



# Gene Model Checker User Guide

---

Wilson Leung

## Table of Contents

**Introduction** ..... 2

**Acknowledgements**..... 2

**Questions about the Gene Model Checker**..... 2

**Availability**..... 2

**General overview** ..... 3

**Configuring the gene model** ..... 4

    Getting help ..... 4

    Available fields in the *Gene Model Checker* configuration panel ..... 4

**Gene Model Checker results** ..... 6

    Checklist ..... 6

    Dot Plot..... 7

    Transcript Sequence ..... 7

    Peptide Sequence ..... 7

    Extracted Coding Exons..... 8

    Downloads..... 8

**Using the Gene Model Checker to identify annotation errors**..... 9

    Submit the initial gene model..... 9

    Examine the Checklist to identify problems in our initial gene model ..... 10

    Use the Surrounding Sequence feature to identify problems with our gene model ..... 11

    Use the Custom Track feature to identify problems with our gene model..... 13

    Use *tblastn* searches and the Genome Browser to identify problems with our gene model..... 15

    Examine the dot plot and alignment of our gene model against the *D. melanogaster* ortholog..... 24

    Use dot plots to identify large insertions and deletions in the proposed gene model ..... 27

    Download the files generated by the *Gene Model Checker* ..... 31

**Verifying gene models with consensus errors**..... 32

    Check the original *tgo* gene model in *Drosophila sechellia* ..... 32

    Check the *tgo* gene model with the modified consensus sequence ..... 34

## Introduction

Before you can submit an annotation project to the Genomics Education Partnership (GEP; <https://thegep.org>), you must first validate the gene models and generate three additional data files for all the gene models in your project [i.e. a file in the General Feature Format (GFF), a file containing all the transcript sequences, and a file containing all the predicted peptide sequences]. Based on the recommendations from GEP faculty members, we have created the *Gene Model Checker* tool to assist annotators with their gene annotation efforts. This user guide contains an overview of the program and some examples on how to use this program in practice.

## Acknowledgements

The *Gene Model Checker* is developed by Wilson Leung at Washington University in St. Louis for the Genomics Education Partnership (GEP).

## Questions about the *Gene Model Checker*

Please contact Wilson ([wleung@wustl.edu](mailto:wleung@wustl.edu)) if you have any questions or encounter any problems with the *Gene Model Checker*.

## Availability

The *Gene Model Checker* is available under the “**Resources & Tools**” section of the [F Element project page](#) and the [Pathways project page](#) on the GEP website.

## General overview

The *Gene Model Checker* interface is divided into two main panels. The left panel allows you to specify the characteristics of the gene model you would like to verify. The right panel displays the results generated by the *Gene Model Checker*. In order to verify a gene model, you must specify the species, genome assembly, scaffold name, the protein ortholog in *D. melanogaster*, coordinates of the coding exons in a comma-separated list, orientation of the gene relative to the scaffold, and the stop codon coordinates (Figure 1).

**Configure Gene Model**

**Project Details**

Species Name:

Genome Assembly:

Scaffold Name:

**Ortholog Details**

Ortholog in *D. melanogaster*:

**Model Details**

Errors in Consensus Sequence?  Yes  No

Coding Exon Coordinates:

Annotated Untranslated Regions?  Yes  No

Orientation of Gene Relative to Query Sequence:  Plus  Minus

Completeness of Gene Model Translation:  Complete  Partial

Stop Codon Coordinates:

**Checklist** | Dot Plot | Transcript Sequence | Peptide Sequence | Extracted Coding Exons | Downloads


Expand All | Collapse All

View	Criteria	Status	Message
<input type="checkbox"/>	Check for Start Codon	Pass	
<input type="checkbox"/>	Acceptor for CDS 1	Skip	Already checked for Start Codon
<input type="checkbox"/>	Donor for CDS 1	Pass	
<input type="checkbox"/>	Acceptor for CDS 2	Pass	
<input type="checkbox"/>	Donor for CDS 2	Pass	
<input type="checkbox"/>	Acceptor for CDS 3	Pass	
<input type="checkbox"/>	Donor for CDS 3	Skip	Already checked for Stop Codon
<input type="checkbox"/>	Check for Stop Codon	Pass	
<input type="checkbox"/>	Additional Checks	Pass	
<input type="checkbox"/>	Number of coding exons matched ortholog	Pass	

**Figure 1.** The graphical interface for the *Gene Model Checker* is divided into two panels. The left panel is used to specify the parameters for the gene model. The right panel displays the results of the *Gene Model Checker* analysis of the proposed gene model.

## Configuring the gene model

### Getting help

A context-sensitive tooltip will appear when you select each field. These tooltips provide additional information (such as the required format) for each field. As you enter values into each field, the *Gene Model Checker* will validate the value. An exclamation point  icon will appear on the right side of each field that failed validation.

### Available fields in the *Gene Model Checker* configuration panel

Field	Description
<b>Species Name</b>	Select the species for the proposed gene model. The options in this drop-down menu are derived from the list of species available on the <a href="#">GEP UCSC Genome Browser</a> .
<b>Genome Assembly</b>	Select the genome assembly for the proposed gene model. The options in the drop-down menu are derived from the list of assemblies available on the <a href="#">GEP UCSC Genome Browser</a> .
<b>Scaffold Name</b>	Enter the name of the scaffold or annotation project where the proposed gene model is located. This value corresponds to the “Position/Search Term” field in the <i>GEP UCSC Genome Browser</i> .  <b>Examples:</b> contig10, chr3R, scaffold_14906
<b>Ortholog in <i>D. melanogaster</i></b>	The name of the <i>D. melanogaster</i> protein isoform that is the putative ortholog of your gene model.
<b>Errors in Consensus Sequence?</b>	Select “Yes” if there are errors in the project consensus sequence that affect the proposed gene model. Otherwise, select “No”.
<b>File with Changes to the Consensus Sequence</b>	This field will appear when you select the “Yes” option for the “Errors in Consensus Sequence?” field. Provide a <a href="#">Variant Call Format</a> (VCF) file that describes the changes to the consensus sequence. You can use the <a href="#">Sequence Updater</a> tool on the GEP website to create the VCF file.
<b>Coding Exon Coordinates</b>	A comma-delimited list of coordinates that corresponds to the translated exons in your gene model. The complete set of coordinates should encompass the region that begins with the start codon and <b><u>ends just before the stop codon</u></b> .  For example the entry for a gene with three coding exons (CDS) will appear as follows: 25673-25835, 27079-27199, 27285-27468  While you can enter the coordinates in any order, we recommend that you list the coordinates in the order in which the exons are translated (i.e. from 5’ to 3’).

<b>Annotated Untranslated Regions?</b>	Select “Yes” if you have annotated the untranslated regions of the exons (UTRs) in addition to the coding regions. Otherwise, select “No”.
<b>Transcribed Exon Coordinates</b>	<p>This field will appear when you select the “Yes” option for the “Annotated Untranslated Regions?” field.</p> <p>Enter a comma-delimited list of coordinates that correspond to the <b>transcribed exons</b> in your gene model. The set of coordinates should encompass the region that begins with the transcription start site and ends with the transcription end site.</p>
<b>Orientation of Gene Relative to Query Sequence</b>	The orientation of the gene relative to the genomic scaffold or the project sequence (“Plus” or “Minus”).
<b>Completeness of Gene Model Translation</b>	<p>Indicate whether the gene model encompasses the entire coding region.</p> <p><b>Complete:</b> The coordinates in the “Coding Exons Coordinates” field span from the start codon to just before the stop codon.</p> <p><b>Partial:</b> The coordinates in the “Coding Exons Coordinates” field only encompass part of the coding region of the gene.</p>
<b>Region Missing</b>	<p>This field will appear when you select the “Partial” option for the “Completeness of Gene Model Translation” field. Use this field to indicate the region(s) that are missing from the proposed gene model:</p> <p><b>Missing 5’ end of translated region</b> Partial gene without the start codon but contains the stop codon.</p> <p><b>Missing 3’ end of translated region</b> Partial gene without the stop codon but contains the start codon.</p> <p><b>Missing both 5’ and 3’ ends of translated region</b> Partial gene with neither the start codon nor the stop codon.</p>
<b>Phase of First Exon</b>	This field will appear when you indicate that the 5’ end of the gene is missing from the proposed gene model in the “Region Missing” field. Specify the number of bases (0, 1, or 2) before the first complete codon in the first CDS of the proposed gene model.
<b>Stop Codon Coordinates</b>	<p>If this field is empty when you select this field, the <i>Gene Model Checker</i> will infer the stop codon coordinates based on the coding exon coordinates (for complete genes and partial genes with only the 5’ end missing).</p> <p>You can change the coordinates by specifying the stop codon coordinates in the following format (start-end). For example, the entry for a stop codon on the positive strand will appear as follows: 27469-27471.</p> <p>Note that there should be <b>no overlap</b> between the coordinates in the “Stop Codon Coordinates field and the “Coding Exon Coordinates” field.</p>

## Gene Model Checker results

The results panel is divided into six sections (Figure 2). You can click on the tabs to switch to each section. For sections that contain a grid (i.e. **Checklist** and the **Extracted Coding Exons** sections), you can use the buttons labeled “Expand All” and “Collapse All” to control the visibility of all the rows within the grid. The expanded sections will provide you with additional information about a feature (e.g., the sequences surrounding a splice site). Some of the grids have a magnifying glass icon next to each row. Clicking on this icon will bring up the corresponding region in the *GEP UCSC Genome Browser*. You can also right click ([control-click on macOS](#)) on the magnifying glass icon and then open the *GEP UCSC Genome Browser* in a new web browser tab or window.

<span>Checklist</span>   <span>Dot Plot</span>   <span>Transcript Sequence</span>   <span>Peptide Sequence</span>   <span>Extracted Coding Exons</span>   <span>Downloads</span>				
<span>Expand All</span>   <span>Collapse All</span>				
	View	Criteria	Status	Message
		Check for Start Codon	Pass	
		Acceptor for CDS 1	Skip	Already checked for Start Codon
		Donor for CDS 1	Pass	
		Acceptor for CDS 2	Pass	
		Donor for CDS 2	Pass	
		Acceptor for CDS 3	Pass	
		Donor for CDS 3	Skip	Already checked for Stop Codon
		Check for Stop Codon	Pass	
		Additional Checks	Pass	
		Number of coding exons matched ortholog	Pass	

Figure 2. The *Gene Model Checker* results panel.

### Checklist

The checklist contains a list of criteria that were used to verify your gene model. For complete genes, the program checks for the presence of the start and stop codons, canonical donor and acceptor splice sites (i.e. GT, AG), and the presence of stop codons in the translated peptide sequence. To provide additional context, the five bases immediately before and after the start codon, splice donor sites, splice acceptor sites, and stop codons are provided in the expanded output of each row. The nucleotides examined by the *Gene Model Checker* are in red and the surrounding sequences are in blue (Figure 3). This expanded section is useful when you are trying to correct minor errors in the exon or CDS coordinates (e.g., correct off-by-one errors).

		Donor for CDS 1	Pass
<b>Feature Coordinates:</b> 25836-25837 <b>Surrounding Sequence:</b> GAGAGGTACGTA			

Figure 3. The splice donor site is highlighted in red and the surrounding five bases are highlighted in blue in the checklist.

Each checklist item has four possible values:

Status	Explanation
Pass	The submitted model passes the criteria.
Warn	The submitted model contains unusual features (e.g., a non-canonical splice donor site) that requires explanations in the annotation report for the GEP project.
Fail	The submitted model fails the criteria. You should provide a detailed explanation in the annotation report as to why the proposed gene model fails the checklist item.
Skip	The <i>Gene Model Checker</i> did not check this criterion. For example, the <i>Gene Model Checker</i> will skip the check for the acceptor site in the first coding exon if it already checks for the presence of the start codon.

**Note:** There are valid reasons for failing some of the tests on the checklist (e.g., genes with non-canonical splice sites, stop codon readthrough, etc.). However, these exceptions will require detailed explanations in the annotation report for the GEP project.

### Dot Plot

This section contains a graphical dot plot of the peptide sequence from your gene model compared against the orthologous peptide sequence from *D. melanogaster*. It contains a link to the *Dot Plot Viewer* where you can adjust the parameters (e.g., scoring matrix, word size) used to create the dot plot. It also contains a link to the global alignment between the protein sequences derived from your gene model and the orthologous protein from *D. melanogaster*.

### Transcript Sequence

The nucleotide sequences corresponding to the exon coordinates specified in the “Transcribed Exon Coordinates” field are extracted and concatenated together. If the proposed gene model does not include untranslated regions, the coordinates in the “Coding Exon Coordinates” field are used to construct the transcript sequence. If the gene model is located on the minus strand relative to the query sequence, this panel will display the reverse complement of the extracted sequence. This is one of the sequences required for annotation project submission.

### Peptide Sequence

A conceptual peptide translation based on the coordinates specified in the “Coding Exon Coordinates” field. Nucleotide sequences corresponding to the set of coding exons are extracted and concatenated together. The concatenated sequence is translated using the [standard genetic code](#). If the gene model is located on the minus strand relative to the query sequence, the reverse complement of the extracted nucleotide sequence is used to produce the peptide sequence. This is one of the sequences required for annotation project submission.

## Extracted Coding Exons

This section contains all the coding exons and their translations based on the set of coordinates in the “Coding Exon Coordinates” field. This section is particularly useful for identifying coding exons with in-frame stop codons that are usually caused by a pair of incompatible donor and acceptor splice sites.

## Downloads

This section allows you to download the files generated by the *Gene Model Checker*. For each project, you will need to **save all three output files** for each of the isoform you have annotated. In preparation for project submission, you can use the [Annotation Files Merger](#) tool (available on the GEP website) to concatenate the GFF files for all the genes and isoforms in your project into a single project GFF file. Similarly, you can use the *Annotation Files Merger* to combine all the transcript files into a project transcript file, and all the peptide sequence files into a project peptide sequence file.



## Using the *Gene Model Checker* to identify annotation errors

This section will illustrate how the *Gene Model Checker* can be used to identify and correct problems in a gene model. Specifically, we will use the *Gene Model Checker* to verify the coding exon coordinates for Rheb-PA on chr3R of the *Drosophila yakuba* May 2011 (WUGSC dyak\_caf1/DyakCAF1) assembly.

### Submit the initial gene model

Open a web browser window and navigate to the [F Element project page](#) on the GEP website. Click on the “*Gene Model Checker*” link under the “Resources & Tools” section, and then enter the following into the *Gene Model Checker* form (Figure 4).

Field	Value
Species Name	<i>D. yakuba</i>
Genome Assembly	May 2011 (WUGSC dyak_caf1/DyakCAF1)
Scaffold Name	chr3R
Ortholog in <i>D. melanogaster</i>	Rheb-PA
Errors in Consensus Sequence?	No
Coding Exon Coordinates	17358666-17358713, 17358842-17358912, 17359013-17359218, 17359279-17359407, 17359470-17359556
Annotated Untranslated Regions?	No
Orientation of Gene Relative to Query Sequence	Plus
Completeness of Gene Model Translation	Complete
Stop Codon Coordinates	17359557-17359559

Click on the “Verify Gene Model” button to check our gene model.

Figure 4. Enter the initial gene model for *D. yakuba* Rheb-PA into the *Gene Model Checker*

### Examine the Checklist to identify problems in our initial gene model

The checklist reports multiple problems in our proposed gene model: it found in-frame stop codons in CDS 2, 3, and 4. It also cannot locate the canonical splice donor sites for CDS 1 and 2, nor the canonical splice acceptor sites for CDS 3 and 4 (Figure 5). Because errors in earlier parts of the gene model will propagate to the rest of the gene model, we will tackle the problems reported by the *Gene Model Checker* starting at the beginning of the checklist.

View	Criteria	Status	Message
	Check for Start Codon	Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Fail	Found non-canonical sequence GG
	Acceptor for CDS 2	Pass	
	Donor for CDS 2	Fail	Found non-canonical sequence AG
	Acceptor for CDS 3	Fail	Found non-canonical sequence CC
	Donor for CDS 3	Pass	
	Acceptor for CDS 4	Fail	Found non-canonical sequence GT
	Donor for CDS 4	Pass	
	Acceptor for CDS 5	Pass	
	Donor for CDS 5	Skip	Already checked for Stop Codon
	Check for Stop Codon	Pass	
	Additional Checks	Fail	Found premature stop codons in translation
	Check for in-frame stop codons in CDS_1	Pass	
	Check for in-frame stop codons in CDS_2	Fail	Found in-frame stop codons
	Check for in-frame stop codons in CDS_3	Fail	Found in-frame stop codons
	Check for in-frame stop codons in CDS_4	Fail	Found in-frame stop codons
	Check for in-frame stop codons in CDS_5	Pass	
	Length of translated region should be multiples ...	Fail	Length of in-phase coding region: 541 Number of extra nucleotides: 1
	Number of coding exons matched ortholog	Pass	

Figure 5. The *Gene Model Checker* identified multiple problems in our initial gene model for Rheb-PA in *D. yakuba*.

Click on the “Peptide Sequence” tab to open the section. Notice that there are multiple stop codons (asterisks) throughout the translated sequence (Figure 6). In-frame stop codons in the translation can often be attributed to incompatible donor and acceptor splice sites (because it introduces a frame shift in the coding exons downstream of the incompatible splice sites). We can use the checklist to help us identify the splice sites that are incompatible with each other.

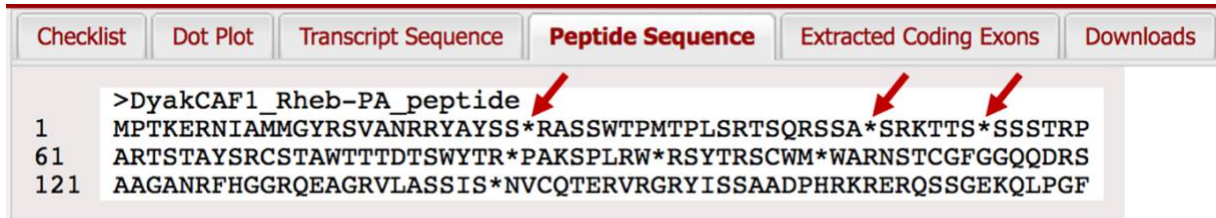


Figure 6. Multiple stop codons (asterisks) in the conceptual peptide translation of our gene model.

### Use the Surrounding Sequence feature to identify problems with our gene model

Click on the “Checklist” tab to return to the checklist section, and then click on the “Expand All” button to examine each of the items in the checklist in detail.

The “Check for Start Codon” checklist item shows that CDS 1 begins with the canonical start codon ATG (methionine), as we would expect for a protein-coding gene. The “Donor for CDS 1” item shows that the two bases after the end of CDS 1 (highlighted in red) is “GG”, instead of the canonical splice donor site GT (Figure 7). However, this “GG” sequence is followed by a “T” (the first nucleotide in blue in Figure 7). If we include the G at 17,358,714 in CDS 1, then the splice donor site at 17,358,715-17,358,716 would have the canonical splice donor site sequence GT. Examining the next item on the checklist, we see that the splice acceptor site for CDS 2 has the canonical sequence AG. Hence examination of the sequences surrounding the end of CDS 1 and the beginning of CDS 2 suggest that we have likely made a mistake when we record coordinates for the end of CDS 1.

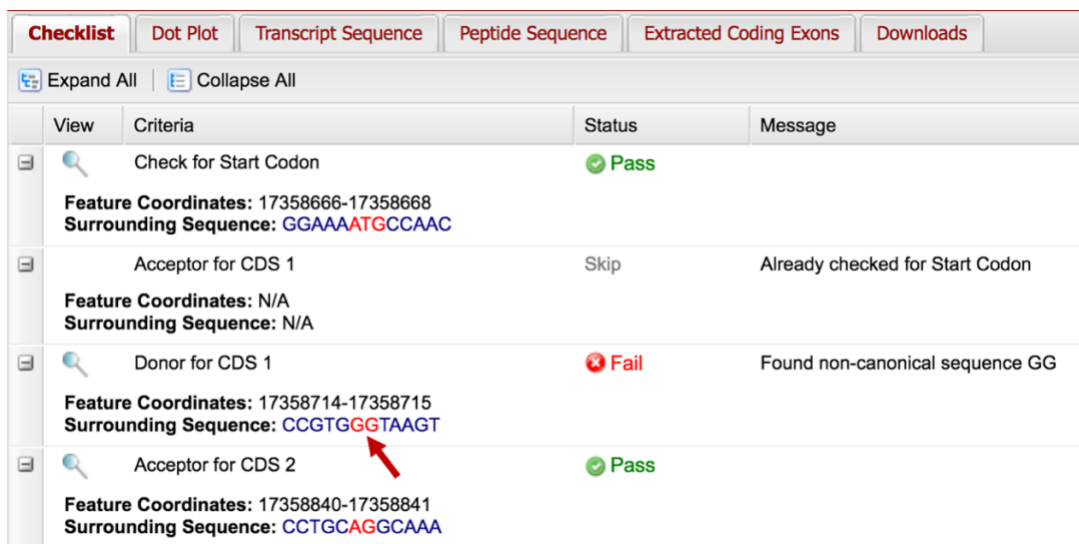


Figure 7. Using the expanded checklist section to identify the incorrect splice donor site for CDS 1.

To verify that the non-canonical splice donor site for CDS 1 has led to the in-frame stop codon in CDS 2, we need to examine the translation of each coding exon. Click on the “Extracted Coding Exons” tab and then click on the “Expand All” button. The “CDS Translation” section for CDS 2 shows an in-frame stop codon in the translation (Figure 8). The fact that CDS 1 is in the correct reading frame while CDS 2 is in the incorrect reading frame suggests that the splice donor site for CDS 1 is incompatible with the splice acceptor site for CDS 2.

Vi...	Exon	Phase	Start	End	Orientation	Length
	CDS_1	0	17358666	17358713	+	48
<b>CDS Translation:</b> MPTKERNIAMMGYRSV						
	CDS_2	0	17358842	17358912	+	71
<b>CDS Translation:</b> ANRRYAYSS*RASSWTPMTPLSR						
	CDS_3	1	17359013	17359218	+	206

Figure 8. The CDS translation for CDS 2 contains an in-frame stop codon (asterisk).

Based on the results of our analysis, we will shift the end coordinate of CDS 1 by one base to 17,358,714. To make this change, go back to the “Configure Gene Model” panel and change the “Coding Exon Coordinates” from “17358666-17358713” to “17358666-17358714” (Figure 9).

GEP Gene Model Checker

**Configure Gene Model**

**Project Details**

Species Name:

Genome Assembly:

Scaffold Name:

**Ortholog Details**

Ortholog in D. melanogaster:

**Model Details**

Errors in Consensus Sequence?  Yes  No

Coding Exon Coordinates:

Figure 9. Change the end coordinate of the first CDS from 17,358,713 to 17,358,714.

Click on the “Verify Gene Model” button to check our revised gene model.

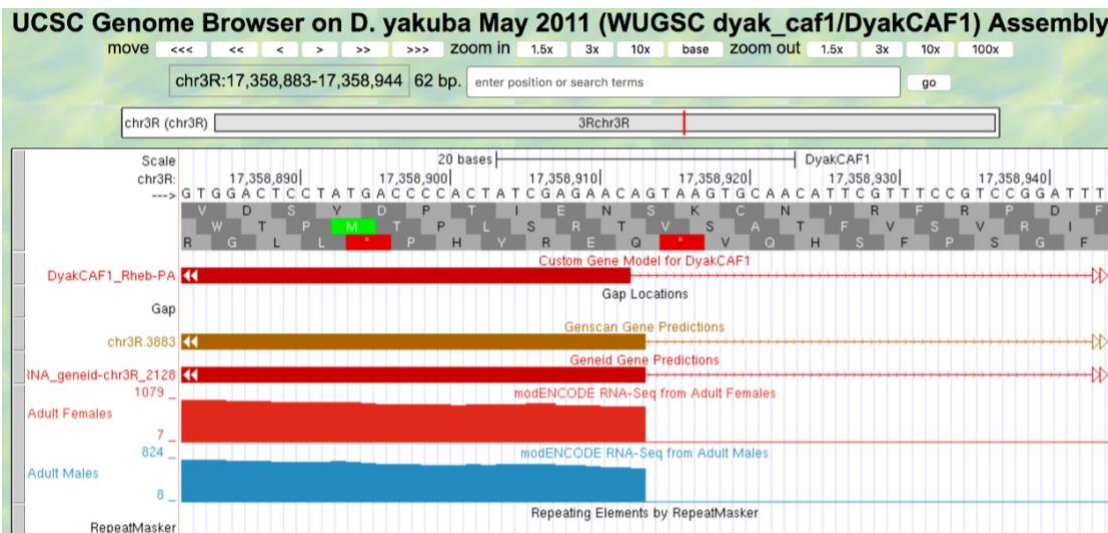
Looking at the updated checklist, we see that the problem with the splice donor site for CDS 1 has been resolved (Figure 10). However, the problem with the splice donor site for CDS 2 remains. Consequently, we will examine the coordinates for CDS 2 next.

View	Criteria	Status	Message
	Check for Start Codon	Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Pass	
	Acceptor for CDS 2	Pass	
	Donor for CDS 2	Fail	Found non-canonical sequence AG
	Acceptor for CDS 3	Fail	Found non-canonical sequence CC
	Donor for CDS 3	Pass	
	Acceptor for CDS 4	Fail	Found non-canonical sequence GT
	Donor for CDS 4	Pass	
	Acceptor for CDS 5	Pass	
	Donor for CDS 5	Skip	Already checked for Stop Codon
	Check for Stop Codon	Pass	
	Additional Checks	Fail	Found premature stop codons in translation
	Check for in-frame stop codons in CDS_1	Pass	
	Check for in-frame stop codons in CDS_2	Pass	
	Check for in-frame stop codons in CDS_3	Pass	
	Check for in-frame stop codons in CDS_4	Pass	
	Check for in-frame stop codons in CDS_5	Fail	Found in-frame stop codons
	Length of translated region should be multiples of 3	Fail	Length of in-phase coding region: 542 Number of extra nucleotides: 2
	Number of coding exons matched ortholog	Pass	

Figure 10. The *Gene Model Checker* checklist after correcting the splice site boundary for CDS 1 shows our gene model contains additional errors.

### Use the Custom Track feature to identify problems with our gene model

We will use the Custom Track feature to diagnose the problem with the splice donor site for CDS 2. Click on the icon next to “Donor for CDS 2” checklist item. A new window will open showing the *GEP UCSC Genome Browser* view of this region (Figure 11).



Custom Gene Model

Figure 11. *GEP UCSC Genome Browser* view of the putative donor site for CDS 2.

We notice from the *UCSC Genome Browser* view that our gene model (shown under the track titled “Custom Gene Model”) differs from the *Genscan* and *Geneid* predictions by a single nucleotide. The splice donor site in the proposed gene model is also inconsistent with the RNA-Seq read coverage from the adult females and adult males samples. To gather additional evidence for the location of the splice donor site, scroll down to the track configuration section and then change the display modes for the following evidence tracks:

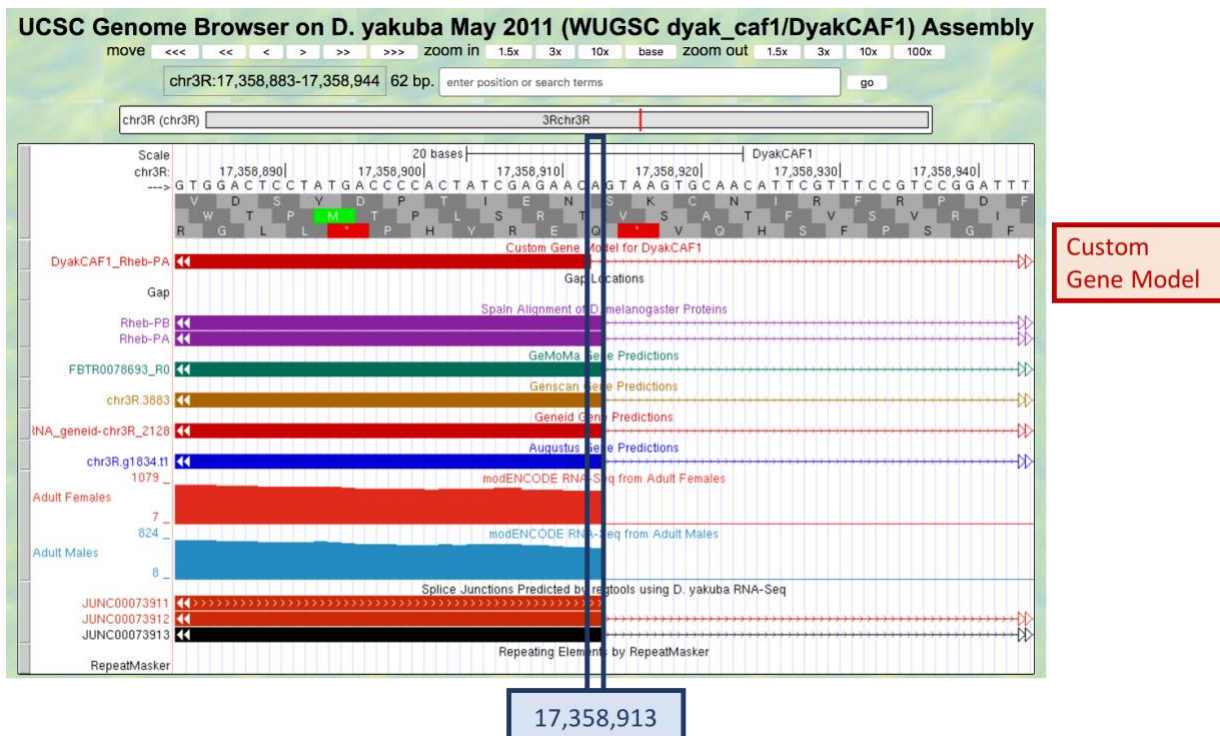
Under “Genes and Gene Prediction Tracks”:

- *D. mel* Proteins: **pack**
- *GeMoMa* Genes: **pack**
- *Augustus*: **pack**

Under “RNA-Seq Tracks”:

- Splice Junctions: **pack**

Click on one of the “refresh” buttons to update the Genome Browser display (Figure 12).



**Figure 12.** The custom gene model placed the end of CDS 2 at 17,358,912, while the protein sequence alignments, multiple gene predictors, and RNA-Seq data placed the end of CDS 2 at 17,358,913.

The *SPALN* alignment to the *D. melanogaster* proteins Rheb-PA and Rheb-PB, predictions from four gene predictors (i.e. *GeMoMa*, *Genscan*, *Geneid*, and *Augustus*), the RNA-Seq read coverage, and the splice junction JUNC00073912 (score = 1511) all support extending the end of CDS 2 to 17,358,913.

To update the proposed gene model, go back to the “Configure Gene Model” panel of the *Gene Model Checker* and change the “Coding Exon Coordinates” from 17358842-17358912 to 17358842-17358913. Click on the “Verify Gene Model” button. The revised gene model gene model has successfully resolved the issue with the splice donor site for CDS 2 (Figure 13).

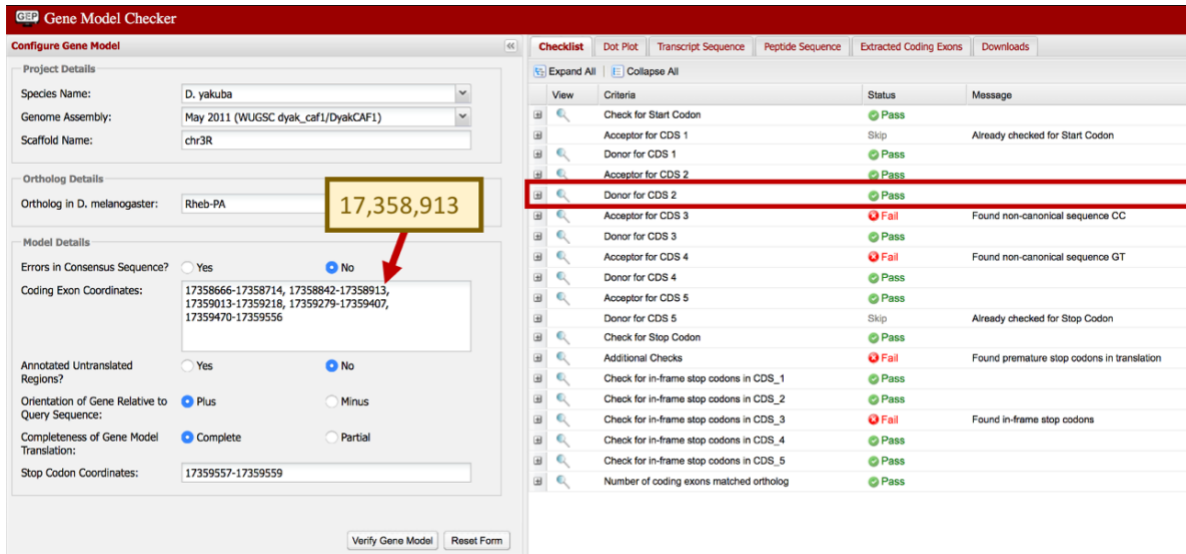


Figure 13. Changing the end coordinate of CDS 2 to 17,358,913 resolves the non-canonical splice donor site at the end of CDS 2 previously reported by the *Gene Model Checker* checklist.

**Use *tblastn* searches and the Genome Browser to identify problems with our gene model**

The *Gene Model Checker* checklist for the revised gene model indicates that the splice acceptor site for CDS 3 has the non-canonical sequence “CC”. This splice acceptor site for CDS 3 is likely incompatible with the splice donor site for CDS 2, leading to the in-frame stop codon within CDS 3 (Figure 14).

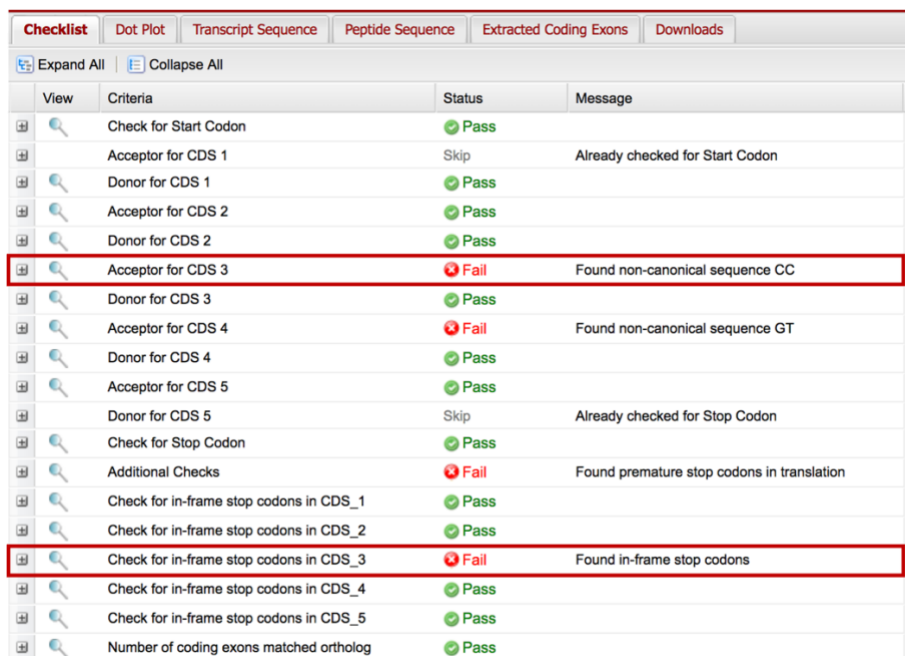


Figure 14. The *Gene Model Checker* checklist reports a non-canonical splice acceptor site and in-frame stop codons in CDS 3.

To diagnose the cause of the error, We will perform *tblastn* searches to verify the reading frames and the locations of CDS 2 and CDS 3. Open a new web browser window and navigate to the [F Element project page](#) on the GEP website. Click on the “Gene Record Finder” link under the “Resources & Tools” section. Enter the name of the gene (*Rheb*) into the text box and then click on the “Find Record” button. The “CDS usage map” under the “Polypeptide Details” section of the *Gene Record Finder* shows that both the A and B isoforms of *Rheb* have the same set of five CDSs (Figure 15).

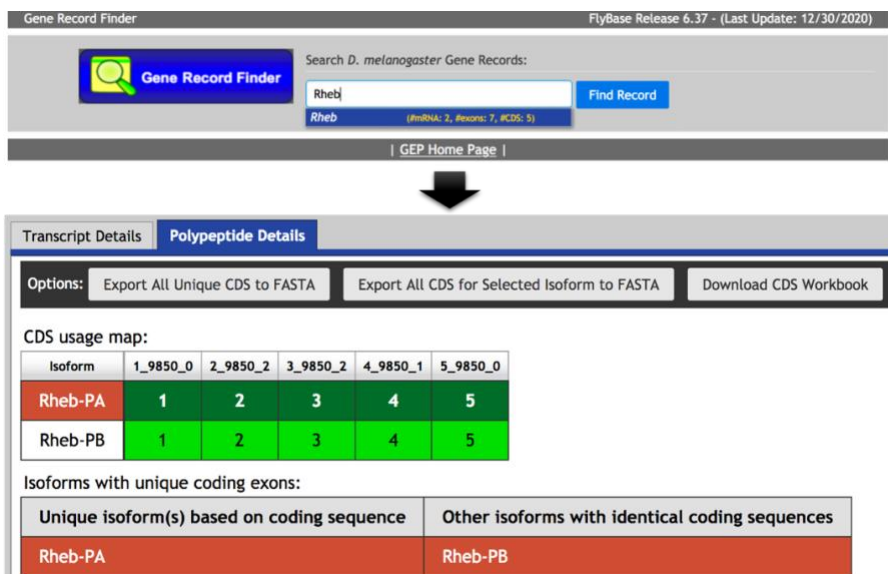


Figure 15. The *Gene Record Finder* record for the *D. melanogaster Rheb* gene shows two isoforms with identical coding sequences.

The CDS table at the bottom of the page shows that CDS 3\_9850\_2 corresponds to the third CDS of *Rheb-PA*. Click on the “3\_9850\_2” row in the CDS table. A “Sequence viewer” window will appear which shows the amino acid sequence for this CDS. Select the sequence (including the header with the > sign), and copy the sequence onto the clipboard.

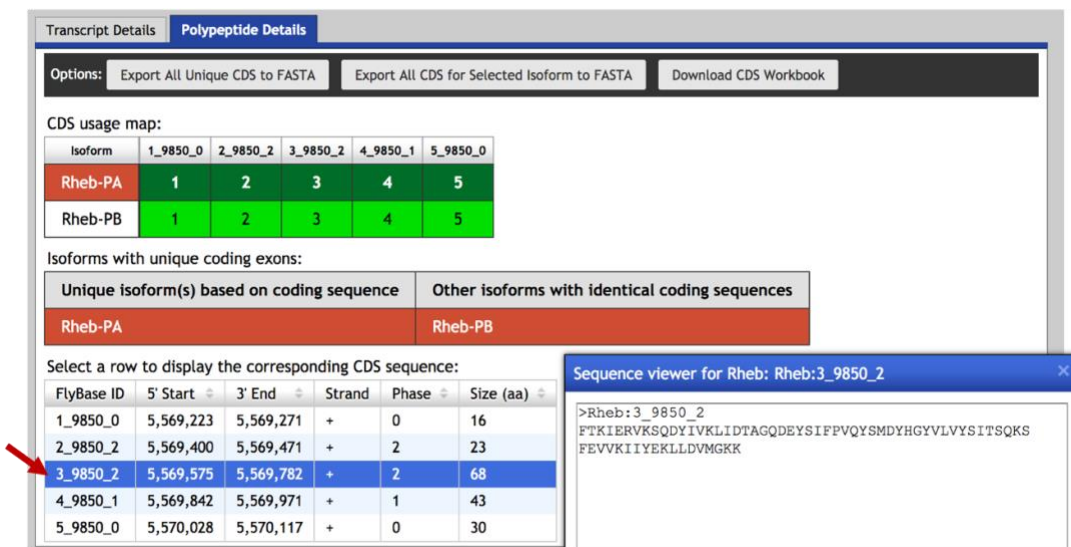


Figure 16. Retrieve the amino acid sequence for CDS 3\_9850\_2 of the *D. melanogaster Rheb* gene from the *Gene Record Finder*.



Open a new web browser window and navigate to the [NCBI BLAST home page](#). Click on the “tblastn” image under the “Web BLAST” section, and then click on the “Align two or more sequences” checkbox. Paste the amino acid sequence for CDS 3\_9850\_2 of the *D. melanogaster* *Rheb* gene into the “Enter Query Sequence” text box (Figure 17).

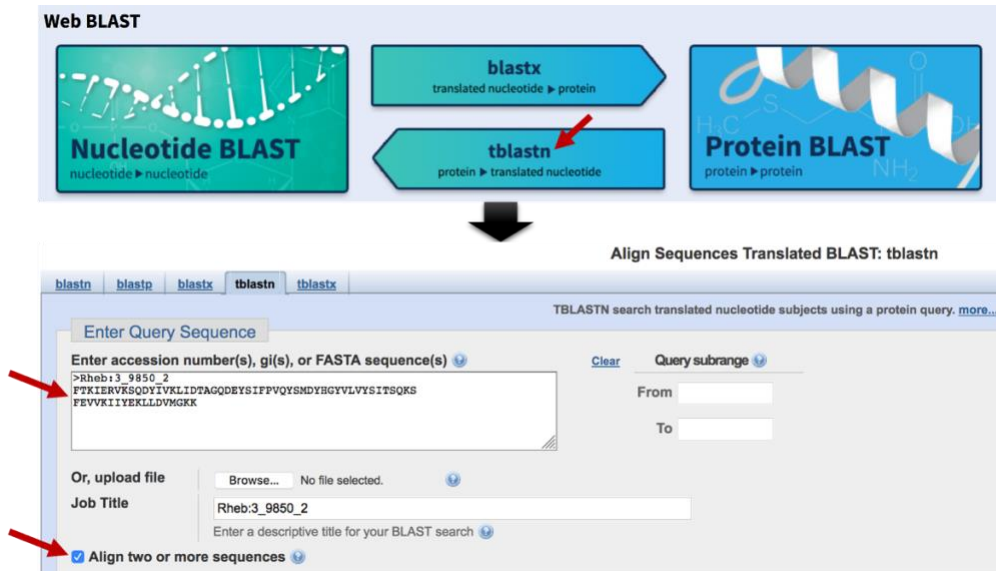


Figure 17. Use the third CDS (3\_9850\_2) of Rheb-PA as the query in the “Align two or more sequences” *tblastn* search.

For the subject sequence of the *tblastn* search, we need to specify the GenBank accession number for the scaffold that corresponds to *D. yakuba* chr3R. Go back to the web browser tab with the *Gene Model Checker* checklist and then click on the icon next to “Acceptor for CDS 3” checklist item. Change the display mode for the “INSDC” track under the “Mapping and Sequencing Tracks” section to **pack**, and then click on one of the “refresh” buttons. The “INSDC” (i.e. International Nucleotide Sequence Database Collaboration) evidence track shows that the GenBank accession number for *D. yakuba* chr3R is **CM000160.2** (Figure 18). In addition, since the coding regions of the *D. yakuba* *Rheb* ortholog is located within the region at 17,358,666-17,359,559 on chr3R, we will limit the search region for our *tblastn* search from 17,300,000 to 17,400,000.

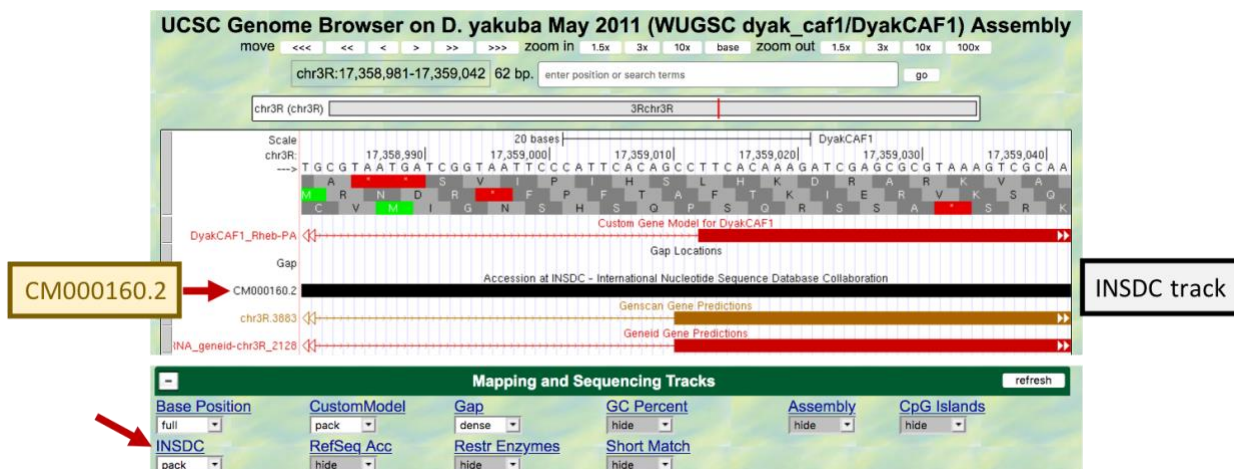


Figure 18. Use the INSDC track to determine the GenBank accession number for the *D. yakuba* scaffold chr3R.

Go back to the web browser tab with the NCBI *tblastn* search interface. Enter “CM000160.2” into the “Enter Subject Sequence” text box. Under “Subject subrange”, enter “17300000” in the “From” field and “17400000” in the “To” field (Figure 19).

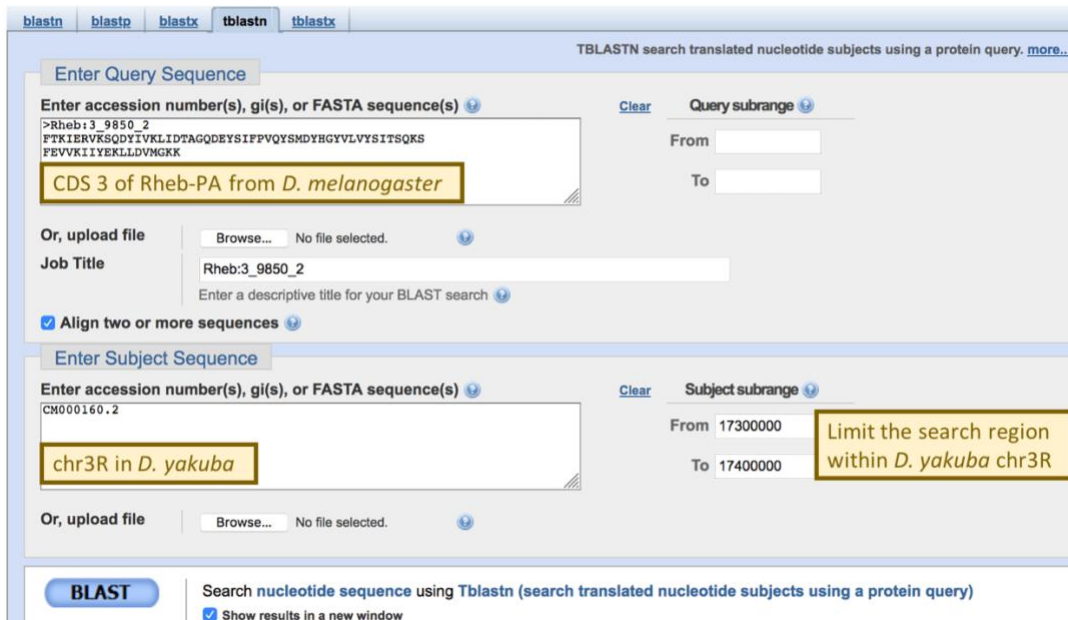


Figure 19. Specify the GenBank accession number for *D. yakuba* chr3R (i.e. CM000160.2) in the “Enter Subject Sequence” text box, and limit the *tblastn* search to the 17300000-17400000 region within this scaffold.

Click on the plus icon next to the “Algorithm parameters” header to expand the section. Under “Scoring Parameters”, select the “No adjustment” option under the “Compositional adjustments” field. Under “Filtering and Masking”, unselect the “Low complexity regions” checkbox to turn off this filter (Figure 20). Click on the “BLAST” button to run the search.

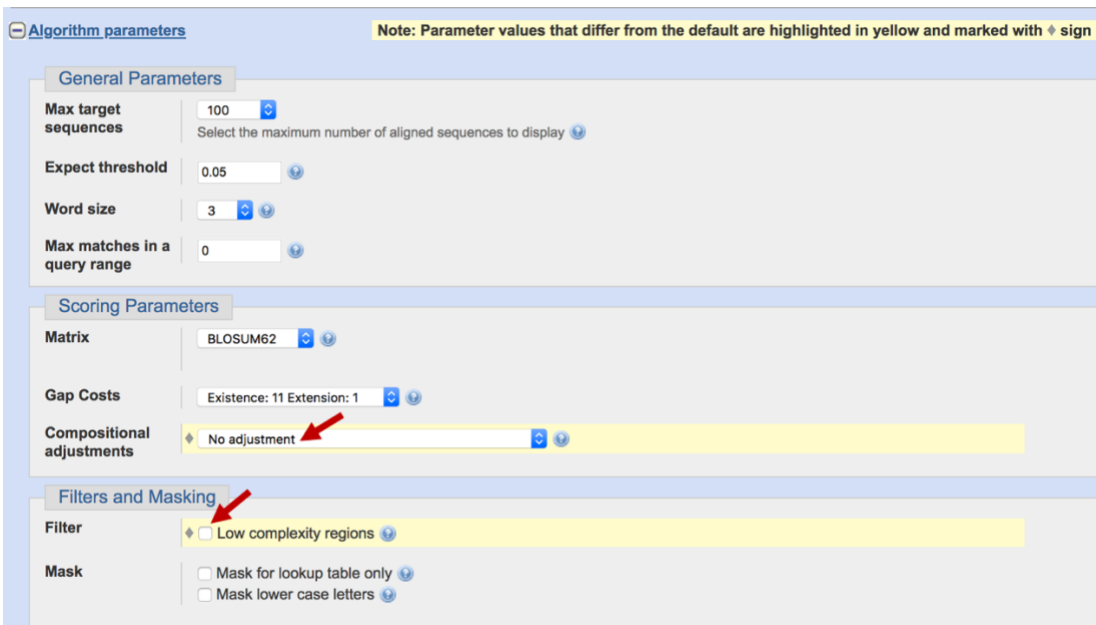



Figure 20. Under “Algorithm parameters”, turn off compositional adjustments and the low complexity filter for the *tblastn* search.

Once the *tblastn* search is complete, click on the “Alignment” tab to examine the *tblastn* alignment. The *tblastn* alignment covers all 68aa of CDS 3\_9850\_2, and it places the CDS at 17,359,013-17,359,216 in frame +2 (Figure 21).

Score	Expect	Identities	Positives	Gaps	Frame
137 bits(344)	4e-42	68/68(100%)	68/68(100%)	0/68(0%)	+2
Query 1	FTKIERVKSQDYIVKLIDTAGQDEYSIFPVQYSMDYHG YVLVYSITSQKSFEVVKIIYEK				60
Sbjct 17359013	FTKIERVKSQDYIVKLIDTAGQDEYSIFPVQYSMDYHG YVLVYSITSQKSFEVVKIIYEK				17359192
Query 61	LLDVMGKK	68	Query Descr Rheb:3_9850_2		
Sbjct 17359193	LLDVMGKK	17359216	Query Length 68		
			Subject ID CM000160.2 (dna)		
			Subject Descr Drosophila yakuba strain Tai18E2 chromosome 3R		
			Subject Length 100000		

Figure 21. The *tblastn* search of CDS 3\_9850\_2 from the *D. melanogaster* *Rheb* gene (query) against the *D. yakuba* scaffold chr3R (CM000160.2; subject) placed this CDS at 17,359,013-17,359,216 in frame +2.

Go back to the web browser tab with the *UCSC Genome Browser* view of the region surrounding the splice acceptor site for CDS 3. (If you have closed the *UCSC Genome Browser* window, go back to the *Gene Model Checker* checklist and click on the  icon next to “Acceptor for CDS 3” checklist item to re-open the Genome Browser view.) Scroll down to the track configuration section and then change the display modes for the following evidence tracks:

Under “Mapping and Sequencing Tracks”:

- INSDC: **hide**

Under “Genes and Gene Prediction Tracks”:

- *D. mel* Proteins: **pack**
- *GeMoMa* Genes: **pack**
- *Augustus*: **pack**

Under “RNA-Seq Tracks”:

- Splice Junctions: **pack**

Click on one of the “refresh” buttons to update the Genome Browser display.

The *SPALN* alignment to the *D. melanogaster* proteins *Rheb*-PA and *Rheb*-PB, predictions from four gene predictors (i.e. *GeMoMa*, *Genscan*, *Geneid*, and *Augustus*), the RNA-Seq read coverage, and the splice junction JUNC00073912 (score = 1511) all support extending the start of CDS 3 by two nucleotides to 17,359,011 (Figure 22).

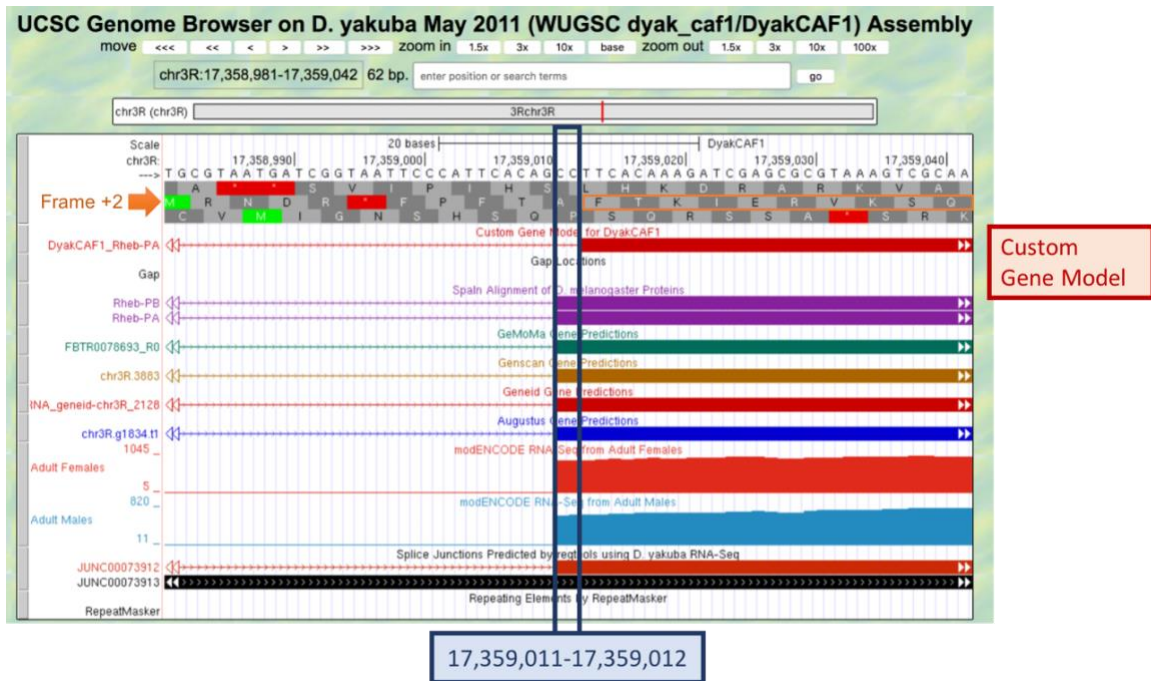


Figure 22. The evidence tracks from protein sequence alignments, gene predictions, and RNA-Seq data support changing the start coordinate of CDS 3 to 17,359,011.

Since the *tblastn* alignment for CDS 3\_9850\_2 begins at 17,359,013 on *D. yakuba* chr3R (Figure 21), there are two nucleotides between the splice acceptor site at 17,359,009-17,359,010 and the start of the first complete codon of CDS 3\_9850\_2 at 17,359,013. Hence CDS 3\_9850\_2 has a phase 2 splice acceptor site (Figure 23).

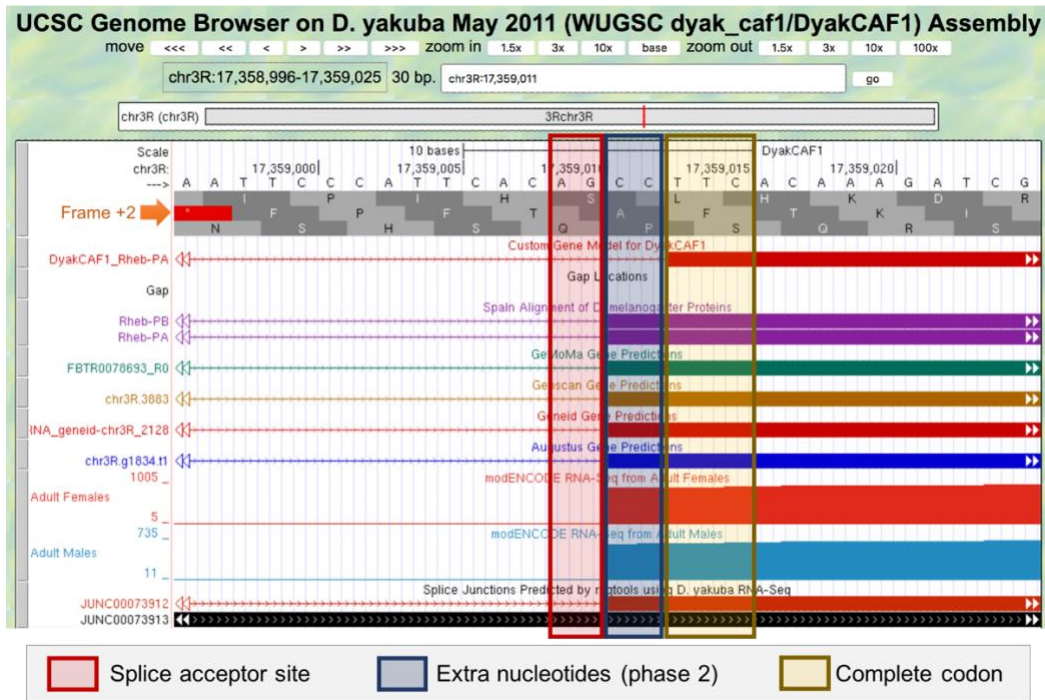


Figure 23. The splice acceptor site for CDS 3\_9850\_2 is in phase 2 relative to frame +2.

If CDS 3\_9850\_2 has a phase 2 splice acceptor site, then the previous CDS (i.e. 2\_9850\_2) must have a phase 1 splice donor site. The *tblastn* search of CDS 2\_9850\_2 against the *D. yakuba* scaffold chr3R (CM000160.2) placed the CDS at 17,358,844-17,358,912 in frame +1 (Figure 24).

**Drosophila yakuba strain Tai18E2 chromosome 3R, whole genome shotgun sequence**

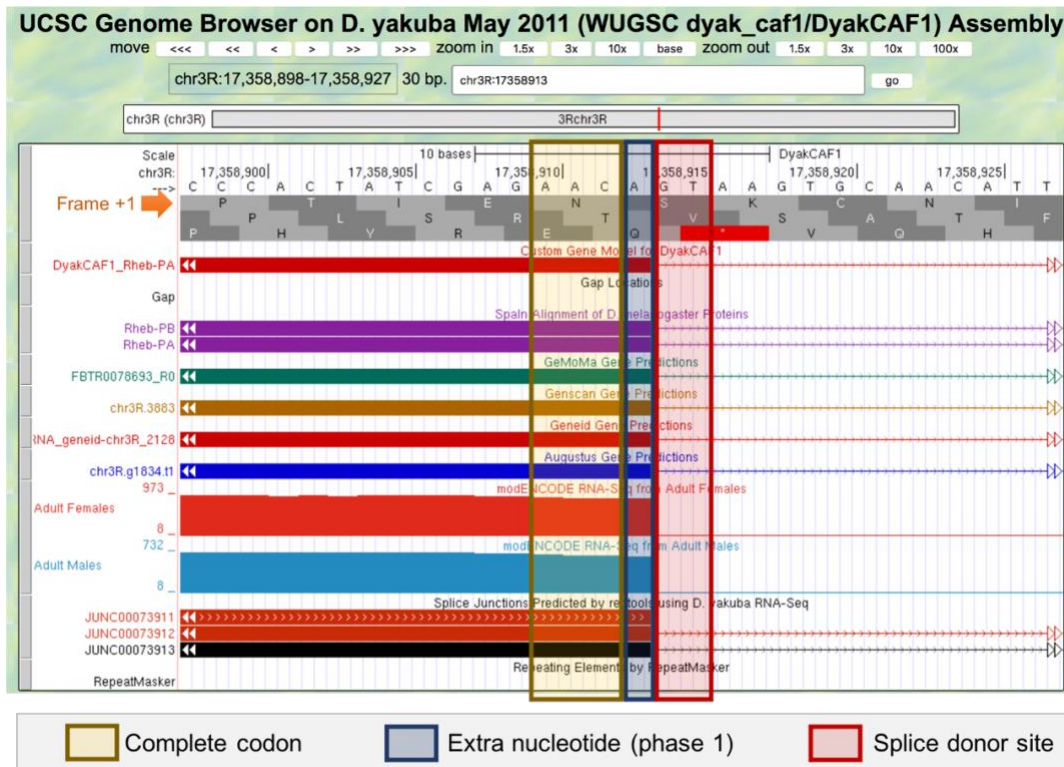
Sequence ID: [CM000160.2](#) Length: 28832112 Number of Matches: 1

Range 1: 17358844 to 17358912 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Positives	Gaps	Frame
51.2 bits(121)	8e-13	23/23(100%)	23/23(100%)	0/23(0%)	+1
Query 1	KSSLCIQFVEGQFVDSYDPTIEN 23			Query Descr	Rheb:2_9850_2
Sbjct	<span style="border: 1px solid red; padding: 2px;">17358844</span>	KSSLCIQFVEGQFVDSYDPTIEN <span style="border: 1px solid red; padding: 2px;">17358912</span>		Query Length	23
				Subject ID	<a href="#">CM000160.2</a> (dna)
				Subject Descr	Drosophila yakuba strain Tai18E2 chromosome 3R
				Subject Length	100000

**Figure 24.** The *tblastn* search of CDS 2\_9850\_2 from the *D. melanogaster* *Rheb* gene (query) against the *D. yakuba* scaffold chr3R (CM000160.2; subject) placed this CDS at 17,358,844-17,358,912 in frame +1.

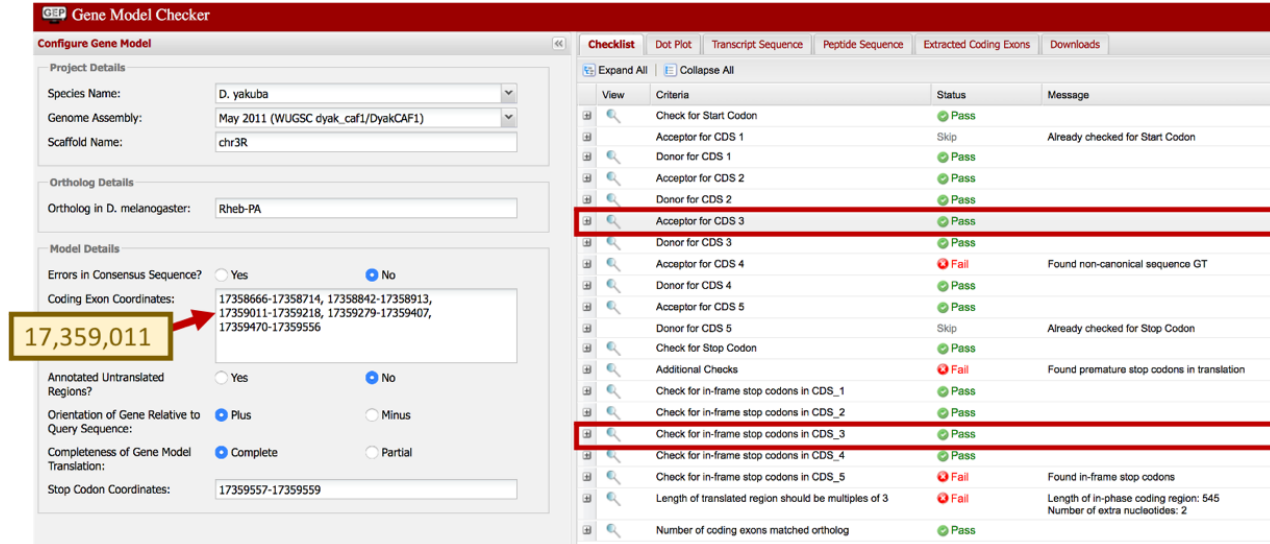
The evidence tracks surrounding the end of the *tblastn* alignment for CDS 2\_9850\_2 supports extending the end of the CDS by one extra nucleotide to 17,358,913. Hence the splice donor site at 17,358,914-17,358,915 will be in phase 1 relative to frame +1 (Figure 25).



**Figure 25.** CDS 2\_9850\_2 has a phase 1 splice donor site relative to frame +1.

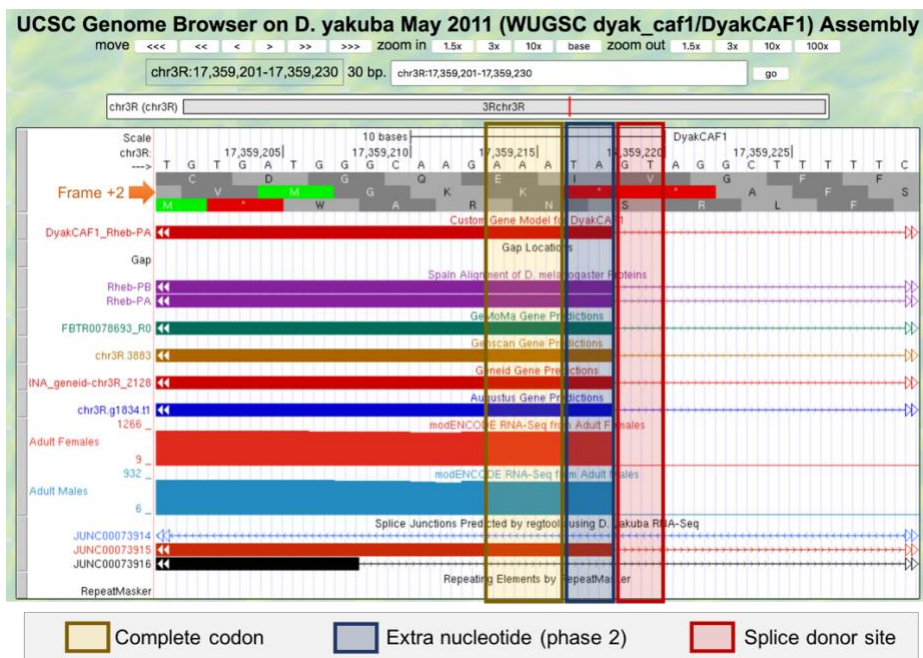
Collectively, the analysis of the splice junction between CDS 2 and 3 supports changing the end coordinate of CDS 2 to 17,358,913 with a phase 1 splice donor site relative to frame +1, and changing the start coordinate of CDS 3 to 17,359,011 with a phase 2 splice acceptor site relative to frame +2.

Go back to the web browser tab with the *Gene Model Checker*. Under the “Coding Exon Coordinates” field, change the coordinates for CDS 3 from “17359013-17359218” to “17359011-17359218”. Click on the “Verify Gene Model” button. The revised gene model gene model has successfully resolved the issue with the splice donor site for CDS 3 (Figure 26).



**Figure 26. Changing the start coordinate of CDS 3 to 17,359,011 resolves the non-canonical splice acceptor site and the in-frame stop codon issues previously reported by the *Gene Model Checker* checklist.**

We can use the same strategy to resolve the non-canonical GT splice acceptor site for CDS 4. The *tblastn* alignment for CDS 3\_9850\_2 ends at 17,359,216 in frame +2 (Figure 21). Examination of the evidence tracks surrounding this position using the *GEP UCSC Genome Browser* supports extending the end of the CDS to 17,359,218 with a phase 2 splice donor site at 17,359,219-17,359,220 (Figure 27).



**Figure 27. CDS 3\_9850\_2 has a phase 2 splice donor site relative to frame +2.**

The *tblastn* search of CDS 4\_9850\_1 against *D. yakuba* chr3R placed the CDS at 17,359,279-17,359,407 in frame +1 (Figure 28).

**Drosophila yakuba strain Tai18E2 chromosome 3R, whole genome shotgun sequence**

Sequence ID: [CM000160.2](#) Length: 28832112 Number of Matches: 1

Range 1: 17359279 to 17359407 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Positives	Gaps	Frame
80.1 bits(196)	1e-22	40/43(93%)	40/43(93%)	0/43(0%)	+1

Query 1 VPVVLVGNKIDLHQERTVSTEEGKLAESWRAAFLETSKQNE 43  
 VPVVLVGNK DL ERTVSTEEGKLAESWRAAFLETSKQNE  
 Sbjct 17359279 VPVVLVGNKTDLPQPERTVSTEEGKLAESWRAAFLETSKQNE 17359407

Query Descr	Rheb:4_9850_1
Query Length	43
Subject ID	<a href="#">CM000160.2</a> (dna)
Subject Descr	Drosophila yakuba strain Tai18E2 chromosome 3R
Subject Length	100000

Figure 28. The *tblastn* search of CDS 4\_9850\_1 from the *D. melanogaster* *Rheb* gene (query) against the *D. yakuba* scaffold chr3R (CM000160.2; subject) placed this CDS at 17,359,279-17,359,407 in frame +1.

Examination of the genomic region surrounding the start of CDS 4 (i.e. 17,359,279) using the *GEP UCSC Genome Browser* shows that the *SPALN* protein alignments, the predictions from multiple gene predictors, the RNA-Seq read coverage, and the splice junction JUNC00073915 (score = 2076) support extending the start of the CDS by one nucleotide to 17,359,278. Hence the splice acceptor site at 17,359,276-17,359,277 is in phase 1 relative to frame +1 (Figure 29).

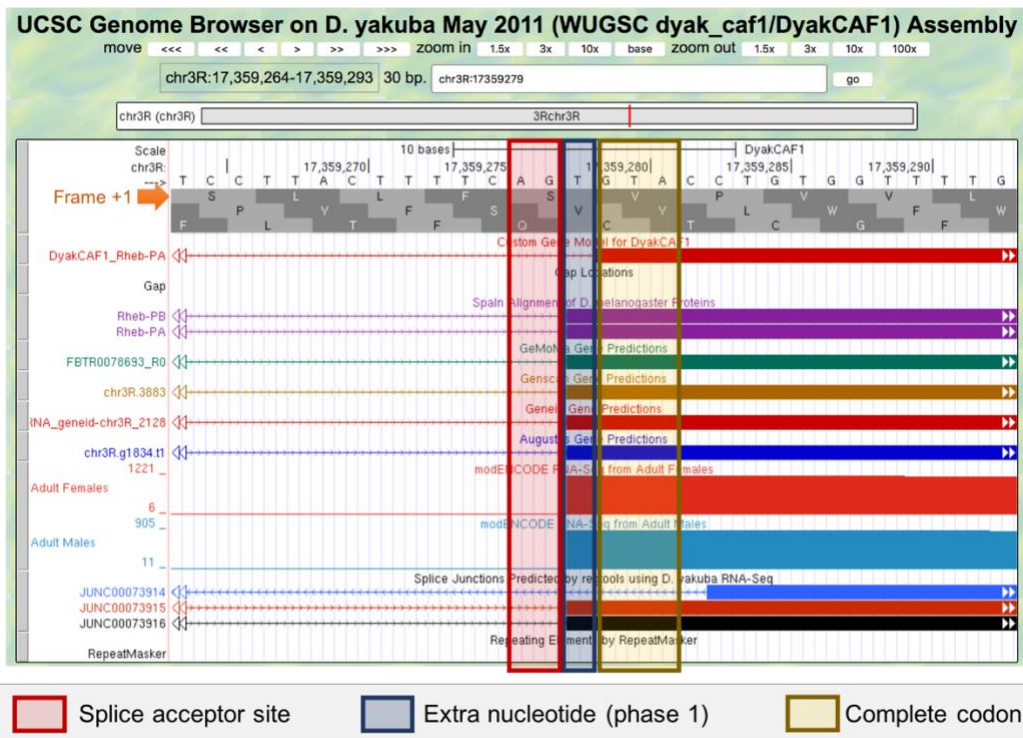


Figure 29. CDS 4\_9850\_1 has a phase 1 splice acceptor site relative to frame +1.

To revise the coordinates for CDS 4, go back to the web browser tab with the *Gene Model Checker*. Under the “Coding Exon Coordinates” field, change the coordinates for CDS 4 from “17359279-17359407” to “17359278-17359407”. Click on the “Verify Gene Model” button.

The *Gene Model Checker* checklist shows that the revised gene model satisfies all the criteria on the *Gene Model Checker* checklist (Figure 30).

View	Criteria	Status	Message
	Check for Start Codon	Pass	
	Acceptor for CDS 1	Skip	Already checked for Start Codon
	Donor for CDS 1	Pass	
	Acceptor for CDS 2	Pass	
	Donor for CDS 2	Pass	
	Acceptor for CDS 3	Pass	
	Donor for CDS 3	Pass	
	Acceptor for CDS 4	Pass	
	Donor for CDS 4	Pass	
	Acceptor for CDS 5	Pass	
	Donor for CDS 5	Skip	Already checked for Stop Codon
	Check for Stop Codon	Pass	
	Additional Checks	Pass	
	Number of coding exons matched ortholog	Pass	

**Figure 30.** After changing the start coordinate for CDS 4 to 17,359,278, the revised gene model satisfies all of the criteria on the *Gene Model Checker* checklist.

### Examine the dot plot and alignment of our gene model against the *D. melanogaster* ortholog

The comparative annotation strategy used by the GEP is based on parsimony with the orthologous protein from the informant genome (i.e. *D. melanogaster*). Hence the proposed gene model should aim to minimize the number of changes compared to the *D. melanogaster* ortholog unless the change in gene structure is supported by experimental evidence (e.g., RNA-Seq data) or by sequence conservation in species more closely-related to the target species.

The *Gene Model Checker* checklist verifies that the proposed gene model satisfies the biological constraints for most eukaryotic genes (e.g., the gene model has a start codon, stop codon, canonical splice donor and acceptor sites, and no in-frame stop codons). However, gene models that satisfy all the criteria on the checklist might not be the most parsimonious gene model compared to the *D. melanogaster* ortholog. The “Dot Plot” panel compares the proposed gene model against the putative *D. melanogaster* ortholog using a dot plot and a pairwise global protein alignment to help annotators identify changes in the proposed gene model compared to the *D. melanogaster* ortholog.

**Note:** Supporting evidence for changes in gene structure compared to the *D. melanogaster* ortholog (e.g., splitting a single exon into multiple exons, merging adjacent exons) should be described in the annotation report for the GEP project.

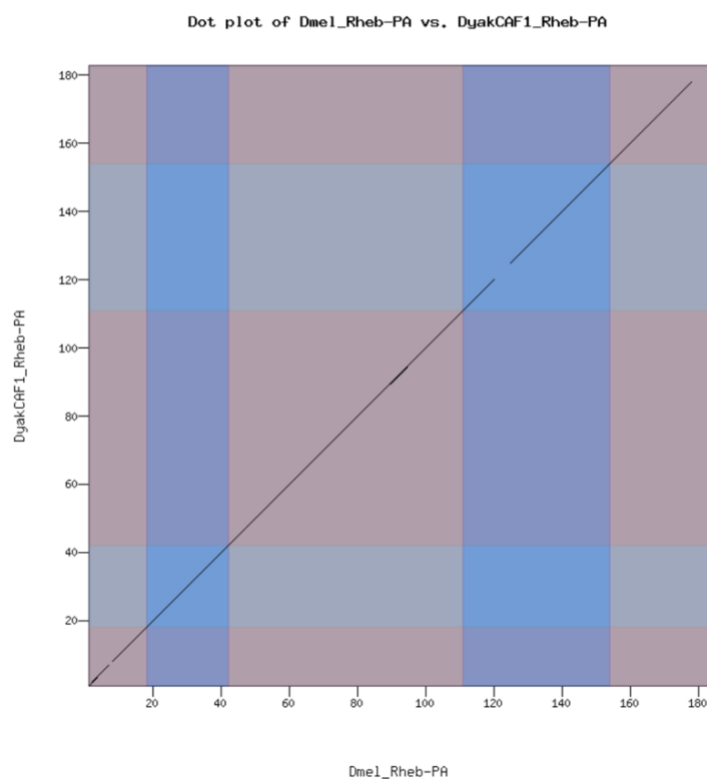


Click on the “Dot Plot” tab in the *Gene Model Checker* output. The panel shows a dot plot between the protein sequence for *D. melanogaster* Rheb-PA (x-axis) against the protein sequence for the submitted gene model for Rheb-PA in *D. yakuba* (y-axis). Identical residues between the protein sequences in the informant and the target genomes are depicted by a dot in the dot plot. The long diagonal line with a slope of approximately 1 indicates that the two sequences have approximately the same lengths, and that the two sequences are similar to each other along the entire length of the two sequences. The alternating colors in the dot plot represent the approximate locations of each CDS. The gaps along the diagonal line in the dot plot indicates residues that differ between the two sequences (Figure 31).



[View protein alignment](#)

[View dot plot in the Dot Plot Viewer](#)



**Figure 31.** The dot plot comparison of the protein sequence for *D. melanogaster* Rheb-PA (x-axis) against the protein sequence for the proposed gene model in *D. yakuba* (y-axis). The rectangles with alternating colors indicate that the Rheb-PA protein in *D. melanogaster* and *D. yakuba* both have five coding exons. The gaps along the diagonal line suggests that the major differences between the two protein sequences are located near the beginning of CDS 4 and the end of CDS 5.

**Note:** For gene models that show weak sequence similarity with the *D. melanogaster* ortholog, you can click on the “**View dot plot in the Dot Plot Viewer**” link to launch the *Dot Plot Viewer*. The *Dot Plot Viewer* allows you to control the alignment parameters (e.g., scoring matrix, word size, and neighborhood) in order to alter the sensitivity and specificity of the dot plot (Figure 32).

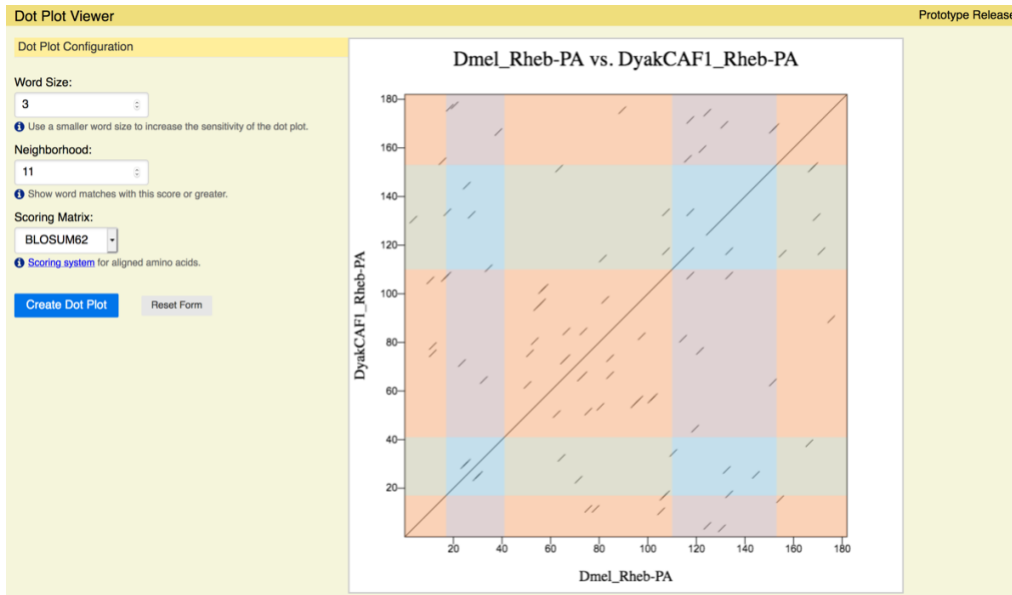


Figure 32. Change the parameters in the *Dot Plot Viewer* to control the sensitivity of the dot plot.

Go back to the *Gene Model Checker* page. Click on the “View protein alignment” link to view the alignment between the two protein sequences. Our observation of the weaker sequence similarity at the beginning of CDS 4 and the end of CDS 5 are consistent with the global alignment (Figure 33).

### Alignment of *Dmel\_Rheb-PA* vs. *DyakCAF1\_Rheb-PA*

[View plain text version](#)  
[Download alignment image](#)

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

```

Dmel_Rheb-PA      1  MPTKERHIAAMGYRSVGVKSSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDT 60
*****:*****
DyakCAF1_Rheb-PA  1  MPTKERHIAAMGYRSVGVKSSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDT 60
*****:*****
Dmel_Rheb-PA      61  AGQDEYSIFPVQYSMDYHGYVLVYSITSQKSFEVVKIIEKLLDVMGKKYVPVVLVGNKI 120
*****:*****
DyakCAF1_Rheb-PA  61  AGQDEYSIFPVQYSMDYHGYVLVYSITSQKSFEVVKIIEKLLDVMGKKYVPVVLVGNKT 120
*****:*****
Dmel_Rheb-PA      121 DLHQERTVSTEEGKKLAESWRAAFLETSAKQNE SVGDIFHQLLLIENENGNPQEKSGCL 180
** : *****
DyakCAF1_Rheb-PA  121 DLQPERTVSTEEGKKLAESWRAAFLETSAKQNE SVGDIFHQLLLIENENGNPQEKSSCL 180
*****:*****
Dmel_Rheb-PA      181  VS 182
**
DyakCAF1_Rheb-PA  181  VS 182

```

Figure 33. The protein alignment between *D. melanogaster* Rheb-PA (top) and its putative ortholog in *D. yakuba* (bottom). The alternating colors represent adjacent CDSs. The protein alignment shows that most of the differences between the two protein sequences (highlighted by purple boxes) are located in CDS 4 and CDS 5.

**Note:** The symbols in the match line corresponds to the level of similarity between the aligned residues. The asterisk (\*) indicates that the aligned residues are identical. The colon (: ) indicates the aligned residues have a score of > 0.5 in the Gonnet PAM 250 scoring matrix. The period (.) indicates that the score of the aligned residue is > 0 and ≤ 0.5. A space indicates that the aligned residues have a score ≤ 0. See the “[Bioinformatics Tools FAQ](#)” on the EMBL-EBI website for additional details.

## Use dot plots to identify large insertions and deletions in the proposed gene model

Large horizontal gaps in the dot plot indicate that there are residues in the *D. melanogaster* protein sequence that are absent from the protein sequence of the putative ortholog in the target species (e.g., *D. yakuba*). In contrast, large vertical gaps in the dot plot indicate that there are extra residues in the protein sequence of the target species compared to the *D. melanogaster* protein. These large-scale changes in CDS sizes are unusual among the different *Drosophila* species, and the annotator should provide detailed explanations of the supporting evidence for these changes in the annotation report for the GEP project. In many cases, there might be an alternative set of splice donor and acceptor sites that would minimize the change in CDS size compared to the *D. melanogaster* ortholog.

To illustrate the impact of a large deletion in the proposed gene model compared to *D. melanogaster* ortholog on the dot plot, we will change the “Coding Exon Coordinates” field to the following set of coordinates:

17358666-17358714, 17358842-17358913, 17359044-17359218, 17359278-17359407,  
17359470-17359556

Click on the “Verify Gene Model” button. The proposed gene model passes all the criteria on the *Gene Model Checker* checklist (Figure 34).

The screenshot shows the Gene Model Checker interface. On the left, the 'Configure Gene Model' section is visible, with the following details:

- Project Details:** Species Name: *D. yakuba*, Genome Assembly: May 2011 (WUGSC dyak\_caf1/DyakCAF1), Scaffold Name: chr3R.
- Ortholog Details:** Ortholog in *D. melanogaster*: Rheb-PA.
- Model Details:** Errors in Consensus Sequence?  Yes  No. Coding Exon Coordinates: 17358666-17358714, 17358842-17358913, 17359044-17359218, 17359278-17359407, 17359470-17359556. Annotated Untranslated Regions?  Yes  No. Orientation of Gene Relative to Query Sequence:  Plus  Minus. Completeness of Gene Model Translation:  Complete  Partial. Stop Codon Coordinates: 17359557-17359559.

On the right, the 'Checklist' tab is active, showing a table of criteria:

View	Criteria	Status	Message
<input type="checkbox"/>	Check for Start Codon	Pass	
<input type="checkbox"/>	Acceptor for CDS 1	Skip	Already checked for Start Codon
<input type="checkbox"/>	Donor for CDS 1	Pass	
<input type="checkbox"/>	Acceptor for CDS 2	Pass	
<input type="checkbox"/>	Donor for CDS 2	Pass	
<input type="checkbox"/>	Acceptor for CDS 3	Pass	
<input type="checkbox"/>	Donor for CDS 3	Pass	
<input type="checkbox"/>	Acceptor for CDS 4	Pass	
<input type="checkbox"/>	Donor for CDS 4	Pass	
<input type="checkbox"/>	Acceptor for CDS 5	Pass	
<input type="checkbox"/>	Donor for CDS 5	Skip	Already checked for Stop Codon
<input type="checkbox"/>	Check for Stop Codon	Pass	
<input type="checkbox"/>	Additional Checks	Pass	
<input type="checkbox"/>	Number of coding exons matched ortholog	Pass	

Figure 34. An alternative gene model for Rheb-PA in *D. yakuba* which passes all the criteria on the *Gene Model Checker* checklist.

Click on the “Dot Plot” tab. The dot plot image in this section shows a large horizontal gap (Figure 35). The alternating color boxes (which corresponds to the individual coding exons) indicate that the extra sequence in *D. melanogaster* Rheb-PA compared to the submitted gene model is located at the beginning of CDS 3.

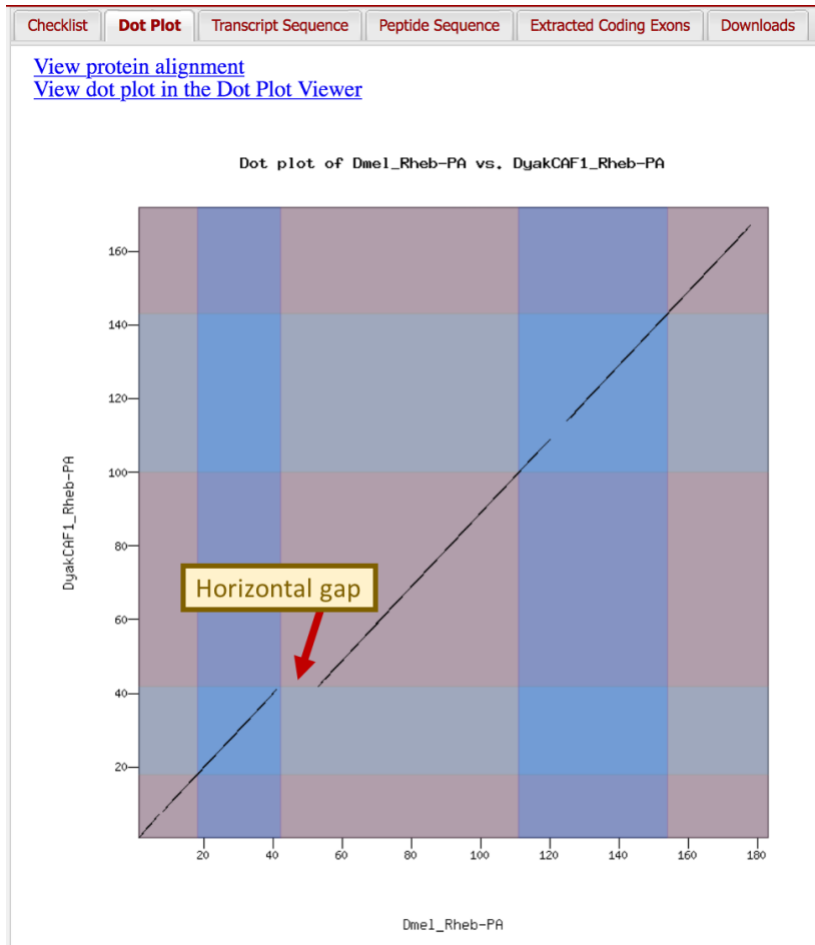


Figure 35. Large horizontal gap indicates that there are extra sequences at the beginning of CDS 3 in *D. melanogaster* Rheb-PA (x-axis) compared to the *D. yakuba* gene model (y-axis).

Click on the “View protein alignment” link at the top of the “Dot Plot” panel. Consistent with the dot plot, the alignment shows 11 residues at the beginning of CDS 3 in *D. melanogaster* Rheb-PA that aligned to gaps in the *D. yakuba* model (Figure 36).

### Alignment of Dmel\_Rheb-PA vs. DyakCAF1\_Rheb-PA

[View plain text version](#)  
[Download alignment image](#)

**Identity:** 165/182 (90.7%), **Similarity:** 167/182 (91.8%), **Gaps:** 11/182 ( 6.0%)

Dmel_Rheb-PA	1	MPTKERHIAMMGYRSVKGSSLCIQFVEGQFVDSYDPTIENFTKIERVKSQDYIIVKLIDT	60
		*****:*****	
DyakCAF1_Rheb-PA	1	MPTKERNIAMMGYRSVKGSSLCIQFVEGQFVDSYDPTIEN-----NYIVKLIDT	49
Dmel_Rheb-PA	61	AGQDEYSIFPVQYSMDYHGCVLVYSITSQKSFEVVKIIEYKLLDVMGKKYVPVVLVGNKI	120
		*****	
DyakCAF1_Rheb-PA	50	AGQDEYSIFPVQYSMDYHGCVLVYSITSQKSFEVVKIIEYKLLDVMGKKYVPVVLVGNKT	109
Dmel_Rheb-PA	121	DLHQERTVSTEEGKKLAESWRAAFLETSAKQNESVGDIHQLLILLENENGNPQEKSGCL	180
		***:*****_*	
DyakCAF1_Rheb-PA	110	DLQPERTVSTEEGKKLAESWRAAFLETSAKQNESVGDIHQLLILLENENGNPQEKSSCL	169
Dmel_Rheb-PA	181	VS	182
		**	
DyakCAF1_Rheb-PA	170	VS	171

Figure 36. The protein alignment shows 11 residues at the beginning of CDS 3 in *D. melanogaster* Rheb-PA (top) that do not align to residues in the *D. yakuba* gene model (bottom).

Previous *tblastn* search of CDS 3\_9850\_2 against *D. yakuba* chr3R shows an alignment that begins at 17,359,013 (Figure 21). If CDS 3 begins at 17,359,044 (as suggested by the proposed gene model), this will lead to the truncation of 11 amino acids that are conserved between *D. melanogaster* and *D. yakuba* (Figure 37).

[Download](#) [GenBank](#) [Graphics](#)

**Drosophila yakuba strain Tai18E2 chromosome 3R, whole genome shotgun sequence**  
 Sequence ID: [CM000160.2](#) Length: 28832112 Number of Matches: 1


Range 1: 17359013 to 17359216 [GenBank](#) [Graphics](#) [Next Match](#) [Previous Match](#)

Score	Expect	Identities	Positives	Gaps	Frame
137 bits(344)	4e-42	68/68(100%)	68/68(100%)	0/68(0%)	+2

Query	1	FTKIERVKSQDYIVKLI	DTAGQDEYSIFPVQYSMDYHGYVLVYSITSQKSF	EVVKIIYEK	60
Sbjct	17359013	FTKIERVKSQDYIVKLI	DTAGQDEYSIFPVQYSMDYHGYVLVYSITSQKSF	EVVKIIYEK	17359192
Query	61	LLDVMGKK	68		
Sbjct	17359193	LLDVMGKK	17359216		

Query Descr	Rheb:3_9850_2
Query Length	68
Subject ID	<a href="#">CM000160.2</a> (dna)
Subject Descr	Drosophila yakuba strain Tai18E2 chromosome 3R
Subject Length	100000

**Figure 37.** The *tblastn* alignment of *D. melanogaster* CDS 3\_9850\_2 (query) against *D. yakuba* scaffold chr3R (CM000160.2; subject) shows that all 68 residues are conserved between the two species. If CDS 3 begins at 17,359,044 (as suggested by the proposed gene model), this would result in the removal of 11 conserved amino acids from the *D. yakuba* ortholog of Rheb-PA compared to *D. melanogaster* (red box).

We can use the *GEP UCSC Genome Browser* to gather additional evidence to support revising the start coordinate of CDS 3. Go back to the *Gene Model Checker* checklist and then click on the  icon next to “Acceptor for CDS 3” checklist item to view the proposed gene model in the *GEP UCSC Genome Browser*. Zoom out 3x and then change the display settings for the following evidence tracks:

Under “Genes and Gene Prediction Tracks”:

- *D. mel* Proteins: **pack**
- *GeMoMa* Genes: **pack**
- *Augustus*: **pack**

Under “RNA-Seq Tracks”:

- Splice Junctions: **pack**

Click on one of the “refresh” buttons to update the Genome Browser display.

The Genome Browser view shows that the start of CDS 3 in the proposed gene model (at 17,359,044) are inconsistent with the available evidence from protein sequence alignments, multiple gene predictors, RNA-Seq read coverage, and the splice junction prediction JUNC00073912 (score = 1511). Hence the available evidence supports changing the start of CDS 3 from 17,359,044 to 17,359,011 (Figure 38).

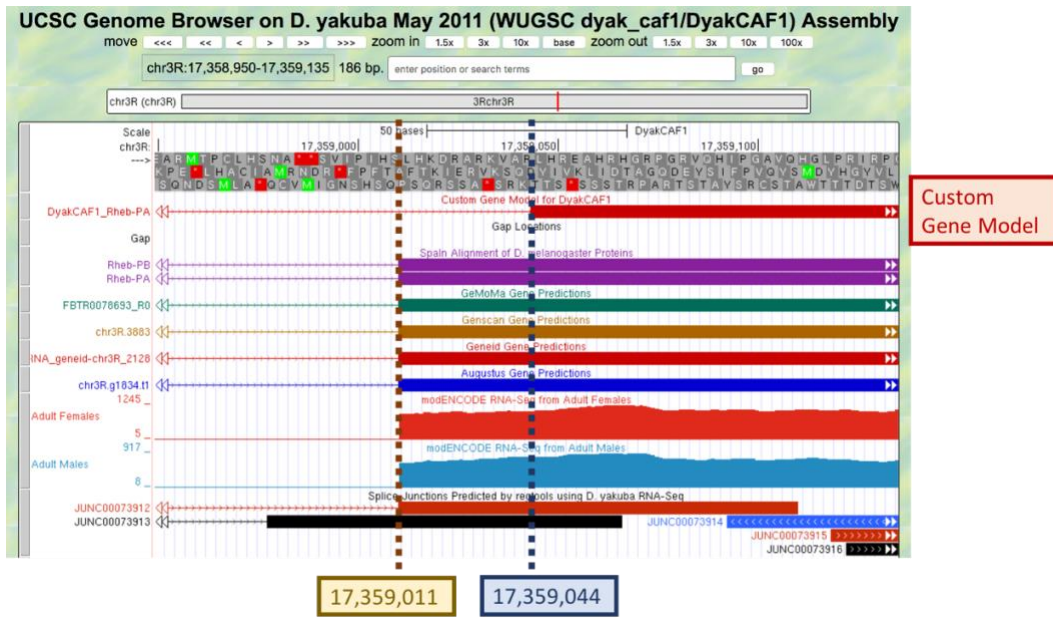
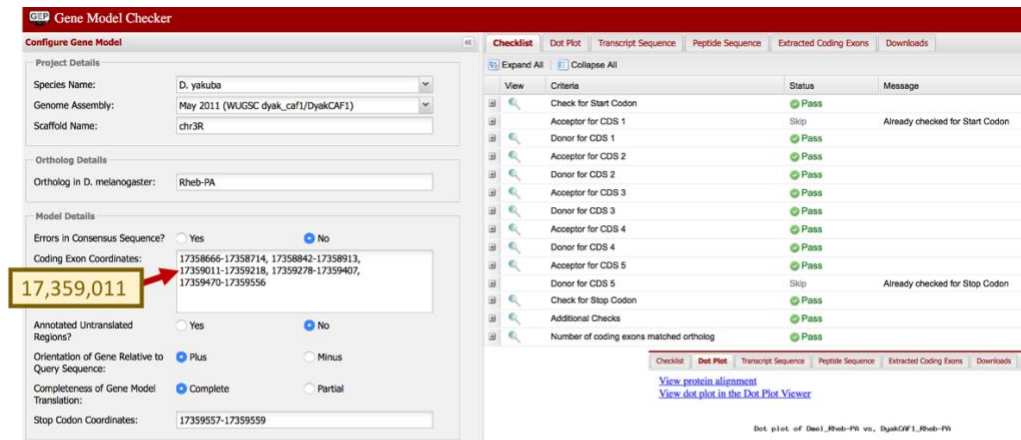


Figure 38. The evidence tracks on the GEP UCSC Genome Browsers support changing the start coordinate of CDS 3 to 17,359,011.

Go back to the “Configure Gene Model” panel of the Gene Model Checker, and change the start coordinate for CDS 3 from “17359044” to “17359011”. Click on the “Verify Gene Model” button. The proposed gene model passes all the criteria on the Gene Model Checker checklist. The dot plot and protein alignment show that the beginning of CDS 3 in the revised *D. yakuba* Rheb-PA gene model are conserved with the orthologous CDS in *D. melanogaster* (Figure 39).



Alignment of Dmel\_Rheb-PA vs. DyakCAF1\_Rheb-PA

[View plain text version](#)  
[Download alignment image](#)

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

```

Dmel_Rheb-PA      1  MPTKERHIANMGYRSVQESTFCQYVQVQVQVYDFFFTFFTFIERVRSQDIVIKLIDR 60
DyakCAF1_Rheb-PA 1  MPTKERHIANMGYRSVQKSSLCIQVEGQVDSYDPIENTFTFFTFIERVRSQDIIVIKLIDR 60
Dmel_Rheb-PA      61  AGQDEYSIFPVOYSMDYHGTVLVYSITSQKSFVVKIIEKLLDVGKGYQFVVLVGNK 120
DyakCAF1_Rheb-PA 61  AGQDEYSIFPVOYSMDYHGTVLVYSITSQKSFVVKIIEKLLDVGKGYQFVVLVGNK 120
Dmel_Rheb-PA      121  DIQPERVSRERCKKTAESWRRAAFLETSSAKQNSVGDIFHQLLLILENENGNPQEKSSCL 180
DyakCAF1_Rheb-PA 121  DIQPERVSRERCKKTAESWRRAAFLETSSAKQNSVGDIFHQLLLILENENGNPQEKSSCL 180
Dmel_Rheb-PA      181  VS 182
DyakCAF1_Rheb-PA 181  VS 182
    
```

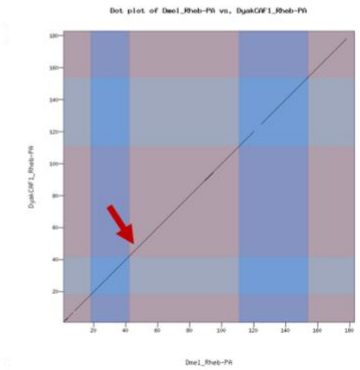
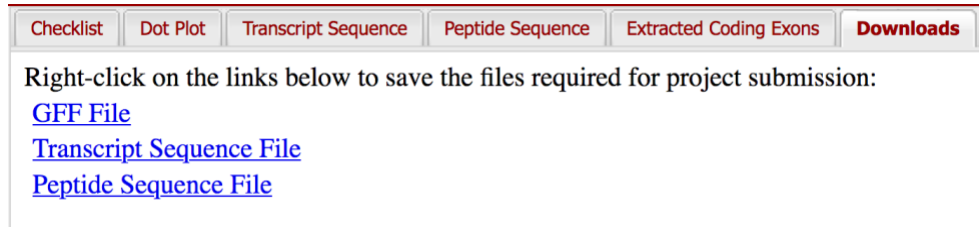


Figure 39. The Gene Model Checker results after changing the start coordinate of CDS 3 to 17,359,011 shows a more parsimonious gene model for Rheb-PA in *D. yakuba* compared to the *D. melanogaster* ortholog.

### Download the files generated by the *Gene Model Checker*

Now that we have verified the *D. yakuba* gene model for Rheb-PA, we should save the files generated by the *Gene Model Checker* in preparation for project submission. Click on the “Downloads” tab to see links to the three files generated by the *Gene Model Checker* (Figure 40).



**Figure 40.** Right-click on the links and choose “Save Link As...” or “Download Linked File As...” to save the files generated by the *Gene Model Checker*.

Right click ([control-click on macOS](#)) on the links and choose “Save Link As...” or “Download Linked File As...” to save the files onto your computer. When preparing files for project submission, you will need to concatenate the output from all the genes you have annotated and generate three files: one file containing all the GFF entries, another file with all the transcript sequences, and the third file with all the peptide sequences. The [Annotation Files Merger](#) tool (available on the GEP website) can be used to create these combined GFF, transcript, and peptide sequence files.

## Verifying gene models with consensus errors

Because the genome assemblies for the GEP projects have not undergone manual sequence improvement, the project sequences might contain consensus errors that impact the coding region annotations. For example, the most common types of errors for the Pacific Biosciences (PacBio) and the Nanopore sequencing platforms are insertions and deletions, which could lead to frame shifts that disrupt the open reading frames of coding exons.

To address this issue, the GEP has developed the [Sequence Updater](#) tool to enable annotators to document errors in the consensus sequence. The *Sequence Updater* will generate a file in the [Variant Call Format](#) (VCF) that describes changes to the original project sequence in a standardized format. The *Gene Model Checker* can take the sequence changes specified in the VCF file into account when it validates a gene model.

### Check the original *tgo* gene model in *Drosophila sechellia*

The [Sequence Updater User Guide](#) includes an example where there is one extra nucleotide in the *Drosophila sechellia* scaffold super\_0 that disrupts the open reading frame of the *tgo* gene. In this tutorial, we will incorporate the VCF file generated by the *Sequence Updater* into our annotation of the *D. sechellia tgo* ortholog.

If we were to check the gene model for *tgo*-PA (which has a single CDS at 17,034,485-17,036,408) using the original super\_0 sequence, we find that the *Gene Model Checker* reported an in-frame stop codon within CDS 1 (Figure 41).

The screenshot shows the Gene Model Checker interface. On the left, the 'Configure Gene Model' section is filled out with the following details:

- Project Details:** Species Name: *D. sechellia*, Genome Assembly: May 2011 (Broad dsec\_caf1/DsecCAF1), Scaffold Name: super\_0
- Ortholog Details:** Ortholog in *D. melanogaster*: tgo-PA
- Model Details:** Errors in Consensus Sequence? No, Coding Exon Coordinates: 17034485-17036408, Annotated Untranslated Regions? No, Orientation of Gene Relative to Query Sequence: Plus, Completeness of Gene Model Translation: Complete, Stop Codon Coordinates: 17036409-17036411

The main panel shows a 'Checklist' with the following items:

View	Criteria	Status	Message
<input type="checkbox"/>	Check for Start Codon	Pass	
<input type="checkbox"/>	Acceptor for CDS 1	Skip	Already checked for Start Codon
<input type="checkbox"/>	Donor for CDS 1	Skip	Already checked for Stop Codon
<input type="checkbox"/>	Check for Stop Codon	Pass	
<input type="checkbox"/>	Additional Checks	Fail	Found premature stop codons in translation
<input type="checkbox"/>	Check for in-frame stop codons in CDS_1	Fail	Found in-frame stop codons
<input type="checkbox"/>	Length of translated region should be multiples of 3	Fail	Length of in-phase coding region: 1924 Number of extra nucleotides: 1
<input type="checkbox"/>	Number of coding exons matched ortholog	Pass	

The 'Peptide Sequence' tab is selected, showing the following sequence:

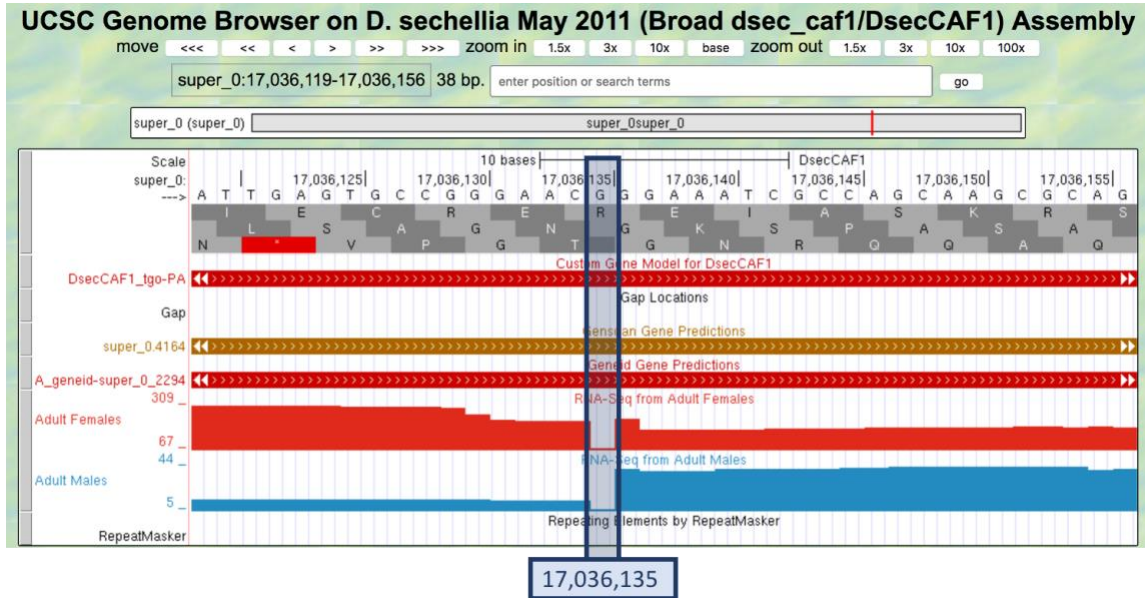
```
>DsecCAF1_tgo-PA_peptide
1 MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVP TCSALARKPKDLTLRMAV
61 AHMKALRG TGNTSSDGYKPSFLTDQELKHLILEAADGFLFVVS CD SGRVIYVSDSVTPV
121 LNYTQSDWYGTSLYEH IHPDDRKIREQLSTQESQNA GRILDLKSGTVKKEGHQSSMRLS
181 MGARRGFICRM RVGNVPESMVSGHLNRLKQRNSLGP SRDGTNYAVVHCTGYIKNWPPTD
241 MFPNMHMERDVDDMSSHCLVAIGRLQVTSTAANDMSG SNNQSEF ITRHAMDGKFTPVDQ
301 RVLNILGYTPTELLGKICYDFHPEDQSHMKESPDQV LKQKQMFSLLYRARAKNSEYVW
361 LRTQAYAF LNPYTDVEYIVCTNSSGKTMHGAPLDAAA AHTPEQVQQQQQQEQHVYVQA
421 APGVYARRELTPVGSATNDGMYQTHMLAMQAPT PQQQQQQRPGSAQTTPVGYTYDTTSS
481 PYSAGGSP LAKIPKSGTSPFPVAPNSWAALRPQQ QQQQQPVTTEGYQYQTSPARSPSG
541 PTYTQLSAG NKSPASAAGSISGGSTTASQCAGN VGLAAGWRSPASAAPNGASAPARS S
601 RWTGRSRTAAGSGV LRYAADVGSHADHV*GSEYQH VQHAVR
```

An arrow points to the asterisk in the sequence at position 541, which is labeled 'Stop codon' in a yellow box.

Figure 41. *Gene Model Checker* detected an in-frame stop codon within CDS 1 of the *tgo* ortholog in *D. sechellia* scaffold super\_0.

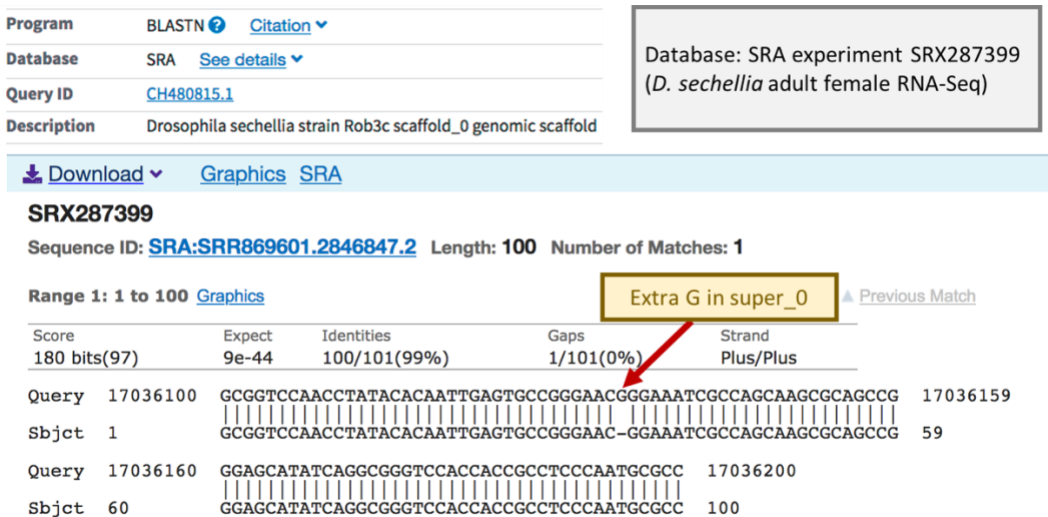


Examination of the RNA-Seq read coverage from the *D. sechellia* adult females and adult males samples show a sudden drop in read coverage at position 17,036,135 on scaffold super\_0 (Figure 42). This drop in coverage indicates a potential discrepancy between the Illumina RNA-Seq reads that have been mapped to this region of the scaffold and the consensus sequence at position 17,036,135.



**Figure 42.** The RNA-Seq coverage tracks for the adult females and adult males samples show a substantial drop in RNA-Seq read depth at position 17,036,135 of the *D. sechellia* scaffold super\_0.

Examination of the alignments of the RNA-Seq reads confirmed that there is an extra G at position 17,036,135 of the super\_0 scaffold compared to the Illumina reads (Figure 43). Hence the G at position 17,036,135 should be removed from the consensus when we verify the gene model. A VCF file which describes this change to the consensus sequence was generated by the *Sequence Updater* (DsecCAF1\_tgo-PA.vcf.txt).



**Figure 43.** Comparison of *D. sechellia* scaffold super\_0 (top) against an Illumina read (SRA:SRR869601.2846847.2) from the NCBI Sequence Read Archive experiment SRX287399 (bottom) shows an extra G at 17,036,135 in super\_0.

### Check the *tgo* gene model with the modified consensus sequence

We will incorporate the VCF file (DsecCAF1\_tgo-PA.vcf.txt) when we check the gene model for *tgo* using the *Gene Model Checker*. The *Gene Model Checker* will automatically adjust the Coding Exons Coordinates so that it is relative to the revised sequence and verify the gene model using these revised coordinates. This allows us to specify the “Coding Exon Coordinates”, the “Transcribed Exon Coordinates” and the “Stop Codon Coordinates” **relative to the original project sequence**.

In this case, we will enter the following into the *Gene Model Checker* form:

Field	Value
Species Name	<i>D. sechellia</i>
Genome Assembly	May 2011 (Broad dsec_caf1/DsecCAF1)
Scaffold Name	super_0
Ortholog in <i>D. melanogaster</i>	tgo-PA
Errors in Consensus Sequence?	Yes
File with Changes to the Consensus Sequence	DsecCAF1_tgo-PA.vcf.txt
Coding Exon Coordinates	17034485-17036408
Annotated Untranslated Regions?	No
Orientation of Gene Relative to Query Sequence	Plus
Completeness of Gene Model Translation	Complete
Stop Codon Coordinates	17036409-17036411

Click on the “Verify Gene Model” button. The *Gene Model Checker* checklist will show a warning with the message “Modified Consensus Sequence” which indicates the original sequence has changed based on the VCF file (Figure 44). Using the revised sequence, the protein alignment and the dot plot show that the entire CDS of tgo-PA is conserved between *D. melanogaster* and its *D. sechellia* ortholog (Figure 45).

In order to maintain the ability for the *Gene Model Checker* to be used as a tool for diagnosing errors in the submitted gene model, the *Gene Model Checker* will report two different sets of coordinates when you provide a VCF file. The GFF and custom tracks will show the coordinates relative to the original project sequence so that they are consistent with the rest of the evidence tracks on the *GEP UCSC Genome Browser*. In contrast, the expanded section of the Checklist and the Extracted Coding Exons will report the coordinates relative to the revised sequence.

Figure 44 Verify the *D. sechellia* tgo-PA gene model with the sequence changes described in the DsecCAF1\_tgo-PA.vcf.txt VCF file.

Alignment of Dmel\_tgo-PA vs. DsecCAF1\_tgo-PA

[View plain text version](#)  
[Download alignment image](#)

Identity: 640/643 (99.5%), Similarity: 640/643 (99.5%), Gaps: 3/643 ( 0.5%)

```

Dmel_tgo-PA      1 MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPITCSALARKPKDLTILRMAV 60
DsecCAF1_tgo-PA 1 MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPITCSALARKPKDLTILRMAV 60

Dmel_tgo-PA      61 AHMKALRGTGNTSSDGTYPKPSFLTDQELKHLILEAADGFLFVVSQDSGRVIYVSDSVPV 120
DsecCAF1_tgo-PA 61 AHMKALRGTGNTSSDGTYPKPSFLTDQELKHLILEAADGFLFVVSQDSGRVIYVSDSVPV 120

Dmel_tgo-PA      121 LNTQSDWYGTSLYEHIIHPDDREKIREQLSTQESQAGRIIDLKSGTVKKEGHQSSMRLS 180
DsecCAF1_tgo-PA 121 LNTQSDWYGTSLYEHIIHPDDREKIREQLSTQESQAGRIIDLKSGTVKKEGHQSSMRLS 180

Dmel_tgo-PA      181 MGARRGFICRMVGNVNPESHVSGHLNRLKQRNSLGPSSRDGTNYAVVHCTGYIKNWPPD 240
DsecCAF1_tgo-PA 181 MGARRGFICRMVGNVNPESHVSGHLNRLKQRNSLGPSSRDGTNYAVVHCTGYIKNWPPD 240

Dmel_tgo-PA      241 MFPNMHMERDVMSSHCCLVAIGRLQVTSANDMSGSNQSEFIRHAMDGKFTFVDQ 300
DsecCAF1_tgo-PA 241 MFPNMHMERDVMSSHCCLVAIGRLQVTSANDMSGSNQSEFIRHAMDGKFTFVDQ 300

Dmel_tgo-PA      301 RVLNLLGYTPTELLGKICYDFHPEDQSHMKESFDQVLKQKQMFLLYRARKNSEYVW 360
DsecCAF1_tgo-PA 301 RVLNLLGYTPTELLGKICYDFHPEDQSHMKESFDQVLKQKQMFLLYRARKNSEYVW 360

Dmel_tgo-PA      361 LRTQAYAFLNPHYTDEVEYIVCTNSSGKTHGAPLDAAAAHTPEVQQQQQQEQHVYVQA 419
DsecCAF1_tgo-PA 361 LRTQAYAFLNPHYTDEVEYIVCTNSSGKTHGAPLDAAAAHTPEVQQQQQQEQHVYVQA 420

Dmel_tgo-PA      420 APGVYARRELTPVGSATNDGMYQTHMLAQPTPQQQQQQQRPQSAQTTPVGYTYDT 479
DsecCAF1_tgo-PA 421 APGVYARRELTPVGSATNDGMYQTHMLAQPTPQQQQQQQ--RQPSAQTTPVGYTYDT 478

Dmel_tgo-PA      480 HSPYSAGGPSPLAKIPKSGTSPFPVAPNSWAALRPQQQQQQQFVTEGYQQQTSPARSP 539
DsecCAF1_tgo-PA 479 HSPYSAGGPSPLAKIPKSGTSPFPVAPNSWAALRPQQQQQQQFVTEGYQQQTSPARSP 538

Dmel_tgo-PA      540 SGPTYTQLSAGNCRQQAQPGAYQAGPPPPNAPGMWDWQAGGHPPHPTAHPHPHPHA 599
DsecCAF1_tgo-PA 539 SGPTYTQLSAGNCRQQAQPGAYQAGPPPPNAPGMWDWQAGGHPPHPTAHPHPHPHA 598

Dmel_tgo-PA      600 HPGGPAGAGQPGQGFSDMLQMLDHTPTTFEDLNINMFSTPFE 642
DsecCAF1_tgo-PA 599 HPGGPAGAGQPGQGFSDMLQMLDHTPTTFEDLNINMFSTPFE 641
    
```

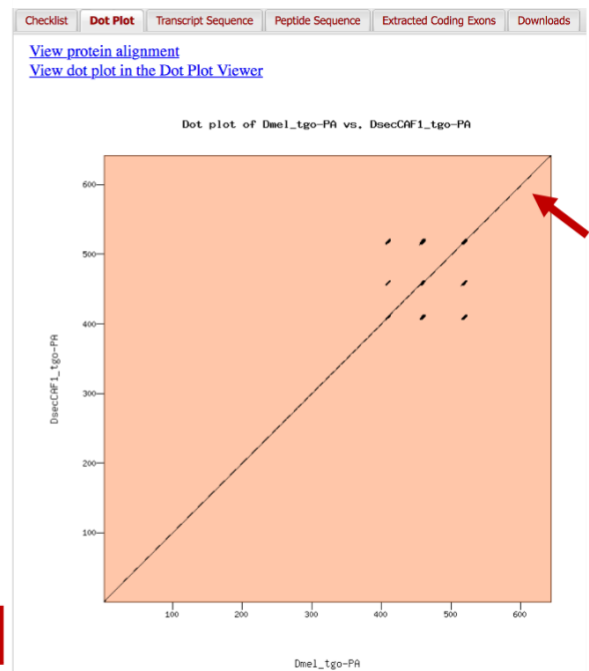


Figure 45. After applying the VCF file to the *D. sechellia* super\_0 sequence, the protein alignment and dot plot comparison of the *D. sechellia* and *D. melanogaster* protein sequences show strong sequence similarity at the C-terminus of tgo-PA (red box and red arrow).