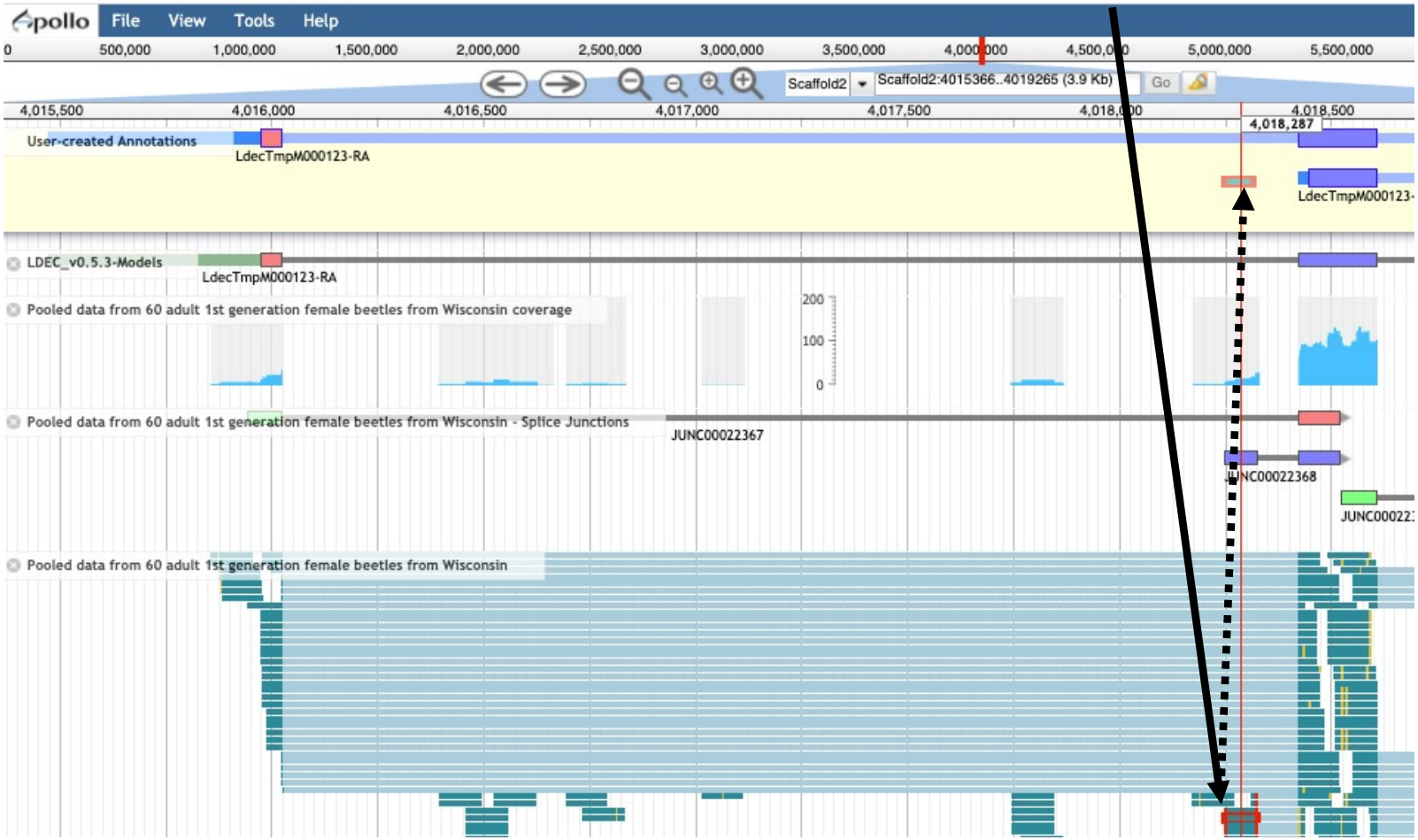
Isoform annotation example

Select and delete 5’ exon from one of the isoforms



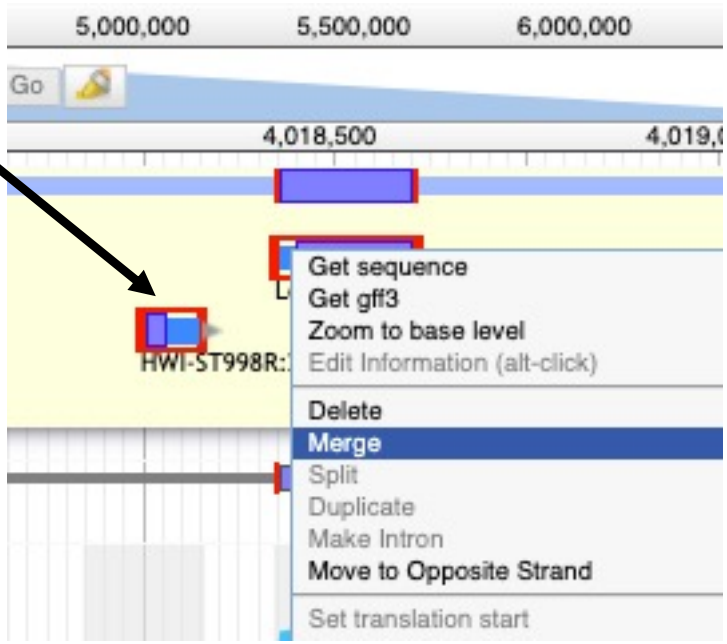
Isoform annotation example

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



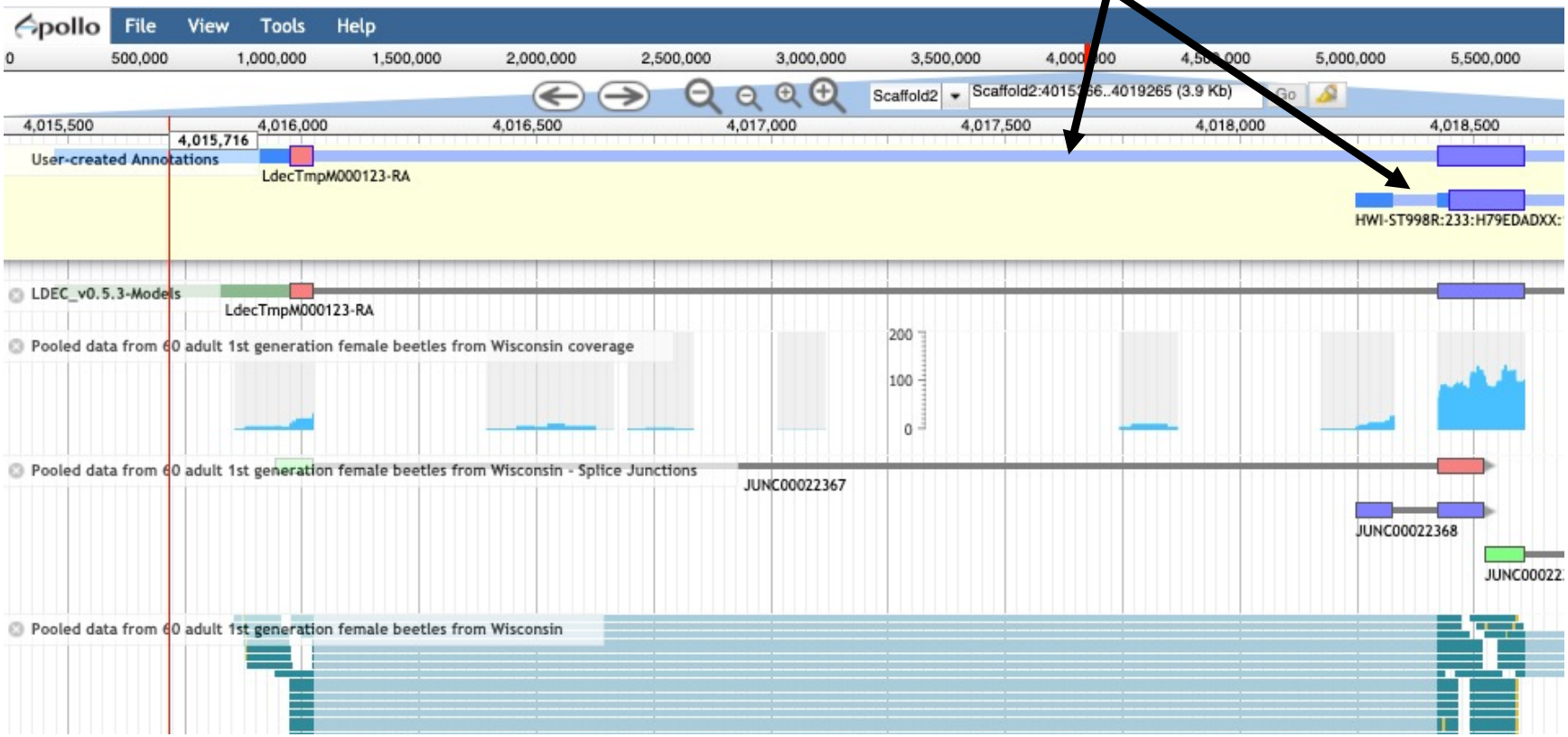
Isoform annotation example

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



Isoform annotation example

2 isoforms supported by RNA-Seq evidence

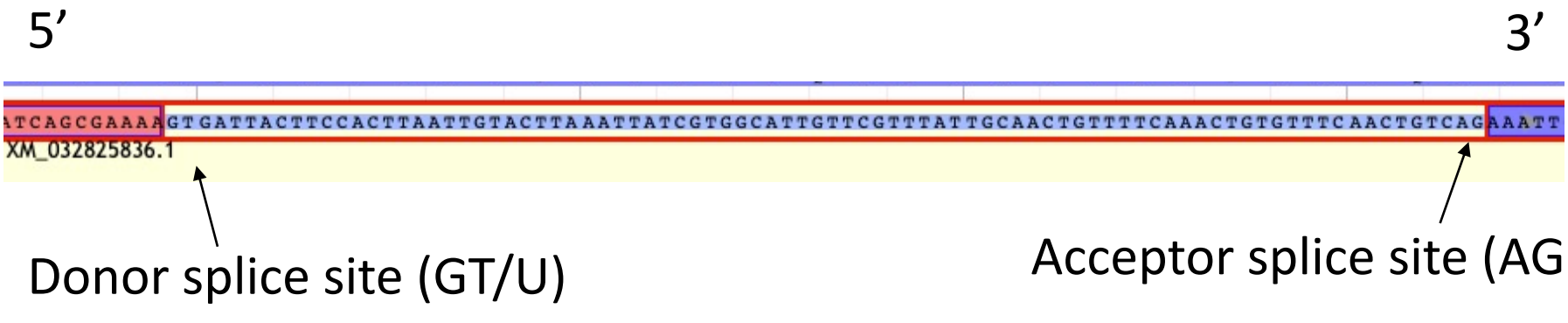


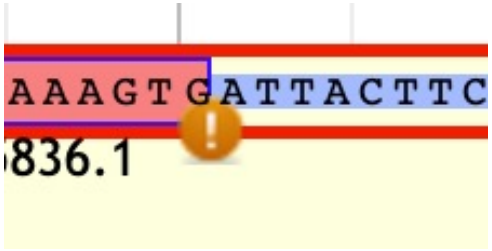
Non-canonical splice sites

Splice sites

Introns are removed from primary transcripts by cleavage at conserved sequences called **splice sites**. These sites are found at the 5′ and 3′ ends of introns.

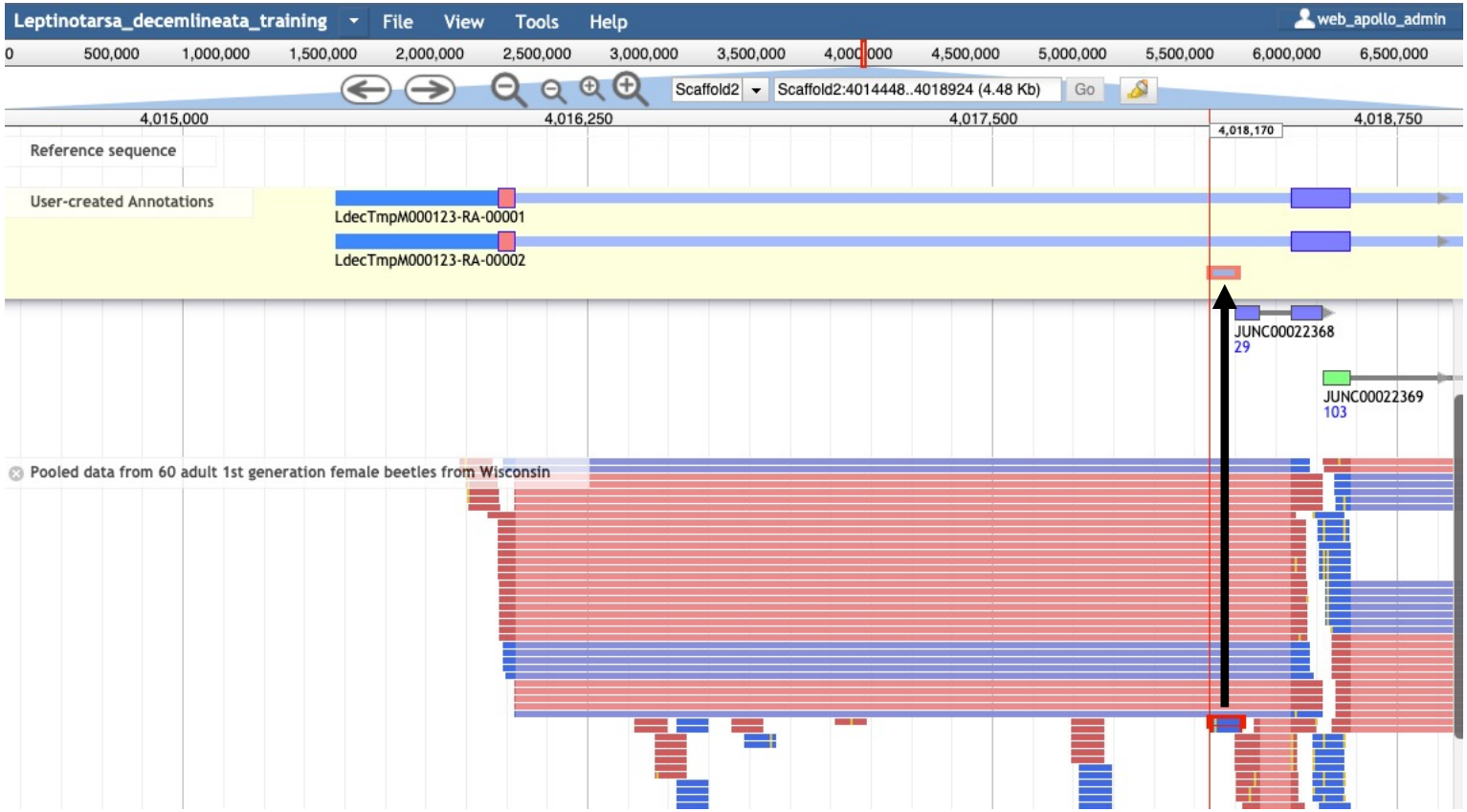
(https://www.nature.com/scitable/topicpage/rna-splicing-introns-exons-and-spliceosome-12375/)





’Non-canonical’ splice sites – non-conserved, and possibly erroneous sites – are marked by an exclamation point in Apollo.

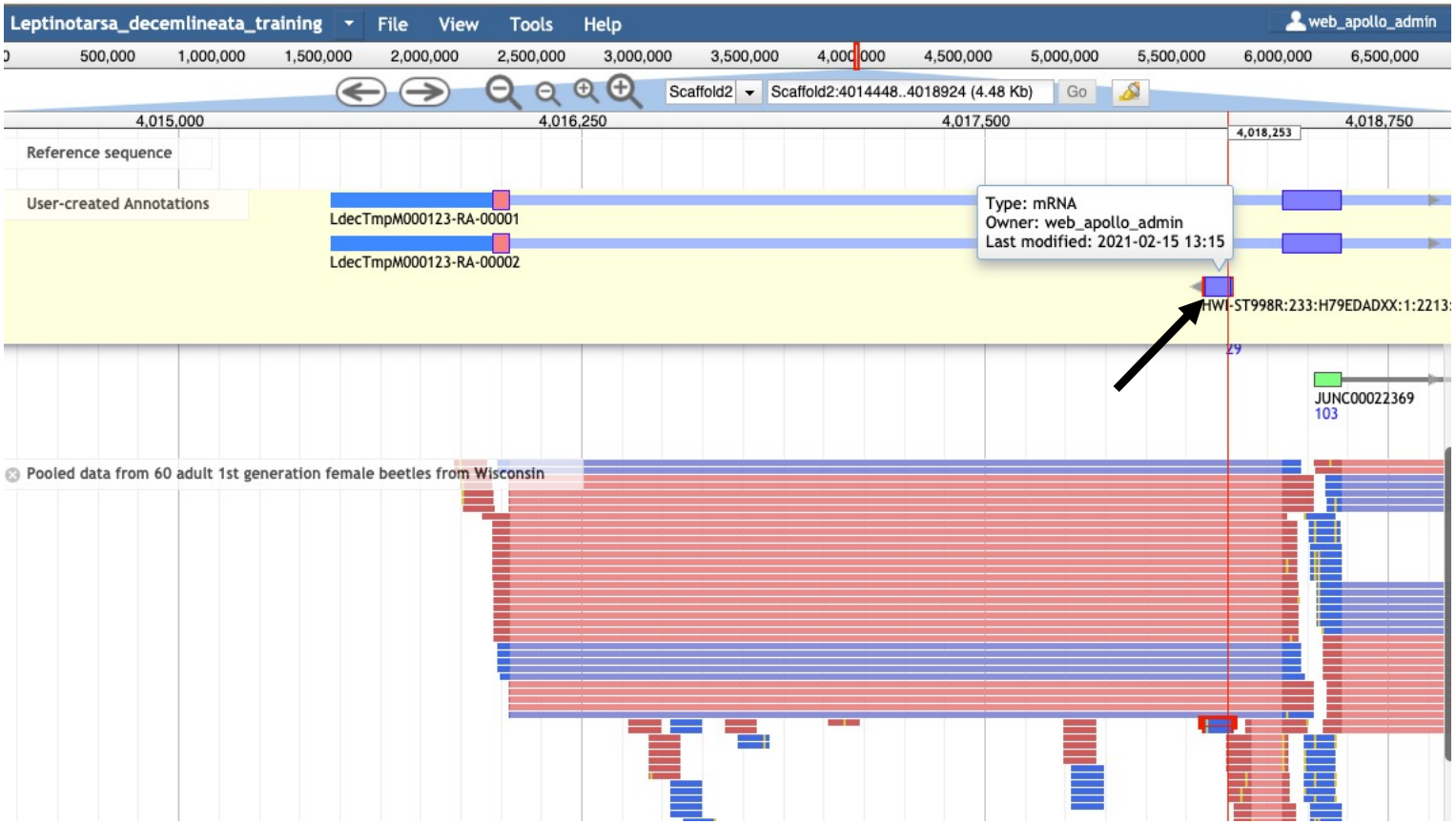
Fixing non-canonical splice sites



Let's revisit the previous 'Isoform annotation' example. Let's add the new exon using different starting evidence.

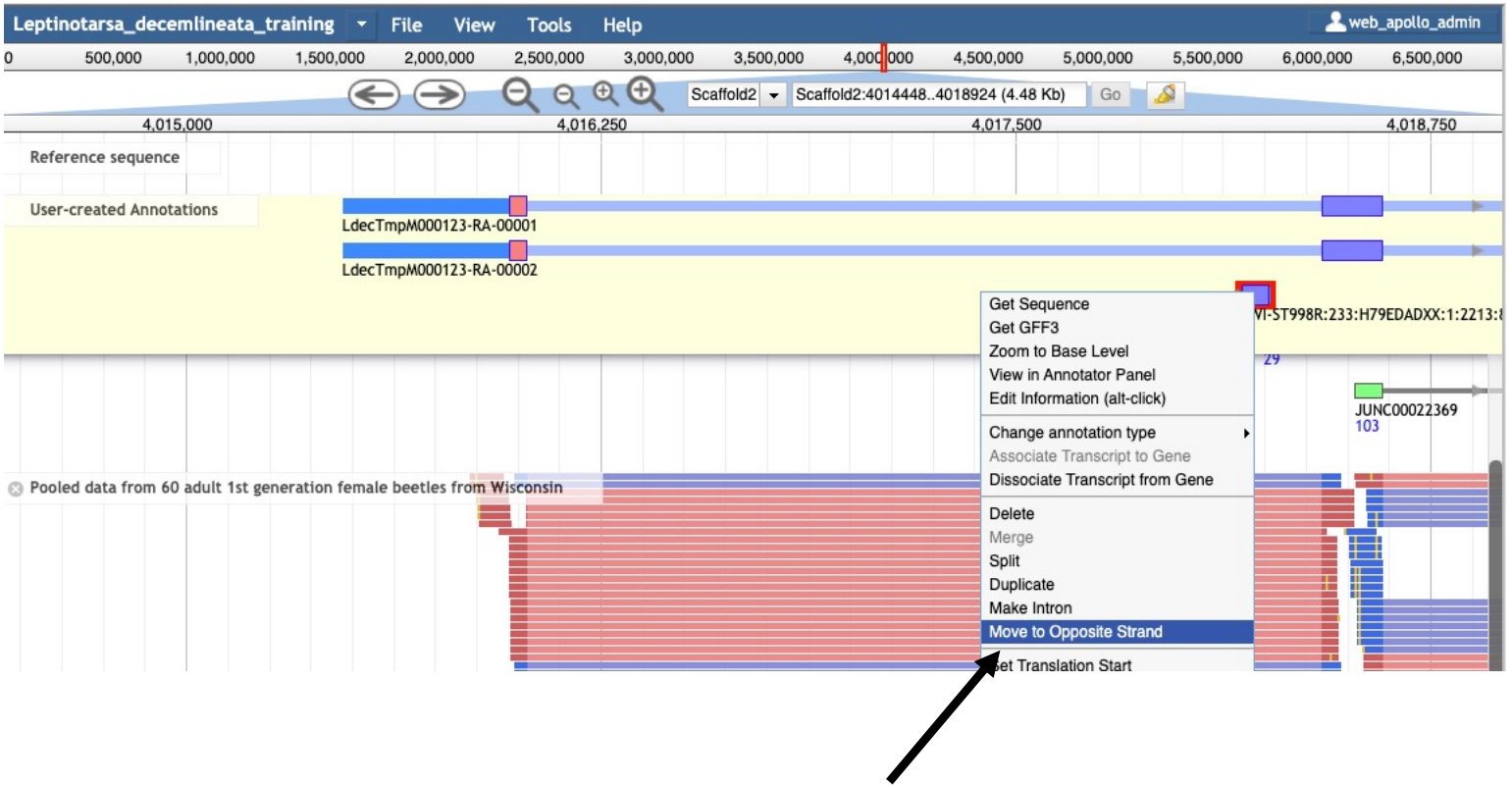
USDA

Fixing non-canonical splice sites



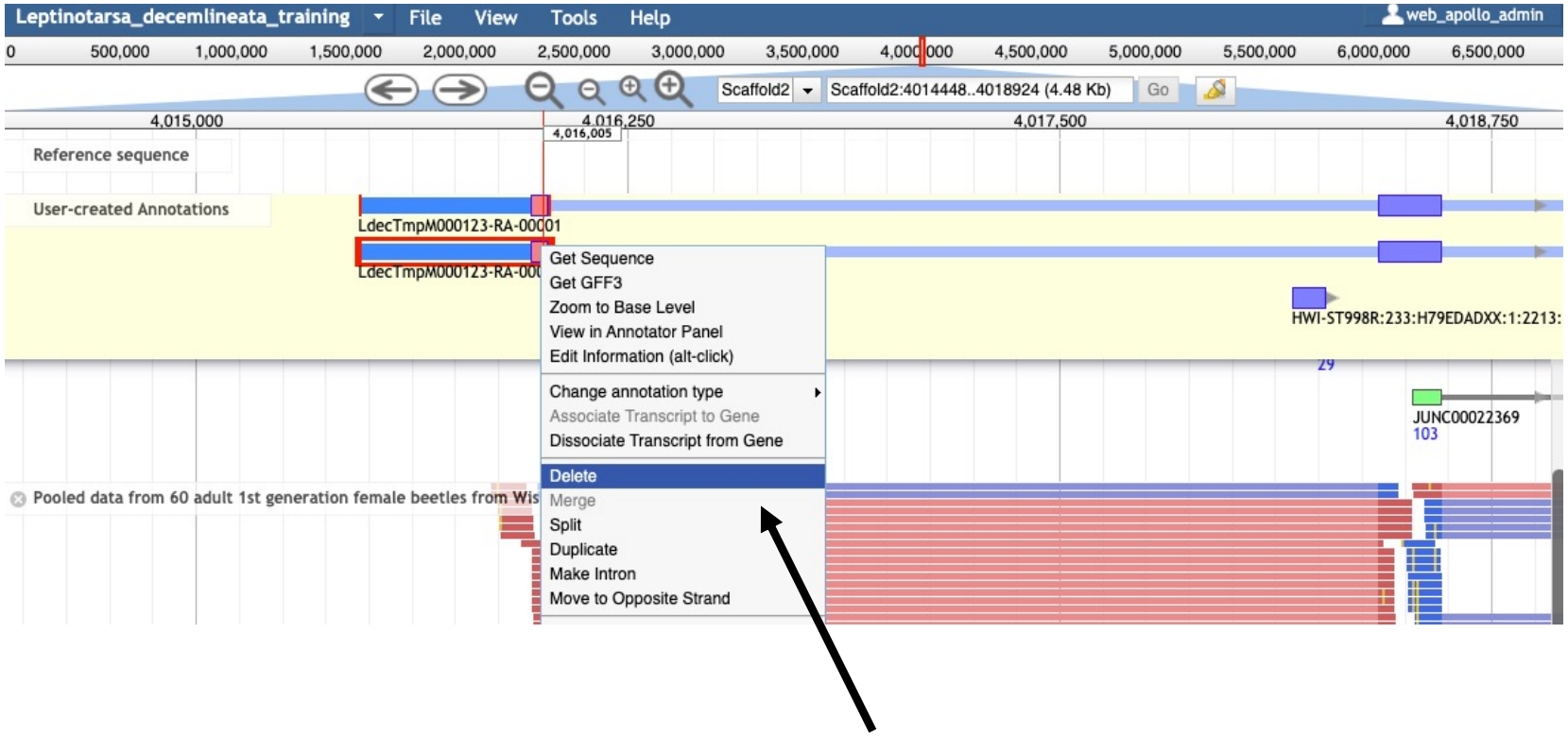
That starting evidence was mapped to the reverse strand – let's flip it to the forward strand

Fixing non-canonical splice sites



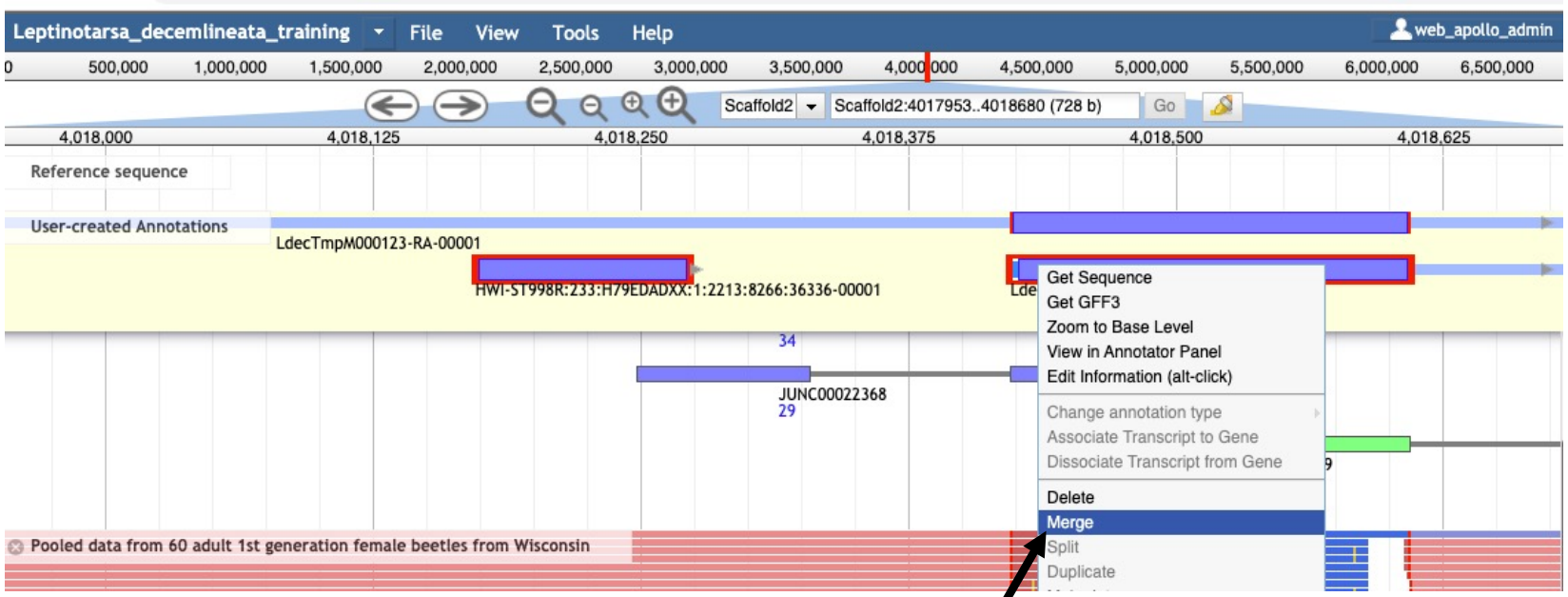
Right-click on the evidence and select ‘Move to Opposite Strand’

Fixing non-canonical splice sites



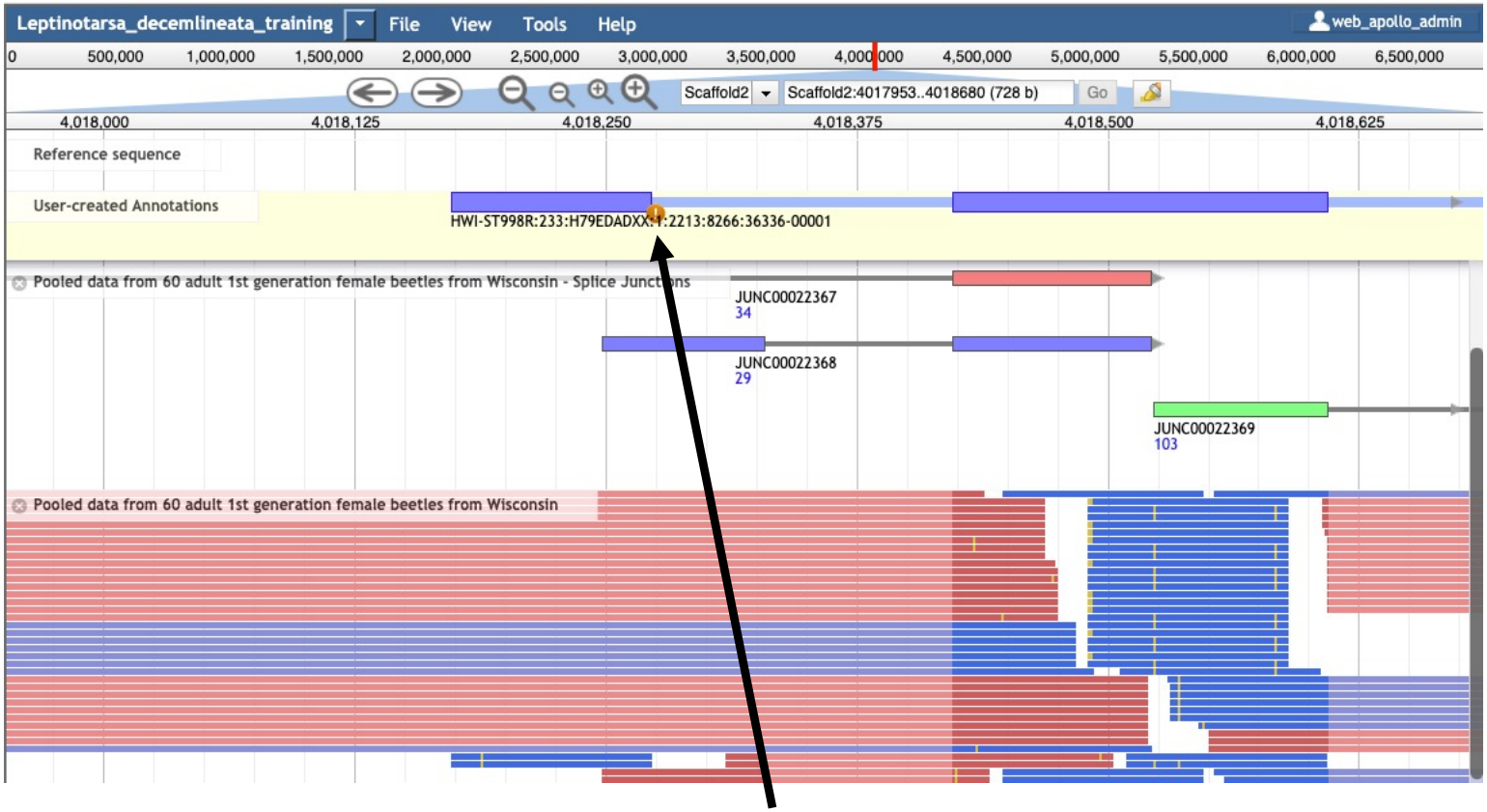
Delete the 5’ exon as previously (no support for it for this isoform)

Fixing non-canonical splice sites



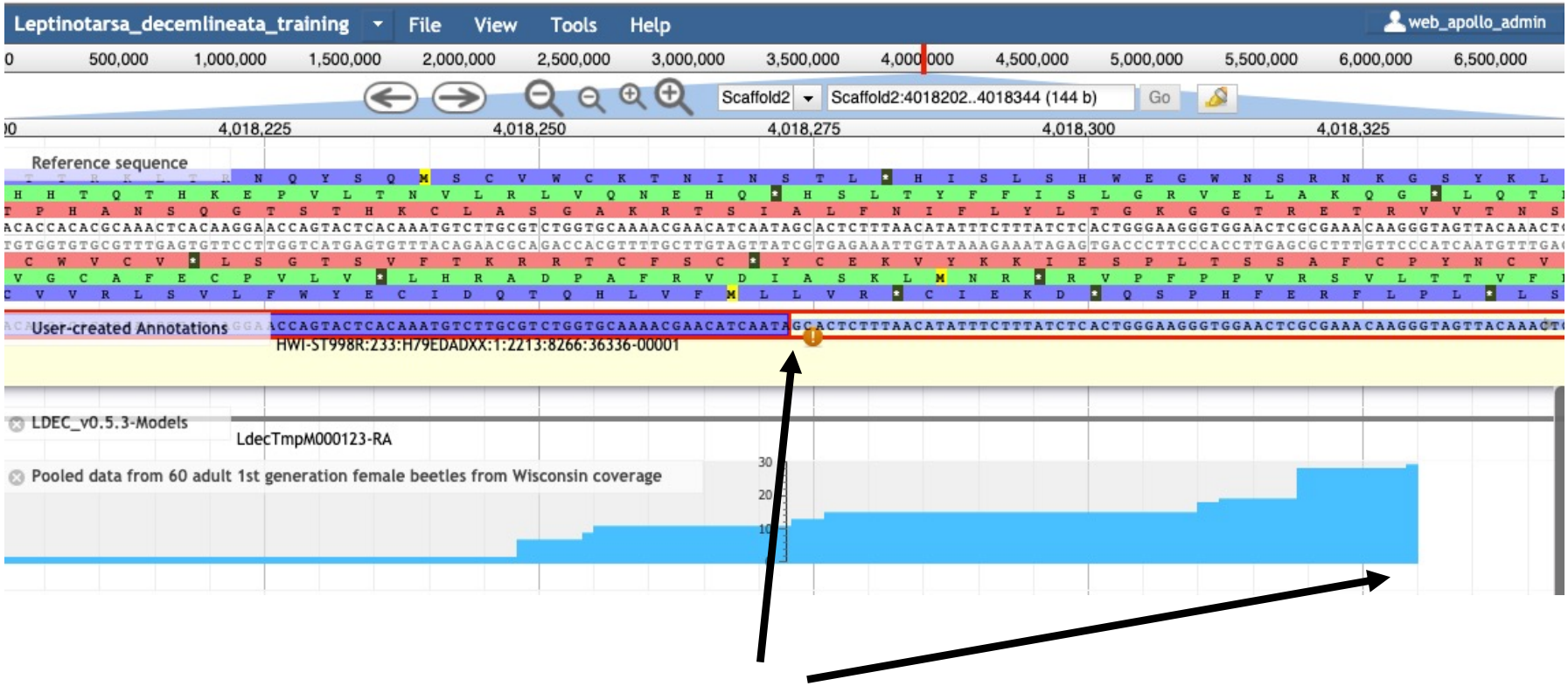
Merge the new 5’ exon with the model

Fixing non-canonical splice sites



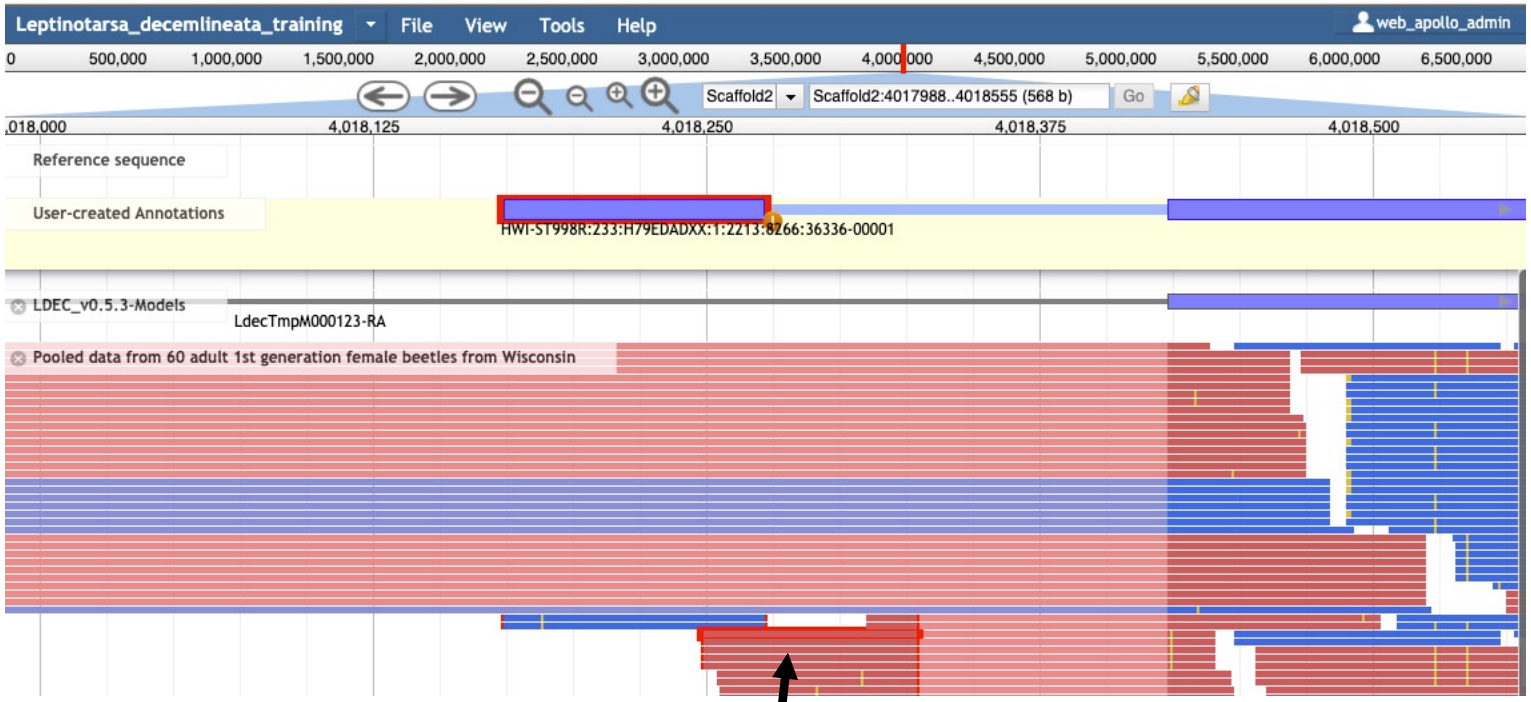
Now that the models are merged, Apollo shows us a non-canonical splice site

Fixing non-canonical splice sites



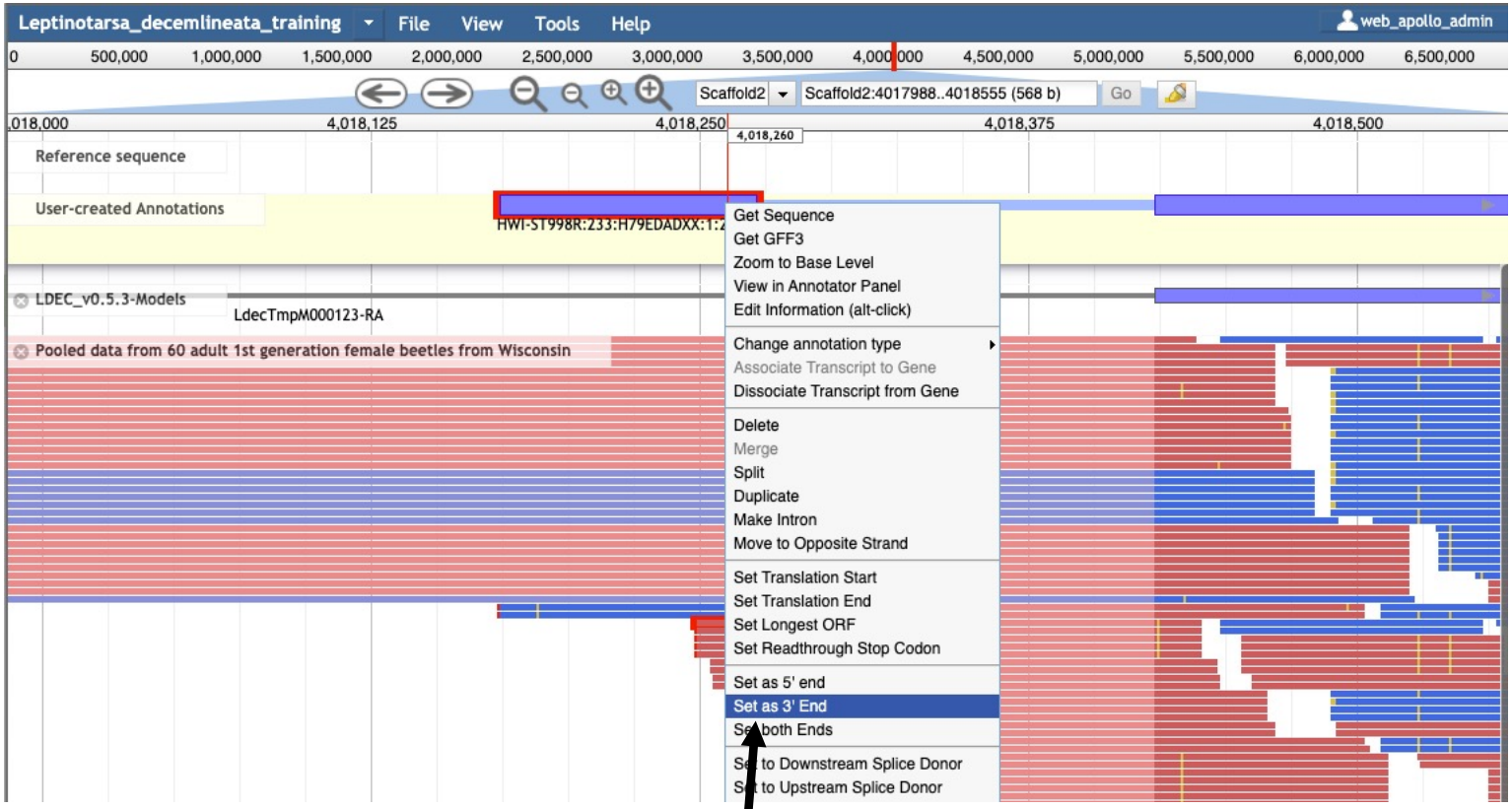
- Canonical donor splice sites should be G(T/U) – this is GC
- The coverage track shows us that the exon needs to extend further to the right

Fixing non-canonical splice sites



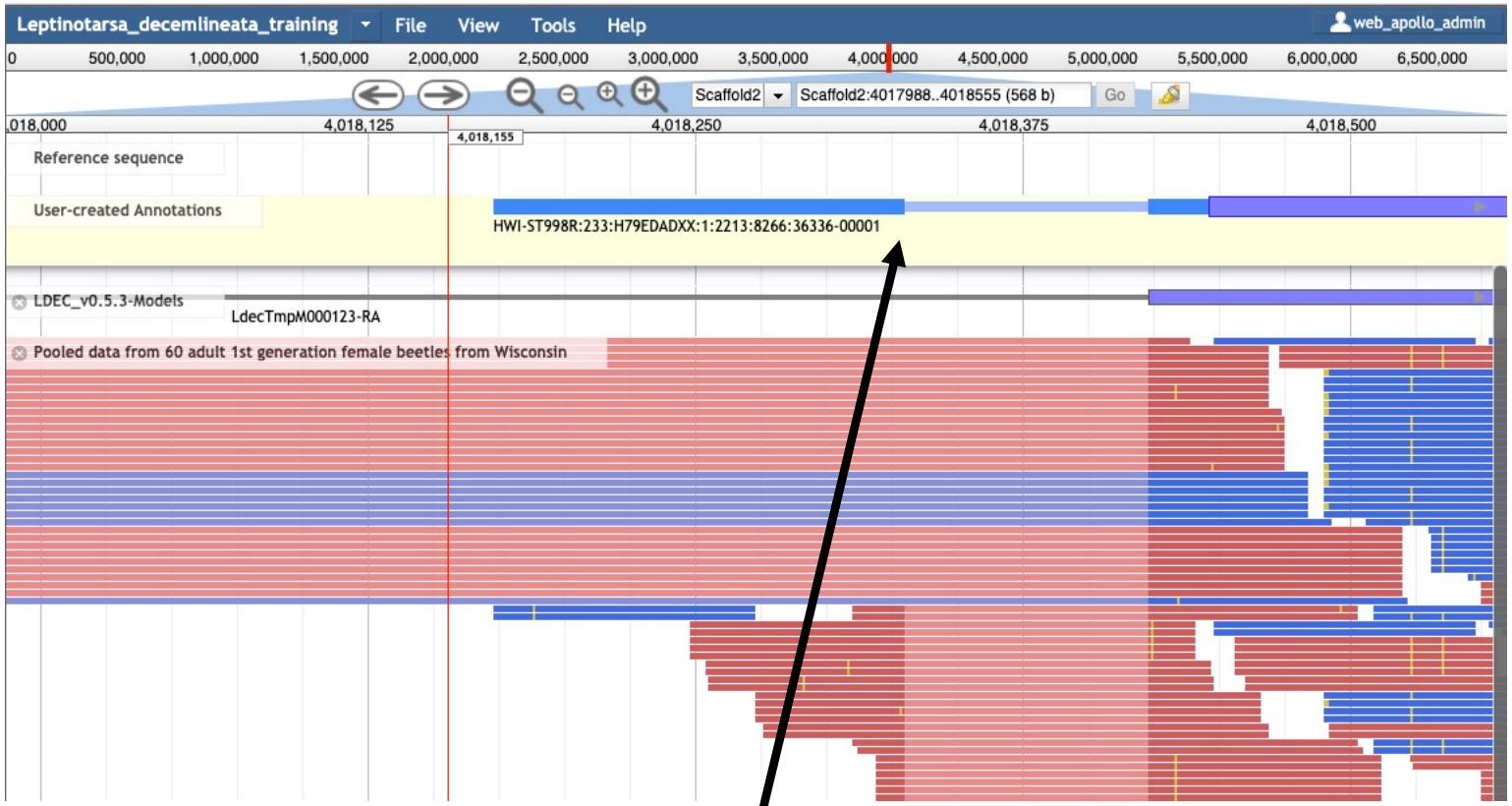
Let’s use another RNA-Seq read to extend the exon to the actual splice site

Fixing non-canonical splice sites



- Shift-click on the model's 5' exon and the RNA-Seq read
  - right-click to open menu
    - Select 'Set as 3' end'

Fixing non-canonical splice sites



Fixed.

Thank you!

The NAL Team

- Chris Childers
- Vern Chapman
- Ming Chan
- Sean Buehler
- Amanda Cooksey

- 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.Childers2@usda.gov

USDA