

# Gene Model Checker User Guide

Wilson Leung

# **Table of Contents**

Introduction
Acknowledgements2
Questions about the Gene Model Checker
Availability2
General overview
Configuring the gene model
Getting help4
Available fields in the Gene Model Checker configuration panel4
Gene Model Checker results
Checklist
Dot Plot
Transcript Sequence
Peptide Sequence
Extracted Coding Exons
Downloads
Using the Gene Model Checker to identify annotation errors9
Submit the initial gene model9
Examine the Checklist to identify problems in our initial gene model
Use the Surrounding Sequence feature to identify problems with our gene model
Use the Custom Track feature to identify problems with our gene model
Use tblastn searches and the Genome Browser to identify problems with our gene model
Examine the dot plot and alignment of our gene model against the <i>D. melanogaster</i> ortholog
Use dot plots to identify large insertions and deletions in the proposed gene model
Download the files generated by the <i>Gene Model Checker</i>
Verifying gene models with consensus errors
Check the original tgo gene model in Drosophila sechellia
Check the <i>tgo</i> gene model with the modified consensus sequence

# Introduction

Before you can submit an annotation project to the Genomics Education Partnership (GEP; <u>https://thegep.org</u>), 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 (<u>wleung@wustl.edu</u>) 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 <u>F Element project</u> page and the <u>Pathways project page</u> 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).

Gene Model Checker	2										
Configure Gene Model			**	C	hecklis	t Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads	
Project Details					Expand All						
Species Name:	D. biarmipes	1	~		View	Criteria		Status	Message		
Genome Assembly:	sembly: Aug. 2013 (GEP/Dot)		~		Q	Check for Sta	irt Codon	© Pass			
Scaffold Name:	contig10					Acceptor for (	CDS 1	Skip	Already checked for Star	t Codon	
					Q	Donor for CD	S 1	© Pass			
Ortholog Details				æ	٩	Acceptor for (	CDS 2	Pass			
Ortholog in D. melanogaster:	CG31997-PB			Ħ	٩	Donor for CD	S 2	© Pass			
					٩	Acceptor for 0	CDS 3	Pass			
Model Details				æ		Donor for CD	S 3	Skip	Already checked for Stop	Codon	
Errors in Consensus Sequence?	Yes O No		#	9	Check for Sto	p Codon	Pass				
Coding Exon Coordinates:	25673-25835, 27079-27199, 27285-27468		H	R	Additional Ch	ecks	@ Pass				
county Exon coordinates.			-	Q	Number of co	ding exons matched orth	nolog OPass				
Annotated Untranslated Regions?	Yes	<ul> <li>No</li> </ul>									
Orientation of Gene Relative to Query Sequence:	Plus	Minus									
Completeness of Gene Model Translation:	<ul> <li>Complete</li> </ul>	O Partial									
Stop Codon Coordinates:	27469-27471										
		Verify Gene Model Re	eset Form								

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 <sup>(1)</sup> icon will appear on the right side of each field that failed validation.

Field	Description
Species Name	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 <u>GEP UCSC</u> <u>Genome Browser</u> .
Genome Assembly	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 <u>GEP UCSC Genome Browser</u> .
Scaffold Name	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> . Examples: contig10, chr3R, scaffold_14906
Ortholog in <i>D. melanogaster</i>	The name of the <i>D. melanogaster</i> protein isoform that is the putative ortholog of your gene model.
Errors in Consensus Sequence?	Select "Yes" if there are errors in the project consensus sequence that affect the proposed gene model. Otherwise, select "No".
File with Changes to the Consensus Sequence	This field will appear when you select the "Yes" option for the "Errors in Consensus Sequence?" field. Provide a <u>Variant Call Format</u> (VCF) file that describes the changes to the consensus sequence. You can use the <u>Sequence Updater</u> tool on the GEP website to create the VCF file.
Coding Exon Coordinates	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 <u>ends just</u> <u>before the stop codon</u> . 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').

## Available fields in the Gene Model Checker configuration panel

Annotated Untranslated Regions?	Select "Yes" if you have annotated the untranslated regions of the exons (UTRs) in addition to the coding regions. Otherwise, select "No".
Transcribed Exon Coordinates	This field will appear when you select the "Yes" option for the "Annotated Untranslated Regions?" field.
	Enter a comma-delimited list of coordinates that correspond to the <b><u>transcribed exons</u></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.
Orientation of Gene Relative to Query Sequence	The orientation of the gene relative to the genomic scaffold or the project sequence ("Plus" or "Minus").
Completeness of Gene Model Translation	Indicate whether the gene model encompasses the entire coding region.
	<b>Complete:</b> The coordinates in the "Coding Exons Coordinates" field span from the start codon to just before the stop codon.
	<b>Partial:</b> The coordinates in the "Coding Exons Coordinates" field only encompass part of the coding region of the gene.
Region Missing	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:
	<b>Missing 5' end of translated region</b> Partial gene without the start codon but contains the stop codon.
	<b>Missing 3' end of translated region</b> Partial gene without the stop codon but contains the start codon.
	Missing both 5' and 3' ends of translated region Partial gene with neither the start codon nor the stop codon.
Phase of First Exon	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.
Stop Codon Coordinates	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).
	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.
	Note that there should be <b>no overlap</b> between the coordinates in the "Stop Codon Coordinates field and the "Coding Exon Coordinates" field.

# 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 con 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.

C	hecklis	t Dot Plot Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads					
ŧ	Expand All   E Collapse All									
	View	Criteria	Status	Message						
Ħ	Q	Check for Start Codon	Pass							
ŧ		Acceptor for CDS 1	Skip	Already checked for Star	t Codon					
Ŧ	Q	Donor for CDS 1	Pass							
ŧ	Q	Acceptor for CDS 2	Pass							
Ŧ	Q	Donor for CDS 2	Pass							
ŧ	Q	Acceptor for CDS 3	Pass							
ŧ		Donor for CDS 3	Skip	Already checked for Stop	Codon					
Ħ	Q	Check for Stop Codon	Pass							
ŧ	Q	Additional Checks	Pass							
±	Q	Number of coding exons matched ortho	log 📀 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).

Ξ	Q Donor for CDS 1	📀 Pass
	Feature Coordinates: 25836-25837 Surrounding Sequence: GAGAG <mark>GT</mark> ACGTA	

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 <u>standard genetic code</u>. 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 inframe 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 <u>Annotation Files Merger</u> 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 <u>Annotation Files Merger</u> 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 <u>F Element project page</u> 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	D. yakuba
Genome Assembly	May 2011 (WUGSC dyak_caf1/DyakCAF1)
Scaffold Name	chr3R
Ortholog in D. melanogaster	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.

Configure Gene Model			~					
Project Details								
Species Name:	D. yakuba	D. yakuba 👻						
Genome Assembly:	May 2011 (WUGSC	Nay 2011 (WUGSC dyak_caf1/DyakCAF1)						
Scaffold Name:	chr3R	hr3R						
Ortholog Details								
Ortholog in D. melanogaster:	Rheb-PA							
Model Details								
Errors in Consensus Sequence?	) Yes	<ul> <li>No</li> </ul>						
Coding Exon Coordinates:	17358666-17358713 17359013-17359218 17359470-17359556	9, 17358842-17358912, 9, 17359279-17359407, 9						
Annotated Untranslated Regions?	) Yes	<ul> <li>No</li> </ul>						
Orientation of Gene Relative to Query Sequence:	• Plus	O Minus						
Completeness of Gene Model Translation:	<ul> <li>Complete</li> </ul>	O Partial						
Stop Codon Coordinates:	17359557-17359559	)						
		Verify Gene Model Reset	t Form					

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.

Checklist		Dot Plot Transcript Sequence Peptide S		Peptide Sequence	Extract	ed Coding Exons	Downloads	
47	Expand A		apse All					
	View	Criteria		Sta	tus	Message		
Ð	Q	Check for S	Start Codon	0	Pass			
Ð		Acceptor fo	r CDS 1	Ski	p	Already chee	cked for Start Codon	
Ð	Q	Donor for C	DS 1	0	Fail	Found non-o	anonical sequence G	G
ŧ	Q	Acceptor fo	r CDS 2	0	Pass			
Ð	Q	Donor for C	DS 2	0	Fail	Found non-c	anonical sequence A	G
Ð	٩	Acceptor fo	r CDS 3	0	Fail	Found non-c	anonical sequence C	С
Ð	Q	Donor for C	DS 3	0	Pass			
Ð	Q	Acceptor fo	r CDS 4	0	Fail	Found non-o	anonical sequence G	т
Ð	Q	Donor for C	DS 4	0	Pass			
Ð	Q	Acceptor fo	r CDS 5	0	Pass			
Ð		Donor for C	DS 5	Ski	p	Already chee	cked for Stop Codon	
Ð	Q	Check for S	Stop Codon	0	Pass			
Ð	Q	Additional C	Checks	0	Fail	Found prema	ature stop codons in t	ranslation
Ð	Q	Check for in	n-frame stop codons in C	DS_1 📀	Pass			
Ð	٩	Check for in	n-frame stop codons in C	DS_2 📀	Fail	Found in-fra	me stop codons	
Ð	Q	Check for in	n-frame stop codons in C	DS_3 📀	Fail	Found in-fram	me stop codons	
Ð	9	Check for in	n-frame stop codons in C	DS_4 🥹	Fail	Found in-fra	me stop codons	
Ð	9	Check for in	n-frame stop codons in C	DS_5 📀	Pass			
	٩	Length of tr	anslated region should b	e multiples 📀	Fail	Length of in- Number of e	phase coding region: xtra nucleotides: 1	541
Ð	0	Number of	coding exons matched or	rtholog 📀 I	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.

Checklist Dot Plot Transcript Sequence Peptide Sequence Extracted Coding Exons Downloads										
ŧ	Expand All E Collapse All									
	View	Criteria	Statu	Message						
	Q	Check for Start Codon	© P	ass						
	Featur Surro	re Coordinates: 17358666-17358668 unding Sequence: GGAAAATGCCAAC								
		Acceptor for CDS 1	Skip	Already ch	ecked for Start Codon					
	Featur Surro	re Coordinates: N/A unding Sequence: N/A								
	Q	Donor for CDS 1	😮 Fa	ail Found non-	-canonical sequence GG					
	Feature Coordinates: 17358714-17358715 Surrounding Sequence: CCGTGGGTAAGT									
	Q	Acceptor for CDS 2	O P	ass						
	Featur Surro	re Coordinates: 17358840-17358841 unding Sequence: CCTGCAGGCAAA								

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.

C	hecklis	t Dot Plot Transcript	Sequence	Peptide Sequence Extra	acted Coding Exons Do	ownloads			
ŧ	Expand All E Collapse All								
	Vi	Exon	Phase	Start	End	Orientation	Length		
	Q	CDS_1	0	17358666	17358713	+	48		
	CDS MPT	S Translation: KERNIAMMGYRSV							
	Q	CDS_2	0	17358842	17358912	+	71		
	CDS ANR	S Translation: RYAYSS*RASSWTPMTPLSR							
	Q	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 "17358666-1735871<u>4</u>" (Figure 9).

nfigure Gene Model		
Project Details		
pecies Name:	D. yakuba	~
enome Assembly:	May 2011 (WUGSC dyak_caf1/DyakCAF1)	~
Scaffold Name:	chr3R	
Ortholog Details	Rheh-PA	
Ortholog Details Ortholog in D. melanogaster:	Rheb-PA	
Ortholog Details Ortholog in D. melanogaster: Iodel Details	Rheb-PA	
Ortholog Details Ortholog in D. melanogaster: Model Details Errors in Consensus Sequence?	Rheb-PA 17,358,714 Yes No	

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.

C	hecklist	Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads				
ę.	Expand All	Colla	apse All							
	View	Criteria			Status	Message				
ŧ	Q	Check for	Start Codon		© Pass					
Ð		Acceptor	for CDS 1		Skip Already checked for Start Codon					
Ŧ	Q	Donor for	CDS 1		Pass					
ŧ	9	Acceptor	for CDS 2		Pass	© Pass				
ŧ	9	Donor for	CDS 2		Fail	Found non-canonical sequence AG				
ŧ	0	Acceptor	for CDS 3		Fail	Found non-canonical sequence CC				
Ð	Q	Donor for	CDS 3		© Pass					
Đ	0	Acceptor	for CDS 4		Fail	Found non-canonical sequence GT				
ŧ	9	Donor for	CDS 4		© Pass					
ŧ	9	Acceptor	for CDS 5		© Pass					
ŧ		Donor for	CDS 5		Skip	Skip Already checked for Stop Codon				
ŧ	9	Check for	Stop Codon		Pass	© Pass				
ŧ	9	Additional	I Checks		Fail	Found premature stop codons in translation				
Ð	9	Check for	in-frame stop codons in	CDS_1	@ Pass					
ŧ	0	Check for	in-frame stop codons in	CDS_2	Pass					
ŧ	0	Check for	in-frame stop codons in	CDS_3	Pass					
ŧ	0	Check for	in-frame stop codons in	CDS_4	Pass					
ŧ	0	Check for	in-frame stop codons in	CDS_5	Fail	Found in-frame stop codons				
Đ	٩	Length of	translated region should	be multiples of 3	Fail	Length of in-phase coding region: 542 Number of extra nucleotides: 2				
21	6	Number	f 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  $\bigcirc$  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).



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 17358842-1735891<u>3</u>. 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).

Gene Model Checker	•								
Configure Gene Model			C	hecklist	Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads
Project Details			5	Expand A	I   E Colla	apse All			
Species Name:	D. yakuba	~		View	Criteria			Status	Message
Genome Assembly:	ome Assembly:         May 2011 (WUGSC dyak_caf1/DyakCAF1)            fold Name:         chr3R				Check for	r Start Codon		Pass	
Scaffold Name:					Acceptor	for CDS 1		Skip	Already checked for Start Codon
			۲	٩	Donor for	CDS 1		Pass	
Ortholog Details			۲	Q	Acceptor	for CDS 2		Pass	
Ortholog in D. melanogaster:	Ortholog in D. malanogastar: Phoh-PA 17.259.012				Donor for	CDS 2		Pass	
or a role of the anti-gastari		17,558,915	۰	Q	Acceptor	for CDS 3		Fail	Found non-canonical sequence CC
Model Details			۲	۹,	Donor for	CDS 3		Pass	
Errore in Concensus Sequence?	Vac	No.	۲	۹,	Acceptor	for CDS 4		Fail	Found non-canonical sequence GT
			۲	۹,	Donor for	CDS 4		Pass	
Coding Exon Coordinates:	17358666-17358714, 17358842-17358913, 17359013-17359218, 17359279-17359407,			٩	Acceptor	for CDS 5		Pass	
	17359470-17359556		۲		Donor for	CDS 5		Skip	Already checked for Stop Codon
			۲	Q.	Check for	r Stop Codon		Pass	
Annotated Untranslated	Yes	O No	۲	۹,	Additional	l Checks		Fail	Found premature stop codons in translation
Regions?			۲	Q,	Check for	r in-frame stop codons in	CDS_1	Pass	
Orientation of Gene Relative to	Plus	Minus	۲	۹,	Check for	r in-frame stop codons in	CDS_2	Pass	
Query Sequence:			۲	۹, –	Check for	r in-frame stop codons in	CDS_3	Fail	Found in-frame stop codons
Completeness of Gene Model Translation:	<ul> <li>Complete</li> </ul>	Partial	۲	Q,	Check for	r in-frame stop codons in	CDS_4	Pass	
Stop Coden Coordinater:	17250557-17250550		۲	۹	Check for	r in-frame stop codons in	CDS_5	Pass	
Stop Couon Coordinates.	1/33933/-1/339339		۲	۹	Number of	of coding exons matched	ortholog	Pass	
		Verify Gene Model Reset Form							

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).

С	hecklist	Dot Plot Transcript Sequence Peptide Sequ	uence Extracted	d Coding Exons Downloads				
ŧ	Expand A	II   E Collapse All						
	View	Criteria	Status	Message				
*	Q	Check for Start Codon	Pass					
+		Acceptor for CDS 1	Skip	Already checked for Start Codon				
*	Q	Donor for CDS 1	Pass					
+	Q	Acceptor for CDS 2	Pass	Pass				
ŧ	Q	Donor for CDS 2	Pass					
ŧ	Q	Acceptor for CDS 3	😮 Fail	Found non-canonical sequence CC				
+	Q	Donor for CDS 3	Pass					
ŧ	Q	Acceptor for CDS 4	🚳 Fail	Found non-canonical sequence GT				
+	Q	Donor for CDS 4	Pass					
ŧ	Q	Acceptor for CDS 5	Pass					
ŧ		Donor for CDS 5	Skip	Already checked for Stop Codon				
ŧ	Q	Check for Stop Codon	Pass					
*	Q	Additional Checks	😮 Fail	Found premature stop codons in translation				
+	Q	Check for in-frame stop codons in CDS_1	Pass					
+	Q	Check for in-frame stop codons in CDS_2	Pass					
+	Q	Check for in-frame stop codons in CDS_3	🚳 Fail	Found in-frame stop codons				
+	Q	Check for in-frame stop codons in CDS_4	Pass					
+	Q	Check for in-frame stop codons in CDS_5	Pass					
ŧ	Q	Number of coding exons matched ortholog	Pass					

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 <u>F Element</u> <u>project page</u> 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).

Gene Record Fin	der						FlyBase Rel	lease	6.37 - (Last Update: 12/30/2020)
			Search D	). melanogast	er Gene Reco	ds:			
	Gene Re	cord Finde	Rheb	Rheb					
			Rheb	(ImR)	VA: 2, Bexons: 7, IK	05: 5)			
				<u>GEP  </u>	Home Page				
					•				
Franscript Det	ails Poly	peptide De	tails						
Options: Ex	port All Unio	ue CDS to I	TASTA	Export All	CDS for Sele	ected Isofo	rm to FAST	٨	Download CDS Workbook
-	pore All only		ASIA	Export Au	cos for sete			<u> </u>	Dominoud CDD Monibook
CDS usage m	nap:								
Isoform	1_9850_0	2_9850_2	3_9850_2	4_9850_1	5_9850_0				
Rheb-PA	1	2	3	4	5				
Bhah DP	1	2	3	4	5				
RIED-PB									
soforms wit	h unique c	oding exo	ns:						
Isoforms wit	h unique c oform(s) b	oding exc ased on c	ns: oding seq	luence	Other is	oforms v	vith ident	ica	l coding sequences

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.

			_	·				
Options: Ex	port All Uniq	ue CDS to F	ASTA	Export All (	CDS for S	lected Isoform to FASTA	Download CDS Work	book
DS usage m	iap:							
lsoform	1_9850_0	2_9850_2	3_9850_2	4_9850_1	5_9850_			
Rheb-PA	1	2	3	4	5			
Rheb-PB	1	2	3	4	5			
Unique iso	oform(s) ba	ased on co	oding sea	uence	Other	soforms with identica	l coding sequences	
Unique iso Rheb-PA	oform(s) ba	ased on co	oding seq	uence	Other Rheb-I	soforms with identica B	l coding sequences	
Unique iso Rheb-PA elect a row	oform(s) ba	ased on co	oding seq	cDS seque	Other Rheb-I	soforms with identica B	l coding sequences	eb:3.9850.2
Unique iso Rheb-PA elect a row FlyBase ID	oform(s) bi to display 5' Start	ased on co the corres 3' End	sponding seq	CDS seque	Other Rheb-I ence: e = S	soforms with identica B ze (aa) 0	l coding sequences viewer for Rheb: Rhe	eb:3_9850_2
Unique iso Rheb-PA elect a row FlyBase ID 1_9850_0	to display 5' Start 5,569,223	the corres 3' End 5,569,2	sponding seq sponding of Strai 71 +	CDS seque nd Phase 0	Other Rheb-I ence: e = S	soforms with identica B ze (aa) Sequence PRt Bay	l coding sequences viewer for Rheb: Rh 9850_2 SODY TVKLI DTAGODE	eb:3_9850_2
Unique iso Rheb-PA elect a row FlyBase ID 1_9850_0 2_9850_2	oform(s) bi to display 5' Start = 5,569,223 5,569,400	ased on cc the corres 3' End 5,569,2 5,569,4	sponding seq sponding of Strain 71 + 71 +	CDS seque nd Phase 0 2	Other Rheb-I ence: ie = S 1 2	soforms with identica B Ze (aa) Sequence PRheb:3 PTKIERVI FEVVKIII	l coding sequences viewer for Rheb: Rh 9850_2 SQDYIVKLIDTAGQDE EKLLDVMGKK	eb:3_9850_2 YSIFPVQYSMDYHGYVLVYSITSQH
Unique isc Rheb-PA elect a row FlyBase ID 1_9850_0 2_9850_2 3_9850_2	5 Start = 5,569,223 5,569,400 5,569,575	ased on cc the corres 3' End 5,569,2 5,569,4 5,569,7	sponding seq	CDS seque nd Phase 2 2	Other Rheb-l ence: ie = 5 1 2 6	soforms with identica B Sequence PRheb:3 PTKIERVI FEVVKII	l coding sequences viewer for Rheb: Rh 9850_2 SQDYIVKLIDTAGQDE EKLLDVMGKK	eb:3_9850_2 YSIFPVQYSMDYHGYVLVYSITSQH
Unique isc Rheb-PA elect a row FlyBase ID 1_9850_0 2_9850_2 3_9850_2 4_9850_1	5 Start 5 5,569,223 5,569,400 5,569,842	ased on cc the corres 3 End 5,569,2 5,569,4 5,569,9 5,569,9	sponding seq sponding 271 + 171 + 171 + 182 + 171 +	CDS seque nd Phase 0 2 2 1	Other Rheb-l ence: :e = S 1 2 6 4	soforms with identica B Ze (aa) PRheb: 3 PrKIERVI FEVVKIIS	l coding sequences viewer for Rheb: Rh 9850_2 SQDYIVKLIDTAGQDE EKLLDVMGKK	eb:3_9850_2 YSIFPVQYSMDYHGYVLVYSITSQP

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 <u>NCBI *BLAST* home page</u>. 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 cicon 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 <u>CM000160.2</u> (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).

Line accession	number(s), gi(s), or FASTA sequence(s) 🥹	Clear	Query	subrange	9
<pre>&gt;Rheb:3_9850_2 FTKIERVKSQDYIVKL FEVVKIIYEKLLDVMG</pre>	IDTAGQDEYSIFPVQYSMDYHGYVLVYSITSQKS KK		From		
CDS 3 of Rh	eb-PA from <i>D. melanogaster</i>	11.	То		
Or, upload file	Browse No file selected.				
Job Title	Rheb:3_9850_2				
	Enter a descriptive title for your BLAST search 😡				
Align two or m	ore sequences 😡				
Align two or m Enter Subject	ore sequences 🛞				
Align two or m Enter Subject Enter accession	ore sequences 😡 Sequence number(s), gi(s), or FASTA sequence(s) 😡	Clear	Subjec	t subrange	9
Align two or m Enter Subject Enter accession CM000160.2	ore sequences 😥 Sequence number(s), gi(s), or FASTA sequence(s) 😡	Clear	Subjec	t subrange 7300000	Limit the search region
✓ Align two or m Enter Subject Enter accession ™000160.2 chr3R in <i>D</i> . 1	ore sequences 😡 Sequence number(s), gi(s), or FASTA sequence(s) 😡	Clear	Subjec From 17 To 17	t subrange 7300000 7400000	Limit the search region within <i>D. yakuba</i> chr31

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.

Algorithm paramete	rs Note: Parameter values that differ from the default are highlighted in yellow and marked with $\phi$ sign
General Parar	neters
Max target sequences	100 C Select the maximum number of aligned sequences to display 🕑
Expect threshold	0.05 😡
Word size	3 3
Max matches in a query range	0
Scoring Paran	neters
Matrix	BLOSUM62 💈 🥹
Gap Costs	Existence: 11 Extension: 1 🛛 😌 😣
Compositional adjustments	No adjustment
Filters and Ma	sking
Filter	♦ Dow complexity regions ⊌
Mask	<ul> <li>Mask for lookup table only </li> <li>Mask lower case letters </li> </ul>

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).

### La Download GenBank Graphics

		and discussion of the second second second				
Drosophila yaku	uba stra	ain Tai18E2 chro	omosome 3R, w	hole genome	e shotgun s	sequence
Sequence ID: CMU	JU100.2	Length: 20032112	Number of Match	es: I		
Range 1: 17359013	8 to 1735	9216 GenBank Gra	phics	V N	lext Match 🔺 P	Previous Match
Score	Expect	Identities	Positives	Gaps	Frame	
137 hits(344)	40-47	68/68(100%)	68/68(100%)	0/68(0%)	+2	
107 010(011)	10 12	00,00(100 /0)	00,00(100,0)	0/00(070)		
Ouerv 1	FTKIE	RVKSODYIVKLIDTA	GODEYSIFPVOYSMD	YHGYVLVYSITS	OKSFEVVKII	YEK 60
~	FTKIE	RVKSQDYIVKLIDTA	GQDEYSIFPVQYSMD	YHGYVLVYSITS	QKSFEVVKII	YEK
Sbjct 17359013	FTKIE	RVKSQDYIVKLIDTA	GQDEYSIFPVQYSMD	YHGYVLVYSITS	QKSFEVVKII	YEK 17359192
Query 61	LLDVM	GKK 68	Query Descr	Rheb-3 9850 2		
Shigt 17250102	LLDVM	GKK 17250216	Query Deser	11165.5_5656_2		
SDJCC 17559195		GKK 17559210	Query Length	68		
			Subject ID	CM000160.2 (dna	)	
			Subject Descr	Drosophila yakub	a strain Tai18E	2 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 sicon next to "Acceptor for CDS 3" checklist item to reopen 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).

	as [	+	
chr3R (chr3	R) 3Rchr3R		
Scale chr3R: > T	20 bases 17,359,990 17,359,000 17,359,010 G C G T A A T G A T C G G T A A T T C A C A G A S V P H S	DyakCAF1 17,359,020  17,359,030  17,359,040  ССТТСАСААА GATC GAGC GC GTAAA GTC GCAA L H K D R A R K V A	
	R N D R F P F T C V M I G N S H S Q Custom Gene t	P S Q R S S A S R K for DyakCAF1	Custom
elenen -uner u A	Gap	Loc. tions	custom
Gap	Spain Alignment of	D. melanogaster Proteins	Gene Mod
Rheb-PB	· · · · · · · · · · · · · · · · · · ·		
Dileber of M	GeMoMa (	ene Predictions	
FBTR0078693_R0 🔇	Genscan (	Predictions	
chr3R.3863			
_geneid-chr3R_2128	Geneid G	ne redictions.	
chr38 g1834 11 (2)	Augustus (	ene Predictions	
1045 _	modENCODE RNA	Sea from Adult Females	
ult Females			
820 _	modENCODE RN.	-Se from Adult Males	
11	Splice Junctions Predicted by	regt ols using D. vakuba RNA-Seg	
JUNC00073912 🛠	and an		
20MC00013313	Depending Flow	nts v DeneatMasker	

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).



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).

Sequence ID: CMO	00160.2	Length: 28832112	2 Number of Matche	es: 1	
Range 1: 1735884	4 to 1735	8912 GenBank Gra	aphics	V Nex	t Match 🔺 Previous Match
Score 51.2 bits(121)	Expect 8e-13	Identities 23/23(100%)	Positives 23/23(100%)	Gaps 0/23(0%)	Frame +1
Query 1	KSSLC KSSLC	IQFVEGQFVDSYDP1 IQFVEGQFVDSYDP1	TIEN 23	Query Descr	Rheb:2_9850_2
Sbjct 17358844	KSSLC.	IQFVEGQFVDSYDP1	TEN 17358912	Query Length	23
				Subject ID	<u>CM000160.2</u> (dna)
				Subject Descr	Drosophila yakuba strain Tai18E2 chromosome 3R
				Subject Length	100000

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

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 "1735901<u>1</u>-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).

Configure Gene Model			44	C	hecklist	Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads	
Project Details				<b>£</b> 2	Expand A	II   💽 Colla	apse All				
Species Name:	D. yakuba		~		View	Criteria			Status	Message	
Genome Assembly:	May 2011 (WUGS	C dyak_caf1/DyakCAF1)	~	Ŧ	٩	Check for	Start Codon		Pass		
Scaffold Name:	chr3R	chr3R				Acceptor for CDS 1		Skip	Already checked for Start Codon		
				*	Q.	Donor for	CDS 1		Pass		
Ortholog Details				Ŧ	Q	Acceptor	for CDS 2		Pass		
Ortholog in D. melanogaster:	plog in D. melanogaster: Rheb-PA			٠	Q.	Donor for	CDS 2		Pass		
gunteri		NIEUTA				Acceptor	for CDS 3		Pass		
Model Details				±	Q	Donor for	CDS 3		Pass		
Errors in Consensus Secuence?	Vec	O No			Q	Acceptor	for CDS 4		Fail	Found non-canonical sequence GT	
Cadica Free Condicate	17050666 17050	VIA 17350042 17350012			٩	Donor for	CDS 4		Pass		
Coding Exon Coordinates:	17358666-17358, 17359011-173592	/19, 1/358892-1/358913, 218, 17359279-17359407,			Q	Acceptor	for CDS 5		Pass		
7 359 011	17359470-173595	556		Ŧ		Donor for	CDS 5		Skip	Already checked for Stop Codon	
,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,,				Ŧ	Q	Check for	Stop Codon		Pass		
Annotated Untranslated	) Yes	O No			Q	Additional	I Checks		Fail	Found premature stop codons in translation	
Regions?		-			Q	Check for	in-frame stop codons in	CDS_1	Pass		
Orientation of Gene Relative to	Plus	Minus		۲	Q	Check for	in-frame stop codons in (	CDS_2	Pass		
Query Sequence:					Q	Check for	in-frame stop codons in	CDS_3	Pass		
Completeness of Gene Model Translation:	<ul> <li>Complete</li> </ul>	Partial		Ŧ	Q.	Check for	in-frame stop codons in	CDS_4	Pass		
Stop Codon Coordinates:	17350557-173505	150		Ŧ	٩	Check for	in-frame stop codons in t	CDS_5	😮 Fail	Found in-frame stop codons	
Stop Codon Coordinates:	1/359557-17359	009 		Ŧ	٩	Length of	translated region should	be multiples of 3	Fail	Length of in-phase coding region: 545 Number of extra nucleotides: 2	
					Q	Number o	of coding exons matched	ortholog	Pass		

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 17359	407 GenBank Gra	aphics	Vex	t Match 🔺 Previous Match
Score 80.1 bits(196)	Expect 1e-22	Identities 40/43(93%)	Positives 40/43(93%)	Gaps 0/43(0%)	Frame +1
Query 1 Sbjct 17359279	VPVVLV VPVVLV	GNKIDLHQERTVST GNK DL ERTVST GNKTDLQPERTVST	'EEGKKLAESWRAAFI 'EEGKKLAESWRAAFI 'EEGKKLAESWRAAFI	JETSAKQNE 43 JETSAKQNE 1735	9407
				Query Descr	Rheb:4_9850_1
				Query Length	43
				Subject ID	CM000160.2 (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 "1735927**8**-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).

GEP Gene Model Checker										
Configure Gene Model		<<	Cł	necklist	Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads	
Project Details			-	Expand All	E Colla	apse All				
Species Name:	D. yakuba 👻			View	Criteria			Status	Message	
Genome Assembly:	May 2011 (WUGSC dyak_caf1/DyakCAF1)			Q	Check for	Start Codon		Pass		
Scaffold Name:	chr3R		Ħ		Acceptor for CDS 1			Skip	Already checked for Start Codon	
					Donor for CDS 1					
Ortholog Details			H	Q	Acceptor	for CDS 2		Pass		
Ortholog in D. melanogaster:	Rheb-PA		H	Donor for CDS 2				Pass		
				Acceptor for CDS 3		Pass				
Model Details			Ħ	Donor for CDS 3		Pass				
Errors in Consensus Sequence2	Ver No		Ħ	Q	Acceptor f	for CDS 4		Pass		
Errors in consensus sequence:			Ħ	0	Donor for	CDS 4		Pass		
Coding Exon Coordinates:	17358666-17358714, 17358842-17358913, 17359011-17359218, 17359278-17359407, 17359470-17359556			9	Acceptor f	for CDS 5		Pass		
					Donor for	CDS 5		Skip	Already checked for Stop Codon	
	17 359 278			Q	Check for	Stop Codon		Pass		
Annotated Untranslated	Ves 0 No			0	Additional	Checks		Pass		
Regions?			Ħ	٩	Number o	f coding exons matched	ortholog	Pass		
Orientation of Gene Relative to Query Sequence:	Plus     Minus									
Completeness of Gene Model Translation:	Complete Partial									
Stop Codon Coordinates:	17359557-17359559									

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).

 Checklist
 Dot Plot
 Transcript Sequence
 Peptide Sequence
 Extracted Coding Exons
 Downloads

 View protein alignment
 View dot plot in the Dot Plot Viewer
 View
 Vie





Dmel\_Rheb-PA

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

#### <u>View plain text version</u> Download alignment image

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



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  $\leq$  0.5. A space indicates that the aligned residues have a score  $\leq$  0. See the "<u>Bioinformatics Tools FAQ</u>" 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).

GEP Gene Model Checker	F)												
Configure Gene Model				Ch	hecklist	Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Coding Exons	Downloads			
Project Details				Expand All   E Collapse All									
Species Name:	D. yakuba		~	View Criteria				Status	Message				
Genome Assembly:	May 2011 (WUGSC dyak_caf1/DyakCAF1)				Check for Start Codon								
Scaffold Name:	chr3R			۲	Acceptor for CDS 1 Skip				Already checked for Sta	Already checked for Start Codon			
					9	Donor for CD							
Ortholog Details						Acceptor for	CDS 2	© Pass					
Ortholog in D. melanogaster:	Rheb-PA		Donor for CDS 2			Pass Pass	© Pass						
					Donor for CDS 3			© Pass	Pass				
Model Details					Q	Acceptor for	CDS 4	© Pass					
Errors in Consensus Sequence?	⊖ Yes	O No			Q	Donor for CD	IS 4	O Pass					
Coding Exon Coordinates:	17358666-17358714, 17358842-17358913, 17359044-17359218, 17359278-17359407.				Q	Acceptor for	CDS 5	Pass					
	17359470-17359556					Donor for CD	S 5	Skip	Already checked for Sto	p Codon			
				۲	Q	Check for Sto	op Codon	Pass					
Annotated Untranslated	) Yes	O No			Q	Additional Ch	necks	Pass					
Regions?					Q	Number of co	oding exons matched orth	holog 📀 Pass					
Orientation of Gene Relative to Query Sequence:	O Plus	O Minus											
Completeness of Gene Model Translation:	<ul> <li>Complete</li> </ul>	Partial											
Stop Codon Coordinates:	17359557-17359559												

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



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).

#### La Download GenBank Graphics

Drosophila yakuba strain Tai18E2 chromosome 3R, whole genome shotgun sequence Sequence ID: <u>CM000160.2</u> Length: 28832112 Number of Matches: 1

Coro		Export	Identities	Pocitivos	Cans	Eramo	
137 hit	s(344)	4e-42 68/68(100%)		68/68(100%)	0/68(0%)	+2	
			00,00(100,0)	00,00(10070)	0,00(0,0)		-
Query	1	FTKIER	VKSQDYIVKLIDTA VKSODYIVKLIDTA	GODEYSIFPVQYSMD CODEYSIFPVQYSMD	YHGYVLVYSITS	OKSFEVVKI	IYEK 60 IYEK
Sbjct	17359013	FTKIER	VKSQDYIVKLIDTA	GQDEYSIFPVQYSMD	YHGYVLVYSITS	OKSFEVVKI	IYEK 17359192
Query	61	LLDVMG	KK 68	Query Descr	Rheb:3_9850_2		
Sbjct	17359193	LLDVMG	KK 17359216	Query Length	68		
				Subject ID	<u>CM000160.2</u> (dna	)	
				Subject Descr	Drosophila yakub	a strain Tai18	E2 chromosome 3R
				Cubicat Longth	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 start 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).

chr3R (chr3R)	3Rchr3R	
Scale chr3R:   > REMITPICE HIS NA REPETUHACIAND ISTONDISMUATO	50 pases 17,359,000 17,350,000 17,350,0	
DyakCAF1_Rheb-PA	Gan Locations	Bana Mod
Gap	Spale Alignment of D. malapaparter Proteins	Gene wood
Rheb-PB	Spain Alignment of D. Petenogaster Proteins	
Rnep-PA	GeMoMa Gene Predictions	
FBTR0078693_R0 <k< td=""><td>Genscan Gene Predictions</td><td></td></k<>	Genscan Gene Predictions	
chr3R_3883	Geneid Gene Predictions	<b>**</b>
_geneld-chr3R_2128	Augustus Gene Predictions	→
chr3R.g1834.t1	modENCODE RNA-Sea from Adult Eamples	
ilt Females		
5	modENCODE RNA-Sea from Adult Males	
ult Males		
6	Splice Junctions Predicted by realools using D, vakuba RNA-Seg	
JUNC00073912	UNC00073914 (****	
	jui	NC00073915 >>>>>>>>

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 "173590<u>11</u>". 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).

				~ (	hecklist	Dot Plot	Transcript Sequer	nce Peptide Se	equence	Extracted Coding Exons	Downloads	
Project Details					Expand Al	Colla	lapse All					
ipecies Name:	D. yakuba		~		View	Criteria				Status	Message	
enome Assembly:	May 2011 (WUGSC dyak	caf1/DyakCAF1)	~		6	Check for	r Start Codon			Pass		
caffold Name:	chr3R					Acceptor	for CDS 1			Skip	Already che	cked for Start Co
					5	Donor for	CDS 1			Pass		
Irtholog Details					e.	Acceptor	for CDS 2			Pass		
tholog in D. melanogaster:	melanogaster: Rheb-PA			Donor for CDS 2					© Pass			
					9	Acceptor	for CDS 3			© Pass		
odel Details					9	Donor for	r CDS 3			© Pass		
rors in Consensus Sequence	? 🔿 Yes	O No		3	6	Acceptor	for CDS 4			© Pass		
ding Exon Coordinates:	17358666-17358714, 13	7358842-17358913,		3	4	Donor for	r CDS 4			© Pass		
	17359011-17359218, 17	7359278-17359407,		3	•	Acceptor	for CDS 5			© Pass		
,359,011	1/3234/0-1/323220			8		Donor for	r CDS 5			Skip	Already che	cked for Stop Co
				3	6	Additional	I Chacks			Pass		
notated Untranslated	O Yes	O No			0	Number o	of coding exons mate	thed ortholog		© Pass		
ientation of Gene Relative to	o O Plus	Minus					0	ecklist Det Plot	Transcript	Sequence Peptide Sequence	e Extracted Cor	Sing Exons Down
and another and a sector and a								the strategy of the				
ompleteness of Gene Model anslation:	<ul> <li>Complete</li> </ul>	O Partial					2	iew dot plot in t	the Dot Pla	N Viewer		
ompleteness of Gene Model ranslation: top Codon Coordinates:	Complete 17359557-17359559	O Partial					3	iew dot plot in t	the Dot Plo Dot. 1	<u>X Viewer</u> plot of Dwol_Mond-PA vs	, DyakOVF1_Rbv	ub-PA
iompiteness of Gene Model ranslation: ignment of Dmm w.plain text version wnload alignment ima ntity: 177/182 (97.39	Complete 17359557-17359559 Hel_Rheb-PA v 1882 (5), Similarity: 178/1	Partial 7 <b>s. DyakCAF</b> 182 (97.8%), <b>Gaps</b>	<b>1_Rheb-</b>	PA %)			3	sew protein any few dot plot in 1 se- se- se- se- se- se- se-	omen the Dot Pic	<u>R Viewer</u> olot. of Deol.,#Heb-FN ve	. Dyssk0f#1.99v	æ-m
mpleteness of Gene Model anslation: gp Codon Coordinates: gpinnent of Dmn v plain text version nload alignment ima tity: 177/182 (97.39 Rheb-PA .CAF1_Rheb-PA	Complete     17359657-17359559      tel_Rheb-PA v      ige     (), Similarity: 178//1      MPPREBNIAMMGYR      MPPREBNIAMMGYR	Partial rs. DyakCAF 182 (97.8%), Gaps svc2551/c101/v/2001	1_Rheb-	PA %)	CRVKSQ SRVKSQ	YIVKLI	D7 60		Dot. 1	8.Viewer	, DysA0F1,99	8-M
molectness of Gene Model instalation: gnment of Dmr plain text version nload alignment ima tity: 177/182 (97.39 _Rheb-PA CAF1_Rheb-PA Bheb-PA	Complete     17359557-17359559  el_Rheb-PA v  ge  ), Similarity: 178//  MPTKERNI AMMOUR  MPTKERNI AMMOUR  ) MPTKERNI  ) MPTKERNI ]	Partial rs. DyakCAF 182 (97.8%), Gaps svo_csurde(a)/2010 svo_csu	1_Rheb- : 0/182 ( 0.0/ NOSYDPTIES NOSYDPTIES	PA %)	ERVKSQ ERVKSQ	DYIVKLI	DT 60 DT 60 DT 60	iew dot plot in J so- so- so- so- so- so- so- so- so- so-	Dot. 1	et Viceour	. 0,04091,99	****
molecteness of Gene Model imilation: cplain text version nload alignment ima tity: 177/182 (97.39 Rheb_PA Rheb_PA Rheb_PA Rheb_PA	Complete     17359657-17359559      IZ359657-17359559      Ige     Similarity: 178//1 <u>MPTKERNTAMMGYR:</u> AGODEYSIFEVOXS     AGODEYSIFEVOXS	Partial rs. DyakCAF 182 (97.8%), Gaps 559753(04/200) 559753(04/200) 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 18076070107521550 1807607010752 1807607010752 1807607010750 1807607010750 18076070 18076070 18076070 18076070 1807607	1_Rheb- : 0/182 ( 0.0 2053/024432 2053/024432 SPEVVKL142 (SPEVVKL142	PA %) FTKI KLLDV	SRVKSQ SRVKSQ 4GKKY 4GKKY	YIVKLI YIVKLI WWWGN	22 60 127 60 120 120 120 120	100 georgia aug 100 dot plot in 1 100 dot plot in 1 100 - 100	Dot. 1	el Licent	. DJAACHTLRH	a-94
mpleteress of Gene Model institution: gnment of Dm plain text version nload alignment ima tity: 177/182 (97.39 _Rheb=PA _Rheb=PA _Rheb=PA _Rheb=PA { _Rheb=PA { _Rheb	Complete     17359557-17359559      tel_Rheb-PA v      tel_Rheb-P	Partial rs. DyakCAF 182 (97.8%), Gaps svc_551240(w/x00) MDHGWLWS17500 MDHGWLWYS17500 MDHGWLWWYS175000 MDHGWLWWYS17500 MDHGWLWWYS175000 MDHGWLWWYS17	1_Rheb- : 0/182 ( 0.0 : 0/182 ( 0/	PA %)	ERVKSQ ERVKSQ 4GKKY 4GKKY 4GKKY 4 ENENG2	YIVKLI YIVKLI WWWWGN	5 60 107 60 107 50 120 55 120 55 120	200 - Rooten ang 200 - Soo plot in 1 200 - 100 200 -	omena the Dot Pic	st Microsoft	. 004001.99	de P3

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 (<u>control-click on macOS</u>) 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 <u>Annotation Files Merger</u> 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 <u>Sequence Updater</u> tool to enable annotators to document errors in the consensus sequence. The <u>Sequence Updater</u> will generate a file in the <u>Variant</u> <u>Call Format</u> (VCF) that describes changes to the original project sequence in a standardized format. The <u>Gene Model Checker</u> 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 <u>Sequence Updater User Guide</u> 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).

onfigure Gene Model		<	C	hecklist	Dot Plot	Transcript Seque	ice Peptide S	Sequence	Extracted Co	oding Exons	Downloads	
Project Details			1	Expand		apse All						
Species Name:	D. sechellia 👻			View	Criteria			Statu	us	Message		
Genome Assembly:	May 2011 (Broad dsec_caf1/DsecCAF1)		Ŧ	Check for Start Codon				OP.	ass			
Scaffold Name:	super_0			Acceptor for CDS 1			Skip		Already che	ecked for Start C	odon	
				0	Donor for Check for	CDS 1 Stop Codon		Skip	2255	Already che	ecked for Stop Ci	odon
Ortholog Details				0	Additional	Checks		ØF	ail	Found pren	nature stop codo	ns in translation
Ortholog in D. melanogaster:	tgo-PA			Q	Check for	n-frame stop codon	in CDS_1	ØF	ail	Found in-fr	ame stop codons	
Model Details				٩	Length of t	ranslated region sh	uld be multiples	of 3 😳 Fa	ail	Length of in Number of	n-phase coding re extra nucleotides	gion: 1924 : 1
Errors in Consensus Sequence?	Yes O No			۹	Number of	coding exons matc	ed ortholog	Ø Pass				
Joing Exon Coordinates.	g Exon Coordinates: 17034485-17036408			klist >Ds	Dot Plot	Transcript Se	uence Po	eptide S	Sequence	Extracte	d Coding Exo	ns Downloa
Annotated Untranslated Regions? Drientation of Gene Relative to Query Sequence: Completeness of Gene Model	Yes O No Plus Minus O Complete Partial	1 MDEANIQDKERFASRENHCEIERRRRNKMTAYITELSDMVPTCSALARKPDKLTILR 61 AHMKALRGTCNTSSDGTYRPSFLTDQELKHLILEAADGFLFVVSCDSGRVITVSDSV 121 LNYTQSDWYGTSLYEHIHPDDRKKIREQGSTQESQNAGRIDLLKSGTVKKEGQSSM 181 MGARRGFICHRRVGNVNPESMVSGHLNRLKQRNSLGPSRDGTNYAVVHCTGYIKNMP 241 MFPNMHMERDVDDMSSHCCLVAIGRLQVTSTAANDMSGSNNQSEFITRHAMGKFTF 301 RVLNILGYTPTELLGKICVDFHPEDQSHMKESFDQVLKQKGQMF5LLYRARAKNSE								RMAV VTPV MRLS PPTD FVDQ EEYVW		
nslation: yp Codon Coordinates: 17036409-17036411				421 AFGVDYARRELTPVGSATNDGMYQTHMLAMQAPTPQQQQQPGSQTTPVGYTVDTTHS PYSAGGPSPLAKIPKSGTSPTPVAPNSWAALRPQQQQQQQPVTEGYQYQQTSPARSPSG 541 PTTTQLSACNGKSPASAACSISGGSTTASQCAGNVGLAAGWKSPASAAPNGASAPSARSS 601 RWTGRSRTAAGSCVLRYAADVGSHADHV*GSEVOHDAVR								TTHS SPSG ARSS
							Stop	codo	n			

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).

Program	BLASTN	Citatio	<u>n</u> 💙								
Database	SRA S	ee details v	•	D	Database: SRA experiment SRX287399 ( <i>D. sechellia</i> adult female RNA-Seq)						
Query ID	CH48081	5.1									
Description	Drosoph	ila sechellia	strain Rob3c scaffold_0 ge	nomic scaffold							
🛓 Down	load •	<u>Graphics</u>	<u>SRA</u>								
SRX28	7399										
Sequenc	e ID: SRA:S	RR86960	1.2846847.2 Length	: 100 Number of M	latches: 1						
						-					
Range 1	: 1 to 100 G	raphics		Ext	ra G in super_0	Previous Match					
Score		Expect	Identities	Gaps	Strand						
180 bits	(97)	9e-44	100/101(99%)	1/101(0%)	Plus/Plus						
Query	17036100	GCGGTCC	ААССТАТАСАСААТТGA	GTGCCGGGAACGGGA	AATCGCCAGCAAGC	GCAGCCG 17036159					
Sbict	1	GCGGTCC			AATCGCCAGCAAGC	 GCAGCCG 59					
	_										
Ouerv	1	aa									
A	17036160	GGAGCAT	ATCAGGCGGGTCCACCA	CCGCCTCCCAATGCG	CC 17036200						

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	D. sechellia
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.

#### **Genomics Education Partnership**

GEP Gene Model Checker												
Configure Gene Model				« C	hecklist	Dot Plot	Transcript Sequence	Peptide Sequence	Extracted Co	ding Exons	Downloads	
Project Details				<b>1</b>	Expand A	VI 🛛 🔳 Colla	pse All					
Species Name:	D. sechellia			View	Criteria			Status	Message			
Genome Assembly:	May 2011 (Broad dsec_caf1	/DsecCAF1)	~	+	Q	Check for S	tart Codon		🕑 Pass			
Scaffold Name:	super_0			۲	Acceptor for CDS 1				Skip	Already c	necked for Sta	t Codon
						Donor for CDS 1 Skij Check for Stop Codon Ol Additional Checks			Skip	Already cl	necked for Sto	Codon
Ortholog Details					0				© Pass			
Ortholog in D. melanogaster:	tgo-PA					Number of o	coding exons matched	ortholog	Pass			
Model Details				۲	Q	Modified Co	nsensus Sequence		🛦 Warn	Updated of	consensus seq	uence based on VCF file
Errors in Consensus Sequence? File with Changes to the Consensus Sequence: Coding Exon Coordinates:	Yes     C:\fakepath\DsecCAF1_tgo     17034485-17036408	No PA.vcf.bxt	Browse	Upl	oad	VCF f	ile			Updat seque	ed co nce w	nsensus arning
Annotated Untranslated Regions?	⊖ Yes	<ul> <li>No</li> </ul>										
Orientation of Gene Relative to Query Sequence:	Plus	O Minus										
Completeness of Gene Model Translation:	<ul> <li>Complete</li> </ul>	Partial										
Stop Codon Coordinates:	17036409-17036411											
		Verify Gene M	odel Reset Form									

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



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).