

Using *Pathways* Appendix

Our main question for the Pathways Project Parts A-C is where in our target species is the orthologous gene to the one found in *D. melanogaster*. Ultimately we are looking for specific positions corresponding to the coding sequences, so our focus is on comparing the amino acid sequence to the genome of our target species. After determining information about our gene of interest and its surrounding regions in *D. melanogaster* (Part A), we then want to use the protein sequence in *D. melanogaster* to search for corresponding nucleotide sequences in your target species (Part B). Thus, the results of Part B should provide us with a good hypothesis on a potential region for the orthologous gene in your target species. Lastly, after determining an accession and region, we must then find additional evidence to support our hypothesis. A major piece of evidence that can support the hypothesis is finding that the neighboring genes around the region of interest are the same neighboring genes found around the target gene in *D. melanogaster* (Part C). If true, then we have **synteny**, or two or more genes lying on the same chromosome. The presence of synteny is strong evidence that the location contains the orthologous gene to the target gene in *D. melanogaster*.

The appendices below walk us through much of the important logic needed to arrive at a conclusion on your target gene for your target species, using the target gene *Rheb* for the target species *D. yakuba*.

Appendix 1. Asking where in our target species is the ortholog of our target gene

Appendix 1 describes how we narrow down to a particular hit from a *BLAST* search to start forming our hypothesis on the location of the orthologous gene in our target species.

When performing a search, BLAST may return any number of matches (often referred to as “hits”) for regions of local similarity between our query sequence and database; however, each hit is not necessarily statistically significant. BLAST provides statistical scores to help us determine which **alignments** between the two sequences are statistically significant and which are **spurious** (i.e., likely occurred by chance alone and, therefore, are not evidence of real biological **conservation**). If BLAST returns multiple good hits (i.e., more than one match with a low **E-value** and a high sequence **identity**), we will need to investigate them all further to determine the most likely ortholog.

Our *tblastn* search found five regions within the translated *D. yakuba* genome that show similarities with the protein sequence of *Rheb* in *D. melanogaster* (Figure 1); however, only one of these is a good hit (*Drosophila yakuba* strain Tai18E2 chromosome 3R, Prin_Dyak_Tai18E2_2.1, whole genome shotgun sequence; sequence identity: 97.14% and E-value: 2e-78). The second hit has a much higher E-value (7e-37) and much lower **percent identity** (43.75%), and this pattern of increasingly lower quality matches continues through the

other three matches. Therefore, we will continue our analysis based on the **hypothesis** that the **putative ortholog** of Rheb-PA in *D. yakuba* is in the scaffold of chromosome 3R.

	Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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
<input checked="" type="checkbox"/>	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

Figure 1 The *tblastn* search of the *D. melanogaster* protein Rheb-PA (query) against the translated *D. yakuba* genome assembly (database) found five regions of similarity. The best match (black rectangle) is located on the “chromosome 3R” scaffold (pink arrow) (accession number: NC_052530; red arrow) of *D. yakuba*.

Appendix 2. Accession Numbers

In Appendix 2, the goal is to better understand accession numbers. Notice that each of the genome regions of our five hits has a unique **accession number**. The accession number for the chromosome 3R scaffold in *D. yakuba* is NC_052530 (Figure 1). Read the content of the red box below for more on accession numbers.

Each sequence record in the NCBI database has a sequence version number consisting of an accession number followed by a dot and a version suffix (e.g., NC_052530.2). The accession number is used by NCBI to identify the sequence record, and the version suffix is used to identify revisions to the sequence record. By convention, an accession number without the version suffix refers to the latest version of the sequence record. **Student annotators should only use the accession number and ignore the version number when navigating in the Genome Browser.** Student annotators will use the accession number (e.g., NC_052530) to navigate to a genomic region in the Genome Browser. For example, entering “NC_052530:100-200” in the search terms of the *D. yakuba* Genome Browser would show us coordinates 100-200 of the chromosome 3R scaffold (NC_052530) in the Genome Browser image.

NC_052530.2

Accession Number

Version

The accession number for the chromosome 3R scaffold of *D. yakuba* is NC_052530.

For more on sequence IDs, you can click the following link (<https://www.ncbi.nlm.nih.gov/genbank/sequenceids/>).

If there seems to be multiple possible accessions giving similar results in *BLAST*, visit the Genome Browser and click on your target species. Then click ‘view sequences’ on the right hand side to see a list of all accessions associated with your target species (Figure 2). Check to make sure that your accession is found in the list of sequences for your target species.

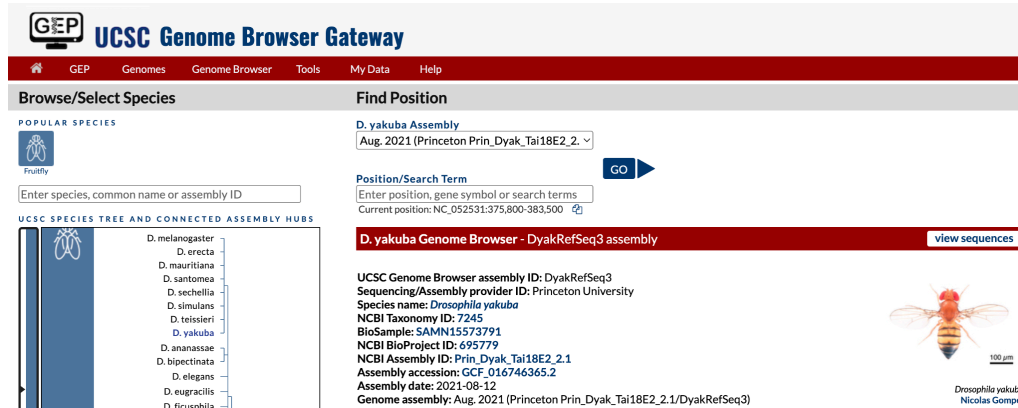


Figure 2 An example of a Genome Browser homepage for the species *D. yakuba*.

Appendix 3. Making a hypothesis about the specific orthologous region on a given scaffold

Our goal in Part B is to figure out the region of DNA in our target species that corresponds to the target gene, i.e. the orthologous region. Similarity to the *D. melanogaster* version of the gene is a strong piece of evidence for locating the orthologous region. The accession number from the *tblastn* search helped us to identify the scaffold (i.e. part of a chromosome) in which we think the orthologous region is located. We next want to narrow in on the set of positions corresponding to the orthologous region on this scaffold. The alignment section of the *BLAST* results for our chosen hit will contain the information on the potential orthologous region.

The different ranges associated with your chosen hit under the ‘Alignments’ tab potentially indicate different coding sequences corresponding to our target gene. However, it is possible that some ranges (i.e., alignment matches or hits) do not correspond to our ortholog; therefore, we need to examine each match for each range more closely. To determine the orthologous region, we will look at all alignments on the scaffold (i.e., the set of ranges for the entire accession) and consider the following:

1. Which ranges have low E-values and high percent identity?
2. Are these ranges collinear, i.e. are they focused on a set of positions that are adjacent to each other?
3. Does the range cover the entire *D. melanogaster* amino acid (i.e. query start to query end covers the length of the amino acid)?
4. Do all ranges appear on the same strand? (The subject frame indicates the reading frame used and the strand, + or -, the reading frame sits on. For one transcript, all corresponding regions must be on the same strand).

Assess the above questions using Figure 3 below, which shows a table for a search of the Rheb-PA isoform in a *D. yakuba* assembly, where the best match was on scaffold chr3R, accession NC_052530. You will ultimately set up a similar table for your target gene, and decide the best range - you will use this to construct your first hypothesis in the Pathways Project, answering the question of where in your target species is a putative ortholog located.

Range	<i>D. melanogaster</i>		Target Species		E-Value	Identities (%)	Subject Frame
	Query Start	Query End	Subject Start	Subject End			
1	55	163	7,623,630	7,623,953	2e-10	36	+3
2	59	128	7,773,176	7,772,970	1e-04	37	-2
3	8	131	11,133,871	11,133,431	9e-10	30	-3
4	54	130	11,148,243	11,147,992	3e-11	40	-1
5	6	44	11,148,568	11,148,431	0.002	46	-3
6	18	108	11,418,759	11,419,094	5e-06	28	+3
7	1	20	19,150,809	19,150,868	2e-78	90	+3
8	16	45	19,150,981	19,151,070	2e-78	83	+1
9	40	109	19,151,150	19,151,359	2e-78	97	+2
10	111	153	19,151,422	19,151,550	2e-78	93	+1
11	153	182	19,151,610	19,151,699	2e-78	93	+3
12	53	173	28,532,182	28,531,790	1e-05	30	-3

**best
collinear
set of
alignments
to Rheb-PA**




Figure 3 Summary of the *tblastn* search results for the 12 matches to Rheb-PA within the NC_052530 scaffold of *D. yakuba*. The best collinear set of alignments to Rheb-PA is located at 19,150,809- 19,151,699. Rheb-PA in *D. melanogaster* is 182 amino acids long and the above collinear set of alignments covers all 182 amino acids (arrows).

A description of the analysis for the region found for *D. yakuba* is listed here:

Since the NC_052530 scaffold in *D. yakuba* is 30,730,773 base pairs (bp) long, it is possible we will have some ranges (i.e., alignment matches or hits) that don't correspond to our ortholog; therefore, we need to examine each match more closely.

Remember that we are looking for matches with low E-values and high sequence identities, and there are five matches (ranges 7 – 11) that fit these criteria (E-value of 2e-78 and sequence identities that range from 83% to 97%). These five alignment matches are also collinear and appear on the same strand of DNA (+ Frame, Figure 3).

The best collinear set of alignments to Rheb-PA is located at 19,150,809 – 19,151,699 on the NC_052530 scaffold of the *D. yakuba* genome assembly and the five alignment matches cover all 182 amino acids of Rheb-PA (Figure 3, arrow). Therefore, we will continue our analysis based on the hypothesis that the putative ortholog of Rheb-PA from *D. melanogaster* is located at approximately 19,150,809 – 19,151,699 on the NC_052530 scaffold of the *D. yakuba* genome assembly.

Appendix 4. Investigate the other *tblastn* alignments to *D. yakuba* chr3R


This section highlights some of the other regions that we did not choose to include in our potential range from Figure 3, to show qualities that can sometimes be detected to explain why they align but should not be considered. We may or may not be using these steps to explain regions not included, but it is useful to read about to know what are potential flags to look for.

In Range 1, the percent identity is low (36%) and the subject positions (7,623-630-7,623,953) are very far from our chosen block of 19,150,809-19,151,699, but the same frame (+) is found (Figure 3). A closer look at the FlyBase Polypeptide Report for Rheb-PA in the 'Protein Domains' section shows that there is a conserved Small_GTPase domain at residues 8-164 aa of the protein (Figure 4). In Range 1 (Figure 5), it can be observed that the region at 5,727,469-5,727,792 aligned with amino acid residues 55-163 of the Rheb-PA protein when it is translated on the positive strand in the first reading frame (+1). This corresponds to this conserved domain, and suggests this might be an amino acid sequence that occurs in more than one protein from more than one gene.

Examination of the *tblastn* alignment blocks also show three cases of ranges where the alignments contain in-frame stop codons (i.e., at Range 3: 11,133,871-11,133,431; Range 6: 11,418,759-11,419,094; and Range 12: 28,532,182-28,531,790). Figure 6 shows the first set, in Range 3. The in-frame stop codons can likely be attributed to homologous overextension of the alignments into the introns. These types of ranges are likely ones we would not include in our hypothesis.

General Information			
Symbol	Dmel\Rheb-PA	Species	<i>D. melanogaster</i>
Annotation Symbol	CG1081-PA	FlyBase ID	FBpp0078342
Associated gene	Dmel\Rheb		
Length (aa)	182	Theoretical pI	5.32
Predicted MW (kDa)	20.7		

...

Protein Domains			
Protein Domains (via Pfam)			
Isoform displayed: Rheb-PA			
			
Pfam protein domains			
InterPro name	classification	start	end
Small_GTPase (Small_GTPase)	Family	8	164

Small_GTPase domain: 8–164aa

Figure 4. The “Protein Domains” section of the FlyBase Polypeptide Report for Rheb-PA ([FBpp0078342](#)). The number of amino acids in the protein and the positioning for the Small_GTPase domain are highlighted.

Range 1: 7623630 to 7623953 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
60.5 bits(145)	2e-10	Compositional matrix adjust.	39/109(36%)	57/109(52%)	1/109(0%)	+3

Features: [ras-related protein rab-11a](#)
[ras-related protein rab-11a](#)

Query	55	VKLIDTAGQDEYSIFPVQYSMDYHGVLVYSITSQKSFEVVKIYEKLLDVMGKKYVPV	114
		++ DTAGQ+ Y Y G +LVY I ++E V+ +L D + V ++	
Sbjct	7623630	AQIWDTAGQERYRAITSAYYRGAVGALLVYDI AKHLTYENVERWLREL RDHADQNIV-IM	7623806
Query	115	LVGNKIDLHQERTVSTEEGKLAESWRAAFLETS AKQNESVGDIFHQLL	163
		LVGNK DL R+V T+E K AE +F+ETSA + +V F +L	
Sbjct	7623807	LVGNKSDLRHLSVPTDEAKLFAERNGLSFIETSA LDSTNVETAFQNIL	7623953

Figure 5. The Range 1 alignment, where examination of the query coordinates of the additional *tblastn* alignments to chr3R shows that they overlap with the Small_GTPase conserved domain. For example, the region at NC_052530: 7,623,630-7,623,953 aligned with amino acid residues 55-163 of the Rheb-PA protein when it is translated on the positive strand in the first reading frame (+3).

Range 3: 9373062 to 9373502 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#) [First Match](#)

Score	Expect	Method	Identities	Positives	Gaps	Frame
58.9 bits(141)	1e-09	Compositional matrix adjust.	45/148(30%)	75/148(50%)	25/148(16%)	-2

Query	8	IAMMGYSVGSLSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDTAGQDEY-	66
		I ++G V GK+ + I + + +F + Y PT+ + V +DY + L DTAGQ++Y	
Sbjct	9373502	ITVVGDMGVMGKTCMLITYTQNEFPPEEYVPTVFDNHACNIAVDDRDNLTLDWTAGQEDYE	9373323
Query	67	SIFPVQY-----SMDYHGVLVYSITSQKSFEVVKIYEKLLDVM	106
		+ P+ Y + + +L YSI+S+ SFE VK + +	
Sbjct	9373322	RLRPLSYPNVSTIALQRDDVIPKLESIA P*TNCFLLCYSISSRTSFENVKSKWWEIRHF	9373143
Query	107	GKKYVPVVLVGNKIDL---HQERTVSTE	131
		+VPVVLVG K+DL + E+ V+T+	
Sbjct	9373142	SA-HVPVVLVGTKLDLRIPNSEKFTTQ	9373062

In-frame stop codon

Figure 6. Example of a stop codon found in Range 3, suggesting it is not a good alignment to include.

Appendix 5. Tracks you will use to examine your target gene in your target species on the Genome Browser (Pertains to Part C and beyond of the Pathways Project)

In Part C, you will be examining data contained in the BLAT Alignment tracks, which are unknown transcript data from your target species that have been computationally mapped against the genomic data from your target species. The BLAT Alignment is our starting point for annotating the gene model in later steps of the Pathway Project, but here we will use the transcript data stored in these tracks to determine what proteins they best match against in *D. melanogaster*. We will ultimately do this for our target gene, two neighboring genes upstream, and two neighboring genes downstream. It is useful to start understanding the different tracks, so they are briefly described below.

In the Genome Browser image with default settings (Figure 7), you are likely to see the following tracks:

- NCBI RefSeq Genes
- Spaln Alignment of *D. melanogaster* Proteins
- Gene Prediction Tracks:
 - GeMoMa Gene Predictions with RNA-Seq
 - N-SCAN PASA-EST Gene Predictions

- o Augustus Gene Predictions
- RNA-Seq for Adult Female
- RNA-Seq for Adult Male

The alignments and prediction tracks use a combination of previously published transcripts and proteins and software to automate annotation through algorithms that use 'rules' for gene structure to predict the properties of different genic regions. While useful because it allows fast annotation of high volumes of genes, we cannot always trust the accuracy of these gene models, particularly for *Drosophila* species that are more distantly related to *D. melanogaster*.

To briefly describe each track, the NCBI RefSeq track shows alignments of transcripts (from processed mRNA) for your target species taken from the NCBI RefSeq database and aligned against the genome assembly of your target species. This alignment is done by a computer, trying to determine the most sensible gene model following a series of rules on gene structure and how the transcripts aligned. It is the most reliable track outside of a custom annotation, but since it is automated by a software, it can have errors!

The Spaln alignment takes known proteins from *D. melanogaster* and compares them against the genome assembly of your target species (similar to *blastx*). GeMoMa and Augustus are similar to the Spaln alignment but use slightly different algorithms. N-Scan compares the genomic sequence against other genomes to look for sequences indicative of a gene and then makes a de novo gene prediction - that is, it does not use any transcript data. The remaining tracks are the RNA-seq data we examined in the UEG modules.

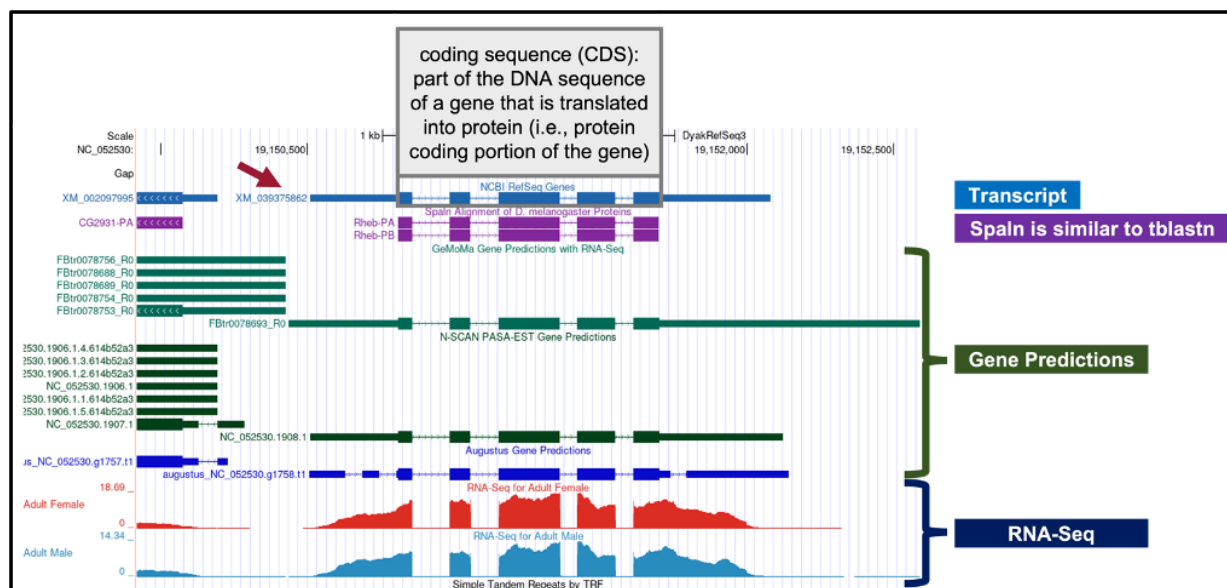


Figure 7 Genome Browser image of NC_052530:19,149,918-19,152,590 in *D. yakuba* (default tracks).

Looking at the NCBI RefSeq Genes track, we notice that the alignment for the *coding sequences* of the *D. yakuba* RefSeq transcript XM_039375862 (Figure 7, red arrow) against the NC_052530 scaffold of *D. yakuba* line up with the Spaln alignment to the *D. melanogaster* proteins Rheb-PA

and Rheb-PB. Furthermore, the coding portions of the five alignment blocks are in congruence with the placements of the five coding exons (CDS's) predicted by GeMoMa, N-SCAN, and Augustus.

According to the RefSeq transcript XM_039375862 (Figure 7, red arrow), the putative (probable) ortholog of Rheb-PA is located at NC_052530:19,150,510-19,152,080 in *D. yakuba* and the region spanning NC_052530:19,150,809-19,151,702, within the putative ortholog, corresponds to the alignment to the *coding* region of the RefSeq transcript.



We cannot always trust that what we see in the Genome Browser is accurate, particularly for *Drosophila* species that are *more distantly related* to *D. melanogaster*. The gene prediction tracks, like their name implies, are predictions; thus, *your role, as a researcher in the Pathways Project*, is to help scientists studying these genes be confident in the specific model for the gene. *Your brain is far superior to a computer algorithm* in weighing conflicting evidence, thus your model will be more reliable than what a computer can produce alone. For example, in your own project, there might be a situation where a gene predictor(s) doesn't show a gene in an area that has an alignment to *D. melanogaster* proteins (or vice versa); therefore, you'd need to investigate that further.

Appendix 6. Data and Explanation for examining the genomic neighborhood around the putative ortholog for *Rheb* in *D. yakuba* and *D. pseudoobscura*

If the genomic neighborhood looks similar between *Rheb* in *D. melanogaster* and the putative ortholog in *D. yakuba* in our Part C analysis, we can be confident we have found the true ortholog. However, if any of the information is inconsistent with this being a syntenic region, we should also inspect our other hits in the *tblastn* search (Part B) to see if a different genomic region is a better match overall. If you think this may be the case for your target gene/species, check with your instructor.

Examination of the genomic regions surrounding the *Rheb* gene in *D. melanogaster* (Figure 8; top) and the putative *Rheb* ortholog in *D. yakuba* (Figure 8; bottom) shows that the relative gene order (i.e., *CG12746*, *CG2931*, *Rheb*, *CRMP*, and *CG2926*) and orientations (+, -, +, +, -) are the same in the two species. Hence the synteny analysis supports the assignment of the *D. yakuba* feature at NC_052530:19,150,510-19,152,080 as an ortholog of *Rheb*.

- **Note:** If your target species' assembly happened to have been numbered from the opposite end of the relevant scaffold, the orientation (+ or - strand) of the orthologs could be the opposite (i.e., -, +, -, -, +) of what you see in *D. melanogaster* but still be syntenic. **This finding of opposite relative strands is NOT a discrepancy you need to discuss in your results unless the relative order does not seem to be followed.**

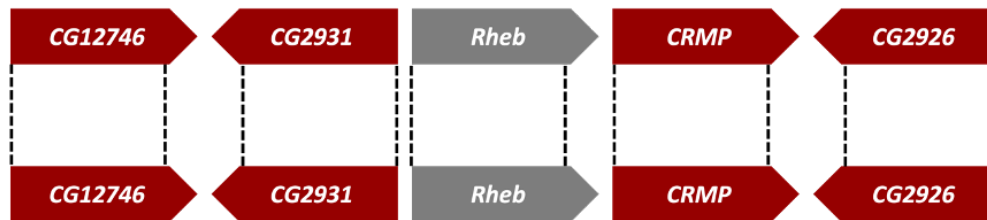
D. melanogaster* (chr3R scaffold)**D. yakuba* (NC_052530 scaffold)**

Figure 8 Comparison of the relative order and orientation of the genomic neighborhoods of *Rheb* in *D. melanogaster* (top) and *D. yakuba* (bottom).

The *Rheb* gene for *D. yakuba* is not very different from the *Rheb* gene in *D. melanogaster*. For a more distantly related species like *D. pseudoobscura*, our argument of synteny is more difficult. The schematic below (Figure 9) shows the final results for *Rheb* in *D. pseudoobscura*. There is synteny for *D. pseudoobscura*, but the results are less clear. Which genes do NOT show results supporting a hypothesis where the chosen region in *D. pseudoobscura* is orthologous to *Rheb* in *D. melanogaster*, because synteny is not observed?

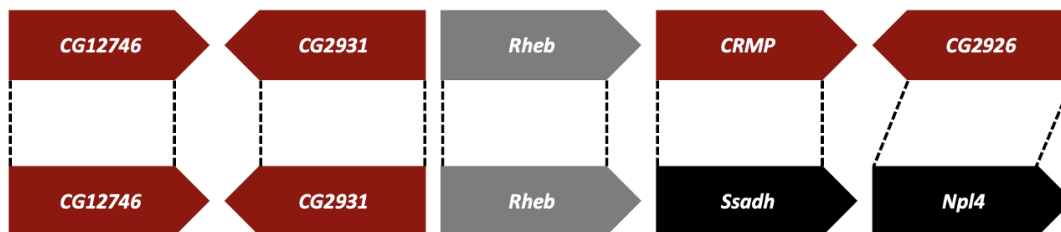
D. melanogaster* (chr3R scaffold)**D. pseudoobscura* (chromosome 2 (CM000070.3) scaffold) *reverse**

Figure 9 Comparison of the relative order and orientation of the genomic neighborhoods of *Rheb* in *D. melanogaster* (top) and *D. pseudoobscura* (bottom). Note here that *D. pseudoobscura* is reversed, that is, it was located on the - strand, but is re-drawn as if it were on the + strand.

Appendix 7. Determining the reading frames for exons and the corresponding phases at the splice donor and splice acceptor sites to manually annotate an isoform

The main part of manual gene annotation is keeping track of the reading frame of each CDS, which involves determining the phasing at the 3' end of an exon (by the splice donor site) and the 5' end of the next exon (by the splice acceptor site) (Figure 10). After accounting for these, we can then establish the coordinates spanning the exon and check that the open reading frames do not contain any stop codons (except at the end of the last CDS). Below is a walkthrough of the findings/rationale for manual annotation of the *Rheb-PA* isoform for *D. yakuba*.

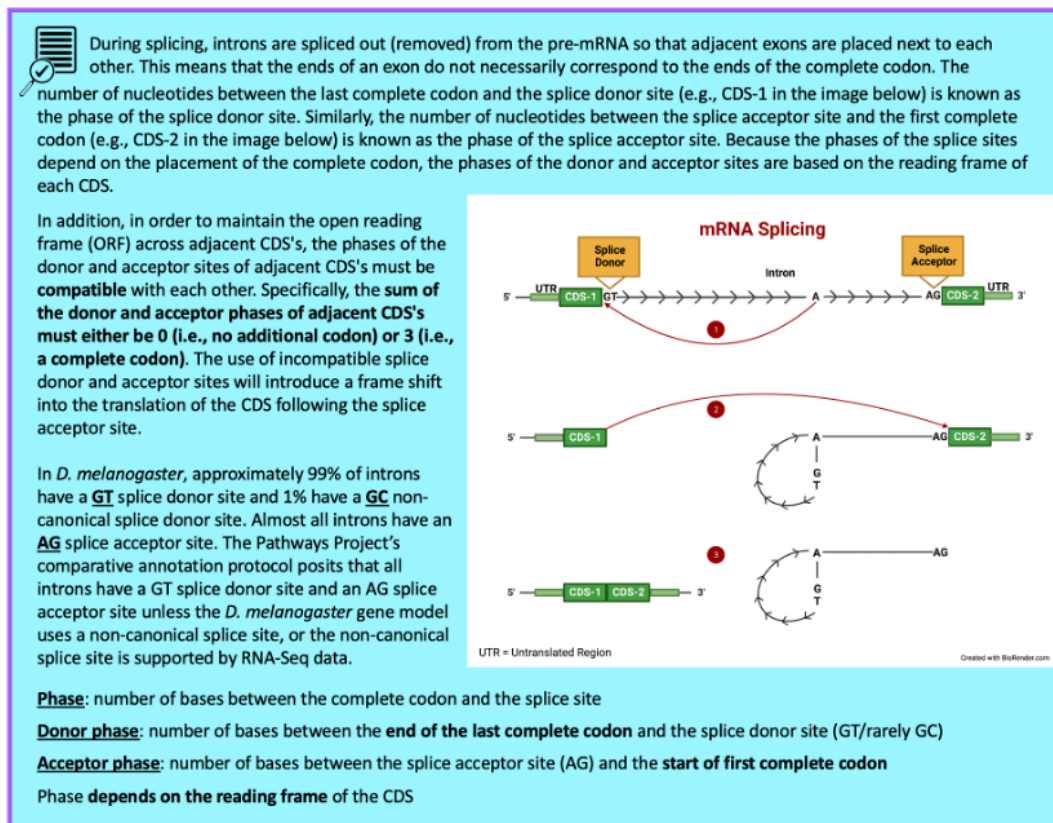


Figure 10 Box describing the role of phasing in exons related to splice donor and splice acceptor sites and diagram depicting splicing

Because the *tblastn* alignment for CDS-1 of *Rheb* terminates at 19,150,856, we expect to find the splice donor site for CDS-1 at around position 19,150,856. The GT splice donor site closest to 19,150,856 (Figure 11A, green box) is located at 19,150,858-19,150,859 (Figure 11A, red box) which is supported by multiple lines of evidence—NCBI RefSeq Genes, Spaln alignment of *D. melanogaster* proteins, and the GeMoMa, N-SCAN, and Augustus gene predictions, and the RNA-Seq read coverage from samples of adult females and adult males.

We visually inspected the region surrounding the approximate end of CDS-1 and placed the splice donor site at 19,150,858-19,150,859, after which we confirmed Frame +3 of CDS-1 has an ORF. Now we need to determine the **phase** of the splice donor site for CDS-1. We already know from the start codon for *D. yakuba* and the Rheb-PA isoform that the first CDS is in Frame +3. There are three possible phases for the splice donor site (0, 1, or 2), which depend on the reading frame (Figure 11B). The splice donor site (at 19,150,858-19,150,859) is a GT, a typical splice donor pattern. We see the last complete codon (i.e., containing three nucleotides) of CDS-1 before the splice donor site codes for the amino acid Tryptophan in Frame +1, Arginine in Frame +2, and Valine in Frame +3.

The splice donor site at 19,150,858-19,150,859 is in phase 0 relative to Frame +1 because CDS-1 ends in a complete codon and thus has no extra nucleotides. In contrast, Frame +2 and Frame +3 don't end in complete codons so they each have extra nucleotides. The splice donor site is in phase 2 relative to Frame +2 because there are 2 extra nucleotides (GG) after the last complete codon. The splice donor site is in phase 1 relative to Frame +3 since there is 1 extra nucleotide (G) after the last complete codon (Figure 11B).

We previously determined CDS-1 is translated in Frame +3 when we determined the start codon; therefore, the last complete codon of CDS-1 (GTG which codes for Valine (V)) is located at 19,150,854-19,150,856 and there is one extra nucleotide (G at 19,150,857) between the last complete codon and the splice donor site. Hence, the CDS-1 splice donor site is in phase 1.

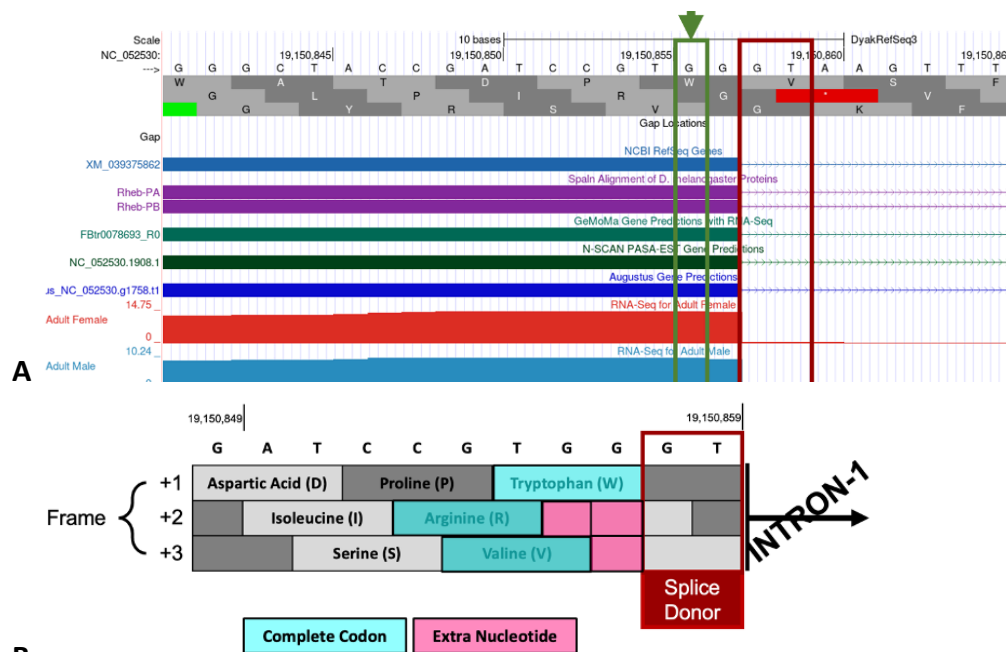


Figure 11 NC_052530:19,150,849-19,150,859. (A) The splice donor site (GT) at 19,150,858-19,150,859 (red box) is supported by multiple lines of evidence and is near the *approximate* end coordinate of CDS-1—at 19,150,856—as determined by *tblastn*. Since the splice donor site (GT) is at 19,150,858-19,150,859 (red box), the last coordinate of CDS-1 is 19,150,857. (B) The CDS-1 splice donor site has three possible phases (0, 1, or 2), which depend on the reading frame. Frame +1 ends in a complete codon, Frame +2 ends with two extra nucleotides, and Frame +3 ends with one extra nucleotide.

Our *tblastn* alignment placed CDS-2 at *approximately* 19,150,987 – 19,151,055; therefore, we expect to find the splice acceptor site for CDS-2 around position 19,150,987.

There is only one potential canonical (“standard rule”) splice acceptor site (AG) in the 30 bp region surrounding the start of the *tblastn* alignment to CDS-2 (Figure 12A, green box). The splice acceptor site is located at 19,150,983-19,150,984 (Figure 12A, red box) and is supported by the NCBI RefSeq Genes and Spaln alignment of *D. melanogaster* proteins, the GeMoMa, N-SCAN, and Augustus gene predictions, and the RNA-Seq read coverage. Thus, CDS-2 starts at 19,150,985.

Now we need to determine the frame in which CDS-2 is translated. Each frame can have extra nucleotides (0, 1, or 2) between the beginning of CDS-2 at 19,150,985 and the first complete codon of each frame (Figure 12B).

Each frame can have extra nucleotides (0, 1, or 2) between the beginning of CDS-2 at 19,150,985 and the first complete codon of each frame (Figure 12B).

Frame +1 has two extra nucleotides (GC) before the first complete codon (Lysine; K), where splice acceptor site is in phase 2 relative to Frame +1. Frame +2 begins with a complete codon (Alanine; A) so it has zero extra nucleotides, where splice acceptor site is in phase 0 relative to Frame +2. Frame +3 has one extra nucleotide (G) before the first complete codon (Glutamine; Q), where splice acceptor site is in phase 1 relative to Frame +3.

Since the CDS-1 splice donor site at 19,150,858 – 19,150,859 is in phase 1 relative to Frame +3 (i.e., CDS-1 ends with one extra nucleotide), the CDS-2 splice acceptor site must be in phase 2 (i.e., CDS-2 must begin with two extra nucleotides) to maintain the ORF after Intron-1 has been removed. Since **CDS-1** had one extra nucleotide (G) between the last complete codon (GTG = Valine; V) and the splice donor site, **CDS-2** must start with **two** extra nucleotides to join the extra one from CDS-1 because we **need 3 nucleotides to make a complete codon**.

Since CDS-2 is in Frame +1 and the first complete codon (AAA codes for K) is located at 19,150,987- 19,150,989, there are **two nucleotides** (GC at 19,150,985- 19,150,986) between the potential splice acceptor site and the first complete codon. Hence, the CDS-2 splice acceptor site is in **phase 2**.

The extra nucleotides near the splice sites (i.e., G + GC) will form an additional amino acid (Glycine/G) after **splicing** (Figure 13). Collectively, our analysis suggests that CDS-1 ends at 19,150,857 with a phase 1 splice donor site and CDS-2 begins at 19,150,984 with a phase 2 splice acceptor site.

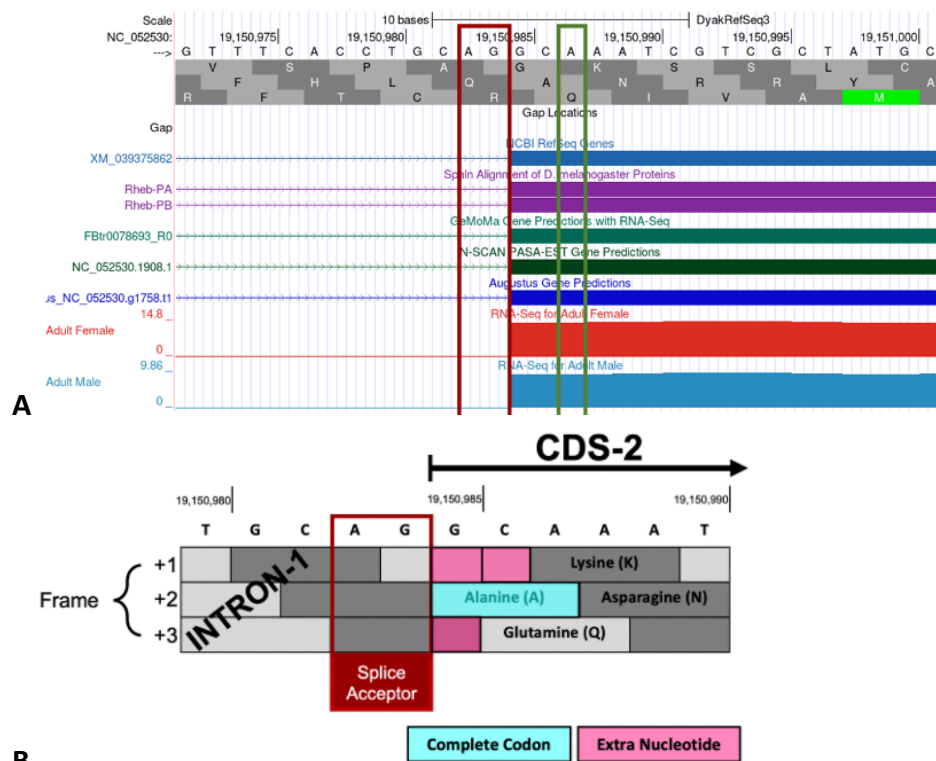


Figure 12 Example showing the region around the beginning of CDS-2. Our *tblastn* search placed the start of CDS-2 at *approximately* 19,150,987. The splice acceptor site (AG) is at 19,150,983-19,150,984 and CDS-2 starts at 19,150,985. (B) The CDS-2 splice acceptor site at 19,150,983-19,150,984 has three possible phases (0, 1, or 2), which depend on the reading frame.

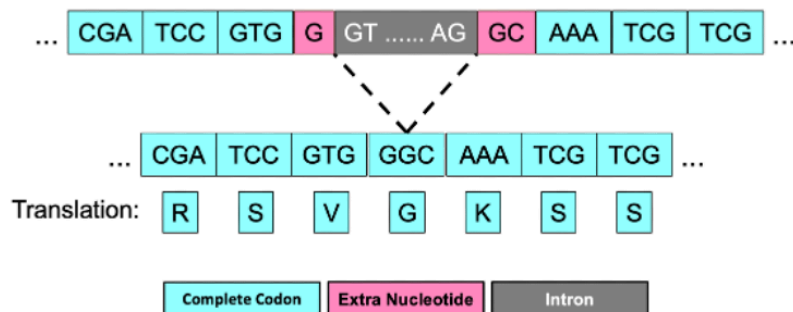


Figure 13 The phase 1 donor site (G) of CDS-1 combines with the phase 2 acceptor site (GC) of CDS-2 to form the codon GGC, which codes for a Glycine (G).

The same annotation strategy can be used to determine the phases for the remaining splice donor and splice acceptor sites between CDS-2 and CDS-3, CDS-3 and CDS-4, and CDS-4 and CDS-5 (Figure 14).

CDS	Frame	Splice Acceptor Phase	Splice Donor Phase
1	+3		1
2	+1	2	1
3	+2	2	2
4	+1	1	0
5	+3	0	

Figure 14 Summary of the phases of each splice donor and acceptor site in Rheb-RA.

Recall that to maintain the open reading frame (ORF) across adjacent CDS's, the phases of the donor and acceptor sites of adjacent CDS's must be compatible with each other (i.e., the sum of the donor and acceptor phases of adjacent CDS's must either be 0 or 3).

Looking at the splice sites between CDS-1 and CDS-2, we see the splice donor phase of CDS-1 is one and the splice acceptor phase of CDS-2 is two. Thus, the sum of the donor and acceptor phases for Intron-1 is three (Figure 14).

Appendix 8. Examining an intron with >1 splice junction

This example walks through Intron-2 and Intron-3 for *D. yakuba* and the Rheb-PA isoform, which has two splice junctions.

For the region surrounding Intron-2, (**NC_052530:19,151,055-19,151,156**), zoomed out by 3x, there are two splice junctions predicted in this region, JUNC00113128 and JUNC00113129 (Figure 15).

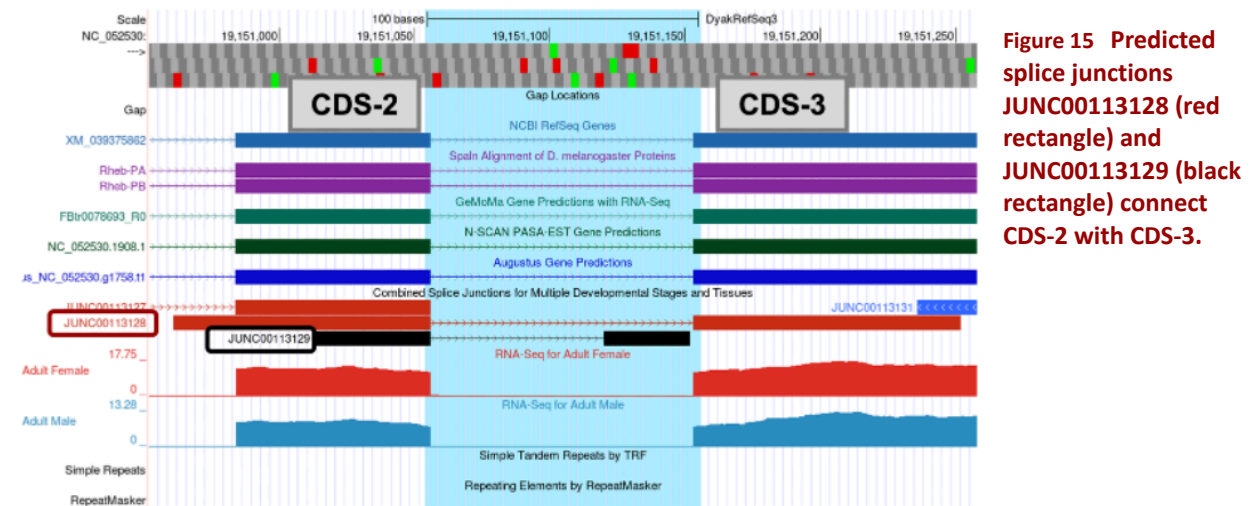


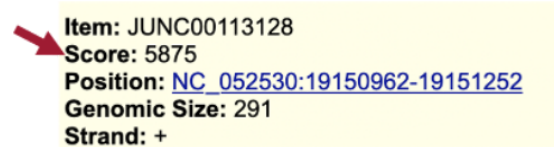
Figure 15 Predicted splice junctions JUNC00113128 (red rectangle) and JUNC00113129 (black rectangle) connect CDS-2 with CDS-3.

The *tblastn* alignment for CDS-2 ends at 19,151,055. The potential splice donor site at 19,151,057-19,151,058 for CDS-2 is supported by the splice junction predictions JUNC00113128 and JUNC00113129, the NCBI RefSeq Genes and Spaln alignment of *D. melanogaster* proteins, gene predictors GeMoMa, N-SCAN, and Augustus, and the RNA-Seq data (Figure 15). There is one nucleotide (A at 19,151,056) between the last complete codon (AAC) and the potential splice donor site. Hence, this splice donor site is in phase 1 relative to Frame +1.

The *tblastn* alignment for CDS-3 spans from 19,151,156-19,151,359 in Frame +2. The potential splice acceptor site at 19,151,152-19,151,353 for CDS-3 is supported by the splice junction prediction JUNC00113128, the NCBI RefSeq Genes and Spaln alignment of *D. melanogaster* proteins, gene predictors GeMoMa, N-SCAN, and Augustus, and the RNA-Seq data (Figure 15). There are two nucleotides (CC at 19,151,154-19,151,355) between the first complete codon (TTC) and the potential splice acceptor site. Hence, this splice acceptor site is in phase 2 relative

to Frame +2. This phase 2 splice acceptor site is compatible with the phase 1 splice donor site for CDS-2.

Clicking on the splice junction name (e.g. “JUNC00113128”) allows us to examine the number of spliced RNA-Seq reads that support this splice junction prediction (Figure 16).



Item: JUNC00113128
 Score: 5875
 Position: [NC_052530:19150962-19151252](#)
 Genomic Size: 291
 Strand: +

Figure 16 The score for JUNC00113128 shows that this junction is supported by 5,875 spliced RNA-Seq reads.

In addition to the splice junction JUNC00113128, which supports the proposed splice acceptor site for CDS-3 at 19,151,152-19,151,353, there is another splice junction which suggests a different splice acceptor site (JUNC00113129, Figure 15).

Thus, we need to investigate whether the predicted junction JUNC00113129 is supported by other lines of evidence. CDS-3 in *D. yakuba* includes two methionine in Frame +2 (at 19,151,255- 19,151,257 and 19,151,348-19,151,350). Hence, the splice junction JUNC00113129 could indicate the presence of a novel isoform of *Rheb* in *D. yakuba*. However, when we assess the number of spliced RNA-Seq reads that support the splice junction JUNC00113129, we find that this junction is weakly supported by only 8 spliced RNA-Seq reads; thus, there is little evidence to postulate a **novel isoform** of *Rheb* in *D. yakuba* based on this splice junction prediction.

Note: In addition to the scores, when analyzing multiple splice junction predictions for an intron, be sure to confirm the predictions are on the same strand as the gene you’re annotating. For example, if a splice junction is predicted in the negative strand and the *Rheb* gene is on the positive strand relative to the *D. yakuba* NC_052530 scaffold, you could eliminate that prediction.