BIOL199 BDB
Fall 2022

Pathways Project Annotation Walkthrough Part G

*Adapted by Melinda A. Yang, from "*Pathways Project: Annotation Walkthrough" by *GEP members Katie Sandlin, Wilson Leung, and Laura Reed, and from "Pathways Project Annotation Notebook" by GEP members Katie Sandlin and Alexa Sawa*

Table of Contents

(Learning) Objectives:

- Understand how to use Gene Model Checker to verify results
- Analyze a protein alignment with a dot plot
- Acquire and understand final datafiles and figures from the Pathways Project

Part G: Verify and submit gene model(s)

Our analysis of the CDS-by-CDS *tblastn* alignments and the evidence tracks on the Genome Browser allow us to precisely define the start and end positions of each of the five coding exons (CDS's) of Rheb-PA. To verify that our proposed gene model satisfies the basic biological constraints (e.g., begins with a start codon, has compatible splice sites, and ends with a stop codon), we will check our gene model coordinates using the Gene Model Checker.

Instructions

1. Open a new web browser tab and navigate to the Gene Model Checker (http://tiny.cc/genemodelchecker) and fill in the following information:
    - Species Name: Your target species
    - Genome Assembly: Your target species genome assembly from Part B
    - Scaffold Name: Your scaffold accession ID from Part B
    - Ortholog in *D. melanogaster*: The protein isoform (e.g. Rheb-PA) for your current manual annotation
    - Errors in Consensus Sequence: 'No'
    - Coding Exon Coordinates: Enter a comma-delimited list of coordinates for every CDS. Do not put commas between your numbers. **(I suggest copying this entry into your lab notebook, in case you need to rerun the Gene Model Checker again)**
    - Annotated Untranslated Regions?: 'No'
    - Orientation of Gene Relative to Query Sequence: Indicate the strand you found for your target gene on your target species
    - Completeness of Gene Model Translation: "Complete"
    - Stop Codon Coordinates: Click within the textbox and the coordinates will automatically populate - check this matches your stop codon coordinates

Comments

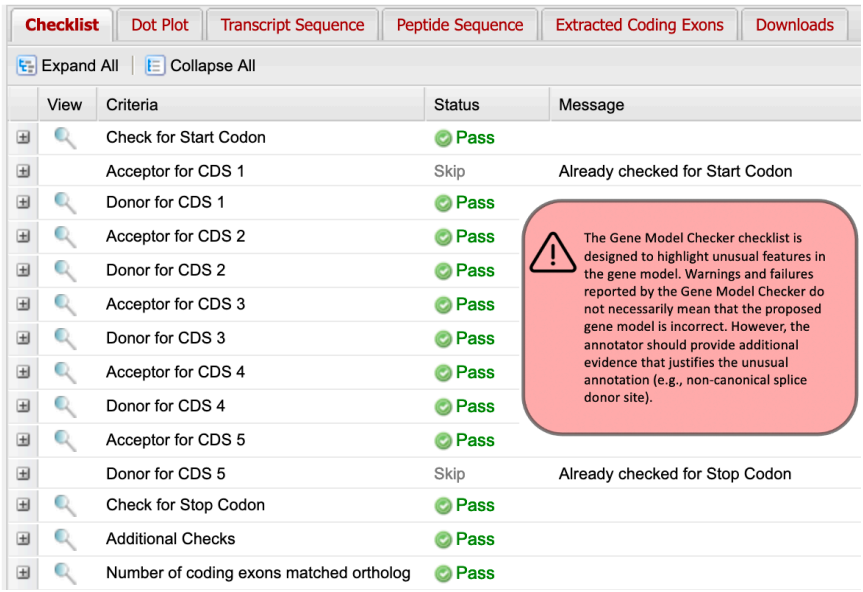
You may find it useful to pull up worksheet, questions B2 and B6 from Labs 7-8.

For the Coding Exon Coordinates for *D. yakuba*'s Rheb-PA, I put "19150809-19150857, 19150985-19151056, 19151154-19151361, 19151421-19151550, 19151613-19151699"

Note that stop codons should NOT be included in the coding exon coordinates section.

- Click ‘Verify Gene Model’ to run the Gene Model Checker

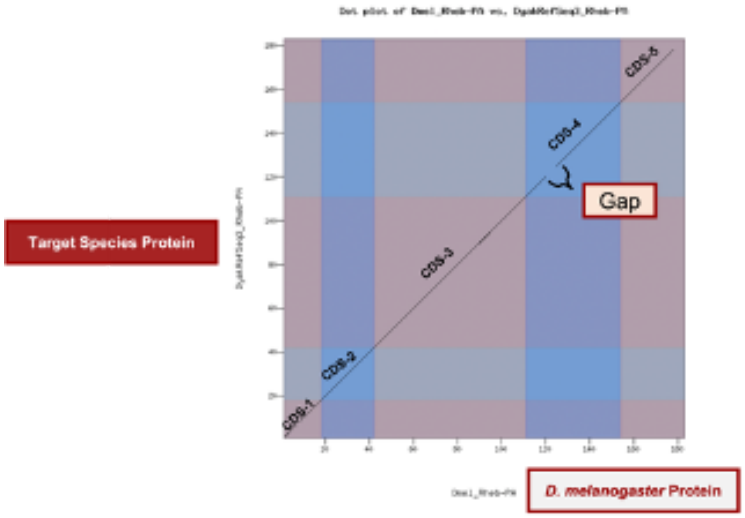
2. Examine the results in the ‘Checklist’ tab to check your gene model. Verify that your proposed gene model follows the list of expected criteria for a gene. For areas there the criteria were not met, check if you already verified that part of the model in your above splice junction verifications.



The Gene Record Finder uses an algorithm and a standard set of criteria typical to genes. There may be exceptions you have found that do not follow these standard criteria that fail the status check but are still correct for the true gene model.

Answer question G1a

3. Click on the “**Dot Plot**” tab to examine the dot plot between the *D. melanogaster* protein (x-axis) and the protein sequence for the submitted model in your target species (y-axis).

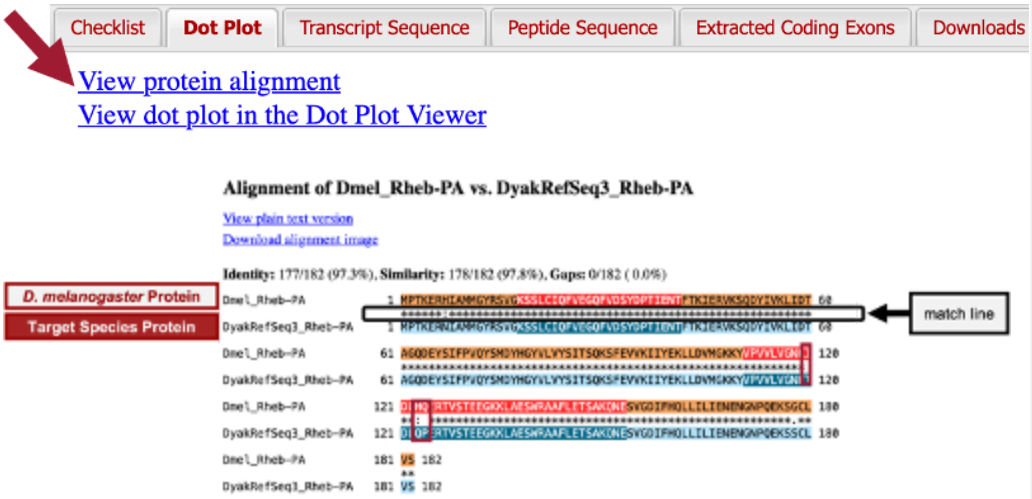


The alternating color boxes in the dot plot correspond to the different CDS’s in the two sequences. Dots in the dot plot correspond to regions of similarity between the *D. melanogaster* protein and the submitted *D. yakuba* gene model for Rheb-PA. If the submitted sequence is identical to the *D. melanogaster* ortholog, then the dot plot will show a straight diagonal line with a slope of 1. Changes in the size of the submitted model compared to the *D. melanogaster* ortholog will alter the slope of this line.

In this figure, the dot plot shows that the five CDS's of Rheb-PA in *D. melanogaster* and *D. yakuba* have similar lengths (compare the length shown on the x-axis to the length shown on the y-axis for each CDS). However, within a small region of CDS-4, the dot plot did not detect sequence similarity between the submitted model for *D. yakuba* and the *D. melanogaster* ortholog.

Answer question G1b

4. Click on the “View protein alignment” link above the dot plot

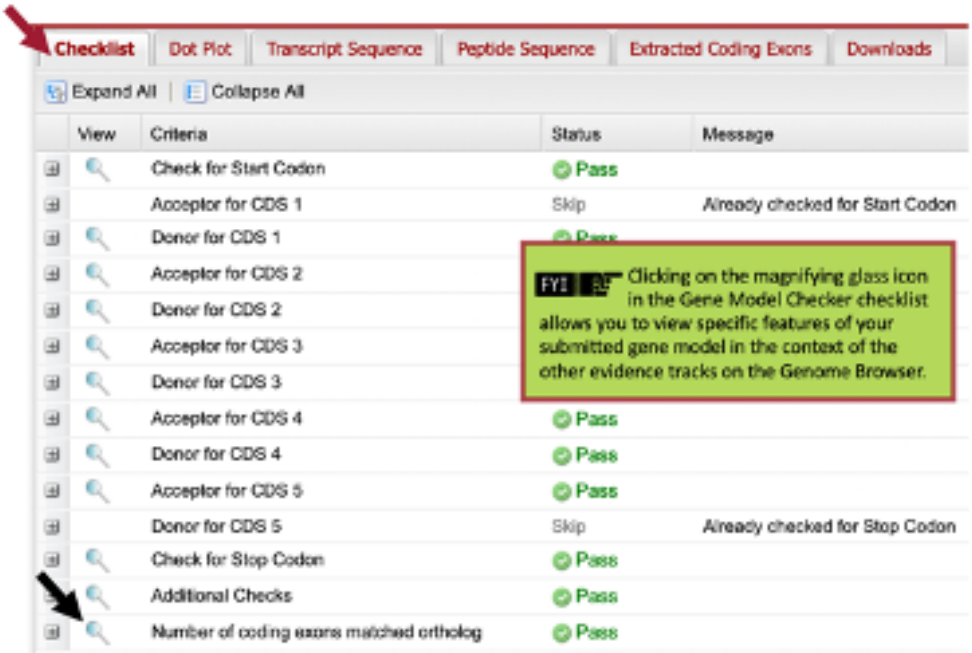


The alignment shows the comparison of the *D. melanogaster* protein against the conceptual translation for the submitted *D. yakuba* gene model. Like the dot plot, alternating colors correspond to the different CDS’s. The symbols in the match line denote the level of similarity (“*” indicates conserved amino acids, “:” denotes amino acids with highly similar chemical properties).

The protein alignment shows that the five CDS's have high levels of sequence similarity between the *D. melanogaster* ortholog and the *D. yakuba* gene model. The tiny gap in the dot plot within CDS-1 can be attributed to similar, but not identical, amino acids near its center.

Answer question G1c

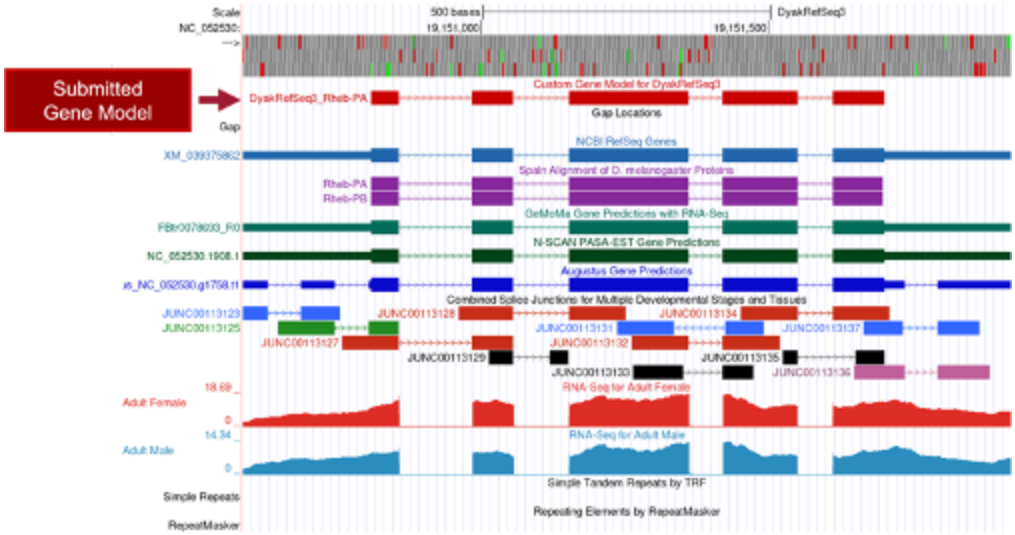
5. We now want to check any discrepancies. For regions with high amino acid dissimilarity, we can examine the Genome Browser. Back in the Checklist tab, we can click on a magnifying glass next to the region we want to examine to open a ‘Custom Genome Browser’ in a new window that includes your gene model.



The protein alignment between the *D. melanogaster* Rheb-PA and the submitted gene model in *D. yakuba* shows that there is a gap within CDS-4 in the dot plot that can be attributed to differences in three amino acid residues. We need to verify the placement of CDS-4.

Wait to do G1d until AFTER you have completed steps 6 and 7 (next two rows)

6. To view the entire gene model, zoom out in the Custom Genome Browser until you can view the entire gene model in red. Use the other evidence tracks in the 'Custom Genome Browser' to help you verify your full gene model.



Looking into the placement of CDS-4 in the Custom Gene Model for Rheb-PA, we see that it is in congruence with the *NCBI RefSeq Genes* and *Spaln* alignments, the *GeMoMa*, *N-SCAN*, and *Augustus* gene predictions, and the RNA-Seq data (exon junctions and histograms).

7. Save an image of the Custom Genome Browser with specific settings. Select the "default tracks" for the region, set the following evidence tracks (if available) to "pack", and then click on "refresh":

1. at least one transcript prediction track (e.g., TransDecoder Transcripts, modENCODE Cufflinks Transcripts)
2. at least one splice-site prediction track (e.g., Combined Splice Junctions, modENCODE TopHat Junctions)
3. a comparative genomics track (e.g., *Drosophila* Chain/Net, or *D. melanogaster* Chain and *D. melanogaster* Net, where any without 'pack' should be set to 'full')

Answer question G1d

8. There are three data files that must be downloaded that contain the records of your submitted gene model – a General Feature Format File (GFF), a Transcript Sequence File (fasta), and a Peptide Sequence File (pep). The Gene Model Checker automatically creates these three files for a specific isoform (e.g., Rheb-PA) when you verify a gene model. To download them for one isoform, click on the "Downloads" tab and then click on each of the links to save each file to your computer.

Note: You don't have to rename these files, you'll only need to rename the merged file.

9. You have now completed everything for one isoform, but you likely have more to download. For each UNIQUE isoform, including ones with identical regions, redo step G1 with the following edits:

- Ortholog in *D. melanogaster*: Your next unique isoform
- Coding Exon Coordinates: Enter a comma-delimited list of coordinates for every CDS for this unique isoform. Do not put commas between your numbers. **(I suggest copying this entry into your lab notebook, in case you need to rerun the Gene Model Checker again)**
- Re-populate the stop codons section
- All other sections should stay the same - if you think anything changes, please let your instructor know.
- For isoforms with unique coding regions, follow steps #5-7 to verify your gene model.
- Make sure in the end, you have GFF/PEP/FASTA files for every UNIQUE isoform (for most of you, this should be TWO unique isoforms).

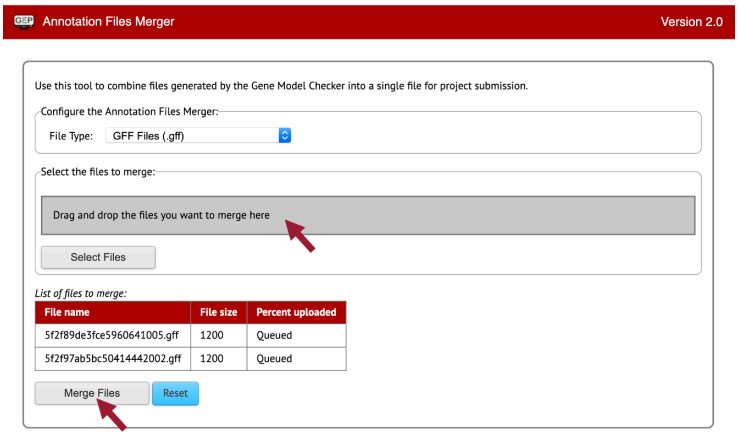
You do NOT need to repeat this for isoforms with identical coding sequences.

Repeat G1a-G1d only for unique isoforms

10. If you have **a single unique isoform**, rename your GFF/PEP/FASTA files so that they have the following naming structure: the shorthand for your target species, an underscore, followed by the gene symbol, followed by the subscript for this filetype (e.g. dyak_Rheb.gff). Skip to step G14.

11. If you have **two or more unique isoforms**, complete steps G11-G13. Open a new web browser tab and navigate to the Annotation Files Merger (http://tiny.cc/annofilemerge) to combine the GFF, transcript sequence, and peptide sequence files for all our isoforms into a single file (you will submit one merged file for each file type). Below are the steps for merging a GFF file.

- Change the “File Type:” to “GFF Files (.gff).”
- Drag all GFF files we downloaded for all isoforms to the “Drag and drop the files you want to merge here” section.
- Click on the “Merge Files” button



12. Download the merged GFF file:

- Right-clicking (control click on macOS) on the “Merged File Link”

- Click ‘Save Link As…”
- Enter a filename - the shorthand for your target species, an underscore, followed by the gene symbol, followed by the subscript for this filetype (e.g. dyak_Rheb.gff)
- Once you click on the “Save” button, the merged GFF file should download onto your computer into your Downloads folder.

13. Repeat #11-12 for the Transcript Sequence Files (.fasta) and the Peptide Sequence Files (.pep), making sure to change the ‘File Type:’ section and adjust filenames appropriately.

14. Open Terminal or MobaXTerm and use the following steps to check your newly generated GFF/PEP/FASTA files.

i. Open a Terminal or MobaX window but do NOT sign into Spydur.

ii. Navigate into your Downloads/ folder where your GFF/PEP/FASTA files are

- Macs: Type cd ~/Downloads/ to change into your Downloads directory.
- Windows: Type cd /cygdrive/c/Users/<username>/Downloads/ to change into your Downloads directory.

iii. Use less <filename> to look inside each file. (and ‘q’ to quit)

iv. For each file, make sure they include data for every unique isoform found in your target species for your target gene.

v. If you have an entry for each unique isoform, you should be good to go.

Answer question G2

15. Examine your dot plot and write legend(s) for your figures for one unique isoform. Note that the dot plot is a visual/graphical representation of the protein alignment showing the position of the amino acids for the *D. melanogaster* protein on the x-axis and the position of the amino acids for your target species on the y-axis; therefore, large gaps, regions with no sequence similarity, and any other anomalies seen in the dot plot can be located within the protein alignment.

Answer questions G3-G4

16. This was the last step for the annotation of your target gene in your target species. Take the time now to finalize your main figures and data files. The Google Drive folder (linked on BB) containing all lab projects is now editable - add in all required files into your group’s folder.

Answer question G5

The final step is to discuss your findings with the other group studying the same target gene. If you do not have a second group, make sure to get the results for a second target species from me for your team to discuss. Discuss the evolutionary relationships between your two target species to *D. melanogaster*. Share your results with each other, making sure to point out anything unusual to get the other group up to date. Compare your dot plots and protein alignments for these two target species' orthologs for the target gene. Discuss and see if mutational patterns inform on the evolutionary patterns that may be at play for your target gene in the context of the three *Drosophila* species examined. **Use this discussion to answer questions G6-G8.**

You should now have all parts of your Pathways Worksheet completed. If you plan to complete a Pathways Annotation Notebook for **Paper6**, your worksheets, screenshots, and images/files should have all the appropriate information to complete the form. In particular, several images from question G5 will need to be pasted into that form.

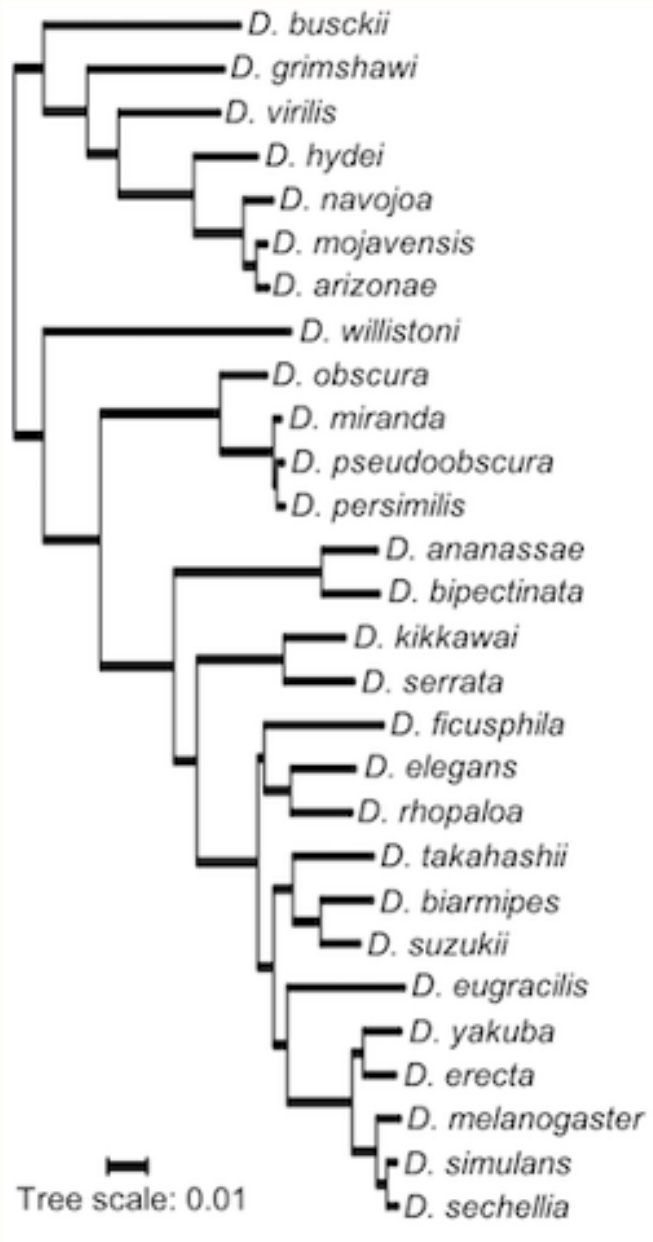


Figure 1. Phylogeny showing the evolutionary relationships between 28 *Drosophila* species. Use this to help you answer question G6.