Pathways Project: Annotation Form

Katie M. Sandlin, Chinmay P. Rele, & Laura K. Reed

Directions: Use this form to document your findings as you locate and annotate your assigned target gene in your target species using the Pathways Project: Annotation Walkthrough as a guide. Instructions on taking a screenshot and how to copy and paste can be found on the Pathways Project page of the GEP website (thegep.org/pathways) in the "Help" section.

Student Details

Student Name(s):	
Student Email(s):	
Instructor:	
Course Name/Number:	
Semester (e.g., Fall 2022):	
College/University:	

Project Details

Target Species (e.g., *D. yakuba*)	
Target Gene's Symbol (e.g., *Rheb*)	

Co-Author Permissions

By submitting this form (via your instructor) to the Genomics Education Partnership (GEP), you acknowledge that you're allowing the data presented here to be published.

If you want to be a co-author on publication(s) arising from this data, you must respond promptly to requests to read and approve the manuscript, and, as part of that review, you will also be required to validate specific data within the manuscript (full instructions

will be provided). If GEP and/or your instructor cannot reach you at the time the publication(s) is ready for review, you won't be listed as a co-author since you aren't able to read and approve the manuscript; instead, you will be listed in the acknowledgements.

	Student #1	Student #2 (if applicable)	Student #3 (if applicable)
Name(s) in the format you want it to be displayed as in publications:			
Permanent email address(es) (e.g., one you will probably use 5 years from now):			
Alternative email address(es) (optional):			
Cell phone number(s) (optional):			
Yes or No: I understand that my instructor may submit this form and supporting documentation to the GEP, who may use this work in a publication. To be a co-author on any publication, I must reply promptly to email from the GEP when the manuscript is ready for review.			

Note: If more than three students contribute to an individual gene annotation as a group project, those students won't be eligible for co-authorship, but their class will be acknowledged.

OPTIONAL A: 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 a **screenshot** 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)[1]. Select "default tracks" for the region and, set a comparative genomics track (e.g., *Drosophila* Conservation (28 Species)) to "pack" and then click on "refresh" prior to taking a screenshot.

3. Sketch the genomic neighborhood (described above) 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 11). Note: You can do this by hand and upload a picture (e.g., taken with a cellphone) of your drawing or you can create one digitally[2 or 3].

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?

OPTIONAL B: Explain which query sequence you are using for this *tblastn* and indicate its length.

OPTIONAL C: Explain which database/subject sequence you are using for this *tblastn*.

OPTIONAL D: When performing a search, BLAST may return any number of matches (often referred to as "hits") for regions of local similarity between the database being searched and the query sequence it's looking for matches to; however, each hit is not necessarily statistically significant. What constitutes a good match/hit?

---

[1] A nested gene, or gene-within-a-gene, refers to a gene that is contained within another external host gene. Most often we find nested genes completely within an intron of, and in the opposite orientation to (i.e., positive vs. negative strand) its host gene. See Kumar (2009) for more information on nested genes.
[2] Using PowerPoint/Google Slides or Word/Google Docs, click on "Insert" in the toolbar ribbon and then click on "Shapes." Under "Block Arrows," select 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.
[3] Templates for sketching the genomic neighborhood of your target gene are provided in PowerPoint and Google Slides.

6. Paste below a **screenshot** of the "Descriptions" panel for the results of your *tblastn* search of the target species' Genome Assembly against the amino acid sequence of the *D. melanogaster* protein-coding isoform for your target gene (Walkthrough Figure 21).

7. Summarize your *tblastn* search results. How many *tblastn* hits to distinct scaffolds were returned? Is there only one very good match, or is there some ambiguity? What scaffold (and accession number) do you think the target gene is on in the target species? (Note: This question is referring to the "Descriptions" panel in the previous question.)

8. After sorting the *tblastn* search results of your <u>best</u> match by "Subject start position," fill in Table 1 below (Walkthrough Figure 25).

	TABLE 1: Summary of the *tblastn* search results for the best scaffold match						
	D. melanogaster		**Target Species**				
Range	**Query Start**	**Query End**	**Subject Start**	**Subject End**	**E-Value**	**Identities (%)**	**Subject Frame**

Depending on your results, you may need to add additional rows by clicking in the bottom-right-most cell and pressing tab (or see Microsoft Support).
You may also need to duplicate this table (via copy and paste) if you have more than one very good match in your *tblastn* search.

9. What coordinates in the target species give the best collinear set of alignments to the protein of your target gene?

OPTIONAL E: Explain why you chose the above coordinates for the best collinear set of alignments.

10. 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

NCBI Taxonomy ID (e.g., 7245)	
NCBI Assembly ID (e.g., Prin_Dyak_Tai18E2_2.1)	
Assembly Accession (e.g., GCF_016746365.2)	
Genome Assembly (e.g., Aug. 2021 (Princeton Prin_Dyak_Tai18E2_2.1/DyakRefSeq3))[4]	

Navigate to the GEP UCSC Genome Browser Gateway page and click on your target species in the "UCSC Species Tree and Connected Assembly Hubs" table. Copy and paste each item found on the right side of the web page to its corresponding row in the table above.

11. Since the "NCBI RefSeq Genes" track indicates predicted mRNA transcripts from RNA-Seq data 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 the ortholog?

12. Looking at the "NCBI RefSeq Genes" track, do any of the closest two upstream or closest two downstream genes have multiple isoforms?

OPTIONAL F: 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 Figure 16 of the Walkthrough.

OPTIONAL G: Explain which query sequence you are using for each of the *blastp* searches.

OPTIONAL H: Explain which database/subject sequence you are using for the *blastp* searches.

---

[4] Each species in the GEP UCSC Genome Browser Gateway has more than one option listed in the "Assembly" drop-down menu. The assembly you should use to annotate your target species is listed in the "Genome Browsers" column of the Pathways Project Genome Assemblies web page.

13. 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* (taxid:7227) reference protein database (refseq_protein)							
		2nd Closest Upstream	Closest Upstream	Nested within or Nesting[1] of Target Gene[5]	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[6]						
	E-Value						
	Percent Identity						
Are the genes in the two species orthologs? (yes/no)							

---

[5] If your target gene is not nested within another gene, or another gene is not nested within your target gene, you can either leave this column blank or put an "X" in each box within the column.

[6] 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 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=".

14. Paste below a **screenshot** 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)[1]. Select "default tracks" for the region and, set a comparative genomics track (e.g., *Drosophila* Chain/Net) (if available) to "pack" and then click on "refresh" prior to taking a screenshot.

15. Sketch the genomic neighborhood (described above) 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 40). Note: You can do this by hand and upload a picture (e.g., taken with a cellphone) of your drawing or you can create one digitally[2, 3].

16. 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.

17. 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* target gene. 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).

18. Describe the presence, or lack, of synteny by explaining whether the genes are orthologous and whether the genes are on the same DNA strand. 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*

19. Paste below a **screenshot** of your gene in the Gene Record Finder. 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 45).

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

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

22. 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**

Depending on your gene, you may need to add additional rows by clicking in the bottom-right-most cell and pressing tab (or see Microsoft Support).

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

OPTIONAL I: Explain which query sequences you are using for each of the *tblastn* searches.

OPTIONAL J: Explain which database/subject sequence you are using for the *tblastn* searches.

23. 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**	

Depending on your results, you may need to add additional rows by clicking in the bottom-right-most cell and pressing tab (or see Microsoft Support).

24. How many <u>total</u> isoforms of your target gene are in your target species? How many <u>unique</u> coding isoforms of your target gene are in your target species?

25. Does your species have the same number of isoforms and isoform structure (i.e., coding exons) as *D. melanogaster*? If yes, type yes as the answer below and proceed to the next question. If no, explain the rationale for the difference (e.g., Is there strong evidence (e.g., extended RNA-Seq, high quality splice sites, conservation across multiple species) for distinct protein coding isoforms present in your species that are <u>not</u> found in *D. melanogaster* (i.e., a possible novel isoform)? If yes, how many novel isoforms? Is your target gene missing an isoform in your target species (i.e., isoform of the target gene in *D. melanogaster* is absent from the target species)? If so, what evidence do you have that supports the hypothesis of a missing isoform (you may include screenshots to illustrate).

NOTE: If your target gene has more than one <u>unique</u> isoform in your target species:

- **Copy and paste** Parts 6-7 for <u>each</u> <u>unique</u> isoform **before** filling in any information.
  - After copying text, scroll to the end of this document and click in the open area following the line divider, and paste.
  - If you have additional unique isoforms, scroll to the end of the document again, and click in the open area following the line divider, then paste again. Repeat this process until you have duplicated Parts 6-7 for <u>each</u> <u>unique</u> isoform.

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

26. Complete Table 5 for the isoform listed in the table's header.

TABLE 5: Gene Model for [INSERT ISOFORM NAME HERE] in target species					
	FlyBase ID	**Frame**		**Coordinates**	**Splice**

9

CDS #			Splice Acceptor Phase	Start	End	Donor Phase
			X			
						X

Depending on your gene, you may need to add additional rows by clicking in the bottom-right-most cell and pressing tab (or see Microsoft Support).

27. Coordinates of the stop codon for the above isoform (e.g., 201-203):

Part 7: Verify and submit gene model(s)

28. Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and paste below a **screenshot** of the "Configure Gene Model" section (Walkthrough Figure 74) and the "Checklist" results (Walkthrough Figure 75).

29. **Copy and paste** the *FINAL* coordinates for your gene model *once you finished the verification process* (i.e., copy the coordinates exactly how you listed them in the "Coding Exon Coordinates" text box of the Gene Model Checker (e.g., 100-200, 300-400).

30. Paste below a **screenshot** of the dot plot generated by the Gene Model Checker against the putative *D. melanogaster* ortholog (Walkthrough Figure 76).

31. Paste below a **screenshot** of the protein alignment generated by the Gene Model Checker (via the "View protein alignment" link under the "Dot Plot" tab; Walkthrough Figure 78)[7].

32. Were there any anomalies on the dot plot and protein alignment (e.g., large gaps, regions with no sequence similarity)? If so, explain how any anomalies are strongly supported by the scientific evidence (you may include screenshots to illustrate)[8].

---

[7] Large gaps, regions with no sequence similarity, and any other anomalies seen in the dot plot can be located within the protein alignment.

[8] **Note: Large <u>vertical and horizontal gaps</u>** near exon boundaries in the dot plot often indicate that an incorrect splice site might have been picked. Re-examine these regions and provide a detailed justification as to why you have selected this set of donor and acceptor sites.

33. 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":
   RNA-Seq Tracks:
      1. at least one transcript prediction track (e.g., TransDecoder Transcripts)
      2. at least one splice-site prediction track (e.g., Combined Splice Junctions)
   Comparative Genomics Tracks
      3. a comparative genomics track (e.g., *Drosophila* Chain/Net)

Paste below a **screenshot** of your gene model as shown on the Genome Browser (including the above listed tracks).

34. Prepare three data files– a General Feature Format File (GFF), a Transcript Sequence File (fasta), and a Peptide Sequence File (pep)—for this unique isoform. The Gene Model Checker automatically creates these three files each time you verify a gene model. (See Part 7.2 in the Walkthrough for instructions on how to obtain these files.)

   Once you obtain the three data files (i.e., GFF, transcript sequence, and peptide sequence files) for each unique isoform, see Part 7.3 of the Walkthrough for instructions on how to merge them into a single file prior to project submission.