Annotation Example

This example shows the annotation of STARD9 in *S. scrofa*, for which there is a split/merge disagreement between Ensembl and RefSeq. This problem was first discovered by viewing the Gene Prediction Problems tracks in chromosome 1. To do this example, you must use the *Sus scrofa* demo browser, which you can access using the Tutorial and Demo pulldown menu (Figure. 1).

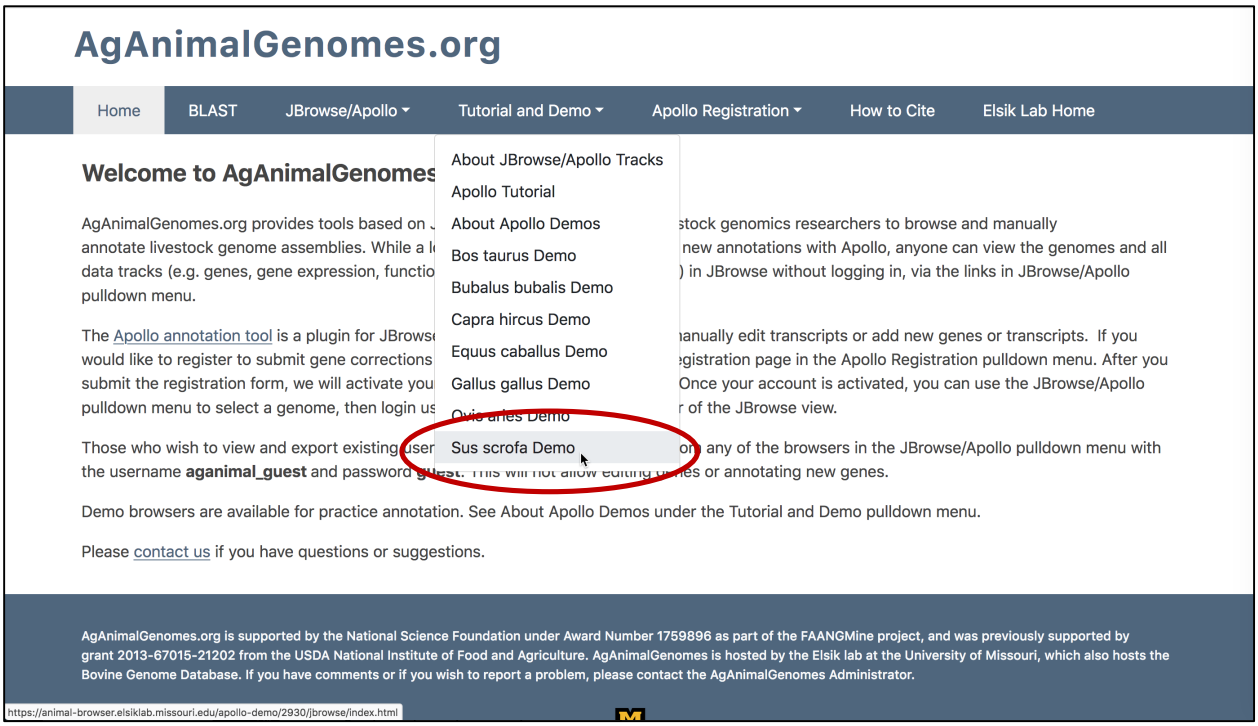

Figure 1. Selecting the *Sus scrofa* demo browser.

After selecting the browser in the pulldown menu (Figure 1), you will see a JBrowse view and will not yet be logged into Apollo (Figure 2). The first time you access the browser, no tracks will be visible. For future access, the last tracks you viewed in the previous session will be visible.

Click the *Select tracks* button (circled in red in Figure 2) to open the Faceted Track Selector. Once in the Faceted Track Selector, click the Gene Prediction Problems category on the left for filter for those tracks, then check the box to the left of both Gene Prediction Problems tracks (Ensembl Protein Coding Discordant and RefSeq Protein Coding Discordant). Checking the box next to Data Type above the table will cause all the rows to be selected. Selected rows will be highlighted with a blue background. Click *Back to Browser* above the table to close the Faceted Track Selector.

Back in the browser you will see the selected tracks. Most likely you will be zoomed out and the tracks will either look like histograms or many small gray arrows, depending on the gene density in each track (Figure 3A). Enter STARD9 in the search box and click Go (Figure 3B) to navigate to this gene. Searching for this gene automatically opens the RefSeq Protein Coding track, because STARD9 is found in this track as well as the RefSeq Protein Coding Discordant track. The RefSeq Protein Coding track is not needed, so remove it by clicking the X in the track label (Figure 3C).

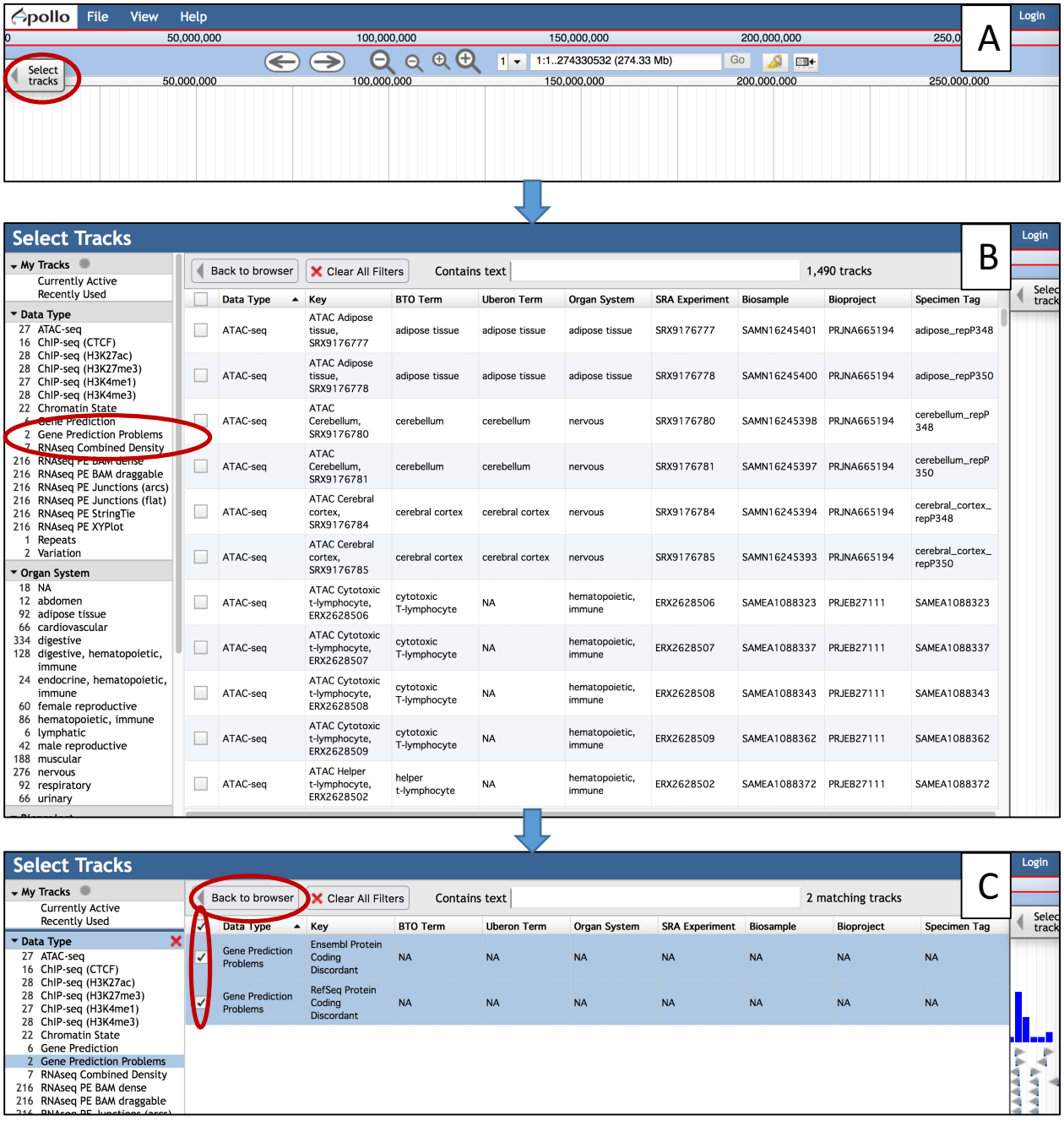

Figure 2. A) The initial JBrowse view showing no tracks. The Faceted Track Selector is opened by clicking *Select tracks* on the left of the browser (circled in red). B) The Faceted Track Selector. Clicking Gene Prediction Problems on the left (circled in red) to filter for those tracks. C) The Faceted Track Selector after filtering for Gene Prediction Problem tracks. Both tracks (Ensembl Protein Coding Discordant and RefSeq Protein Coding Discordant) are selected. Once a track is selected, its row in the table is highlighted in blue. The Faceted Track Selector is closed by clicking *Back to browser* (circled in red).

Figure 3. A) The view after selecting the Gene Prediction problem tracks. When you are zoomed out the tracks with a high density of features, similar to the Ensembl Protein Coding Discordant track, look like histograms representing gene density, similar to Ensembl Protein Coding Discordant track. Tracks with lower density, such as the RefSeq Protein Coding Discordant track show individual genes appearing as gray arrows. B) Entering a gene symbol (STARD9) in the search box (circled in red) to navigate to that gene. C) The result of navigating to the STARD9 gene. The RefSeq Protein Coding Gene track was opened automatically because STARD9 is also found in that track. The track is not needed and can be removed by clicking the X in the track label (circled in red).

Figure 4A shows the gene region of interest. Verify that the Ensembl transcripts are considered two different genes by viewing the details for each one. For each transcript, right-click the transcript and then select *View details*. Figures 4B and 4C show that the transcripts are from two different genes. This is an example of a split/merge disagreement between Ensembl and RefSeq and should be annotated with Apollo. Click Login at the upper right of the browser window to login to Apollo (Figure 4A). The credentials for the Sus scrofa demo browser are username: aganimal_demo_guest and password: aganimal_demo.

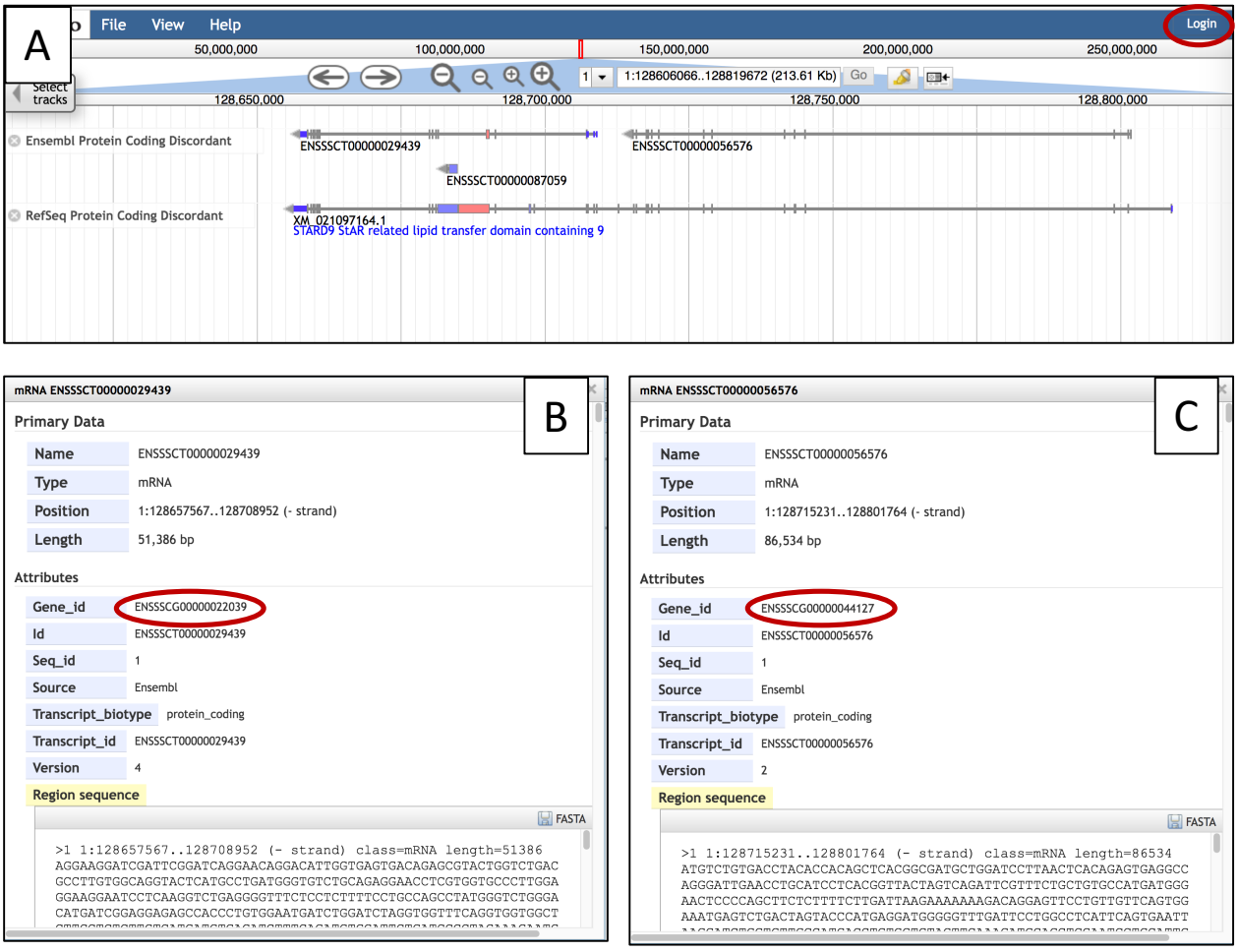

Figure 4. A) JBrowse view after removing the RefSeq Protein Coding track. The login button is circled in red on figure right. B) The information panel shown when right-clicking the Ensembl transcript on the right (ENSSSCT00000029439), showing the gene id ENSSSCG00000022039. C) The information panel shown when right-clicking the Ensembl transcript on the right (ENSSSCT00000056756), showing the gene id ENSSSCG0000005676, confirming that these transcripts are considered to be encoded by different genes.

After logging into to Apollo, the view is split between the browser on the left and the Information panel on the right. Also the Select tracks button is no longer visible (Figure 5A). Click the small list icon in the upper left of the Information Panel to bring back the Select tabs button (Figure 5A). Although the Information Panel includes a Tracks tab that lists the tracks, this list is not easily viewed or search. It is recommended that you use the Faceted Track Selector rather than the Tracks tab in the Information Panel.

After bringing back the Select tracks tab, click the red X in the upper left of the Information Panel to close the panel (Figure 5B).

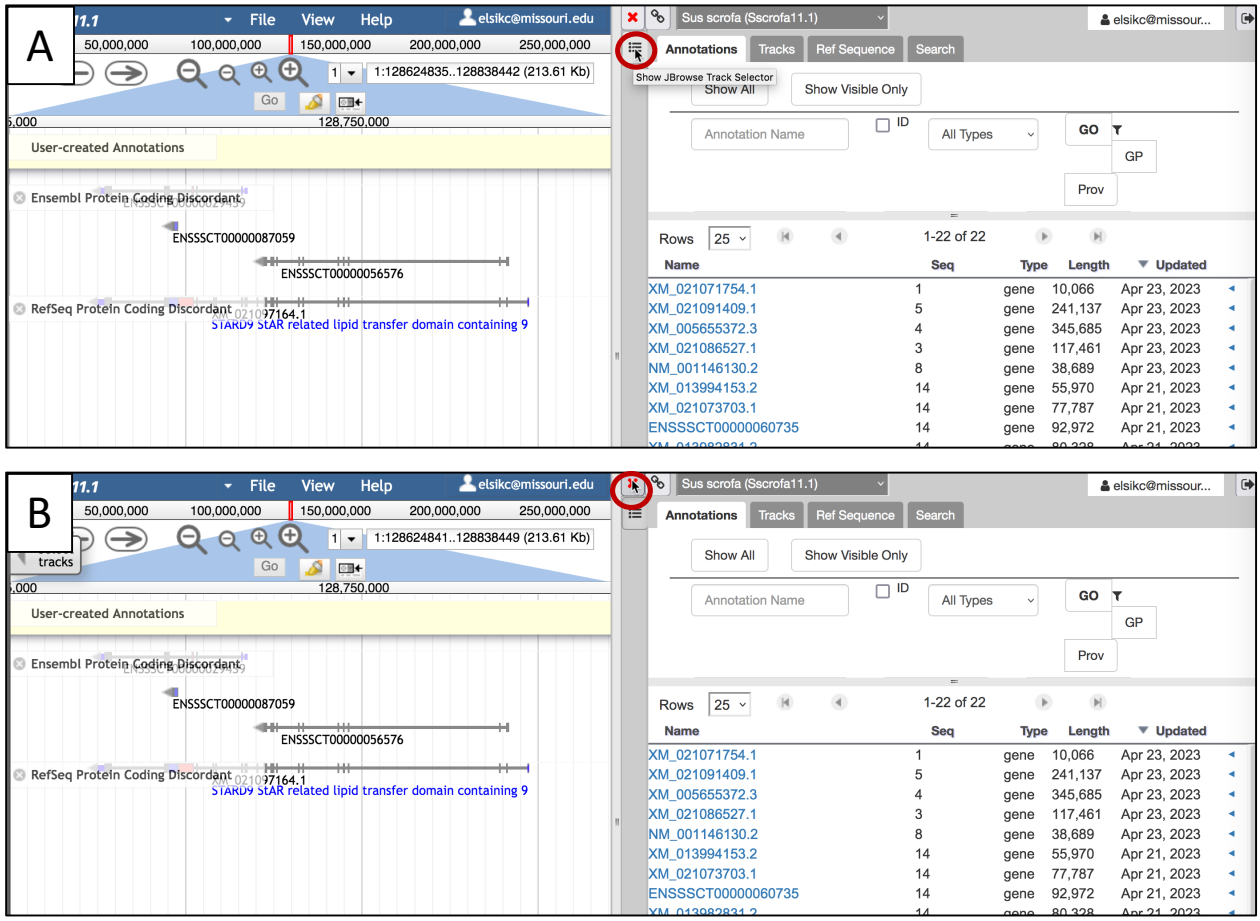

Figure 5. Apollo view after logging in. The view is split between browser on the left and the Information Panel on the right. Information Panel has tabs for Annotations, Tracks, Ref Sequence (chromosome ids) and BLAT Search. The JBrowse Select tracks button that provides access to the Faceted Track Selector can be toggled on by clicking the list icon right below the red X (circled in red on the figure). B) The Information Panel can be hidden by clicking the red X in the upper left of the panel (circled in red on the figure). When the Information Panel is closed, the red X is replaced by a green-bordered square icon that can be used to re-open the Information Panel (not shown).

Once you have closed the Information Panel, you can zoom further into the gene region as shown in Figure 6.

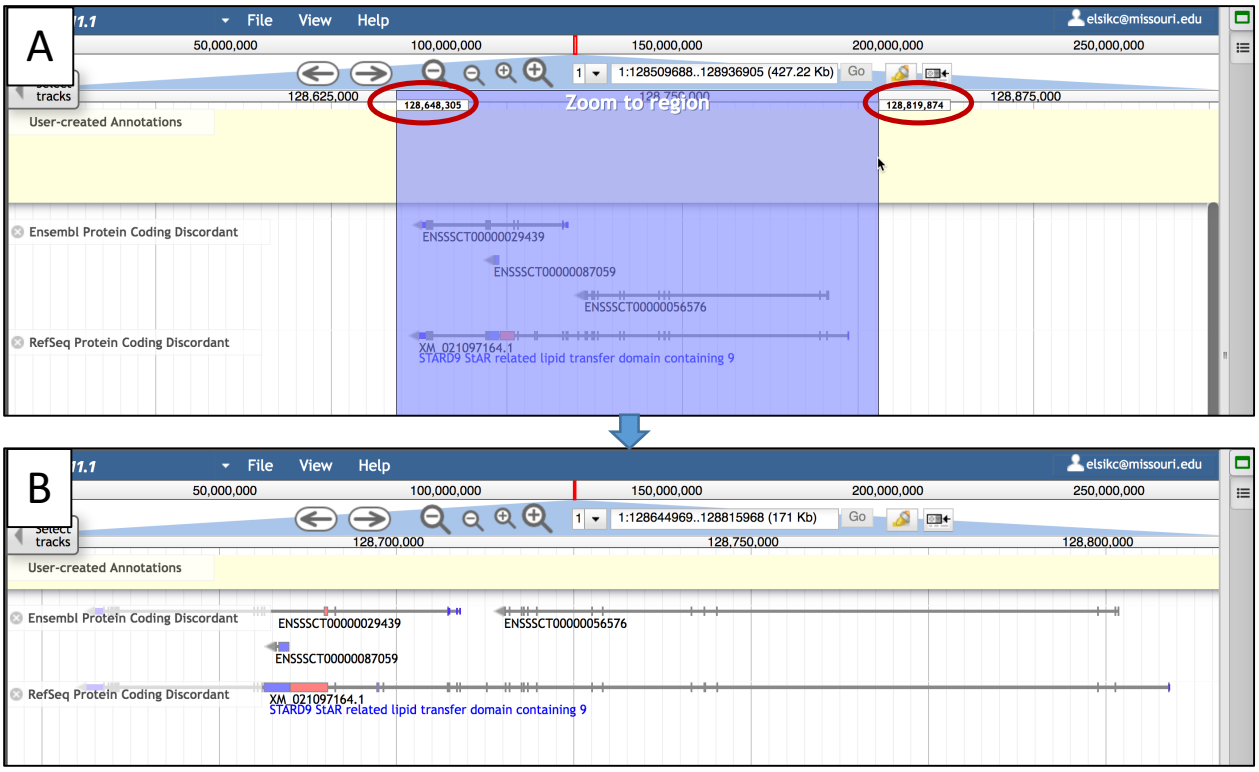

Figure 6. A) The browser after closing the Information Panel. The view is split vertically between the Editing Area with a light yellow background on the top and the Evidence Area with a white background where JBrowse tracks are viewed on the bottom. We can zoom to the region of interest by placing the cursor in the lower number line and dragging it across the region of interest. While doing so, the coordinates of the zoomed view show just below the number line (circled in red on the figure). B) The zoomed-in view.

The next step is to determine which RNAseq experiments will be useful in resolving the split merge disagreement. First, we will use one or more RNAseq Combined Density tracks to identify RNAseq experiments to view. Use the *Select tracks* button to open the Faceted Track Selector. Click RNAseq Combined Density on the left to filter for those tracks (Figure 7A). If the Gene Prediction Problem category is still highlighted, you can click it to un-highlight it. Doing so will remove those tracks from the Faceted Track Selector table, but will not remove them from the browser. The RNAseq Combined Density Tracks must be viewed one-at-a-time because they are large and computationally intensive. We select the track at the top of the list to view first, and then click *Back to browser*. Figure 7B shows the selected track which looks like heatmaps, with darker blue indicating higher read density. On the left side each heatmap is a green rectangle that can be moused over to view the tissue and experiment accession (Figure 7B). We are particularly interested in experiments showing darker blue matching the exons in question (those present in RefSeq but not Ensembl).

After taking note of an informative RNAseq experiment as well as the Bioproject of the combined density track, close the RNAseq Combined Density track. If there were no informative experiments in this track, we would select the next combined density track in the Faceted Track Selector table to view. However, we did find informative experiments, so we will not look at another combined density track. Use the Select tracks tab to open the Faceted Track Selector. Un-highlight *RNAseq Combined Density* and

then click on the Data Type RNAseq PE Junctions (arcs) and the Bioproject PRJEB14330. Enter “lung” in the search box and select both tracks (Figure 8A), then close the Faceted Track Selector. In the browser, zoom in to the discordant region using the small plus sign above the browser (Figure 8B). The arcs appear to be connecting the exons in question (Figure 8C).

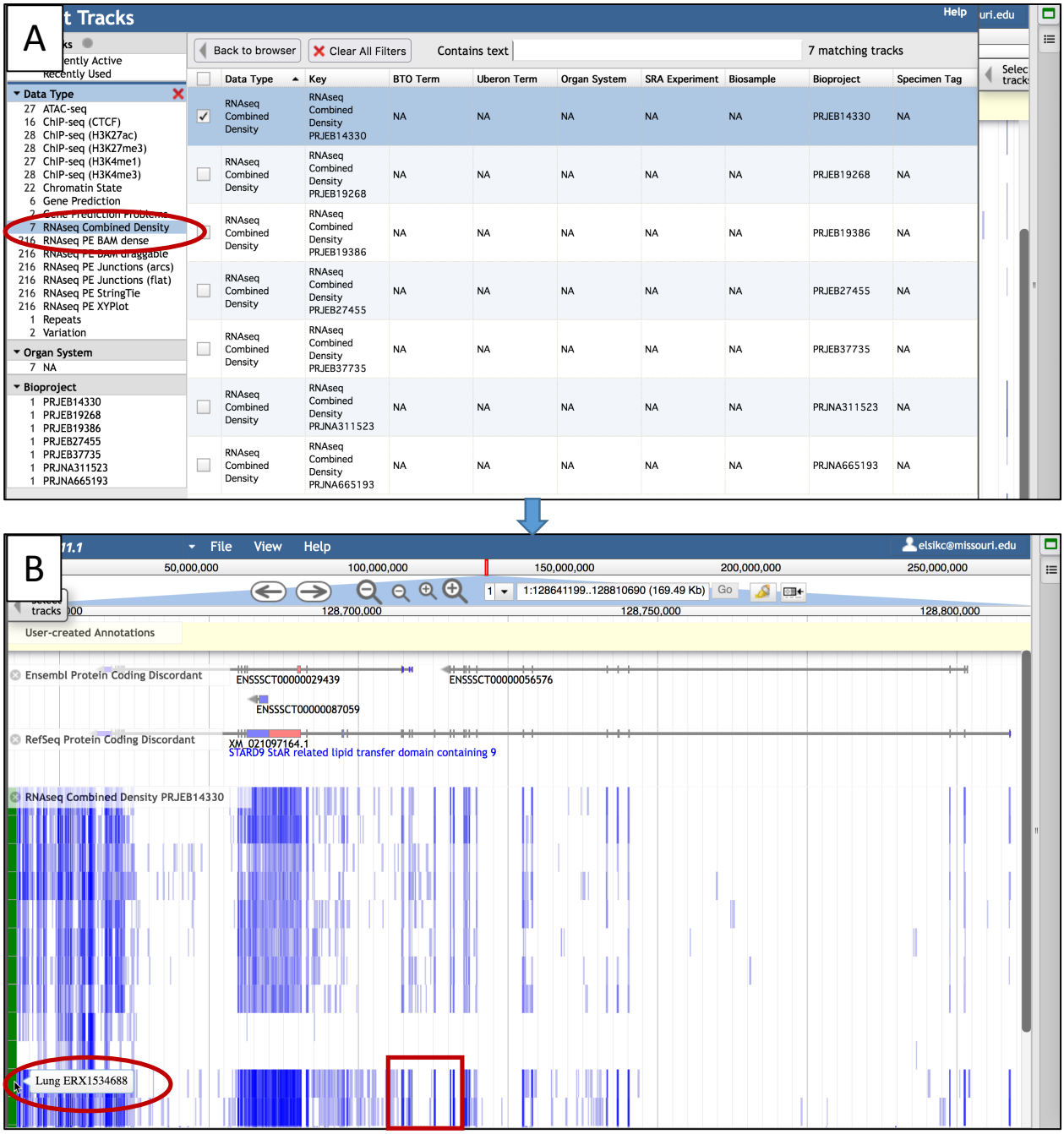

Figure 7. A) The Faceted Track Selector showing the RNAseq Combined Density category selected on the left and one track selected in the table on the right. The combined density tracks are large tracks, and only one can be viewed at a time. B) The highlighted experiment in the track, Lung ERX1534688, and the experiment below it both look informative because there are dark blue marks corresponding to exons that are present in the RefSeq transcript, but not Ensembl (shown in red square).

Figure 8. A) Selecting RNAseq Junctions (arcs) tracks for Bioproject PRJEB14330, filtered for by entering “lung” in the search box (circled in red). B) JBrowse view showing the arc junction tracks. The thickness and color of the arcs represents the numbers of reads supporting the junction, with lighter green/yellow colors supported by fewer reads and darker green supported by a larger number of reads. The light thin arcs in the upper track may be spurious alignments. C) JBrowse view zoomed in more to the exon that is present in RefSeq but not Ensembl.

The arc junction tracks show support for the RefSeq transcript. For an alternative view, we will look at the same experiments in the dense RNAseq BAM format. Remove the arcs tracks by clicking the X in the track labels. Open the Faceted Track Selector. Un-highlight *RNAseq PE Junctions (arcs)* and highlight *RNAseq PE BAM dense*. Enter "lung" in the search box if it is not already entered (Figure 9A). After closing the Faceted Track Selector you will see the dense BAM tracks (Figure 9B). If you are zoomed-out too far, you will get a "Too much data to show" error. In that case, zoom in until the error goes away.

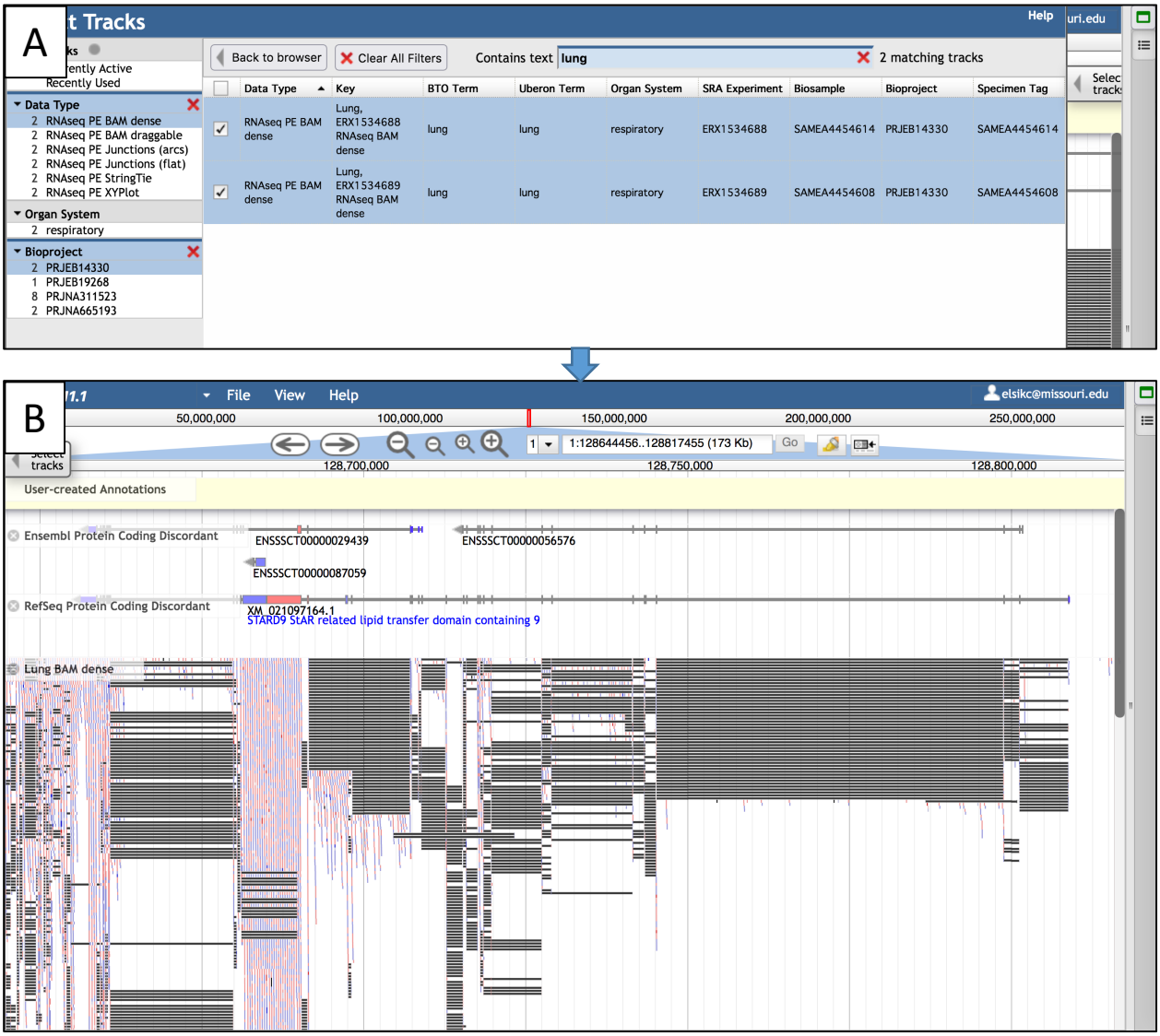

Figure 9. A) Selecting RNAseq BAM (dense) tracks for Bioproject PRJEB14330, filtered for by entering "lung" in the search box. B) JBrowse view showing the dense BAM tracks. The small red and blue rectangles are the aligned portion of the reads. Splice junctions are shown as dark gray lines.

It is easier to focus on splice junctions if you remove unspliced reads from dense BAM tracks. Do so by clicking the downward arrow that becomes visible when you mouse over a track label, and select Hide unspliced reads (Figure 10)

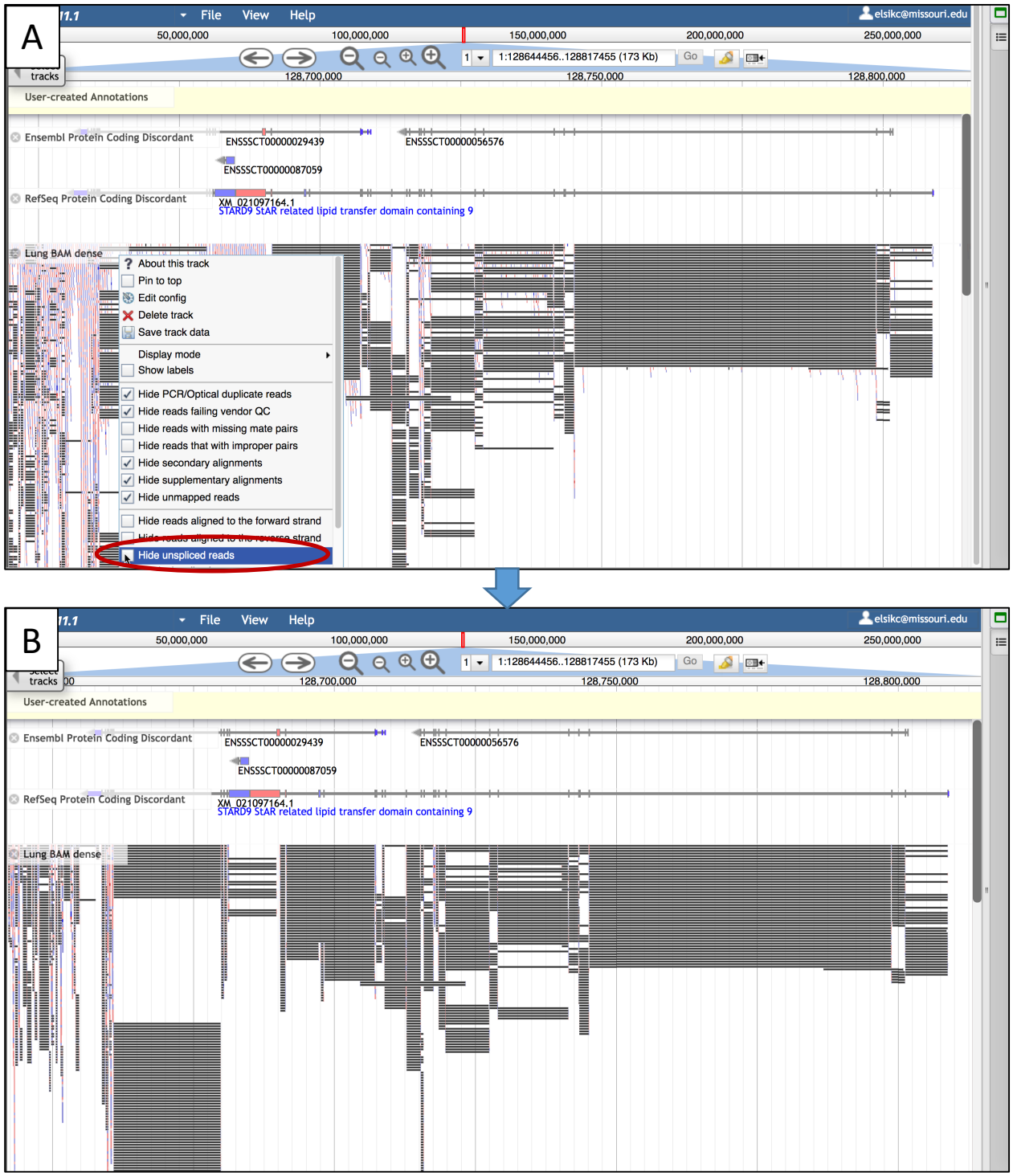

Figure 10. A) Hiding unspliced read alignments in the dense BAM track by clicking the track label and selecting Hide unspliced reads in the pulldown menu (circled in red). B) JBrowse view after hiding unspliced reads.

Zoom in further to investigate the exons and introns present in the RefSeq but not Ensembl (Figure 11). The dense BAM track provides strong support for the RefSeq transcript.

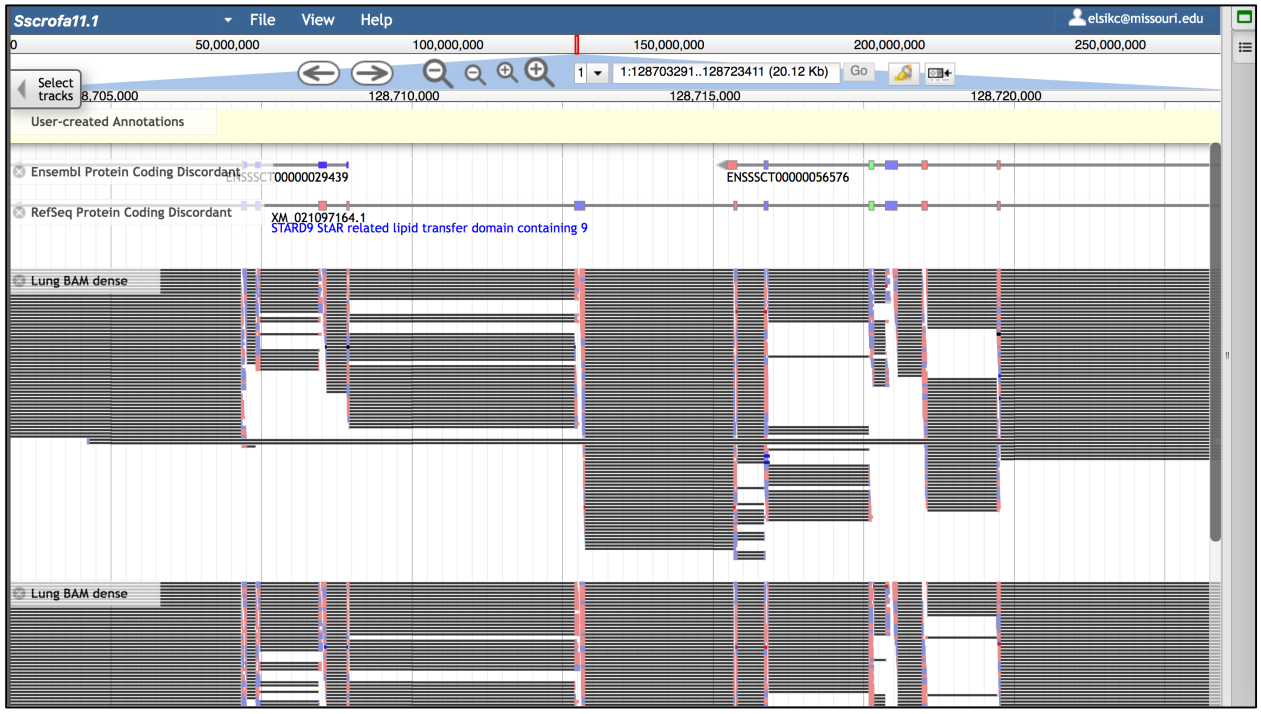

Figure 11. A zoomed-in view of dense RNAseq BAM tracks with unspliced reads hidden. These tracks and the previously viewed arc junctions tracks support the RefSeq gene model as being correct.

Based on the RNAseq evidence, we decide to use the RefSeq transcript to initialize the annotation. Click the RefSeq transcript and drag it up to the Editing Area (Figure 12). The transcript in the editing area is now called an *annotation*.

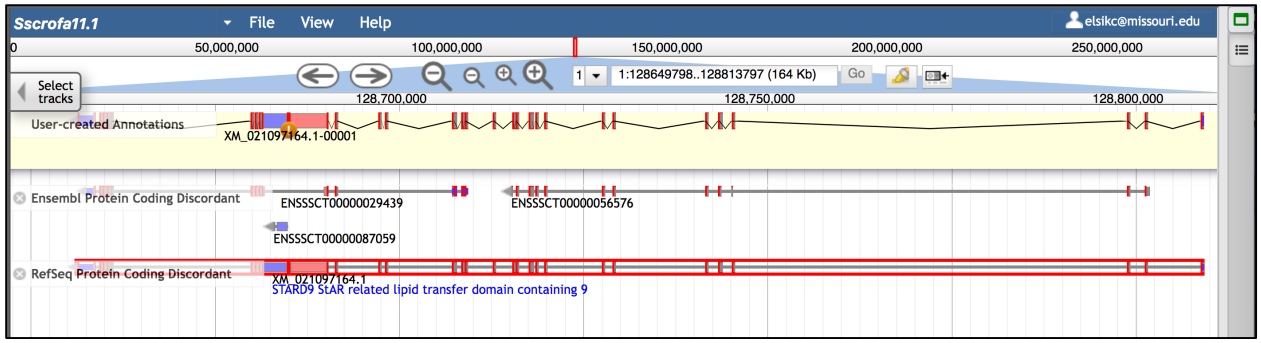

Figure 12. The RefSeq transcript is clicked, then dragged to the Editing Area. Clicking the transcript causes it to be surrounded by a red box and causes the exon edges to be highlighted in red. The edges of exons in other transcripts that match the RefSeq exon edges are also highlighted in red. Notice that the id of the annotation is the id of the original transcript with an extension added. The exclamation mark (seen near the id) indicates a non-canonical splice site.

Once the transcript has been dragged to the editing area, we wish to confirm that the junctions in the RNAseq data perfectly match the introns in the annotation. We will use flat RNAseq junctions to do so. Remove the dense BAM tracks from the browser by clicking the X in the track labels. Open the faceted track selector. Un-highlight *RNAseq PE BAM dense* and highlight *RNAseq PE Junctions (flat)*. Make sure that "lung" is entered in the search box. Select both tracks and close the Faceted Track Selector (Figure 13A) to return to the browser (Figure 13B). Zoom in to the first intron of interest (Figure 13C).

Identify a junction that appears to correspond with the intron in the annotation (Figure 13C). Right click the selected junction to view details. Note the *Score*, which is the number of reads supporting the junction. Also note the start and end coordinates of the junction (Figure 14A). Determine whether the exon boundaries in the annotation is perfectly abutted with junction by placing the cursor in the lower number line and moving it until the red line perfectly aligns with the exon edge, and note the number shown in the small bow next to the cursor (Figure 14B). We see that the exon edge on the left is one less than the junction start coordinate, so it the exon does perfectly abut with the junction. Repeating the same procedure for the exon edge on the right, we see that it also is also perfectly abutted with the junction (Figure 14C). This procedure is repeated for all introns that are present in RefSeq but not Ensembl, and exons that disagree between the two gene sets.

After checking splice junctions, we look for additional issues. Zooming in to the 3' end of the transcript (the far right) we can see that the start codon in the annotation is not in the same location as the original transcript (Figure 15A). Apollo computes start codons based on the largest open reading frame, and does not always reproduce the original coding sequence. Zoom into region as far as possible and click the annotation so you can see the bases in the annotation. The color of the exon in the annotation matches the color of the translated reference sequence above. This transcript is on the minus strand, so we look at the peach colored reading frame below the reference sequence DNA (Figures 15B and 15C).

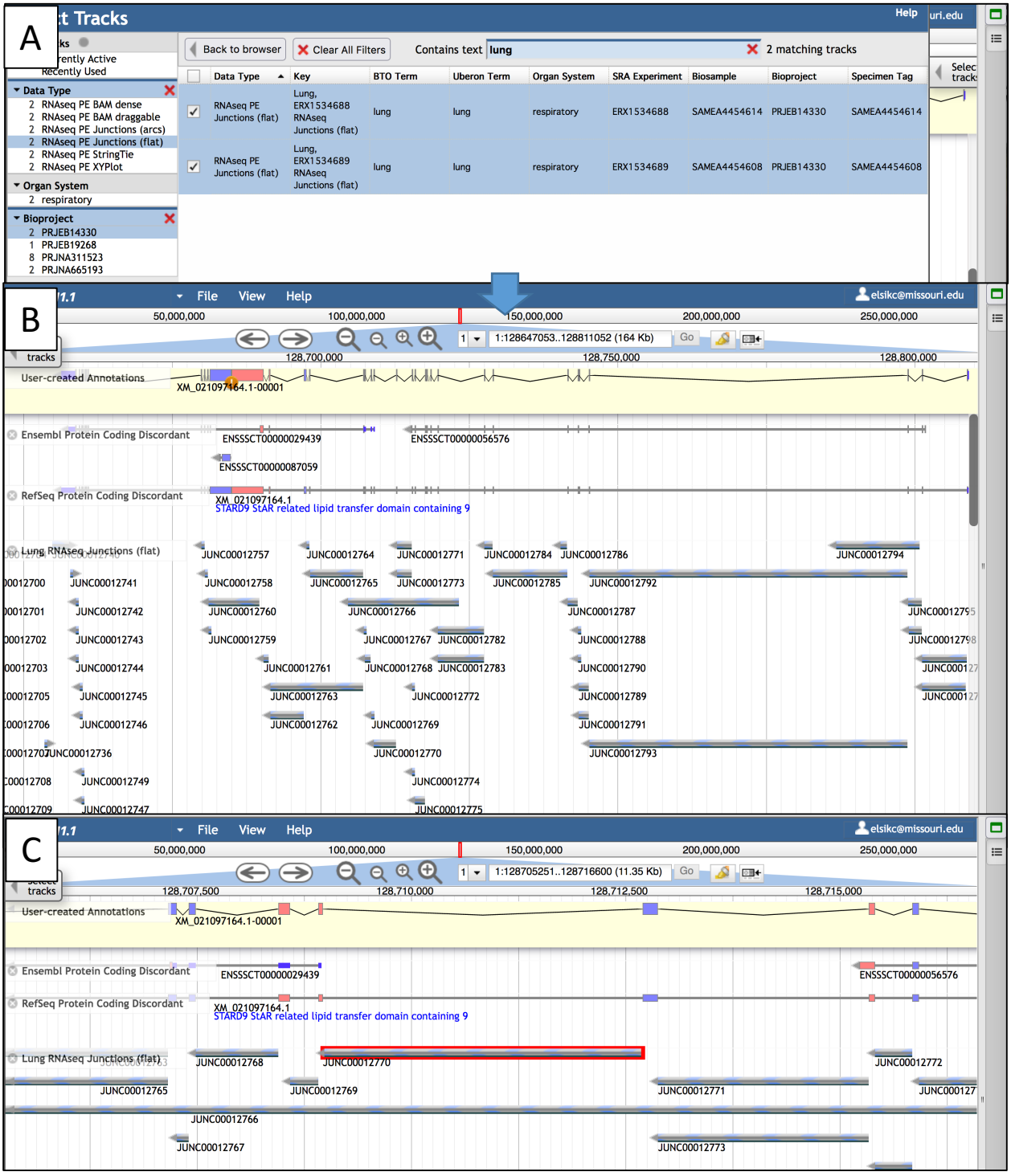

Figure 13. A) Selecting RNAseq Junctions (flat) tracks for Bioproject PRJEB14330, filtered for by entering “lung” in the search box. B) JBrowse view showing the flat junctions tracks. C) Zoomed in view showing a junction matching an intron that is present in the RefSeq, but not Ensembl, transcript.

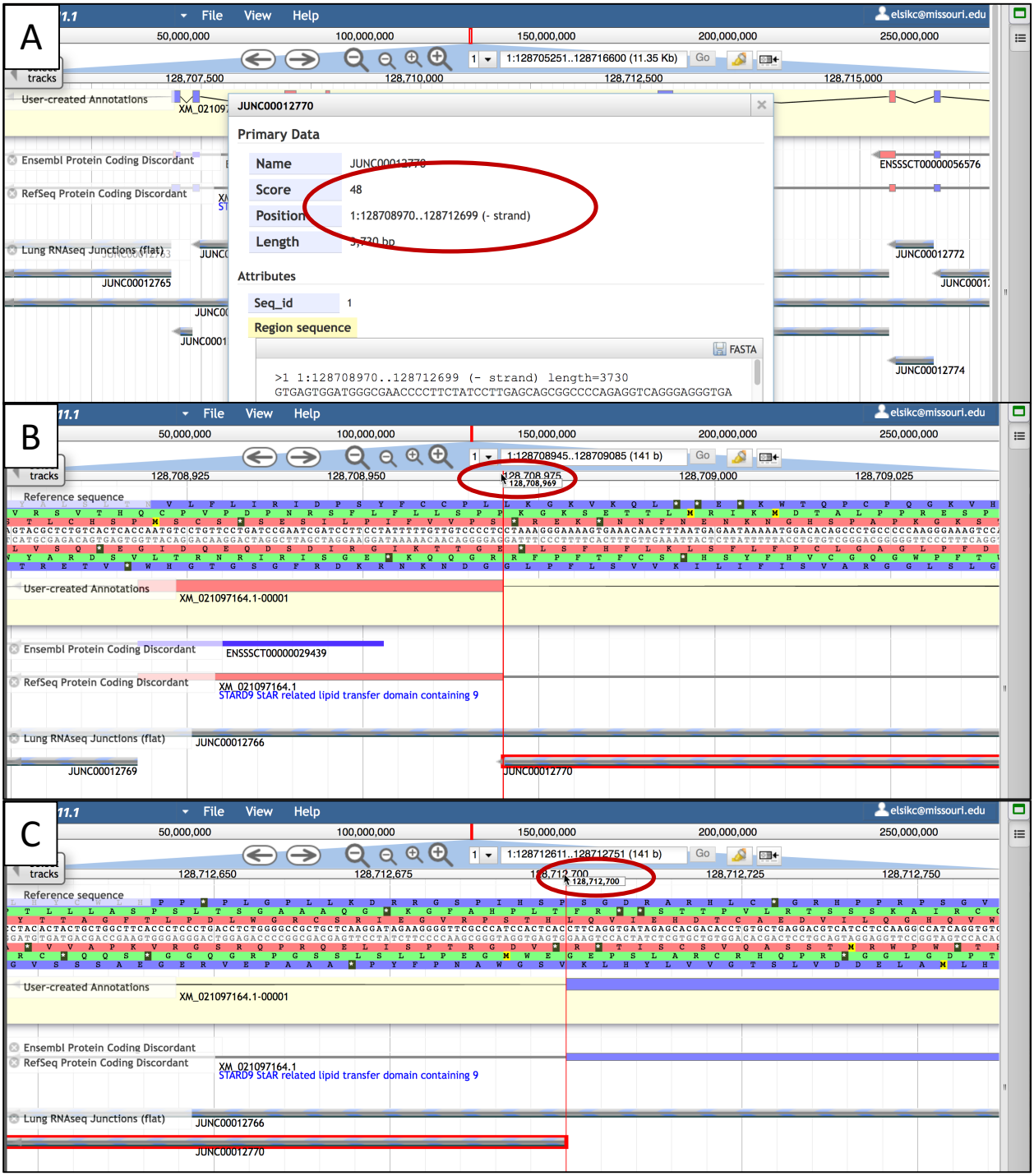

Figure 14. A) Right-clicking the junction shown in Figure 13 allows you to view details. The *Score* is the number of reads supporting that junction. We can see that the junction starts at coordinate 128,708,970 and ends at 128,712,699. B) A zoomed-in view showing the exon boundary at the left of the junction. Placing the cursor in the lower number line causes a red line to appear. The cursor is positioned so that the red line perfectly matches the exon edge. The small box next to the cursor shows that the last exon coordinate (128,708, 969) perfectly abuts the first junction coordinate. C) A zoomed-in view showing the exon boundary at the right of the junction showing that the exon first exon coordinate (128,712,700) perfectly abuts the last junction coordinate. Thus the annotation intron is validated by the RNAseq.

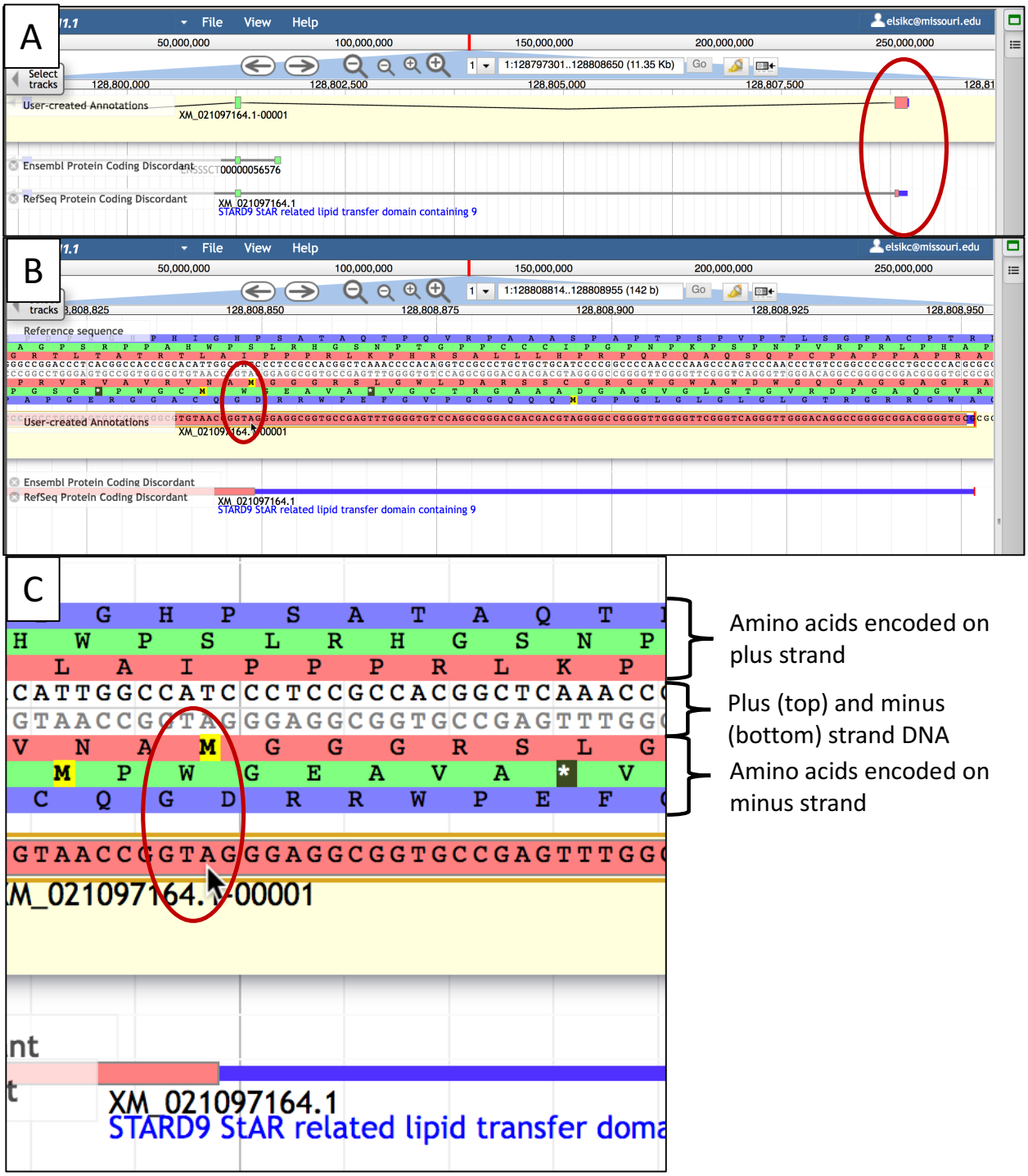

Figure 15. A) Apollo view with the terminal 3' exon in the annotation and original transcript circled in red. Notice that the coding region in the annotation is larger than the coding region in the original transcript. B) Apollo view zoomed-in enough to see the DNA. Clicking on the annotation shows the DNA in it. The Methionine (M) and the start codon (ATG shown in reverse) in the annotation that correspond with the start codon in the original transcript are circled in red. C) The figure is zoomed in to show the cursor pointing to the translation start that corresponds with the original transcript.

The translation start site in the annotation is reset by right-clicking the “A” of the “ATG” in the desired start codon, and then selecting Set Translation Start in the pulldown menu (Figure 16).

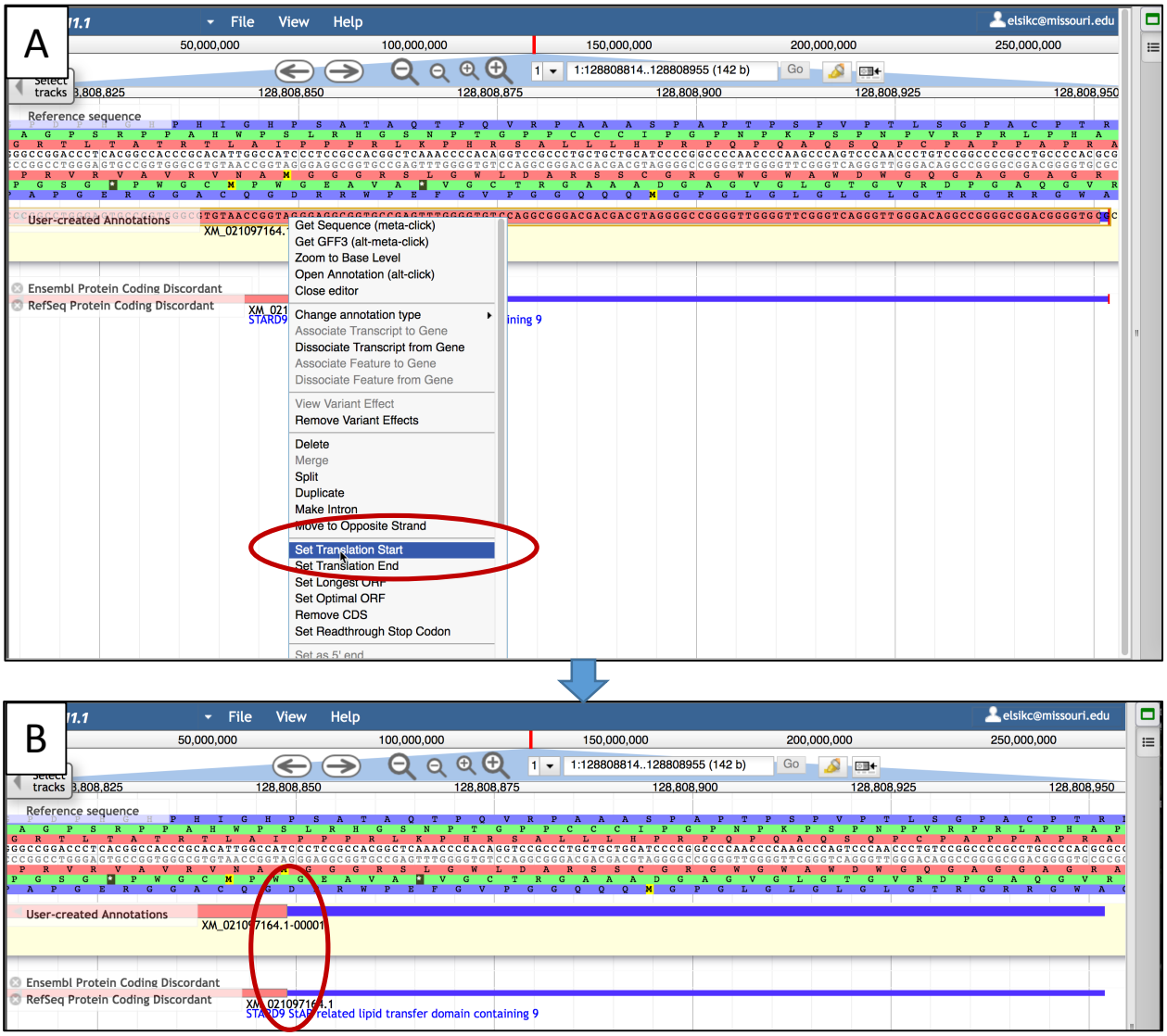

Figure 16. A) Setting the translation start by right-clicking the “A” in “ATG” (shown in reverse), then selecting *Set Translation Start* in the pulldown menu. B) A view showing the corrected translation start site.

The final step is to use the sequence of the annotation in a search of a database of well-known proteins, such as UniProtKB/Swissprot. Obtain the protein sequence of the annotation by right-clicking the annotation and selecting "Get Sequence" in the pulldown menu (Figure 17). Use the NCBI BLAST website (https://blast.ncbi.nlm.nih.gov/Blast.cgi) and select *Protein BLAST* to search a protein database.

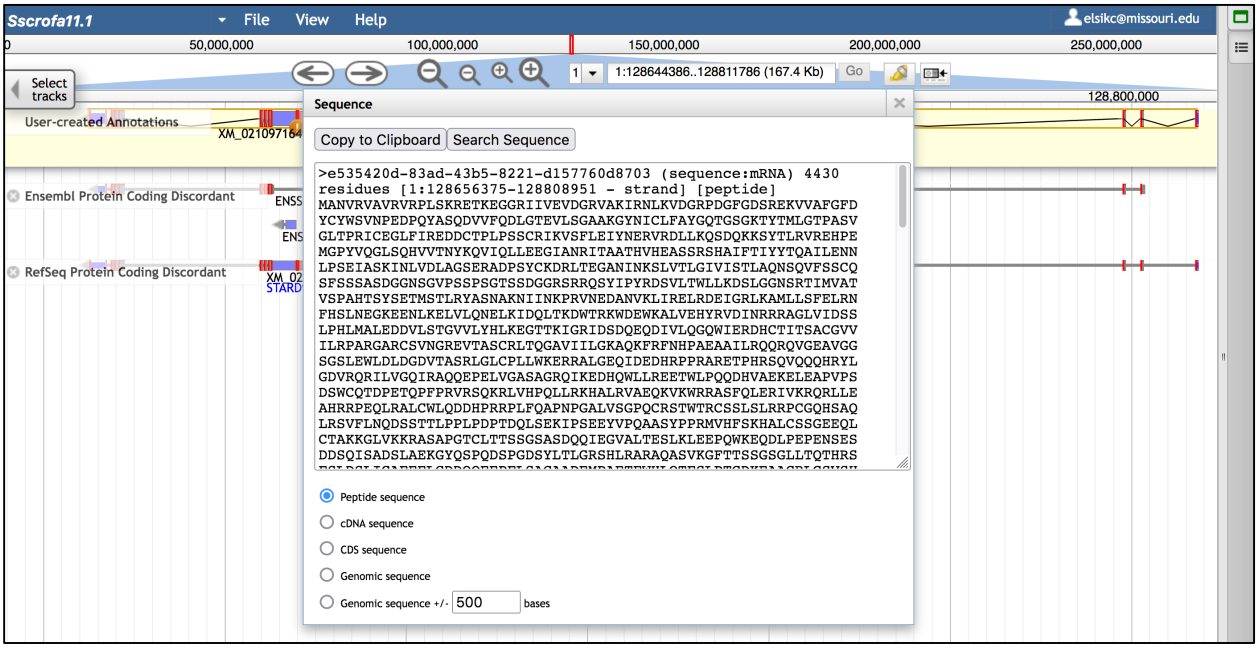

Figure 17. The Sequence panel after right clicking the sequence and selecting *Get Sequence* in the pulldown menu.