Pathways Project: Annotation Walkthrough

Katie Sandlin, Wilson Leung, & Laura Reed

Prerequisites

- Understanding Eukaryotic Genes (modules 1, 4, 5, and 6)
- RNA-Seq Primer
- An Introduction to NCBI BLAST

Resources & Tools

- All links for this walkthrough, including the recommended prerequisites listed above, can be found on the Pathways Project page of the GEP website (thegep.org/pathways).
- Pathways Project: Annotation Notebook
- Pathways Project: Annotation Report
- Pathways Project: Annotation Report Exemplar
- Pathways Project: Annotation Videos
- Pathways Project: Annotation Workflow

Table of Contents

Introduction .................................................................................................................................................................... 2

Part 1: Examine genomic neighborhood surrounding target gene in *D. melanogaster* ....................................................... 3

Part 2: Identify genomic location of ortholog in target species ......................................................................................... 7
  Part 2.1: Retrieve protein sequence for Rheb-PA in *D. melanogaster* ........................................................................... 7
  Part 2.2: Perform a BLAST search of Rheb-PA against the *D. yakuba* genome assembly ............................................... 8
  Part 2.3: Summarize the tblastn results for Rheb-PA on *D. yakuba* chr3R................................................................. 11

Part 3: Examine genomic neighborhood of putative ortholog in target species ............................................................... 13
  Part 3.1: Examine evidence for a protein-coding gene in region of *D. yakuba* surrounding the tblastn alignment ............. 14
  Part 3.2: Use synteny to gather additional evidence for the ortholog assignment............................................................ 17

Part 4: Determine structure of target gene in *D. melanogaster* ....................................................................................... 23

Part 5: Determine approximate location of coding exons (CDS’s) in target species .......................................................... 25

Part 6: Refine coordinates of coding exons (CDS’s) ......................................................................................................... 30
  Part 6.1: Verify start codon coordinates for Rheb-PA ....................................................................................................... 31
  Part 6.2: Verify stop codon coordinates for Rheb-PA ....................................................................................................... 33
  Part 6.3: Determine phases of donor and acceptor splice sites ........................................................................................ 35
Introduction

The Pathways Project is focused on annotating genes found in well characterized signaling and metabolic pathways across the *Drosophila* genus. The current focus is on the insulin signaling pathway which is well conserved across animals and critical to growth and metabolic homeostasis. The long-term goal of the Pathways Project is to analyze how the regulatory regions of genes evolve in the context of their positions within a network.

This walkthrough illustrates how to apply the GEP annotation strategy for the Pathways Project to construct a gene model for the Ras homolog enriched in brain (Rheb) gene (target gene) in *Drosophila yakuba* (target species). This walkthrough focuses on annotation of the coding regions, so we won’t annotate the untranslated regions (UTRs) nor the transcription start site (TSS).

It is important to note that the Rheb gene in *D. yakuba* is relatively straightforward in comparison to what your own project might be. Some of the steps in this walkthrough might appear to be excessive, but keep in mind that you will potentially have a more complex project, 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.

We recommend you follow the parts in the order they are presented, and then refer to parts when needed for your own project. Note that the figures have been configured to fit this document while still maintaining readability; therefore, your screen may differ slightly. Commas have been included with most coordinates to improve readability, but you do not have to enter them in the Genome Browser (navigation will work the same with or without commas).
Part 1: Examine genomic neighborhood surrounding target gene in *D. melanogaster*

1. Navigate to the GEP UCSC Genome Browser Gateway Page.
2. Click on “D. melanogaster” in the “REPRESENTED SPECIES” table.
3. Ensure “Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)” under the “D. melanogaster Assembly” field is selected.
4. Enter “**Rheb**” under the “Position/Search Term” field.
5. Click on the “Go” button (Figure 1).

![Figure 1](image)

**Figure 1** Navigate to the Rheb gene in the *D. melanogaster* Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) assembly.

6. On the following “FlyBase Protein-Coding Genes” page, click on “Rheb-RA at chr3R:5568921-5570491” (Figure 2).

![Figure 2](image)

**Figure 2** Navigate to Rheb-RA on the right arm of chromosome 3 (chr3R) in *D. melanogaster*.

As shown in Figure 2, *Rheb* has two isoforms in *D. melanogaster*, and they both span the same coordinates (i.e., both Rheb-RA and Rheb-RB are located on the scaffold of the right arm of chromosome 3 (chr3R) at 5,568,921 – 5,570,491).

Therefore, it doesn’t matter which isoform we selected here, but **if the coordinates of the isoforms in your own project are different, you should choose the longest isoform for this step.**

7. Because the Genome Browser remembers our previous track display settings, click on “default tracks” in the display configuration buttons below the Genome Browser image (Figure 3, arrow).
In the “FlyBase Protein-Coding Genes” track, we should now see the gene structure for the two isoforms of the *Rheb* gene (i.e., Rheb-RA and Rheb-RB). Notice that *Rheb* is on the “chr3R” scaffold (Figure 4).

8. Zoom out until you can see two genes on either side of *Rheb*.

We are now viewing the genomic neighborhood of the *Rheb* gene in *D. melanogaster* (i.e., region of the chr3R scaffold containing the *Rheb* gene and its two closest upstream and two closest downstream genes) (Figure 5).
FYI: **Upstream:** located on the 5’ side of the target gene  
**Downstream:** located on the 3’ side of the target gene

As we learned in Understanding Eukaryotic Genes: Module 1, genes on the same DNA molecule may be transcribed in opposite directions (i.e., genes on the positive strand of DNA are transcribed from left to right and genes on the negative strand of DNA are transcribed from right to left). Since genes have directionality, the areas of DNA we call upstream and downstream change depending on the orientation of the target gene (i.e., upstream and downstream is determined in relation to the directionality of your target gene).

Since river streams have directionality (i.e., flow in only one direction), to help you orient yourself, think of your target gene as a raft floating down a stream. Wherever the raft just floated from is “upstream” of its current location and wherever the raft is floating to is “downstream” of its current location. Notice how the areas of DNA we call upstream and downstream to the target change change depending upon whether the target gene is on the positive or negative strand of DNA.

Genomic Neighborhood of Target Gene on Positive Strand of DNA

Genomic Neighborhood of Target Gene on Negative Strand of DNA
Now we need to draw a sketch of the genomic neighborhood of *Rheb* in *D. melanogaster* (Figure 6). To do so, we must identify the direction of transcription for each gene by zooming in on an intron of each gene.

9. Start by zooming in on an intron of *Rheb*.
10. Since the arrows within the introns of *Rheb* point to the right, in our sketch we need to draw an arrow pointing to the right and label it *Rheb*.
11. Repeat Steps 9-10 for the two closest genes on both sides of *Rheb* (i.e., CG12746, CG2931, CRMP, and CG2926).
   • Note: If a gene only has one exon, and therefore no introns, zoom into the exon to identify the direction of transcription.

The arrows within the introns of *Rheb* point to the right; therefore, we know *Rheb* is transcribed from left to right and located on the positive strand of DNA in *D. melanogaster*. Since *Rheb* is our target gene (i.e., gene we are annotating), the two closest genes on its 5’ side are considered upstream and the two closest genes on its 3’ side are considered downstream.

- The two closest upstream genes to *Rheb* in *D. melanogaster* are CG2931 and CG12746.
- The two closest downstream genes to *Rheb* in *D. melanogaster* are CRMP and CG2926.

**D. melanogaster (chr3R scaffold)**

![Figure 6 Sketch of the genomic neighborhood of Rheb in D. melanogaster. Since the direction of the arrows within their introns point to the right, Rheb, CG12746, and CRMP are on the positive strand. CG2931 and CG2926 are on the negative strand since the direction of the arrows within their introns point to the left.](image)

Each protein-coding gene annotated by FlyBase in *D. melanogaster* has an annotation symbol that begins with the prefix “CG” (i.e., Computed Gene). Unless genes are characterized experimentally and formally named, they are referred to by this symbol. For example, the two neighboring upstream genes (i.e., CG2931 and CG12746) and one downstream gene (i.e., CG2926) have not yet been named. However, the Ras homolog enriched in brain and Collapsin Response Mediator Protein genes have been characterized by past scientific studies and so they are referred to by their gene symbols *Rheb* and *CRMP*, respectively.
Part 2: Identify genomic location of ortholog in target species

Now that we’ve examined the genomic neighborhood of Rheb in *D. melanogaster*, we need to identify the location of Rheb in *D. yakuba*.

Part 2.1: Retrieve protein sequence for Rheb-PA in *D. melanogaster*

1. In the “FlyBase Protein-Coding Genes” track, click on “Rheb-RA” (Figure 7).

   ![Figure 7](image_url) Click on “Rheb-RA” to view details regarding this protein-coding gene that was annotated by FlyBase.

2. Under the “Links to sequence” heading, click on the “Translated Protein” link (Figure 8, left).

   We are now viewing the sequence for the 182 amino acids in the translated protein of Rheb-RA in *D. melanogaster* (Figure 8, right).

3. Copy the entire sequence (including the header) so we can use it in our *tblastn* search.

![Figure 8](image_url) Click on the “Translated Protein” link for the Rheb-RA feature to obtain the sequence of the translated protein.
What is the difference between Rheb-RA and Rheb-PA?

The prefix "Rheb" corresponds to the gene symbol. The "R" or "P" in the suffix designates the associated transcript (mRNA) or protein-product of the gene, respectively. The “A” in the suffix refers to the A isoform of the gene.

Gene symbols (e.g., Rheb) are italicized while their mRNA and protein products are not (e.g., Rheb-RA and Rheb-PA).

Part 2.2: Perform a BLAST search of Rheb-PA against the D. yakuba genome assembly

The Basic Local Alignment Search Tool (BLAST) finds regions of local similarity between biological sequences by comparing nucleotide or protein sequences to sequence databases (or to an individual nucleotide or protein sequence) and calculates the statistical significance of each match. The statistical values we will focus on are:

- **Expect Value (E-value):** describes the expected number of BLAST hits with this alignment score or better due to chance
  - The lower the E-value (i.e., the closer it is to zero), the more significant the match.

- **Percent Identity:** describes how similar the query sequence is to the database (or subject) sequence (i.e., how many characters in each sequence are identical)
  - The higher the Percent Identity (i.e., the closer it is to 100), the more significant the match.

The five traditional BLAST programs:
- **blastn** program searches nucleotide databases using a nucleotide query
- **blastp** program searches protein databases using a protein query
- **blastx** program searches protein databases using a translated nucleotide query
- **tblastn** program searches translated nucleotide databases using a protein query
- **tblastx** program searches translated nucleotide databases using a translated nucleotide query

<table>
<thead>
<tr>
<th>BLAST Program</th>
<th>Query (sequence to match)</th>
<th>Database/Subject (searching for match)</th>
<th>Function</th>
<th>Common Use Cases</th>
</tr>
</thead>
<tbody>
<tr>
<td>blastn</td>
<td>nucleotide</td>
<td>nucleotide</td>
<td>searching with shorter queries, cross-species comparison</td>
<td>map mRNAs against genomic assemblies</td>
</tr>
<tr>
<td>blastp</td>
<td>protein</td>
<td>protein</td>
<td>general sequence identification and similarity searches</td>
<td>search for proteins similar to predicted genes</td>
</tr>
<tr>
<td>blastx</td>
<td>nucleotide → protein</td>
<td>protein</td>
<td>identifying potential protein products encoded by a nucleotide query</td>
<td>map proteins/CDS against genomic sequence</td>
</tr>
<tr>
<td>tblastn</td>
<td>protein</td>
<td>nucleotide → protein</td>
<td>identifying database sequences encoding proteins similar to query</td>
<td>map proteins against genomic assemblies</td>
</tr>
<tr>
<td>tblastx</td>
<td>nucleotide → protein</td>
<td>nucleotide → protein</td>
<td>identifying nucleotide sequences similar to the query based on their coding potential</td>
<td>identify genes in unannotated sequences</td>
</tr>
</tbody>
</table>

Arrows indicate the BLAST program translates the nucleotide sequence before performing the search.

See An Introduction to NCBI BLAST for additional details.

In Part 2.1, we retrieved the protein sequence for Rheb-PA in D. melanogaster. Now we need to perform a BLAST search of Rheb-PA against the D. yakuba genome assembly.

Since we are looking for the orthologous region of Rheb in D. yakuba, we want BLAST to search the entire genome of D. yakuba to identify regions of local similarity with the protein sequence of Rheb in D. melanogaster we obtained in Part 2.1. In other words, we need to BLAST our Rheb protein sequence
from *D. melanogaster* against the entire genome of *D. yakuba* to narrow down the possible regions where *Rheb* could be in *D. yakuba*.

For this **BLAST** search, we will use the *tblastn* program to search the translated nucleotide database of our target species, *D. yakuba*, using the protein sequence of *Rheb* in *D. melanogaster* as our query.

- **Query**: sequence we are looking to match (i.e., protein sequence of *Rheb* in *D. melanogaster*)
- **Database (Subject)**: collection of sequences we are searching for matches (**BLAST** will translate the entire genome of *D. yakuba* before searching for a match to the Rheb-PA sequence from *D. melanogaster*)

1. Navigate to the Pathways Project Genome Assemblies page.
2. Click on the “Genome BLAST” link for *D. yakuba* (Figure 9).
   - Note: Here we are selecting the entire genome of *D. yakuba* as our database to search.

![Pathways Project Genome Assemblies](image)

**Figure 9** Click on the “Genome BLAST” link for *D. yakuba* in the “NCBI BLAST Link” column.

3. Make sure the “*tblastn*” tab is selected (Figure 10).
4. Paste the Rheb-PA sequence we copied from Part 2.1 into the “Enter Query Sequence” text box. The “Job Title” text box should then automatically populate.
5. Select the check box next to “Show results in a new window.”
6. Click on the “BLAST” button.

---

**FYI**

The National Center for Biotechnology Information (NCBI) periodically updates the genome assemblies used for **BLAST** searches that can cause a host of issues when they occur in the middle of a semester. Using the “Genome BLAST” links on the “Pathways Project Genome Assemblies” page ensures student annotators are navigating to the correct genome assembly database when performing their **BLAST** search against the whole genome of their target species. However, **this circumvention will only be needed in Part 2 of the Pathways Project protocol**. In Parts 3 and 5, we will navigate directly to the NCBI BLAST home page.
Figure 10 Configure tblastn to compare the *D. melanogaster* protein Rheb-PA (query) against the *D. yakuba* whole genome assembly (database).

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 if we have the best ortholog.

Our tblastn search found six regions within the *D. yakuba* genome that show similarities with the protein sequence of Rheb in *D. melanogaster* (Figure 11); however, only one of these is a good hit (*Drosophila yakuba* strain Tai18E2 chromosome 3R, whole genome shotgun sequence; sequence identity: 97.14% and E-value: 2e⁻⁷⁸). The second hit has a much higher E-value (8e⁻³⁷) and much lower percent identity (43.75%), and this pattern of increasingly lower quality matches continues throughout the other four matches as well. Therefore, we will continue our analysis based on the hypothesis that the putative (probable) ortholog of Rheb-PA in *D. yakuba* is located somewhere in the scaffold of chromosome 3R (chr3R).

Notice that each of the genome regions of our six hits has a unique accession number. The accession number for the chromosome 3R scaffold in *D. yakuba* is CM000160.2 (Figure 11, right arrow).
Figure 11  Our tblastn search with the *D. melanogaster* protein Rheb-PA (query) against the *D. yakuba* whole genome assembly (database) found six regions of similarity. The best match is located on the “chromosome 3R” scaffold (accession number: CM000160.2) of *D. yakuba*.

Part 2.3: Summarize the tblastn results for Rheb-PA on *D. yakuba* chr3R

1. Under the “Descriptions” tab, uncheck the box next to “select all” (Figure 12).
   - Note: This will hide the alignments from chromosomes 3L, X, 2L, and 2R since we only want to look at the best match (i.e., chromosome 3R).
2. Click on the “*Drosophila yakuba* strain Tai18E2 chromosome 3R, whole genome shotgun sequence” link in the “Description” column to navigate to the alignment.

Figure 12  Uncheck “select all” and then click on the link for the best match to navigate to the alignment.

3. In the blue toolbar for the BLAST hit, select the “Subject start position” option from the drop-down menu of the “Sort by” field to order the matches based on the start coordinates on *D. yakuba* chromosome 3R in ascending order (Figure 13).
Figure 13  In the blue toolbar for the BLAST hit, select the “Subject start position” option from the drop-down menu of the “Sort by” field to order the matches based on the start coordinates of the D. yakuba chromosome 3R scaffold in ascending order.

We should now see 12 ranges listed in ascending order. Each range corresponds to a contiguous portion of the subject sequence (i.e., D. yakuba genome) that shows significant sequence similarity (i.e., matches) with a portion of the query sequence (i.e., D. melanogaster Rheb-PA). Here, the 12 ranges show matches between Rheb-PA in D. melanogaster (Query) and a range of coordinates from chromosome 3R in D. yakuba (Sbjct). For example, Range 1 shows a match between Rheb-PA in D. melanogaster (Query: 55 – 163) and coordinates 5,727,469 – 5,727,792 of chromosome 3R in D. yakuba when translated in Frame +1. Range 1 has an E-value of 2e-10 and a sequence identity of 36% (Figure 14).

Figure 14  The query coordinates (55 – 163) and subject coordinates (5,727,469 – 5,727,792) are shown in red and blue, respectively.

Since chromosome 3R in D. yakuba is 28,832,112 bp long (Figure 14, blue circle), it is likely we will have some ranges (i.e., alignment matches or hits) that do not 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 15).
The best collinear set of alignments to Rheb-PA is located at 17,358,666 – 17,359,556 on the chromosome 3R scaffold of the *D. yakuba* genome assembly and the five alignment matches cover all 182 amino acids of Rheb-PA (Figure 15). Therefore, we will continue our analysis based on the hypothesis that the putative ortholog of Rheb-PA is located at approximately 17,358,666-17,359,556 on the chromosome 3R scaffold of the *D. yakuba* genome assembly. See Appendix A for information on investigating the other *tblastn* alignments to *D. yakuba* chromosome 3R scaffold.

<table>
<thead>
<tr>
<th>Range</th>
<th>D. melanogaster</th>
<th>D. yakuba</th>
<th>E-Value</th>
<th>Identities (%)</th>
<th>Subject Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Query Start</td>
<td>Query End</td>
<td>Subject Start</td>
<td>Subject End</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>55</td>
<td>163</td>
<td>5,727,469</td>
<td>5,727,792</td>
<td>2e-10</td>
</tr>
<tr>
<td>3</td>
<td>8</td>
<td>131</td>
<td>9,373,502</td>
<td>9,373,062</td>
<td>1e-09</td>
</tr>
<tr>
<td>4</td>
<td>54</td>
<td>130</td>
<td>9,387,874</td>
<td>9,387,623</td>
<td>4e-11</td>
</tr>
<tr>
<td>5</td>
<td>6</td>
<td>44</td>
<td>9,388,199</td>
<td>9,388,062</td>
<td>0.002</td>
</tr>
<tr>
<td>6</td>
<td>18</td>
<td>108</td>
<td>9,658,240</td>
<td>9,658,575</td>
<td>5e-06</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>20</td>
<td>17,358,666</td>
<td>17,358,725</td>
<td>2e-78</td>
</tr>
<tr>
<td>8</td>
<td>16</td>
<td>45</td>
<td>17,358,838</td>
<td>17,358,927</td>
<td>2e-78</td>
</tr>
<tr>
<td>9</td>
<td>40</td>
<td>109</td>
<td>17,359,007</td>
<td>17,359,216</td>
<td>2e-78</td>
</tr>
<tr>
<td>10</td>
<td>111</td>
<td>153</td>
<td>17,359,279</td>
<td>17,359,407</td>
<td>2e-78</td>
</tr>
<tr>
<td>11</td>
<td>153</td>
<td>182</td>
<td>17,359,467</td>
<td>17,359,556</td>
<td>2e-78</td>
</tr>
<tr>
<td>12</td>
<td>53</td>
<td>173</td>
<td>26,681,803</td>
<td>26,681,411</td>
<td>1e-05</td>
</tr>
</tbody>
</table>

Figure 15  Summary of the *tblastn* search results for the 12 matches to Rheb-PA within chromosome 3R of *D. yakuba*. The best collinear set of alignments to Rheb-PA is located at 17,358,666-17,359,556. As we saw on the right side of Figure 8, Rheb-PA in *D. melanogaster* is 182 amino acids long and the above collinear set of alignments covers all 182 amino acids.

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

In Part 1, we sketched the genomic neighborhood of *Rheb* in *D. melanogaster*. Here we will examine the genomic neighborhood of *Rheb* in *D. yakuba* and then compare the order and orientation of these genes to what we found in *D. melanogaster*.

Based on parsimony, the genes surrounding *Rheb* in *D. yakuba* should be identical or very similar in function to the genes in the genomic neighborhood of *Rheb* in *D. melanogaster*. Additionally, the neighboring genes should also match in orientation (look at the direction of transcription of the neighboring genes; this is called synteny). Since *Rheb* and CG2926 are on the positive and negative strand, respectively, in *D. melanogaster*, these two genes should also be on different strands in *D. yakuba*. 
Part 3.1: Examine evidence for a protein-coding gene in region of *D. yakuba* surrounding the *tblastn* alignment

1. Navigate to the [Genome Browser](#).
2. Click on “D. yakuba” in the “REPRESENTED SPECIES” table.
3. Under the “D. yakuba Assembly” field, confirm that “May 2011 (WUGSC dyak_caf1/DyakCAF1)” is selected.
4. In Part 2, we determined the putative ortholog of Rheb-PA is located at approximately 17,358,666 – 17,359,556 on the chromosome 3R (chr3R) scaffold of the *D. yakuba* genome assembly. Enter “chr3R:17,358,666-17,359,556” under the “Position/Search Term” field to examine this region.
5. Click on the “Go” button (Figure 16).
The accession number for the *D. yakuba* chromosome 3R scaffold (CM000160.2) and scaffold name (chr3R) are synonymous (i.e., both are different ways to refer to the same region of the *D. yakuba* genome).

In the *D. yakuba* Genome Browser, “CM000160.2:17,358,666-17,359,556” OR “chr3R:17,358,666-17,359,556” could be entered in the “Position/Search Term” field text box and we would still navigate to the same genomic region.

Annotators can use the accession number to navigate to a genomic region in all the Genome Browsers for the Pathways Project. Don’t be alarmed by the black bar in the Genome Browser image that will appear in either the "INSDC" or the "RefSeq Acc" evidence track when you navigate to a region using the accession number. In the Genome Browser image, you can hide the track by either 1.) right-clicking on the “INSDC” or “RefSeq Acc” track and then clicking on “hide” or 2.) clicking on “default tracks”.

It is *not* important for you to understand the technical minutiae of why the black bar appears nor why it’s fine to ignore it. However, if you’re interested to learn more, please see Appendix B.

**CAVEAT:** Genome Browsers vary depending on the *Drosophila* species, so navigating with the scaffold name may not always work and you may only be able to use the accession number for navigation.

See Appendix B for more information about accession numbers.

---

Figure 16  Navigate to the region surrounding the best collinear set of alignments to the *D. melanogaster* Rheb-PA protein in the *D. yakuba* May 2011 (WUGSC dyak_caf1/DyakCAF1) assembly (i.e., chr3R:17,358,666-17,359,556).

The Genome Browser Gateway should default to the correct assembly once you click on the *Drosophila* species in the left-hand menu. To double check though, you can see which assembly you should be using via the "Gene Browsers" column of the "Pathways Project Genome Assemblies" web page. For example, *D. yakuba* has two different assembly options to choose from and we should use the "May 2011 (WUGSC dyak_caf1/DyakCAF1)" assembly.

6. In the list of buttons below the Genome Browser image, click on “default tracks” (Figure 3).
7. Zoom out 3x.
In the Genome Browser image, we should now see the following tracks (Figure 17):

- BLAT Alignments of NCBI RefSeq Genes
- Spaln Alignment of *D. melanogaster* Proteins
- Gene Prediction Tracks:
  - GeMoMa Gene Predictions
  - Geneid Gene Predictions
  - Augustus Gene Predictions
- modENCODE RNA-Seq from Adult Females
- modENCODE RNA-Seq from Adult Males

Notice that the BLAT alignment for the *coding regions* of the *D. yakuba* RefSeq transcript XM_039375862 against the chr3R 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 first four alignment blocks are in congruence with the placements of the first four coding exons (CDS's) predicted by GeMoMa, Geneid, and Augustus. The coding portions of the last alignment block are in congruence with the predictions by GeMoMa and Augustus.

According to the BLAT alignment for the *RefSeq* transcript XM_039375862, the putative (probable) ortholog of Rheb-PA is located at chr3R:17,358,367-17,359,937 in *D. yakuba* and the region spanning chr3R:17,358,666-17,359,559, within the putative ortholog, corresponds to the alignment to the coding region of the *RefSeq* transcript.

![UCSC Genome Browser on D. yakuba May 2011 (WUGSC dyak_caf1/DyakCAF1) Assembly](image)

*Figure 17  Genome Browser image of chr3R:17,357,775-17,360,447 in *D. yakuba* (default tracks).*
Part 3.2: Use synteny to gather additional evidence for the ortholog assignment

1. In the “BLAT Alignments of NCBI RefSeq Genes” track shown in the Genome Browser image, click on “XM_039375862” (Figure 17, arrow).
   - Note: XM_039375862 is the accession number for the D. yakuba RefSeq mRNA transcript that is aligned to this region of the D. yakuba chr3R scaffold by BLAT.

Notice the position of XM_039375862 is “chr3R:17,358,367-17,359,937” (Figure 18).
2. Scroll to the bottom of the GenBank Record window to the “translation” sequence within the “CDS” section (Figure 19).

We are now viewing the computationally predicted protein sequence of the putative ortholog to Rheb-PA in *D. yakuba*.

3. Copy the accession number for the translated protein sequence (see arrow in Figure 19).

![BLAT Alignments of NCBI RefSeq Genes (XM_039375862)](image)

When we run BLAST for this protein, we can enter the accession number “XP_039231796.1” and the program will use the translated protein sequence shown in the grey box in Figure 19. Note: We recommend using the accession number instead of the actual sequence (i.e., MPTKERNI... here because the formatting may get a bit mixed up depending on the web browser or computer you use.

---

The NCBI database of reference sequences (RefSeq) is a curated, non-redundant collection of naturally occurring DNA, RNA, and protein sequences. The RefSeq database includes both known (manually reviewed by NCBI staff or collaborators) and computationally predicted sequences.

The two-letter prefix (followed by an underscore) of RefSeq accession numbers has an implied meaning with respect to the type of molecule it represents (e.g., known or predicted model, genomic scaffold, mRNA).

<table>
<thead>
<tr>
<th>Accession Prefix</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XM_</td>
<td>mRNA (predicted model)</td>
</tr>
<tr>
<td>XP_</td>
<td>protein (predicted model)</td>
</tr>
<tr>
<td>NP_</td>
<td>protein (known)</td>
</tr>
</tbody>
</table>

A complete list of RefSeq accession number prefixes is available in the NCBI Handbook.
4. Navigate to NCBI BLAST.
5. Click on the “Protein BLAST” button (Figure 20).
   - Note: This is a blastp search (Part 2.2, review box).

6. Paste the accession number for the translated protein sequence we copied from the GenBank Record for XM_039375862 (i.e., “XP_039231796.1”) into the “Enter Query Sequence” text box (Figure 21). The “Job Title” text box should then automatically populate.
8. In the “Organism” text box, enter “Drosophila melanogaster (taxid:7227).”
9. Select the check box next to “Show results in a new window”.
10. Click on the “BLAST” button.
Our *blastp* search found 53 proteins within the *D. melanogaster* reference proteins database that show similarities with the protein sequence of the putative ortholog of Rheb-PA in *D. yakuba* (XP_039231796.1); however, only one of these is a good hit (Ras homolog enriched in brain, isoform A [Drosophila melanogaster]; accession: NP_730950.2, sequence identity of 97.25%, and an E-value of 6e-131). The remaining hits had much higher E-values (5e-41 to 0.002) and much lower percent identities (43.75% to 23.48%) (Figure 22). Therefore, the *blastp* search result shows that the *D. yakuba* RefSeq prediction XP_039231796.1 is most similar to the A isoform of Rheb among all the annotated proteins in *D. melanogaster*. Hence, based on parsimony, the *blastp* search result supports the hypothesis that XP_039231796.1 is the putative ortholog of Rheb in *D. yakuba*.

**Figure 22** The best *blastp* match to the putative ortholog of Rheb-PA in *D. yakuba* is “Ras homolog enriched in brain, isoform A [Drosophila melanogaster]” (Accession: NP_730950.2) with an E-value of 6e-131 and a sequence identity of 97.25%.

To help us stay oriented to the location of our target gene while looking at the neighboring genes, we can right-click on the the RefSeq transcript XM_039375862 and then select “Highlight XM_039375862.” We should now see the region of our target gene highlighted in blue. When we zoom out to examine the neighboring transcripts, we can easily visualize where our target gene is because it will remain highlighted. To clear, right-click on the highlighted area and then select “Remove highlight.”

Note: When we right-click on the highlighted area, notice that we could also click on “Zoom to highlighted” to quickly navigate to the region.
11. Now we need to repeat the *blastp* search for the transcripts of the two neighboring genes on both sides of XM_039375862 (i.e., XM_002097995, XM_015192882, XM_002097997 and XM_002097998) (i.e., repeat Steps 1-10 for the neighboring genes (Figure 24)). See Appendix C for instructions on how to combine (or batch) multiple sequences so that you only must run one BLAST search instead of four.

If the coding isoforms of your gene differ, choose the longest coding isoform that is best supported by the other lines of evidence as the query for your *blastp* search.

Now that we’ve examined the genomic neighborhood of the putative ortholog, we need to identify the direction of transcription for the putative ortholog of Rheb-PA and the neighboring genes in *D. yakuba* and then use this information to draw a sketch of the genomic neighborhood. To do so, we need to repeat the process we followed in Part 1 for the putative ortholog of Rheb-PA on each of the neighboring genes in *D. yakuba* identified in our *blastp* search (i.e., zoom into an intron of each gene and draw arrows in the correct directions on our sketch).

12. Use the strategy described in Steps 10 – 11 of Part 1 to determine the orientations of the putative Rheb ortholog (XM_039375862) and the transcripts of the two neighboring genes on both sides (i.e., XM_002097995, XM_015192882, XM_002097997 and XM_002097998) (Figure 23).

**D. yakuba (chr3R scaffold)**

![Figure 23: Sketch of the genomic neighborhood of the putative Rheb ortholog (on positive strand) in D. yakuba. The putative orthologs of CG12746 and CG2931 are located upstream of Rheb, and the putative orthologs of CRMP and CG2926 are located downstream of Rheb in the D. yakuba chr3R scaffold, and each gene is on the +, -, +, and - strands, respectively.](image-url)
Figure 24  Summary of the blastp search results for the protein sequences of the two nearest predicted upstream and downstream neighbors to Rheb-PA in D. yakuba against the D. melanogaster reference proteins database.

<table>
<thead>
<tr>
<th>Target Species</th>
<th>2nd Closest Upstream</th>
<th>Closest Upstream</th>
<th>Target Gene</th>
<th>Closest Downstream</th>
<th>2nd Closest Downstream</th>
</tr>
</thead>
<tbody>
<tr>
<td>D. melanogaster Gene Symbol</td>
<td>CG12746</td>
<td>CG2931</td>
<td>Rheb</td>
<td>CRMP</td>
<td>CG2926</td>
</tr>
<tr>
<td>Strand (+/-)</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>NCBI RefSeq Gene (mRNA) Accession</td>
<td>XM_015192882</td>
<td>XM_002097995</td>
<td>XM_039375862</td>
<td>XM_002097997</td>
<td>XM_002097998</td>
</tr>
<tr>
<td>NCBI RefSeq Protein Accession</td>
<td>XP_015048368.1</td>
<td>XP_002098031.1</td>
<td>XP_039231796.1</td>
<td>XP_002098033.2</td>
<td>XP_002098034.1</td>
</tr>
<tr>
<td>Strand (+/-)</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Accession</td>
<td>NP_649551.4</td>
<td>NP_649552.1</td>
<td>NP_730950.2</td>
<td>NP_730954.2</td>
<td>NP_649554.1</td>
</tr>
<tr>
<td>D. melanogaster Gene Symbol</td>
<td>CG12746</td>
<td>CG2931</td>
<td>Rheb</td>
<td>CRMP</td>
<td>CG2926</td>
</tr>
<tr>
<td>E-Value</td>
<td>0.0</td>
<td>0.0</td>
<td>6e-131</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Percent Identity</td>
<td>84.30%</td>
<td>96.72%</td>
<td>97.25%</td>
<td>99.66%</td>
<td>86.18%</td>
</tr>
</tbody>
</table>

If the genomic neighborhood looks similar between Rheb in D. melanogaster (Figure 6) and the putative ortholog in D. yakuba (Figure 23), 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 inspect our other hits in the tblastn search (Part 2) to see if a different genomic region is a better match overall.

Examination of the genomic regions surrounding the Rheb gene in D. melanogaster (Figure 25; top) and the putative Rheb ortholog in D. yakuba (Figure 25; 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 chr3R:17,358,666-17,359,556 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.
Part 4: Determine structure of target gene in *D. melanogaster*

In Part 4 we will use the Gene Record Finder, which is a web tool that enables us to quickly identify a unique set of exons for a given gene and to retrieve their **Coding DNA Sequences (CDS’s)**, also referred to as **coding exons**.

The Gene Record Finder will also provide details, such as number of isoforms, exon-intron structure and their coordinates, and transcript and protein information of the gene in question (in this case *Rheb*). **It is important to remember that the details provided by the Gene Record Finder are for the gene in the reference species, *D. melanogaster*.** We will use the details from *Rheb* in *D. melanogaster* to assist us with creating a gene model for *Rheb* in *D. yakuba*.

Before we can construct the orthologous gene model, we need to ascertain the gene structure (e.g., number of isoforms and CDS’s) of the *D. melanogaster Rheb* gene using the Gene Record Finder.

1. Open a new web browser tab and navigate to the Gene Record Finder.
2. Enter “Rheb” into the text box.
3. Click on the “Find Record” button (Figure 26).

If we enter “rheb” in the Gene Record Finder, we will get the following error:

```text
Cannot complete the request because of the following error: Cannot find the FlyBase gene record: rheb
```

This is because gene symbols in *Drosophila* are **case-sensitive**. For example, *Tor* and *tor* correspond to two different genes in *D. melanogaster*. 

---

*Figure 25  Comparison of the relative order and orientation of the genomic neighborhoods of *Rheb* in *D. melanogaster* (top) and *D. yakuba* (bottom).*
The Gene Record Finder shows that Rheb has two isoforms (A and B) in *D. melanogaster*. A graphical overview of the two isoforms is shown in the “mRNA Details” panel. The “CDS usage map” (under the “Polypeptide Details” tab) shows that both isoforms have the same set of coding exons (CDS’s) (i.e., 1_9845_0, 2_9845_2, 3_9845_2, 4_9845_1, and 5_9845_0). (The coding exons are ordered from 5’ to 3’ (from left to right) in the CDS usage map.) Hence the differences between these two isoforms are limited to the untranslated regions (UTRs) (Figure 27).
Based on parsimony (i.e., minimizing the number of changes compared to *D. melanogaster*), we expect to find both the A and B isoforms of *Rheb* in our *D. yakuba* genome sequence. For this walkthrough, we will only focus on annotating the CDS’s, which do not include the UTRs. Consequently, we only need to determine the coordinates of the five CDS’s for one of the isoforms (e.g., isoform A) because the set of coding exons for both the A and B isoforms are the same. **If the coding sequences for the two isoforms weren’t identical, we’d need to annotate both isoforms.**

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

The initial *tblastn* search we performed in Part 2 helped define the search *region* for the putative ortholog within the genomic scaffold of the target genome, *D. yakuba*. The approximate coordinates of each CDS can now be determined by aligning each CDS of the gene in *D. melanogaster* against this search region of the target genome.

The next step in our analysis is to determine the approximate coordinates of each coding exon (CDS) of Rheb-PA in *D. yakuba*. 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. 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.

In addition to comparing a query sequence against a collection of subject sequences in a database (e.g., Part 2 of this walkthrough), the NCBI **BLAST** web service also allows us to compare two or more sequences against each other (using the program *bl2seq* (**BLAST** 2 **sequences**)).

To map the amino acid sequences of each *D. melanogaster* CDS against the *D. yakuba* scaffold, we must translate the entire *D. yakuba* scaffold sequence in all six reading frames (i.e., three reading frames in the positive strand and three reading frames in the negative strand) and then compare each **conceptual translation** against each CDS sequence from *D. melanogaster*. This means that we can use either *tblastn* or *blastx* to perform this search (depending on whether we treat the CDS sequence as the query or the subject sequence, respectively).

Here, we will perform five *tblastn* searches— using each *D. melanogaster Rheb* CDS sequence as the query and the accession number (CM000160.2) we identified for the *D. yakuba* chr3R scaffold in Part 2.2 as the database (subject) sequence.
1. Scroll down to the CDS usage map (under the “Polypeptide Details” tab). Since *Rheb* has 5 CDS’s, we will need to run five different *tblastn* searches, one for each of our CDS’s. Let’s start with CDS-1 (Figure 28).

2. To view the protein sequence for CDS-1, select row 1 (FlyBase ID: 1_9845_0).

3. Copy the protein sequence (including the header) shown in the pop-up window.

4. To setup the *tblastn* search, navigate to the [NCBI BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi) website.

5. Click on the “*tblastn*” image under the “Web BLAST” section (Figure 29).

![Figure 28](image1.png) **Use the Gene Record Finder to retrieve the amino acid sequence for CDS-1 (FlyBase ID: 1_9845_0). To obtain the sequence for CDS-1, select row 1 (left) and then copy the protein sequence shown in the pop-up window (right). The “Size (aa)” column tells us how many amino acids (aa) are in each CDS (circle) (e.g., CDS-1 is 16 amino acids long).**

![Figure 29](image2.png) **Navigate to the [NCBI BLAST](https://blast.ncbi.nlm.nih.gov/Blast.cgi) website, and then click on the “*tblastn*” image.**
6. Paste the sequence for CDS-1 into the “Enter Query Sequence” text box (Figure 30).
7. Select the “Align two or more sequences” checkbox.
   - Note: NCBI BLAST is using the bl2seq (BLAST 2 sequences) program.
8. In the “Enter Subject Sequence” text box, enter the Accession Number for the D. yakuba chr3R scaffold (i.e., “CM000160.2”) we identified in Part 2.2.

The average novel contains roughly 400,000 characters (not including spaces). The chr3R scaffold in D. yakuba is 28,832,112 bp (or characters) long, which is roughly equivalent to 72 novels worth of information. Furthermore, since BLAST must translate the entire D. yakuba scaffold sequence in all six reading frames, that would be more like searching 432 novels to find a few pages of important information. While the English alphabet consists of 26 letters, the genomic alphabet only consists of four (A, T, G, and C). With 432 novels worth of only A’s, T’s, C’s, and G’s BLAST must search, we’re likely to get several spurious alignments; however, limiting the search region on chr3R will increase the statistical power of the search and reduce the number of spurious matches.

Think back to our findings in Part 2.2 and 2.3. We found the best collinear set of alignments to Rheb-PA in D. yakuba at 17,358,666-17,359,556 on the chr3R scaffold. Thus, we can narrow down the 432 novels worth of information to less than one novel by limiting the subrange of the chr3R scaffold (CM000160.2) to 100,000 bp (i.e., searching chr3R: 17,300,000-17,400,000).

Note: the subrange can be larger than 100,000 bp. Whatever you choose, be sure to generously round (like you wish your instructor would your grade ☺) in both directions.

9. Based on our analysis in Part 2.3, limit the “Subject subrange” by entering from “17300000” to “17400000” (Figure 30).
   - Note: Do NOT include commas or BLAST will search outside the subject subrange.

The default NCBI BLAST parameters are optimized for searching the query sequence against a large collection of sequences in a database. When we are using BLAST to compare only two sequences against each other, we need to change some of the algorithm parameters because the default parameters could potentially mask the conserved regions of the coding exon.

10. Click on the “+” icon next to “Algorithm parameters” to expand the section (Figure 31).
11. In the “Scoring Parameters” section, change the “Compositional adjustments” field to “No adjustment.”
12. In the “Filters and Masking” section, uncheck the “Low complexity regions” checkbox in the “Filter” field.
13. Select the check box next to “Show results in a new window” then click on the “BLAST” button.
Figure 30  Configure tblastn to compare the *D. melanogaster* CDS-1 (query) against the *D. yakuba* chr3R scaffold (Accession: CM000160.2; subject). Since the tblastn search of Rheb-PA against the *D. yakuba* assembly placed the putative Rheb ortholog at approximately 17,358,666-17,359,556 on scaffold chr3R, the “Subject subrange” is used to limit the search region from 17,300,000 to 17,400,000.

Figure 31  Adjust “Algorithm parameters” to decrease the likelihood that conserved regions of the coding exon will be masked.

The tblastn results show a single match (E-value: 5e-07; sequence identity: 93.75%) to CDS-1.

14. Click on the “Alignments” tab to view the corresponding tblastn alignment (Figure 32).
Figure 32 The **tblastn** alignment between the *D. melanogaster* CDS-1 (query) against the *D. yakuba* chr3R scaffold (subject) within the subject subrange of 17,300,000-17,400,000 is located at 17,358,666-17,358,713 in the chr3R scaffold of *D. yakuba* (blue boxes) when the sequence is translated in Frame +3. The **tblastn** alignment has an E-value of 5e-07 and 94% sequence identity. This alignment covers all 16 aa of CDS-1 (red boxes).

The “Query” coordinates show that the alignment covers all 16 amino acids (aa) of CDS-1 (Figure 32).

- **NOTE:** We can find the length (in aa) of CDS-1 in *D. melanogaster* using the Gene Record Finder (“Size (aa)” column under the “Polypeptide Details” tab; Figure 28, circle) or at the top of the **tblastn** search results.

The “Subject” coordinates correspond to the region within chr3R scaffold of *D. yakuba* (i.e., 17,358,666 – 17,358,713) that shows sequence similarity to CDS-1 from *D. melanogaster* when it is translated in the third reading frame in the positive strand (i.e., Frame +3). Hence, we can place CDS-1 of *Rheb* at approximately “chr3R:17,358,666-17,358,713” in *D. yakuba*.

We can apply this same procedure to place the other four CDS’s on the chr3R scaffold of *D. yakuba*. See Appendix C for instructions on how to combine (or batch) multiple sequences so you only must run one BLAST search instead of four more.

15. Copy the CDS-2 sequence (along with the header) from the Gene Record Finder.
16. Return to the **tblastn** search web browser tab and delete the CDS-1 sequence from the “Enter Query Sequence” textbox.
17. Paste the CDS-2 sequence in the textbox.
Leave everything else the same as we had it for CDS-1.

18. Click on the “BLAST” button to run the tblastn search.
19. Click on the “Alignments” tab to view the corresponding tblastn alignment.
20. Repeat Steps 15-19 to BLAST the remaining CDS’s (Figure 33).

<table>
<thead>
<tr>
<th>CDS</th>
<th>FlyBase ID</th>
<th>Query Length Size (aa)</th>
<th>D. melanogaster</th>
<th>Target Species</th>
<th>Subject Frame</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Query Start</td>
<td>Query End</td>
<td>Subject Start</td>
</tr>
<tr>
<td>1</td>
<td>1_9845_0</td>
<td>16</td>
<td>1</td>
<td>16</td>
<td>17,358,666</td>
</tr>
<tr>
<td>2</td>
<td>2_9845_2</td>
<td>23</td>
<td>1</td>
<td>23</td>
<td>17,358,844</td>
</tr>
<tr>
<td>3</td>
<td>3_9845_2</td>
<td>68</td>
<td>1</td>
<td>68</td>
<td>17,359,013</td>
</tr>
<tr>
<td>4</td>
<td>4_9845_1</td>
<td>43</td>
<td>1</td>
<td>43</td>
<td>17,359,279</td>
</tr>
<tr>
<td>5</td>
<td>5_9845_0</td>
<td>30</td>
<td>1</td>
<td>30</td>
<td>17,359,470</td>
</tr>
</tbody>
</table>

Figure 33  Summary of the tblastn CDS-by-CDS search results.

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 placed on the positive strand and the subject ranges for the CDS’s are in ascending order. Consequently, the CDS-by-CDS search results support the hypothesis that the putative (probable) ortholog of Rheb-PA is located at approximately 17,358,666-17,359,559 on the chr3R scaffold of the D. yakuba genome assembly (i.e., chr3R:17,358,666-17,359,559).

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

Now that we’ve mapped each CDS separately to determine their approximate locations (Figure 33), we will further refine the CDS boundaries by searching for compatible splice donor and acceptor sites by visual inspection using the Genome Browser.

As part of the modENCODE project, RNA-Seq data for D. yakuba was generated using samples from adult females and males. These RNA-Seq reads (100–125 bp in length) are derived primarily from processed mRNA (i.e., after the introns have been removed). Hence, genomic regions with RNA-Seq read coverage usually correspond to transcribed exons, which include both the translated and untranslated regions.

The “modENCODE RNA-Seq” tracks correspond to the samples from adult females (red) and males (blue) where RNA-Seq data is available (Figure 34). The height of the histograms within each track corresponds to the number of RNA-Seq reads that have been mapped to each position of the D.
yakuba scaffold. By default, the scale of the “RNA-Seq Coverage” track will change automatically based on the minimum and maximum read depth within the genomic region being viewed.

If the number of RNA-Seq reads differs between samples (e.g., adult males vs. adult females), that would indicate alternative splicing.

See Understanding Eukaryotic Genes Module 6 to review alternative splicing. See RNA-Seq Primer for an overview of the different types of RNA-Seq data available on the Genome Browser.

Part 6.1: Verify start codon coordinates for Rheb-PA

Introns are removed from the pre-mRNA prior to translation. The processed mRNA also includes a 5’ cap at the 5’ end and a poly-A tail at the 3’ end. This walkthrough focuses on annotation of the coding regions so we will not annotate the untranslated regions (UTRs) or the transcription start site (TSS).

- **codon**: sequence of three nucleotides of DNA (A, T, G, and C) or RNA (A, U, G, and C) that corresponds to a specific amino acid
- **start codon**: first codon of a coding sequence; signals the ribosome to begin translation of the mRNA; ATG (AUG) codes for Methionine (M)
- **stop codon**: signals the termination of translation; sometimes called “nonsense codons,” since they do not specify any amino acid; TAA (UAA), TAG (UAG), TGA (UGA)

We need to ascertain whether the *tblastn* alignment for CDS-1 at 17,358,666-17,358,713 (Figure 34) is supported by the RNA-Seq data and determine the location of the start codon.

1. Return to the **Genome Browser** for *D. yakuba*.
2. To examine the region of the *tblastn* alignment for CDS-1, enter “chr3R:17,358,666-17,358,713” into the “enter position or search terms” text box.
3. Under the “Mapping and Sequencing Tracks,” change the “Base Position” track to “full.”
4. Click on any “refresh” button.
The RNA-Seq tracks for both samples show high RNA-Seq read depth within the tblastn alignment block (17,358,666-17,358,713), consistent with the hypothesis that this region is being transcribed in *D. yakuba*.

5. To examine the region surrounding the start of the tblastn alignment to CDS-1, enter “chr3R:17,358,666” into the “enter position or search terms” text box.
6. Click on the “go” button.
7. Zoom out 3x and another 10x.

In Part 5, we found our tblastn alignment for CDS-1 begins at 17,358,666 when translated in Frame +3 (Figure 32). Examination of this region using the Genome Browser shows us that a start codon at this location is supported by multiple evidence tracks—the BLAT Alignments of NCBI RefSeq Genes and Spaln alignment of *D. melanogaster* proteins, as well as the GeMoMa, Geneid, and Augustus gene predictions (Figure 34).

Since the RNA-Seq read coverage from the adult females and adult males samples extend upstream beyond the start codon, we need to determine if there are alternative start codon candidates available that could serve as the translation initiation site for Rheb-PA.

8. Zoom out 3x and another 10x.

The contiguous RNA-Seq read coverage extends upstream of the start codon at 17,358,666-17,358,668 to position 17,358,293 in the adult males’ sample (Figure 35). Examination of the “Base Position” track shows that the first stop codon upstream from 17,358,666-17,358,668 occurs at 17,358,477-
17,358,479 (remember that we are looking at Frame +3 and that a codon consists of three amino acids). Since there are no alternative start codons in Frame +3 between this stop codon and the start codon at 17,358,666-17,358,668, the most likely translation initiation site for the Rheb-PA ortholog in \textit{D. yakuba} is assigned to the position 17,358,666-17,358,668 on chr3R. Consistent with the gene model for Rheb-RA in \textit{D. melanogaster}, the RNA-Seq read coverage extending upstream of the start codon likely corresponds to the 5' untranslated region (UTRs) of the initial exon of \textit{Rheb}.

![Image of UCSC Genome Browser](image)

\textbf{Figure 35} The contiguous RNA-Seq read coverage extends upstream of the start codon at chr3R:17,358,666-17,358,668 to position 17,358,293 in the adult males’ sample (blue box). Within the region from the start of the RNA-Seq read coverage in the adult males’ sample to just before the start codon (i.e., 17,358,293-17,358,665), the first in-frame stop codon is located at 17,358,477-17,358,479 (red box). Since there are no alternative start codons in Frame +3 between this stop codon and the start codon at 17,358,666-17,358,668, the translation initiation site for the Rheb-PA ortholog in \textit{D. yakuba} is assigned to the position 17,358,666-17,358,668 on chr3R.

\textbf{Part 6.2: Verify stop codon coordinates for Rheb-PA}

In Part 5, we found our \textit{tblastn} alignment for CDS-5 ends at 17,359,559 when translated in Frame +3. The alignment covers all 30 amino acids of CDS-5 and ends with a stop codon (Figure 36).
Figure 36  The **tblastn** alignment of CDS-5 of Rheb-PA (query) against the *D. yakuba* chr3R scaffold (subject) placed the CDS at 17,359,470 – 17,359,559 when the sequence is translated in Frame +3. The **tblastn** alignment covers all 30 amino acids of CDS-5, and it ends with a stop codon (‘*‘; red arrow).

1. To examine the genomic region surrounding the end of the **tblastn** alignment to CDS-5, enter “chr3R:17,359,559” into the “enter position or search terms” text box.
2. Click on the “go” button.
3. Zoom out 3x and another 10x (Figure 37).

The stop codon at 17,359,557-17,359,559 is consistent with the **BLAT** Alignments of NCBI RefSeq Genes, **Spaln** alignment of *D. melanogaster* proteins (stop codons don’t become part of the protein rather they signal when translation should terminate and the newly made protein should be released), and the **GeMoMa** and **Augustus** gene predictions; however, the stop codon is inconsistent with the **Geneid** prediction mRNA_geneid-chr3R_2128, which predicted an intron in this region. However, the coding span for this **Geneid** prediction extends from 17,358,666-17,366,011, which encompasses regions with sequence similarity to the coding exons of the **CRMP** gene. 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 17,359,557-17,359,559, and the last codon (S; Serine) before the stop codon ends at 17,359,556 (Figure 37).

Note that the RNA-Seq read coverage tracks for both samples indicate that transcription extends beyond the stop codon. Based on the gene structure of Rheb-RA in *D. melanogaster*, 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*. 
Figure 37 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 17,359,557 – 17,359,559, and the last codon (S; Serine) before the stop codon ends at 17,359,556.

Part 6.3: Determine phases of donor and acceptor splice sites

During splicing, introns are spliced out (removed) from the pre-mRNA so that adjacent exons are placed next to each other. This means that the ends of an exon do not necessarily correspond to the ends of the complete codon. The number of nucleotides between the last complete codon and the splice donor site is known as the phase of the splice donor site. Similarly, the number of nucleotides between the splice acceptor site and the first complete codon is known as the phase of the splice acceptor site. Because the phases of the splice sites depend on the placement of the complete codon, the phases of the donor and acceptor sites are based on the reading frame of each CDS.

In addition, in order 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. Specifically, the sum of the donor and acceptor phases of adjacent CDS’s must either be 0 (i.e., no additional codon) or 3 (i.e., a complete codon). The use of incompatible splice donor and acceptor sites will introduce a frame shift into the translation of the CDS following the splice acceptor site.

In D. melanogaster, approximately 99% of introns have a GT splice donor site and 1% have a GC non-canonical splice donor site. Almost all introns have an AG splice acceptor site. The GEP comparative annotation protocol posits that all introns have a GT splice donor site and an AG splice acceptor site unless the D. melanogaster gene model uses a non-canonical splice site, or the non-canonical splice site is supported by RNA-Seq data.

See Understanding Eukaryotic Genes Module 4 to further review splicing.
Because the tblastn alignment for CDS-1 of Rheb terminates at 17,358,713 and the alignment includes the last amino acid of the CDS (Figure 32), we expect to find the splice donor site for CDS-1 at around position 17,358,713.

1. To examine the genomic region surrounding the splice donor site of CDS-1, enter “chr3R:17,358,713” into the “enter position or search terms” text box.
2. Click on the “go” button.
3. Zoom out 3x and another 10x to examine the 30 bp surrounding this position.

The GT splice donor site closest to 17,358,713 is located at 17,358,715 – 17,358,716. This splice donor site is in phase 0 relative to Frame +1 (Figure 38, top), in phase 2 relative to Frame +2 (Figure 38, middle), and in phase 1 relative to Frame +3 (Figure 38, bottom). This splice donor site is supported by multiple lines of evidence—BLAT Alignments of NCBI RefSeq Genes, Spaln alignment of D. melanogaster proteins, and the GeMoMa, Geneid, and Augustus gene predictions, as well as the RNA-Seq read coverage from samples of adult females and adult males.

Figure 38  There are three possible phases for the splice donor site (0, 1, or 2), which depend on the reading frame. This splice donor site (at 17,358,715 – 17,358,716) is supported by multiple lines of evidence. Since CDS-1 is in Frame +3 and the last complete codon (GTG which codes for V) is located at 17,358,711 – 17,358,713, there is one nucleotide (G at 17,358,714) between the last complete codon and the splice donor site. Hence, this splice donor site is in phase 1 (bottom).
Based on our tblastn alignment in Part 5, CDS-1 is in Frame +3.

1. Zoom out far enough to see the entire length of CDS-1 and confirm that Frame +3 has an open reading frame (ORF) (i.e., no stop codons are shown within Frame +3 of CDS-1).

Since CDS-1 is in Frame +3, the splice donor site at 17,358,715 – 17,358,716 is in phase 1. This means that the splice acceptor site of CDS-2 must be in phase 2 to maintain the ORF. Based on our tblastn alignment in Part 5, we placed CDS-2 at 17,358,844 – 17,358,912 in Frame +1 (Figure 33).

2. To examine the genomic region surrounding the splice acceptor site of CDS-2, enter “chr3R:17,358,844” into the “enter position or search terms” text box.
3. Click on the “go” button.
4. Zoom out 3x and another 10x to examine the 30 bp surrounding this position (Figure 39).

There is only one potential canonical splice acceptor site (AG) in the 30 bp region surrounding the start of the tblastn alignment to CDS-2. This potential splice acceptor site, located at 17,358,840 – 17,358,841, is supported by the BLAT and Spaln alignments, and the GeMoMa, Geneid, and Augustus gene predictions, as well as the RNA-Seq read coverage from samples of adult females and adult males.

Since CDS-2 is in Frame +1 and the first complete codon (AAA codes for K) is located at 17,358,844 – 17,358,846, there are two nucleotides (GC at 17,358,842 – 17,358,843) between the potential splice acceptor site and the first complete codon. Hence, this splice acceptor site is in phase 2.

---

**Figure 39** This splice acceptor site (at 17,358,840-17,358,841) is supported by multiple lines of evidence. Since CDS-2 is in Frame +1 and the first complete codon (AAA which codes for K) is located at 17,358,844-17,358,846, there are two nucleotides (GC at 17,358,842-17,358,843) between the splice acceptor site and the first complete codon. Hence, this splice acceptor site is in phase 2.
Since the splice donor site for CDS-1 is in phase 1 (relative to Frame +3), the splice acceptor site for CDS-2 must be in phase 2 (relative to Frame +1) to maintain the ORF after the intron has been removed. The extra nucleotides near the splice sites (i.e., G + GC) will form an additional amino acid (Glycine/G) after splicing (Figure 40). Collectively, our analysis suggests that CDS-1 ends at 17,358,714 with a phase 1 splice donor site and CDS-2 begins at 17,358,842 with a phase 2 splice acceptor site.

![Diagram](image)

**Figure 40** 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 41).

![Table](image)

**Figure 41** Summary of the phases of each splice donor and acceptor site in Rheb-RA.

**Part 6.4: Use spliced RNA-Seq reads to verify coordinates for first intron**

1. Under the “RNA Seq Tracks,” change the “Splice Junctions” track to “pack.”
   - Note: The “Splice Junctions” track identifies the splice junctions predicted by the `regtools` program.
2. Click on the “refresh” button.
3. To examine the region surrounding Intron-1 (i.e., intron between CDS-1 and CDS-2), enter “chr3R:17,358,715-17,358,841” into the “enter position or search terms” text box.
• Note: We found these coordinates in Part 6.3. Since our analysis in Part 6.3 suggested that CDS-1 ends at 17,358,714, Intron-1 should begin at 17,358,715; since we found that CDS-2 begins at 17,358,842, Intron-1 should end at 17,358,841.

4. Click on the “go” button then zoom out 3x to examine the 381 bp surrounding this position (Figure 42).

See **Understanding Eukaryotic Genes Module 4** to further review splice junction predictions.

Figure 42  The splice junction JUNC00073911 predicted by regtools connects CDS-1 with CDS-2. The predicted splice donor and acceptor sites are indicated by the green and blue arrows, respectively.

There is only one splice junction predicted in this region.

5. To examine the splice donor site predicted by the splice junction JUNC00073911, zoom into the region surrounding the beginning of the intron predicted by this junction (Figure 42, green arrow).
6. To examine the splice acceptor site predicted by the splice junction JUNC00073911, zoom into the region surrounding the end of the intron predicted by this junction (Figure 42, blue arrow).

The splice junction between the phase 1 donor site of CDS-1 and the phase 2 acceptor site of CDS-2 is supported by the BLAT and Spaln alignments, the GeMoMa, Geneid, and Augustus gene predictions, and the RNA-Seq data from samples of adult females and males.

7. We can gather additional evidence to support this splice junction by clicking on “JUNC00073911” and then examining the “Score” field (Figure 43, top).

The score tells us how many spliced RNA-Seq reads there are that support a predicted splice junction. Since JUNC00073911 has a score of 1694, that splice junction prediction is supported by 1,694 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, JUNC00073911 will be red in the Genome Browser image since greater than 1,000 spliced RNA-Seq reads support the feature (Figure 43, bottom).

8. Click on the back button of the web browser to return to the Genome Browser image.

Examination of the features in the “Splice Junctions Predicted by regtools using D. yakuba RNA-Seq” track shows that the splice junction prediction JUNC00073911 is consistent with our splice donor site for CDS-1 at 17,358,715 – 17,358,716 and our splice acceptor site for CDS-2 at 17,358,840 – 17,358,841 that we annotated in Part 6.3.
Part 6.5: Use splice junction predictions to verify coordinates for second intron

The same annotation strategy can be used to determine the coordinates for Intron-2 between CDS-2 and CDS-3.

1. To examine the region surrounding Intron-2, enter “chr3R:17,358,912-17,359,013” into the “enter position or search terms” text box.
   • Note: These coordinates can be found in the table in Figure 33.
2. Click on the “go” button.
3. Zoom out 3x to examine the 306 bp surrounding this position (Figure 44).

![Figure 44](image)

There are two splice junctions predicted in this region, JUNC00073912 and JUNC00073913.

The `tblastn` alignment for CDS-2 ends at 17,358,912 (Figure 33). The potential splice donor site at 17,358,914-17,358,915 for CDS-2 is supported by the splice junction predictions JUNC00073912 and JUNC00073913, the `BLAT` and `Spaln` alignments, the `GeMoMa`, `Geneid`, and `Augustus` gene predictions, and the RNA-Seq data from samples of adult females and males. There is one nucleotide (A at 17,358,913) 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 17,359,013-17,359,216 in Frame +2 (Figure 33). The potential splice acceptor site at 17,359,009-17,359,010 for CDS-3 is supported by the splice junction prediction JUNC00073912, the `BLAT` and `Spaln` alignments, the `GeMoMa`, `Geneid`, and `Augustus` gene
predictions, and the RNA-Seq data from samples of adult females and males. Note that this potential splice acceptor site is not supported by JUNC00073913. There are two nucleotides (CC at 17,359,011-17,359,012) 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.

4. Click on the “JUNC00073912” feature to determine the number of spliced RNA-Seq reads that support this splice junction prediction (Figure 45).

5. Click on the back button of the web browser to return to the Genome Browser image.

![Figure 45](image)

The score for JUNC00073912 shows that this junction is supported by 1,511 spliced RNA-Seq reads.

In addition to the splice junction JUNC00073912, which supports the proposed splice acceptor site for CDS-3 at 17,359,009-17,359,010, there is another splice junction which suggests a different splice acceptor site (JUNC00073913) (Figure 46).

CDS-3 in *D. yakuba* includes two methionine in Frame +2 (at 17,359,112-17,359,114 and 17,359,205-17,359,207). Hence, the splice junction JUNC00073913 could indicate the presence of a novel isoform of *Rheb* in *D. yakuba* where the translation initiation site is located within CDS-3, and the region from the end of the splice junction (i.e., 17,358,978) to the start codon (e.g., 17,359,111) would be part of the 5’ UTR.

To assess whether there is sufficient evidence to propose a novel isoform of *Rheb* in *D. yakuba*, we will need to determine the number of spliced RNA-Seq reads that support the splice junction JUNC00073913.

6. Click on the splice junction “JUNC00073913” and then examine the “Score” field (Figure 47).

7. Click on the back button of the web browser to return to the Genome Browser image.

Since the splice junction JUNC00073913 is only weakly supported by the available RNA-Seq data, there is insufficient evidence to postulate a novel isoform of *Rheb* in *D. yakuba* based on this splice junction prediction. A score filter can be applied to the “Splice Junctions” evidence track to hide the splice junction predictions that are supported by a small number of spliced RNA-Seq reads.

8. Scroll down to the “RNA Seq Tracks” section and then click on the “Splice Junctions” link.

9. Change the “Show only items with score at or above” to “10.”

10. Click on the “Submit” button (Figure 48).
Figure 46  Splice junction prediction JUNC00073913 proposes an alternative splice acceptor site at 17,358,976-17,358,977. Since CDS-3 in *D. yakuba* includes two methionine in Frame +2, JUNC00073913 could indicate the presence of a novel isoform of *Rheb* in *D. yakuba* that is not found in *D. melanogaster*.

Figure 47  The score for JUNC00073913 shows that this junction is only supported by 5 spliced RNA-Seq reads.

Figure 48  Configure the Splice Junctions track to only display splice junctions that are supported by at least 10 spliced RNA-Seq reads.

11. To examine the intron between CDS-2 and CDS-3 after applying the filter, enter “chr3R:17,358,914-17,359,010” into the “enter position or search terms.”
12. Click on the “go” button.
13. Zoom out 1.5x.

After applying to filter, we no longer see splice junction prediction JUNC00073913 in the Genome Browser image.
Part 6.6: Use splice junction predictions to verify coordinates for third intron

The **tblastn** alignment for CDS-3 spans 17,359,013-17,359,216 in Frame +2 (Figure 33).

1. To examine the region surrounding the splice donor site for CDS-3, enter “**chr3R:17,359,216**” into the “enter position or search terms” text box.
2. Click on the “go” button.
3. Zoom out 3x and another 10x (Figure 49).

![Image of UCSC Genome Browser](image)

**Figure 49** The potential splice donor site at 17,359,219-17,359,220 for CDS-3 is supported by the **BLAT** and **Spaln** alignments, the **GeMoMa**, **Geneid**, and **Augustus** gene predictions, the RNA-Seq data from samples of adult females and males, and the splice junction prediction JUNC00073915. There are two nucleotides (TA at 17,359,217-17,359,218) between the last complete codon (AAA) and the potential splice donor site. Hence, this splice donor site is in phase 2 relative to Frame +2.

4. To examine the genomic region surrounding the start of the **tblastn** alignment to CDS-4 (Figure 33), enter “**chr3R:17,359,279**” into the “enter position or search terms” text box.
5. Click on the “go” button.
6. Zoom out 3x and another 10x (Figure 50).

The genomic region also contains an additional splice junction prediction JUNC00073914. Since this feature is in blue, this junction is supported by 10–49 spliced RNA-Seq reads. By contrast, the splice junction prediction JUNC00073915 is in red, indicating that the splice junction is supported by more than 1,000 spliced RNA-Seq reads (Figure 51).

7. Click on the “**JUNC00073914**” feature to learn more about this splice junction prediction.
Figure 50  The potential splice acceptor site at 17,359,276-17,359,277 for CDS-4 is supported by the BLAT and Spaln alignments, the GeMoMo, Geneid, and Augustus gene predictions, the RNA-Seq data from samples of adult females and males, and the splice junction prediction JUNC00073915. There is one nucleotide (T at 17,359,278) between the first complete codon (GTA) and the potential splice acceptor site. Hence, this splice acceptor site is in phase 1 relative to Frame +1. This phase 1 splice acceptor site is compatible with the phase 2 splice donor site for CDS-3.

Figure 51  The splice junction prediction JUNC00073914 is supported by 18 spliced RNA-Seq reads. However, the splice junction is in the minus strand whereas the Rheb gene is on the plus strand relative to D. yakuba chr3R scaffold.
Because the *Rheb* ortholog is in the plus strand relative to *D. yakuba* chr3R and the splice junction JUNC00073914 is in the minus strand, this splice junction is unlikely to be an intron for the *Rheb* gene (Figure 51). This splice junction could be caused by RNA-Seq reads that have been misplaced, or it could have been derived from a non-coding RNA that overlaps with *Rheb*.

Based on the low score (18) of the JUNC00073914 splice junction, it being in the minus strand, requiring the use of non-canonical splice donor and acceptor sites, and truncating conserved amino acids in CDS-3 and CDS-4, the splice acceptor site for CDS-4 is assigned to 17,359,276-17,359,277.

**Part 6.7: Use splice junction predictions to verify change in donor site sequence for fourth intron**

Return to the [Gene Record Finder](#) record for the *D. melanogaster Rheb* gene.

The “Introns with Non-canonical Splice Sites” panel shows that the intron with the FlyBase ID “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 (Figure 52).

Figure 52  The “Introns with Non-canonical Splice Sites” panel of the Gene Record Finder shows that the intron between the transcribed exons Rheb:6 and Rheb:7 begins with a non-canonical GC splice donor site in *D. melanogaster*. 
1. To determine the IDs of the CDS’s that overlap with the transcribed exons Rheb:6 and Rheb:7, click on the Genome Browser image in the “mRNA Details” panel of the Gene Record Finder.

Comparison of the “FlyBase Transcribed Exons” and “FlyBase Coding Exons” tracks in the Genome Browser for *D. melanogaster* shows that the coding exon CDS-4 (Rheb:4_9845_1) overlaps with the transcribed exon Rheb:6, and coding exon CDS-5 (Rheb:5_9845_0) overlaps with the transcribed exon Rheb:7.

2. Zoom in to the end of CDS-4 to verify that it has a non-canonical GC donor site (Figure 53).

Figure 53 Examination of the region surrounding the end of CDS-4 in *D. melanogaster* shows that it has a non-canonical GC splice donor site.

Consequently, if the splice sites are conserved between *D. melanogaster* and *D. yakuba*, then the splice donor site for CDS-4 would also use a non-canonical GC donor site in *D. yakuba*.

3. Go back to the web browser tab with the Genome Browser for *D. yakuba*.
The `tblastn` alignment for CDS-4 ends at 17,359,407 in Frame +1 (Figure 33).

4. To examine the genomic region surrounding the end of the `tblastn` alignment to CDS-4, enter “chr3R:17,359,407” into the “enter position or search terms” text box.

5. Click on the “go” button.

6. Zoom out 3x and another 10x (Figure 54).

7. To examine the genomic region surrounding the start of the `tblastn` alignment to CDS-5, enter “chr3R:17,359,470” into the “enter position or search terms” text box.

8. Click on the “go” button.

9. Zoom out 3x and another 10x (Figure 55).
The potential splice acceptor site at 17,359,468-17,359,469 for CDS-5 is supported by the BLAT and Spaln alignments, the GeMoMa and Augustus gene predictions, the RNA-Seq data from samples of adult females and males, and the splice junction prediction JUNC00073917. The Geneid prediction mRNA_geneid-chr3R_2128 is inconsistent with the potential splice acceptor site. However, per the previous analysis described in Part 6.2, this Geneid prediction might have combined two adjacent genes (Rheb and CRMP) into a single prediction. There are no nucleotides between the potential splice acceptor site and the first complete codon (TCC). Hence, this splice acceptor site is in phase 0 relative to Frame +3. This phase 0 splice acceptor site is compatible with the phase 0 splice donor site for CDS-4.

Taken together, the analysis described above would produce the CDS coordinates for Rheb-PA in *D. yakuba* chr3R summarized below (Figure 56):

<table>
<thead>
<tr>
<th>CDS</th>
<th>FlyBase ID</th>
<th>Frame</th>
<th>Splice Acceptor Phase</th>
<th>Coordinates</th>
<th>Splice Donor Phase</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1_9845_0</td>
<td>+3</td>
<td></td>
<td>17,358,666</td>
<td>17,358,714</td>
</tr>
<tr>
<td>2</td>
<td>2_9845_2</td>
<td>+1</td>
<td>2</td>
<td>17,358,842</td>
<td>17,358,913</td>
</tr>
<tr>
<td>3</td>
<td>3_9845_2</td>
<td>+2</td>
<td>2</td>
<td>17,359,011</td>
<td>17,359,218</td>
</tr>
<tr>
<td>4</td>
<td>4_9845_1</td>
<td>+1</td>
<td>1</td>
<td>17,359,278</td>
<td>17,359,407</td>
</tr>
<tr>
<td>5</td>
<td>5_9845_0</td>
<td>+3</td>
<td>0</td>
<td>17,359,470</td>
<td>17,359,556</td>
</tr>
</tbody>
</table>

**Figure 56** Summary of the five CDS’s in Rheb-PA. Stop codon is located at 17,359,557-17,359,559.
Part 7: Verify and submit gene model(s)

Now that we’ve completed the annotation of Rheb-PA in *D. yakuba*, we need to verify our proposed gene model and prepare the corresponding sequence files for submission.

Part 7.1: Verify gene model of Rheb-PA

Our analysis of the CDS-by-CDS *tblastn* alignments and the evidence tracks on the Genome Browser allow us to precisely define the start and end positions of each of the five coding exons (CDS’s) of Rheb-PA. To verify that our proposed gene model satisfies the basic biological constraints (e.g., begins with a start codon, has compatible splice sites, and ends with a stop codon), we will check our gene model coordinates using the Gene Model Checker.

1. Open a new web browser tab and navigate to the Gene Model Checker (Figure 57).
2. In the “Project Details” section:
   - Species Name: select “D. yakuba”
   - Genome Assembly: select “May 2011 (WUGSC dyak_caf1/DyakCAF1)”
   - Scaffold Name: enter “chr3R”
3. In the “Ortholog Details” section:
   - Ortholog in *D. melanogaster*: enter “Rheb-PA”
4. In the “Model Details” section:
   - Errors in Consensus Sequence?: select “No”
   - Coding Exon Coordinates: enter a comma-delimited list of coordinates for the five CDS’s “17358666-17358714, 17358842-17358913, 17359011-17359218, 17359278-17359407, 17359470-17359556”
   - Annotated Untranslated Regions?: select “No”
   - Orientation of Gene Relative to Query Sequence: select “Plus” since Rheb-PA is on the positive strand relative to chr3R scaffold
   - Completeness of Gene Model Translation: select “Complete”
   - Stop Codon Coordinates: click within the textbox and the coordinates will automatically populate
5. Click on the “Verify Gene Model” button to run the Gene Model Checker.

Note that the coordinates for the “Coding Exon Coordinates” field **do not include the stop codon**. We will enter the stop codon coordinates separately in the “Stop Codon Coordinates” field.
Figure 57  Verify the D. yakuba gene model for Rheb-PA using the Gene Model Checker.

Once the analysis is complete, the right panel of the Gene Model Checker contains the results. The “Checklist” tab enumerates the list of criteria that have been checked by the Gene Model Checker (Figure 58). For example, the Gene Model Checker verifies that our proposed gene model begins with a start codon and ends with a stop codon. It also verifies that the splice junctions contain the standard splice donor and acceptor sites. Some of the items on the checklist have been skipped because they do not apply to a complete gene (e.g., Acceptor for CDS 1).

In addition to verifying the basic gene structure, the Gene Model Checker also compares the proposed gene model against the putative D. melanogaster ortholog using a protein alignment and a dot plot.

6. Click on the “Dot Plot” tab to examine the dot plot between the D. melanogaster protein (x-axis) and the protein sequence for the submitted model in D. yakuba (y-axis) (Figure 59).

The alternating color boxes in the dot plot correspond to the different CDS’s in the two sequences. Dots in the dot plot correspond to regions of similarity between the D. melanogaster protein and the submitted D. yakuba gene model. If the submitted sequence is identical to the D. melanogaster ortholog, then the dot plot will show a straight diagonal line with a slope of 1. Changes in the size of the submitted model compared to the D. melanogaster ortholog will alter the slope of this line. In this case, the dot plot shows that the five CDS’s of Rheb-PA in D. melanogaster and D. yakuba have similar lengths (compare the length shown on the x-axis to the length shown on the y-axis for each
CDS). However, within a small region of CDS-4, the dot plot did not detect sequence similarity between the submitted model for *D. yakuba* and the *D. melanogaster* ortholog (Figure 59).

Figure 58 The Gene Model Checker checklist shows that the proposed gene model for Rheb-PA satisfies the biological constraints of most protein-coding genes (e.g., canonical start codon, stop codon, splice sites).

Figure 59 The dot plot comparing the *D. melanogaster* Rheb-PA (x-axis) with the submitted gene model in *D. yakuba* (y-axis) shows that the main differences between the two protein sequences are in CDS-4 and CDS-5.
To further investigate the dot plot, we will examine the protein alignment between the two sequences.

7. Click on the “View protein alignment” link above the dot plot (Figure 60).

![View protein alignment](image)

Figure 60 The dot plot comparing the *D. melanogaster* Rheb-PA (x-axis) with the submitted gene model in *D. yakuba* (y-axis) shows that the main differences between the two protein sequences are in CDS-4 and CDS-5.

The alignment shows the comparison of the *D. melanogaster* protein against the conceptual translation for the submitted *D. yakuba* gene model. Like the dot plot, alternating colors correspond to the different CDS’s (Figure 61).

The protein alignment shows that the five CDS’s have high levels of sequence similarity between the *D. melanogaster* ortholog and the *D. yakuba* gene model. The symbols in the match line denote the level of similarity (“*” indicates conserved amino acids, “;” denotes amino acids with highly similar chemical properties). Hence, the tiny gap in the dot plot within CDS-1 can be attributed to similar, but not identical, amino acids near its center.

![Figure 61](image)

Figure 61 The protein alignment between *D. melanogaster* Rheb-PA versus the submitted gene model in *D. yakuba* shows that the tiny gap within CDS-4 in the dot plot can be attributed to three amino acid residues that differ between these two species (red boxes).

The protein alignment between the *D. melanogaster* Rheb-PA and the submitted gene model in *D. yakuba* shows that the gap within CDS-4 in the dot plot can be attributed to differences in three amino acid residues. We need to verify the placement of CDS-4.
8. Click on the “Checklist” tab.
9. Click on the magnifying glass icon next to the “Acceptor for CDS 4” entry (Figure 62).
   • Note: A new window containing the Genome Browser will appear with our submitted gene model shown in the red “Custom Gene Model” track.

![Checklist](image)

Figure 62  To verify the placement of CDS-4, click on the magnifying glass icon next to the “Acceptor for CDS 4” entry.

10. Zoom out about 1.5x and drag the Genome Browser image to the left so that we can examine the genomic region surrounding CDS-4 of Rheb-PA in Frame +1.

The placement of CDS-4 in the Custom Gene Model for Rheb-PA is in congruence with the BLAT and Spaln alignments, the GeMoMa, Geneid, and Augustus gene predictions, and the RNA-Seq data from samples of adult females and males (Figure 63). The red arrows in Figure 63 demarcate the three amino acid residues that differ between CDS-4 in *D. yakuba* and *D. melanogaster*.

![UCSC Genome Browser](image)

Figure 63  The placement of CDS-4 in the Custom Gene Model for Rheb-PA is in congruence with multiple lines of evidence. The red arrows demarcate the three amino acid residues that differ between CDS-4 in *D. yakuba* and *D. melanogaster*. 
To verify our previous observations, we can view the submitted gene model within the context of the other evidence tracks in the Genome Browser.

11. Zoom out until we see our entire submitted gene model for Rheb-PA in *D. yakuba* (Figure 64).

![Figure 64](image)

**Figure 64** Our submitted gene model for Rheb-PA in *D. yakuba* is shown under the track title “Custom Gene Model for DyakCAF1.”

### Part 7.2: Download files required for project submission

In addition to the Pathways Project Annotation Report, you must prepare three additional data files to submit a project to the GEP – a General Feature Format File (GFF), a Transcript Sequence File (fasta), and a Peptide Sequence File (pep). The Gene Model Checker automatically creates these three files for a specific isoform (e.g., Rheb-PA) when you verify a gene model.

1. You can download these files by clicking on the “Downloads” tab and then clicking on each of the links to save each file to your computer (Figure 65).
   - Note: You don’t have to rename these files, you’ll only need to rename the merged file.

![Figure 65](image)

**Figure 65** In preparation for project submission, click on the “Downloads” tab and save the GFF, transcript sequence, and peptide sequence files for the gene model to your computer.
Part 7.3: Verify gene model of remaining isoform(s)

Recall that \textit{Rheb} in \textit{D. yakuba} has two isoforms; the files we just downloaded were for the Rheb-PA isoform. Since the coding exons (CDS’s) for both of our isoforms are identical, we only need to change the name of the ortholog in the Gene Model Checker and then verify our second isoform.

1. Change the name of the ortholog in the “Ortholog in \textit{D. melanogaster}” field to “Rheb-PB” (Figure 66).
   - Leave everything else the same as we had it for Rheb-PA. Note: If the CDS’s in our project weren’t identical, we’d need to change out the coordinates in addition to the “Ortholog in \textit{D. melanogaster}” field.
2. Click on “Verify Gene Model” to create a new set of files for Rheb-PB.

![Figure 66](Image) Change the name of the ortholog in the “Ortholog in \textit{D. melanogaster}” field to “Rheb-PB” and then click on “Verify Gene Model.”

We can check that the Rheb-PB isoform was verified by looking at the axes’ titles under the “Dot Plot” tab.

3. Click on the “Downloads” tab and then click on each of the links to save each file to your computer.

Part 7.4: Merge project files

Now we need to combine the GFF, transcript sequence, and peptide sequence files for all our isoforms into a single file prior to project submission (we’ll submit one merged file for each file type).

1. Open a new web browser tab and navigate to the \textit{Annotation Files Merger}.

Let’s merge our two isoform GFF files first.
2. Change the “File Type:” to “GFF Files (.gff).”
3. Drag the two GFF files we downloaded for Rheb-PA and Rheb-PB to the “Drag and drop the files you want to merge here” section.
4. Click on the “Merge Files” button (Figure 67).

5. Download the merged GFF file by right-clicking (control click on macOS) on the “Merged File Link” (Figure 68).
6. Click on “Save Link As...”
7. Enter “dyak_Rheb.gff” as the file name.
8. Once you click on the “Save” button, the merged GFF file should download onto your computer.
Now we need to merge our two isoform Transcript Sequence Files (.fasta).

9. Click on the “Merge another set of files” button.
10. Change the “File Type:” to “Transcript Sequence Files (.fasta).”
11. Drag the two fasta files we downloaded for Rheb-PA and Rheb-PB to the “Drag and drop the files you want to merge here” section.
12. Repeat steps 4 – 6.
13. Enter “dyak_Rheb.fasta” as the file name.
14. Once you click on the “Save” button, the merged fasta file should download onto your computer.

Lastly, we need to merge our two isoform Peptide Sequence Files (.pep).

15. Click on the “Merge another set of files” button.
16. Change the “File Type:” to “Peptide Sequence Files (.pep).”
17. Drag the two pep files we downloaded for Rheb-PA and Rheb-PB to the “Drag and drop the files you want to merge here” section.
18. Repeat steps 4 – 6.
19. Enter “dyak_Rheb.pep” as the file name.
20. Once you click on the “Save” button, the merged pep file should download onto your computer.

Appendix

Appendix A. Investigate the other tblastn alignments to D. yakuba chr3R

The “Protein Domains” section of the FlyBase Polypeptide Report for Rheb-PA (FBpp0078342) shows a conserved Small_GTPase domain at residues 8-164 aa of the protein.
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 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).

Examination of the *tblastn* alignment blocks also shows three cases where the alignments contain in-frame stop codons (i.e., at 9,373,502-9,373,062, 9,658,240-9,658,575, and 26,681,803-26,681,411). The in-frame stop codons can likely be attributed to homologous overextension of the alignments into the introns.

**Appendix B. Accession Numbers**

Each sequence record in GenBank has a sequence version number consisting of an accession number followed by a dot and a version suffix (e.g., CM000160.2). 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.

The naming conventions for the scaffold names in the GEP UCSC Genome Browser may vary depending on the *Drosophila* species. Scaffold names for the Comparative Analysis Freeze 1 (CAF1) assemblies (e.g., *D. yakuba*) and the modENCODE assemblies follow the naming conventions established by the *Drosophila* 12 Genomes Consortium and the modENCODE project, respectively. For the remaining species (e.g., *D. pseudoobscura*), the scaffold names correspond to the GenBank accession number for the scaffold (e.g., CM000070).

The Genome Browser also supports navigation by the sequence version number. When you enter the sequence version number (e.g., CM000160.2) into the "Position/Search Term" text box, the Genome Browser searches the "INSDC" ([International Nucleotide Sequence Database Collaboration](https://www.ncbi.nlm.nih.gov/Insdc/)) and the
"RefSeq Acc" evidence tracks for an exact match to the sequence version number. If there is a match, the Genome Browser will substitute the sequence version number with the corresponding scaffold name and navigate to the region.

For example, the CM000160.2 sequence version number corresponds to the scaffold name chr3R in the D. yakuba DyakCAF1 assembly. When you search for "CM000160.2:17,358,666-17,359,556" in the D. yakuba DyakCAF1 assembly, the Genome Browser will substitute "CM000160.2" with "chr3R". Hence either "CM000160.2:17,358,666-17,359,556" or "chr3R:17,358,666-17,359,556" could be entered in the "Position/Search Term" text box, and we would still navigate to the same genomic region.

By default, the Genome Browser will display the evidence track which contains the feature that matches the search term. For example, when you search for the sequence version number "CM000160.2", the Genome Browser will display the "CM000160.2" feature in the INSDC evidence track. Since this feature spans the entire length of chr3R, the "CM000160.2" feature will appear as a black bar across the Genome Browser image.

Appendix C. Combining (or Batching) BLAST Searches

In Part 3.2, instead of running four individual blastp searches of the protein IDs of the nearest two neighbors upstream and downstream of the target gene, we could have combined all these searches (i.e., batched them) into a single search. Please note that this time saving method may be confusing for novice annotators.

Enter Accession Numbers (one per line) in the "Enter Query Sequence" text box to run multiple query sequence searches at once.
Top: View of the **blastp** results for XP\_015048368.1 (see the “Query ID” (red arrow) to determine which alignment you are viewing). Click on the “Results for” box (purple arrow) to see a drop-down menu that will allow you to view the other search results.

Bottom: Click on the results for “XP\_002098031.1” (green arrow) to view the sequences that match the closest upstream neighbor of our target gene.

In Part 5, instead of running five individual **tblastn** searches (one for each CDS), we could have entered all five CDS sequences into the “Enter Query Sequence” text box and performed all five searches at once.

Copy and paste the header and sequence for each CDS into the “Enter Query Sequence” text box.