

## Manual Curation Exercise II

Protein analysis for the gene prediction with protein id 1231 shows that the gene prediction is incomplete at both the 5' and 3' ends. It is common for gene predictions based on homology, as this one is, to be incomplete at the 5' and 3' ends due to poor conservation at the terminal ends of the proteins. However, the protein alignments for this prediction are at rather low coverage, 47-66%, and the gene prediction is missing both the 5' Met and the 3' stop suggesting it is incomplete. EST evidence aligned to the genomic sequence provides supporting evidence to extend this model further at the 3' end. Using the track editor tool, create a new model that captures both the 5' Met and 3' stop and captures the Autophagy protein domain.

1. Gene prediction analysis
  - a. Find the protein page for protein id 1231
  - b. Click on "Search" at the top of the portal page.
  - c. Scroll down to the bottom of the page to the "Gene Models" search function and choose "Protein id" from the drop-down menu. Type 1231 into the search field and click "Search Models"
  - d. "Model Search Results" appear at the bottom of the page. Click on the "P" link in the "Model Id" field to go to the protein analysis page for this gene prediction.
  - e. Is this model complete? Does it capture the 5' Met and 3' stop?
    - i. Click on the green horizontal bar which represents the protein translation of this gene prediction.
    - ii. A new page will appear that has the amino acid sequence of this gene prediction in fasta format. This prediction is incomplete.
    - iii. Click the back button on your web browser to go back to the protein page.
  - f. How good are the protein alignments?
    - i. Take a look at the first protein in the alignment view, the *Ustilago maydis* protein. The alignment is represented by a grey bar.
    - ii. The alignment starts at position 201 of the *U.maydis* protein and ends at position 655. The total length of the *U.maydis* protein is 788 amino acids. 58% of the *U.maydis* protein is covered by the gene prediction at 44% identity.
    - iii. This protein alignment and the others below also suggest that this gene prediction is incomplete.
  - g. Take a look at the gene prediction and supporting evidence in the genome browser viewer. Click on the "To Genome Browser" link just above the protein alignments. The browser viewer will open in a new window.
  - h. Zoom the browser view out 3x to see more of the gene region by clicking on the "-3x" button at the top of the browser viewer page.
  - i. Notice that there is EST evidence at the 3' end that could support extension of the prediction at the 3' end. Expand the EST cluster track by clicking on the label "EST cluster consensi BLAT alignment"

2. Gene prediction editing
  - a. Login to the track editor tool.
    - i. Click on the “Open/Close Toolbar” near the top of the page just under the “Feature:” field.
    - ii. The toolbar will open on the left side of the browser.
  - b. Click the “Enable Edit Mode” button in the “Track Editor” section of the toolbar. Login using your personal UserID and password.
  - c. Close toolbar
  - d. This gene prediction is on the bottom strand which is more difficult to edit. For easier editing, flip the strand by clicking on the “(+) Ref. Strand” button found in the navigation controls at the top.
  - e. Copy the gene prediction to the “User Models” track for editing.
    - i. Click on the gene prediction, mouse over “Copy model to track” and click on “JAM\_UserModels”.
    - ii. The browser will refresh and your copy will appear in the “User Models” track. Expand the “User Models” track by clicking on the label that says “User Models”.
    - iii. Small black arrows on either side of the label confirm that track is expanded.
  - f. The new model is named by using the first three letters of your UserID appended to the original model name. You should see this name at the far left side of the browser.
  - g. Add the EST exons to the gene model. The exons will have to be added one at a time
    - i. Click on the first exon.
    - ii. Mouse over “Add exon to model”.
    - iii. Click on the name of the model you copied to the track.
    - iv. The page will refresh and your model will now have the extra exon at the 3’ end. The model will also turn red indicating that the structure has been edited from the original.
    - v. Repeat two more times for the other two EST exons.
  - h. Check the protein sequence of your newly created model
    - i. Click on your edited model in the “User Models” track.
    - ii. Click on “Protein sequence” in the drop down menu.
    - iii. A new window opens with the protein sequence in fasta format.
    - iv. The protein sequence shows that the EST alignments extend the 3’ exon past the stop codon into the UTR.
    - v. Close the protein sequence window.
  - i. Modify the CDS on the 3’ end.
    - i. Click on your edited model. Mouse over “Zoom browser into” and click on “Exon 9” the most 3’ exon.
    - ii. Determine which frame this exon is in. Click on the “Base Position track at the top of the browser at position 58350 to center the browser at the 5’ end of the exon and to zoom in 3x.

- iii. Click on the 3' exon of your model 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 "A" in the first frame.
- iv. Click on "Model menu" from the drop down menu and then choose "Zoom browser into" and "Exon 9".
  - v. Zoom browser into the 3' end of the exon by clicking in the "Base Position" track at the 57750 position.
  - vi. Click on your model and choose "CDS coordinates" from 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.
  - vii. A small box also appears displaying the base position of the CDS boundary. Type 57781 into the small box and click on "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.
- j. Find the 5' Methionine.
  - i. Click on your model, mouse over "Zoom browser into" and click on "Model".
  - ii. Scroll down the browser viewer to see if any other models have a 5' exon that captures the 5' Met.
  - iii. Expand the Genewise1Plus track by clicking on the "Genewise1Plus" label.
  - iv. Click on the Genewise1Plus model, e\_gw1.1.215.1, and choose "Model web page" from the drop down menu.
  - v. Click on the green horizontal bar to view the amino acid translation and confirm that this model captures the 5' Met.
  - vi. Click the back button on your browser twice to go back to the browser viewer page.
- k. Confirm that the Genewise1Plus 5' exon is in the same frame as your user model.
  - i. Click on the Genewise1Plus model, mouse over "Copy model to track" and click on "JAM\_UserModels".
  - ii. Click on the 5' exon of your newly created model, which should be blue indicating that it was copied but not edited, and choose "Exon menu". Small boxes will appear displaying the coordinates of the boundaries of the 5' exon, 60592 and 60306.
  - iii. Close the drop down menu by clicking on the "X" in the upper right hand corner of the menu box.
  - iv. Click on the 5' exon of your already edited model and choose "Exon menu". The exon boundary coordinates are again displayed in the small boxes, 60418 and 60306. The 3' boundary of the exon has the same coordinates, 60306, as the Genewise1Plus model that you copied to the "User Models" track for comparison, confirming that the 5' exons of both models are in the same frame.

- v. Modify the 5' boundary of the exon by typing 60592 into the left-hand box and click "Save".
- l. Confirm that the edited model is complete
  - i. Click on the edited model and choose "Protein sequence".
  - ii. A new window opens up displaying the protein sequence in fasta format.
  - iii. The protein translation contains the 5' Met, the 3' stop and is free of internal stops.
  - iv. Close the protein sequence window.
- m. Release the edited model.
  - i. Click on your edited model and choose "Release user model" from the drop down menu.
  - ii. If the model has a CDS that is divisible by 3 the message will display "No errors" and you can click on "Process" to continue releasing the model.
  - iii. The model is now green indicating that it has been released.
- n. Click on the model and choose "Model web page" This will take you to the protein page for this model.
- o. To add your edited model to the Gene Catalog click on "View/modify manual annotation", login to annotate and set the "Disposition" field to "Catalog".