

Using Apollo at the i5k Workspace@NAL

Monica Poelchau, USDA-ARS NAL

August 18th, 2020



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
- Sequence alterations and stop-codon readthroughs
- Annotating Non-coding features

Other resources

- An additional Apollo webinar with more background:
<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:
 - <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 Apollo training webinar:
<https://eupathdb.org/eupathdb/webinars.jsp#apollo>

Basic RNA-Seq evaluation

RNA-Seq tracks

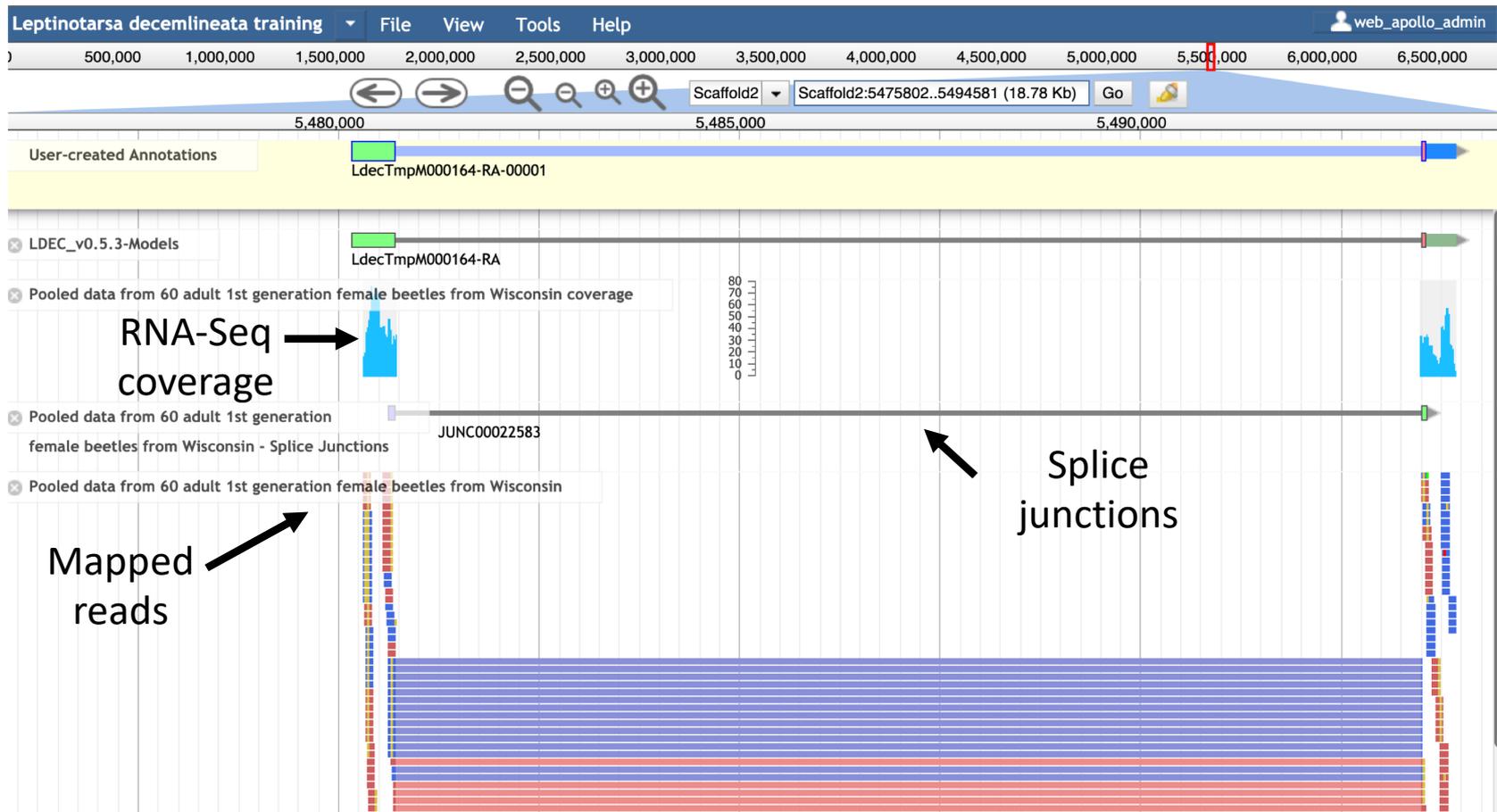
- **Coverage plots:** Histogram of the number of mappings at each nucleotide; hover over the blue area to see the value
- **Mapped reads:** Individual glyphs of each mapped read. Show mapped and spliced areas, and SNPs/indels. Informative, but hard to work with when zoomed out.
- **Junction reads:** Useful combined with coverage plots; show where mapped reads are spliced. Control-click on read and look under 'score' to see how many mapped reads support the splice junction.

The screenshot shows the 'Available Tracks' panel in a genome browser. It is organized into several sections:

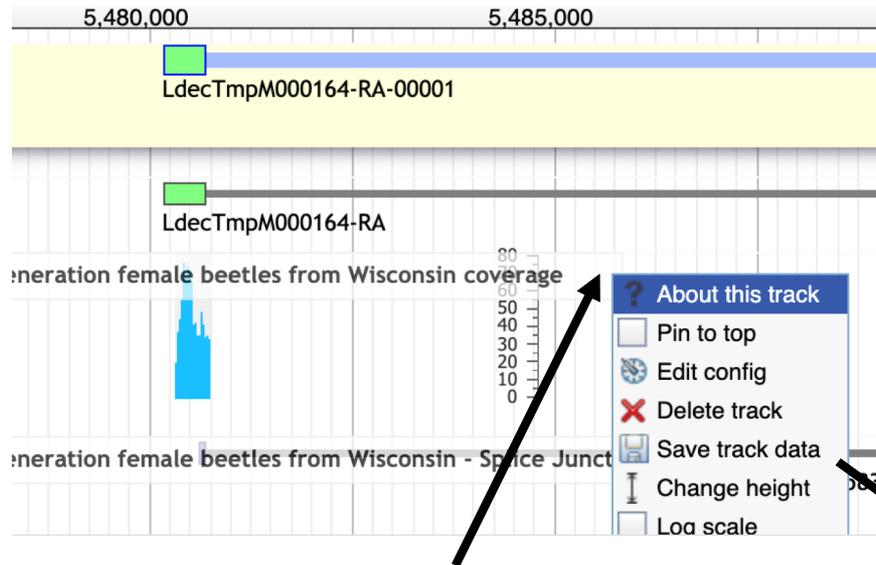
- 0. Reference Assembly** (3 tracks):
 - GC Content
 - Gaps in assembly
 - BLAST+ Results
- NCBI Annotation Release 100** (2 tracks):
 - NCBI_Annotation_Release_100_Gene
 - NCBI_Annotation_Release_100_Pseudogene
- RNA-Seq** (25 tracks):
 - Coverage Plots** (8 tracks):
 - Egg, coverage
 - Larva, coverage
 - Male, coverage
 - Non-reproductive Adult Worker, coverage
 - Odobru_Obru_v1_RNA-Seq-alignments_2020-06-02_coverage
 - Pupa, coverage
 - Queen, coverage
 - Reproductive Adult Worker, coverage
 - Mapped Reads** (8 tracks):
 - Egg
 - Larva
 - Male
 - Non-reproductive Adult Worker
 - Odobru_Obru_v1_RNA-Seq-alignments_2020-06-02
 - Pupa
 - Queen
 - Reproductive Adult Worker
 - Splice junctions** (8 tracks):
 - Egg, junction reads
 - Larva, junction reads
 - Male, junction reads
 - Non-reproductive Adult Worker, junction reads
 - Odobru_Obru_v1_RNA-Seq-alignments_2020-06-02_junctions
 - Pupa, junction reads
 - Queen, junction reads
 - Reproductive Adult Worker, junction reads
- Transcriptome Assembly** (1 track)

Three black arrows point to the 'Coverage Plots', 'Mapped Reads', and 'Splice junctions' sections respectively.

A simple case



A simple case



Select 'About this track' from drop-down menu

Information about methods

About track: Pooled data from 60 adult 1st generation female beetles from Wisconsin coverage

Name Pooled data from 60 adult 1st generation female beetles from Wisconsin coverage

Publication status Unpublished - please follow Toronto/Ft. Lauderdale conditions of data re-use.

File provider Justin Clements and Dr. Sean Schoville UW Madison

Data provider Justin Clements

Sequencing platform Illumina Hi-seq 200 bp

Alignment method Tophat2

Data source NA

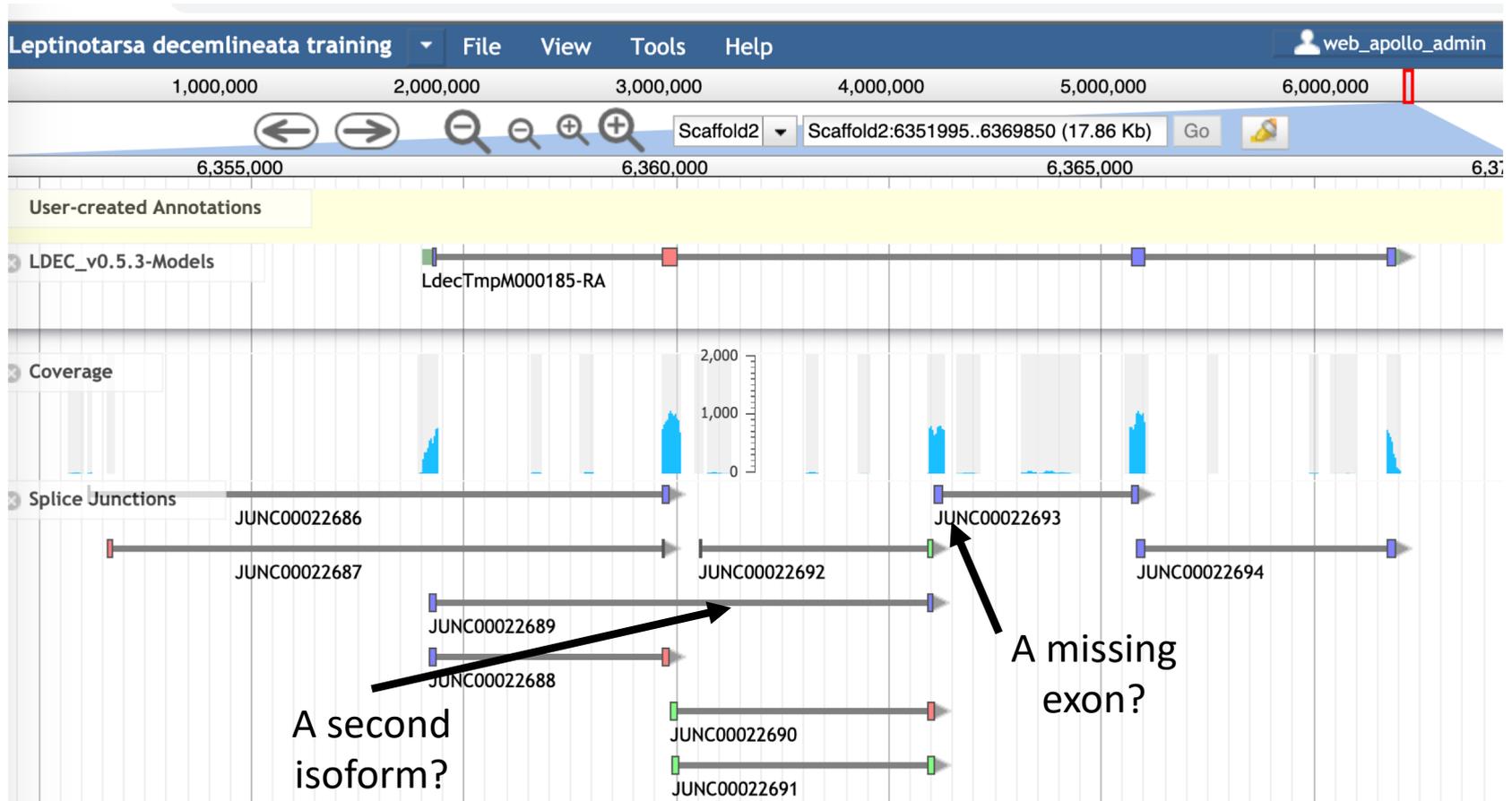
Track type JBrowse/View/Track/Wiggle/XYPlot

Category Transcriptome/Coverage plots (BigWig)

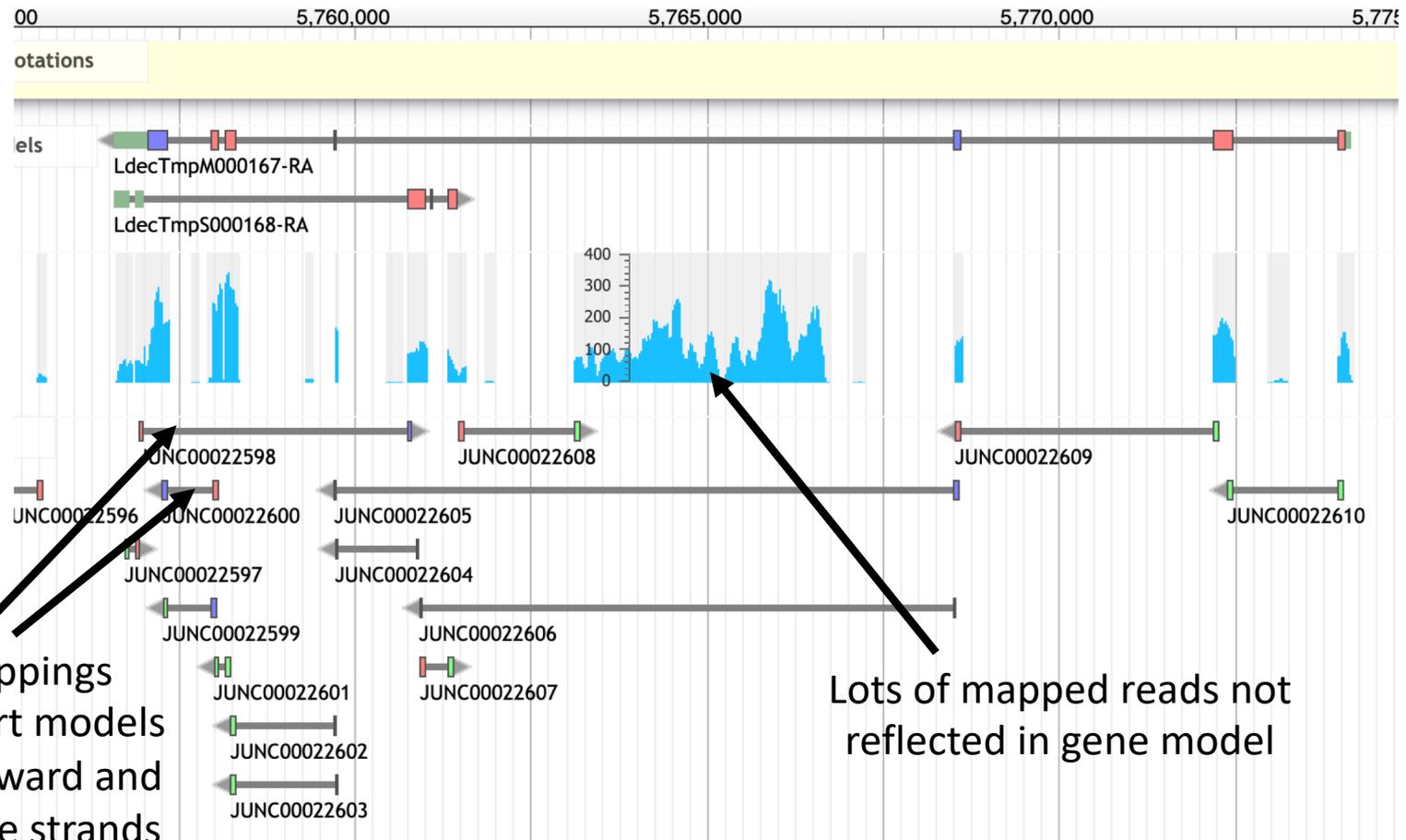
Stats (current reference sequence) (7)

Name	Value
basesCovered	61640461
scoreMax	6163708
scoreMean	249.53225598036977
scoreMin	1

A more complex case



A really messy case

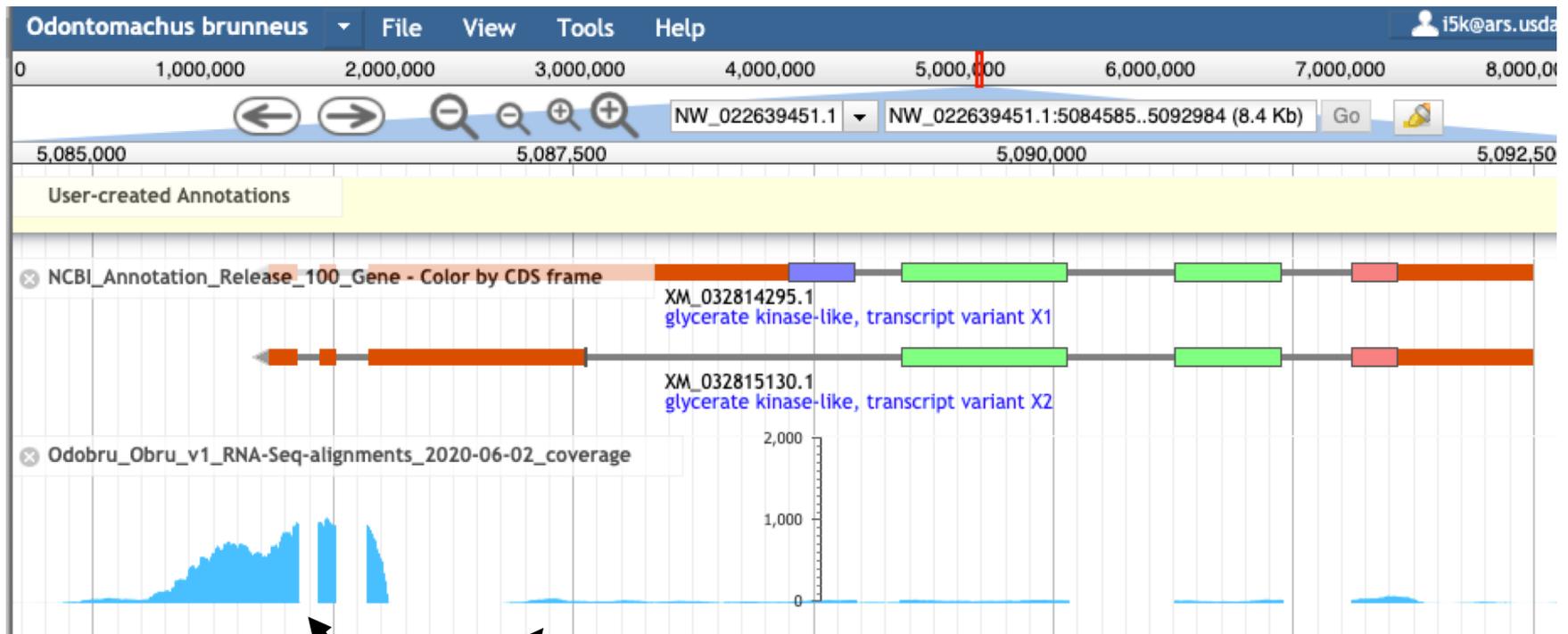


Basic structural changes –
splitting and merging a
model, adding and
removing exons

Annotation Example

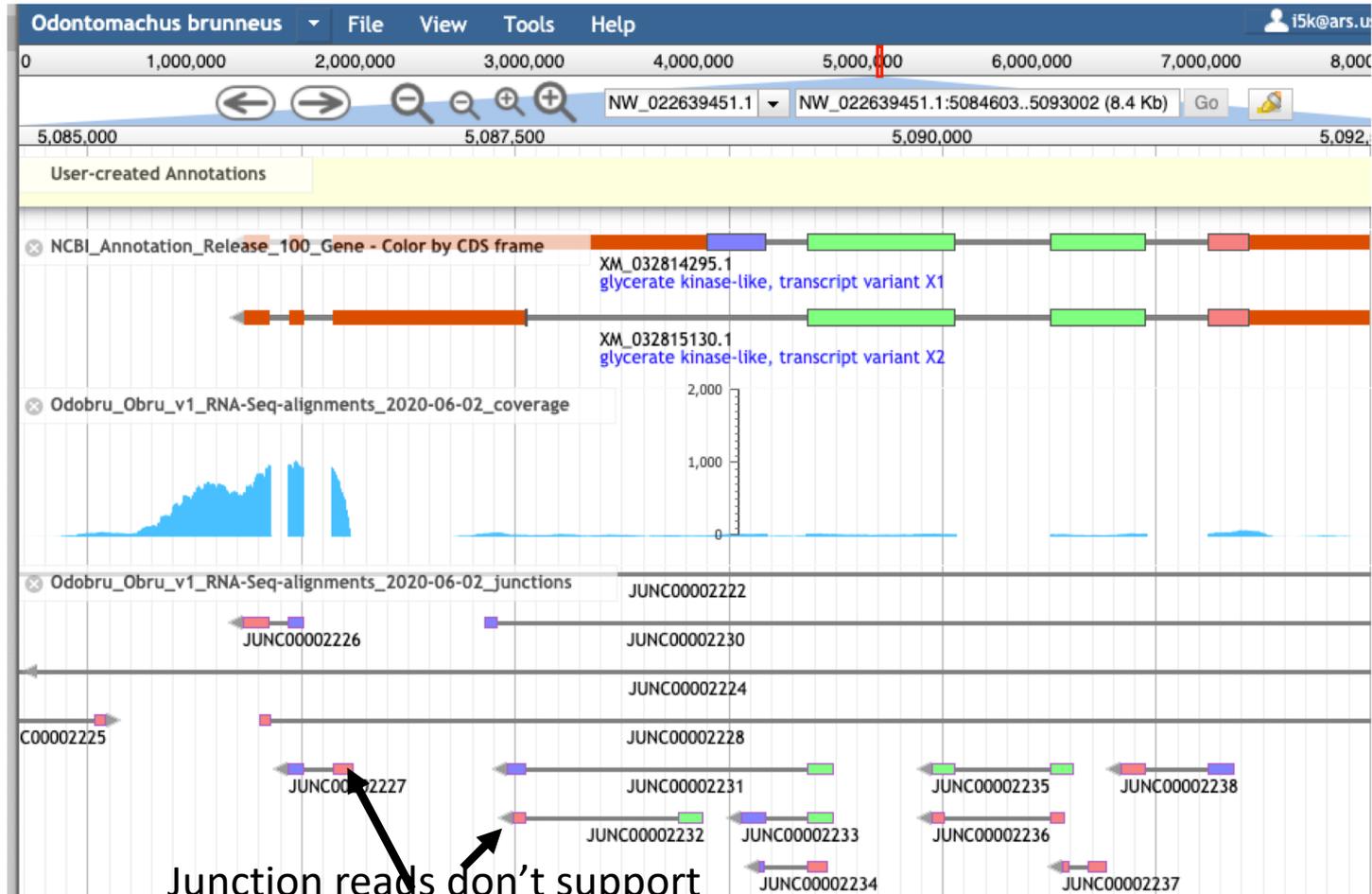
- Glycerate kinase-like in the trap-jaw ant *Odontomachus brunneus*
- More information about the trap-jaw ant genome assembly: <https://i5k.nal.usda.gov/odontomachus-brunneus>
- *Odontomachus brunneus* Apollo URL: https://apollo.nal.usda.gov/apollo/4006447/jbrowse/index.html?loc=NW_022639451.1%3A5084490..5093717&tracks=DNA%2CAnnotations%2CNCBI_Annotation_Release_100_Gene-CBT&highlight=

RNA-Seq evaluation



Very different coverage
between UTR and CDS

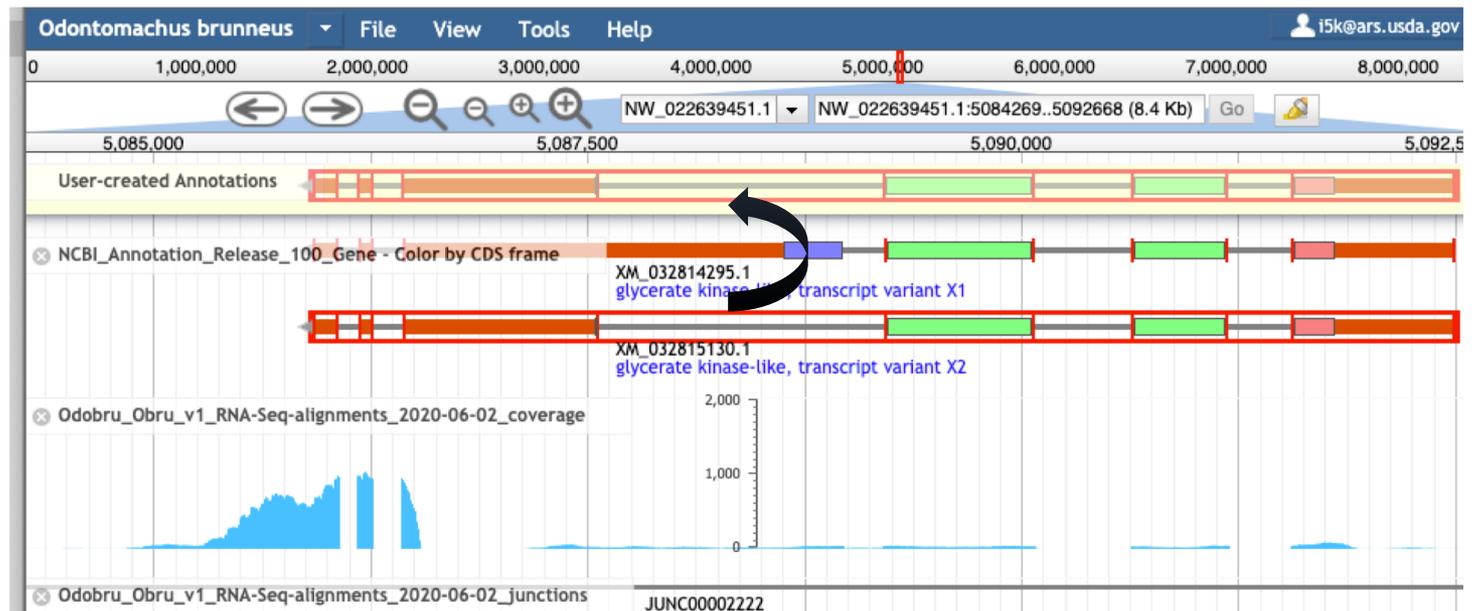
RNA-Seq evaluation



Junction reads don't support
connection between the two
expressed regions

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

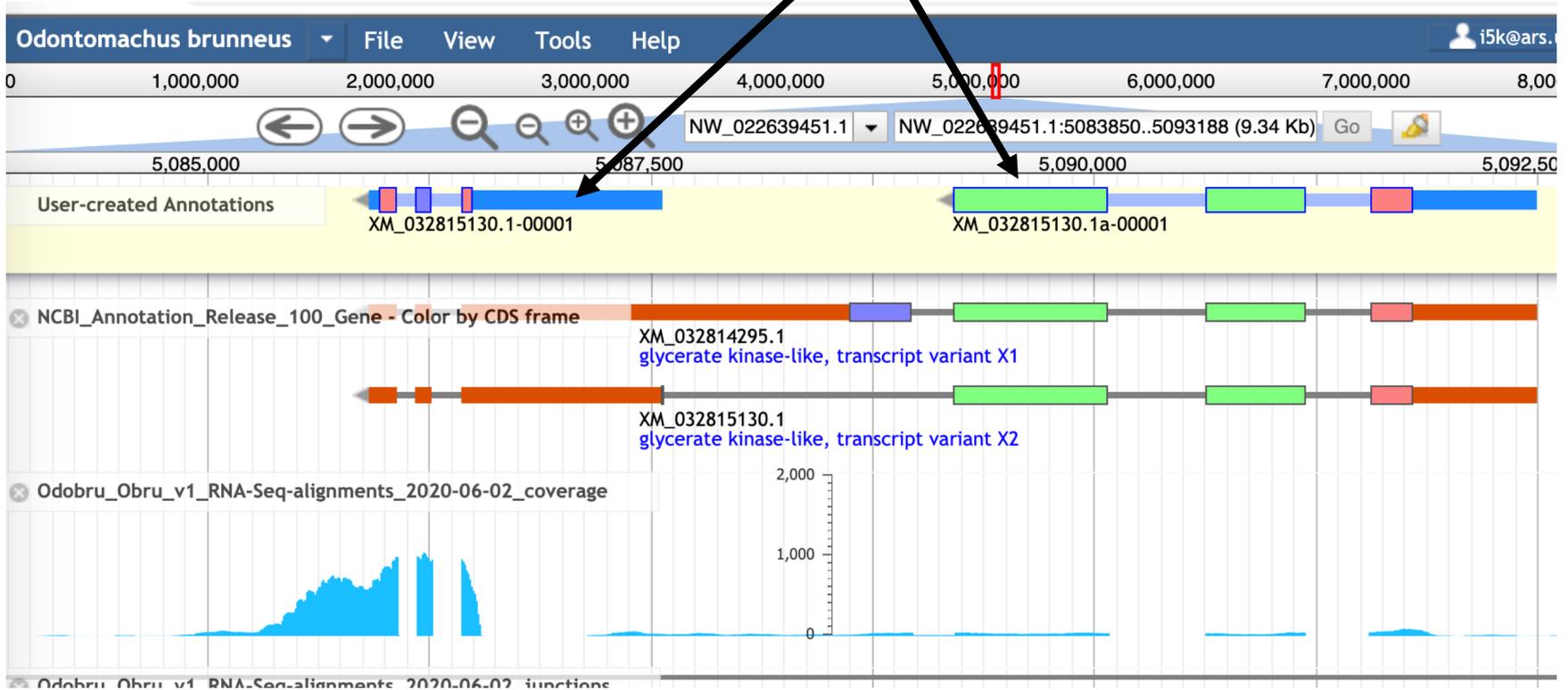
Right-click on the model while continuing to hold shift to get the drop-down menu

The screenshot displays the genome browser interface for *Odontomachus brunneus*. The top menu bar includes 'File', 'View', 'Tools', and 'Help'. The main view shows genomic coordinates from 1,000,000 to 8,000,000. A specific region is highlighted with a red vertical line at approximately 5,090,000. Below the coordinates, there are several tracks: 'User-created Annotations' showing a blue bar for XM_032815130.1-00001; 'NCBI_Annotation_Release_100_Gene - Color by CDS frame' showing orange and green bars for XM_032814295.1 (glycerate kinase-like, transcript variant X1) and XM_032815130.1 (glycerate kinase-like, transcript variant X2); and 'Odobru_Obru_v1_RNA-Seq-alignments_2020-06-02_coverage' showing a blue histogram. A context menu is open over the transcript model, listing options such as 'Get Sequence', 'Get GFF3', 'Zoom to Base Level', 'View in Annotator Panel', 'Edit Information (alt-click)', 'Change annotation type', 'Associate Transcript to Gene', 'Dissociate Transcript from Gene', 'Delete', 'Merge', 'Split', 'Duplicate', 'Make Intron', 'Move to Opposite Strand', 'Set Translation Start', and 'Set Translation End'. The 'Split' option is highlighted in blue.

Select 'split'

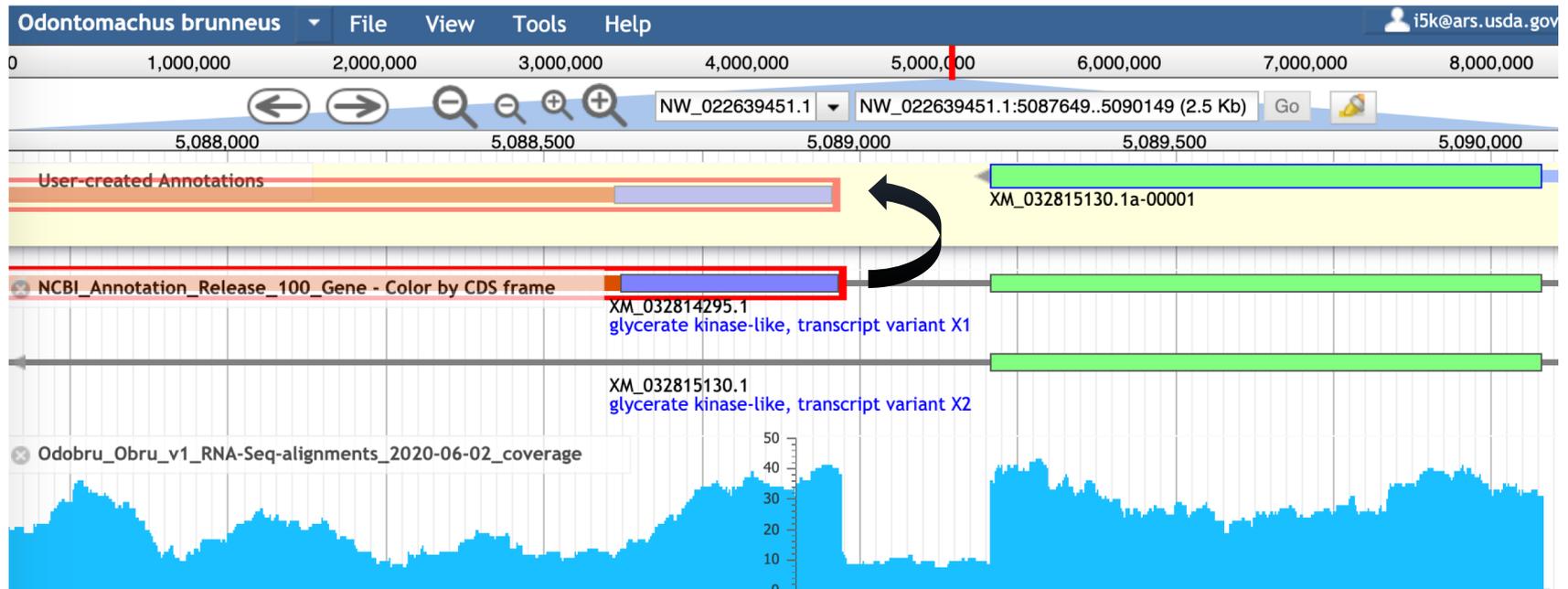
Split model

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

The screenshot displays the Genes and Genomes browser interface for *Odontomachus brunneus*. The top navigation bar includes 'File', 'View', 'Tools', and 'Help'. The main view shows a genomic track with coordinates from 1,000,000 to 8,000,000. A specific region is zoomed in, showing coordinates from 5,088,000 to 5,090,000. The track includes 'User-created Annotations', 'NCBI_Annotation_Release_100_Gene - Color by CDS frame', and 'Odobru_Obru_v1_RNA-Seq-alignments_2020-06-02_coverage'. Two exons are highlighted with red boxes: one from transcript XM_032814295.1 (blue) and another from transcript XM_032815130.1 (green). A context menu is open over the green exon, with the 'Merge' option selected. The menu items include: Get Sequence, Get GFF3, Zoom to Base Level, View in Annotator Panel, Edit Information (alt-click), Change annotation type, Associate Transcript to Gene, Dissociate Transcript from Gene, Delete, Merge, Split, Duplicate, Make Intron, Move to Opposite Strand, Set Translation Start, and Set Translation End.

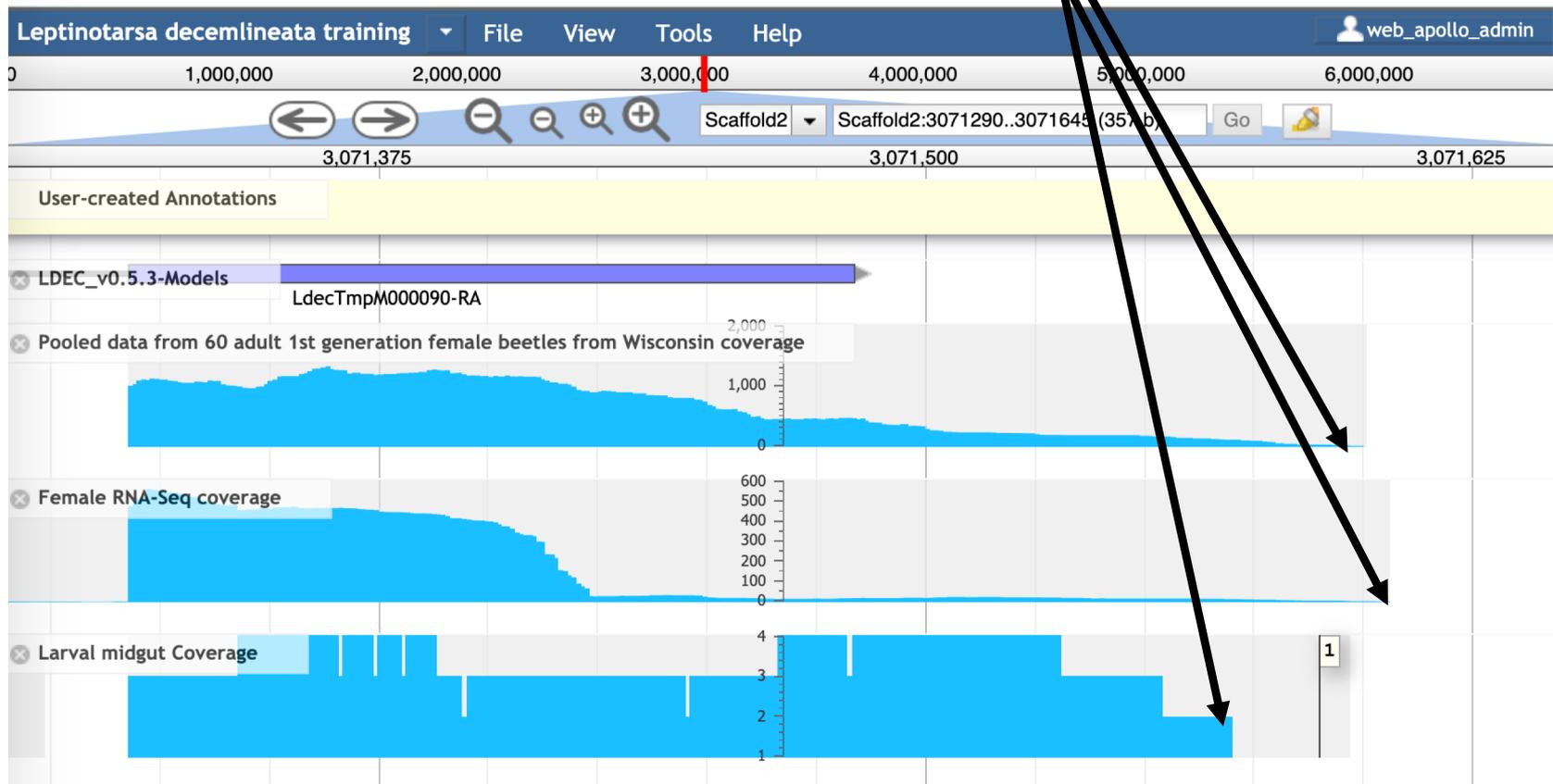
UTRs – how and when to
add or adjust

Adding or adjusting UTR boundaries

- When should you add or change UTRs?
 - Only if you have RNA-Seq evidence with sufficient coverage (e.g. > 50 reads)
 - 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

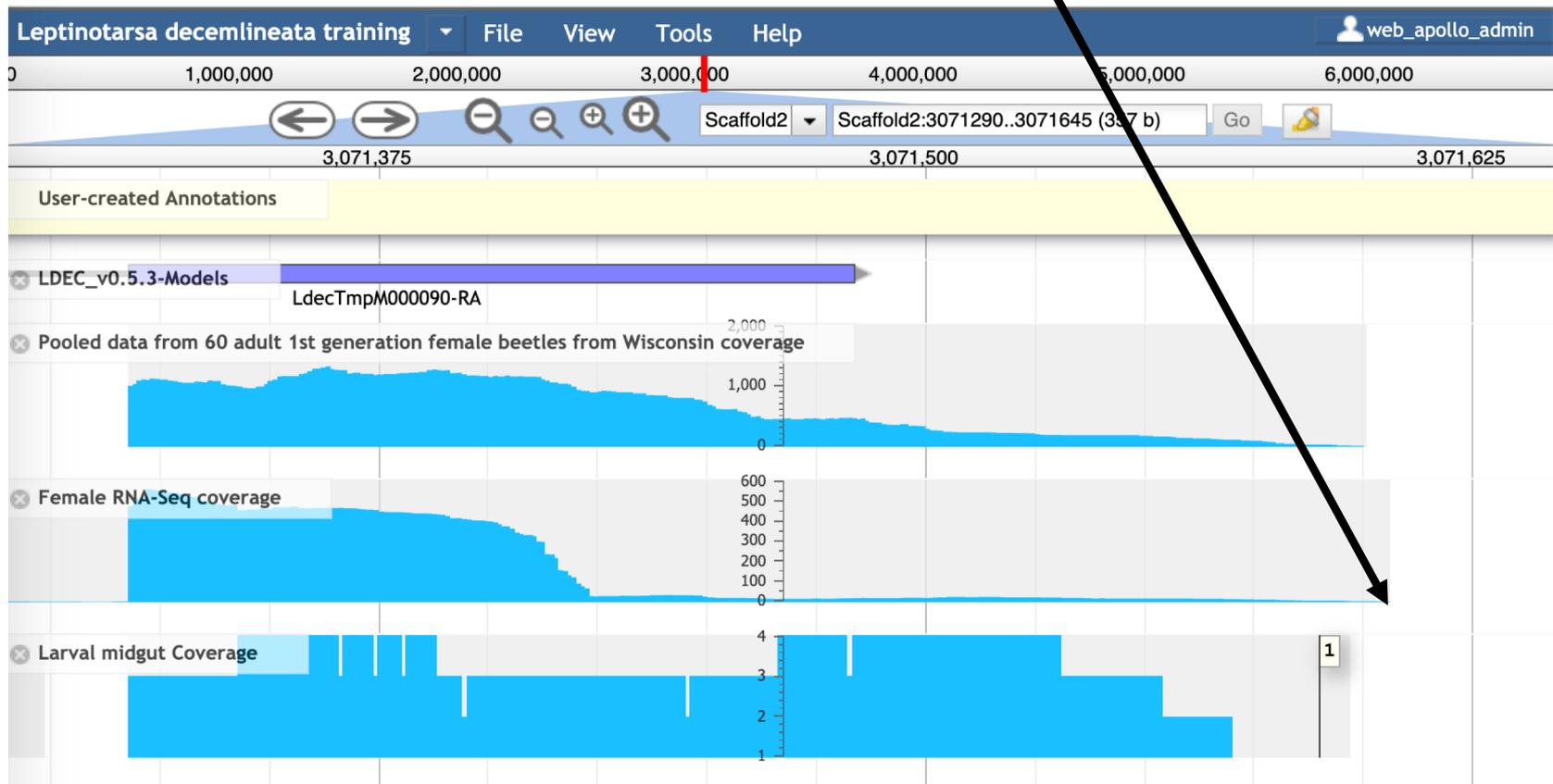
Adjusting gene boundaries

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

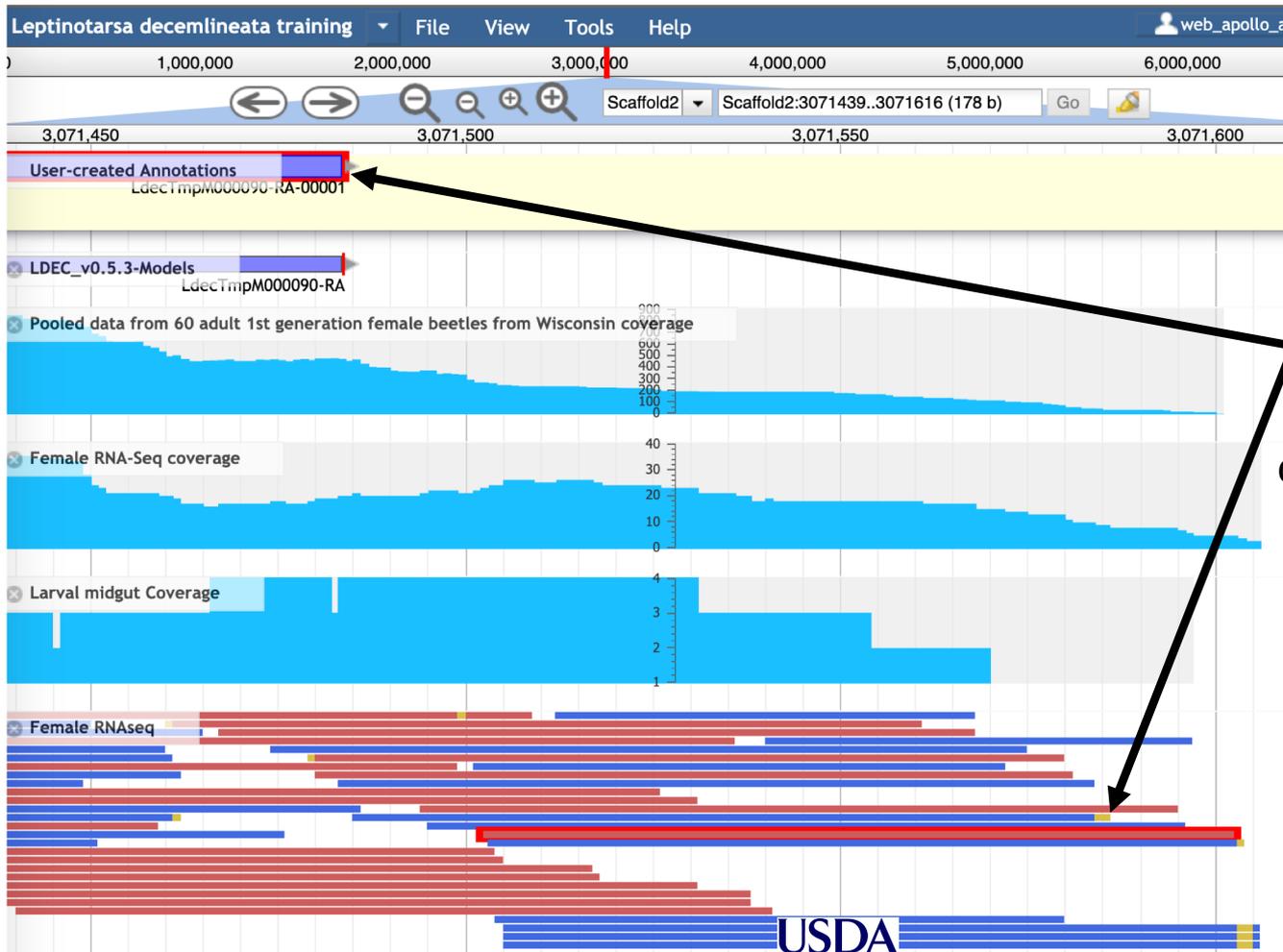


Adjusting gene boundaries

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



Adjusting gene boundaries



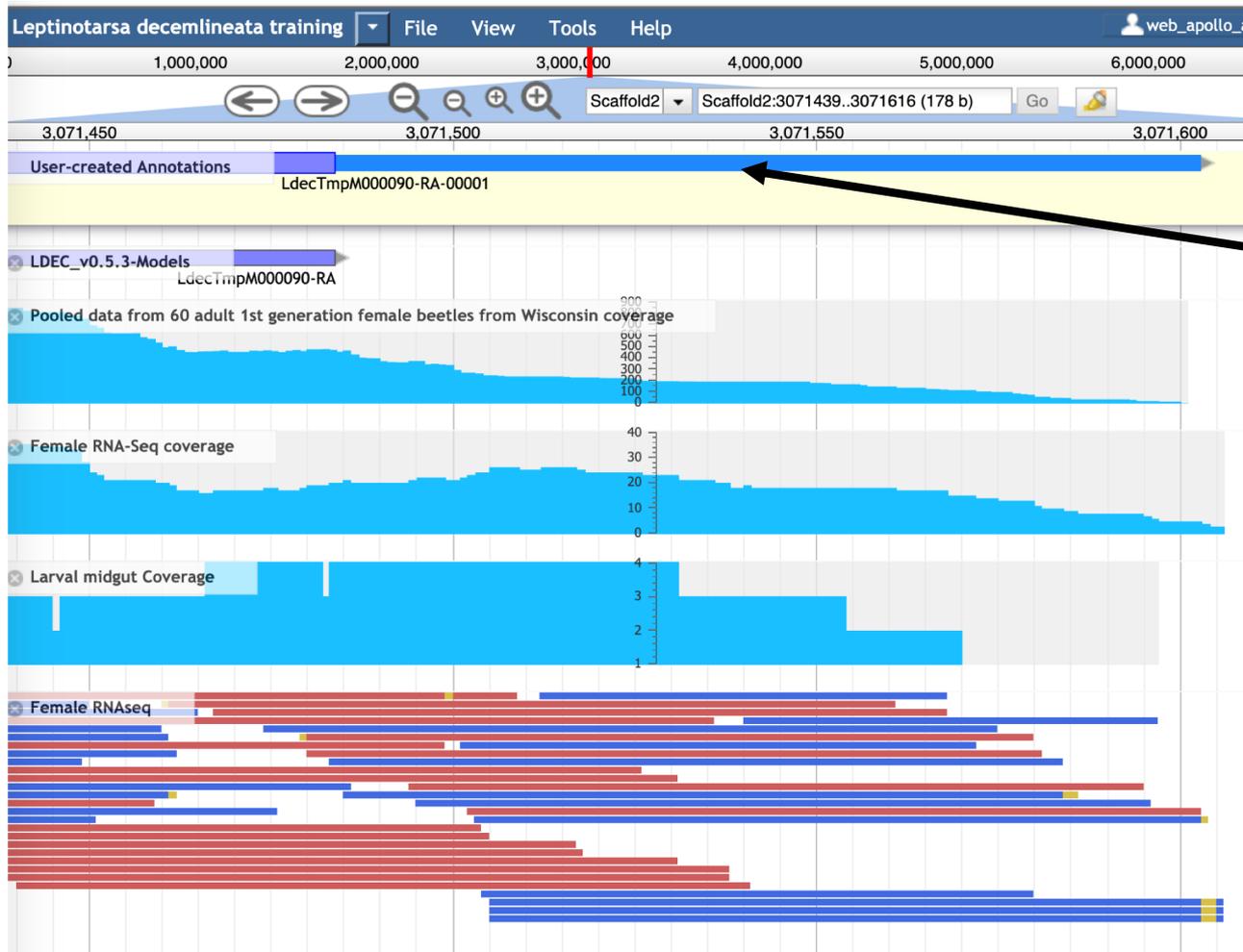
One way to change the boundary: find a mapped read on the same strand as the model; hold shift and click on the read and the model to highlight them both

Adjusting gene boundaries

The screenshot displays the Apollo genome browser interface for *Leptinotarsa decemlineata*. The top navigation bar includes 'File', 'View', 'Tools', and 'Help'. The main view shows a genomic track for Scaffold2, with coordinates ranging from 1,000,000 to 6,000,000. A specific region is zoomed in, showing coordinates from 3,071,450 to 3,071,600. The 'User-created Annotations' track is highlighted in yellow, and a right-click context menu is open over it. The menu options include: Get Sequence, Get GFF3, Zoom to Base Level, View in Annotator Panel, Edit Information (alt-click), Change annotation type, Associate Transcript to Gene, Dissociate Transcript from Gene, Delete, Merge, Split, Duplicate, Make Intron, Move to Opposite Strand, Set Translation Start, Set Translation End, Set Longest ORF, Set Readthrough Stop Codon, Set as 5' end, Set as 3' End (highlighted), Set both Ends, Set to Downstream Splice Donor, Set to Upstream Splice Donor, Set to Downstream Splice Acceptor, Set to Upstream Splice Acceptor, Undo, Redo, and Show History. A black arrow points from the text on the right to the 'Set as 3' End' option in the menu.

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

The screenshot displays the Apollo genome browser interface for *Leptinotarsa decemlineata*. Two 'Information Editor' windows are open, showing the 'Attributes' section with 'Tag' and 'Value' fields, and 'Add' and 'Delete' buttons. The right window also shows 'PubMed IDs', 'Gene Ontology IDs', and 'Comments' sections. A comment is added to the 'Comments' field: 'Added 3' UTR based on Female RNA-Seq'. An arrow points from the text 'Add comment explaining UTR addition' to this comment.

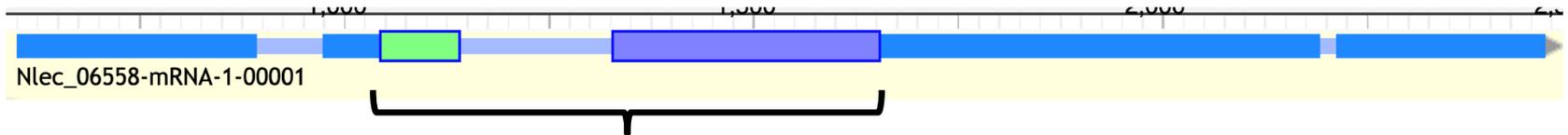
Add comment explaining UTR addition

Starts, stops, open
reading frames

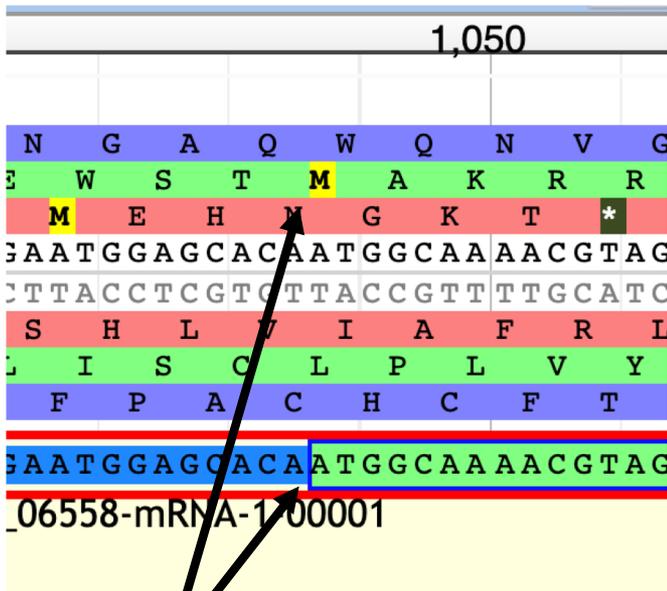
Setting the sequence start, stop, and open reading frame (ORF)

- Apollo will automatically calculate the longest possible ORF that includes canonical 'Start' and 'Stop' signals
(<https://genomearchitect.readthedocs.io/en/latest/UsersGuide.html>)
- However, in some fringe cases, you will need to double-check
- You can change this if needed

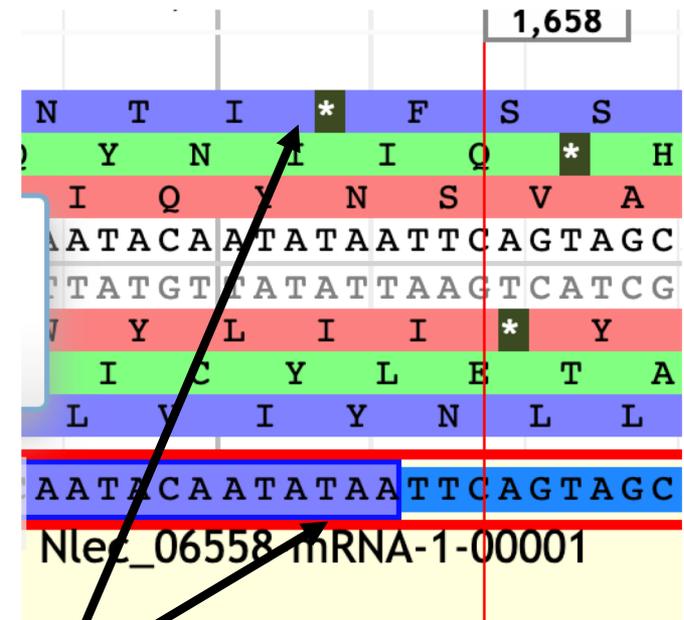
Starts, stops, ORFs



Open reading frame (ORF): translated region

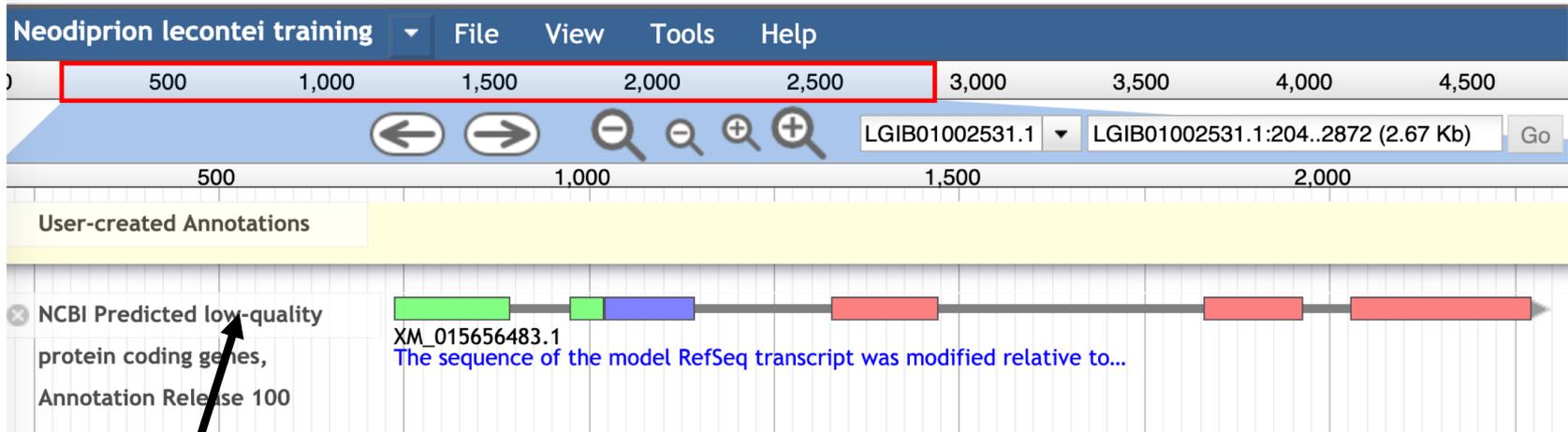


Translation start at Methionine



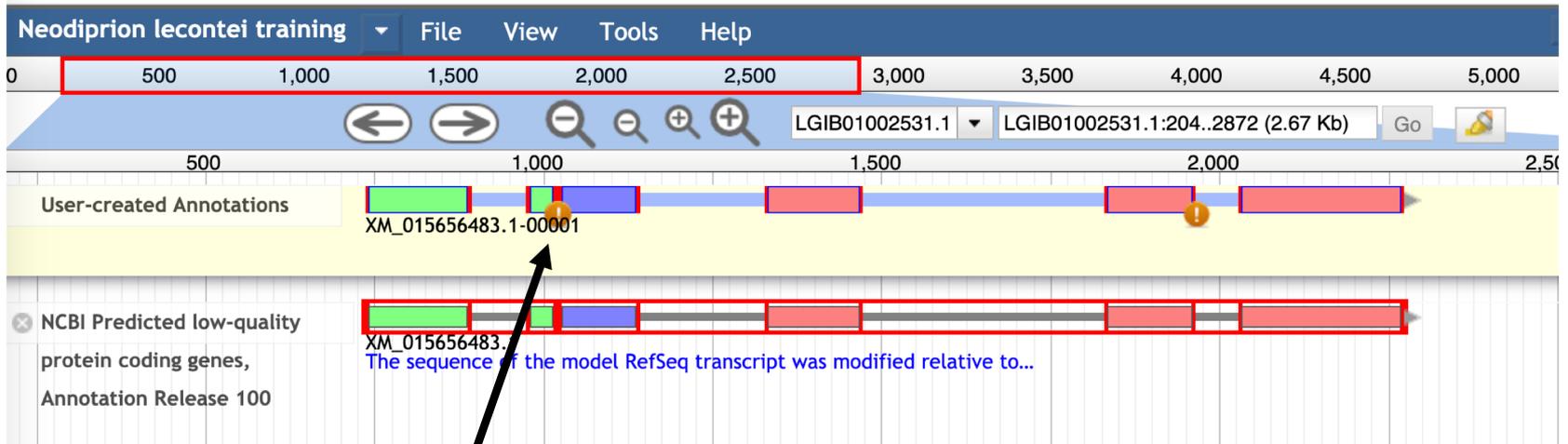
Translation stop at stop codon

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

Neodiprion lecontei training File View Tools Help web_apollo_admin

0 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000

← → 🔍 - +

LGIB01002531.1 LGIB01002531.1:703..832 (131 b) Go

725 750 775 800 825

Reference sequence

V R L N Y Y P R Y L R Y G C Q Y C D T Q I E A R R A V * G C * N I D C L * F M E N V

T P E L I L S T L S Q I R M P I L R Y A N * G P E S C I R L L E Y * L P L I H G K C

Y A * I D I I H A I S D T D A N T A I R K L R P G E L Y K V A R I L T A S D S W K M L

GTACGCCTGAATTGATATTATCCACGCTATCTCAGATACGGATGCCAACTACTGCGATACGCAATTGAGGCCCGGAGAGCTGTATAAGGTTGCTAGAAATATTGACTGCCTCTGATTCATGGAAATGT?

CATCGGACTTAACATAAATAGGTGCGATAGAGTCTATGCCTACGGTTATGACGCTATGCCCTAACTCCGGGCCTCTCGACATATTCCAACGATCTTATAACTGACGGAGACTAAGTACCTTTTACA?

L V G S N I N D V S D * I R I G I S R Y F Q P G S L Q I L N S S Y Q S G R I * P F H

Y A Q I S I I W A I E S V S A L V A I T L N L G P S S Y L T A L I N V A E S E H F I

T R R F Q Y * G R * R L Y P H W Y Q S Y C I S A R L A T Y P Q * F I S Q R Q N M S F T

User-created Annotations

ATACGGATGCCAACTACTGCGATACGCAAAATTGAGGCCCGGAGAGCTGTATAAGGTTGCTAGAAATATTGACTGCCTCTGATTCATGGAAATGT?

XM_015656483.1-00001

NCBI Predicted low-quality protein coding genes, Annotation Release 100

XM_015656483.1

The sequence of the model RefSeq transcript was modified relative to...

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

Sequence 750

```
>88802400-725a-4c8b-9ec9-1943fde9749c (sequence:exon) 10 residues [LGIB01002531.1:737-893 + strand]
[peptide]
IRMPILRYAN
```

Sure enough, the protein sequence is suspiciously short

Peptide sequence
 cDNA sequence
 CDS sequence
 Genomic sequence
 Genomic sequence +/- bases

Starts, stops, ORFs

The screenshot shows a genome browser interface for 'Neodiprion lecontei training'. The top menu includes 'File', 'View', 'Tools', and 'Help'. A scale bar at the top indicates positions from 0 to 3,000. The reference sequence is displayed with a pink reading frame starting at position 739. A context menu is open over the sequence, with the option 'Set Translation Start' highlighted. The menu also includes options like 'Get Sequence', 'Zoom to Base Level', 'View in Annotator Panel', 'Edit Information (alt-click)', 'Change annotation type', 'Associate Transcript to Gene', 'Dissociate Transcript from Gene', 'Delete', 'Merge', 'Split', 'Duplicate', 'Make Intron', 'Move to Opposite Strand', 'Set Translation End', 'Set Longest ORF', 'Set Readthrough Stop Codon', and 'Set as 5' end'.

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

Neodiprion lecontei training | File View Tools Help | web_apollo_admin

0 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000

← → 🔍 🔍

LGIB01002531.1 | LGIB01002531.1:703..832 (131 b) | Go

725 750 775 800 825

Reference sequence

V R L N Y P R Y L R Y G C Q Y C D T Q I E A R R A V * G C * N I D C L * F M E N V

T P E L I L S T L S Q I R M P I L R Y A N * G P E S C I R L L E Y * L P L I H G K C

Y A * I D I I H A I S D T D A N T A I R K L R P G E L Y K V A R I L T A S D S W K M L

G T A C G C C T G A A T T G A T A T T A T C C A C G C T A T C T C A G A T A C G G A T G C C A A T A C T G C G A T A C G C A A A T T G A G G C C C G G A G A G C T G T A T A A G G T T G C T A G A A T A T T G A C T G C C T C T G A T T C A T G G A A A A T G T

C A T G C G G A C T T A A C T A T A A T A G G T G C G A T A G A G T C T A T G C C T A C G G T T A T G A C G C T A T G C G T T T A A C T C C G G G C C T C T C G A C A T A T T C C A A C G A T C T T A T A A C T G A C G G A G A C T A A G T A C C T T T T A C A

L V G S N I N D V S D * I R I G I S R Y A F Q P P G S L Q I L N S S Y Q S G R I * P F H

Y A Q I S I I W A I E S V S A L V A I R L N L G P S S Y L T A L I N V A E S E H F I

T R R F Q Y * G R * R L Y P H W Y Q S V C I S A R L A T Y P Q * F I S Q R Q N M S F T

User-created Annotations

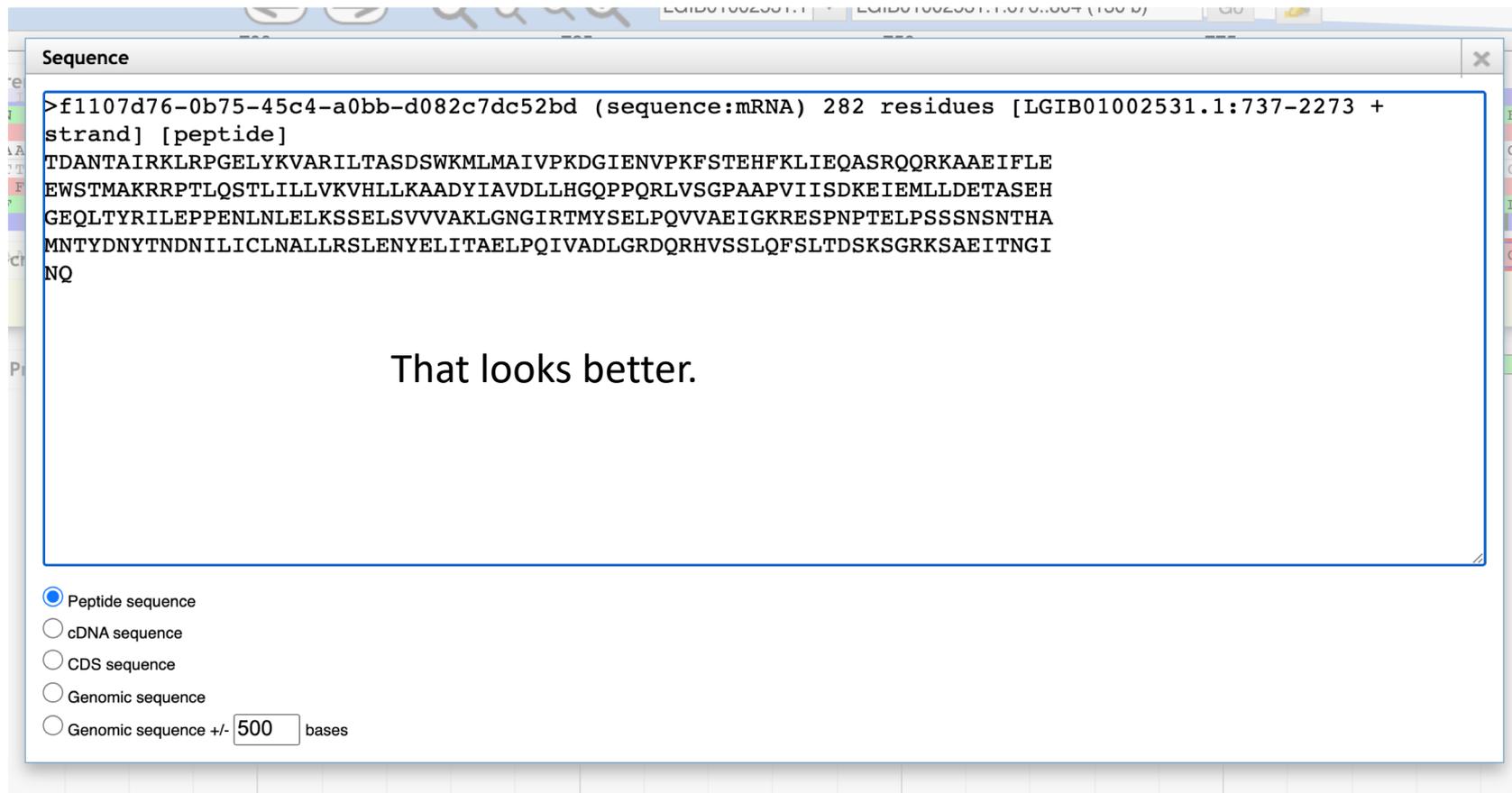
XM_015656483.1-00001

NCBI Predicted low-quality protein coding genes, Annotation Release 100

XM_015656483.1
The sequence of the model RefSeq transcript was modified relative to...

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

Starts, stops, ORFs



The screenshot shows a software window titled "Sequence" with a close button in the top right corner. The main content area contains the following text:

```
>f1107d76-0b75-45c4-a0bb-d082c7dc52bd (sequence:mRNA) 282 residues [LGIB01002531.1:737-2273 + strand] [peptide]  
TDANTAIRKLRPGELYKVARILTASDSWKMLMAIVPKDGIENVPKFSTEHFKLIEQASRQQRKAAEIFLE  
EWS'TMAKRRPTLQSTLILLVKVHLLKAADYIAVDLLHGQPPQRLVSGPAAPVVISDKEIEMLLDETASEH  
GEQLTYRILEPPENLNLELKSSELSVVAKLNGGIRTMYSELPQVVAEIGKRESPNPTELPSSSNSNTHA  
MNTYDNYTNDNILICLNALLRSLENYELITAE LPQIVADLGRDQRHVSSLQFSLTDSKSGRKS AEITNGI  
NQ
```

Below the sequence, there is a radio button menu with the following options:

- Peptide sequence
- cDNA sequence
- CDS sequence
- Genomic sequence
- Genomic sequence +/- bases

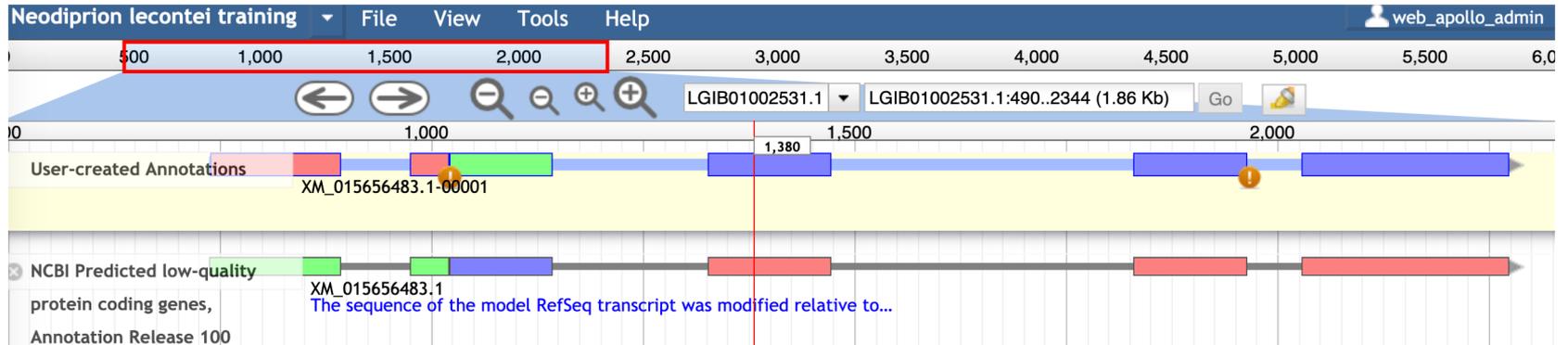
In the center of the main content area, the text "That looks better." is displayed.

Starts, stops, ORFs

The screenshot shows a genome browser interface for 'Neodiprion lecontei training'. The top menu bar includes 'File', 'View', 'Tools', and 'Help'. A scale bar at the top indicates positions from 0 to 3,500. Below the scale, a sequence alignment is shown with a red box highlighting a region between approximately 500 and 2,000. A context menu is open over the alignment, listing various actions such as 'Get Sequence', 'Zoom to Base Level', and 'Set Longest ORF'. The 'Set Longest ORF' option is highlighted in blue. The menu also includes options for 'Delete', 'Merge', 'Split', 'Duplicate', 'Make Intron', 'Move to Opposite Strand', 'Set Translation Start', 'Set Translation End', and 'Set Readthrough Stop Codon'. The background shows 'User-created Annotations' and 'NCBI Predicted low-quality protein coding genes, Annotation Release 100'.

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



Sequence

```
>9190e15d-dcee-45c8-9236-5d7babfca448 (sequence:mRNA) 282 residues [LGIB01002531 strand] [peptide]
TDANTAIRKLRPGELYKVARILTASDSWKMLMAIVPKDGIENVPKFSTEHFKLIEQASRQQRKAAEIFLE
EWSTMAKRRPTLQSTLILLVKVHLLKAADYIAVDLLHGQPPQRLVSGPAAPVIISDKEIEMLLDETASEH
GEQLTYRILEPPENLNLELKSSSELSVVVAKLGNIGRTMYSELPQVVAEIGKRESPNPTELPSSSNSNTHA
MNTYDNYTNDNILICLNALLRSLENYELITAELPQIVADLGRDQRHVSSLQFSLTDSKSGRKSAEITNGI
NQ
```

This also fixed the reading frame.

Starts, stops, ORFs

The screenshot shows a genome browser interface for the Neodiprion lecontei training dataset. The top menu bar includes 'File', 'View', 'Tools', and 'Help'. A scale bar at the top indicates positions from 0 to 4,000. The reference sequence is displayed in a multi-line format with color-coded nucleotides. A context menu is open over a specific nucleotide, listing various actions such as 'Get Sequence', 'Zoom to Base Level', and 'Set Translation End'. The 'Set Translation End' option is highlighted in blue. Below the reference sequence, there are annotations including 'User-created Annotations' and 'NCBI Predicted low-quality protein coding genes'. The 'User-created Annotations' section shows a yellow bar for 'XM_015656483.1-00001'. The 'NCBI Predicted low-quality' section shows a red bar for 'XM_015656483.1' with the description 'The sequence of the model RefSeq'.

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

Neodiprion lecontei training File View Tools Help

0 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500

← → 🔍 - 🔍 + 🔍 + LGIB01002531.1 LGIB01002531.1:2182..2311 (131 b)

2,200 2,225 2,250 2,275

Reference sequence

D Q R H V S S L Q F S L T D S K S G R K S A E I T N G I N Q * V S

S T P R * F F A I Q S H R F K I W K K I C R N Y * R Y * P I G V

I N A T L V L C N S V S P I Q N L E E N L P K L L T V L T N R C L

GATCAACGCCACGTTAGTTCTTTGCAATTCAGTCTCACCGATTCAAAATCTGGAAGAAAATCTGCCGAAATTACTAACGGTATTAACCAATAGGTGTC

CTAGTTGCGGTGCAATCAA GAAACGTTAAGTCAGAGTGGCTAAGTTTTAGACCTTCTTTTAGACGGCTTAAATGATTGCCATAATTGGTTATCCACAGI

P D V G R * N K A I * D * R N L I Q F I Q R F * * L R Y * G I P T

I L A V N T R Q L E T E G I * F R S S F R G F N S V T N V L L H I

S * R W T L E K C N L R V S E F D P L F D A S I V L P I L W Y T D

User-created Annotations

XM_015656483.1-00001

NCBI Predicted low-quality

protein coding genes, Annotation Release 100

XM_015656483.1
The sequence of the model RefSeq transcript was modified relative to...

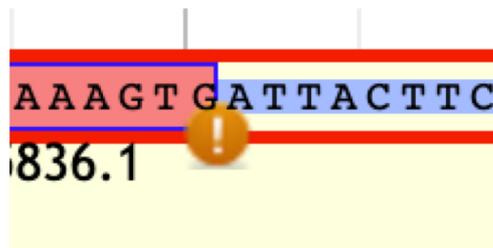
Now the sequence after the translation end is 3' UTR.

Non-canonical splice sites

Splice sites

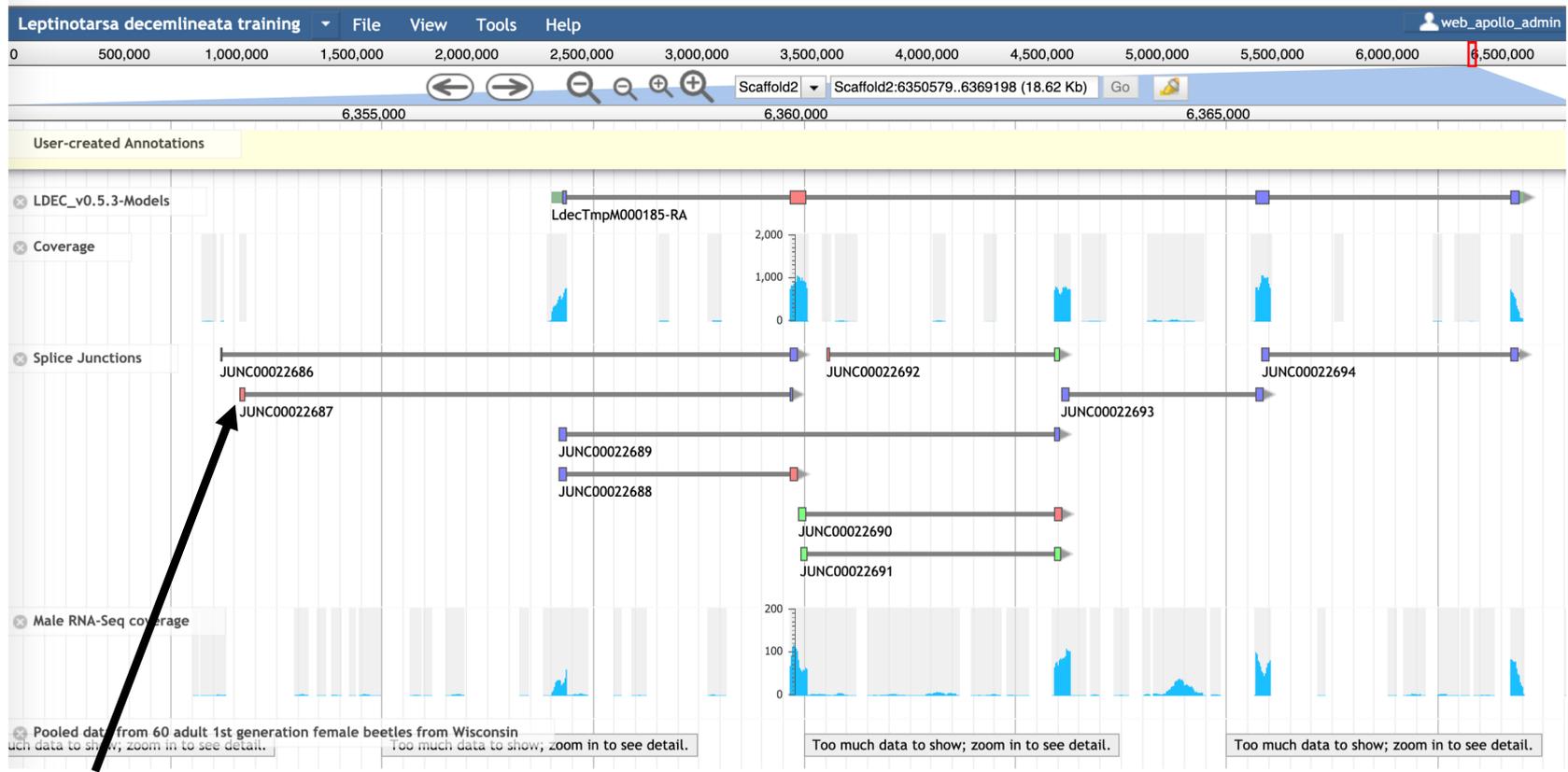
Introns are removed from primary transcripts by cleavage at conserved sequences called **splice sites**. These sites are found at the 5' and 3' ends of introns.

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



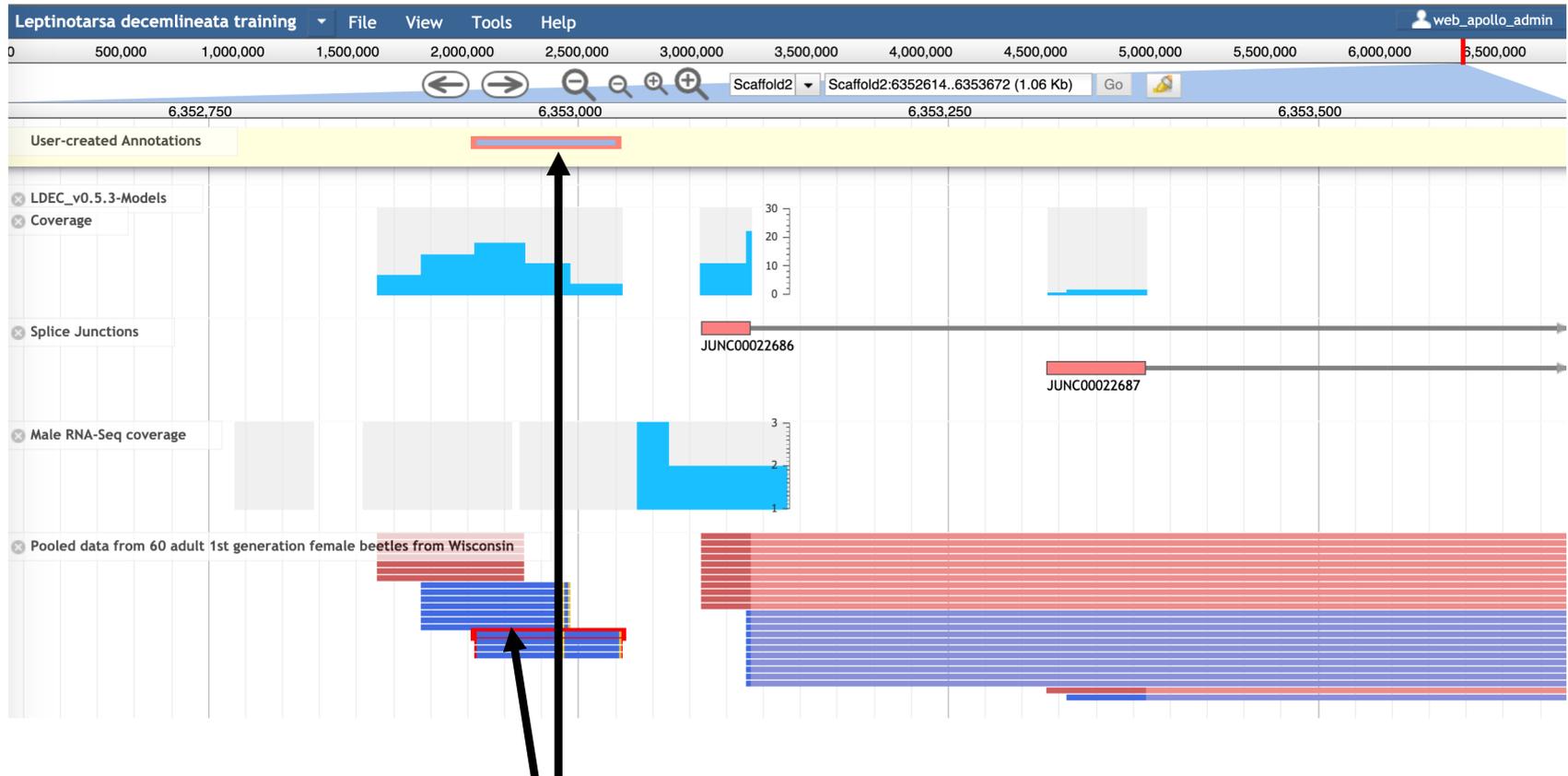
'Non-canonical' splice sites – non-conserved, and possibly erroneous sites – are marked by an exclamation point in Apollo.

Fixing non-canonical splice sites



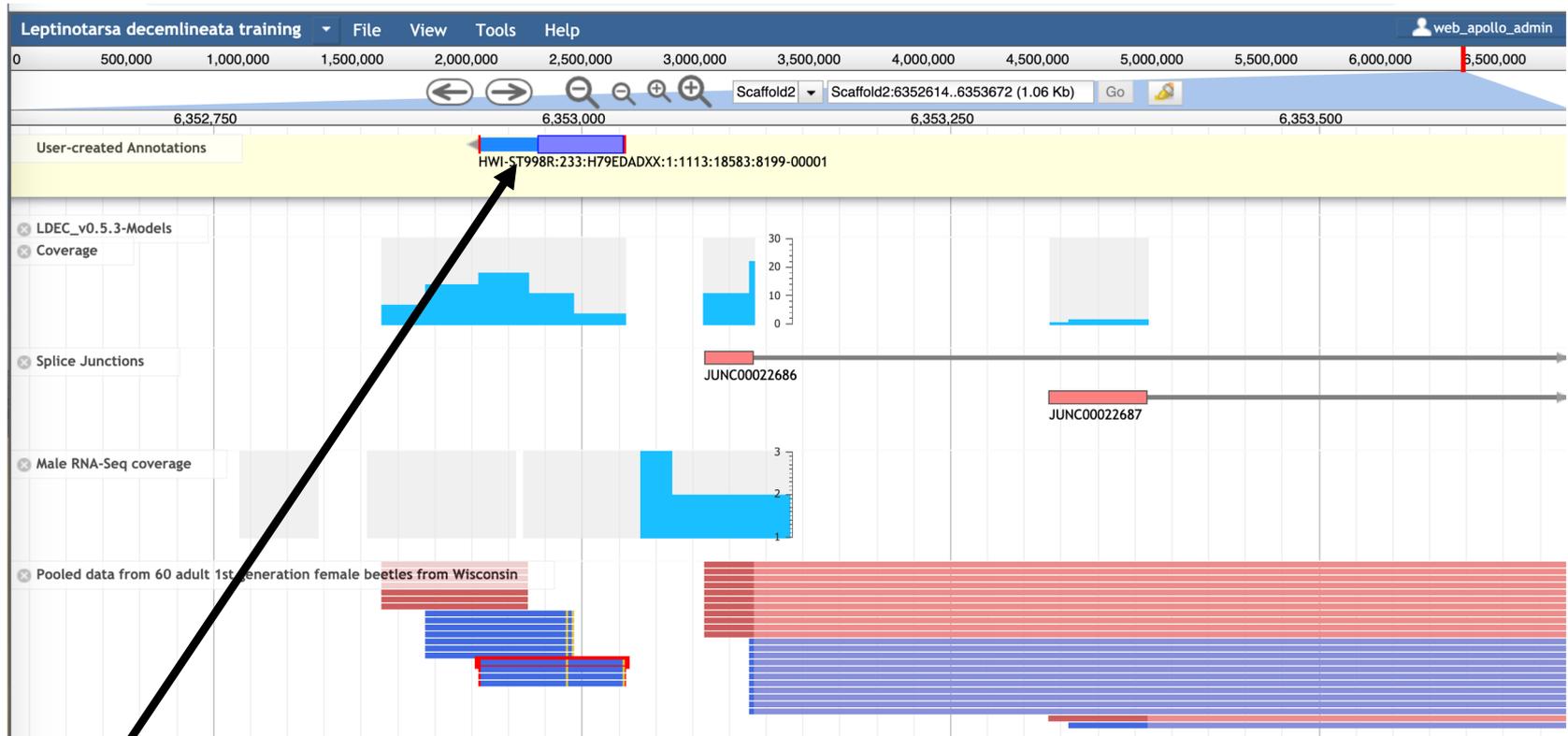
Looks like there's another isoform here – let's add an exon

Fixing non-canonical splice sites



Drag evidence to UcA track to add 5' exon

Fixing non-canonical splice sites



Oops, wrong strand – our model is on the forward strand

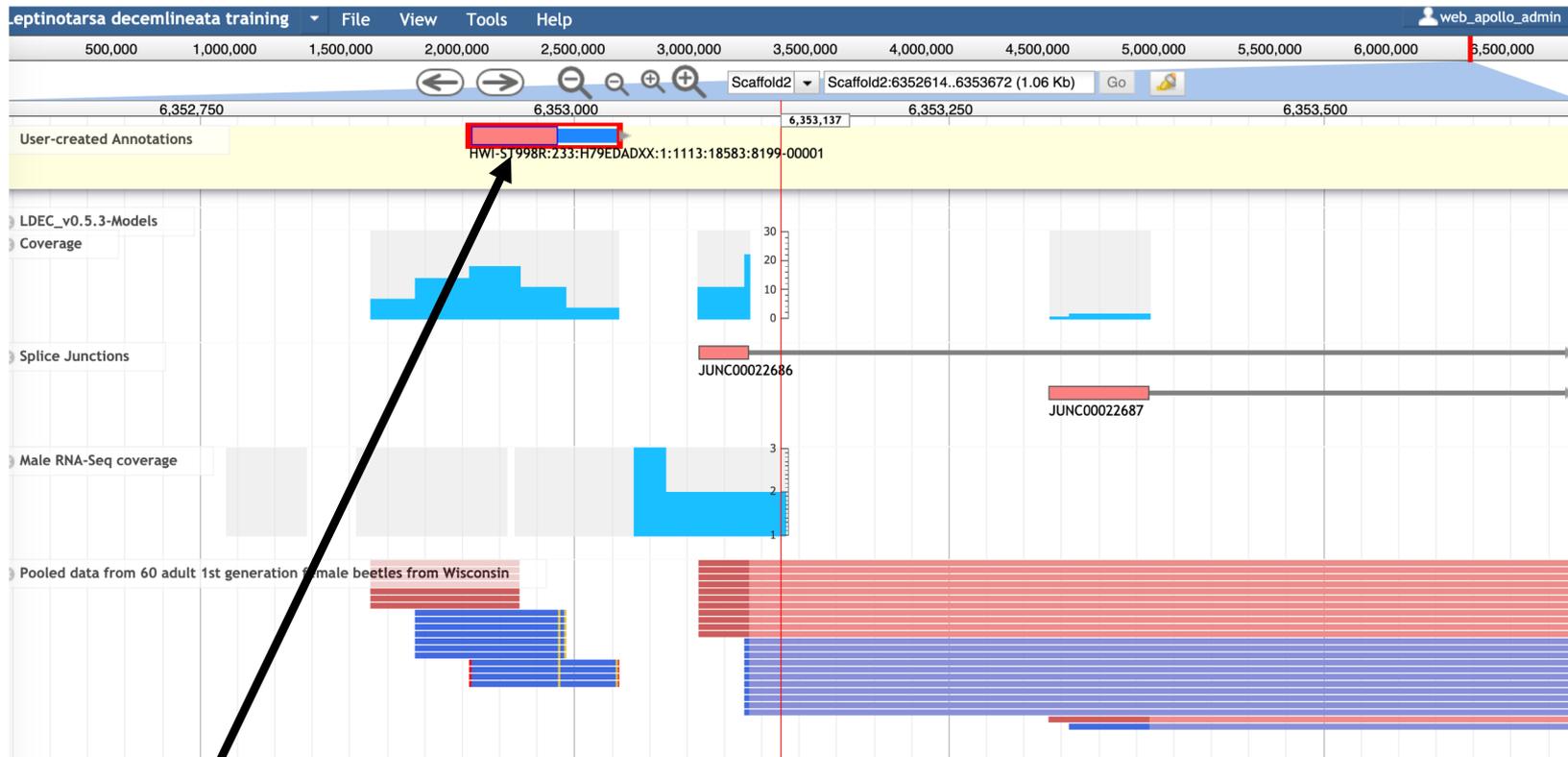
Fixing non-canonical splice sites

The screenshot displays the Leptinotarsa decemlineata training interface. The top menu bar includes 'File', 'View', 'Tools', and 'Help'. The main window shows a genomic track for Scaffold2:6352614..6353672 (1.06 Kb). A context menu is open over a splice junction, with the 'Move to Opposite Strand' option highlighted. The menu options are:

- Get Sequence
- Get GFF3
- Zoom to Base Level
- View in Annotator Panel
- Edit Information (alt-click)
- Change annotation type
 - Associate Transcript to Gene
 - Dissociate Transcript from Gene
- Delete
- Merge
- Split
- Duplicate
- Make Intron
- Move to Opposite Strand**
- Set Translation Start
- Set Translation End
- Set Longest ORF
- Set Readthrough Stop Codon
- Set as 5' end
- Set as 3' End
- Set both Ends
- Set to Downstream Splice Donor
- Set to Upstream Splice Donor
- Set to Downstream Splice Acceptor
- Set to Upstream Splice Acceptor
- Undo
- Redo
- Show History

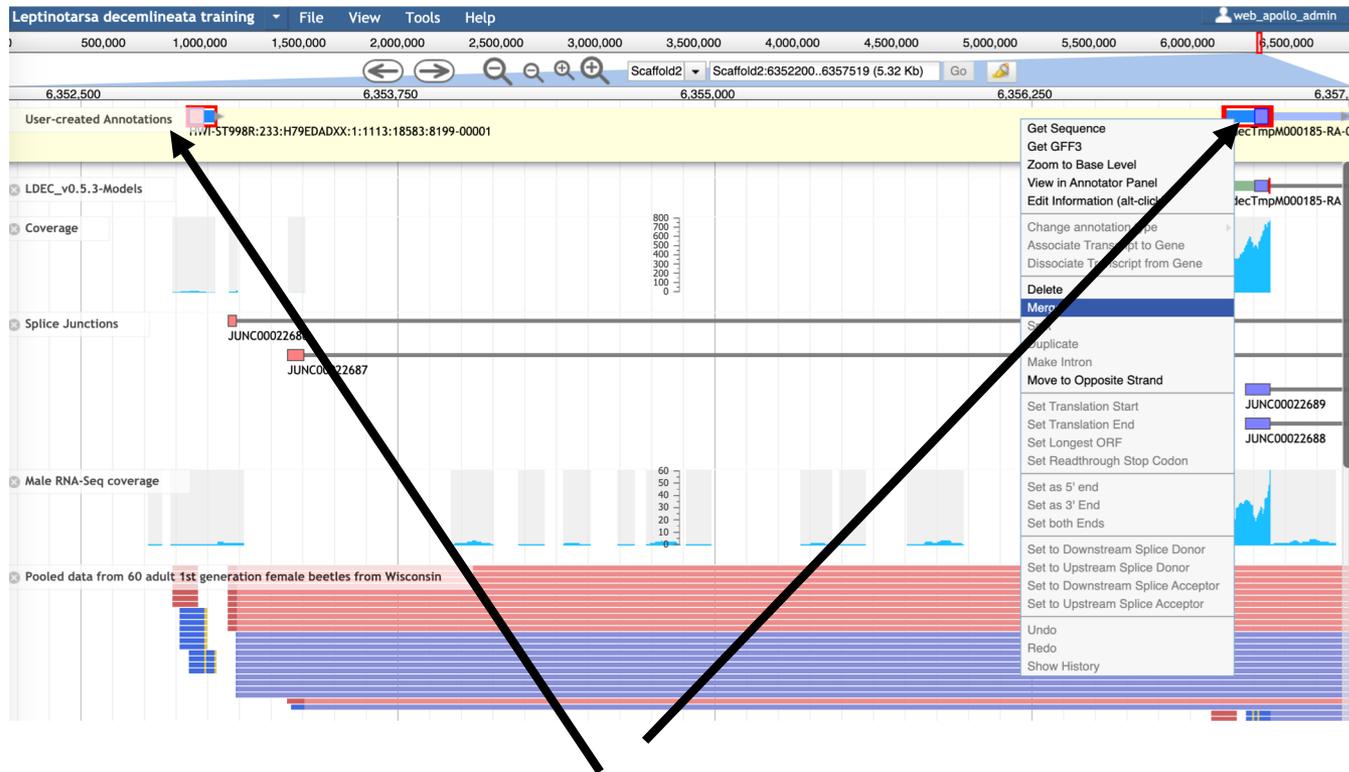
Let's flip it to the opposite strand

Fixing non-canonical splice sites



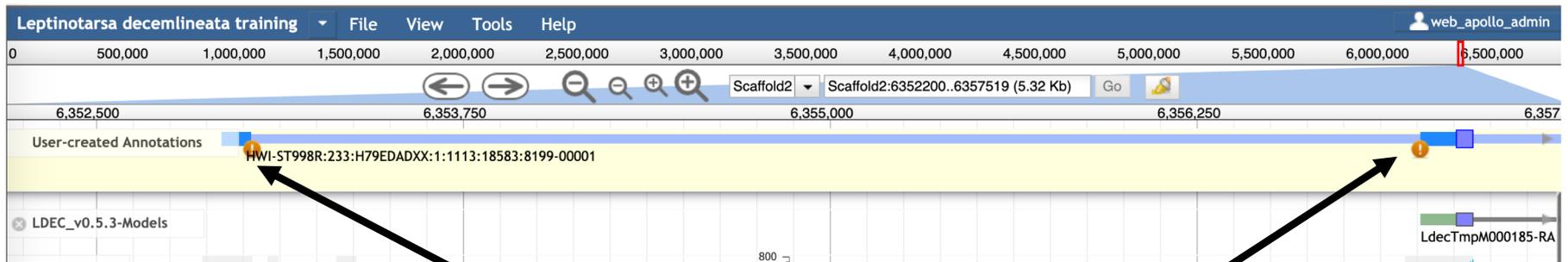
That's better.

Fixing non-canonical splice sites



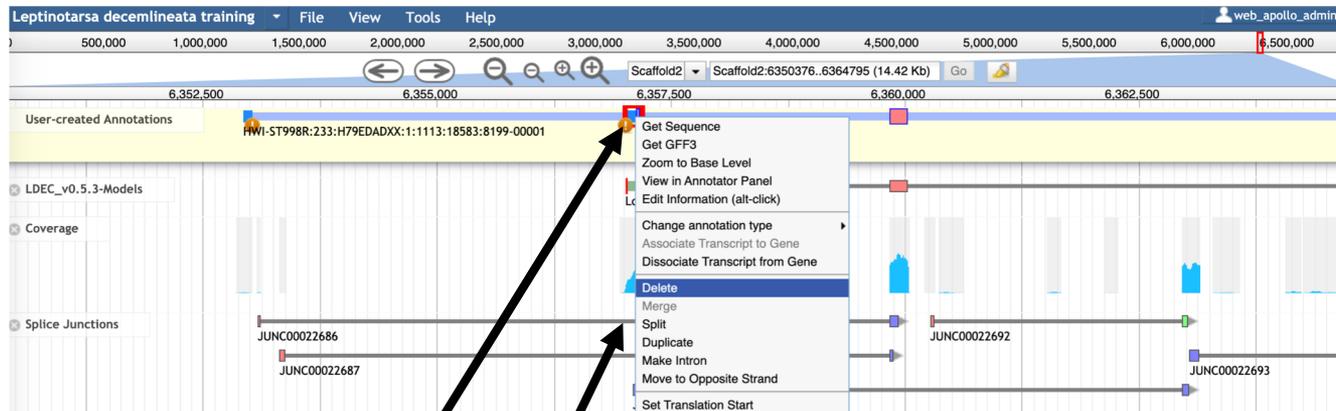
Merge the new exon to the gene model

Fixing non-canonical splice sites



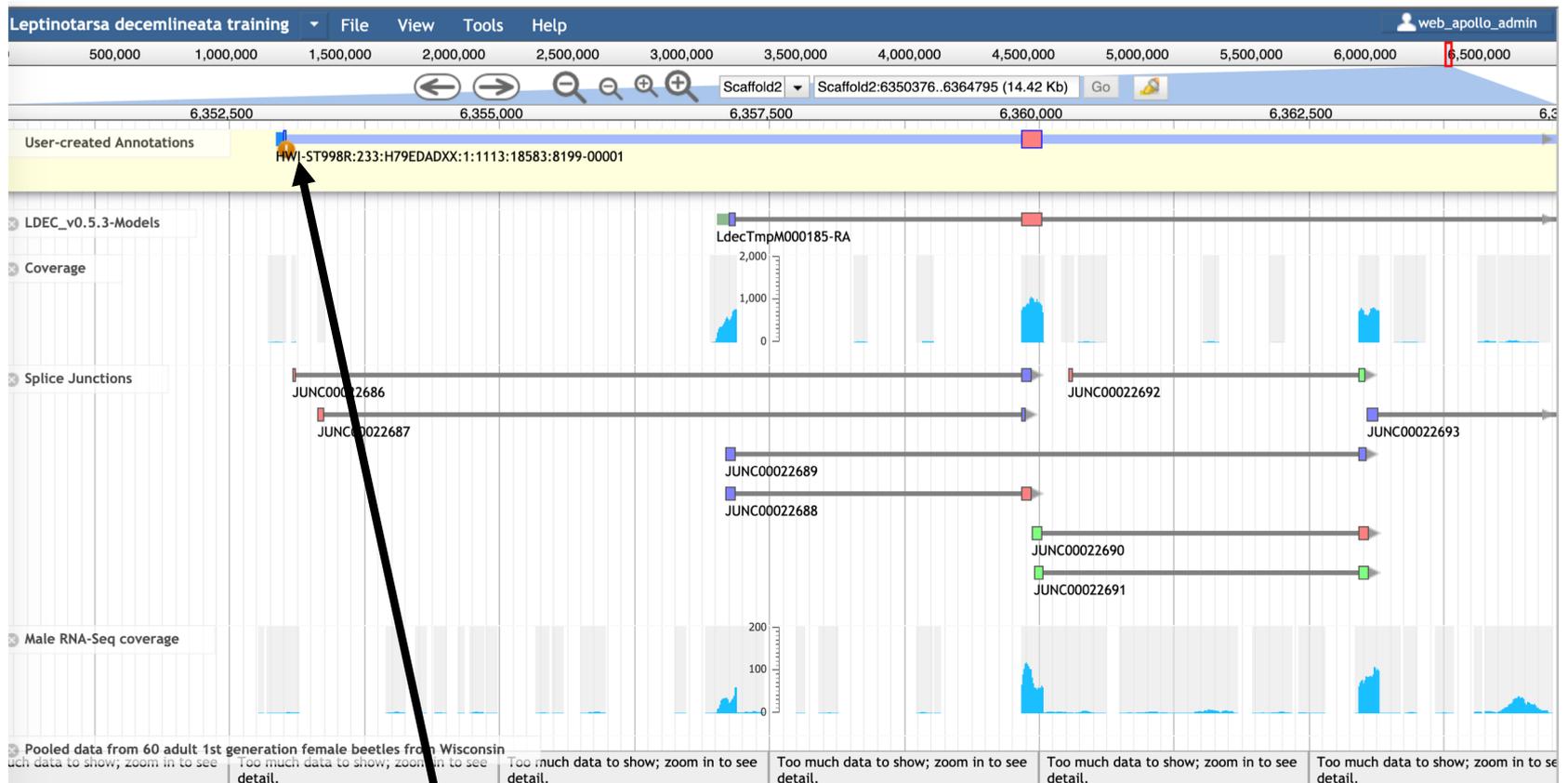
Non-canonical splice sites in merged model

Fixing non-canonical splice sites



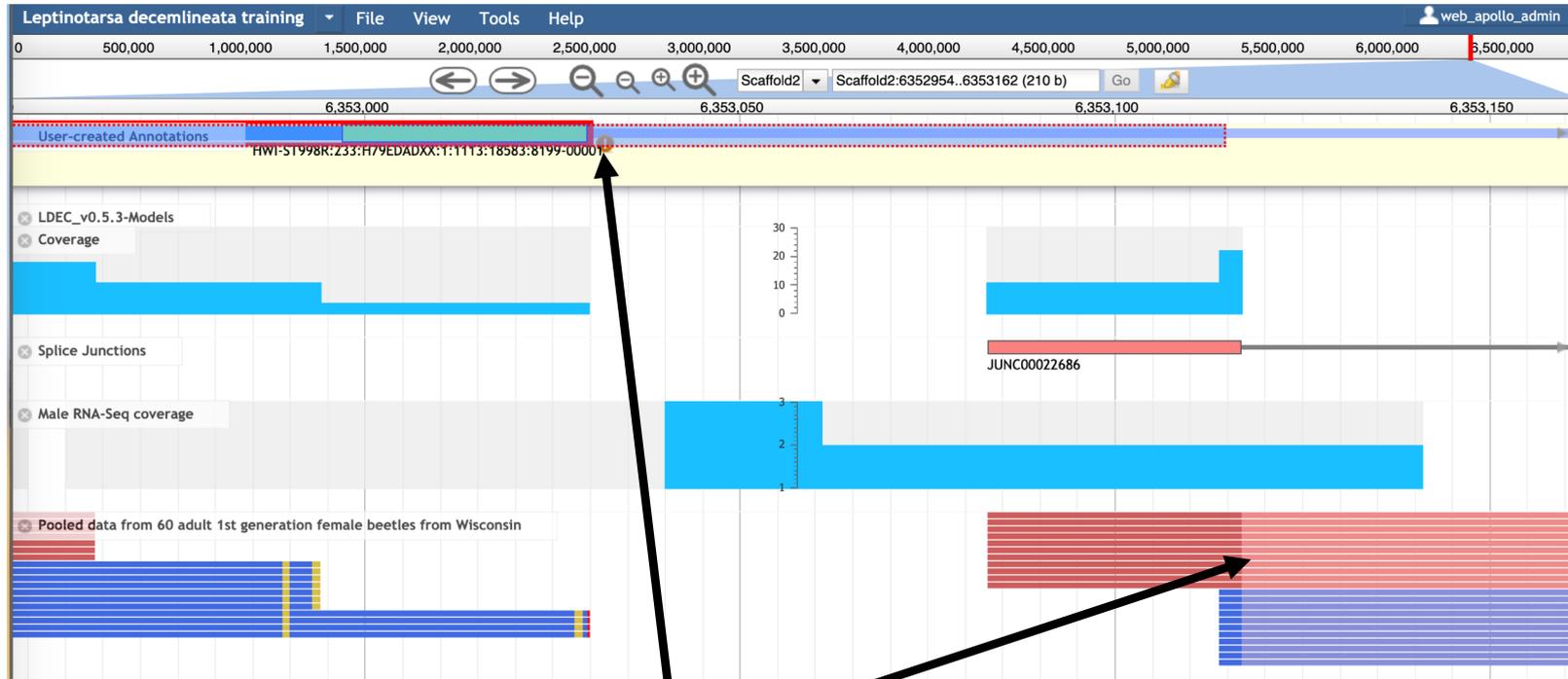
The splice junction reads don't support the 2nd exon with the new 5' exon, so let's remove it

Fixing non-canonical splice sites



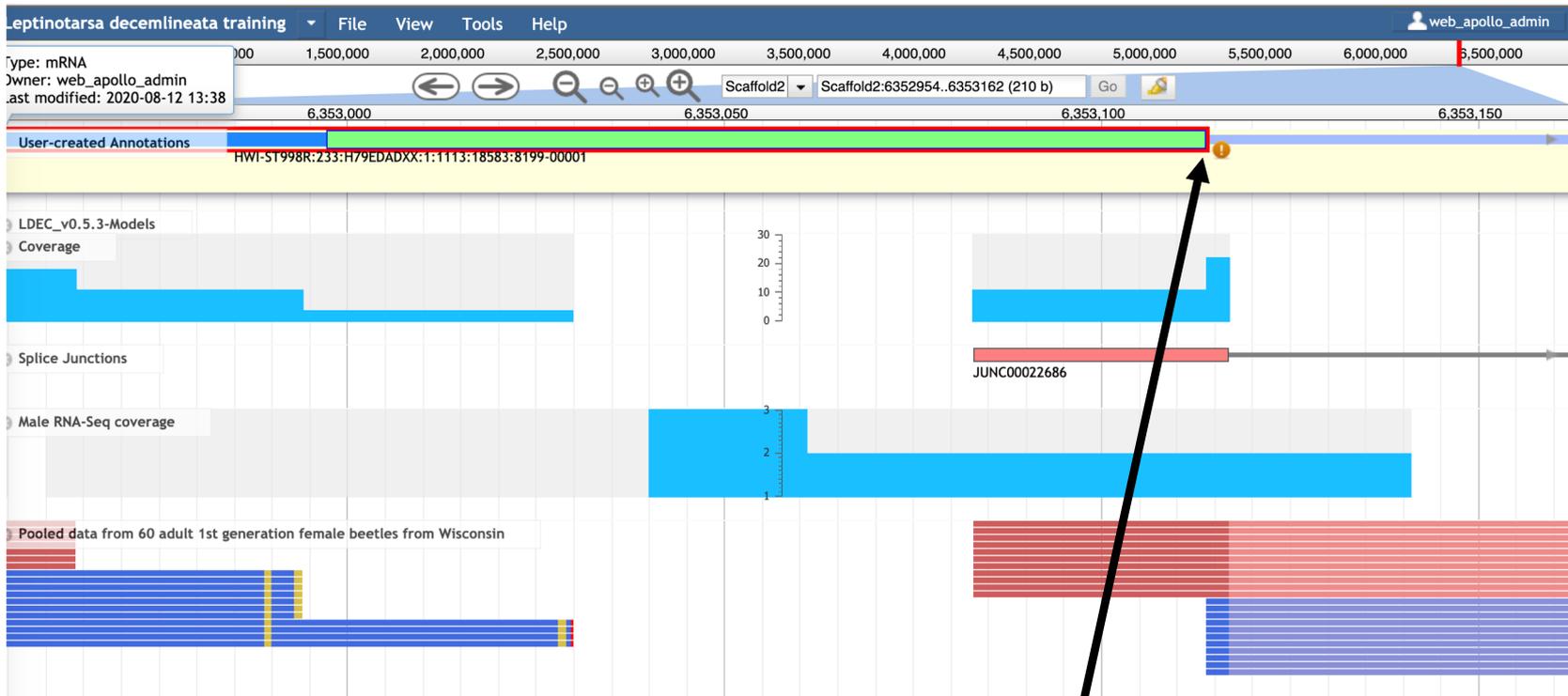
Only 1 non-canonical splice site left to fix

Fixing non-canonical splice sites



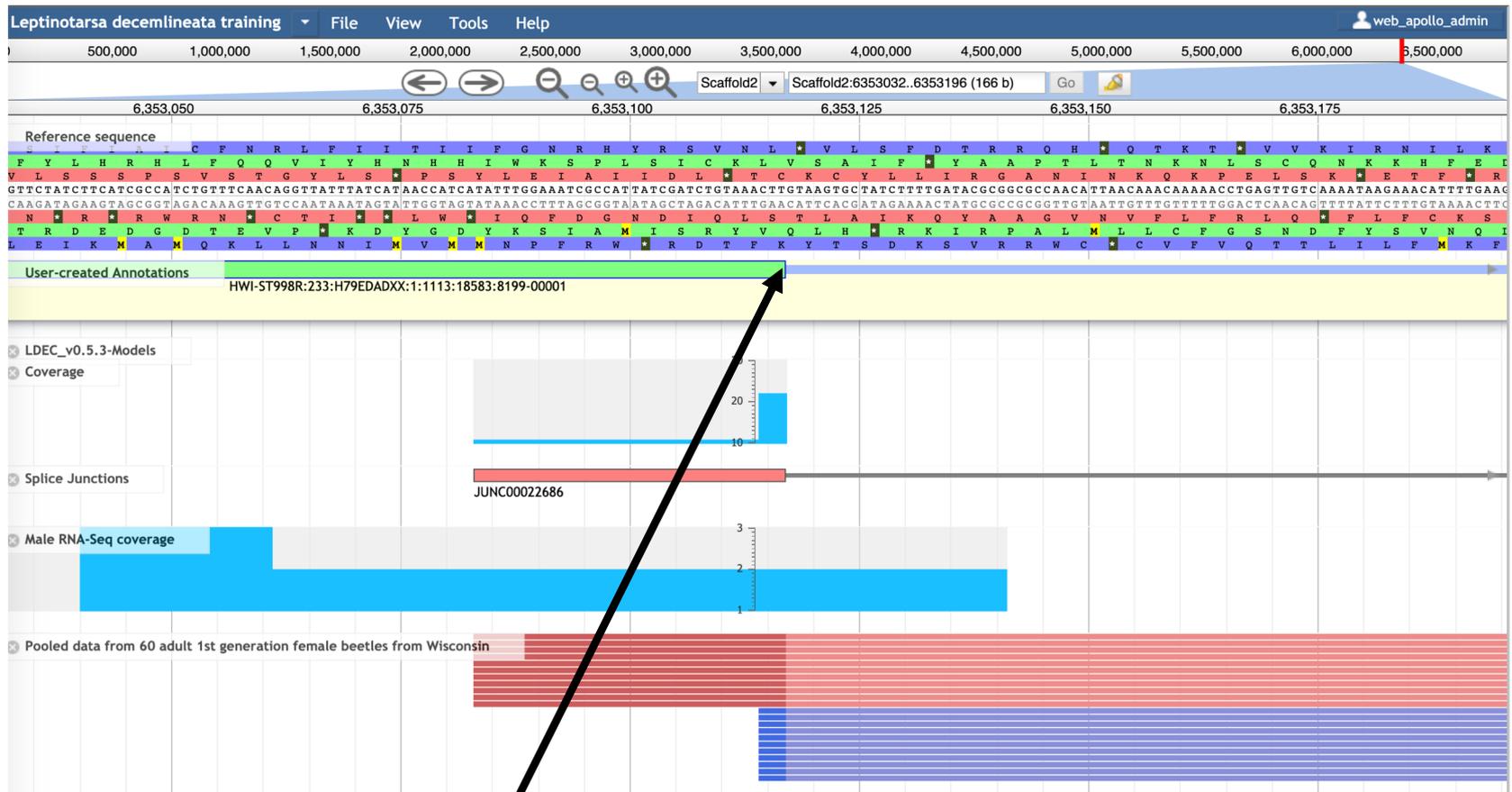
Extend exon to RNA-Seq boundary

Fixing non-canonical splice sites



Still not quite right

Fixing non-canonical splice sites



Fixed!

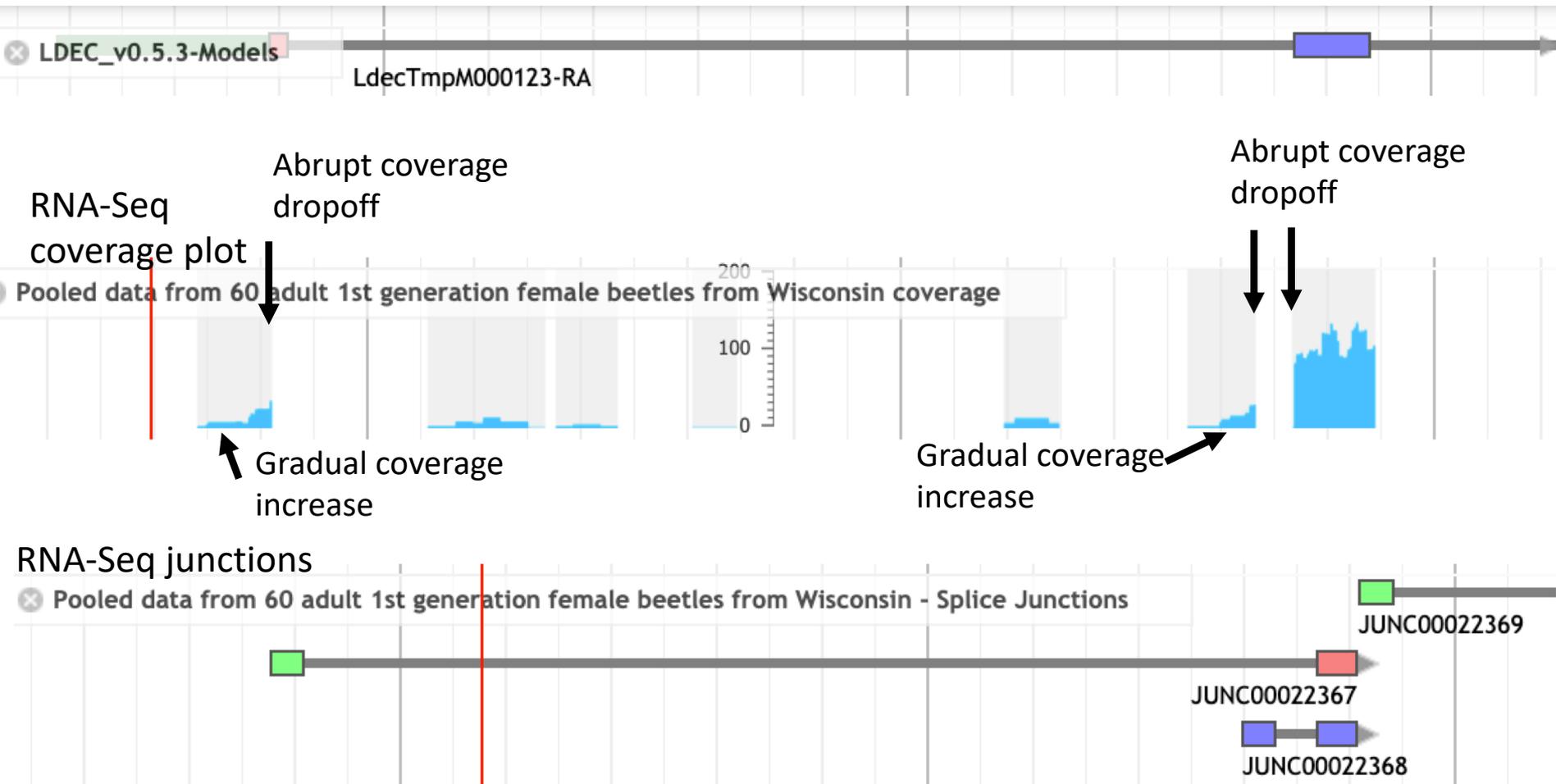
Annotating isoforms

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

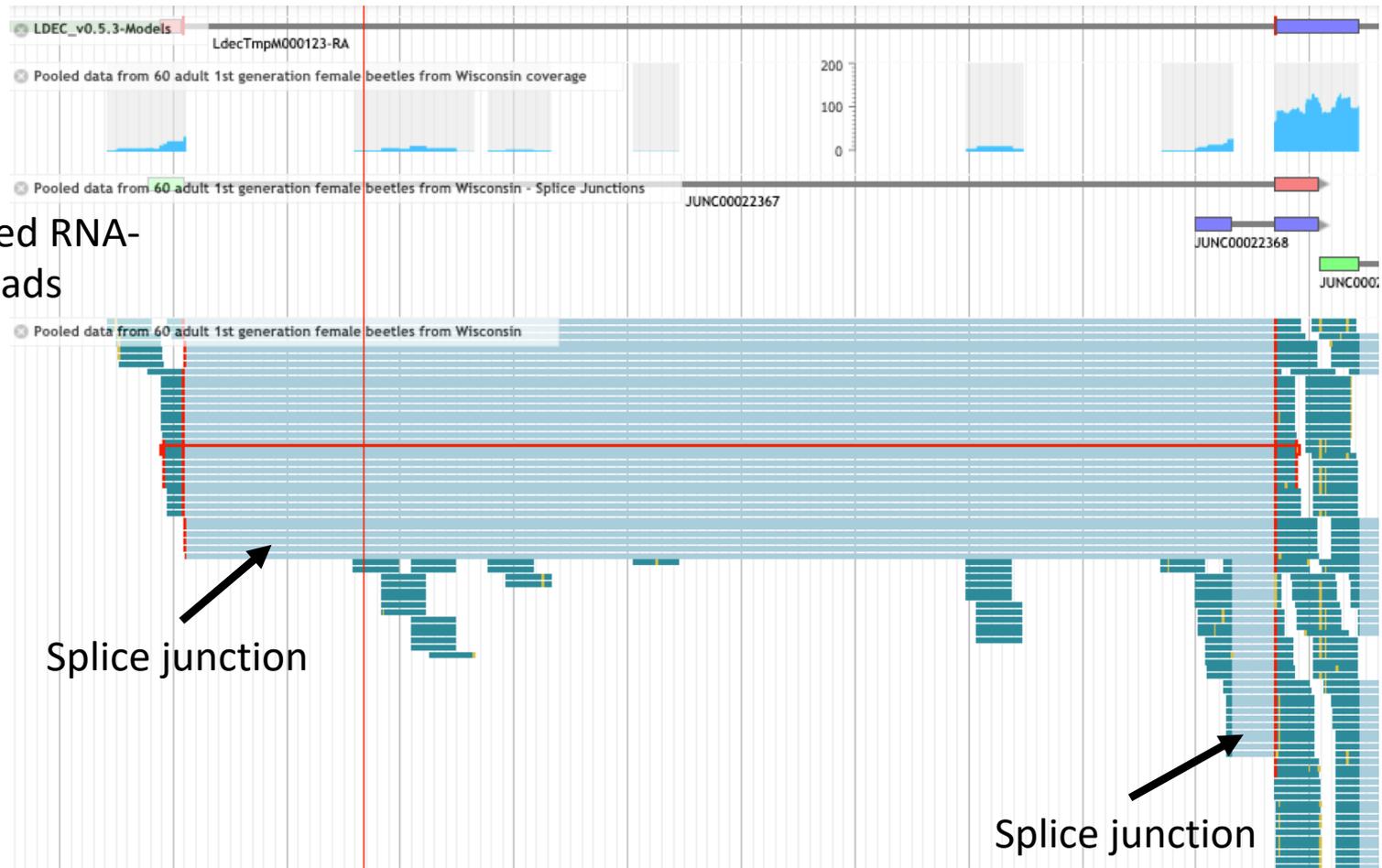
Isoform annotation example

5' end of MAKER tyrosine
protein kinase gene prediction



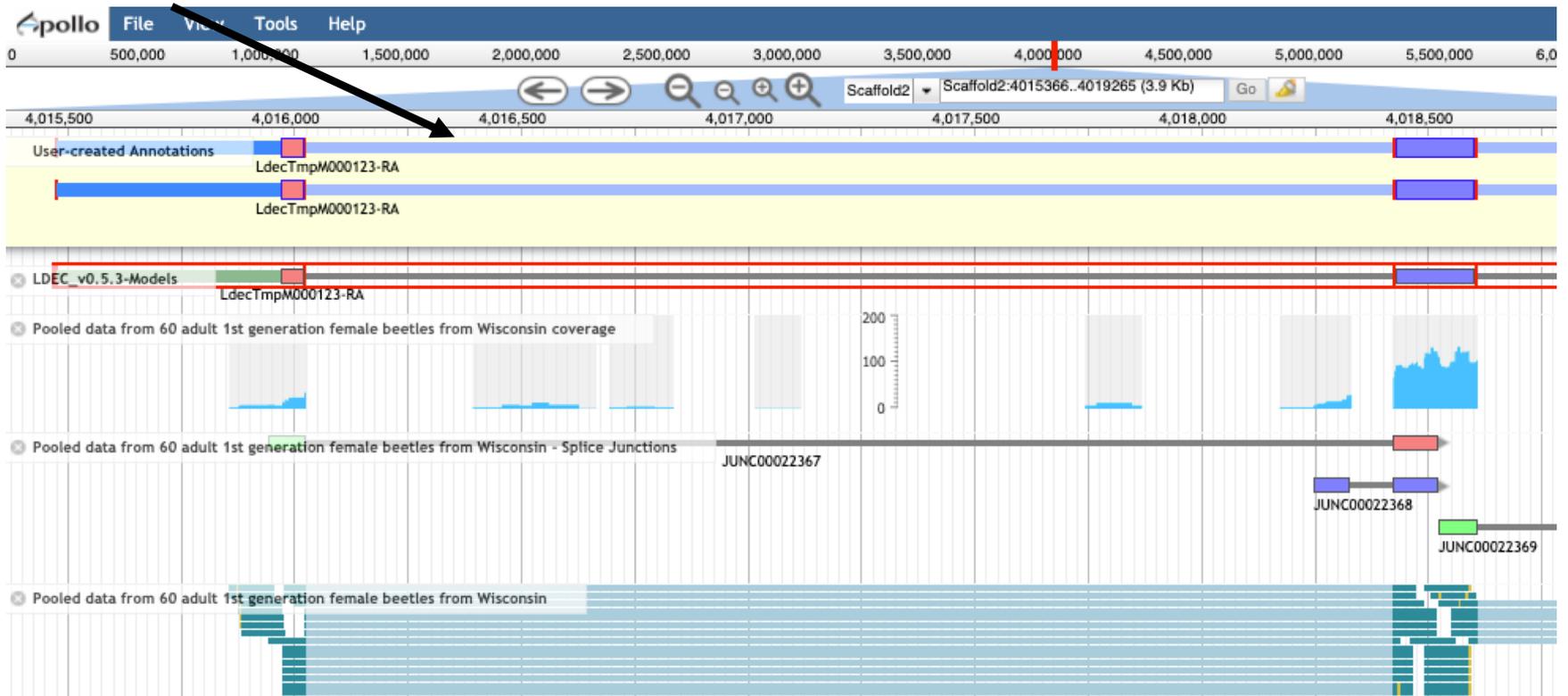
Isoform annotation example

Mapped RNA-Seq reads



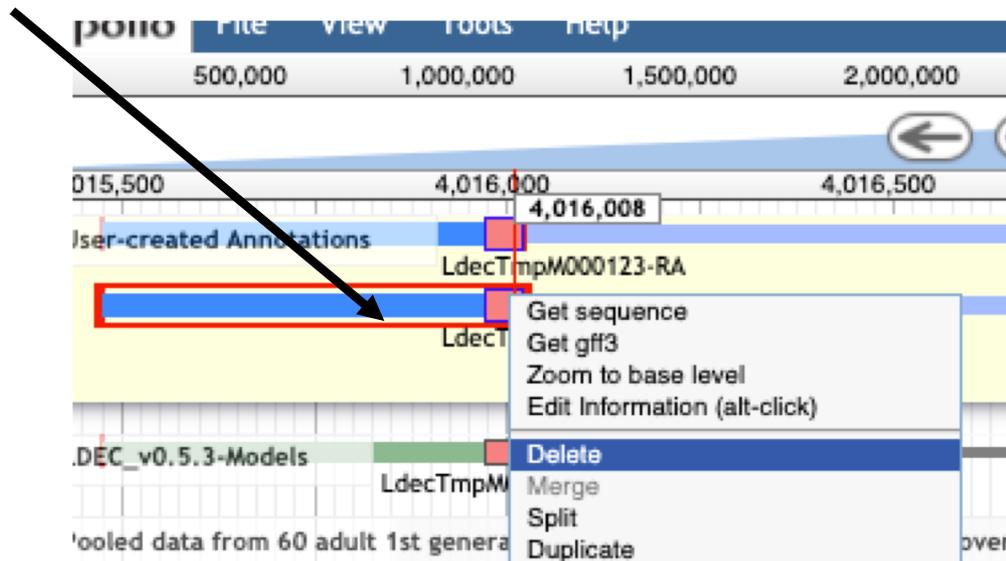
Isoform annotation example

Create 2 isoforms
from Maker model



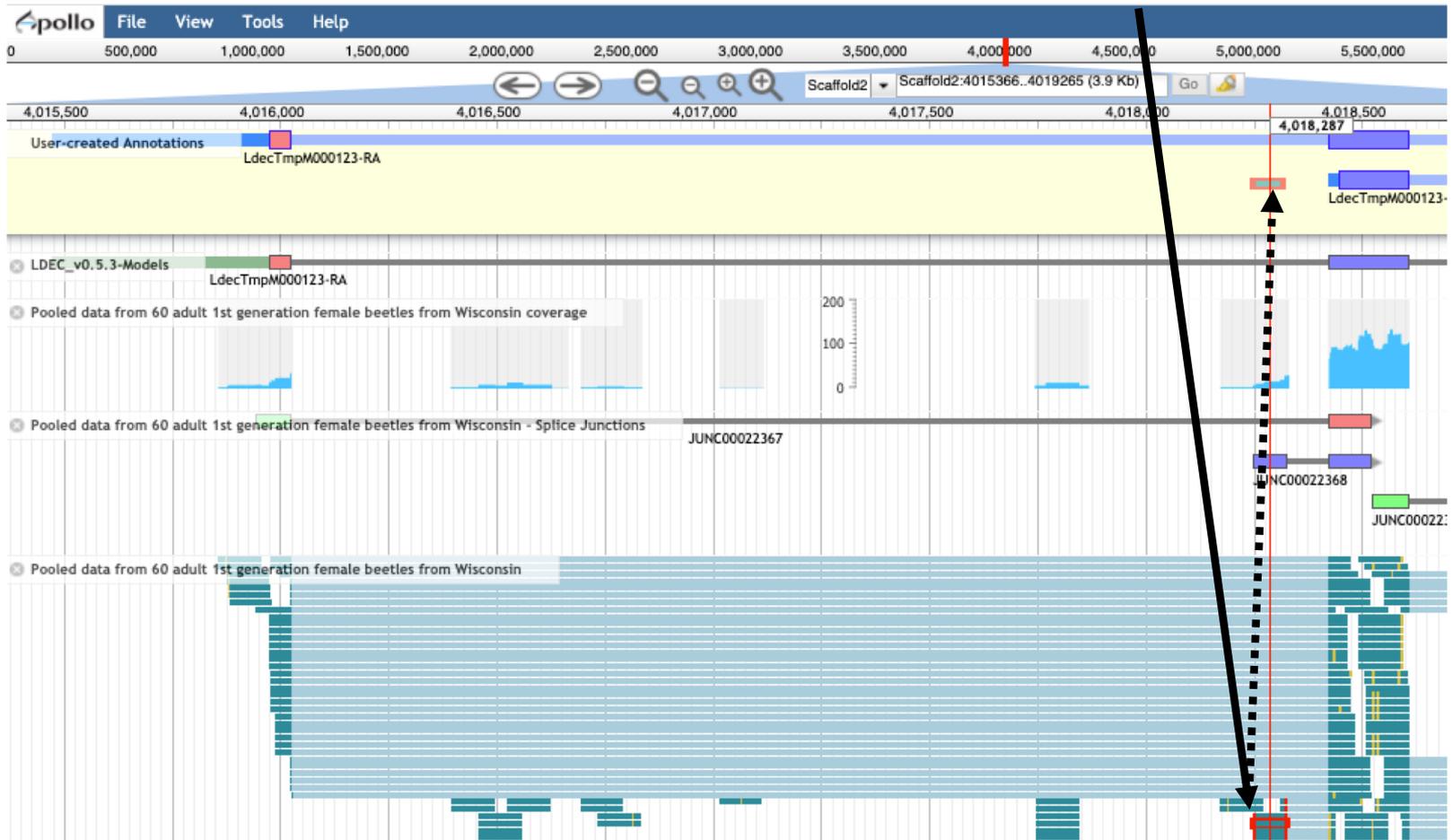
Isoform annotation example

Select and delete 5' exon from one of the isoforms



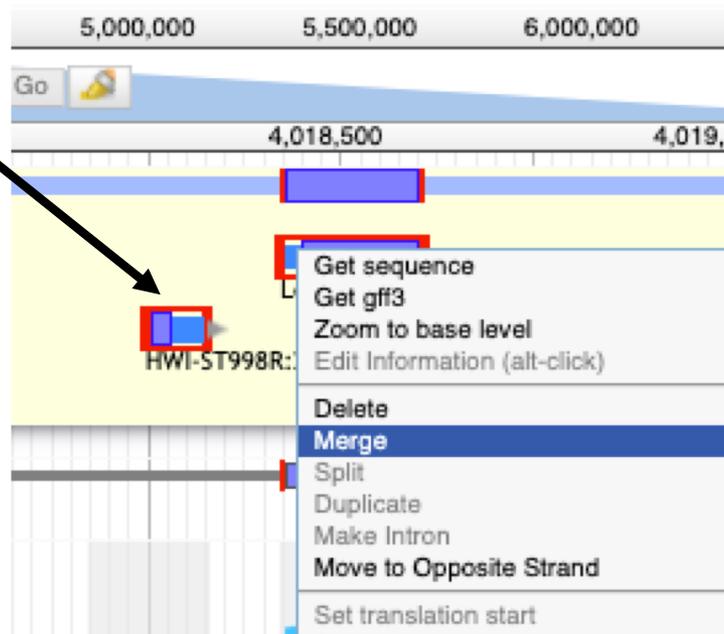
Isoform annotation example

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



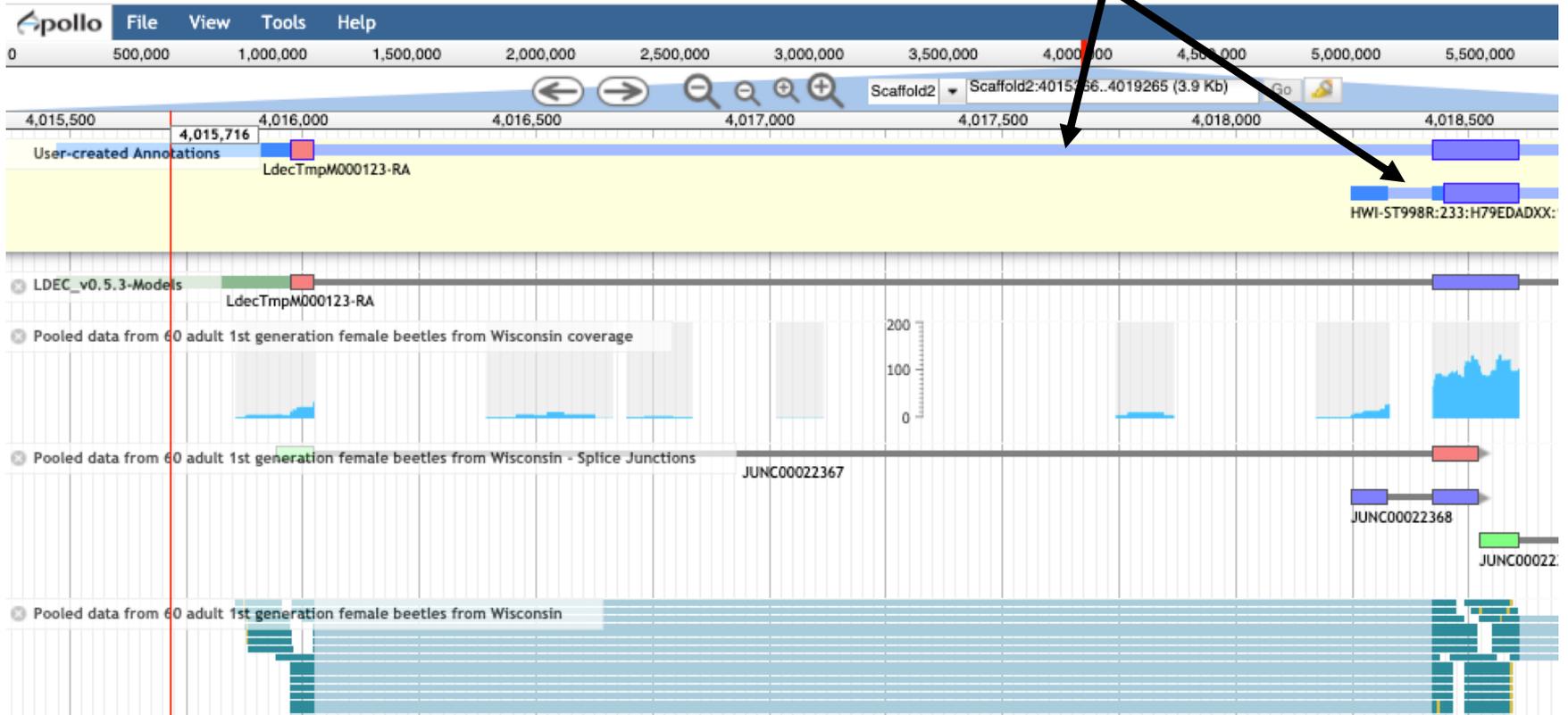
Isoform annotation example

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



Isoform annotation example

2 isoforms supported by RNA-Seq evidence



Sequence alterations and stop-codon readthroughs

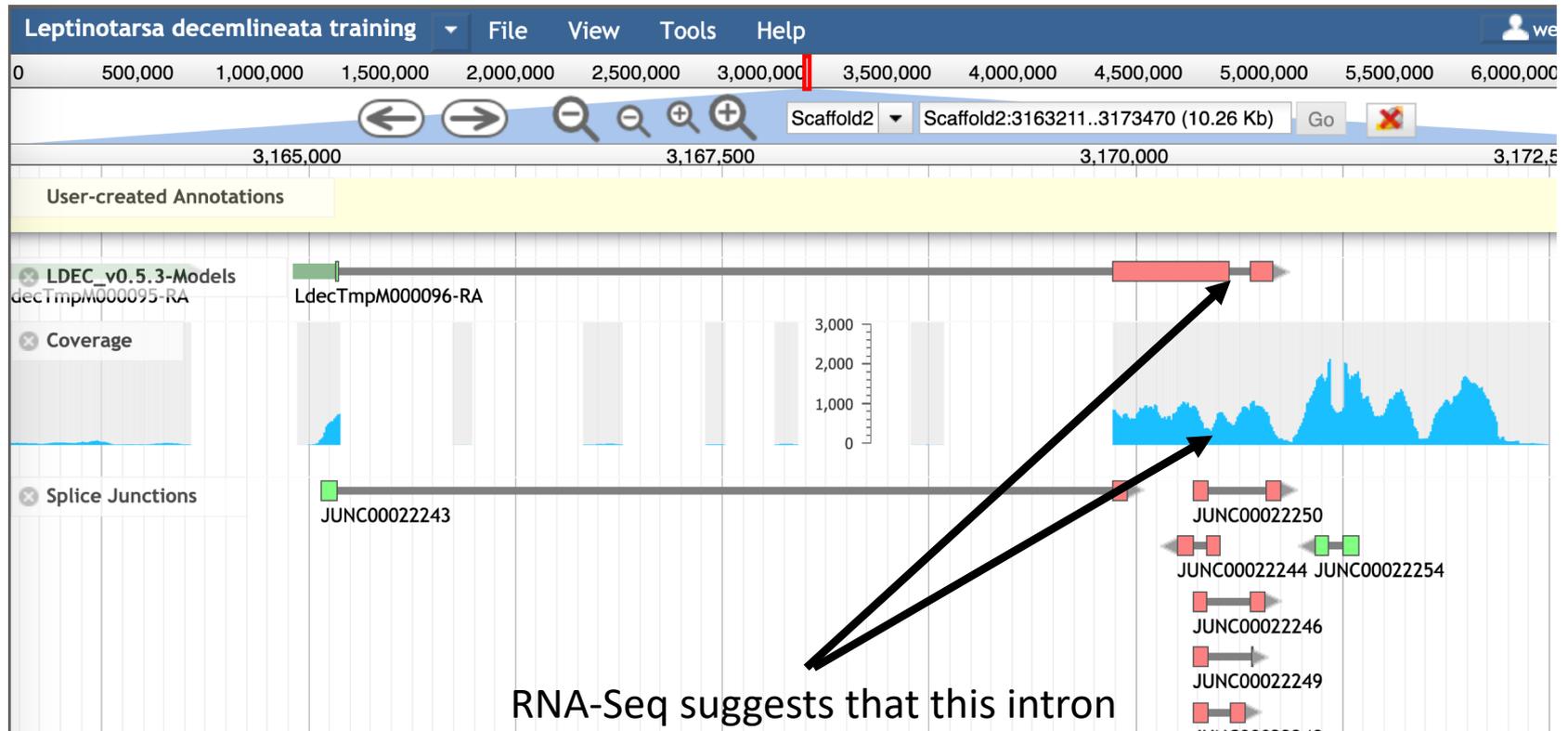
Sequence alterations

- Apollo supports annotating the genome sequence with insertions, deletions, and substitutions
- Note that this will not change the genome fasta in the sequence export – but it will allow Apollo to re-calculate a gene model's sequence
- Only add a sequence alteration in Apollo if there is evidence for it – e.g. SNPs in mapped RNA-Seq

Stop-codon readthroughs

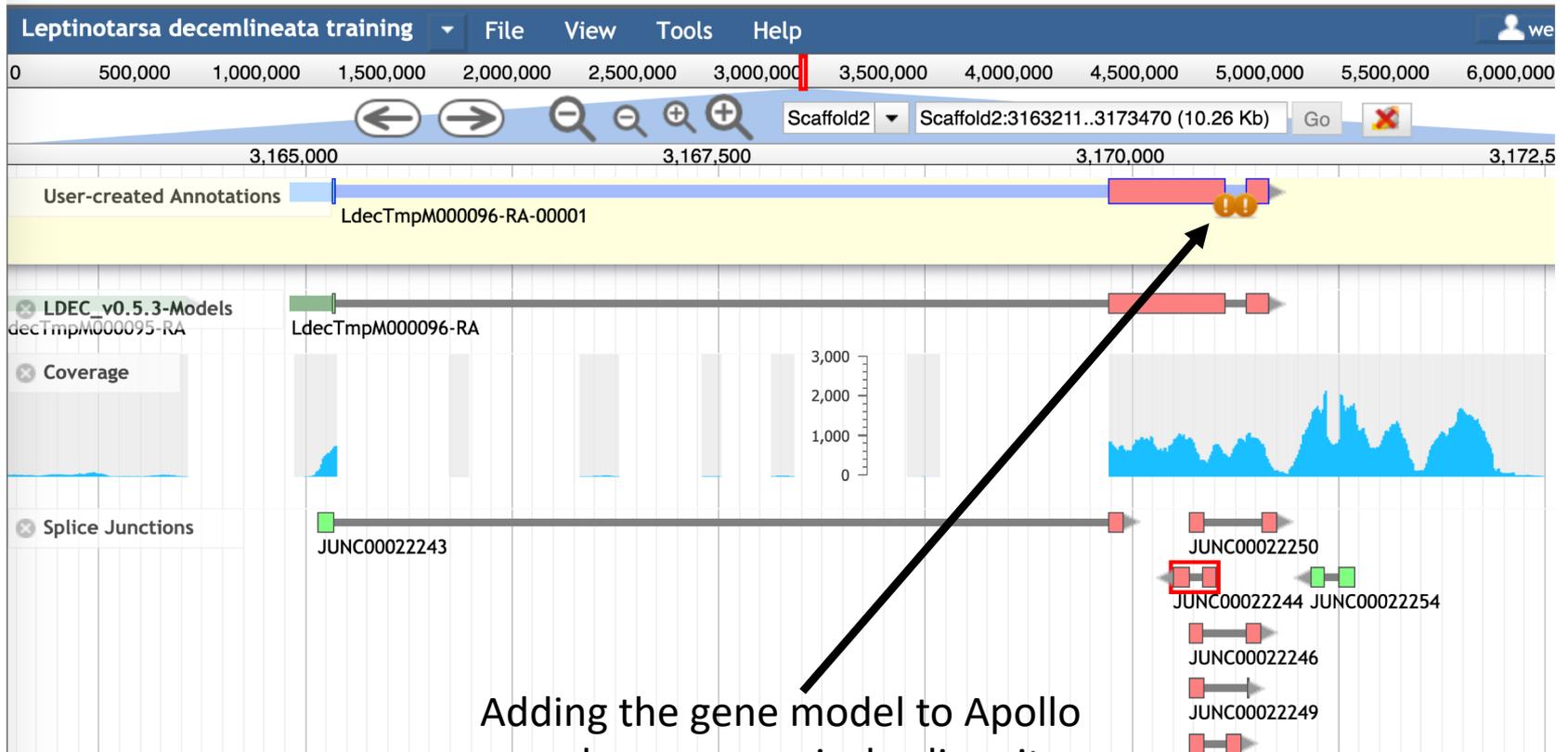
- Apollo allows you to annotate stop-codon readthrough features on the coding sequence of a gene model
- This is a special case for selenocysteine-containing proteins.
- This feature can be used in other cases – e.g. if you have evidence of errors in the genome assembly - but we don't recommend it

Sequence alterations

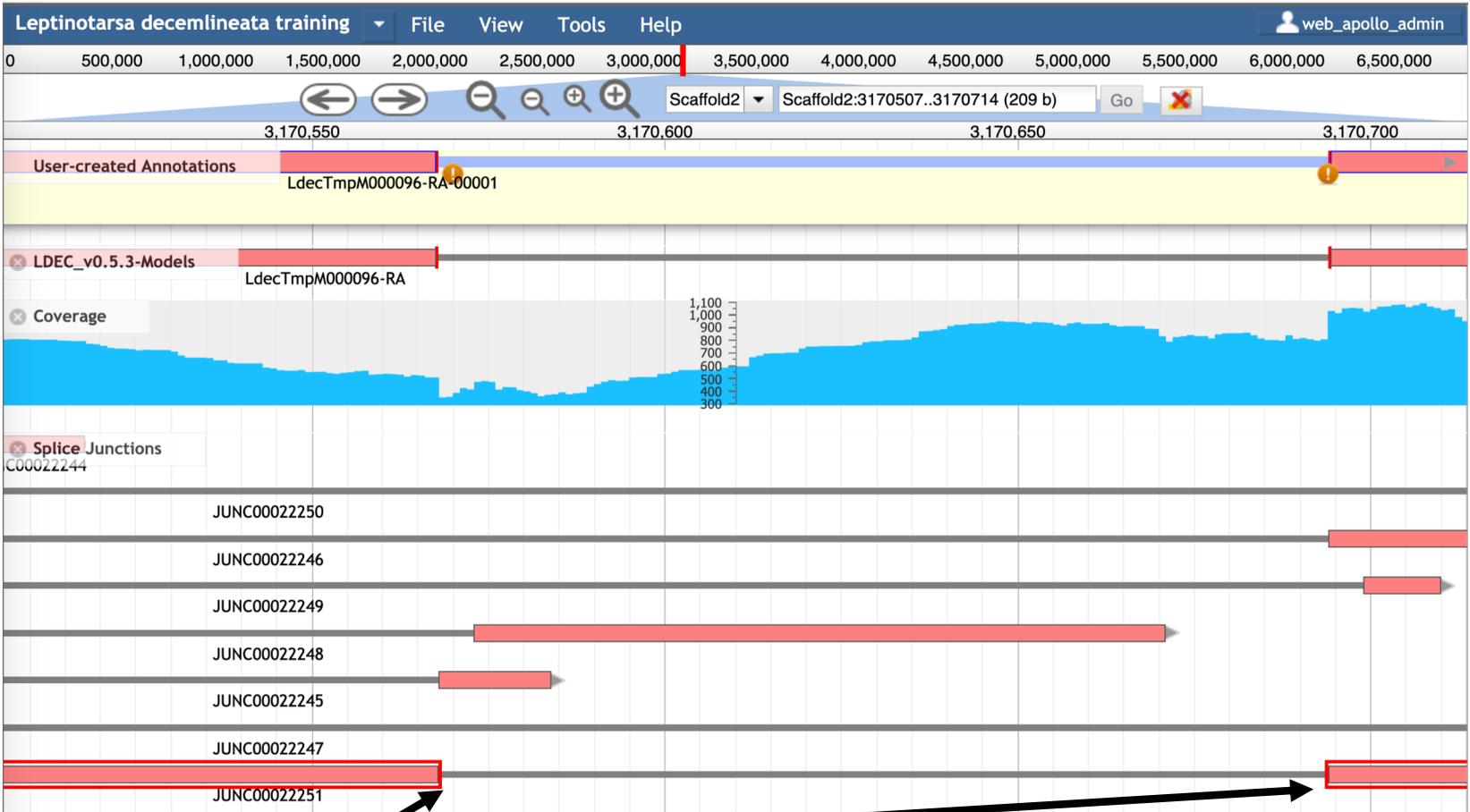


RNA-Seq suggests that this intron may not exist

Sequence alterations

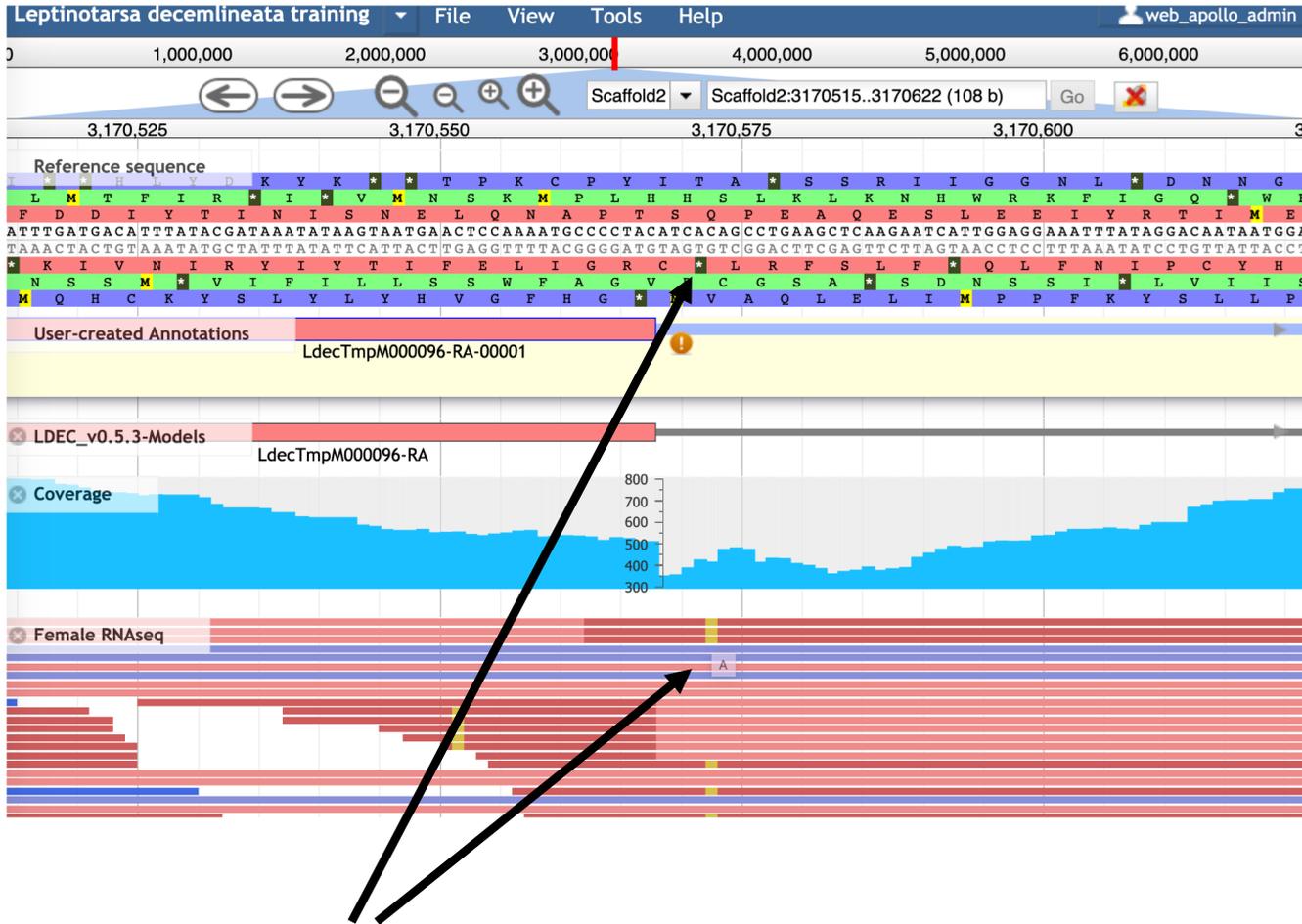


Sequence alterations



There is some support for a spliced isoform, but the RNA-Seq also suggests contiguous coding sequence

Sequence alterations



Zooming in, we see a stop codon in the 'pink' frame on the reverse strand, but SNPs in all the RNA-Seq reads

Sequence alterations

The screenshot displays a genome browser interface for *Leptinotarsa decemlineata*. The top menu bar includes 'File', 'View', 'Tools', and 'Help'. The search bar shows 'Scaffold2:3170515..3170622 (108 b)'. The main view area shows a reference sequence with a context menu open over a nucleotide. The context menu options are: 'Toggle Reverse Strand', 'Toggle Protein Translation', 'Create Genomic Insertion', 'Create Genomic Deletion', and 'Create Genomic Substitution'. An arrow points from the 'Create Genomic Substitution' option to the corresponding nucleotide in the Reference sequence track.

Right-click on the corresponding nucleotide in the *genome assembly* and select 'Create Genomic Substitution'

Sequence alterations

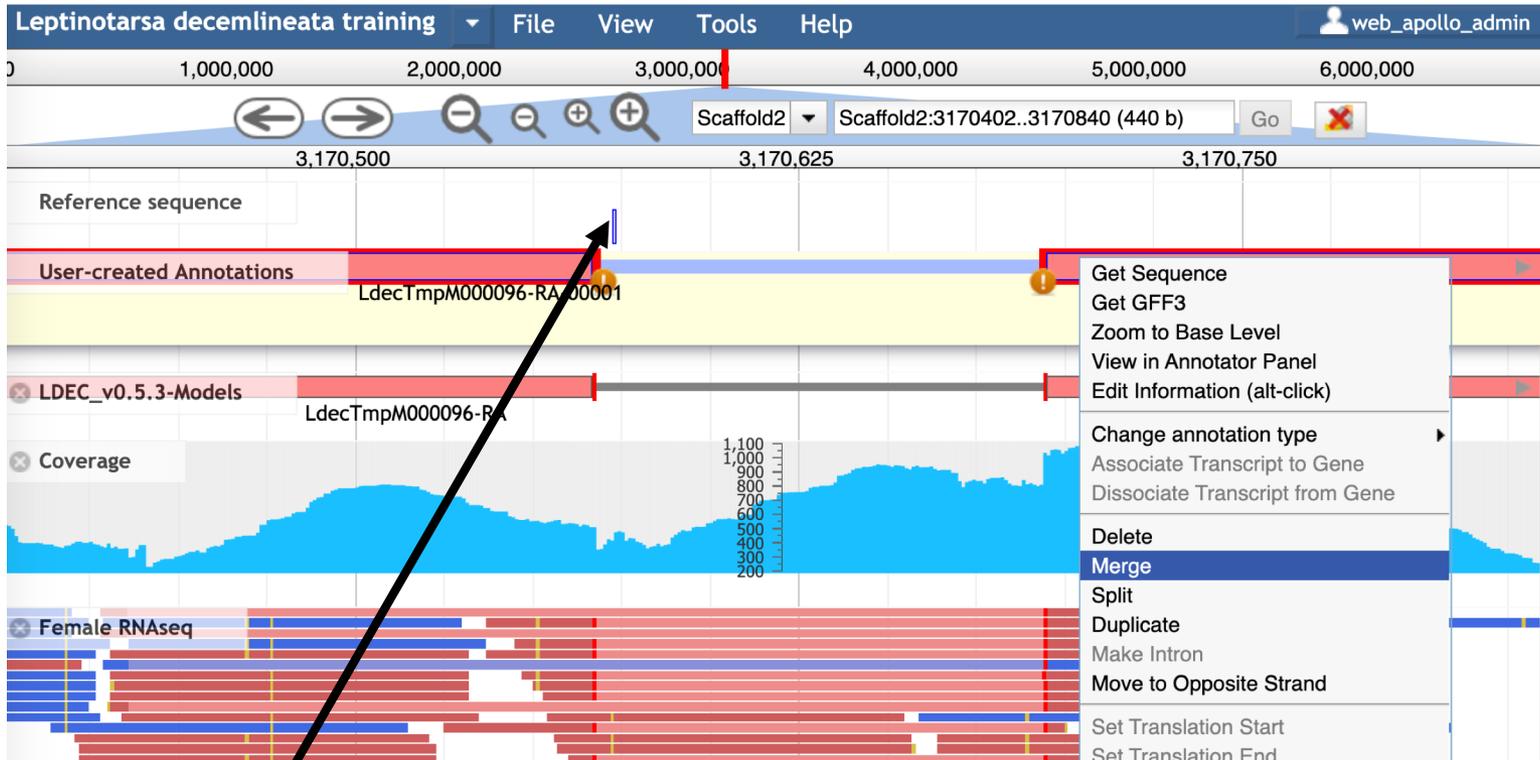
The screenshot displays a genome browser interface for *Leptinotarsa decemlineata*. The top menu includes "File", "View", "Tools", and "Help". The main view shows a reference sequence for Scaffold2:3170515..3170622 (108 b). A dialog box titled "Add Substitution" is open, with the following fields:

- + strand: A
- strand: T
- Comment: All available RNA
- Buttons: Add

The background shows various tracks: "User-created Annotations" (LdecTmpM000096-RA-00001), "LDEC_v0.5.3-Models" (LdecTmpM000096-RA), "Coverage" (blue bar chart), and "Female RNAseq" (red and blue bars).

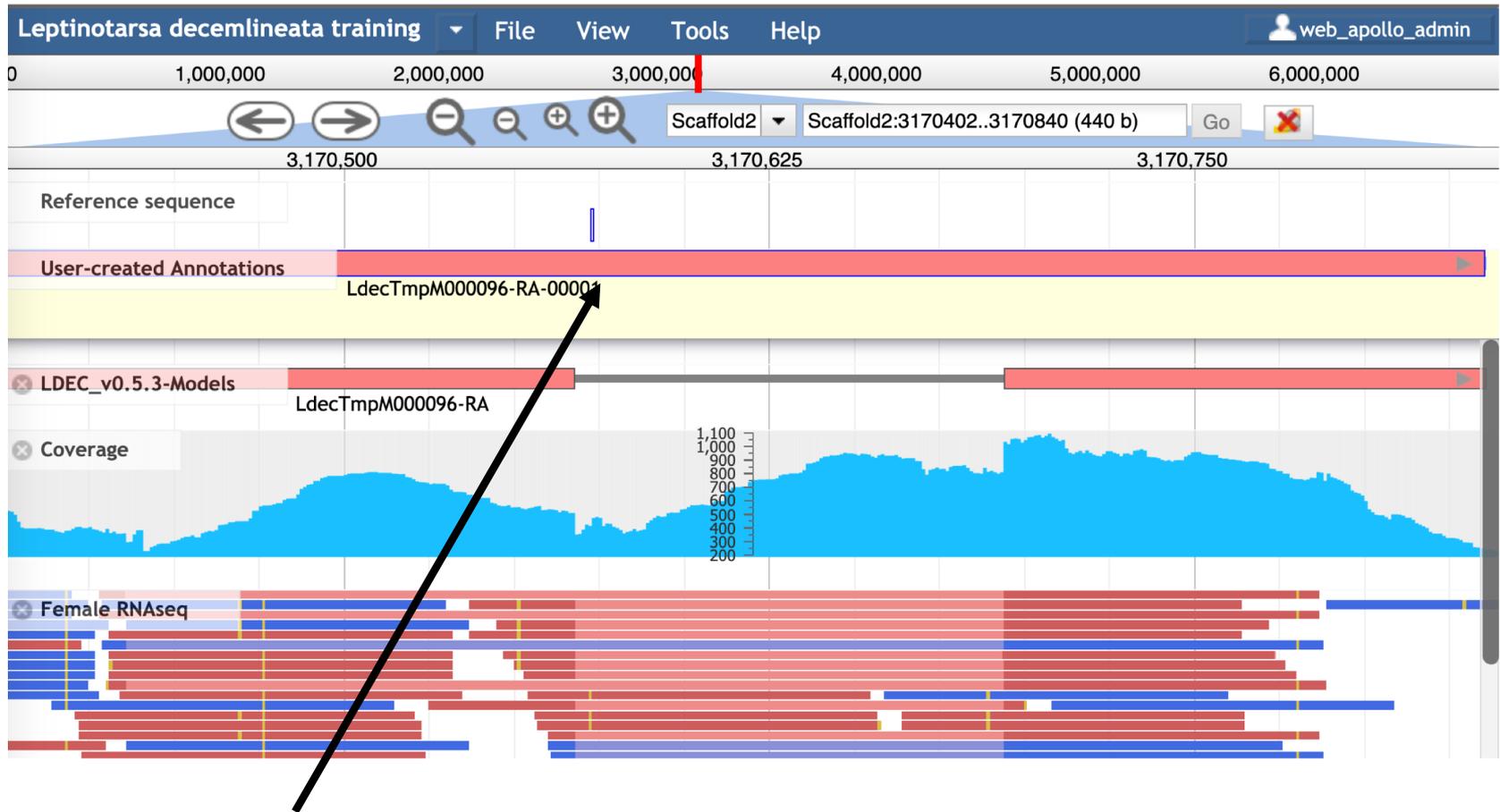
Add the substitution, and a (required) justification

Sequence alterations



Apollo added the substitution! Now, let's merge the CDS regions

Sequence alterations



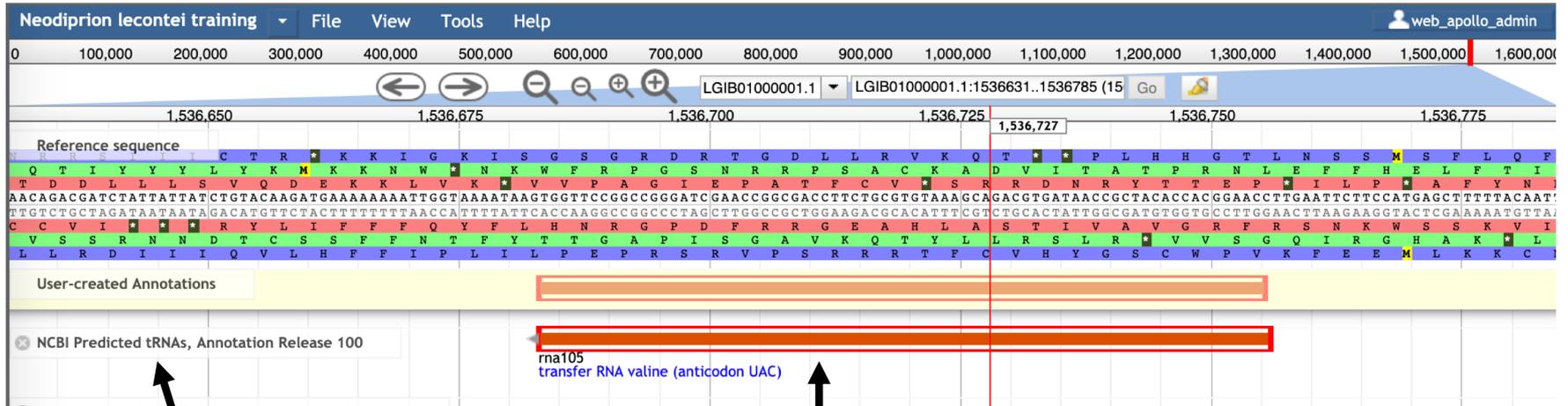
The sequence merged!

Non-coding features

Non-coding features

- Apollo supports protein-coding and non-coding features
- By default, Apollo will create protein-coding features
- For non-coding features, you can set the feature type before or after setting up the model

Non-coding features



This feature is a tRNA

Let's drag it to the UcaA track to modify it

Non-coding features

The screenshot displays the Apollo genome browser interface for the *lecontei* training set. The top navigation bar includes 'File', 'View', 'Tools', and 'Help'. The main view shows a reference sequence with a color-coded alignment. A tooltip for a user-created mRNA annotation is visible, showing the following details:

- Type: mRNA
- Owner: web_apollo_admin
- Last modified: 2020-08-14 12:51

Below the reference sequence, there are three tracks:

- User-created Annotations: A blue bar representing the mRNA annotation, labeled 'rna105-0001'.
- NCBI Predicted tRNAs, Annotation Release 100: A red bar representing a tRNA annotation, labeled 'rna105 transfer RNA valine (anticodon UAC)'.
- NCBI Predicted pseudogenes, Annotation Release 100: This track is currently empty.

Apollo turned it
into an mRNA –
let's fix that

Non-coding features

The screenshot shows the Neodiprion lecontei training genome browser interface. The top navigation bar includes 'File', 'View', 'Tools', and 'Help'. The main display area shows a reference sequence with a user-created annotation 'rna105-00001' highlighted in yellow. A context menu is open over this annotation, with 'Change annotation type' selected, and a sub-menu showing 'tRNA' as the chosen option. A black arrow points from the text below to the 'rna105-00001' annotation. Below the main sequence, there are sections for 'NCBI Predicted tRNAs, Annotation Release 100' and 'NCBI Predicted pseudogenes, Annotation Release 100'. The sequence itself is color-coded by amino acid, and the annotation is a yellow bar with the text 'rna105-00001' and 'ACCAAGGCCCGCCCTAGCTTGGCC'.

Right-click on feature, select
'Change annotation type', then
'tRNA'

Non-coding features

The screenshot displays a genome browser interface for the Neodiprion lecontei training dataset. The top navigation bar includes 'File', 'View', 'Tools', and 'Help' menus, along with the user 'web_apollo_admin'. A coordinate scale at the top shows positions from 100,000 to 1,600,000. The main view shows a reference sequence with a tooltip for a tRNA feature. The tooltip text is: 'Type: tRNA', 'Owner: web_apollo_admin', and 'Last modified: 2020-08-14 12:52'. Below the reference sequence, a 'User-created Annotations' section shows a green bar for 'rna105-00001'. A table below lists 'NCBI Predicted tRNAs, Annotation Release 100' and 'NCBI Predicted pseudogenes, Annotation Release 100'. The table entry for 'rna105' is highlighted in orange and reads: 'rna105 transfer RNA valine (anticodon UAC)'. A black arrow points from the text below to the 'rna105' entry in the table.

Now we have the correct feature type

Non-coding features

The screenshot displays the Apollo genome browser interface for the Neodiprion lecontei training dataset. The top navigation bar includes 'File', 'View', 'Tools', and 'Help'. The main area shows a reference sequence with coordinates from 100,000 to 1,600,000. Below the sequence, there are tracks for 'User-created Annotations', 'NCBI Predicted tRNAs, Annotation Release 100', and 'NCBI Predicted pseudogenes, Annotation Release 100'. A red bar highlights a region in the 'User-created Annotations' track, and a context menu is open over it, showing options like 'View details', 'Highlight this tRNA', and 'Create new annotation'. A dropdown menu under 'Create new annotation' lists various feature types, with 'tRNA' selected. An arrow points from the text below to the 'tRNA' option in the menu.

Another way – right-click on model *before* adding it to the Uca track, select ‘Create new annotation’, then select ‘tRNA’

Non-coding features

The screenshot shows a genome browser interface for *Neodiprion lecontei*. The top navigation bar includes a menu (File, View, Tools, Help) and a user profile (web_apollo_admin). A scale bar at the top indicates genomic coordinates from 100,000 to 1,600,000. Below this, a search bar shows the current view: LGIB01000001.1:1536631..1536785 (15). The reference sequence is displayed with amino acid translations above it. A tooltip is visible over the sequence, containing the following information:

- Type: tRNA
- Owner: web_apollo_admin
- Last modified: 2020-08-14 12:51

Below the reference sequence, there are three tracks of annotations:

- User-created Annotations:** A green bar labeled `rna105-0001`.
- NCBI Predicted tRNAs, Annotation Release 100:** An orange bar labeled `rna105 transfer RNA valine (anticodon UAC)`.
- NCBI Predicted pseudogenes, Annotation Release 100:** This track is currently empty.

Now it's a tRNA!

Thank you!

The NAL Team

- Chris Childers
- Vern Chapman
- Ming Chen
- Susan McCarthy
- Shang-Yu Chang
- Hsiu-Kang Huang

- 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