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\_university.edu](mailto:red@pallet_university.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) | 2021-02-01 |

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@indigo.league](mailto:red@indigo.league) |
| Alternative email address (optional): | [red@kanto.trainers](mailto:red@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 as a whole will be acknowledged.

|  |  |
| --- | --- |
| Names 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.

|  |  |
| --- | --- |
|  | By checking this box, 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](https://support.office.com/en-us/article/use-the-navigation-pane-in-word-394787be-bca7-459b-894e-3f8511515e55?ui=en-US&rs=en-US&ad=US) 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](http://gander.wustl.edu/~wilson/genechecker/index.html). For example, use 10000 instead of 10,000.

Screenshot in 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.

A screenshot of a cell phone

Description automatically generated

Pathways Project: Annotation Videos

Refer to the [annotation videos](https://thegep.org/lessons/pathways-project-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](https://thegep.org/taschedule/); obtain the Zoom link from your instructor.

# A. Gene Report Form

**Inspect the region around your gene in *D. melanogaster.* Record the names of the nearest two protein coding genes upstream and the two downstream of your gene in *D. melanogaster* and in your target species including the *blastp* results that support your target species Gene Symbol.Note if your gene is nested within another gene, or another gene is nested within your gene, and indicate whether the genes are orthologs between the two species.**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  |  | **Most Upstream** | **Nearest Upstream** | **Nested Gene[[1]](#footnote-1)** | **Target Gene** | **Nearest Downstream** | **Most Downstream** |
| *D. melanogaster* Gene Symbol | CG12746 | CG2931 |  | Rheb | CRMP | CG2926 |
| Strand (+/-) in *D. melanogaster* | + | - |  | + | + | - |
| Target NCBI RefSeq Accession | XM\_002097994 | XM\_002097995 |  | XM\_002097996 | XM\_002097997 | XM\_002097998 |
| Strand (+/-) in  Target Species | + | - |  | + | + | - |
| **BEST HIT** | 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 | 0.0 | 0.0 |  | 6e-131 | 0.0 | 0.0 |
| Percent Identity | 84.30% | 96.72% |  | 97.25% | 99.66% | 86.18% |
| **2nd BEST HIT[[2]](#footnote-2)** | *D. melanogaster* Gene Symbol | WUN | Trv |  | Rap1 | CG6106 | Tou |
| E-value | 3E-16 | 2e-13 |  | 5e-41 | 5e-13 | 2e-12 |
| Percent Identity | 25.21% | 39.51% |  | 43.75% | 24.81% | 48.00% |
|  | Orthologs? (Yes/No) | YES | YES |  | YES | YES | YES |

**Paste below an image of the genomic neighborhood of your target gene in *D. melanogaster* from the GEP UCSC Genome Browser including both of the nearest two upstream and two downstream genes.** 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 target species from the GEP UCSC Genome Browser including both of the nearest two upstream and two downstream genes.** 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 |
| A screenshot of a cell phone  Description automatically generated |

**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 |
|  |

**Explain what evidence supports your hypothesis that you have located the correct genomic neighborhood in the target species, and are therefore annotating the ortholog to the *D. melanogaster* gene.** Be sure to also explain any discrepancies found in the *BLAST* results or genomic neighborhood described above (e.g., regions of the protein are missing, synteny, number of exons).

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.

# 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 the [Coding Sequence (CDS) Report Form](#_B._Coding_Sequence) section (but do NOT delete this section).

**All of 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 screenshots 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](http://gander.wustl.edu/~wilson/sequence_updater/index.html), 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.

# B. Coding Sequence (CDS) Report Form

Number of isoforms in *D. melanogaster:* 2

Number of isoforms in this project: 2

**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 of 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*** | **Select if coding isoform is likely present in target species.[[3]](#footnote-3)** |
| Rheb-PA | PB |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

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).

**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.)

## 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.

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

## Rheb-PA – 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.***)***[[4]](#footnote-4)***:

### 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](http://gander.wustl.edu/~wilson/genechecker/index.html) and **paste a screenshot of the checklist results and data entry section below:**

|  |
| --- |
| Gene Model Checker Screenshot |
|  |

**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 page12image1817136704 icon next to “Number of coding exons matched ortholog,” 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)

**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 |
|  |

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

**Paste below a screenshot of the protein alignment generated by the Gene Model Checker (available through the “View protein alignment” link under the “Dot Plot” tab):**

|  |
| --- |
| Protein Alignment of Isoform |
|  |

### 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 |
|  |

**Provide an explanation for any anomalies on the dot plot (e.g., large gaps, regions with no sequence similarity) in the textbox below. 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 vertical and horizontal gaps** 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.

# C. Transcription Start Site (TSS) Report Form

**Note:** Complete this section if you have annotated the Coding Isoforms for the gene above. This section is **OPTIONAL** based on your instructor’s expectations. If more isoforms exist, please add more rows by going to the bottom-right-most cell and pressing tab.

|  |  |
| --- | --- |
| **Name(s) of transcript isoform(s) with unique TSS in *D. melanogaster*** | **List of transcript isoforms with identical TSS in *D. melanogaster*** |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

## Missing Transcription Start Sites

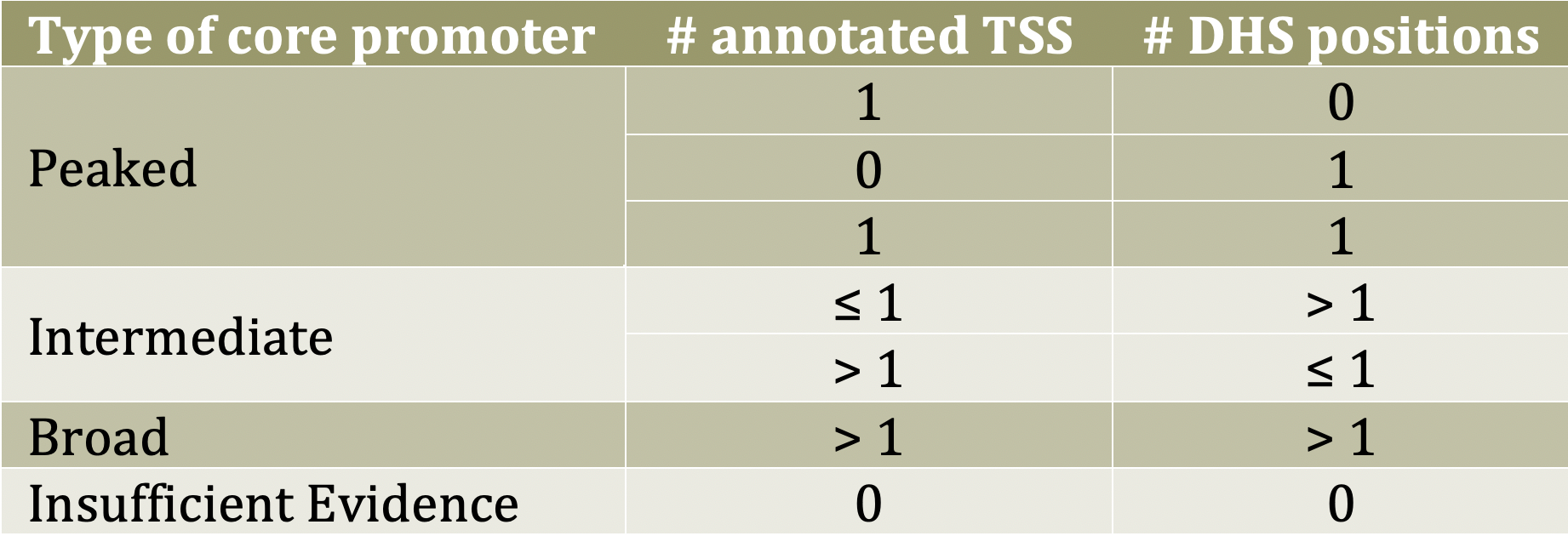
Missing TSS in your target species:  Yes  No

If you think all TSSs within the D. melanogaster genome exist within your target species, please select “No”. However, if you believe there are TSSs within the D. melanogaster genome that do not exist within your target species, please select “Yes” and provide explanations with screenshots and text below.

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

## **[TYPE TRANSCRIPT ISOFORM NAME HERE] – TSS Annotation** *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 section first, and then populate it.***):



Type of core promoter in *D. melanogaster*:

Peaked

Intermediate

Broad

Insufficient Evidence

Coordinate(s) of the TSS position(s) or likely region(s) for the TSS (enter “NA” if data not applicable). Also, enter the reason you think these are the best coordinates in the bottom cell of the table:

|  |  |
| --- | --- |
| Based on blastn alignment | Proposed coordinate(s) |
| Based on core promoter motifs (e.g., Inr) | Proposed coordinate(s) |
| Based on RNA-seq, TopHat, Cufflinks tracks | Proposed coordinate(s) |
| Based on Conservation tracks | Proposed coordinate(s) |
| Based on other evidence (please specify) | Proposed coordinate(s) |
| Reason for proposed coordinate(s). | |

**Note:** If the *blastn* alignment for the initial transcribed exon is a partial alignment, you can **extrapolate the TSS position** based on the number of nucleotides that are missing from the beginning of the exon. (Enter “Insufficient evidence” if you cannot determine the TSS position based on the available evidence.)

What do you believe is/are the best coordinate(s) or coordinate range for the TSS position(s) for this isoform given all the evidence?

Best proposed coordinate(s)

**Provide an explanation if the best TSS coordinate(s) or coordinate range is inconsistent with at least one of the evidence types specified above. You can add screenshots to further explain your logic below the textbox:**

Perform a *blastn* alignment of the initial transcribed exon in *D. melanogaster* against the genomic sequence of your target species. Remember to increase sensitivity by changing the “Word size” parameter to **7**, the “Match/Mismatch Scores” to “**1, -1**”, the “Gap Costs” to “**Existence: 2 Extension: 1**”, and **uncheck the filter for “Low complexity regions.”** **Paste a screenshot of the *blastn* alignment below:**

|  |
| --- |
| *blastn* Alignment |
|  |

**Paste an image of the Genome Browser region surrounding the putative TSS (±300 bp) with the following evidence tracks:**

1. RNA-Seq Alignment Summary
2. Splice Junctions
3. Short Match results for the Inr motif (TCAKTY)
4. A comparative genomics track (e.g., Drosophila Chain/Net)

|  |
| --- |
| Region around TSS in Genome Browser |
|  |

### *Search for core promoter motifs*

The consensus sequences for the *Drosophila* core promoter motifs are available at <http://gander.wustl.edu/~wilson/core_promoter_motifs.html>

Use the "Short Match" functionality in the GEP UCSC Genome Browser to search for each of the core promoter motifs listed below **in the region surrounding the TSS (±300 bp)in your project and in the *D. melanogaster* ortholog**. For TSS annotations where you can only define a TSS search region, you should report all motif instances ±300 bp of either edge of your search region.

#### Coordinates of the motif search region

Your project (e.g., scaffold\_1234:1000-1600): Start - Stop

Orthologous region in *D. melanogaster*: Start - Stop

Record the **orientation and the start coordinate** (e.g., +10000) of each motif match below. (Enter "**NA**" if there are no motif instances within the search region.)

**Note:** Highlight (in yellow) the motif instances that support the TSS annotation(s) above.

|  |  |  |
| --- | --- | --- |
| Core promoter motif | Your project | *D. melanogaster* |
| BREu | Start coordinate | Start coordinate |
| TATA Box |  |  |
| BREd |  |  |
| Inr |  |  |
| MTE |  |  |
| DPE |  |  |
| Ohler\_motif1 |  |  |
| DRE |  |  |
| Ohler\_motif5 |  |  |
| Ohler\_motif6 |  |  |
| Ohler\_motif7 |  |  |
| Ohler\_motif8 |  |  |

# D. Transcript Report Form

**Note:** Complete this section if you have annotated the transcript for the gene above. This section is **OPTIONAL** based on your instructor’s expectations. If more isoforms exist, please add more rows by going to the bottom-right-most cell and pressing tab.

|  |  |
| --- | --- |
| **Name(s) of transcript isoform(s) in *D. melanogaster*** | **Do you think this isoform is present in your target species (select if Yes)?** |
| Transcript isoform |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Is there strong evidence for distinct transcript isoforms present in your species that are not found in *D. melanogaster* (yes/no)?  Yes  No

If yes, how many? *Enter number*

\*Note that if you have a novel coding region you should also have at least one novel transcript.

If yes, create additional transcript isoform reports for those transcript sequences and name the isoforms “-RAA,” “-RAB,” etc. (e.g. dyak\_Rheb-RAA).

\*If these new transcripts include novel CDS’s, then make sure to match the isoform naming used for the coding region above.

## Missing Transcript Isoforms

Missing Isoforms in your target species:  Yes  No

If you think all transcript isoforms within the D. melanogaster genome exist within your target species, please select “No”. However, if you believe there are transcript 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.

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

## [TYPE TRANSCRIPT NAME HERE] - Transcript 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 section first, and then populate it.***):

Enter the coordinates for the **transcribed exons** below(e.g., 100-200). Add more rows if needed. Also mark whether the exon contains UTRs (UnTranslated Regions). Make sure to enter the coordinates without commas.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Exon #** | **Acceptor Phase** | **Coordinates (Start – Stop)** | **Donor Phase** | **Check box if contains UTR** |
| Exon 1 | 0 |  |  | ☐ |
| Exon 2 |  |  |  | ☐ |
| Exon 3 |  |  |  | ☐ |
| Exon 4 |  |  |  | ☐ |
| Exon 5 |  |  |  | ☐ |
| Exon 6 |  |  |  | ☐ |
| Exon 7 |  |  |  | ☐ |
| Exon 8 |  |  |  | ☐ |
| Exon 9 |  |  |  | ☐ |
| Exon 10 |  |  |  | ☐ |
| Exon 11 |  |  |  | ☐ |
| Exon 12 |  |  |  | ☐ |
| Exon 13 |  |  |  | ☐ |
| Exon 14 |  |  |  | ☐ |
| Exon 15 |  |  |  | ☐ |
| Exon 16 |  |  |  | ☐ |
| Exon 17 |  |  |  | ☐ |
| Exon 18 |  |  |  | ☐ |

Use the [Gene Model Checker](http://gander.wustl.edu/~wilson/genechecker/index.html) and **make sure to click “yes” for “Annotated Untranslated Regions?”.**

Enter the coordinates of your final gene model for this isoform into the [Gene Model Checker](http://gander.wustl.edu/~wilson/genechecker/index.html) and **paste an image of the checklist results and data entry section below:**

|  |
| --- |
| Isoform Model with UTRs on Gene Model Checker |
|  |

**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 the 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 page12image1817136704icon next to “Number of coding exons matched ortholog,” 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)

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

|  |
| --- |
| Isoform Model with UTRs on Gene Model Checker |
|  |

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

Show an alignment between the nucleotide sequence for your gene model and the nucleotide sequence from the putative *D. melanogaster* ortholog. You can generate a new alignment using the “Align two or more sequences” feature (*bl2seq*) at the NCBI BLAST web site. Remember to increase sensitivity by changing the “Word size” parameter to **7**, the “Match/Mismatch Scores” to “**1, -1**”, the “Gap Costs” to “**Existence: 2 Extension: 1**”, and **uncheck the filter for “Low complexity regions”.**

If no alignment results, record “NA”. **Paste a screenshot of the nucleotide alignment below:**

|  |
| --- |
| Nucleotide Alignment of isoform |
|  |

# E. 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](http://gander.wustl.edu/~wilson/submissionhelper/index.php). 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](http://gander.wustl.edu/~wilson/submissionhelper/index.php). **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 faculty.

|  |
| --- |
| Annotation File Merger Screenshot for VCF |
|  |

1. Leave column blank if target gene is not nested within another gene or another gene is not nested within target gene [↑](#footnote-ref-1)
2. Not an isoform of best hit [↑](#footnote-ref-2)
3. Leave this blank if this isoform is novel (i.e., does not exist in *D. melanogaster*). [↑](#footnote-ref-3)
4. Note: If you are not doing the TSS and full transcript annotations, skip to section [E. Preparing Project for Submission](#_E._Preparing_the). [↑](#footnote-ref-4)