Transcription Start Sites Project Report

Student name: _____
Student email: _____
Faculty advisor: _____
College/university: _____

Project details

Project name: _____
Project species: _____
Date of submission: _____
Number of genes in project: _____

Does this report cover TSS annotations for all of the genes or is it a partial report? _____
If this is a partial report, please indicate the region of the project covered by this report:
From base _____ to base _____

Note: In some cases, the reconciled gene models (available under "Genes and Gene Prediction Tracks" ➔ "Reconciled Gene Models" on the GEP UCSC Genome Browser) might be incorrect because of misannotations or because of updates to the *D. melanogaster* gene models from FlyBase. This could result in situations where you will need to construct a new gene model for the coding region prior to performing the TSS annotation. If you find one or more genes with this problem, you should fully document the new gene annotation(s) by completing the "**Revised gene models report form**" found on page 8 of this report.

Transcription start sites (TSS) report form

Complete this report form for each gene in your project. Copy and paste this form to create as many copies as needed.

Gene name (*e.g.*, *D. biarmipes eyeless*): _____
Gene symbol (*e.g.*, *dbia_ey*): _____

Name(s) of isoform(s) with unique TSS	List of isoforms with identical TSS

Names of the isoforms with unique TSS in *D. melanogaster* that are absent in this species:

_____

Isoform TSS report

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 TSS annotated by the Celniker group at modENCODE and the number of 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

Coordinates of the first transcribed exon based on blastn alignment:

_____

Coordinate(s) of the TSS position(s):

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

Coordinate(s) of the TSS search region(s):

_____

Note: If part of the TSS search region is only weakly supported by the available evidence, then please specify both a **wide** and a **narrow** search region. For example, if the region at 1500-2000 shows high RNA-Seq read coverage but there is very low RNA-Seq coverage from 1000-1499, then you will report "**1000-2000 (wide)**" as the wide search region and "**1500-2000 (narrow)**" as the narrow search region.

Describe the evidence used to define the TSS search region(s) (*e.g.*, RNA-Seq and Conservation tracks in this species, RAMPAGE data from *D. melanogaster*):

_____

1. Evidence that supports the TSS annotation postulated above

Were you able to define the TSS position(s) based on the blastn alignment? _____

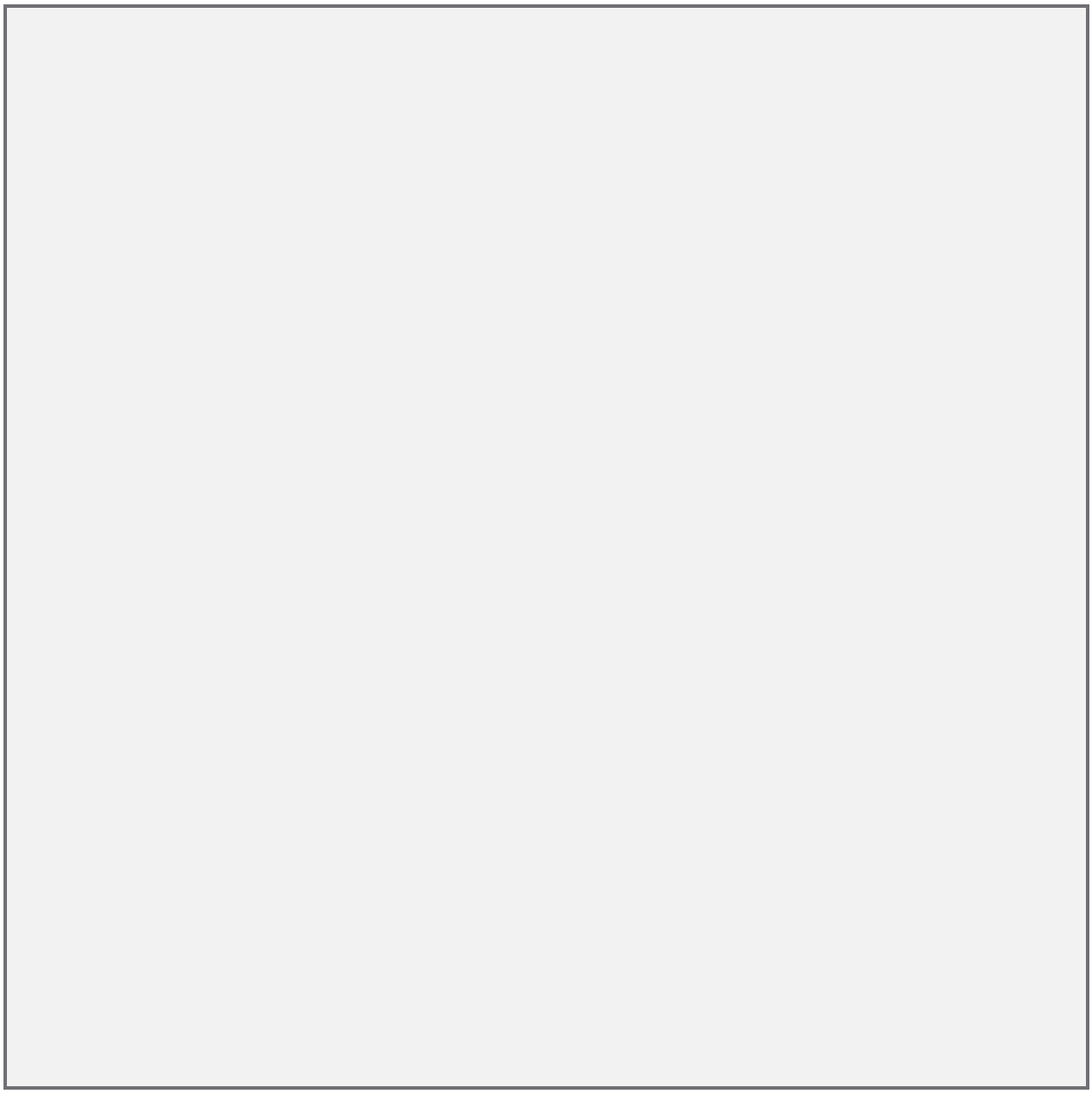
If so, indicate whether the evidence listed below support the TSS position(s).
If not, indicate whether the evidence listed below support the TSS search region(s).

Evidence type	Support	Refute	Neither
blastn alignment of the initial exon from *D. melanogaster*	☐	☐	☐
RNA PolII ChIP-Seq	☐	☐	☐
RNA-Seq coverage and TopHat splice junctions	☐	☐	☐
Core promoter motifs	☐	☐	☐
Sequence conservation with other *Drosophila* species (*e.g.*, "Conservation" track on the Genome Browser)	☐	☐	☐
Other (please specify)	☐	☐	☐

Note: The evidence type refutes the TSS annotation only if it **suggests an alternate TSS position**. For example, the presence of RNA-Seq read coverage upstream of the annotated TSS indicates that the TSS is located further upstream and it would be considered to be evidence against (i.e. Refute) the annotated TSS. In contrast, the lack of RNA-Seq read coverage is a negative result and it neither supports nor refutes the TSS annotation (i.e. Neither).

Provide an explanation if the TSS annotation is inconsistent with at least one of the evidence types specified above:

If the TSS annotation is supported by blastn alignment of the initial transcribed exon against the contig sequence, **paste a screenshot of the blastn alignment into the box below:**



If the TSS annotation is supported by core promoter motifs, RNA PolII ChIP-Seq, or RNA-Seq data, **paste a Genome Browser screenshot of the region surrounding the TSS (±300bp) with the following evidence tracks:**

1. RNA PolII Peaks
2. RNA-Seq Alignment Summary
3. RNA-Seq TopHat
4. Short Match results for the Inr motif (TCAKTY)

If the TSS annotation is supported by sequence conservation with other *Drosophila* species, **paste a screenshot of the pairwise alignment (*e.g.*, from blastn) or the multiple sequence alignment (*e.g.*, from Clustal Omega, ROAST) into the box below:**

2. 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 **within the narrow TSS search region**. (Note that the narrow TSS search region differs from the TSS search region only when you have defined both a wide and a narrow TSS search region.)

Coordinates of the motif search region

Your project (*e.g.*, contig10:1000-1600): _____

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 above.

Core promoter motif	Your project	*D. melanogaster*
BRE^{u}		
TATA Box		
BRE^{d}		
Inr		
MTE		
DPE		
Ohler_motif1		
DRE		
Ohler_motif5		
Ohler_motif6		
Ohler_motif7		
Ohler_motif8		

Have you annotated all the TSS that are in your project?

Use the GEP UCSC Genome Browser for *D. melanogaster* to identify the genes located adjacent to the first and last reconciled genes in your project. For each unique TSS of these *D. melanogaster* genes, perform a blastn search of the initial transcribed exon against your project using the more sensitive search parameters (i.e. Word size = 7; Match/Mismatch Scores = 1, -1; Gap Costs = Existence: 2 Extension: 1; turn off the low complexity filter). **Paste the screenshots of the blastn search results in the box below.** Provide an explanation for any significant (E-value < 1e-2) hits and why these hits do not correspond to real transcribed exons in your project.

Name of the *D. melanogaster* gene adjacent to the first reconciled gene in your project:

_____

Name of the *D. melanogaster* gene adjacent to the last reconciled gene in your project:

_____

Screenshots of the blastn search results:

Revised gene models report form

Complete this section if there are any changes to the coding regions of the reconciled gene models. Copy and paste this form to create as many copies as needed within this report.

Gene name (*e.g.*, *D. biarmipes eyeless*): _____
Gene symbol (*e.g.*, *dbia_ey*): _____
FlyBase release (*e.g.*, 6.25): _____

Name(s) of isoforms that have been removed from the current FlyBase release:

_____

Name(s) of new or revised isoform(s) with unique coding sequences	List of isoforms with identical coding sequences

Names of the isoforms with unique coding sequences in *D. melanogaster* that are absent in this species:

_____

Note: For each revised gene model listed above, you should use the Gene Model Checker to create the corresponding GFF, transcript, and peptide sequence files. You should use the Annotation Files Merger to combine the files for all the revised gene models into a single project GFF, transcript, and peptide sequence file prior to project submission.

Consensus sequence errors report form

Complete this section if there are any errors within the consensus sequence that affect the revised gene models or the Transcription Start Sites (TSS) annotations. **The coordinates reported in this section should be relative to the coordinates of the original project sequence.**

Location(s) within 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

Using the Sequencer Updater (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. **Paste a screenshot with the list of sequence changes into the box below:**

Revised isoform report form

Complete this report form for each unique isoform where the proposed gene model differs from the reconciled gene model. Copy and paste this form to create as many copies as needed within this report.

Gene-isoform name (*e.g.*, dbia_ey-PA): _____
Names of the isoforms with identical coding sequences as this isoform:
_____

Is the 5' end of this isoform missing from the end of project? _____
If so, how many exons are missing from the 5' end: _____
Is the 3' end of this isoform missing from the end of the project? _____
If so, how many exons are missing from the 3' end: _____

Describe the evidence used to support the proposed changes to the reconciled gene model:

1. Gene Model Checker checklist

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

2. 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.*, 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)

Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:

3. 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 into the box below:**

4. 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) into the box below. **Provide an explanation for any anomalies** on the dot plot (*e.g.*, large gaps, regions with no sequence similarity).

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