BIOL199 BDB
Fall 2022

Pathways Project Annotation Walkthrough Parts D-F

*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:

- Use the Genome Browser and Gene Record Finder to determine the structure of your target gene in *D. melanogaster*
- Apply skills in using the Genome Browser and *BLAST* to identify and confirm positions, frames, and phases for annotating your target gene
- Identify support and discrepancy in different lines of evidence to discuss in relation to your gene annotation

Pre-lab Assignments:

1. Read through the lab protocol and review Appendix 7-8.
2. Complete the pre-lab assignment on the hard copy I handed out in class

Introduction

By this point, you should have determined a region in your target species orthologous to your target gene, and accrued evidence to support this hypothesis. In the remainder of the project, our goals are to:

1. Understand the structure of our gene in *D. melanogaster* (Part D)
2. Determine the approximate location of the coding sequences of the *D. melanogaster*'s isoforms in our target species (Part E)
3. Use our understanding of codons, reading frames, and phasing to specify important coordinates of the target gene in our target species (i.e. **manual annotation**) (Part F, pre-interlude)
4. Verify these coordinates using RNA-Seq data, exon junctions, and gene model predictions (Part F, post-interlude)
5. Complete our gene model and use the results to compare the target gene in our target species to the target gene in *D. melanogaster*. (Comparison is Part G, not included in this protocol)

We will again use the *Rheb* gene in *D. yakuba* as our example in the following steps. Remember that the *Rheb* gene in *D. yakuba* is relatively straightforward in comparison to what your own project might be. Some steps might appear to be excessive, but keep in mind that your gene/species may be much more complex, so it is important to follow this protocol as we have tried to equip you with most of what you will need should you encounter complexities.

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

You have completed this as a pre-lab, a link to the protocol and worksheet are linked here (http://tiny.cc/PathwaysD). Compare what you wrote to that of your group mates, and if everyone is comfortable with the steps in Part D, move on to Part E!

End of Part D DISCUSS and CHECK-IN: Compare your partially filled tables with your partner or group, and check understanding of the number of isoforms vs. number of unique isoforms.

Part E: Determine approx. location of coding sequences in target species

In part B, we used *tblastn* to determine the region and genomic scaffold of our putative ortholog in our target species. Here, we will use *tblastn* to determine the approximate coordinates of each coding sequence (CDS), by aligning each CDS of the gene in *D. melanogaster* against this search region of the target genome. This step is similar to the instructions corresponding to Table 3 in CP1.

Because the ***BLAST*** **algorithm does not take the positions of potential splice sites within a complete protein sequence into account when it generates the alignment**, *BLAST* often extends the alignment beyond the coding exon boundary and into the intron (this is related to the discrepancy we observed in Table 2 of CP1). To ameliorate this issue, the GEP annotation protocol recommends mapping each coding exon separately to determine their approximate locations and then further refine the exon boundaries by searching for compatible splice donor and acceptor sites through visual inspection using the Genome Browser (further refining is the objective of Part F).

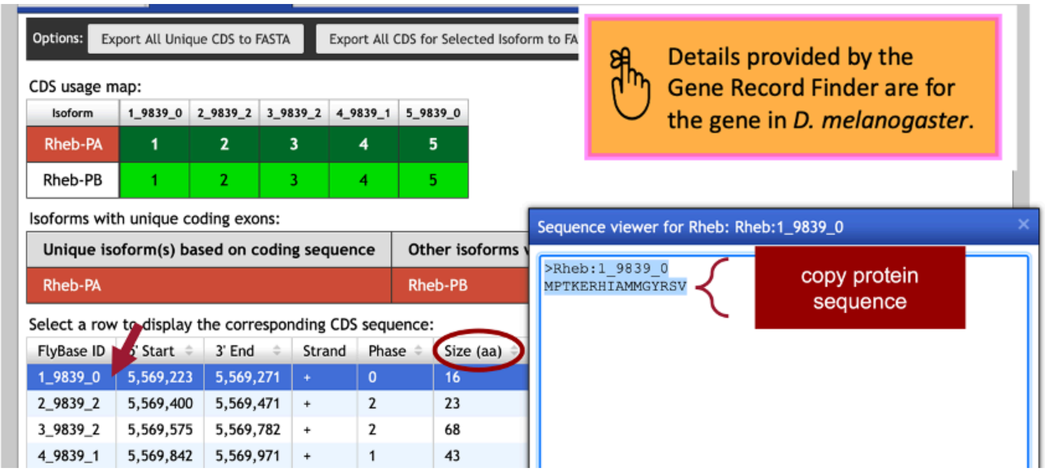
Here, we will perform a *tblastn* search for each coding sequence, using the 'comparing two or more sequences against each other' option (i.e. using *bl2seq* or *BLAST* 2 sequences). Each CDS for *D. melanogaster* of your target gene is thus your query sequence, and the accession number for our target species' scaffold is the database (subject) sequence.

Instructions	**Comments**
1. If not already open, open a new web browser tab and navigate to the Gene Record Finder (http://tiny.cc/generecordfinder). Enter your target gene and click on the 'Find Record' button.	
2. You likely have two isoforms with unique CDS that you listed in Tables 3-4. Start with the isoform you listed in Table 3 and note that you will fill out similar information for Table 4 as well (though results for the same coding sequence in	

both isoforms can be copied from your previous table). This means you will repeat steps #3-8 for each CDS - the number of times you repeat these steps will be dependent on the number of CDS you have.

3. Under the 'Polypeptide Details' tab, scroll to the bottom with a table starting with a column titled 'FlyBase ID'. Click the first row of this table to view the protein sequence for CDS-1 for your first isoform and copy the entire protein sequence (including the header) in the pop-up window.

Remember to only look at the coding sequences in the isoform you are starting to manually annotate.



Options: Export All Unique CDS to FASTA | Export All CDS for Selected Isoform to FA

Details provided by the Gene Record Finder are for the gene in *D. melanogaster*.

CDS usage map:

Isoform	1_9839_0	2_9839_2	3_9839_2	4_9839_1	5_9839_0
Rheb-PA	1	2	3	4	5
Rheb-PB	1	2	3	4	5

Isoforms with unique coding exons:

Unique isoform(s) based on coding sequence	Other isoforms v
Rheb-PA	Rheb-PB

Select a row to display the corresponding CDS sequence:

FlyBase ID	5' Start	3' End	Strand	Phase	Size (aa)
1_9839_0	5,569,223	5,569,271	+	0	16
2_9839_2	5,569,400	5,569,471	+	2	23
3_9839_2	5,569,575	5,569,782	+	2	68
4_9839_1	5,569,842	5,569,971	+	1	43

Sequence viewer for Rheb: Rheb:1_9839_0

```
>Rheb:1_9839_0
MPTKERHIAMMGYRSV
```

copy protein sequence

For *Rheb* in *D. melanogaster*, CDS-1 is 1_9839_0, a CDS found in Rheb-PA and Rheb-PB.

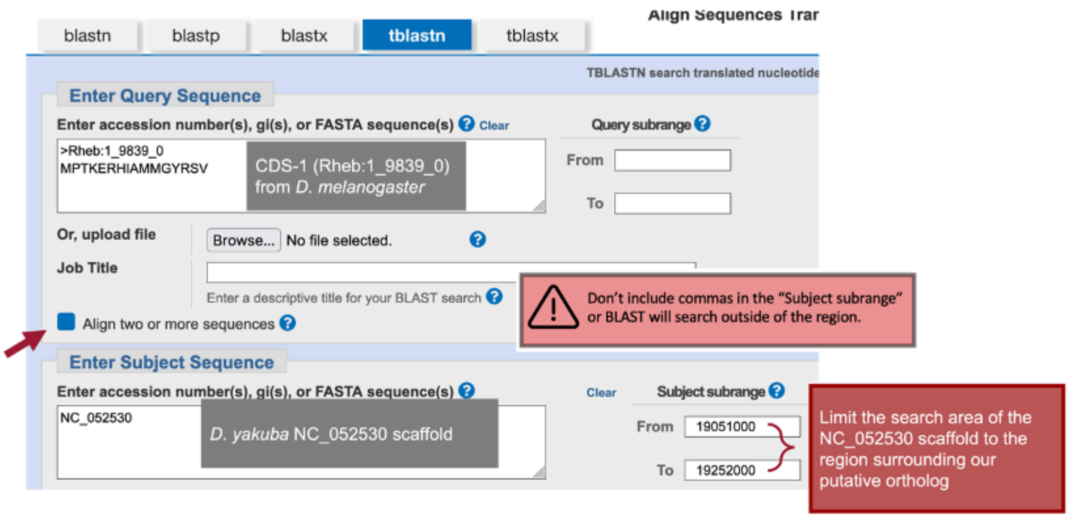
4. Navigate to the NCBI *BLAST* website (http://tiny.cc/blastpage) and click on the *tblastn* image under the Web BLAST section.

5. Fill in the tblastn search using the following instructions:
- For "Enter Query Sequence", paste the sequence for CDS-1.
- Select the "Align two or more sequences" checkbox.
- For "Enter Subject Sequence", enter the accession number identified for your target species (see Pathways2 question B6).
- For subject subrange, **provide a range containing your putative ortholog region, ±50,000-100,000 base pairs**. Round numbers are fine, but make sure commas are NOT included.
- Click on the "+" icon next to "Algorithm parameters" to expand the section
- In the "Scoring Parameters" section, change the "Compositional adjustments" field to "No adjustment."
- In the "Filters and Masking" section, uncheck the "Low complexity regions" checkbox in the "Filter" field.
- Select the check box next to "Show results in a new window".
- Click on the "BLAST" button.

Accession number and putative ortholog region coordinates can be found on Pathways2 worksheet, question B6.

The putative ortholog of Rheb-PA is located at *approximately* 19,150,809 – 19,151,699 on the NC_052530 scaffold for *D. yakuba*. The "Subject subrange" was used to limit the search region to 19,051,000-19,252,000, which was determined by subtracting 100,000 from the smaller coordinate (19,150,809), adding 100,000 to the larger coordinate (19,151,699), and then rounding each to the nearest thousandth.

Images on the next page show what the *BLAST* form should look like.


Align Sequences Trar

blastn | blastp | blastx | tblastn | tblastx

TBLASTN search translated nucleotide

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

>Rheb:1_9839_0
MPTKERHIAMMGYRSV

CDS-1 (Rheb:1_9839_0) from *D. melanogaster*

Query subrange

From

To

Or, upload file: Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Don't include commas in the "Subject subrange" or BLAST will search outside of the region.

Enter Subject Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

NC_052530

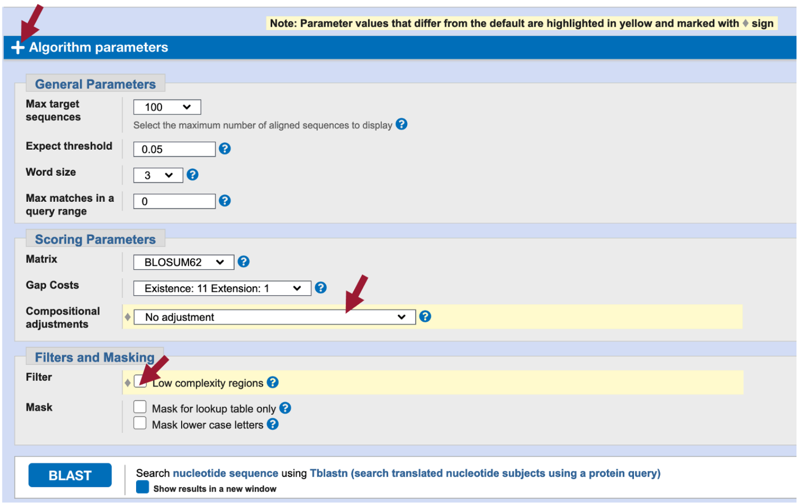
D. yakuba NC_052530 scaffold

Subject subrange

From 19051000

To 19252000

Limit the search area of the NC_052530 scaffold to the region surrounding our putative ortholog


Note: Parameter values that differ from the default are highlighted in yellow and marked with ♦ sign

+ Algorithm parameters

General Parameters

Max target sequences: 100

Select the maximum number of aligned sequences to display

Expect threshold: 0.05

Word size: 3

Max matches in a query range: 0

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: ♦ No adjustment

Filters and Masking

Filter: ♦ □ Low complexity regions

Mask: □ Mask for lookup table only

□ Mask lower case letters

BLAST

Search nucleotide sequence using Tblastn (search translated nucleotide subjects using a protein query)

Show results in a new window

6. Check the number of hits on the following results page. Hopefully there is only one match, or if not one match, only one GOOD match. If there is ambiguity, check with your instructor or TA.

Remember to use E-values and percent identity as your first check for a good match.

7. Click on the "Alignments" tab to view the corresponding *tblastn* alignment. Use the provided information to fill out the rest of Table 3 for CDS-1.

Answer questions E1d-E1f for the corresponding row. Make sure the 'Subject frame' column includes both the symbol (+/-) and the number.

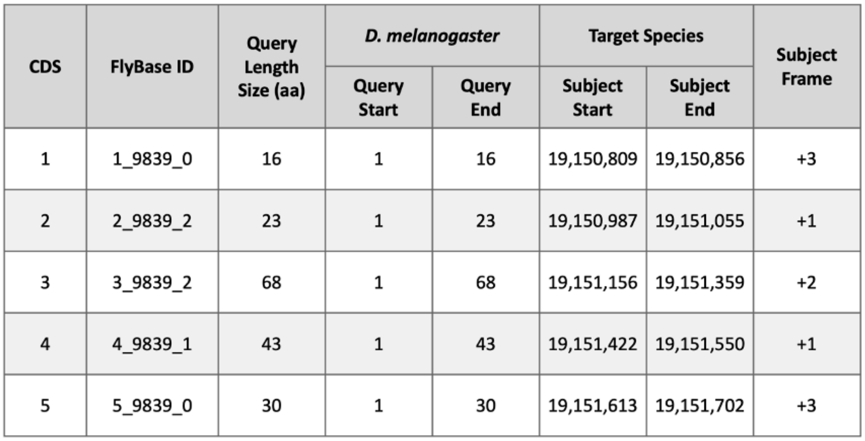
CDS	FlyBase ID	Query Length Size (aa)	*D. melanogaster*		Target Species		Subject Frame
			Query Start	Query End	Subject Start	Subject End	
1	1_9839_0	16	1	16	19,150,809	19,150,856	+3

DISCUSS and CHECK-IN: Compare what you find with your partner or group to make sure you have the same values. Call your instructor or TA over to share what data you put in your table and justify why you chose that alignment.

8. Take screenshots of both the Descriptions and Alignments tabs to put in your lab notebook. Don’t forget to label the screenshots on your notebook with the FlyBase ID (e.g. “<FlyBaseID> Description” and “FlyBaseID Alignment”) to refer back to the *tblastn* search easily if there is anything you need to review.

This way, you can hopefully just refer to your notebook rather than running through the *tblastn* search again.

9. Repeat the above for each of the remaining CDS’s from the Gene Record Finder for your isoform, and fill out the rest of Table 3.



CDS	FlyBase ID	Query Length Size (aa)	*D. melanogaster*		Target Species		Subject Frame
			Query Start	Query End	Subject Start	Subject End	
1	1_9839_0	16	1	16	19,150,809	19,150,856	+3
2	2_9839_2	23	1	23	19,150,987	19,151,055	+1
3	3_9839_2	68	1	68	19,151,156	19,151,359	+2
4	4_9839_1	43	1	43	19,151,422	19,151,550	+1
5	5_9839_0	30	1	30	19,151,613	19,151,702	+3

A table showing the results for the *Rheb* gene comparing *D. melanogaster’s Rheb* CDS’s to the scaffold and putative ortholog region for *D. yakuba*. Examination of the subject ranges for the *tblastn* alignments of the five CDS’s of *Rheb* in *D. yakuba* shows that they are collinear—CDS’s 1-5 are on the + strand and ranges are in ascending order. The CDS-by-CDS search results support the hypothesis that the putative ortholog of Rheb-PA is located at ~19,150,809 – 19,151,702 on the NC_052530 scaffold of the *D. yakuba* genome assembly (i.e., **NC_052530:19,150,809-19,151,702**).

Note that you can return to your original *tblastn* search web browser tab and replace the previous CDS in the ‘Enter Query Sequence’ textbox, leaving everything else the same for a faster BLAST search.

Fill out the rest of Table 3 in question E1.

10. Assess your table. Consider whether the CDS’s are on the same strand (all + or all -), whether the subject ranges all ascend or all descend, and whether the ranges overlap the putative region you hypothesized in question B6.

DISCUSS and CHECK-IN: Present your table on the location of each CDS, and your support on whether you have found all the CDS in this region to your instructor or TA. This could be additional support (or discrepancy) for your hypothesis from Pathways Part B, if you wanted to add it!

See the description of the table in step 9 for an example of the logic used to support such a hypothesis.

Answer question E2.

11. **Save this step for AFTER you complete Part F, as it just repeats for the next isoform - better to work through the whole process for one isoform and then return to the next one after you are comfortable with all steps.** Repeat steps as needed from above for your remaining isoform in Table 4. Remember that coding sequences that match what you found already in Table 3 can be copied over, and only new coding sequences require the *tblastn* search (and screenshot documentation in your lab notebook).

Answer the remaining parts of questions E1-E2.

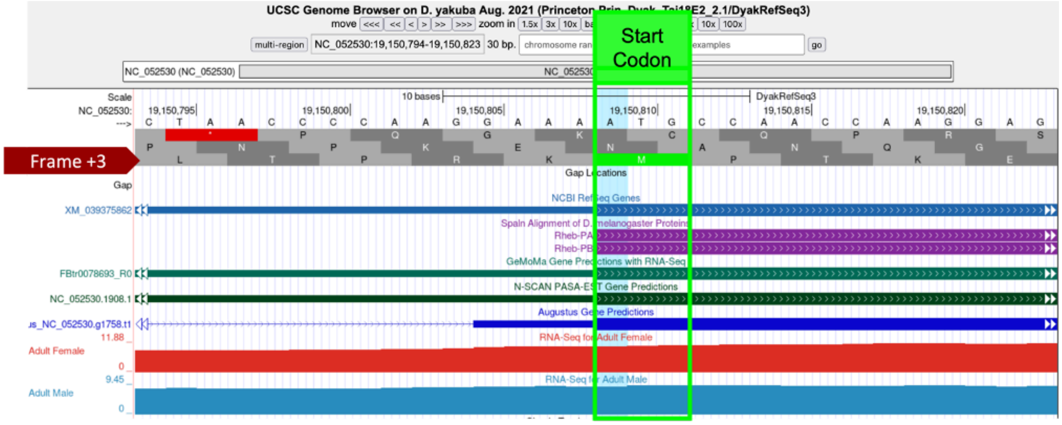
In this section, we mapped each CDS separately to determine their approximate locations and establish with more confidence that we have found the orthologous region to the target gene.

Before beginning Part F, take a break if you haven’t done so already. Discuss with your group the length of your break (~3-5 minutes suggested), and agree on a time to come back. Take a moment to step outside or go for a walk around Gottwald before coming back to work on lab materials.

Part F: Refine coordinates of coding sequences in exons (CDS's)

In this section, we will further refine the CDS boundaries by searching for compatible splice donor and acceptor sites by visual inspection using the Genome Browser. We will use RNA-Seq data generated for your target species and our understanding of gene structure to manually annotate our CDS's, similar to the steps you did in the Understanding Eukaryotic Genes (UEG) Lab from Weeks 4-5. Note that we will repeat the steps below for each unique isoform, and instructions below are assuming you are looking at the unique isoform listed in Table 3.

Instructions	**Comments**
1. Return to the Genome Browser (http://tiny.cc/GEP_BrowserGateway) for your target species.	
2. To examine the region of the *tblastn* alignment for CDS-1, in the "enter position or search terms" text box, enter the accession number followed by the coordinate span for CDS-1 for the isoform you examined in Table 3.	For *D. yakuba* and CDS-1 for Rheb-PA, this is 'NC_052530: 19,150,809-19,150,856'. **Fill in the title for Table 5 in question F1.**
3. For the resulting Genome Browser window, update the following settings: ● Under the "Mapping and Sequencing Tracks," change the "Base Position" track to "full." ● Make sure the track for RNA-Seq coverage is set to "full." ● Click on a 'refresh' button to update any changes.	Note that you should see 2-3 histograms: a red one for adult females, a blue one for adult males, and in some cases, a purple one for mixed embryos.
4. We first must verify our start codon. To do this, we must examine the region around the suspected start of the first coding sequence and find support through multiple evidence tracks. Do the following: ● In the 'enter position or search terms' text box, enter your accession followed by the first position of CDS-1, and then press 'go;. ● Zoom out 3x and then another 10x. ● Make sure that the coding strand is represented by the bases given at the top (i.e., make sure the arrow on the left points in the same direction as the direction of transcription). ● Check that a start codon is available and determine the frame. ○ Check if the reading frame is consistent with the frame from the *tblastn* results in Table 3 (frame for first CDS). ○ Check that the start codon aligns with multiple gene alignment and prediction tracks (e.g. BLAT, Spaln, GeMoMa, Geneid, Augustus). ○ If there are discrepancies, try to resolve where you believe the start codon is.	For *D. yakuba* and CDS-1 for *Rheb*, search for 'NC_052530:19,150,809'.



5. To make sure there is not an alternative start codon further upstream, do the following:

- Check the number of amino acids in CDS-1 in Table 3 for *D. melanogaster*, and see if that corresponds to the number of amino acids in CDS-1 in your target species.
- If they are similar, then you likely found the correct start codon. If they are not similar, complete the remaining steps.
- Zoom out 3x and another 10x to observe a larger region upstream of the putative start codon where evidence of transcription has occurred (i.e. where the RNA-Seq histograms show data).
- Check whether there are additional start codons in this upstream region. If yes, examine whether any of those start codons are more appropriate than your original start codon.

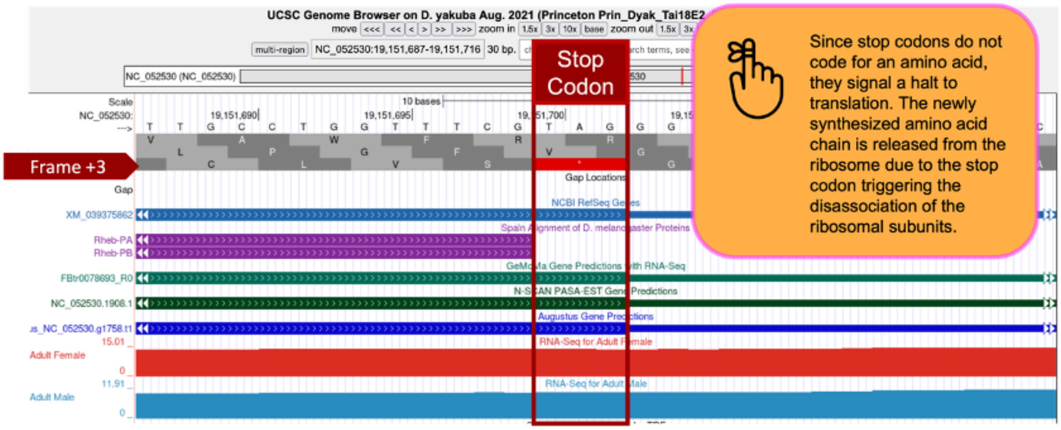
The *tblastn* alignment for CDS-1 (CDS 1_9839_0) of *Rheb* in *D. melanogaster* against the *D. yakuba* scaffold encompasses all 16 amino acids of CDS-1, and the alignment begins with a start codon at 19,150,809-19,150,811. Hence the most **parsimonious** gene model for Rheb-PA in *D. yakuba* would use the start codon at 19,150,809-19,150,811 in scaffold NC_052530.

6. When you have found the start codon you believe is associated with your target gene, record its starting coordinate and the associated frame in Table 5.

Answer questions F1a-F1b by filling them in from Table 3. Then, answer questions F1c and F1e for CDS-1 using your current findings.

7. Now let’s verify our stop codon coordinates. To do this, we must examine the region around the suspected end of the last coding sequence and find support through multiple evidence tracks. Do the following:

- Determine the last CDS for the isoform you are examining in Table 3 and copy the final position in the last CDS.
- In the ‘enter position or search terms’ text box, enter your accession followed by the last position of your final CDS, and then press ‘go;.
- Zoom out 3x and then another 10x.
- Check that a stop codon is available and determine the frame.
    - Check if the reading frame is consistent with the frame from the *tblastn* results in Table 3 (frame for last CDS).
    - Check that the stop codon aligns with multiple gene alignment and prediction tracks (e.g. BLAT, Spaln, GeMoMa, Geneid, Augustus).
    - If there are discrepancies, try to resolve where you believe the stop codon is.


UCSC Genome Browser on D. yakuba Aug. 2021 (Princeton Prin_Dyak_Tai18E2
move <<< << < > >> >>> zoom in 1.5x 3x 10x base zoom out 1.5x 3x
multi-region NC_052530:19,151,687-19,151,716 30 bp.
NC_052530 (NC_052530)
Stop Codon
Frame +3
Scale
10 bases
NC_052530:
19,151,690
19,151,695
19,151,700
T T G C C T G G T T T C G T A G G G
V A W F R R
L P G F V G
C L V S G
Gap Locations
Gap
NCBI RefSeq Genes
XM_039375862
Spaln Alignment of D. melanogaster Proteins
Rheb-PA
Rheb-PB
GeMoMa Gene Predictions with RNA-Seq
FBtr0078693_R0
N-SCAN PASA-EST Gene Predictions
NC_052530.1908.1
Augustus Gene Predictions
us_NC_052530.g1758.t1
RNA-Seq for Adult Female
15.01
Adult Female
0
RNA-Seq for Adult Male
11.91
Adult Male
0

Since stop codons do not code for an amino acid, they signal a halt to translation. The newly synthesized amino acid chain is released from the ribosome due to the stop codon triggering the disassociation of the ribosomal subunits.

Based on the *tblastn* alignment for CDS-5 and the available evidence on the Genome Browser, the stop codon for the Rheb-PA ortholog is placed at 19,151,700-19,151,702, and the last codon (S; Serine), before the stop codon, ends at 19,151,699.

For *D. yakuba* and the Rheb-PA isoform, the last CDS is CDS-5. From Part E, we would search for ‘NC_052530:19,151,702’.

Note that the RNA-Seq read coverage tracks for both adult females and adult males indicate that transcription extends beyond the stop codon.

The region with RNA-Seq read coverage that extends beyond the stop codon likely corresponds to the 3’ untranslated region (UTR) of the last exon in *Rheb*.

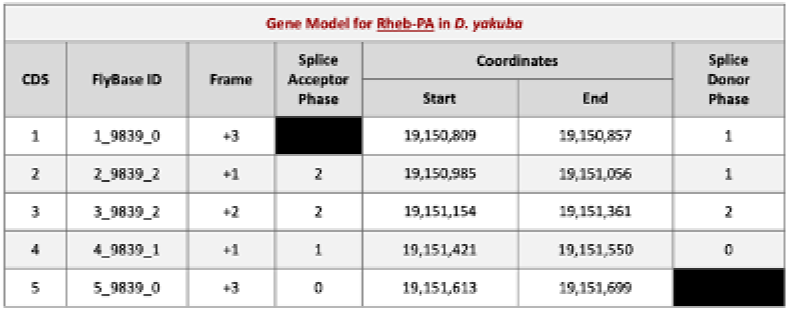
8. When you have found the stop codon you believe is associated with your target gene, record the last coordinate **before** the stop codon and the associated frame in Table 5.

Answer questions F1c and F1f for the last CDS using your current findings. Then, answer question F1h.

9. Knowing the location of the stop and start codons, we can now examine splice donors and acceptors at the beginning and end of each intron, to determine the coordinates and reading frames associated with exons.

- To examine the genomic region surrounding the splice donor site of CDS-1, enter the last position of CDS-1 into the “enter position or search terms” text box and click on the ‘go’ button.
- Zoom out 3x and then 10x to examine the 30 bp around this position.
- Use the reading frame you found for Table 5c to determine the splice donor phase at the end of CDS-1.
- Use the phase you found to determine the splice acceptor phase at the start of CDS-2.
- Note the coordinate of the last base pair in CDS-1 and the phases at the splice donor and acceptor in Table 5.
- Zoom out far enough to see the entire length of CDS-1 and confirm that your CDS-1 has an open reading frame (ORF), i.e., no stop codons are shown within the frame of CDS-1).
- To examine the genomic region surrounding the splice acceptor site of CDS-2, enter the first position for CDS-2 from Table 3 into the “enter position or search terms” text box and press ‘go’.
- Zoom out 3x and another 10x to examine the 30 bp surrounding this position.
- Based on the splice acceptor phase you just calculated from the CDS-1 splice donor phase, determine the reading frame for CDS-2.
- Note the reading frame and start coordinate of CDS-2 in Table 5.
- Use the same strategy above for the remaining CDS’s, making sure to check ORFs to ensure there are no stop codons.
- Make sure your final CDS frame matches what you expected from when you analyzed the stop codon.
- A link explaining the results for Rheb-PA in *D. yakuba* is in Appendix 7.

10. Check that everything is consistent in your table, particularly that your frames match what you found in Table 3 from the BLAST search, and your phases add up appropriately to make full non-contiguous codons.



Gene Model for Rheb-PA in *D. yakuba*

CDS	FlyBase ID	Frame	Splice Acceptor Phase	Coordinates		Splice Donor Phase
				Start	End	
1	1_9839_0	+3		19,150,809	19,150,857	1
2	2_9839_2	+1	2	19,150,985	19,151,056	1
3	3_9839_2	+2	2	19,151,154	19,151,361	2
4	4_9839_1	+1	1	19,151,421	19,151,550	0
5	5_9839_0	+3	0	19,151,613	19,151,699	

DISCUSS and CHECK-IN: Make sure your values are similar to that of your partner or group and show your results to your instructor or TA to make sure they make sense.

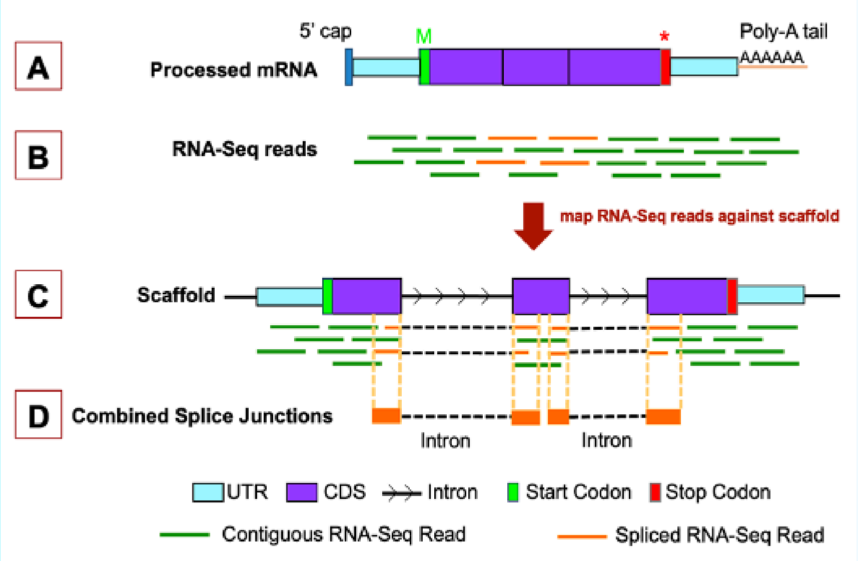
In *D. melanogaster*, ~99% of introns have a GT splice donor site and ~1% have a GC non-canonical splice donor site. Most introns have a GA splice acceptor site. There are occasionally other splice donors and acceptors.

Answer the rest of question F1 and take initial notes in F1i on unusual features such as non-canonical splice donors/acceptors. If there are no non-canonical splice donors and acceptors, then note that instead.

We will return to the F1i box again in our verification steps in a little bit.

Interlude: You should now have a gene model for your first unique coding isoform! The next step is to verify coordinates using splice junctions (i.e. the exon junctions you examined in the UEG lab). Remember that splice junction data look like a bowtie, where the black rectangles indicate the ends and beginnings of exons, and the line connecting the rectangles indicate the span of an intron. Thus, the middles of exons are usually not shown. We will use these to make sure our coordinates fall on splice junctions for RNA-Seq data. The figure below summarizes how RNA-Seq and splice junctions help to establish the exon-intron boundaries.

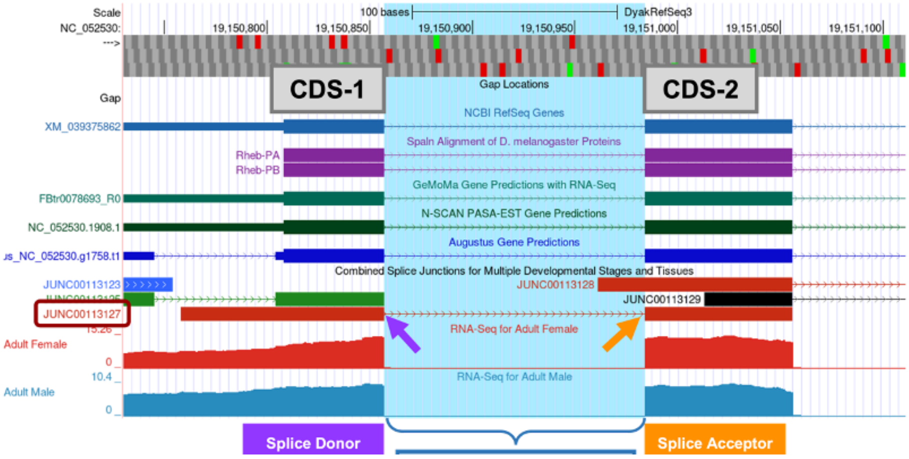
Since RNA-Seq reads are derived primarily from processed mRNAs (where the introns have been removed), the **subset of RNA-Seq reads that span multiple exons (i.e., spliced RNA-Seq reads)** can provide us with additional evidence for a splice junction. When we map a spliced RNA-Seq read against the scaffold, part of the spliced read will map to one CDS while the rest of the read will map to another CDS. The region between these two alignment blocks corresponds to the intron. Splice site prediction tools such as TopHat or regtools can recognize this distinct mapping pattern of spliced RNA-Seq reads in order to infer the possible locations of the splice junction.

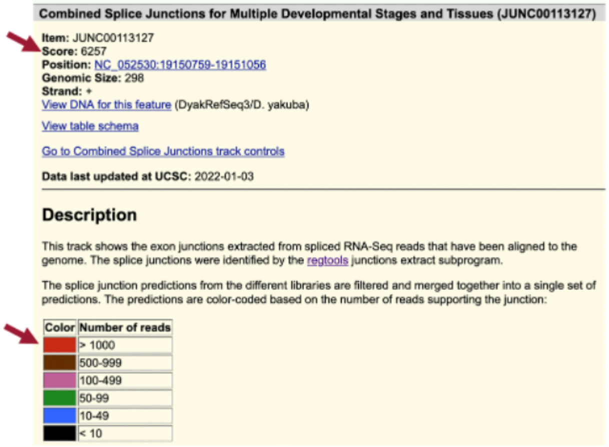


A) "Processed mRNA" that had its introns spliced out and, to stabilize the molecule for translation into a protein, had a 5' cap (acts like a flashing neon sign for the ribosome to recognize the mRNA and begin translating it into a protein) and Poly-A tail (long chain of Adenine nucleotides) added to the 5' and 3' end, respectively. **B)** Processed mRNA is fragmented into smaller pieces called "reads" which are then sequenced (i.e., determine order of bases). **C)** Map the reads against a genomic scaffold looking for overlaps in the coding sequences. **D)** A computer program analyzes the mapping data and infers the possible locations of the splice junctions.

Instructions	Comments
11. Let's verify coordinates for the first intron in your isoform from Table 3. In your Genome Browser, under the "RNA Seq Tracks," change the "Combined Splice Junctions" track to "pack.", and click 'refresh'.	
12. In the "enter position or search term" text box, input the range spanning your first intron. From Table 4, use the region spanning your first intron (the position after CDS-1 to the position right before CDS-2). Click 'go'.	For *D. yakuba* and Rheb-PA, we found that CDS-1 ends at 19,150,857. Thus, Intron-1 should begin at 19,150,858; furthermore, we found that CDS-2 begins at 19,150,985. Thus, Intron-1 should end at 19,150,984.
13. Zoom out 3x to view the larger region and examine the splice junctions. ● If there is only one splice junction, check if it aligns with your gene model	**The splice junction JUNC00113127 (red) connects CDS-1 with CDS-2.**

- If there are more than one splice junctions, click on the junction name. On the new page, examine the ‘Score’ field.
- For the multiple splice junctions, use the number in the ‘Score’ field to determine which splice junction has more support. If both have similar levels of support, reach out to the TA or instructor for more clarity.
- Click on the back button of the web browser to return to the Genome Browser image.
- A more in-depth description of dealing with multiple splice junctions can be found in Appendix 8.





Combined Splice Junctions for Multiple Developmental Stages and Tissues (JUNC00113127)

Item: JUNC00113127
Score: 6257
Position: NC_052530:19150759-19151056
Genomic Size: 298
Strand: +
View DNA for this feature (DyakRefSeq3/D. yakuba)

View table schema

Go to Combined Splice Junctions track controls

Data last updated at UCSC: 2022-01-03

Description

This track shows the exon junctions extracted from spliced RNA-Seq reads that have been aligned to the genome. The splice junctions were identified by the regtools junctions extract subprogram.

The splice junction predictions from the different libraries are filtered and merged together into a single set of predictions. The predictions are color-coded based on the number of reads supporting the junction:

Color	Number of reads
🟥	> 1000
🟫	500-999
🟪	100-499
🟩	50-99
🟦	10-49
⬛	< 10

The score tells us how many spliced RNA-Seq reads there are that support a predicted splice junction. Since JUNC00113127 has a score of 6257, that splice junction prediction is supported by 6,257 spliced RNA-Seq reads. Splice junction predictions are color-coded based on the number of spliced RNA-Seq reads that support the junction (i.e., their scores). Based on the color-coded table in the “Description” section, JUNC00113127 will be red in the Genome Browser image since greater than 1,000 spliced RNA-Seq reads support the feature.

If there are any discrepancies (no splice junction with high score where you’d expect one, multiple splice junctions of high score causing ambiguity), note this information in question F1i. If not, note that there are no splice junctions causing ambiguity.

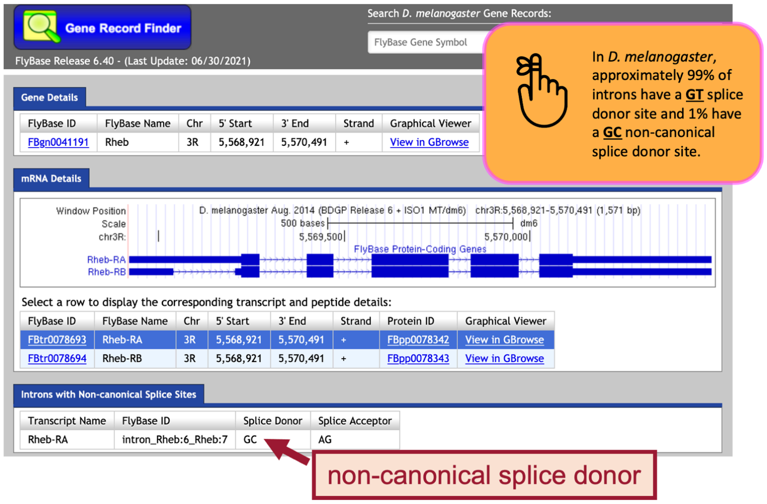
14. Next, zoom in and verify the bases for the splice donor, followed by the splice acceptor (see step 15 below for non-canonical Splice Sites). At the same time, verify whether your RNA-Seq data and gene model predictions and NCBI RefSeq/Spaln Alignments correspond as well to the start and end coordinates of your CDS’s.

If non-canonical splice sites are found or gene model predictions/alignments in the Browser do not align with your specified start and end coordinates, note that in question F1i.

15. If they are not our typical GT and AG, we must check whether *D. melanogaster* also has a non-canonical Splice Site. If you find a non-canonical Splice Site, do the following:

The “Introns with Non-canonical Splice Sites” panel shows that the intron with the FlyBase ID

- Return to the Gene Record Finder (http://tiny.cc/generecordfinder) record for the target gene in *D. melanogaster*.
- Check if there is a panel under ‘mRNA Details’ titled “Introns with Non-canonical Splice Sites”. If it exists, check for the intron you are examining, and whether there is a non-canonical splice acceptor or donor matching the one you observed in your target species.
- To determine if the ID corresponds to your intron, click on the Genome Browser image in the “mRNA Details” panel of the Gene Record Finder and compare the “FlyBase Transcribed Exons” and “FlyBase Coding Exons” tracks in the resulting Genome Browser to determine which CDS correspond to the region listed under “Introns with Non-canonical Splice Sites.”


Gene Record Finder
Search *D. melanogaster* Gene Records:
FlyBase Gene Symbol
Window Position
Scale
chr3R:
D. melanogaster Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) chr3R:5,568,921-5,570,491 (1,571 bp)
500 bases
dm6
5,569,500
5,570,000
FlyBase Protein-Coding Genes
Rheb-RA
Rheb-RB

FlyBase Release 6.40 - (Last Update: 06/30/2021)

In *D. melanogaster*, approximately 99% of introns have a **GT** splice donor site and 1% have a **GC** non-canonical splice donor site.

Gene Details

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Graphical Viewer
FBgn0041191	Rheb	3R	5,568,921	5,570,491	+	View in GBrowse

mRNA Details

Select a row to display the corresponding transcript and peptide details:

FlyBase ID	FlyBase Name	Chr	5' Start	3' End	Strand	Protein ID	Graphical Viewer
FBtr0078693	Rheb-RA	3R	5,568,921	5,570,491	+	FBpp0078342	View in GBrowse
FBtr0078694	Rheb-RB	3R	5,568,921	5,570,491	+	FBpp0078343	View in GBrowse

Introns with Non-canonical Splice Sites

Transcript Name	FlyBase ID	Splice Donor	Splice Acceptor
Rheb-RA	intron_Rheb:6_Rheb:7	GC	AG

non-canonical splice donor

“intron_Rheb:6_Rheb:7” has a GC splice donor in *D. melanogaster*. The FlyBase ID for an intron begins with the prefix “intron_”, followed by the names of the two transcribed exons that flank the intron. Hence, the FlyBase ID “intron_Rheb:6_Rheb:7” indicates that the splice donor site between the transcribed exons Rheb:6 and Rheb:7 has the non-canonical sequence GC.

Comparison of the “FlyBase Transcribed Exons” and “FlyBase Coding Exons” tracks in the Genome Browser for *D. melanogaster* (click gene model image in Gene Record Finder to view) shows that the coding exon CDS-4 (Rheb:4_9847_1) overlaps with the transcribed exon Rheb:6, and coding exon CDS-5 (Rheb:5_9847_0) overlaps with the transcribed exon Rheb:7.

16. When you have confirmed all splice junctions and all splice donors/acceptors, you have completed the gene model for that isoform. Make sure you have adjusted any positions in Table 5 as needed to align with the evidence.
DISCUSS: Make sure you and your group agree on the evidence and any unusual features. Note that if there are no discrepancies, you should note that accordingly for each CDS. You want enough notes here so you feel confident describing this region in your paper or a final report, so someone who hasn’t gone through the protocol will feel confident in your final conclusions. No discrepancies for the entire gene annotation would presume there are canonical splice donors/acceptors, and that splice junctions, RNA-Seq coverage, and Spaln/gene model predictions all align with your chosen coordinates. Confirm with your instructor or TA you have completed Table 5 and ask any remaining questions.

17. Repeat the above process in Parts E and F for any remaining isoforms with unique CDS (you should have at least one more, to fill out in Table 6, using information from Table 4). *Note that those CDS that are the same for the second unique isoform should keep the same phasing and reading frame.*

Answer remaining questions for the second isoform.

Congrats! You’ve now done a full manual annotation of your gene. Next week, we will generate figures and alignments comparing your gene model to the one for *D. melanogaster*.

DISCUSS and CHECK-IN: Let me know when you have coordinates for both isoforms. I will give you the first page of the Pathways G protocol, for you to input your numbers, so we can check if your coordinates follow the rules for gene structure.