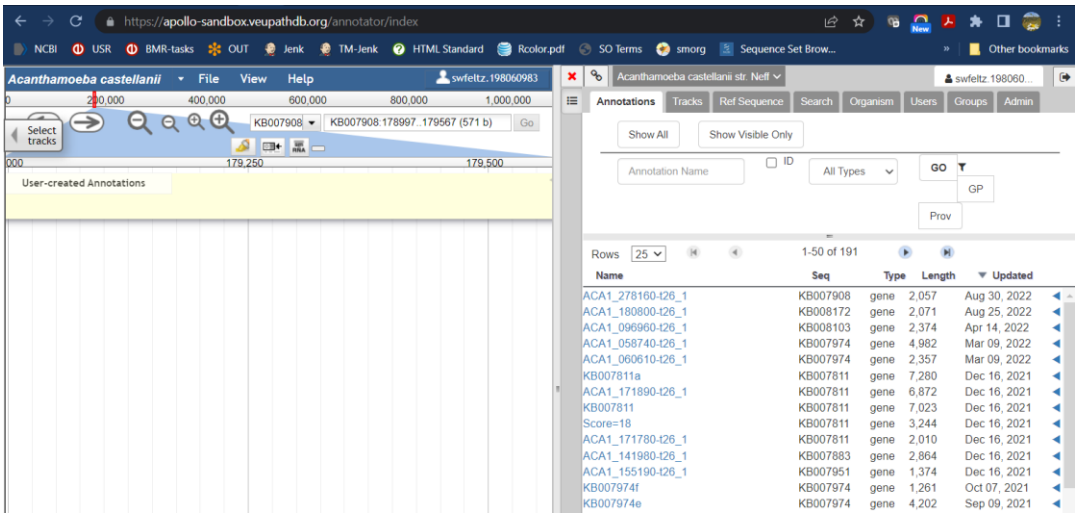
Genome annotation in Apollo
Optional exercise

Apollo is a real-time collaborative genome annotation and curation platform. More information can be found in this publication: https://pubmed.ncbi.nlm.nih.gov/30726205/

Learning objectives:

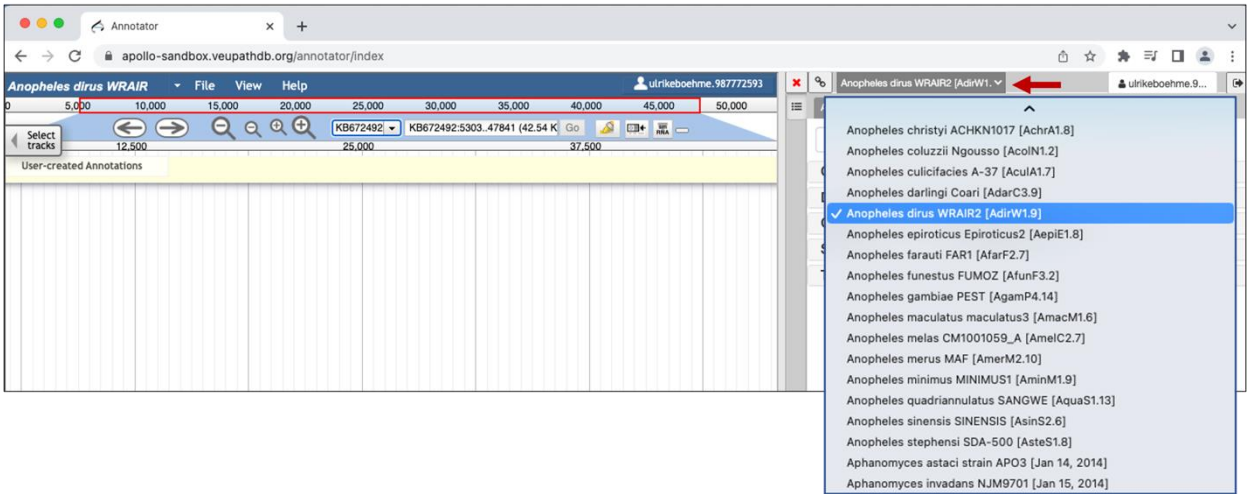
- Accessing Apollo Sandbox (**https://apollo-sandbox.veupathdb.org/annotator/index**)
- Use the menu and navigation bar of Apollo
- Add pre-loaded data tracks
- Changing gene structures in Apollo

1. **Accessing Apollo:** For this optional exercise we will use the **VEuPathDB Sandbox Apollo** instance. This Apollo instance is available to get familiar with all menus and tools. The changes you make during your work in the Sandbox will not affect any of the organism's official gene set, neither will they be preserved.
   a. **Access the Sandbox**. **https://apollo-sandbox.veupathdb.org/annotator/index** To use Apollo you need to be logged into VEuPathDB. If you have not done so yet, log now into VEuPathDB with your user ID and password.
      i. If you are logged into VEuPathDB, the link above will take you to the sandbox. The Apollo interface has a JBrowse view on the left with menus and tabbed pages on the right.



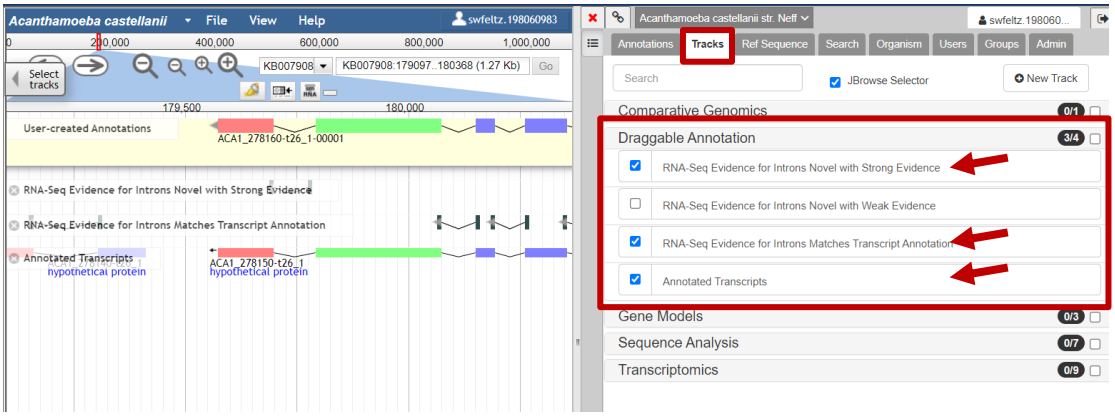
      ii. If you are not logged into VEuPathDB, you will be asked to log in. If you do not have an account, you will need to register. You can also reach the registration page here: https://veupathdb.org/veupathdb/app/user/registration

2. **Choose a Gene model to correct**: Genome annotation forms the backbone of bioinformatics since functional data such as RNA sequence is aligned to the genome before quantification. For example, an incorrect gene models can skew expression values for a gene when aligned RNA sequence reads are counted for the incorrect transcript boundaries (coordinates). Some well-studied genomes have very mature annotation while others lag behind, usually due to poor funding. Community annotation platforms like Apollo allow for annotation improvements and are a great help to the community.
   a. Please choose an incorrect gene model from the table: https://docs.google.com/spreadsheets/d/1LoFdWsD5luVH6V7QVTH8vsRQYcv4PhmodqXoxDKPOro/edit?usp=sharing
   b. In the Apollo sandbox, choose the organism for your gene from the drop-down menu on the right-hand side.

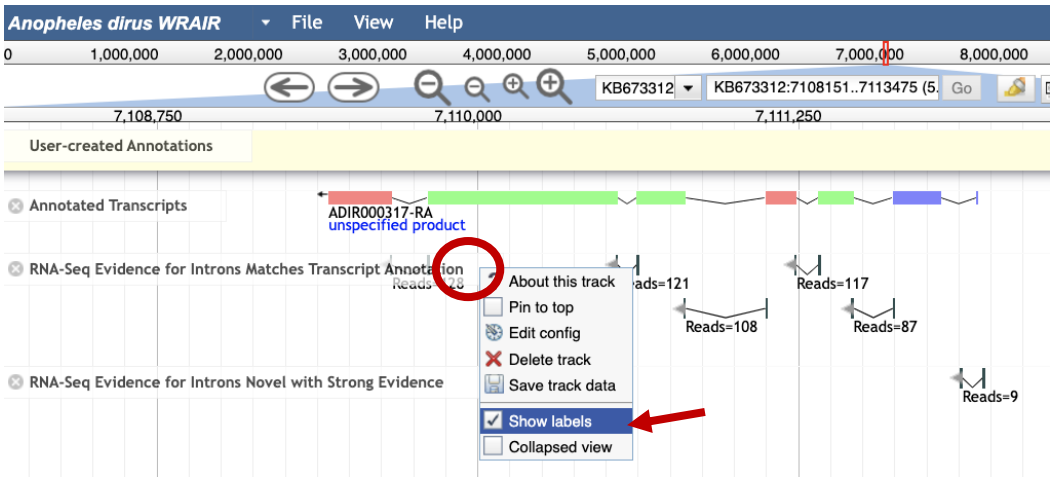


3. **Add draggable annotation tracks**: Draggable annotation tracks allow you to transfer exactly the coordinates of an intron or other feature to the new/corrected gene model that you are creating.
   a. Click on Tracks and select from Draggable Annotation
      i. Annotated Transcripts
      ii. RNA-Seq evidence for Introns Matches Transcript Annotation
      iii. RNA-Seq Evidence for Introns Novel with Strong Evidence.

   The tracks appear automatically in the left panel where you will make your changes to the structural annotation.



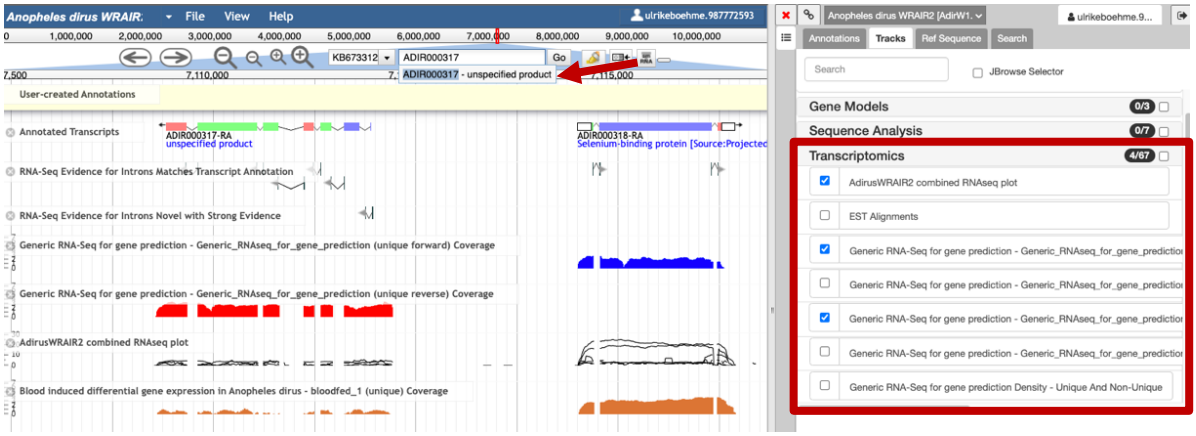
b. Label the intron tracks with the number of reads supporting the intron. All tracks have a drop-down menu that is revealed when you hover of the end of the title. Choose **Show labels** from the drop-down menu for the tracks **RNA-Seq evidence for Introns Matches Transcript Annotation** and **RNA-Seq Evidence for Introns Novel with Strong Evidence** to see the number of reads.



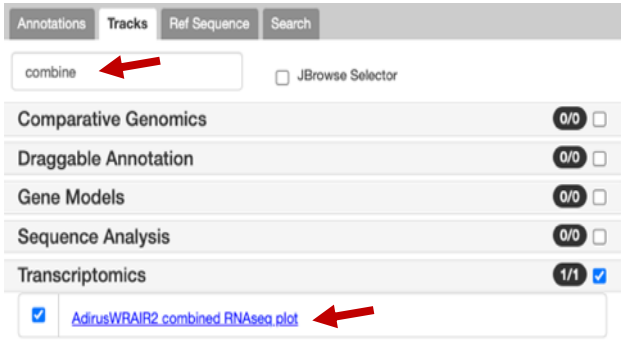
4. Find your gene or location of interest and add additional evidence (functional data).

a. Find your gene of interest by typing the gene ID or location into the search bar.

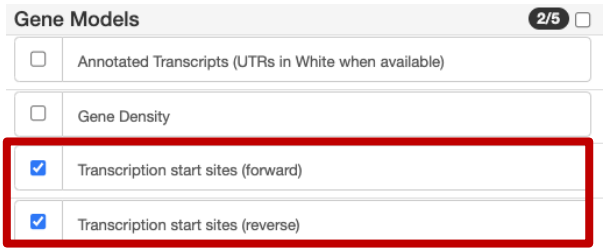
b. Choose from the Transcriptomics menu additional evidence tracks.

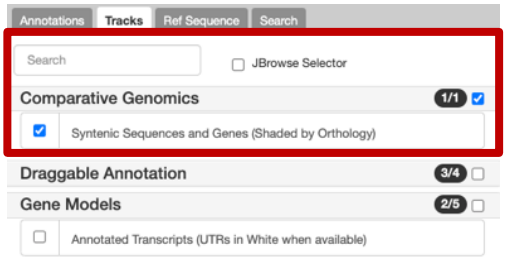


i. Hint: A track that is useful is the combined RNAseq plot. You can find the track by typing into the Search bar the word **combine**.



ii. Other useful tracks are the **synteny track** in comparative genomics and the **transcription start site tracks** in gene models. Not all organisms have transcription start site data, so don't be alarmed if you cannot find the track.





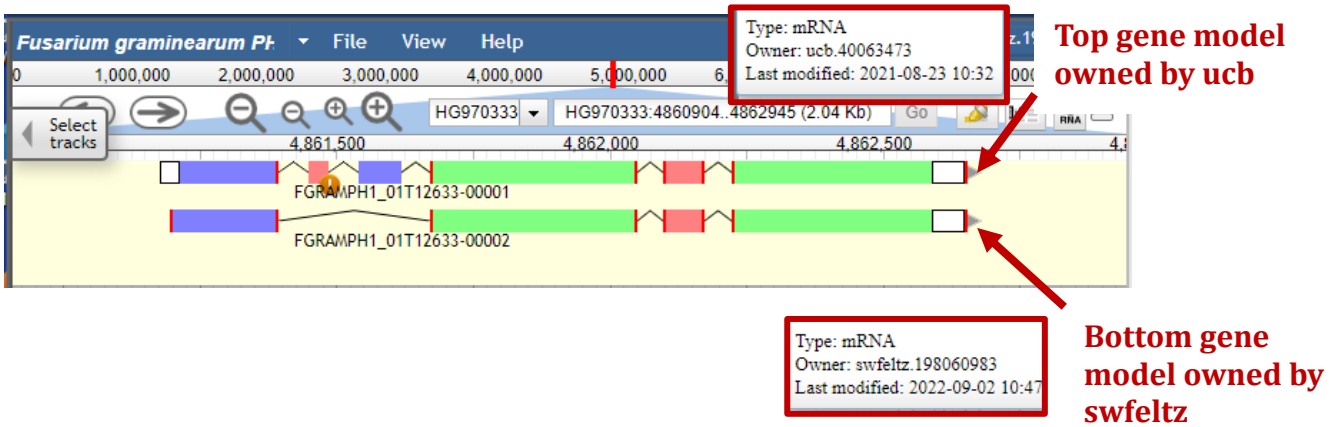
5. Add the gene model into the User-created Annotations area

a. Select the gene model by clicking on one of the introns or double-click on an exon.

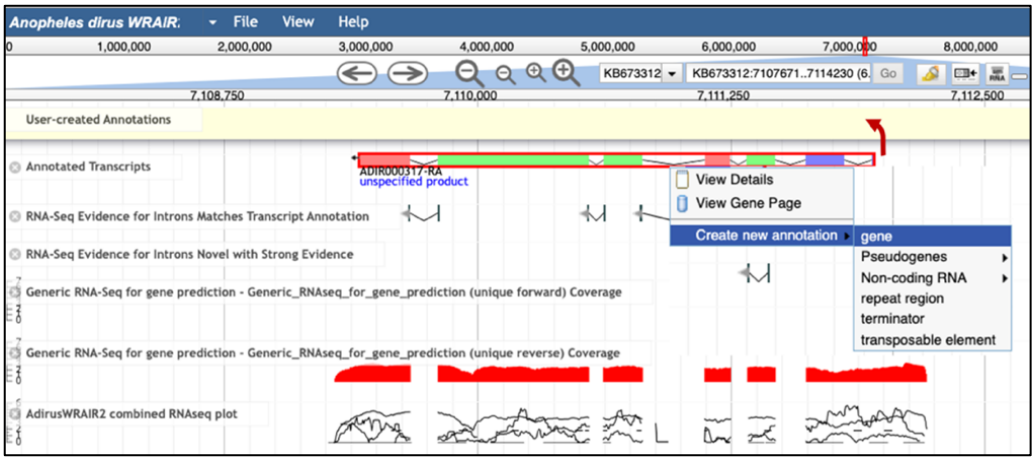
b. Once you see a red-border around the gene you can drag and drop the gene model into the pale-yellow User-created Annotations track.

i. Instead of dragging the gene model, you can also right click on the gene in the Annotated Transcripts track, select from the drop-down menu **Create new annotation** > **gene**.

Note: If other users are correcting the same gene model, you may see their genes models in the User-created Annotations track. Hover over the gene model for a popup of details which contain the owner ID. You will want to make your changes on your own model 😉.

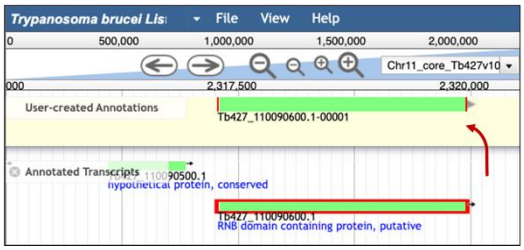


Note: You cannot drag up a single exon. Always select the whole gene and add it into the yellow User-created Annotations area.

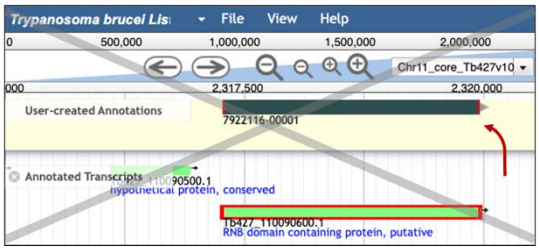


In the event that your gene has a single exon, you need to double-click on the gene. Once you see the red border drag the gene into the pale-yellow User-created Annotations track. The gene model should have a red, green or blue colour indicating the different frames. It should not be a grey box.

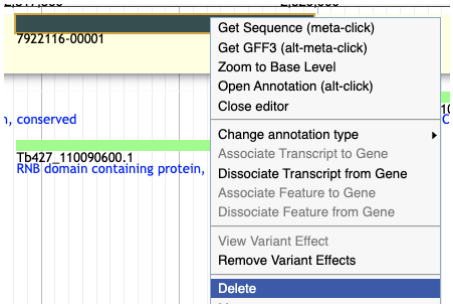
Correct



Incorrect



If you see a grey box when you are trying to drag a single exon gene, you need to delete it and try again. Select the grey box. With a right-click open the menu and choose Delete.



6. **Modify the gene model:** Not every gene model has the same problem. Some may have exon boundaries that do not match the coordinates in the JBrowse tracks. Some may need to be split into two genes, or merged into one gene. Some will have a combination of problems.
   a. **INSPECT YOUR GENE MODEL!** Look at the data tracks such as the intron evidence tracks and the other RNA seq tracks you may have turned on in step 4. What errors do you think need to be corrected?

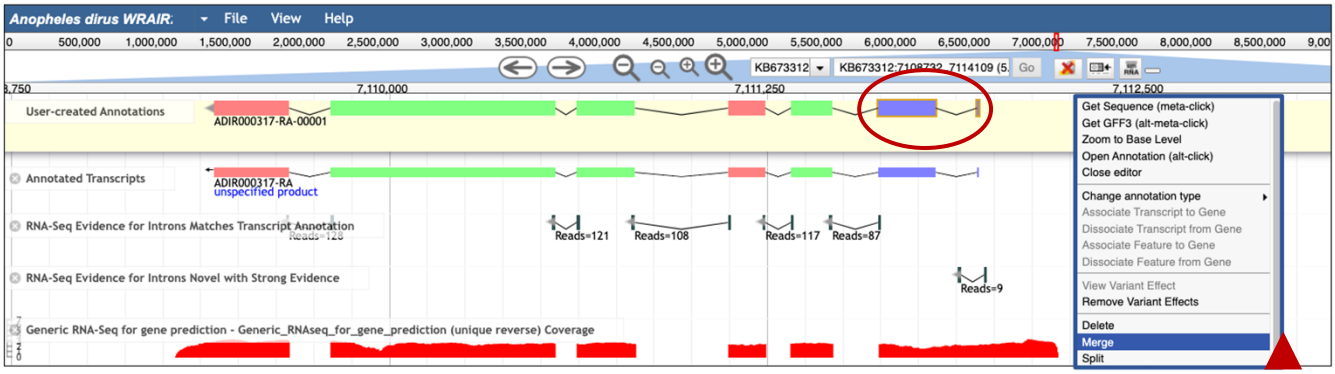
Please note: Always inspect the **RNA-Seq Evidence for Introns Novel with Strong Evidence** track. This track indicates if your gene of interest has possible missing exons, incorrect exon/intron boundaries or a possible alternative transcript.

b. Below are examples of 10 types of gene problems you may encounter when fixing your genes. Each offers screenshots illustrating how to make edits in Apollo. Choose a hyper link in the list below or scroll down to view the examples and screenshots
   - Incorrect intron (see 6.1)
   - Incorrect exon/intron boundary (see 6.2)
   - Missing exon (see 6.3)
   - Missing several exons (see 6.4)
   - Genes need to be merged (see 6.5)
   - Genes need to be split (see 6.6)
   - Missing alternative transcript (see 6.7)
   - Missing gene model (see 6.8)
   - Incorrect start codon (see 6.9)
   - Missing UTRs or incorrect UTR boundaries (see 6.10)

c. You can also have a look at the VEuPathDB Apollo YouTube channel. There are short screencasts showing how to modify gene models in Apollo: https://www.youtube.com/playlist?list=PLWzQB3i5sYALdtuACxZRowVoghLimhwx

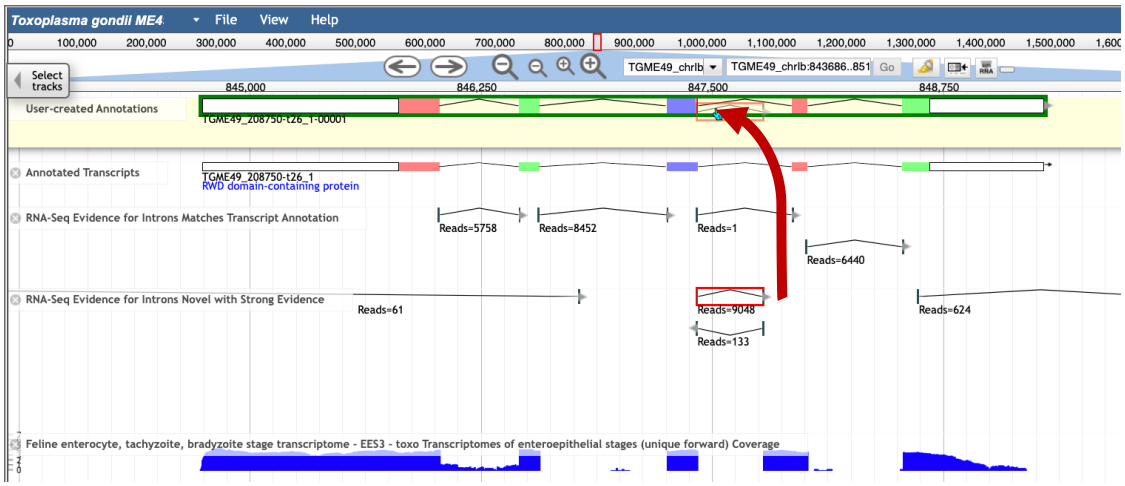
6.1) Incorrect intron

Select the gene model in the pale-yellow User-created Annotations area and with a right-click access the menu. If an intron is not supported select the two exons surrounding the intron by using the shift key and select merge from the menu. Alternatively, select one of the exons, go to the edge of the exon until a little arrow appears and then extend the exon until it overlaps with the second exon.

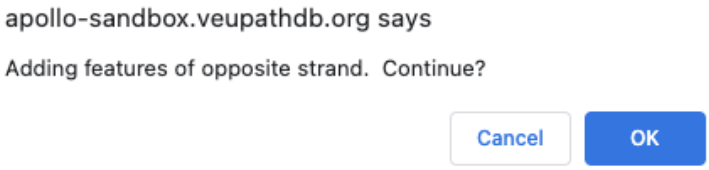


6.2) Incorrect exon/intron boundary

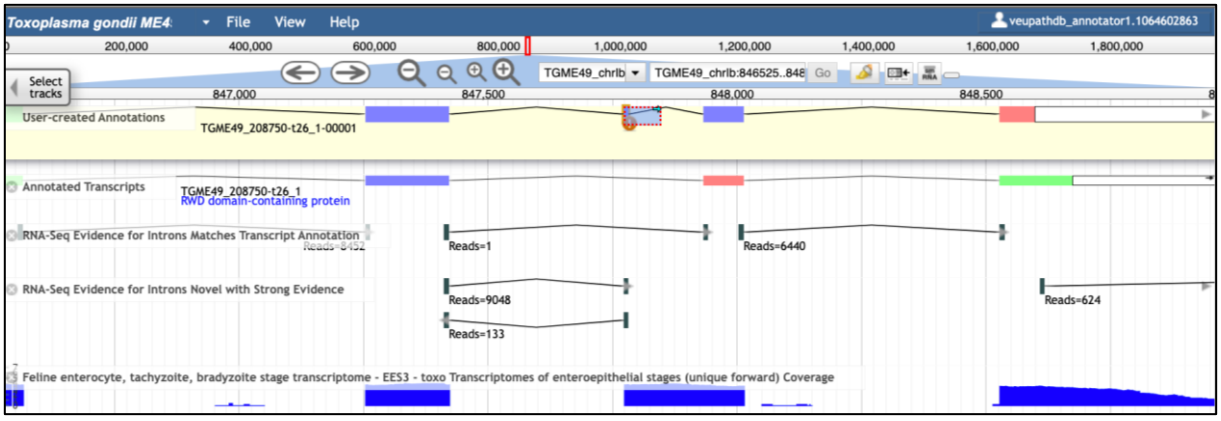
6.2.a Select the intron evidence, drag it into the user-created Annotations area and drop it onto the gene model. The gene model will get a green border.



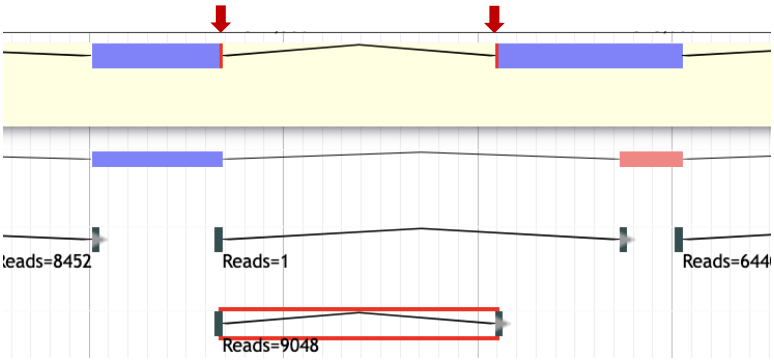
Select OK when you get the message "Adding features of opposite strand".



6.2.b Select one of the exons, go to the edge of the new exon until a little arrow appears and extend the exon until it overlaps with the second exon. Alternatively, you can also use the shift key to select both exons, with a right-click open the menu and choose merge.

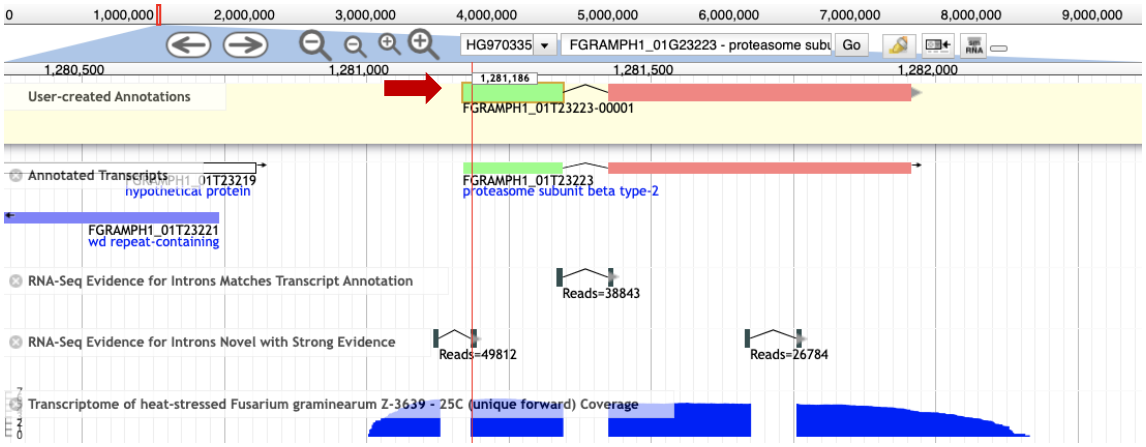


You can recheck if your new gene model is using the correct splice sites by clicking on the intron junctions in the RNA-Seq Evidence for Introns tracks. Red exon borders show up if the splice site corresponds to the evidence.

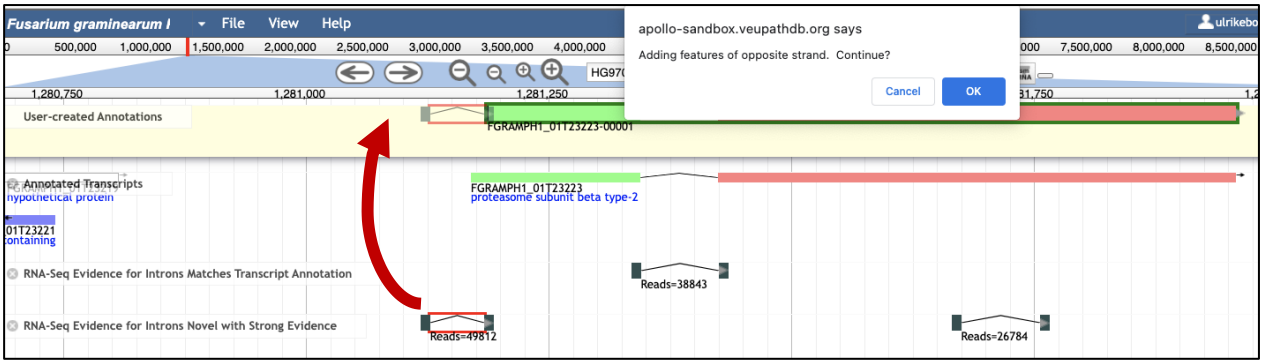


6.3) Adding an exon. In this example, the RNA seq data and the Intron Evidence tracks indicate that there should be another exon. The exon that is annotated is too long and there should be a second exon upstream of that. These instructions outline how to shorten an exon and add a new one.

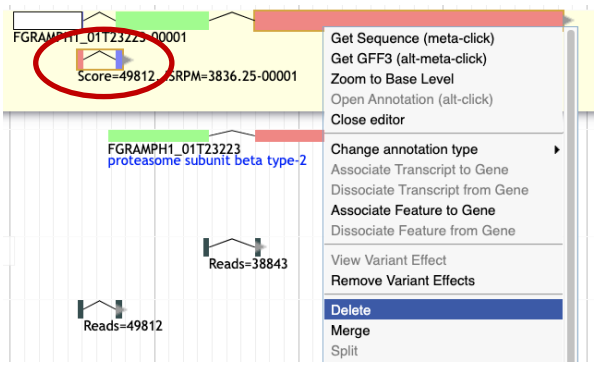
6.3.a Select the existing exon. Point your mouse at the edge of the exon and move the exon boundary so that it fits with the RNA-Seq evidence.



6.3.b Select the intron evidence. Drag it into the user-created Annotations area and drop it onto the gene model. The gene model will get a green border. Select OK when you get the message "Adding features of opposite strand".

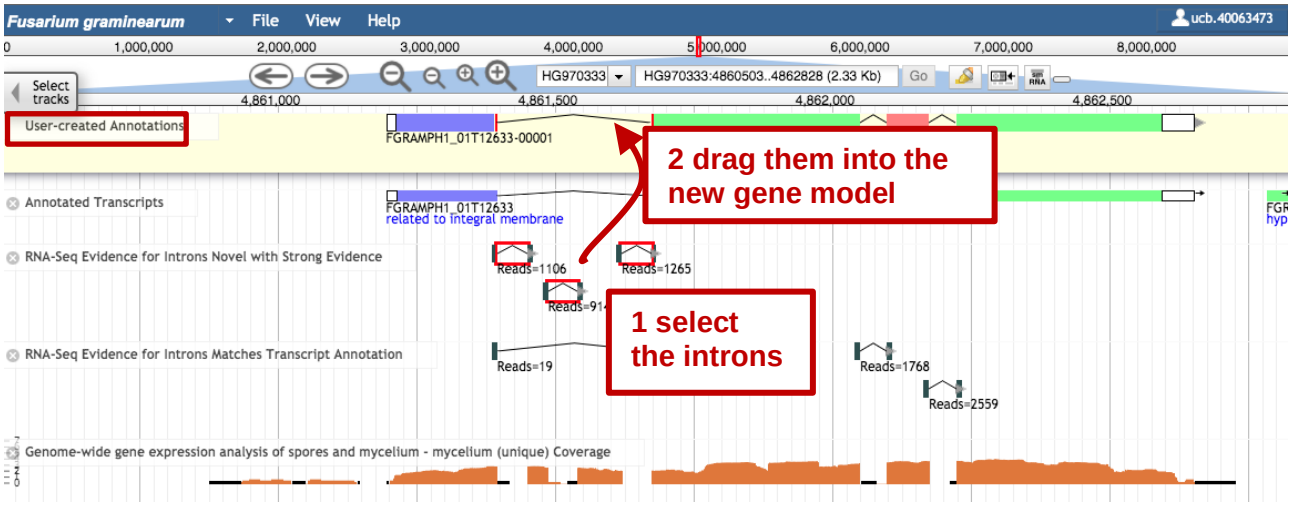


If the intron is being displayed as a separate unit within the User-created Annotation space, right click to delete the intron from the User-created Annotation space and try again. Make sure to drag and drop the intron evidence directly onto the gene. The gene will get a green border.

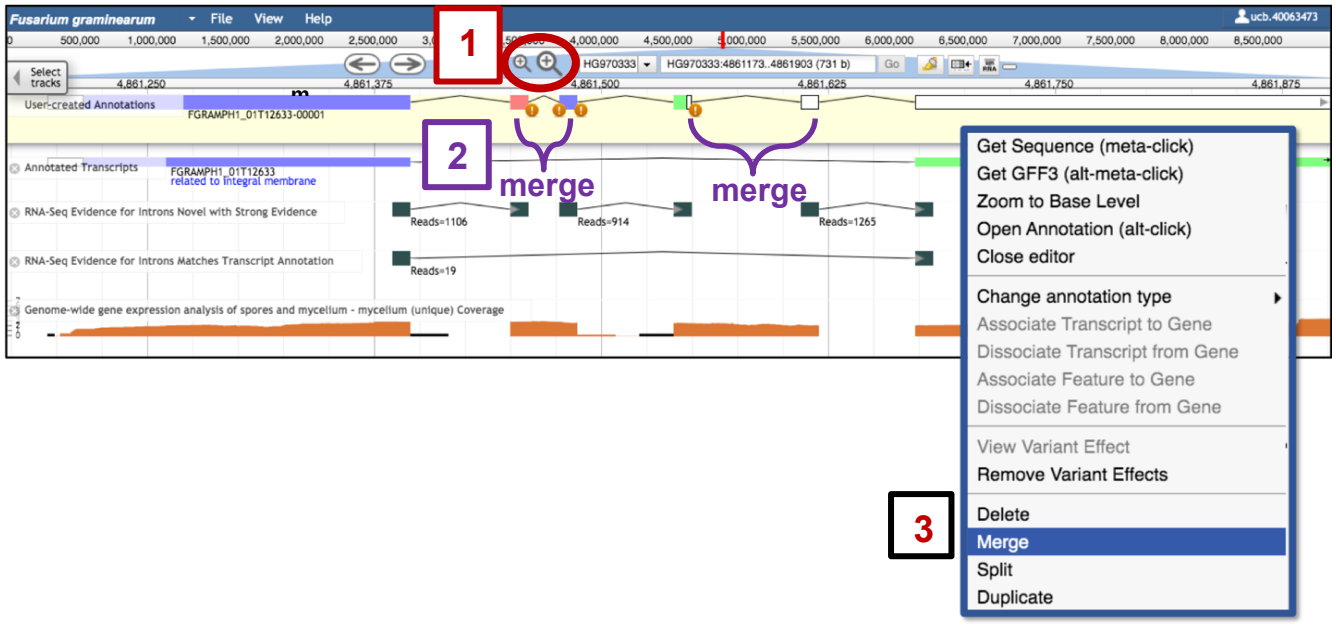


6.4) Adding two or more exons

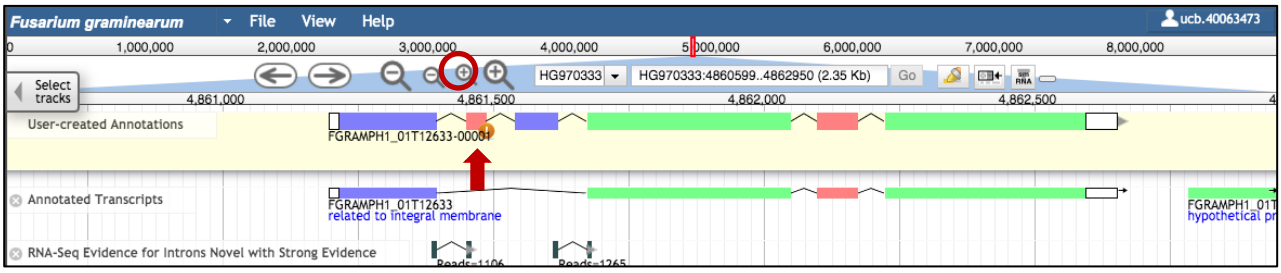
6.4.a To create new exons drag and drop the intron junctions into the User-created Annotations area. You can either select the intron junctions individually or hold down the shift key and select all intron junctions with strong evidence (**1**), drag and drop them into the gene model (**2**). The gene boundaries will show up in green when dragging and dropping.



6.4.b Zoom in by clicking on the **+** sign on the top (**1**). Press the Shift key and select the exons that should be merged (**2**). With a right-click open the drop-down menu and choose **Merge** (**3**). Alternatively, select one of the exons you would like to merge, go to the edge of the feature until a little arrow appears and extend the exon until it overlaps with the second exon.



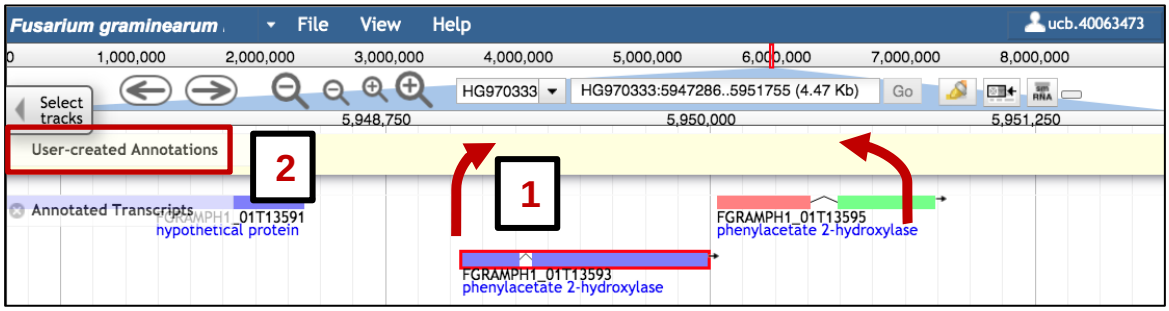
The exclamation marks at the bottom edge of some exons informs about non-canonical splice sites, i.e. GC splice sites. Zoom in by clicking on the + sign and the sequence should become visible. You can check if it is a valid splice site.



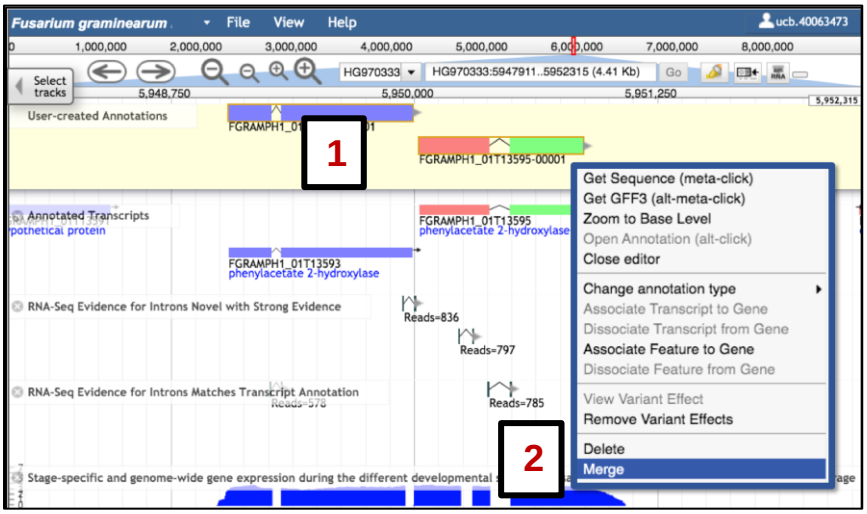
More information on splice sites can be found on this FAQ:
https://veupathdb.org/veupathdb/app/static-content/faq.html#apollo6

6.5) Merging two genes

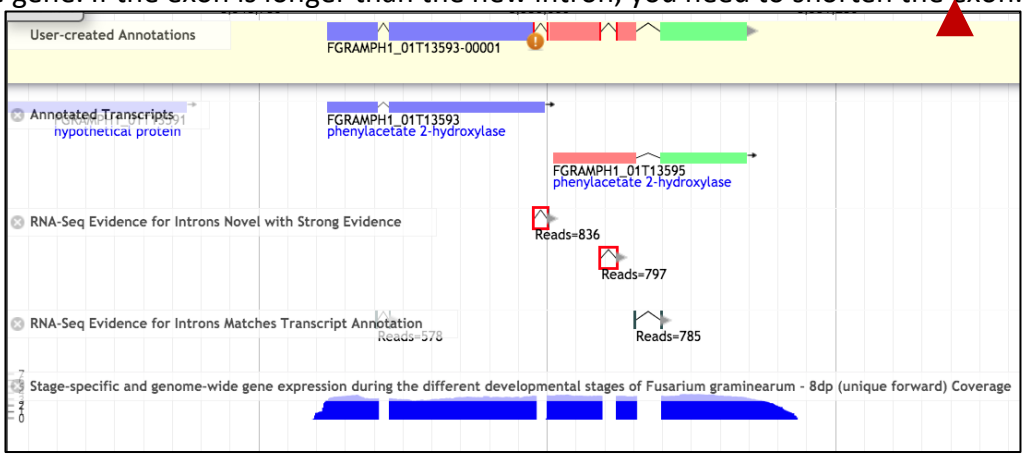
6.5.a Select the gene models that you would like to merge (**1**). Drag and drop the genes into the User-created Annotations track (**2**).



6.5.b Now that both genes are in the User-created Annotations track, hold down the shift key and select both gene models (**1**) in the yellow User-created Annotations area. With a right-click open the drop-down menu and choose **Merge** (**2**).

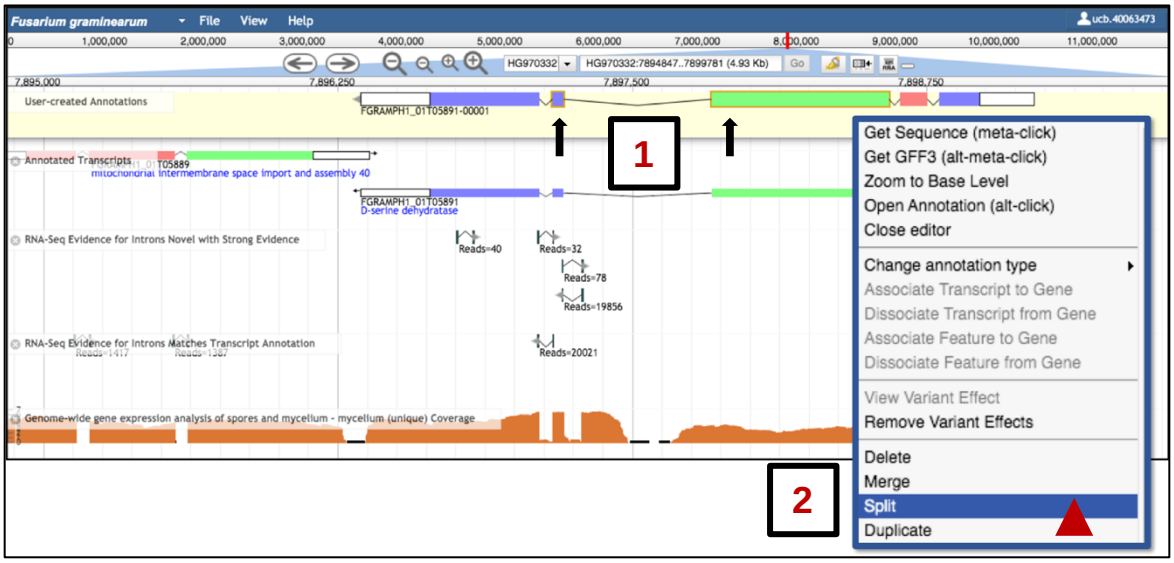


6.5.c Drag and drop the intron junctions into the User-created Annotations area and merge them with the gene. If the exon is longer than the new intron, you need to shorten the exon.

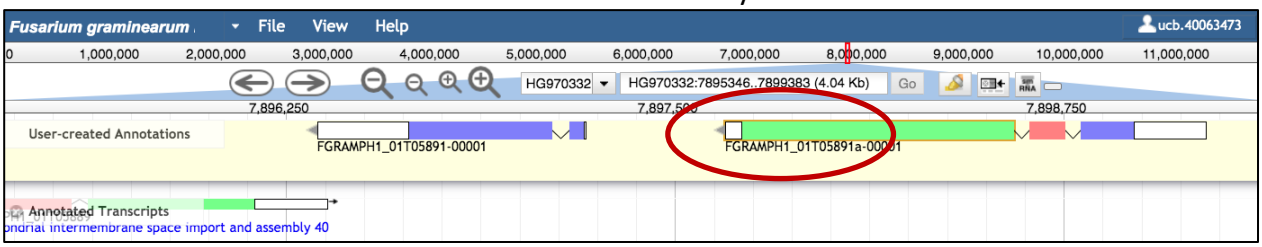


6.6) Splitting genes

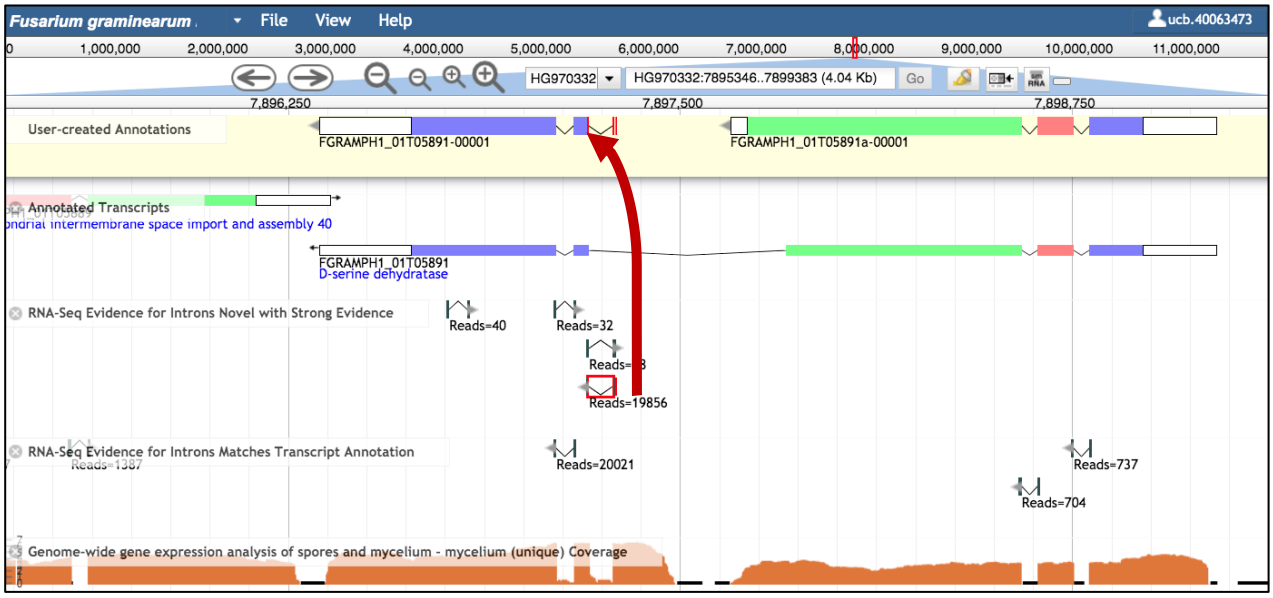
6.6.a To split the gene model, hold down the shift key, mark the two exons that border the intron that should be split (**1**). With a right-click open the annotation drop-down menu and select **Split** (**2**).



6.6.b Now that the gene model has been split, the newly created genes need a correct start codon and stop codon. To create the stop, click with your mouse at the end of the gene model and extend it. The 3'UTR will be created automatically!

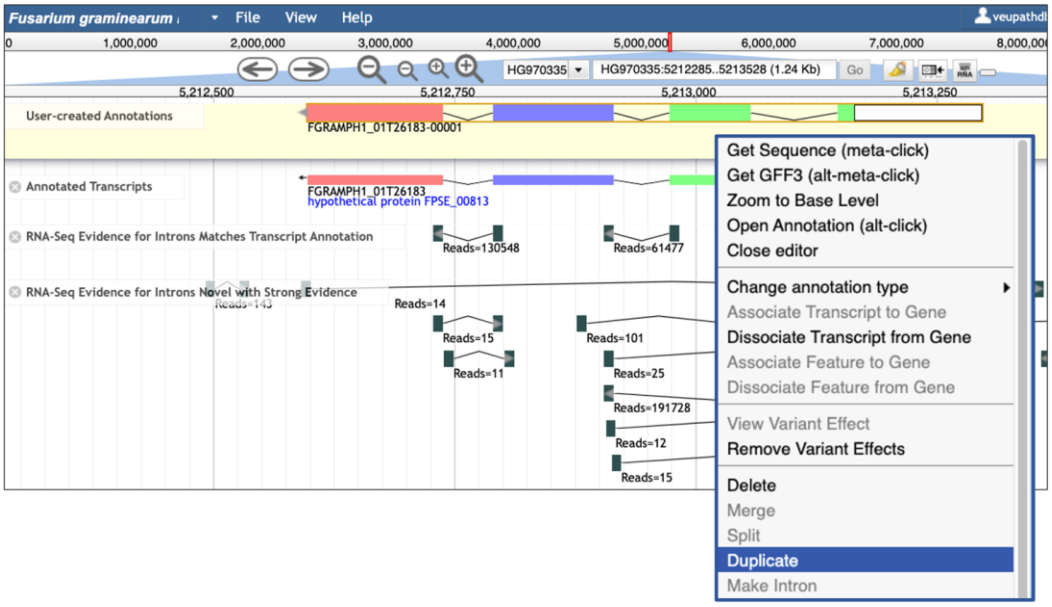


6.6.c The gene model on the left, needs a start codon. Drag and drop the splice junction into the annotation area. Move it up and hover the splice junction over the gene. A green border will show up which indicates that the intron junction has been merged with the gene. Now extend the first exon to include a UTR.

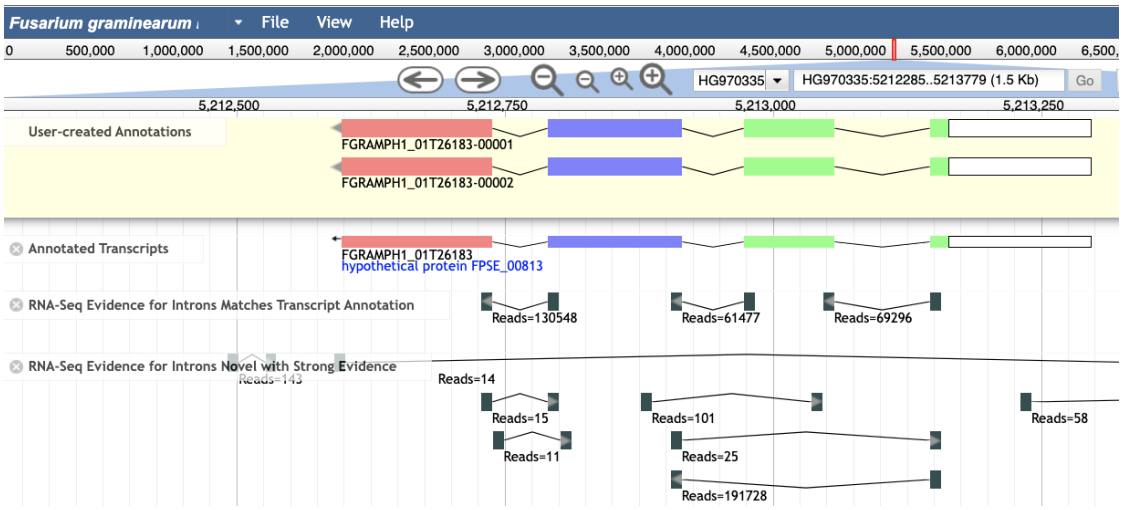


6.7) Adding alternative transcripts

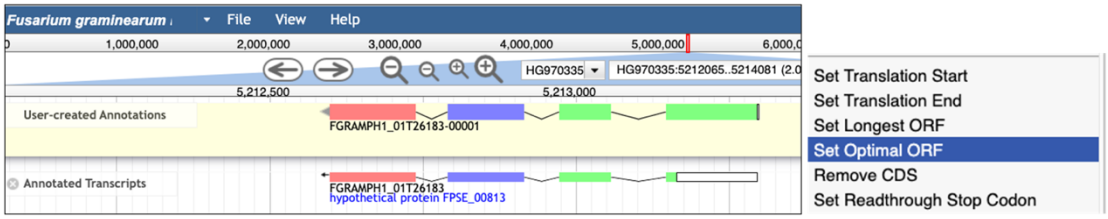
6.7.a Select the gene in the User-created Annotations area, with a right-click open the drop-down menu and choose duplicate.



6.7.b There are now two transcripts with different transcript ID extensions: 00001 and 00002. You can now start to modify the alternative transcript.

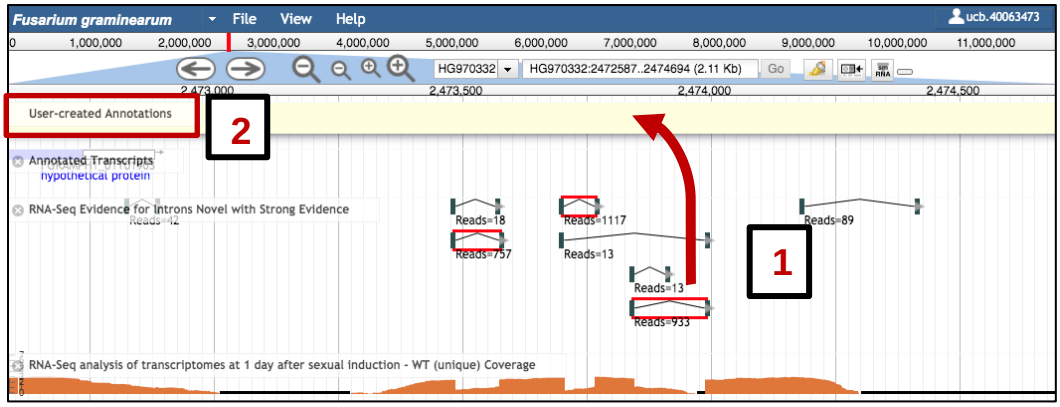


Please note, when adding the gene model into the user-created annotations area, Apollo sometimes modifies the gene so that the start codon is lost. To correct this, use the option **Set Optimal ORF** from the right-click menu. This option will automatically create the longest ORF.

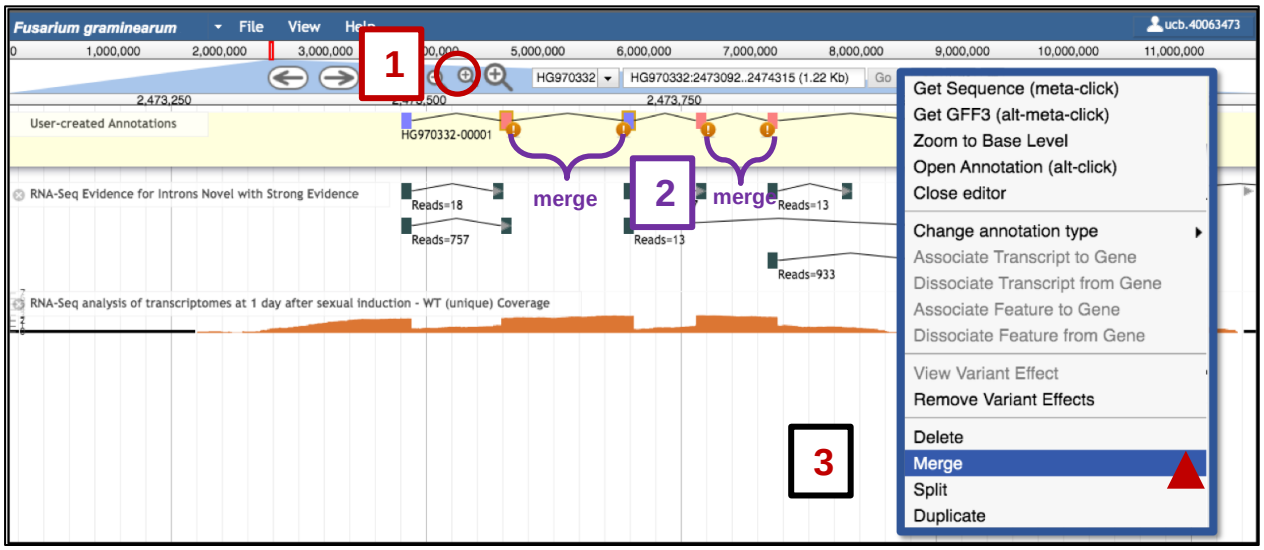


6.8) Creating a new gene

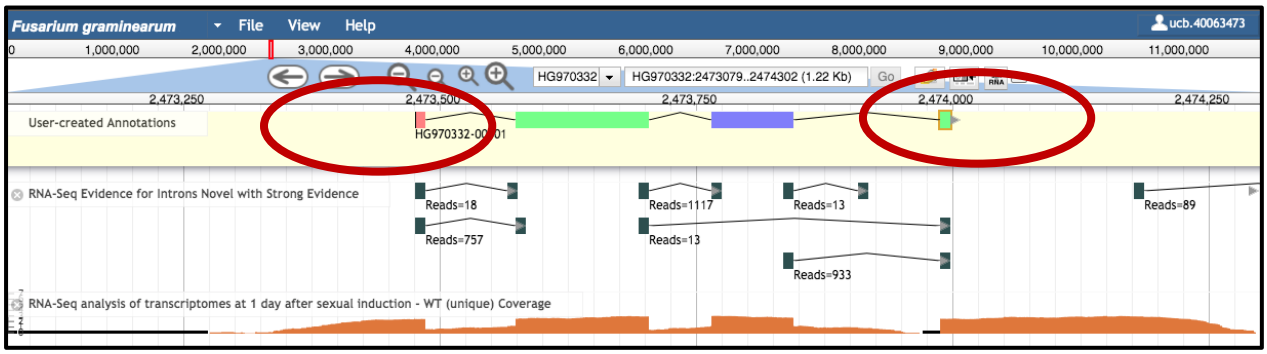
6.8.a Select the supporting evidence, i.e. intron junctions (**1**). Use the shift key to select more than one intron junction. The selected intron junctions will show up with a red border. Drag and drop them into the User-created Annotations track (**2**).



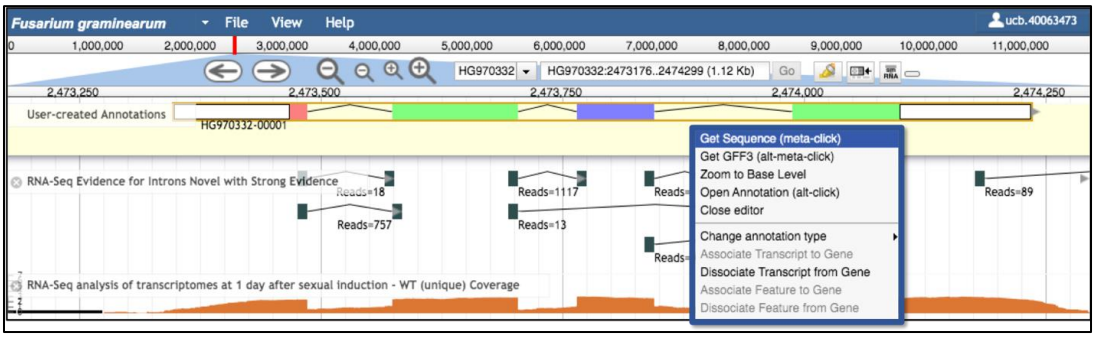
6.8.b Zoom in by clicking on the **+** sign on the top (**1**). Press the Shift key and select the exons that should be merged (**2**). With a right-click open the drop-down menu and choose **Merge** (**3**). Alternatively, select one of the exons you would like to merge, go to the edge of the feature until a little arrow appears and extend the exon until it overlaps with the second exon.



6.8.c Select the first and the last exon, go to the edge of the exon until a little arrow appears and extend it to the start/end. UTRs will be created automatically.

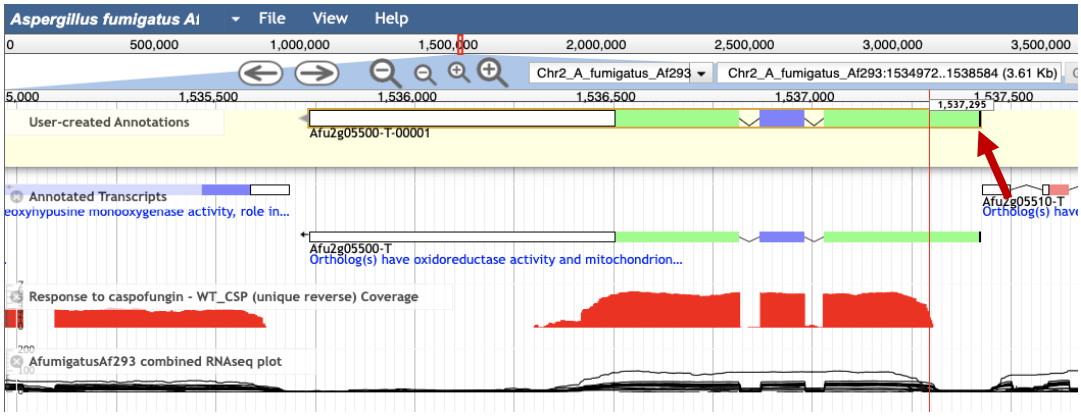


6.8.d Select the new gene model, with a right-click open the annotation drop-down menu and select **Get Sequence**. Copy the sequence, run blast (https://blast.ncbi.nlm.nih.gov) and Interpro (https://www.ebi.ac.uk/interpro) to get more information about the new gene.

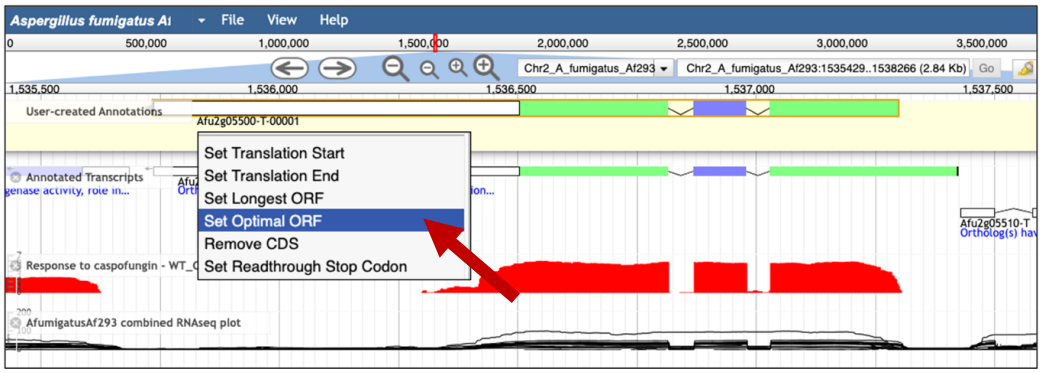


6.9) Incorrect start codon

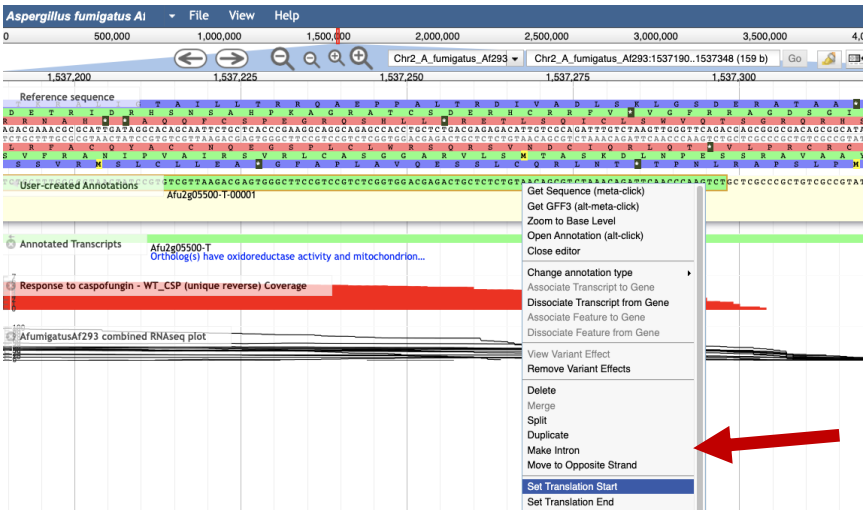
6.9.a Select the gene model in the user-created Annotations area, hover at the end of the first exon and move the exon boundary, so that it fits with the transcript evidence.



6.9.b Open the right-click menu and choose **Set Optimal ORF**. With this option Apollo automatically creates the longest CDS.

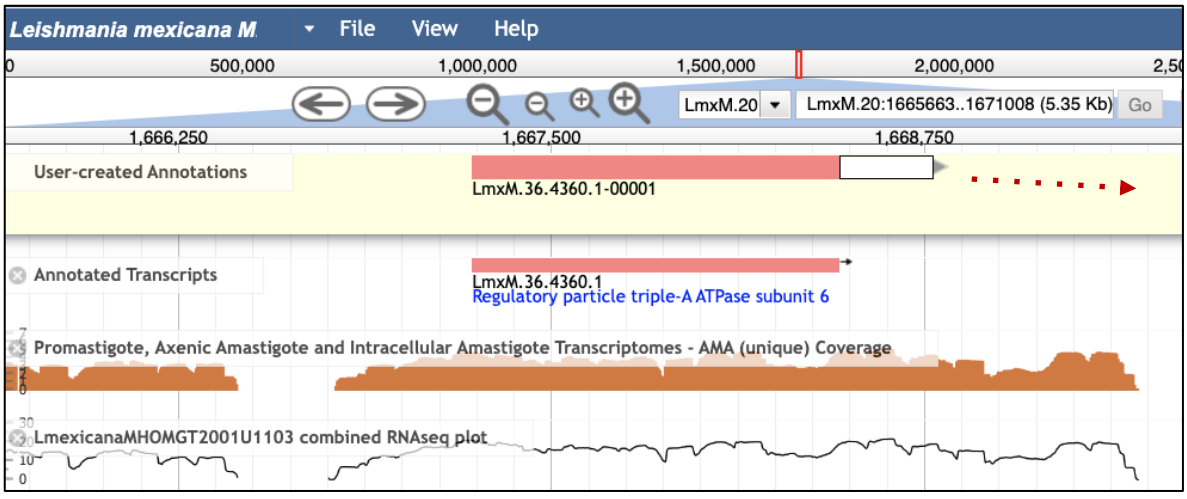


6.9.c Alternatively, you can also zoom in, click on the A of the ATG that you would like as start. Use the option **Set Translation Start** from the right-click menu.

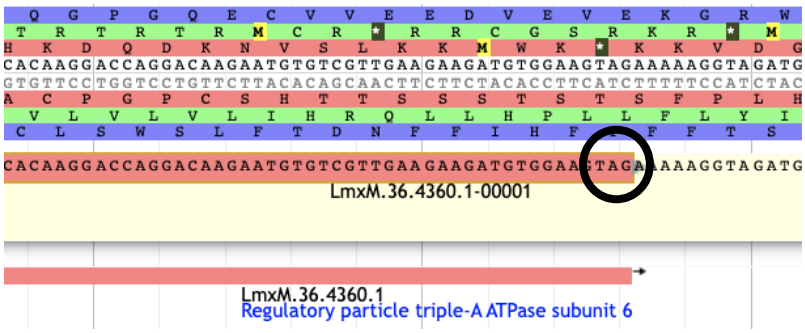


6.10) Adding UTRs

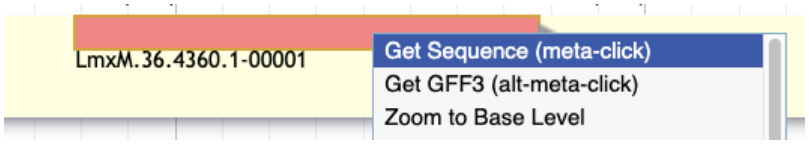
6.10.a Point your mouse at the edge of the feature, a little arrow will appear. Extend the exon to the transcription start/end. Apollo will automatically create UTRs (shown in white).



If you are adding UTRs you don't have to be concerned that your gene model is missing a start and stop codon. If you are not including UTRs always make sure there is a start (ATG) and stop codon (TAA, TAG, TGA). To recheck zoom into the sequence and check.

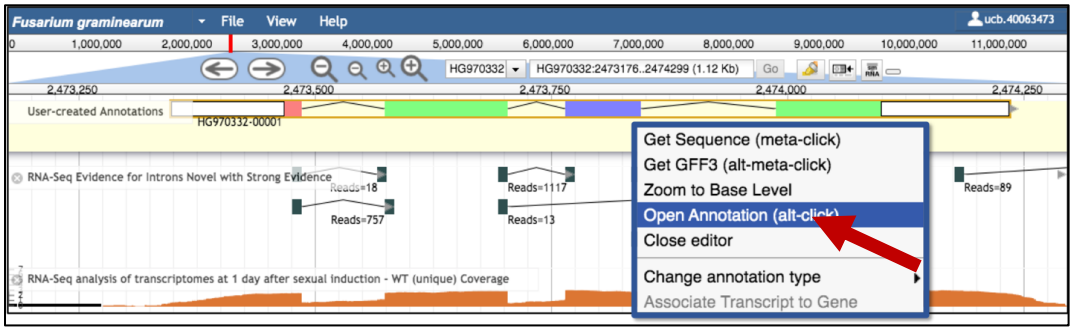


To recheck about the start, you can use the right-click menu and select **GET Sequence** to see if your gene starts with a Methionine.

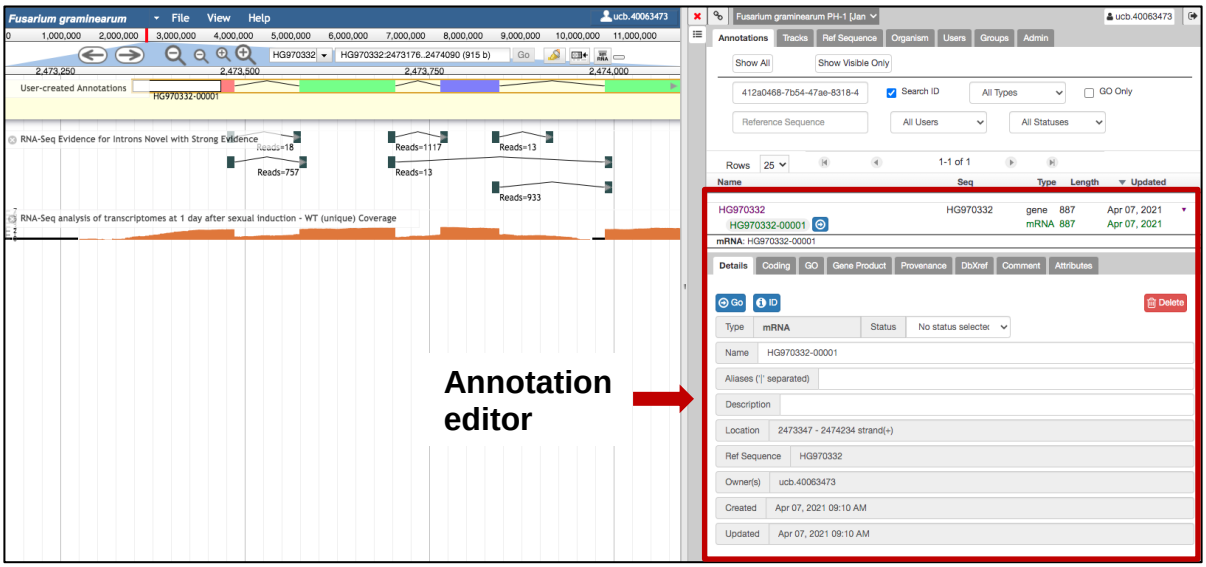


7) Opening of the Annotation editor window

7a Select the gene in the User-created Annotation track and with a right-click open the drop-down menu and choose **Open Annotation**. Alternatively, you can use the short cut **alt-click**.

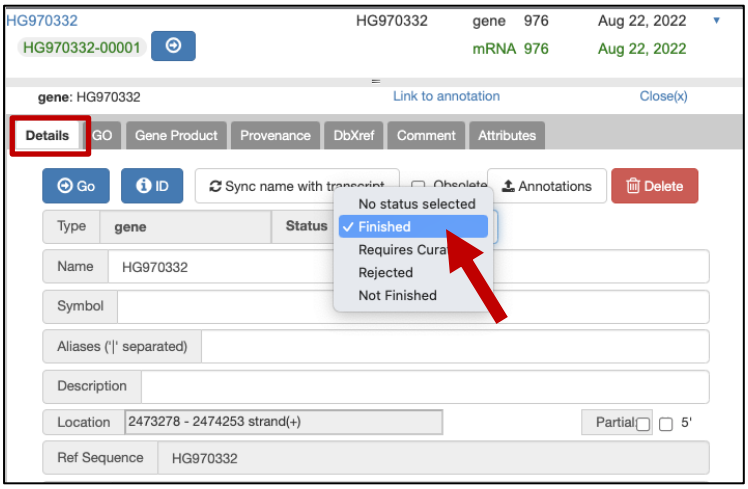


The annotation editor window is now shown on the right-hand side.



8) Finalising the structural annotation

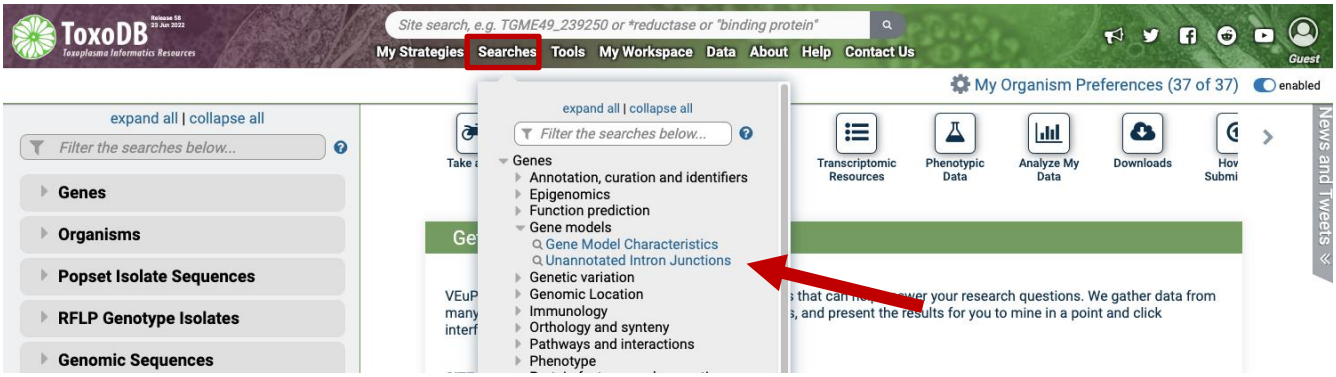
8.a Add a product description for new genes, split and merged genes. Finally go to the Details tab and select the status **Finished**.



Done! Additional information, i.e. tutorials can be found on the following site:
https://veupathdb.org/veupathdb/app/static-content/apollo_help.html

Unannotated Intron Junction search

Want to edit more genes in your favourite model organism? Use the Unannotated Intron Junction search to find genes that have high confidence novel intron junction spanning RNA Sequencing reads.



Information on this search can be found here:
https://static-content.veupathdb.org/documents/Intron_junctions_search_07_08_2022.pdf