Pathways Project: Annotation Form Exemplar

Katie M. Sandlin

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):	Bryce Young and Will Anderson
Student Email(s):	byou@ua.edu and wand@ua.edu
Instructor:	Silverstone
Course Name/Number:	Integrated Genomics/BSC 452
Semester (e.g., Fall 2022):	Spring 2022
College/University:	The University of Alabama

Project Details

Target Species (e.g., *D. yakuba*)	*D. yakuba*
Target Gene's Symbol (e.g., *Rheb*)	*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:	Bryce C. Young	William Anderson Jr.	
Permanent email address(es) (e.g., one you will probably use 5 years from now):	young@gmail.com	will@gmail.com	
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.	Yes	Yes	

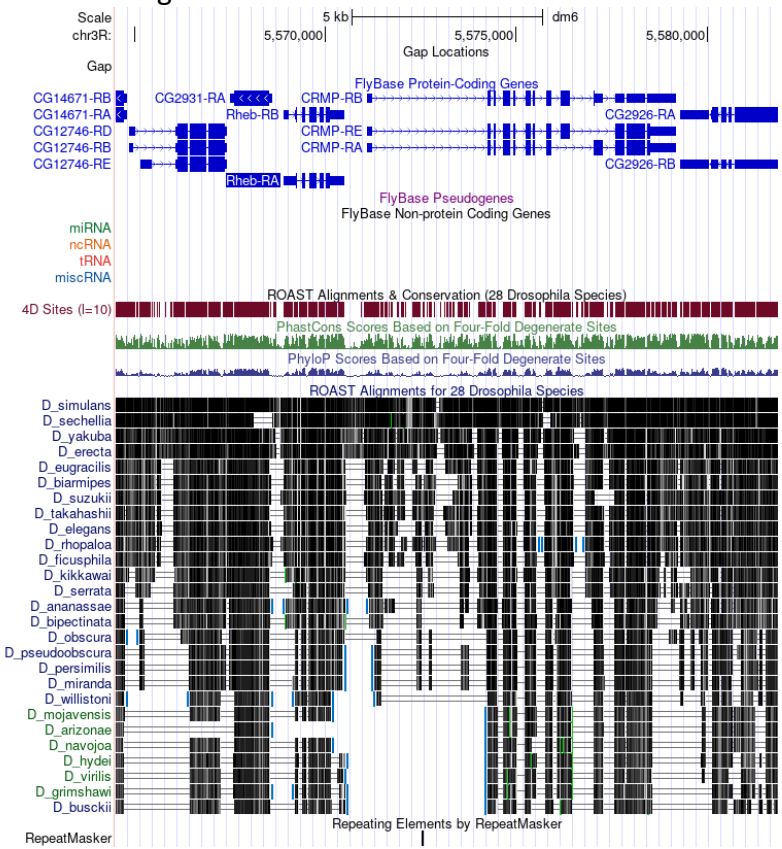
Note: If more than three students contribute to an individual gene annotation as a group project, those students won't be eligible for co-authorship, but their class will be acknowledged.

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

1. Full name and symbol (both italicized) of your target gene in *D. melanogaster*:

Ras homolog enriched in brain (*Rheb*)

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.



[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.

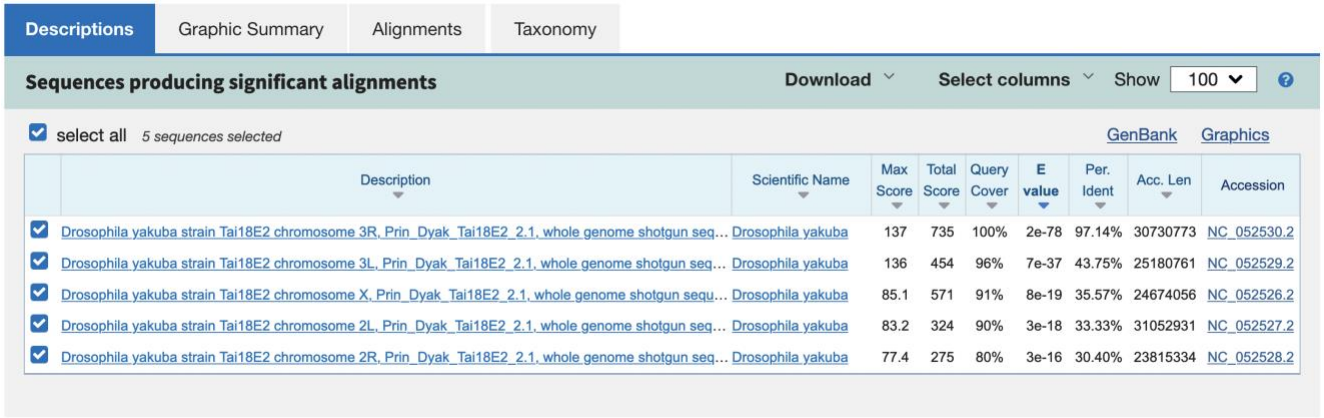
3. Sketch the genomic neighborhood (described above) of your target gene in *D. melanogaster*. Be sure your sketch includes the names and/or gene symbols of the surrounding genes and indicates their orientation (Walkthrough Figure 11). Note: You can do this by hand and upload a picture (e.g., taken with 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):
Drosophila yakuba

5. Does your target gene have multiple isoforms in *D. melanogaster*? If so, how many?
Yes, *Rheb* in *D. melanogaster* has two isoforms (A and B).

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).



Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Drosophila yakuba strain Tai18E2 chromosome 3R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	137	735	100%	2e-78	97.14%	30730773	NC_052530.2
Drosophila yakuba strain Tai18E2 chromosome 3L, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	136	454	96%	7e-37	43.75%	25180761	NC_052529.2
Drosophila yakuba strain Tai18E2 chromosome X, Prin_Dyak_Tai18E2_2.1, whole genome shotgun sequ...	Drosophila yakuba	85.1	571	91%	8e-19	35.57%	24674056	NC_052526.2
Drosophila yakuba strain Tai18E2 chromosome 2L, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	83.2	324	90%	3e-18	33.33%	31052931	NC_052527.2
Drosophila yakuba strain Tai18E2 chromosome 2R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun seq...	Drosophila yakuba	77.4	275	80%	3e-16	30.40%	23815334	NC_052528.2

[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.

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.)

There were five hits to the distinct scaffolds of chromosomes 3R, 3L, X, 2L, and 2R. "*Drosophila yakuba* strain Tai18E2 chromosome 3R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun sequence" (Accession: NC_052530) was the best *tblastn* hit due to it covering 100% of the query (i.e., *D. melanogaster Rheb* protein), having the lowest E-value (2e-78) and highest percent identity (97.14%) among all the *tblastn* hits. The second-best hit was for the chromosome 3L scaffold and, although it had a significant E-value (7e-37) and covered most of the query (96%), the match was only 43.75% identical. Therefore, the *tblastn* search result supports the hypothesis that *Rheb* is located on the chromosome 3R (Accession: NC_052530) scaffold in *D. yakuba*.

8. After sorting the *tblastn* search results of your best 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

Range	*D. melanogaster*		Target Species		E-Value	Identities (%)	Subject Frame
	Query Start	Query End	Subject Start	Subject End			
1	54	130	11,148,243	11,147,992	3e-11	40	-1
2	6	44	11,148,568	11,148,431	0.002	46	-3
3	18	108	11,418,759	11,419,094	1e-06	28	+3
4	111	121	11,419,160	11,419,192	1e-06	73	+2
5	1	20	19,150,809	19,150,868	2e-78	90	+3
6	16	45	19,150,981	19,151,070	2e-78	83	+1
7	40	109	19,151,150	19,151,359	2e-78	97	+2
8	111	153	19,151,422	19,151,550	2e-78	93	+1
9	153	182	19,151,610	19,151,699	2e-78	93	+3

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?
NC_052530:19,150,809–19,151,699

10. Summarize your findings from Part 2 by recording the following information for your target gene in the target species:

Scaffold: chromosome 3R
Scaffold accession number: NC_052530
Approximate coordinates: 19,150,809–19,151,699

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

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

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. NCBI RefSeq Genes" track shows the predicted mRNA transcripts derived from computational predictions and RNA-Seq data isolated from the target species. If there is evidence from the "NCBI RefSeq Genes" track that the region of DNA you identified in Part 2 is being expressed in the target species, record the RefSeq accession number (starts with "XM_") for the possible transcript.

There is only one transcript present (XM_039375862) in this region of the *D. yakuba* NC_052530 scaffold. Expression in this region is supported by the Spaln Alignment of *D. melanogaster* Proteins, the GeMoMa, N-SCAN PASA-EST, and Augustus gene predictions, and RNA-Seq data.

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

According to the NCBI RefSeq genes track, the closest upstream transcript of the putative ortholog is XM_002097995, which is the only transcript in that region. The 2nd closest upstream transcript has four isoforms XM_015192882, XM_015192884, XM_015192883, and XM_002097994.

The closest downstream transcript of the putative ortholog has four isoforms XM_039375860, XM_002097997, XM_039375858, and XM_039375859, and the 2nd closest downstream transcript XM_002097998 is the only transcript in that region.

[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:

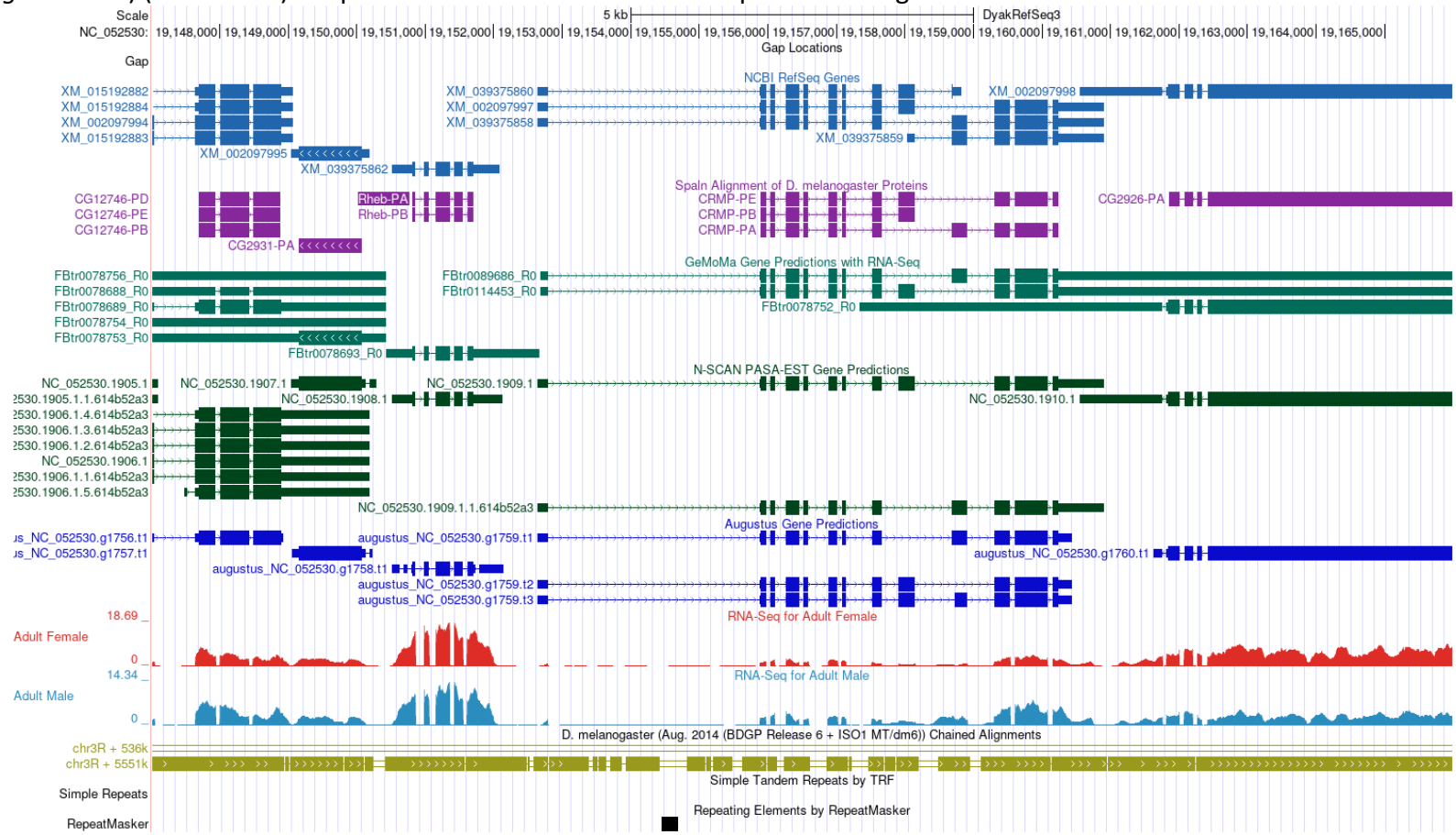
TABLE 2: *blastp* search results for the protein sequences of the genomic neighborhood of the target gene in the target species against the *D. melanogaster* (taxid:7227) reference protein database (refseq_protein)

		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	*CG12746*	*CG2931*	*X*	*Rheb*	*CRMP*	*CG2926*
	Strand (+/-)	+	-	X	+	+	-
Target Species	NCBI RefSeq Gene (mRNA) Accession	XM_015192882	XM_002097995	X	XM_039375862	XM_002097997	XM_002097998
	NCBI RefSeq Protein Accession	XP_015048368	XP_002098031	X	XP_039231796	XP_002098033	XP_002098034
	Strand (+/-)	+	-	X	+	+	-
Best *blastp* Result	Accession	NP_649551	NP_649552	X	NP_730950	NP_730954	NP_649554
	D. melanogaster Gene Symbol[6]	*CG12746*	*CG2931*	*X*	*Rheb*	*CRMP*	*CG2926*
	E-Value	0.0	0.0	X	6e-131	0.0	0.0
	Percent Identity	84.30%	96.72%	X	97.25%	99.66%	86.18%
Are the genes in the two species orthologs? (yes/no)		yes	yes	X	yes	yes	yes

[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=". When in doubt, check FlyBase to confirm gene symbol.

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 for *Drosophila melanogaster* (e.g., *D. melanogaster* Chain or *D. melanogaster* 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.

Comparing *D. melanogaster* chr3R with the *D. yakuba* scaffold NC_052530, *Rheb* and its two closest upstream neighbors (i.e., *CG2931* and *CG12746*) and two closest downstream neighbors (i.e., *CRMP* and *CG2926*) are orthologs in both species.

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. S*ummarize* the information from Table 2 above—be sure to describe any discrepancies found in the BLAST results (Part 2) or genomic neighborhood (Part 3).

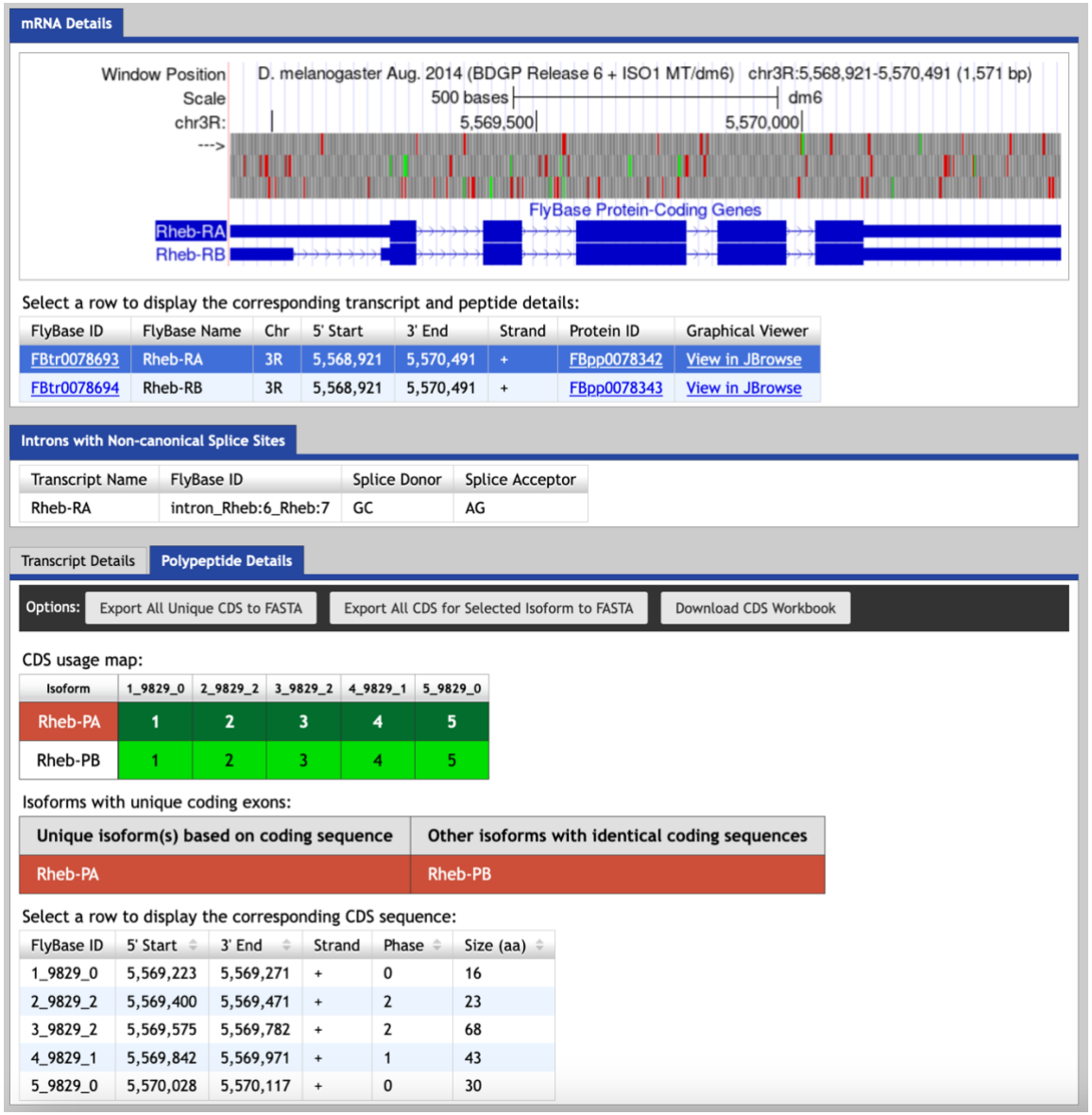
The hypothesis that the XM_039375862 feature is the putative ortholog of *Rheb* in *D. yakuba* is supported by three lines of evidence. First, the tblastn search shows that the best match to the *D. melanogaster Rheb* gene is located at a region within scaffold NC_052530, which contains the RefSeq transcript XM_039375862. Second, the blastp search shows that the best match to the protein product derived from XM_039375862 (i.e., XP_039231796) is to the *Rheb* gene in *D. melanogaster*. Third, the blastp searches showed the two upstream and two downstream neighbors of *Rheb* in both *D. yakuba* and *D. melanogaster* are orthologs.

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.

Comparing *D. melanogaster* chr3R with the *D. yakuba* NC_052530 scaffold, *Rheb* and its two closest upstream neighbors (i.e., *CG2931* and *CG12746*) two closest downstream neighbors (i.e., *CRMP* and *CG2926)* are locally syntenic and orthologous. Furthermore, the orientation of the target gene (+) and its two closest upstream (+, -) and two closest downstream (+, -) neighbors in *D. yakuba* are consistent with *D. melanogaster* (+, -, +, +, -).

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)?

Rheb in *D. melanogaster* has two isoforms, A and B. Both isoforms have identical coding sequences so there is one unique isoform based on coding sequences.

21. How many unique coding exons (CDS's) does your target gene have in *D. melanogaster*? Hint: Look at the "CDS usage map."
Five: 1_9829_0, 2_9829_2, 3_9829_2, 4_9829_1, and 5_9829_0

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**
Rheb-PA	Rheb-PB

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

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

TABLE 4: Summary of *tblastn* CDS-by-CDS search results
"Compositional adjustments"- No adjustment and uncheck box for filtering "Low complexity regions"

CDS	FlyBase ID	Query Length Size (aa)	*D. melanogaster*		Target Species		Subject Frame
			Query Start	Query End	Subject Start	Subject End	
1	1_9829_0	16	1	16	19,150,809	19,150,856	+3
2	2_9829_2	23	1	23	19,150,987	19,151,055	+1

3	3_9829_2	68	1	68	19,151,156	19,151,359	+2
4	4_9829_1	43	1	43	19,151,422	19,151,550	+1
5	5_9829_0	30	1	30	19,151,613	19,151,702	+3

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 total isoforms of your target gene are in your target species and what are their letters? How many unique coding isoforms of your target gene are in your target species and what are their letters?

2 total isoforms (A and B); 1 unique coding isoform (A)

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 not 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).

Yes

NOTE: If your target gene has more than one unique isoform in your target species:

- **Copy and paste** Parts 6-7 below for each unique 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 each unique isoform.

Part 6: Refine coordinates of coding exons (CDS's) {Isoform A}

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

TABLE 5: Gene Model for Rheb-PA in target species

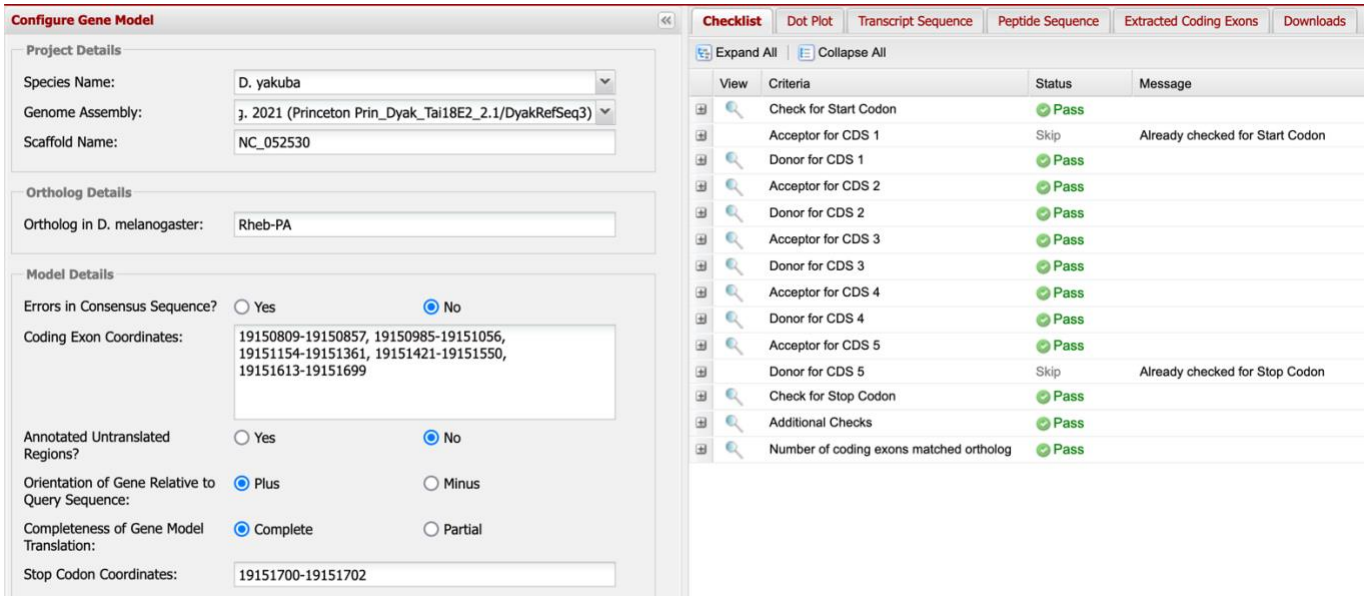
CDS #	FlyBase ID	Frame	Splice Acceptor Phase	Coordinates		Splice Donor Phase
				Start	End	
1	1_9829_0	+3	X	19,150,809	19,150,857	1
2	2_9829_2	+1	2	19,150,985	19,151,056	1
3	3_9829_2	+2	2	19,151,154	19,151,361	2
4	4_9829_1	+1	1	19,151,421	19,151,550	0
5	5_9829_0	+3	0	19,151,613	19,151,699	X

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): 19,151,700-19,151,702

Part 7: Verify and submit gene model(s) {Isoform A}

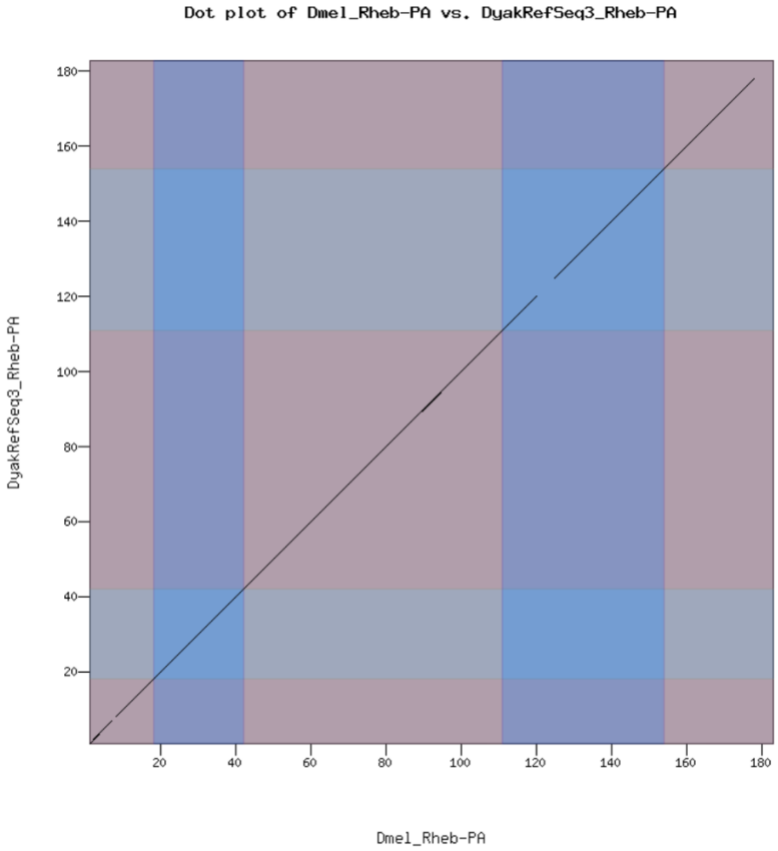
28. Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and paste below a **screenshot** of the "Configure Gene Model" section (Walkthrough Figure 74) and the "Checklist" results (Walkthrough Figure 75).



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).

19150809-19150857, 19150985-19151056, 19151154-19151361, 19151421-19151550, 19151613-19151699

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].

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

Identity: 177/182 (97.3%), **Similarity:** 178/182 (97.8%), **Gaps:** 0/182 (0.0%)

```
Dmel_Rheb-PA          1 MPTKERHIAMMGYRSVGKSSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDT 60
                        ******:**************************************************
DyakRefSeq3_Rheb-PA   1 MPTKERNIAMMGYRSVGKSSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDT 60

Dmel_Rheb-PA         61 AGQDEYSIFPVQYSMDYHGYVLVYSITSQKSFEVVKIIYEKLLDVMGKKYVPVVLVGNKI 120
                        ************************************************************
DyakRefSeq3_Rheb-PA  61 AGQDEYSIFPVQYSMDYHGYVLVYSITSQKSFEVVKIIYEKLLDVMGKKYVPVVLVGNKT 120

Dmel_Rheb-PA        121 DLHQERTVSTEEGKKLAESWRAAFLETSAKQNESVGDIFHQLLILIENENGNPQEKSGCL 180
                        **: ***************************************************.**
DyakRefSeq3_Rheb-PA 121 DLQPERTVSTEEGKKLAESWRAAFLETSAKQNESVGDIFHQLLILIENENGNPQEKSSCL 180

Dmel_Rheb-PA        181 VS 182
                        **
DyakRefSeq3_Rheb-PA 181 VS 182
```

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].

The protein alignment of the Rheb-PA isoform shows that the Rheb-PA ortholog in *D. yakuba* (DyakRefSeq3_Rheb-PA) has 97.3% identity, 97.8% similarity, and 0.0% gaps to the Rheb-PA isoform of *D. melanogaster* (Dmel_Rheb-PA).

The Dot Plot shows a few regions lacking sequence similarity throughout the Rheb-PA isoform, most of which are in CDS-4; however, as shown in the protein alignment, the two regions of CDS-4 that don't show similarity are only one amino acid (or three base pairs) in length.

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."

[8] **Note: Large vertical and horizontal gaps** 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.

Select the "default tracks" for the region, set the following evidence tracks (if available) to "pack", and then click on "refresh":

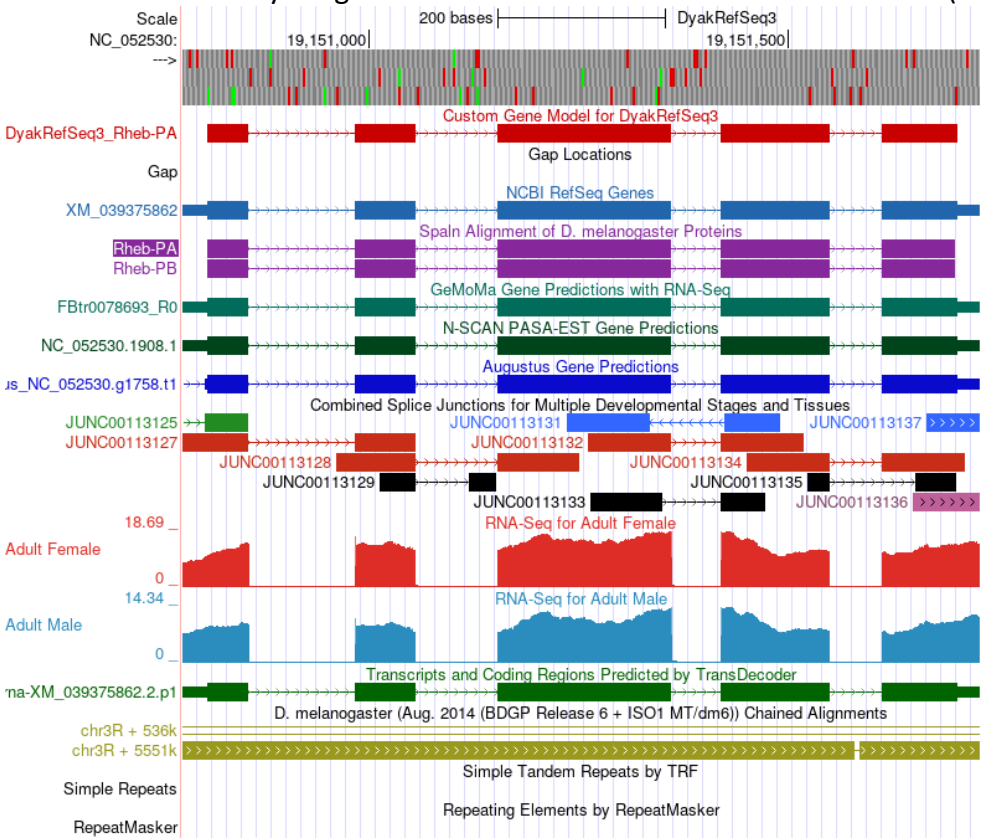
RNA-Seq Tracks:

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 <u>each</u> <u>unique</u> isoform, see Part 7.3 of the Walkthrough for instructions on how to merge them into a single file prior to project submission. To name these files, use your assigned species ("d" followed by the first three letters of your species) + "_gene.filetype" as the format. For example, if you're annotating *Ilp8* in *D. grimshawi*, the merged GFF would be titled "dgri_Ilp8.gff"
