Pathways Project: Annotation Report

Note: You should also prepare the corresponding **GFF, transcript, and peptide sequence files** as part of your submission.

Student Name(s):	Red Akai
Student Email(s):	red@pallet_universty.edu
Faculty Instructor:	Samuel Oak
Course Name/Number:	POK898: How to Annotate Them All
College/University:	Pallet University, Kanto

Project Details

Project Species (e.g., *D. yakuba*)	D. yakuba
NCBI Taxonomy ID (e.g., 7245)	7245
NCBI Assembly ID (e.g., dyak_caf1)	dyak_caf1
Assembly Accession (e.g., GCA_000005975.1)	GCA_000005975.1
Genome Assembly (e.g., May 2011 (WUGSC dyak_caf1/DyakCAF1))	May 2011 (WUGSC dyak_caf1/DyakCAF1) Assembly
Scaffold Name (e.g., chr3R)	chr3R
Scaffold Accession (e.g., CM000160.2)	CM000160.2
Gene ID in target species (e.g., dyak_Rheb)	dyak_Rheb

Gene ID in *D. melanogaster* (e.g., dmel_Rheb)	dmel_Rheb
Accession Number of Ortholog in *D. melanogaster* (e.g., NT_033777)	NT_033777
Chromosome of Ortholog in *D. melanogaster*	3R

Date of Submission (YYYY-MM-DD)	2022/03/011

The data from this document will eventually be published, so we will need some contact information from you as well as permissions:

Permanent email address (e.g., one you will probably use 5 years from now):	red@indogo.league
Alternative email address (optional):	read@kanto.trainers
Cell phone number (optional):	+25 16913 1143

If you choose to be a co-author(s), you will have to respond promptly to requests to read and approve the manuscript, and, as part of that review, you will also be required to validate some specific data within the manuscript (full instructions will be provided). We estimate that you would contribute 3-5 hours of your time to the manuscript preparation process.

If you want to be a co-author(s) on a publication, and we cannot reach you at the time the publication is ready for your review, you will no longer be a co-author(s) on the publication that arises from this data because you are not able to read and approve the manuscript.

If you would like to be a co-author(s), please enter your name(s) in the format you want it/them to be displayed as in publications in the table below. Note: If more than three students contribute to an individual gene annotation report as a group project, none of those students are eligible for co-authorship but the class will be acknowledged.

Name(s) you want on the publication(s)	Isamu Akai

By submitting this report to the GEP, you acknowledge that you are allowing the data presented here to be published.

[✓ Yes]/No

By highlighting "Yes", I/we understand that my instructor may submit my/our Annotation Report and supporting documentation to the Genomics Education Partnership (GEP). The Report contains my/our name(s) and contact information. The Partnership may use my/our work in a publication. To be a co-author on any possible publications, I need to reply promptly to an email from the GEP when the manuscript is ready for review; if I do not, there is a chance my name will not be included as a co-author.

Navigation Pane

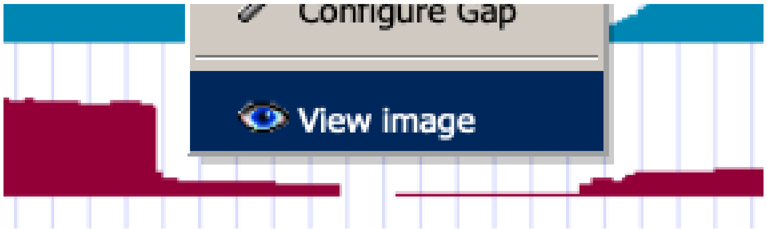
You will find it helpful to use the *Navigation Pane* function of Word while editing this document. It will be much easier to move around sections this way. Since Word versions vary in their layout, please follow the Use the Navigation pane in Word webpage to enable viewing of the navigation pane in your version of Word. Make sure to select the “Headings” tab within the pane to see the different sections in the document. Note that your version of MS Word might be different than what the website is portraying.

Entering Coordinates

When entering coordinates, do not use commas to separate place values – this will make it easier to move data to the Gene Model Checker. For example, use 10000 instead of 10,000.

Avoid Screenshots in the GEP UCSC Genome Browser

Avoid taking a screenshot of the GEP UCSC Genome Browser. We have incorporated a way to export the image. Right-click within the Genome Browser and select “View Image”. This should automatically take you to a new tab. If it doesn’t, make sure that popups are enabled for the GEP UCSC Genome Browser from within your internet browser.



Pathways Project: Annotation Videos

Refer to the annotation videos for details on filling out this report form. GEP Virtual TAs also provide real time support seven days a week for cases where your model is non-standard (TA schedule; obtain the Zoom link from your instructor).

Part A. Gene Report Form

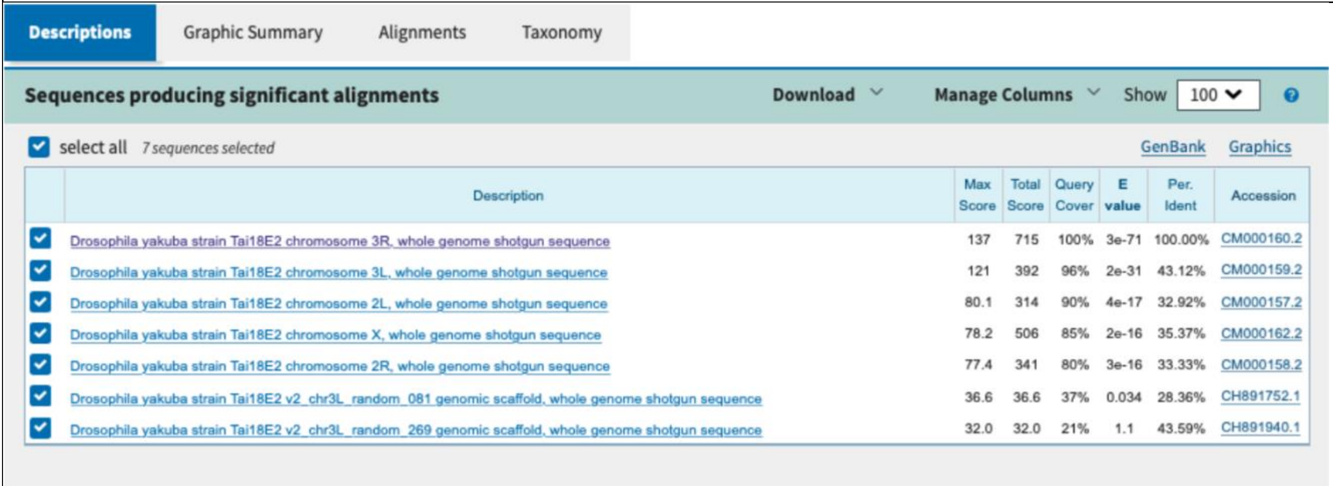
Welcome to the first part of the Annotation Report Form! In the following you will examine the genomic neighborhood of your putative ortholog within your target species and compare it to the genomic neighborhood of your target gene in *D. melanogaster* (also known as Synteny).

For a refresher on how to examine the genomic neighborhood in your target species, refer to the Pathways Project: Annotation Walkthrough Part 3: Examine genomic neighborhood of putative ortholog in target species.

For a short refresher on Synteny and how to analyze it refer to the brief About Synteny Analysis document provided by the GEP.

Paste below a screenshot of the "Descriptions" panel of your *tblastn* results of the amino acid sequence of the *D. melanogaster* protein coding isoform for your target gene against the genome assembly in your target species.

tblastn Screenshot

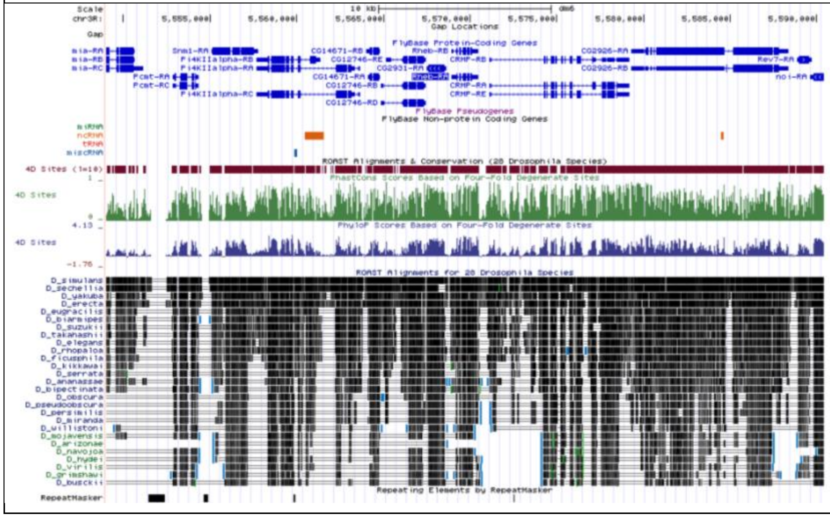


Sequences producing significant alignments

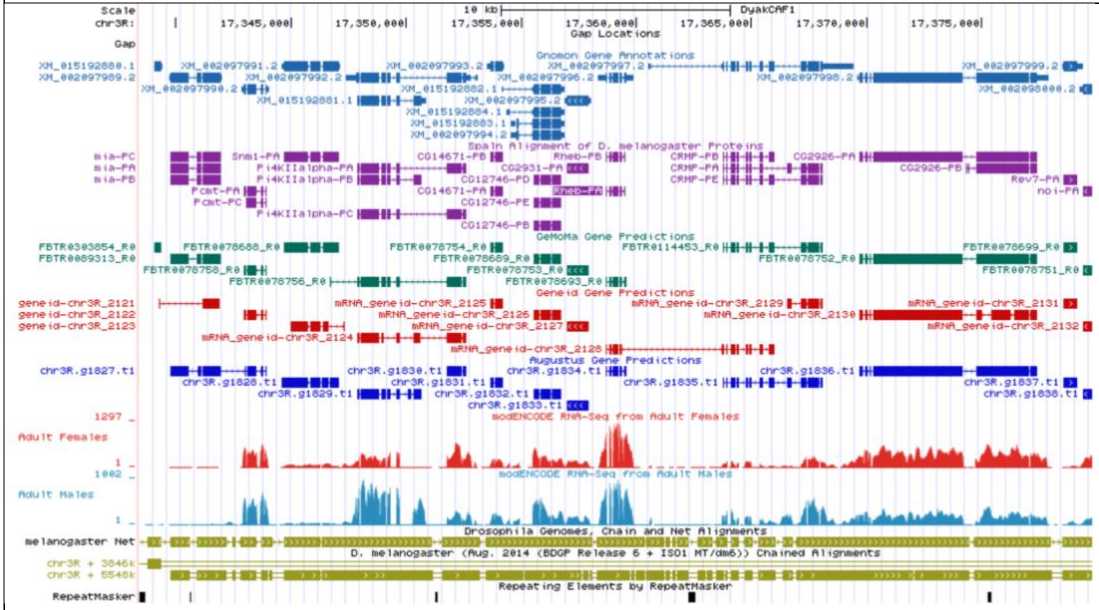
☑ select all *7 sequences selected*

	Description	Max Score	Total Score	Query Cover	E value	Per. Ident	Accession
☑	Drosophila yakuba strain Tai18E2 chromosome 3R, whole genome shotgun sequence	137	715	100%	3e-71	100.00%	CM000160.2
☑	Drosophila yakuba strain Tai18E2 chromosome 3L, whole genome shotgun sequence	121	392	96%	2e-31	43.12%	CM000159.2
☑	Drosophila yakuba strain Tai18E2 chromosome 2L, whole genome shotgun sequence	80.1	314	90%	4e-17	32.92%	CM000157.2
☑	Drosophila yakuba strain Tai18E2 chromosome X, whole genome shotgun sequence	78.2	506	85%	2e-16	35.37%	CM000162.2
☑	Drosophila yakuba strain Tai18E2 chromosome 2R, whole genome shotgun sequence	77.4	341	80%	3e-16	33.33%	CM000158.2
☑	Drosophila yakuba strain Tai18E2 v2_chr3L_random_081 genomic scaffold, whole genome shotgun sequence	36.6	36.6	37%	0.034	28.36%	CH891752.1
☑	Drosophila yakuba strain Tai18E2 v2_chr3L_random_269 genomic scaffold, whole genome shotgun sequence	32.0	32.0	21%	1.1	43.59%	CH891940.1

Paste below an image of the genomic neighborhood of your target gene in *D. melanogaster* from the GEP UCSC Genome Browser including the nearest two upstream and two downstream protein-coding genes and nested/nesting genes (if present). Select the "default tracks" for the region, and then set a comparative genomics track (e.g., Drosophila Conservation (28 Species)) to "pack".

D. melanogaster Genomic Neighborhood


Paste below an image of the genomic neighborhood of your target gene in the <u>target species</u> from the GEP UCSC Genome Browser including both of the two nearest two upstream and two downstream protein-coding genes and any nested/nesting genes (if present). Select the "default tracks" for the region, and then set a comparative genomics track (e.g., Drosophila Chain/Net) (if available) to "pack".

Target Species Genomic Neighborhood


Inspect the region around your gene in *D. melanogaster.* Record the names of the nearest two protein coding genes upstream and the nearest two downstream of your gene in *D. melanogaster* and in your target species including the *blastp* results that support your target species Gene Symbol.

- **Database select: Reference Proteins (refseq_proteins); Organism: *Drosophila melanogaster* (taxid:7227)**

IMPORTANT: The genomic neighborhood surrounding your target gene in your target species is based on the **BLAT Alignments of NCBI RefSeq Genes** predictive track displayed in the GEP UCSC Genome Browser. Furthermore, upstream and downstream is determined in relation to the directionality of your target gene, NOT in relation to the scaffold your target gene is found on. Nested genes are genes which nest a gene(s) within its intron(s) (Note: this does not apply to introns separating UTRs or UTR from CDS).

		2nd closest Upstream	Closest Upstream	Nested Gene[1]	Target Gene	Closest Downstream	2nd closest Downstream
D. melanogaster	Gene Symbol	CG12746	CG2931		Rheb	CRMP	CG2926
	DNA Strand (+/-) in *D. melanogaster*	+	-		+	+	-
Target Species	NCBI RefSeq Gene (mRNA) Accession	XM_002097994	XM_002097995		XM_002097996	XM_002097997	XM_002097998
	NCBI RefSeq Protein Accession	XP_002098030.1	XP_002098031.1		XP_002098032.1	XP_002098033.2	XP_002098034.1
	DNA Strand (+/-) in Target Species	+	-		+	+	-
Best *blastp* Result	Accession	NP_64951.4	NP_649552.1		NP_730950.2	NP_730954.2	NP_649554.1
	D. melanogaster Gene Symbol	CG12746	CG2931		Rheb	CRMP	CG2926
	E-value	3e-16	2e-3		5e-41	5e-13	2e-12
	Percent Identity	25.21%	39.51%		43.75%	24.81%	48.00%
Are the genes in the two species orthologs? (Yes/No)		Yes	Yes		Yes	Yes	Yes

[1] Leave column blank if target gene is not nested within another gene or another gene is not nested within target gene.

Explain what evidence supports your hypothesis that you have located the correct genomic neighborhood in the target species (based on your *tblastn* result) and are therefore annotating the ortholog to the *D. melanogaster* gene. Summarize the information from your table above: Be sure to describe or address any discrepancies found in the *BLAST* results or genomic neighborhood.

- Synteny: Explain whether the genes are orthologous and if the genes are on the same strand or not. If one or more genes are non-orthologous to the expected *D. melanogaster* gene(s), explain why you still think you found the ortholog of your target gene within your target species.

1. The synteny seems to suggest that this is the ortholog in DyakCAF1.
2. The next best blast hits have significantly lower e-values, query covers and percent identities.

Part B. Coding Sequence (CDS) Report Form

Number of isoforms in *D. melanogaster:* Enter number
Number of isoforms in this project: Enter number

Note: If more isoforms exist than there is space for in the table, please add more rows by going to the bottom-right-most cell and pressing tab.

Complete the following table for all the *D. melanogaster* isoforms in this project:

Name(s) of unique isoform(s) in *D. melanogaster* based on coding sequence	**List of isoforms with identical coding sequences in *D. melanogaster***	**Enter "Yes" to specify if coding isoform is likely present in target species.[2]**
Rheb-PA	PB	Yes

Is there strong evidence (e.g., extended RNA-Seq, high quality splice sites, conservation across multiple species) for distinct protein coding isoforms present in your species that are not found in *D. melanogaster*? Yes/[✓ No]

If yes, how many? Enter number

If yes, create additional isoform reports for those coding sequences and name the isoforms "-PAA," "-PAB," etc. (e.g., dyak_Rheb-PAA). Validate your novel isoform using the closest isoform present in your target species as a proxy in the Gene Model Checker.

Note: For isoforms with identical coding sequences, you only need to complete the Isoform Report Form for one of these isoforms (i.e., using the name of the isoform listed in the left column of the table above). However, you should **generate GFF, transcript, and peptide sequence files for ALL isoforms**, regardless of whether they have identical coding sequences as other isoforms. If an isoform is not present in your target species, do not complete an isoform report for that isoform.

When copying sections for additional isoforms, copy the entire section first, and then populate it. (Copying individual pages will alter the Navigation Pane.)

[2] Leave this blank if this isoform is novel (i.e., does not exist in *D. melanogaster*).

Missing CDS Isoforms

Missing Isoforms in your target species: Yes/[✓ **No**]

If you think all CDS isoforms within the D. melanogaster genome exist within your target species, please select "No". However, if you believe there are CDS isoforms within the D. melanogaster genome that do not exist within your target species, please select "Yes", and provide explanations with screenshots and text below to support your hypothesis of a missing isoform. Note: if you suspect an isoform is missing, contact your instructor or reach out to the Virtual GEP TAs.

Please do not try to create page-breaks or section breaks within your explanation.

[TYPE GENE ISOFORM NAME HERE] – Coding Isoform Report Form

Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed; when copying sections for additional isoforms, copy the entire blank section first, and then populate it.)[3]:

Gene Model Checker Checklist

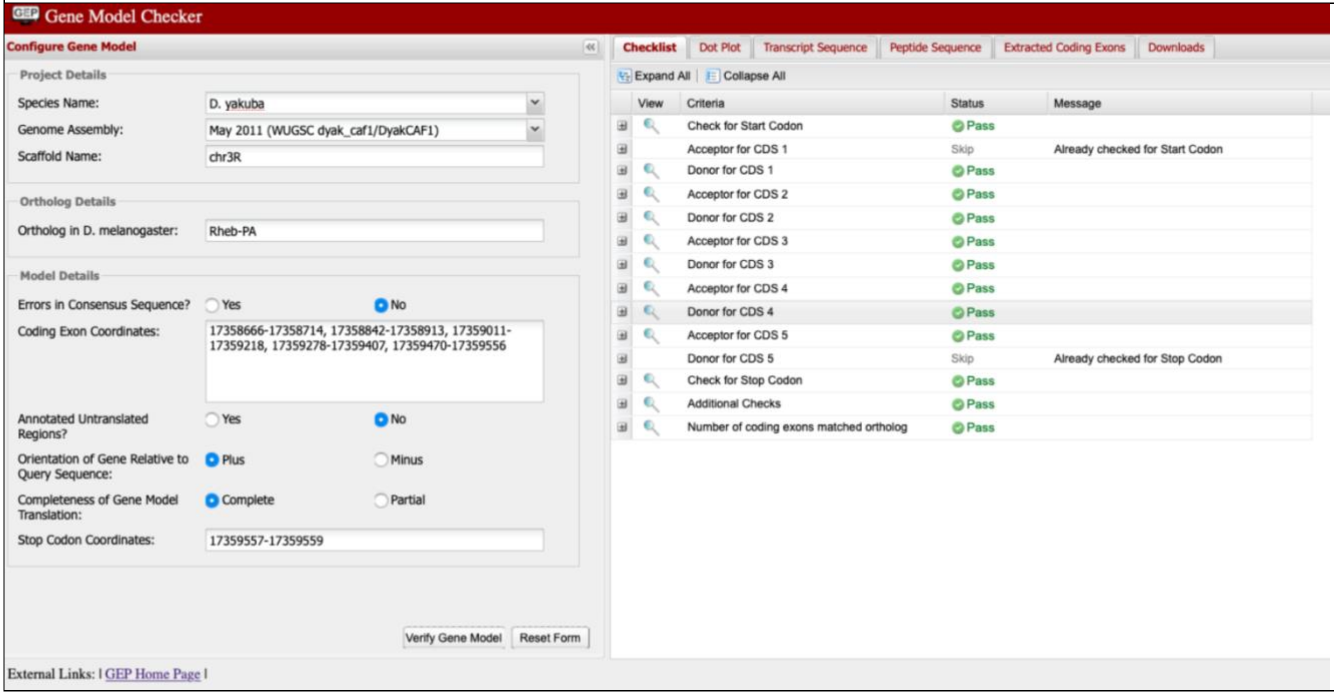
Enter the coordinates for the coding exons below (e.g., 100-200). Add more rows if needed. All rows do not need to be populated. Remember to leave out commas when typing the coordinates. If more CDSs exist, please add more rows by going to the bottom-right-most cell and pressing tab.

CDS Number	Acceptor Phase	CDS Coordinates (Start – Stop)	Donor Phase
CDS 1	0	17358666-17358714	1
CDS 2	2	17358842-17358913	1
CDS 3	2	17359011-17359218	2
CDS 4	1	17359278-17359407	0
CDS 5	0	17359470-17359556	0
CDS 6			
CDS 7			
CDS 8			
CDS 9			
CDS 10			
CDS 11			
CDS 12			
CDS 13			
CDS 14			
CDS 15			
CDS 16			
CDS 17			
CDS 18			

Stop Codon Coordinates (e.g., 201-203): 17359557-17359559

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and paste a screenshot of the checklist results and data entry section below (Note: make sure to include the "Configure Gene Model" tab (Project Details, Ortholog, Details, and Model Details) in the screenshot):

Gene Model Checker Screenshot



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, 17359011-17359218, 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

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	

External Links: | GEP Home Page |

Note: For projects with consensus sequence errors, report the exon coordinates relative to the **original project sequence**. Include the VCF file you have generated above when you submit the gene model to the Gene Model Checker. The Gene Model Checker will use this VCF file to automatically revise the submitted exon coordinates.

View the gene model on Gene Model Checker

Use the custom track feature from the Gene Model Checker to capture an image of your gene model shown on the Genome Browser for your project.

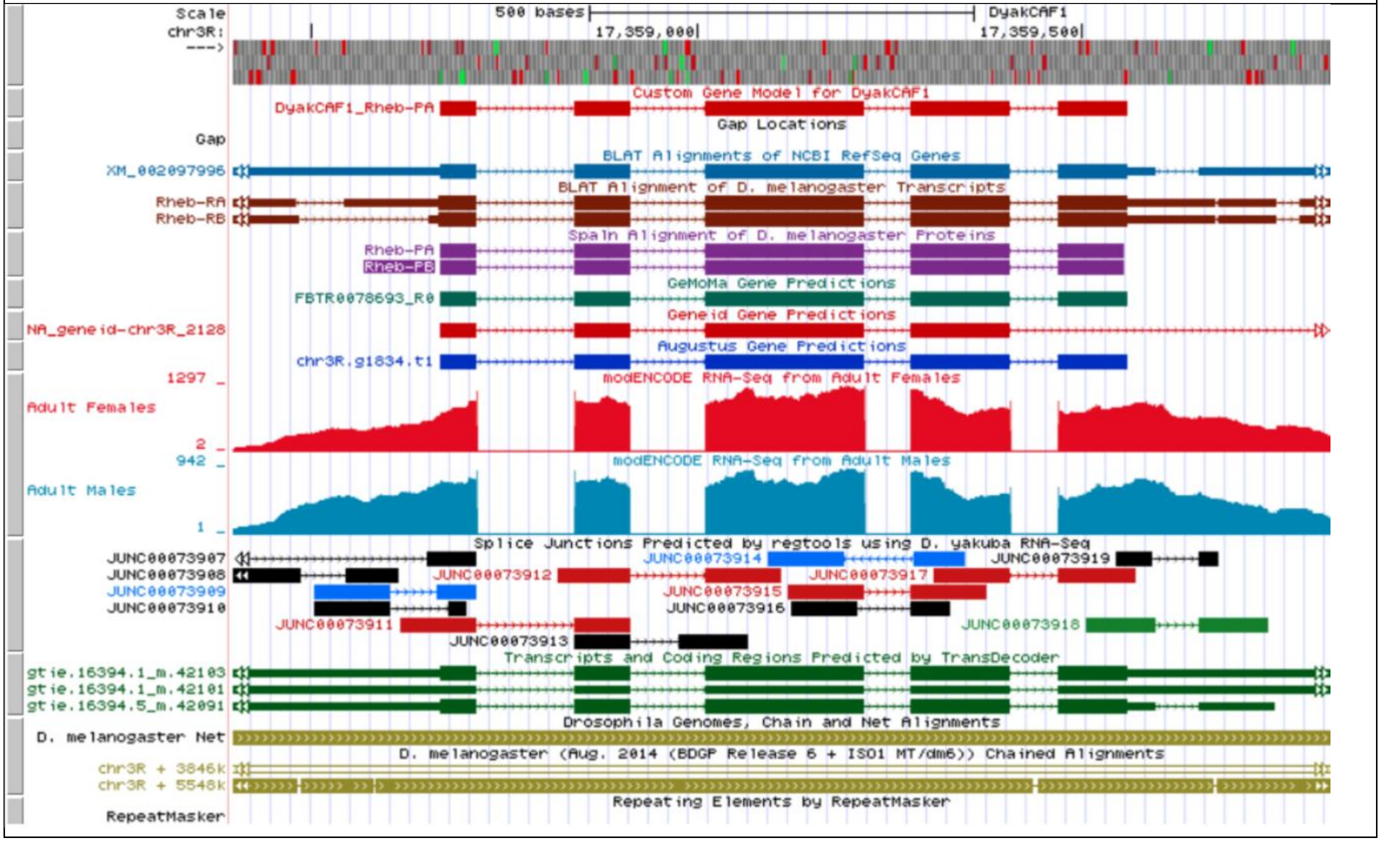
In the checklist results of the Gene Model Checker, click on the 🔍 icon next to "Number of coding exons matched ortholog" within the checklist, and a new window will open showing the Genome Browser view of this region. Your gene model will be shown under the track title "Custom Gene Model."

Select the "default tracks" for the region, and then set the following evidence tracks (if available) to "pack":

1. at least one transcript prediction track (e.g., TransDecoder Transcripts, modENCODE Cufflinks Transcripts)
2. at least one splice-site prediction track (e.g., Splice Junctions, modENCODE TopHat Junctions)
3. a comparative genomics track (e.g., Drosophila Chain/Net)

Click on "refresh" after enabling the above tracks.

Paste below an image of your gene model as shown on the Genome Browser (including the above listed tracks):

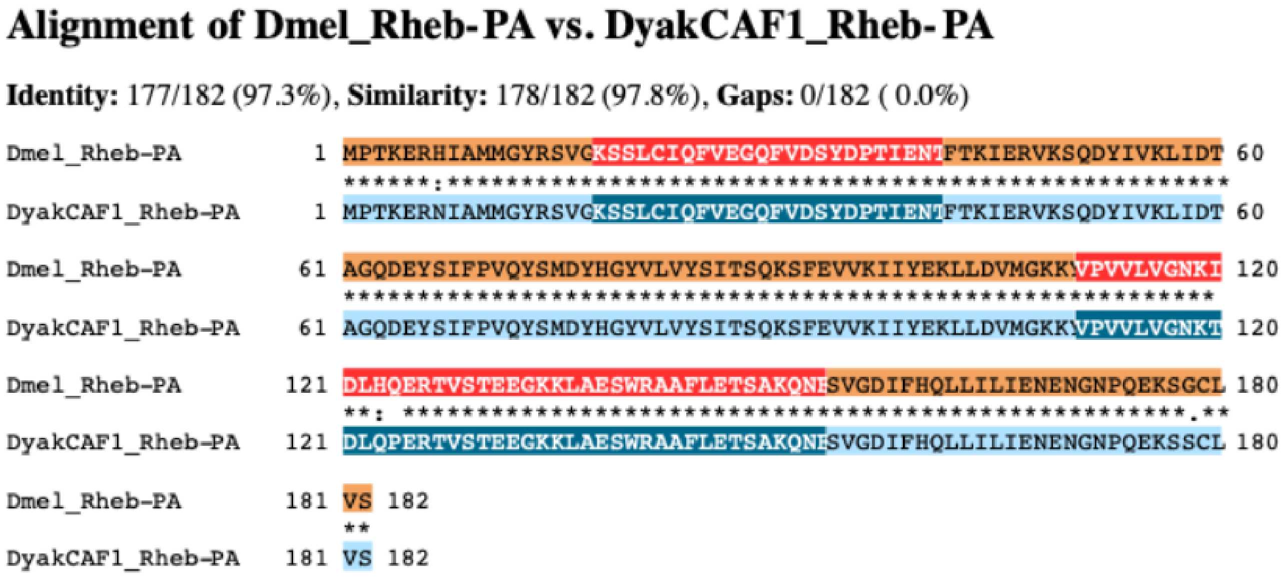
Custom Gene Model on Gene Model Checker


Scale
chr3R:
500 bases
DyakCAF1
17,359,000
17,359,500
Custom Gene Model for DyakCAF1
DyakCAF1_Rheb-PA
Gap
Gap Locations
BLAT Alignments of NCBI RefSeq Genes
XM_002097996
BLAT Alignment of D. melanogaster Transcripts
Rheb-RA
Rheb-RB
Spaln Alignment of D. melanogaster Proteins
Rheb-PA
Rheb-PB
GeMoMa Gene Predictions
FBTR0078693_R0
Geneid Gene Predictions
NA_geneid-chr3R_2128
Augustus Gene Predictions
chr3R.g1834.t1
1297
modENCODE RNA-Seq from Adult Females
Adult Females
2
942
modENCODE RNA-Seq from Adult Males
Adult Males
1
Splice Junctions Predicted by regtools using D. yakuba RNA-Seq
JUNC00073907
JUNC00073908
JUNC00073909
JUNC00073910
JUNC00073911
JUNC00073912
JUNC00073913
JUNC00073914
JUNC00073915
JUNC00073916
JUNC00073917
JUNC00073918
JUNC00073919
Transcripts and Coding Regions Predicted by TransDecoder
gtie.16394.1_m.42103
gtie.16394.1_m.42101
gtie.16394.5_m.42091
Drosophila Genomes, Chain and Net Alignments
D. melanogaster Net
D. melanogaster (Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6)) Chained Alignments
chr3R + 3846k
chr3R + 5548k
Repeating Elements by RepeatMasker
RepeatMasker

Alignment between the submitted model and the *D. melanogaster ortholog*

Paste below an image of the protein alignment generated by the Gene Model Checker (available through the "View protein alignment" link under the "Dot Plot" tab → click on "Download alignment image"):

Protein Alignment of Isoform



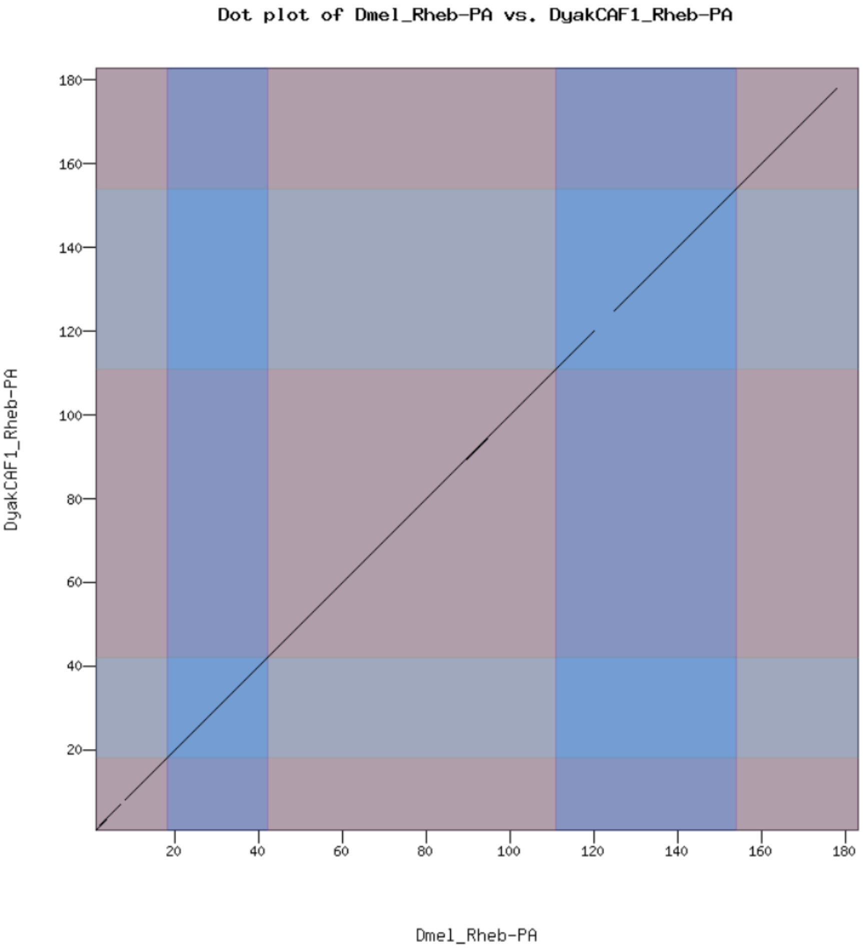
Alignment of Dmel_Rheb-PA vs. DyakCAF1_Rheb-PA

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

```
Dmel_Rheb-PA          1 MPTKERHIAMMGYRSVGKSSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDT 60
                        ******:*****************************************************
DyakCAF1_Rheb-PA      1 MPTKERNIAMMGYRSVGKSSLCIQFVEGQFVDSYDPTIENTFTKIERVKSQDYIVKLIDT 60

Dmel_Rheb-PA         61 AGQDEYSIFPVQYSMDYHGYVLVYSITSQKSFEVVKIIYEKLLDVMGKKYVPVVLVGNKI 120
                        *********************************************************** 
DyakCAF1_Rheb-PA     61 AGQDEYSIFPVQYSMDYHGYVLVYSITSQKSFEVVKIIYEKLLDVMGKKYVPVVLVGNKT 120

Dmel_Rheb-PA        121 DLHQERTVSTEEGKKLAESWRAAFLETSAKQNESVGDIFHQLLILIENENGNPQEKSGCL 180
                        **: *****************************************************.**
DyakCAF1_Rheb-PA    121 DLQPERTVSTEEGKKLAESWRAAFLETSAKQNESVGDIFHQLLILIENENGNPQEKSSCL 180

Dmel_Rheb-PA        181 VS 182
                        **
DyakCAF1_Rheb-PA    181 VS 182
```

Dot plot between the submitted model and the *D. melanogaster ortholog*

Paste below a screenshot of the dot plot of your submitted model against the putative *D. melanogaster* ortholog (generated by the Gene Model Checker):

Dot Plot of Isoform


Address any anomalies on the dot plot and protein alignment in the textbox below (Note: the dot plot is a visual/graphical representation of the protein alignment showing the position of the amino acids for the *D. melanogaster* protein on the *x*-axis and the position of the amino acids for your target species on the *y*-axis; therefore, large gaps, regions with no sequence similarity, and any other anomalies seen in the dot plot can be located within the protein alignment). **Also propose why you think they might be valid. You can include screenshots to illustrate the logic of your explanation below the text box.**

There seems to be a rather large lack of sequence similar at the start of exon 4 in the dot plot, but that only comes to three dissimilar amino acids since the protein is rather short

Note: Large <u>vertical and horizontal gaps</u> near exon boundaries in the dot plot often indicate that an incorrect splice site might have been picked. Please re-examine these regions and provide a detailed justification as to why you have selected this particular set of donor and acceptor sites.

If you need to use images to explain anomalies within your dot plot, please paste screenshots here.

Part C. Consensus Sequence Errors Report Form (ONLY FILL OUT IF CONSENSUS ERRORS EXIST)

Only complete this section if you have identified errors in the project consensus sequence, otherwise move to Part D: Preparing the Project for Submission (but do NOT delete section). If you suspect a potential Consensus Error, please contact your instructor, or reach out to the Virtual GEP TAs for guidance.

All the coordinates reported in this section should be relative to the coordinates of the original genomic sequence.

Location(s) in the project sequence with consensus errors:

locus			

1. Evidence that supports the consensus errors postulated above

Note: Evidence that could be used to support the hypothesis of errors within the consensus sequence include CDS alignment with frame shifts or in-frame stop codons, multiple RNA-Seq reads with discrepant alignments compared to the project sequence, and multiple high-quality discrepancies in the Consed assembly.

If you need to use images to explain evidence of consensus errors, please paste them below.

2. Generate a VCF file which describes the changes to the consensus sequence

If your target species is available in the Sequencer Updater tool, create a Variant Call Format (VCF) file that describes the changes to the consensus sequence you have identified above. **Paste a screenshot with the list of sequence changes below.**

VCF File image

Note: If your target species is not available in the Sequence Updater tool, see the VCF work-around instructions attached at the end of “How to annotate genes in other Drosophila species” instructions.

Please make sure to also save the VCF file as plain text because you will be submitting it along with the other files for this model.

Part D. Preparing the Project for Submission

For **each gene**, you should prepare the project GFF, transcript, and peptide sequence files for **ALL** isoforms along with this report. You can combine the individual files of one type (e.g., GFF) for all isoforms for one gene generated by the Gene Model Checker into a single file using the Annotation Files Merger. You should have a total of three files, one GFF, one transcript, and one peptide for each gene.

Name the files using **species_gene.filetype**
So, if you have a PEP file for *Rheb* in *D. yakuba*, the filename would be "**dyak_Rheb.pep**". The same goes for the other file types (FASTA and GFF).

For only projects with multiple errors in the consensus sequence, you should combine all the VCF files into a single project VCF file using the Annotation Files Merger. **Paste a screenshot (generated by the Annotation Files Merger) with all the consensus sequence errors you have identified in your project.** Please also submit the VCF file generated to your instructor.

Annotation File Merger Screenshot for VCF