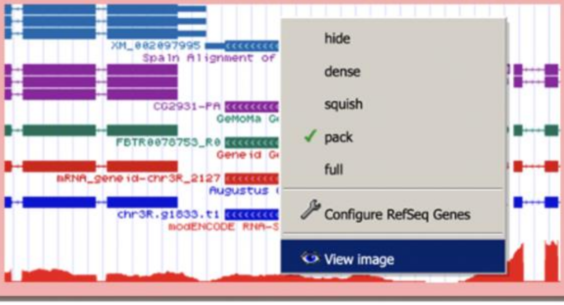
Pathways Project: Annotation Notebook

Katie Sandlin and Alexa Sawa

Directions: Use this document to keep track of your data as you locate and annotate your assigned target gene in your target species using the Pathways Project: Annotation Walkthrough as a guide. All embedded links in this document can be found on the Pathways Project page of the GEP website (thegep.org/pathways).

Avoid taking a screenshot of the GEP UCSC Genome Browser. Instead, you should export the image by right-clicking within the Genome Browser image and selecting "View image". This should automatically take you to a new tab (if it doesn't, make sure that pop-ups are enabled). Right-click on the image in the new browser tab, and then click on "Save Image As..."



Project Details

Target gene:

Target species:

Briefly describe the function of your assigned gene in the insulin signaling pathway. Note: You can get this from the "Gene Summary" on FlyBase (enter gene symbol in the "Jump to Gene" text box in the top right-hand corner of the FlyBase home page) or perform a literature search (e.g., Google Scholar or PubMed).

Part 1: Examine genomic neighborhood surrounding target gene in *D. melanogaster*

1. Full name (not italicized) and symbol (italicized) of your target gene in *D. melanogaster*:

2. Paste below an image of the genomic neighborhood of your target gene in *D. melanogaster* from the GEP UCSC Genome Browser including both nearest two upstream and two downstream genes and nested/nesting gene(s) (if present). Select "default tracks" for the region and, while we didn't discuss this in the walkthrough, set a comparative genomics track (e.g., Drosophila Conservation (28 Species)) to "pack" and then click on "refresh" prior to saving your image so you can submit this in your Annotation Report Form later.

3. Sketch the genomic neighborhood of your target gene in *D. melanogaster*. Be sure your sketch includes the names and/or gene symbols of the surrounding genes and indicates their orientation (Walkthrough Figure 6). Note: You can do this by hand and upload a picture of your drawing or you can create one in Word. (Click on "Insert" in the toolbar ribbon and then click on "Shapes." Under "Block Arrows," click on any style you like. Click on the location in the document where you want the shape to be, then hold and drag your pointer to a different location (determines how large the shape will be) and release the mouse button when it's the desired size.)

Part 2: Identify genomic location of ortholog in target species

4. Target species in which you intend to annotate your target gene (don't forget to capitalize and italicize like a scientist):

5. Does your target gene have multiple isoforms in *D. melanogaster*? If so, how many and which is the longest isoform? Note: Be sure to use the longest isoform for your *tblastn* against the entire genome of your target species.

6. Explain what query sequence you are using for this *tblastn* and indicate how many amino acids long it is.

7. Explain what database/subject sequence you are using for this *tblastn*.

8. Paste below a screenshot of the "Descriptions" panel of your *tblastn* results of the amino acid sequence of the *D. melanogaster* protein coding isoform for your target gene against the genome assembly in your target species (Walkthrough Figure 11).

9. Summarize your *tblastn* search results. What scaffold (and accession number) do you think your target gene is on in the target species? Is there only one very good match, or is there some ambiguity about where your assigned gene is located in the target species? (Note: This question is referring to the "Descriptions" panel in Question 8.)

10. After sorting the *tblastn* search results for your best match by "Subject start position", fill in Table 1 below to help you identify the best collinear set of alignments to the protein of your target gene in the target species' genome (Walkthrough Figure 15).

TABLE 1: Summary of the *tblastn* search results for the best scaffold match

Range	*D. melanogaster*		Target Species		E-Value	Identities (%)	Subject Frame
	Query Start	Query End	Subject Start	Subject End			

Note: Depending on your results, you may have to insert extra rows into this table (see directions at Microsoft Support). **You may also need to duplicate this table if you have more than one very good match in your *tblastn* search.**

11. What coordinates give you the best collinear set of alignments to the protein of your target gene in the target species' genome? Do these coordinates cover the entire length of the query (if not, explain what isn't covered and what might explain the lack of coverage)?

12. Summarize your findings from Part 2 by recording the following information for your target gene in the target species:
Scaffold:
Scaffold accession number:
Approximate coordinates:

Part 3: Examine genomic neighborhood of putative ortholog in target species

13. Since the "BLAT Alignments of NCBI RefSeq Genes" track indicates sequences which are present in the mRNA transcripts isolated from the target species, determine if there is evidence that the region of DNA you identified in Part 2 is being expressed in the target species. If there is only one possible transcript, record the RefSeq accession number (starts with “XM_”).

> If more than one transcript is present in this region, which one(s) is likely your putative ortholog? Are there any transcripts in this region that aren’t likely your putative ortholog (if so, explain why you don’t think they are your putative ortholog, e.g., is it on the opposite DNA strand or is it a nested/nesting gene)?

14. Looking at the BLAT Alignments of NCBI RefSeq Genes track, do any of the closest two upstream or closest two downstream genes have multiple isoforms? (If so, be sure to choose the longest coding isoform that is best supported by the other lines of evidence for your *blastp*.) Why do you think you need to use the longest coding isoform?

15. If a neighboring gene had multiple isoforms but all the isoforms had identical coding sequences, would it matter which one you chose for the *blastp* search? Why or why not? Hint: Review the BLAST programs in the blue box in Part 2.2 of the Walkthrough.

16. Explain what query sequences you are using for each of these *blastp* searches.

17. Explain what database/subject sequence you are using for these *blastp* searches.

18. Fill in Table 2 below to summarize your results for Parts 1-3:

TABLE 2: *blastp* search results for the protein sequences of the genomic neighborhood of the target gene in the target species against the *D. melanogaster* reference protein database (refseq_protein)

		2nd Closest Upstream	Closest Upstream	Nested[1] Gene	Target Gene	Closest Downstream	2nd Closest Downstream
D. melanogaster	**Gene Symbol**						
	Strand (+/-)						
Target Species	**NCBI RefSeq Gene (mRNA) Accession**						
	NCBI RefSeq Protein Accession						
	Strand (+/-)						
Best *blastp* Result	**Accession**						
	***D. melanogaster* Gene Symbol[2]**						
	E-Value						
	Percent Identity						
Are the genes in the two species orthologs? (yes/no)							

[1] Leave column blank if target gene is not nested within another gene or another gene is not nested within your target gene.

[2] If the *D. melanogaster* gene symbol of the best match is not provided in the "Description" column of the *blastp* results, you can identify it by clicking on the Accession Number (listed in the "Accession" column). In the new internet browser window that opens, scroll to the "FEATURES" section near the bottom. Next to the "CDS" feature you'll see the gene symbol listed after "/gene=".

19. Paste below an image of the genomic neighborhood of your target gene in the target species from the GEP UCSC Genome Browser including both nearest two upstream and two downstream genes and any nested/nesting genes (if present). Select "default tracks" for the region and, while we didn't discuss this in the walkthrough, set a comparative genomics track (e.g., Drosophila Chain/Net) (if available) to "pack" and then click on "refresh" prior to saving your image so you can submit this in your Annotation Report Form later.

20. Sketch the genomic neighborhood of your target gene in the target species. Be sure your sketch includes the names and/or gene symbols of the surrounding genes and indicates their orientation (Walkthrough Figure 23). Note: You can do this by hand and upload a picture of your drawing or you can create one in Word. (Click on "Insert" in the toolbar ribbon and then click on "Shapes." Under "Block Arrows," click on any style you like. Click on the location in the document where you want the shape to be, then hold and drag your pointer to a different location (determines how large the shape will be) and release the mouse button when it's the desired size.)

21. Compare your sketches of the genomic neighborhoods of your target gene in both *D. melanogaster* (Part 1) and the target species (Part 3). Are there any differences between the two? If so, explain.

22. Explain what evidence supports your hypothesis that you have located the correct genomic neighborhood in the target species (based on your *tblastn* result from Part 2) and are therefore annotating the ortholog to the *D. melanogaster* gene. Then summarize the information from Table 2 above: Be sure to describe any discrepancies found in the *BLAST* results (Part 2) or genomic neighborhood (Part 3).

> Synteny: Explain whether the genes are orthologous and if the genes are on the same strand or not. If one or more genes are non-orthologous to the expected *D. melanogaster* gene(s), explain why you still think you found the ortholog of your target gene within your target species.

Part 4: Determine structure of target gene in *D. melanogaster*

23. Paste a screenshot of your gene in the Gene Record Finder below. Make sure you can see the diagram of the gene under "mRNA details" and the "CDS usage map" (the table that shows all the unique coding exons and which isoforms use each CDS) under the "Polypeptide Details" tab (Walkthrough Figure 27).

24. How many isoforms does your target gene have in *D. melanogaster*? List the name(s) of the isoform(s). Do any of the isoforms have identical coding sequences (if so, how many unique isoforms based on coding sequence are there)?

25. If your target gene in *D. melanogaster* has multiple isoforms, how do they differ (e.g., different coding exons (CDS's) or untranslated regions (UTRs) of different lengths).

26. How many unique coding exons (CDS's) does your target gene have in *D. melanogaster*? Hint: Look at the "CDS usage map".

27. Summarize your findings from Part 4 in Table 3 below:

TABLE 3: Isoforms with unique coding sequences in *D. melanogaster*

Unique isoform(s) based on coding sequence	Other isoforms with identical coding sequences

Note: Depending on your gene, you may have to insert extra rows into this table (see directions at Microsoft Support).

Part 5: Determine approximate location of coding exons (CDS's) in target species

28. Explain what query sequences you are using for each of these *tblastn* searches.

29. Explain what database/subject sequence you are using for these *tblastn* searches.

30. Summarize your findings from Part 5 in Table 4 below:

TABLE 4: Summary of *tblastn* CDS-by-CDS search results

CDS	FlyBase ID	Query Length Size (aa)	*D. melanogaster*		Target Species		Subject Frame
			Query Start	Query End	Subject Start	Subject End	

Note: Depending on your results, you may have to insert extra rows into this table (see directions at Microsoft Support).

Part 6: Refine coordinates of coding exons (CDS's)

31. Complete Table 5 for each unique isoform of your target gene (copy and paste to create as many copies of this table as you need).

TABLE 5: Gene Model for [insert isoform name here] in target species

CDS	FlyBase ID	Frame	Splice Acceptor Phase	Coordinates		Splice Donor Phase
				Start	End	
			0			

Note: Depending on your gene, you may have to insert extra rows into this table (see directions at Microsoft Support).

Coordinates of the stop codon for the above isoform:

Part 7: Verify and submit gene model(s)

Note: Answer all the following questions for each unique isoform.

32. Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and paste a screenshot of the checklist results and data entry section below (Note: make sure to include the “Configure Gene Model” tab (Project Details, Ortholog, Details, and Model Details) in the screenshot).

33. Paste below a screenshot of the dot plot of your submitted model against the putative *D. melanogaster* ortholog (generated by the Gene Model Checker).

34. Paste below an image of the protein alignment generated by the Gene Model Checker (available through the “View protein alignment” link under the “Dot Plot” tab → click on “Download alignment image”).

35. Were there any anomalies on the dot plot and protein alignment (e.g., large gaps, regions with no sequence similarity)? If so, address any anomalies on the dot plot and protein alignment below (Note: 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). Also propose why you think they might be valid. You can include screenshots to illustrate the logic of your explanation.

36. In the “Checklist” results of the Gene Model Checker, click on the 🔍 icon next to “Number of coding exons matched ortholog” within the checklist, and a new window will open showing the Genome Browser view of this region. Your gene model will be shown under the track title “Custom Gene Model.”

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., Splice Junctions, modENCODE TopHat Junctions)
3. a comparative genomics track (e.g., Drosophila Chain/Net)

Paste below an image of your gene model as shown on the Genome Browser (including the above listed tracks).