GEP Annotation Report: Drosophila Pathway Genes

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

Student name: Savanah Fondse

Student email: sfondse@csustan.edu

Faculty instructor: James Youngblom

College/university: CSU Stanislaus

# Project details

Assigned Gene Name in *D. melanogaster (e.g. eyeless)*: \_\_\_\_\_\_\_\_\_\_*Tsc1*\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Project species: D. yakuba

Ortholog ID in your species if known (GLEAN-R ID): \_\_\_\_\_\_\_\_\_\_\_Dyak\GE23448\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Gene name (*e.g.*, *D. biarmipes* *eyeless*): *D. yakuba Tsc1*

Gene symbol (*e.g.*, *dbia\_ey*): *dyak\_Tsc1*

Scaffold Name: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_chr3R\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Genome Assembly (use the most recent assembly available in Gander):

\_\_May 2011 (WUGSC dyak\_caf1/DyakCAF1)\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

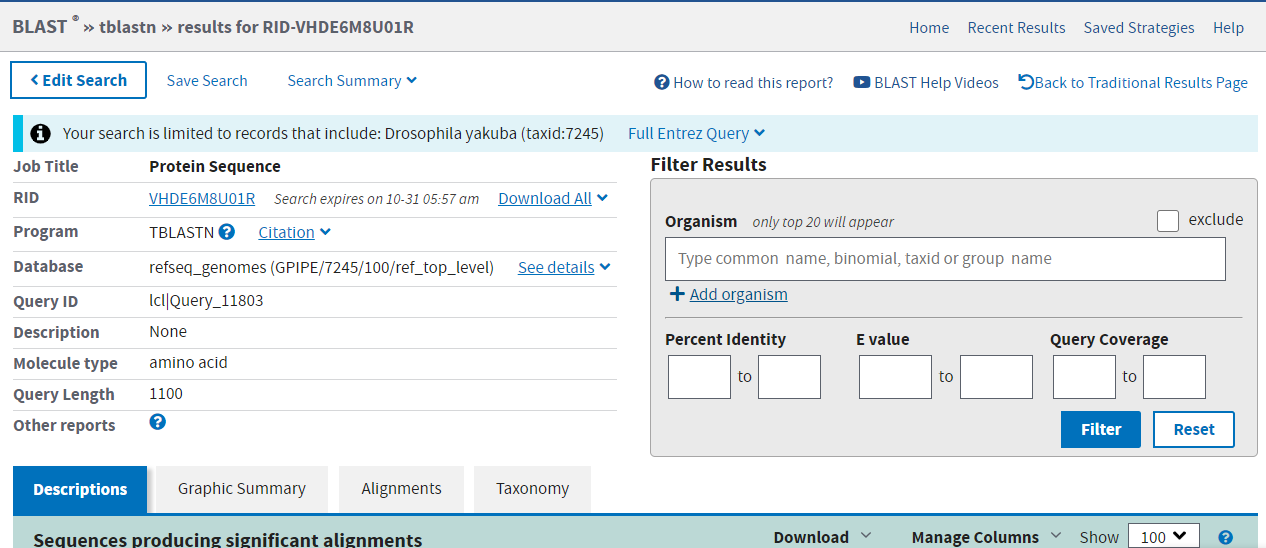
Date of submission: 10/29/19

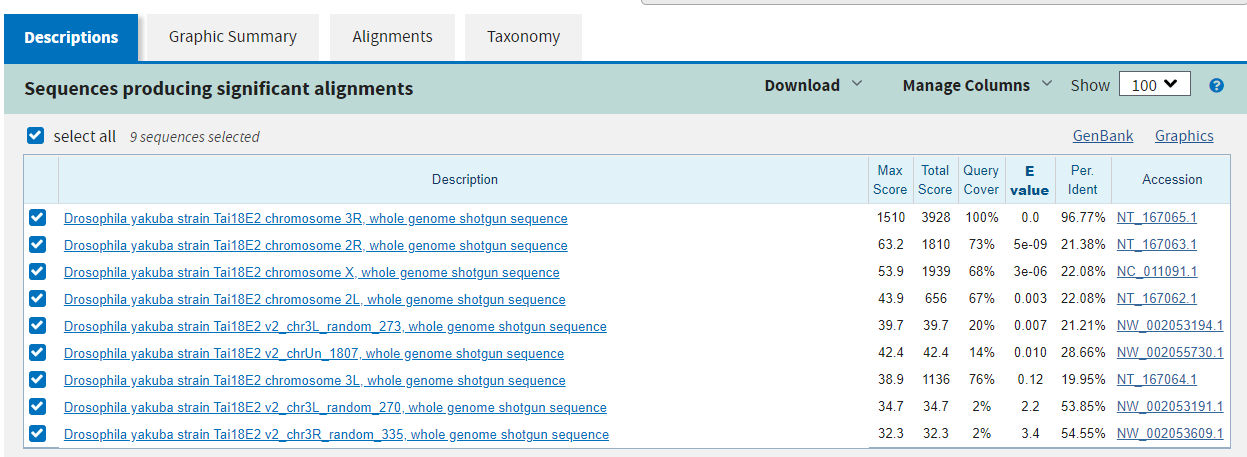
# Instructions for insulin pathway project

**Please see the “How to annotate genes in other Drosophila species” file.**

Using tblastn, BLAST the amino acid sequence of the longest *D. melanogaster* protein coding isoform for your gene against your target species in the RefSeq Genome Database.

**Paste a screen shot of the “Descriptions” panel below:**



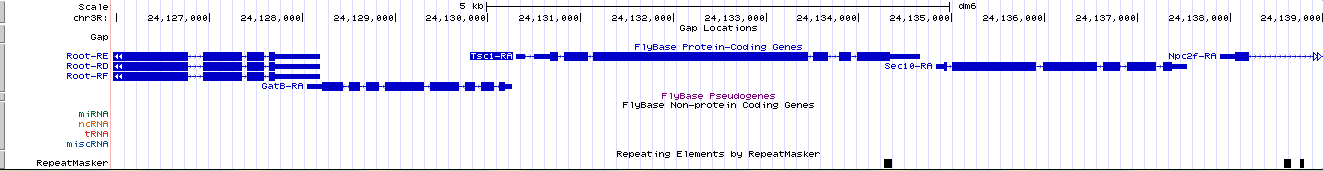


Inspect the region around your gene in *D. melanogaster.* **Record the names of the two protein coding genes upstream and the two downstream of your gene in *D. melanogaster* and in your target species.Also note if your gene is nested within another gene, or another gene is nested within your gene. Indicate whether the genes are orthologs between the two species in the table below** (*i.e.* Does the tblastn for the protein coding sequence of each neighboring gene *D. melanogaster* have a best hit to the neighboring protein coding gene in your species?)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Most upstream | Closest upstream | Nested (if appropriate) | Closest downstream | Most downstream |
| D. mel Gene ID | Root-Rf | GatB-RA |  | Sec10-RA | Npc2f-RA |
| Target Species GLEAN-R ID | Root-PE  Dyak\GE23447 | GatB-PA  Dyak\GE10827 |  | Sec10-PA  Dyak\GE23449 | Npc2f-PA  Dyak\GE23450 |
| Orthologs? (yes, no) | yes | yes |  | yes | yes |

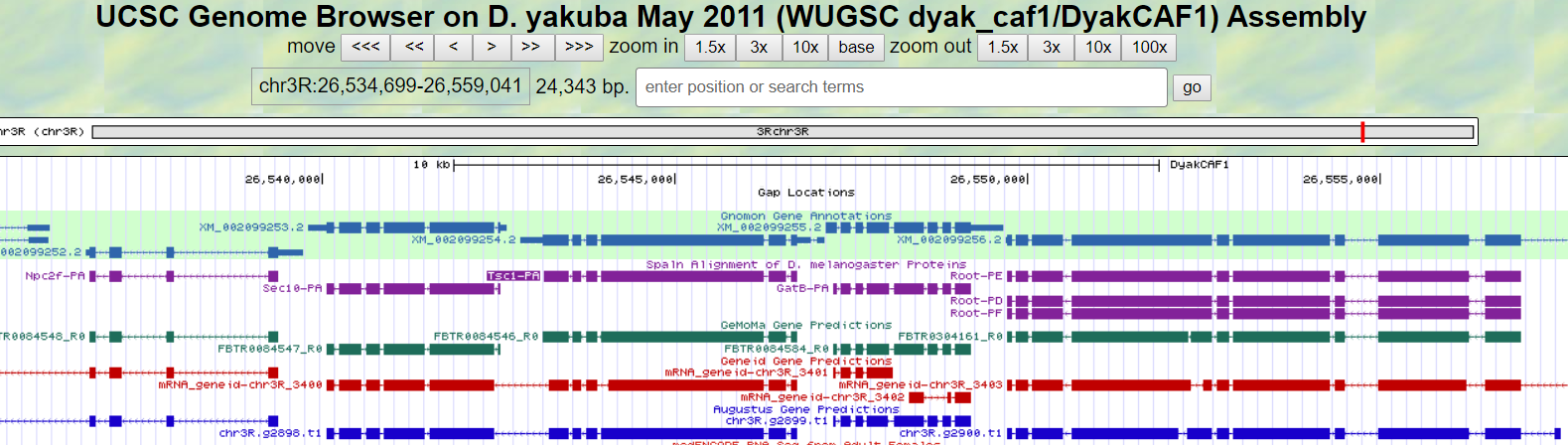
**Paste below a screen shot of the genomic neighborhood for *D. melanogaster* from Gander including both of the two upstream and two downstream genes.**

***D. melanogaster:***



**Paste below a screen shot of the genomic neighborhood for target species from Gander including both of the two upstream and two downstream genes**

***D. yakuba:***



**Explain why you believe you are annotating the ortholog to the *D. melanogaster* gene. Explain any discrepancies found in the BLAST results or genomic neighborhood described above.**

I believe I annotated the ortholog to the *D. melanogaster* gene because it has the same neighboring genes and very little difference in the alignment of the protein sequences between the two species of *Drosophila*.

## Consensus sequence errors report form

Complete this section if you have identified errors in the project consensus sequence.

**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:

### 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.

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

If your target species is available in the Sequencer Updater tool (available through the GEP web site under “Projects”  “Annotation Resources”), 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:**

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.

Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms.

# Coding Sequence Gene report form

Number of isoforms in *D. melanogaster:* 1

Number of isoforms in this project: 1

**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*** | **Isoform coding sequence likely present in target species (yes/no)?** |
| Tsc1-PA | None | yes |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

\* Complete only one isoform report for multiple isoforms that are identical. If an isoform is not present in your target species do not complete an isoform report for that isoform.

Is there strong evidence for distinct protein coding isoforms present in your species but not found in *D. melanogaster* (yes/no)?\_\_\_\_\_\_No\_\_\_\_ How many?\_\_\_\_\_\_0\_\_\_\_\_

If yes, create additional isoform reports for those coding sequences and name the isoforms -PAA, -PAB *etc*. (*e.g.* dbia\_eye-PAA).

**Note:** For isoforms with identical coding sequence, 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**, irrespective of whether they have identical coding sequences as other isoforms.

## 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):

Gene-isoform name (*e.g.*, dbia\_ey-PA): D. yakuba\_Tsc1-PA

Names of the isoforms with identical coding sequences as this isoform:

None

Is the 5’ end of this isoform missing from the end of the project? No

If so, how many exons are missing from the 5’ end: 0

Is the 3’ end of this isoform missing from the end of the project? 0

If so, how many exons are missing from the 3’ end: 0

### Gene Model Checker checklist

Coding Exon Coordinates (e.g. 100-200). Exon numbering as in the gene record finder for the polypeptide. Add more columns if needed.

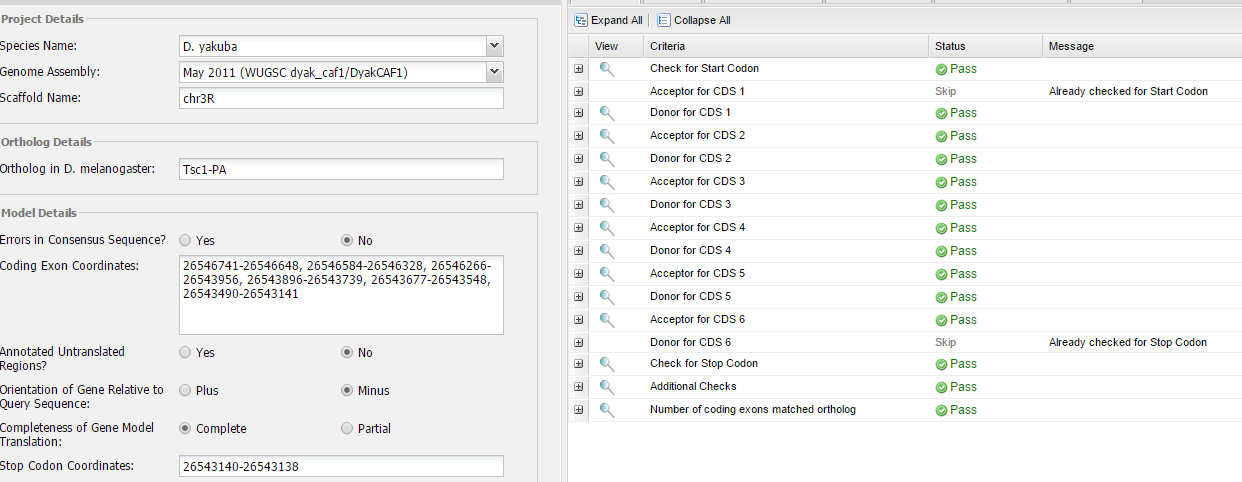
|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exon 1 | Exon 2 | Exon 3 | Exon 4 | Exon 5 | Exon 6 | Exon 7 | Exon 8 | Exon 9 |
| 26546741-26546648 | 26546584-26546328 | 26546266-26543956 | 26543896-26543739 | 26543677-26543548 | 26543490-26543141 |  |  |  |

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

26543140-26543138

• Use the Gene Model Checker for whole genome assemblies that is found at <http://gander.wustl.edu/~wilson/genechecker-ucsc/index.html>

• Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and **paste a screenshot of the checklist results below**



**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 Genome Browser

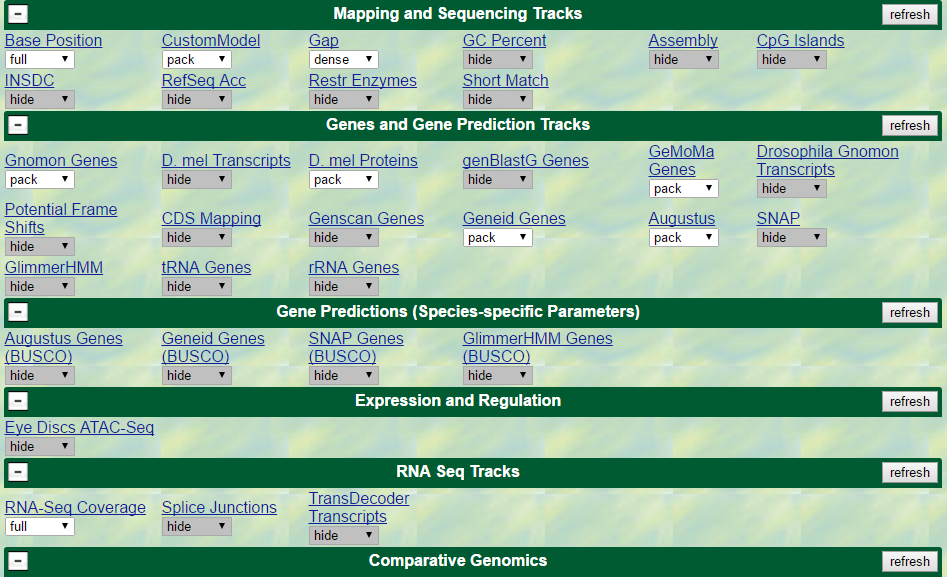
Use the custom track feature from the Gene Model Checker to capture a screenshot of your gene model shown on the Genome Browser for your project. Zoom in so that only this isoform is in the screenshot. (See page 12 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help”  “Documentations”  “Web Framework” on the GEP website at <http://gep.wustl.edu>.)

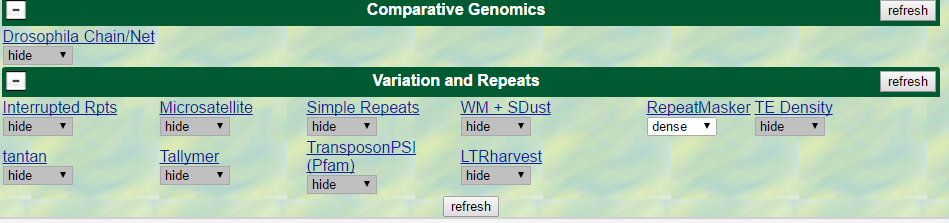
Include the following evidence tracks in the screenshot if they are available.

1. A sequence alignment track (D. mel Proteins or Other RefSeq)
2. At least one gene prediction track (*e.g.*, GLEAN-R Genes)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Alignment Summary)
4. A comparative genomics track (*e.g.*, Conservation, D. mel. Net Alignment)
5. At least one transcript prediction track (*e.g.* modENCODE Cufflinks Transcripts)
6. At least one splice-site prediction track (*e.g.* modENCODE TopHat Junctions)

**Paste a screenshot of your gene model as shown on the Genome Browser below:**

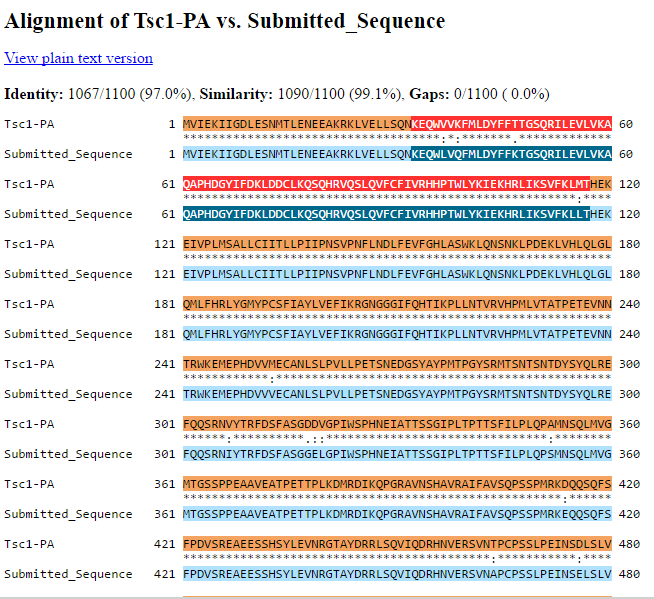
### 

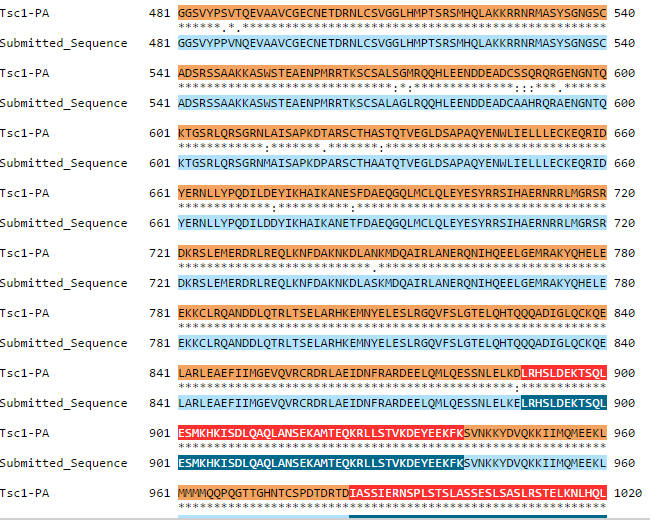


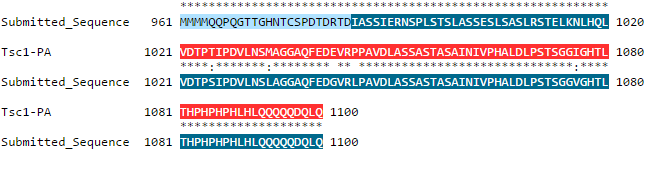


### Alignment between the submitted model and the *D. melanogaster ortholog*

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can either use the protein alignment generated by the Gene Model Checker (available through the “**View protein alignment**” link under the “Dot Plot” tab) or you can generate a new alignment using the “Align two or more sequences” feature (*bl2seq*) at the NCBI BLAST web site. **Paste a screenshot of the protein alignment below:**

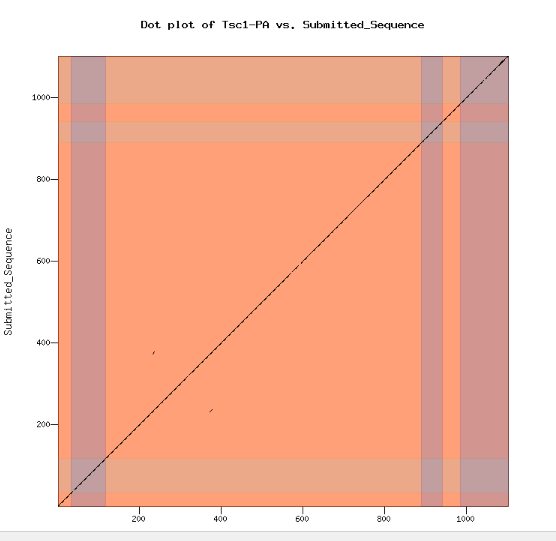






### Dot plot between the submitted model and the *D. melanogaster ortholog*

**Paste a screenshot of the dot plot** of your submitted model against the putative *D. melanogaster* ortholog (generated by the Gene Model Checker). **Provide an explanation for any anomalies** on the dot plot (*e.g.*, large gaps, regions with no sequence similarity).



**Note: If you are not doing the TSS and full transcript annotations skip to step D.**

**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 detail justification as to why you have selected this particular set of donor and acceptor sites.

## Transcription start sites (TSS) report form (optional for some classes)

**Note:** Complete this section if you have annotated the TSS for the gene above. This section is **OPTIONAL** based on your instructor’s expectations.

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

Complete this report form for each unique TSS listed in the table above. Copy and paste this form to create as many copies as needed within this report.

Gene-isoform name (*e.g.*, *dbia\_ey-RA*):

Names of the isoforms with the same TSS as this isoform:

Type of core promoter in *D. melanogaster*

(Peaked / Intermediate / Broad / Insufficient Evidence):

The type of core promoter is defined by the number of annotated TSS and DHS positions:

|  |  |  |
| --- | --- | --- |
| Type of core promoter | # annotated TSS | # DHS positions |
| Peaked | 1 | 0 |
| 0 | 1 |
| 1 | 1 |
| Intermediate | ≤ 1 | > 1 |
| > 1 | ≤ 1 |
| Broad | > 1 | > 1 |
| Insufficient Evidence | 0 | 0 |

Coordinate(s) of the TSS position(s) or likely region(s) for the TSS (enter NA if data not applicable):

Based on blastn alignment:

Based on core promoter motifs (*e.g.*, Inr):

Based on RNA-seq, TopHat, Cufflinks tracks: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Based on Conservation tracks: \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Based on other evidence (please specify):

**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?

Provide an explanation if the best TSS coordinate(s) or coordinate range is inconsistent with at least one of the evidence types specified above:

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:**

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

1. RNA-Seq Alignment Summary
2. RNA-Seq TopHat
3. RNA-Seq Cufflinks
4. Short Match results for the Inr motif (TCAKTY)
5. A comparative genomics track (*e.g.*, Conservation, D. mel. Net Alignment)

### *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 (±300bp)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 ±300bp of either edge of your search region.

#### Coordinates of the motif search region

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

Orthologous region in *D. melanogaster*:

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 |  |  |
| TATA Box |  |  |
| BREd |  |  |
| Inr |  |  |
| MTE |  |  |
| DPE |  |  |
| Ohler\_motif1 |  |  |
| DRE |  |  |
| Ohler\_motif5 |  |  |
| Ohler\_motif6 |  |  |
| Ohler\_motif7 |  |  |
| Ohler\_motif8 |  |  |

## Transcript report form (optional for some classes)

|  |  |
| --- | --- |
| **Name(s) of transcript isoform(s) in *D. melanogaster*** | **Do you think this isoform is present in your target species (yes/no)?** |
|  |  |
|  |  |
|  |  |
|  |  |

Is there strong evidence for distinct transcript isoforms present in your species but not found in *D. melanogaster* (yes/no)?\_\_\_\_\_\_\_\_\_\_ How many?\_\_\_\_\_\_\_\_\_\_\_

\*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 *et.* (*e.g.* dbia\_eye-RAA).

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

## 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 Transcript Isoform Report Form as needed):

Transcript-isoform name (*e.g.*, *dbia\_ey-RA*):

**Coding Exon Coordinates** (e.g. 100-200). Exon numbering as in the gene record finder for the polypeptide. Add more columns if needed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exon 1 | Exon 2 | Exon 3 | Exon 4 | Exon 5 | Exon 6 | Exon 7 | Exon 8 | Exon 9 |
|  |  |  |  |  |  |  |  |  |

**Transcribed Exon Coordinates** (e.g. 100-200). Exon numbering as in the gene record finder for the transcript. Add more columns if needed.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Exon 1 | Exon 2 | Exon 3 | Exon 4 | Exon 5 | Exon 6 | Exon 7 | Exon 8 | Exon 9 |
|  |  |  |  |  |  |  |  |  |

• Use the Gene Model Checker for whole genome assemblies that is found at <http://gander.wustl.edu/~wilson/genechecker-ucsc/index.html>

**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 and **paste a screenshot of the checklist results below:**

**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.

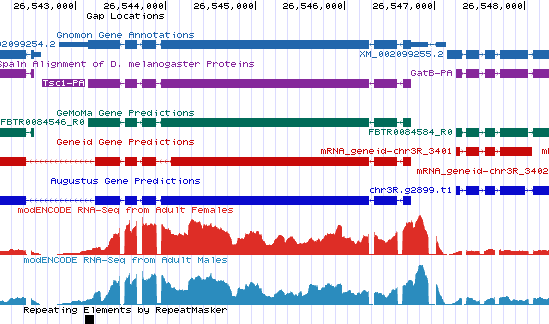
### *View the gene model on the Genome Browser*

Use the custom track feature from the Gene Model Checker to capture a screenshot of your gene model shown on the Genome Browser for your project. Zoom in so that only this isoform is in the screenshot. (See page 12 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help”  “Documentations”  “Web Framework” on the GEP website at <http://gep.wustl.edu>.)

Include the following evidence tracks in the screenshot if they are available.

1. A sequence alignment track (D. mel Proteins/Transcripts or Other RefSeq)
2. At least one gene prediction track (*e.g.*, Genscan)
3. At least one RNA-Seq track (*e.g.*, RNA-Seq Alignment Summary)
4. A comparative genomics track (*e.g.*, Conservation, D. mel. Net Alignment)
5. At least one transcript prediction track (*e.g.* modENCODE Cufflinks Transcripts)
6. At least one splice-site prediction track (*e.g.* modENCODE TopHat Junctions)

**Paste a screenshot of your gene model as shown on the Genome Browser below:**



### *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:**

# 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.

This is optional and will be looked at on a case-by-case basis:

For 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 (see the Annotation Files Merger User Guide for details). **Paste a screenshot (generated by the Annotation Files Merger) with all the consensus sequence errors you have identified in your project.**