

Manual Curation Exercise I

For this example the user will search for the ribonucleotide reductase genes in the Tutorial genome using a gene sequence from the KEGG database, AO090023000916. Using the “Search” feature to find gene predictions that align to this KEGG protein returns a prediction that appears to be two genes merged into one. A view of the genome browser shows that another gene prediction method has accurately predicted the ribonucleotide reductase gene without merging it with another gene upstream but there is no gene predicted at the upstream locus despite EST evidence and homology support. Using the track editor tool create a model at the locus upstream from the ribonucleotide reductase gene and promote it to the gene catalog.

1. Search for the ribonucleotide reductase gene.
 - a. Click on the “Search” link at the top of the portal page.
 - b. Using the “Gene Model Hits” search function search for a hit description that contains “AO090023000916” where the database equals “KEGG”. The search results will appear at the bottom of the page.
 - c. Click on the Name link for gene #5, estExt_fgeneshl_pg.C_10019. This link points to the protein analysis page for this gene prediction.
2. Assess the protein alignments.
 - a. The default alignment view shows clearly that the 3’ half of the gene prediction codes for a single gene.
 - b. Change the alignment view to display all of the proteins that align to this gene prediction
 - i. Scroll to the top of the page and look at the “total hits(shown)” field. There are 177 proteins aligned to this gene prediction and by default only 10 are showing.
 - ii. Scroll to the bottom of the page. Enter 177 into the “Hits” field and click “Apply Filter”.
 - iii. Once the page refreshes, scroll through the protein alignments. The alignments towards the bottom of the viewer display confirm that this gene prediction is two separate genes merged into one gene. This gene model needs to be split.
3. Assess gene predictions at this locus generated by other tracks as potential replacements.
 - a. Assess the locus represented by the 3’ half of the merged gene.
 - i. Scroll to the top of the page and click on “To Genome Browser” link just above the protein alignment display. The genome browser viewer will open in a new window.
 - ii. The prediction being analyzed will be highlighted in yellow in the “Fgenes ab initio models, EST-extended” track.
 - iii. Check to see what gene predictions have been promoted to the Gene Catalog track at this locus. Expand the GeneCatalog track by clicking on the label that says “XXX transcripts in catalog...” A shorter gene prediction is already in the catalog at this locus.

- iv. Click on the model in the gene catalog to view the protein analysis page. The domain and protein alignments for this gene prediction are very good, including the alignment of the protein we began our search with, AO090023000916, the *Aspergillus oryzae* ribonucleotide reductase protein from the KEGG database.
 - v. Click the back button on the web browser to go back to the genome browser viewer.
 - vi. The gene represented by the 3' half of the merged prediction is already represented in the gene catalog by a gene prediction from another track.
 - b. Assess the locus represented by the 5' half of the merged gene.
 - i. There is no gene prediction in the gene catalog at this locus.
 - ii. Looking at the other gene prediction tracks the only gene prediction that covers that locus is the merged gene prediction.
 - iii. Expand the EST cluster track by clicking on the label "EST cluster consensi BLAT alignment".
 - iv. There is EST evidence at this locus that supports the existence of an expressed gene.
 - v. Scrolling further down on the viewer we see that there is also homology support displayed in the blue protein BLASTx alignment tracks.
 - vi. This is strong supporting evidence that a gene exists at this locus but since none was predicted you will have to use the track editor tool to create one.
4. Create a gene model at the upstream locus.
 - a. Login to the track editor tool.
 - i. Click on the "Open/Close Toolbar" link at the top of the page just under the "Feature" field.
 - ii. Login to the track editor tool using your personal UserID and password.
 - b. Create a new model from the EST cluster.
 - i. Click on the EST cluster, mouse over "Copy model to track" and click on "JAM_UserModels".
 - ii. The page will refresh and the copy of the EST cluster will appear in the "User Models" track. The name of the model will contain the first three letters of your UserID attached to the original name of the EST cluster. The color of the user model will be blue indicating that it has been copied but not yet edited.
 - iii. Click on the user model you've created and choose "Protein sequence" from the drop-down menu.
 - iv. Check for the 5' Met and the 3' stop codons. ESTs may contain UTR sequence and may extend beyond the start and stop codons. In this instance there is a stop codon in the protein sequence which needs to be set as the end of the coding sequence.
 - c. Set the termination of the coding sequence.
 - i. Close the protein sequence window.

- ii. Click on the user model, mouse over “Zoom browser into” and click on “exon 8”.
 - iii. Click on the exon and choose “Exon menu” from the drop-down menu. A pink vertical line lines up with the left edge of the amino acid in frame. In this instance, the pink vertical line lines up with the “Y” in the second frame.
 - iv. Click on “Model menu” in the drop-down menu.
 - v. Click on “CDS coordinates” in the drop-down menu. A purple vertical line shows you the boundaries of the coding sequence for your model. In this case, the CDS boundary is the same as the exon boundary.
 - vi. Type 111355 in the small box at the 3’ end of the exon and click “Save”. The page will refresh and the exon of your model will now have UTR which is represented by the narrow portion of the exon. The CDS is the thicker part of the exon.
 - vii. Click on the model, which is now red indicating that it has been edited, and choose “Protein sequence” from the drop-down menu.
 - viii. Confirm that the protein translation terminates at the stop codon.
- d. Set the start of the coding sequence.
- i. Before closing the protein sequence window scan the sequence for a start codon. In this case, the 16th amino acid residue is a Methionine. Note that the Methionine follows the LRRG amino acids. Keep the protein sequence window open for reference.
 - ii. Go back to the browser viewer window, click on the user model, mouse over “Zoom browser into” and choose “Exon 1”. Note that the LRRG amino acids are coded for at the 3’ end of the exon. The Methionine residue is not represented in the three-frame translation display because the codon is split by the intron. At this time, the track editor is not able to set the CDS at a codon that is interrupted by an intron so the CDS start will have to be set at the second codon. The amino acids following the 5’ Met are “STV”.
 - iii. Click on the user model, mouse over “Zoom browser into” and click on “Model”.
 - iv. Click on the exon and choose “CDS coordinates” from the drop-down menu. Reset the 5’ CDS boundary by clicking on the “-3” in the small box at the 5’ end of the exon. This changes the CDS boundary by 3 nucleotides at a time. Keep clicking the “-3” until the vertical purple bar is lined up with the second exon in the model. The coordinates in the small box at the 5’ end of the exon should read 109816. Click “Save” in the small box. Again, the page will refresh and the exon of your model will now have UTR which is represented by the narrow portion of the exon.
 - v. Click on the user model, mouse over “Zoom browser into” and click on “Exon 2”.
 - vi. Reset the 5’ CDS boundary by clicking on the “-1” in the small box at the 5’ end of the exon. This changes the CDS boundary by

3 nucleotides at a time. Click on the “-1” twice so that the vertical purple bar aligns with the left edge of the “S” residue in the third frame. The coordinates in the small box at the 5’ end of the exon should read 109818. Click “Save” in the small box.

- vii. Confirm that the protein translation begins with “STV”, does not have any internal stops and terminates at the stop codon.
 - viii. Copy the protein sequence from the protein sequene page and run a BLASTp alignment against the nr database to confirm homology support for this user created model.
- e. Release the user model.
- i. Click on the model and choose “Release user model” from the drop-down menu.
 - ii. A new page will display a “No errors” message which means that it has checked the coding sequence and determined that it is divisible by 3. Click on “Process”.
 - iii. The page will refresh and the model will now be green indicating that it has been released. If you wish to continue editing a released model, click on it and choose “Edit User Model” from the drop-down menu. The page will refresh, the model will again be red and will be ready for further editing.
5. Promote the new model to the gene catalog.
- a. Click on the released model and choose “Model web page” from the menu.
 - b. From the protein page, click on the “View/modify manual annotation” link in the middle of the page.
 - c. If a “Login Here” button appears at the top of the annotation page then you must login to annotate using your personal UserID and password.
 - d. Click on the “add” link next to “Disposition”. A new window will appear. To promote this model to the gene catalog the Disposition must equal “Catalog”. Click on “Save Item”. The annotation page will refresh and a new value will appear in the Disposition row with your UserID next to it indicating that you added this annotation. Your UserID is linked to the email address that you entered when you registered to annotate so that other annotators may contact you if they have questions about your curations.
 - e. Click on the “add” link next to “Defline”. A small window will appear where you can enter the Defline for the model. In this case enter “hypothetical protein” into the field and click on “Save Item”. Again, The annotation page will refresh and a new value will appear in the Defline row with your UserID next to it.
 - f. Click on the “add” link next to “EST evidence”, select the radio button under “Yes” and then click “Save Item”. This will indicate that there is EST support for this gene prediction which gives it a greater degree of confidence.
 - g. Click on the “add” link next to “Model Notes” and choose “editing needed – 5’ only” from the drop-down menu in the small window. You may also

wish to make a note that the protein translation starts with the second amino acid.