



# F Element Project: Transcription Start Sites Report

Faculty instructor(s): \_\_\_\_\_

College/university: \_\_\_\_\_

Course number: \_\_\_\_\_

Course name: \_\_\_\_\_

## Authorship Information for GEP Scientific Publications

**Initials:** By entering my/our initials, I/we grant permission for the Genomics Education Partnership (GEP) to use the annotation data produced in this report in future scientific publications.  
\_\_\_\_\_

**Note:** Please skip the rest of this section if **more than three students** contribute to this annotation report. When more than three students contribute to an annotation project, the class as a whole will be acknowledged in future GEP scientific publications.

## Co-authors Responsibilities

In order to be a co-author on a GEP publication, you must review, critique, and approve the final gene models and manuscript, responding promptly to requests to read and approve. As part of the preparations for the microPublication article, co-authors are required to validate specific data within the manuscript, supplemental materials, and GenBank submission (the specific details will depend on each annotation project). In most cases, the manuscript preparation process will take approximately 3–5 hours of your time.

The above requirements mean that we must be able to contact you when the GEP microPublication, and later, the scientific paper with meta-analysis, is ready for your review and approval. **If we cannot reach you at that time, you will not be a co-author on our GEP scientific publications**, as scientific journals require all co-authors to have read and approved the manuscript.

Please provide your contact information below. Note that your name and contact information will be publicly available through the scientific publication and the GenBank record (this is standard for all scientific publications.). Please list the authors in ascending alphabetical order by last name. (The actual order of the student co-authors in the scientific publication will be determined by a random number generator.)

**Contact information for Author #1**

**First name** \_\_\_\_\_

**Middle initials** \_\_\_\_\_

**Last name** \_\_\_\_\_

**Author name** \_\_\_\_\_

(name that will appear on the publication):

**Permanent Email address** \_\_\_\_\_

(one you will use five years from now):

**Alternative Email address (optional):** \_\_\_\_\_

**Enter your initials to indicate that you have read  
and accept the co-authors responsibilities** \_\_\_\_\_

**Contact information for Author #2**

**First name** \_\_\_\_\_

**Middle initials** \_\_\_\_\_

**Last name** \_\_\_\_\_

**Author name** \_\_\_\_\_

(name that will appear on the publication):

**Permanent Email address** \_\_\_\_\_

(one you will use five years from now):

**Alternative Email address (optional):** \_\_\_\_\_

**Enter your initials to indicate that you have read  
and accept the co-authors responsibilities** \_\_\_\_\_

### Contact information for Author #3

First name \_\_\_\_\_

Middle initials \_\_\_\_\_

Last name \_\_\_\_\_

Author name  
(name that will appear on the publication): \_\_\_\_\_

Permanent Email address  
(one you will use five years from now): \_\_\_\_\_

Alternative Email address (optional): \_\_\_\_\_

Enter your initials to indicate that you have  
read and accept the co-authors responsibilities \_\_\_\_\_

### Project Details

Project name: \_\_\_\_\_

Project species: \_\_\_\_\_

Date of submission: \_\_\_\_\_

Size of project in base pairs: \_\_\_\_\_

Number of genes in project: \_\_\_\_\_

Does this report cover 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 12 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. ananassae eyeless*): \_\_\_\_\_  
Gene symbol (e.g., *dana\_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:  
\_\_\_\_\_

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (e.g., changes in canonical splice sites, gene structure, etc.):

## Isoform TSS Report

Complete an Isoform TSS report (through page 12) for each unique TSS listed in the table above. Copy and paste this form to create as many copies as needed.

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

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

\_\_\_\_\_

Examine the peaks in the “**TSRchitect Combined RAMPAGE TSS (Replicate 1)**” track [available under the “Updated Transcriptome Tracks” section of the Genome Browser for the *D. melanogaster* Aug. 2014 (BDGP Release 6 + ISO1 MT/dm6) assembly].

### If the promoter of the isoform overlaps with a RAMPAGE peak:

Coordinates of the TSS position based on position with the highest RAMPAGE read density

\_\_\_\_\_

Coordinates of the narrow TSS search region based on RAMPAGE peaks

\_\_\_\_\_

Shape Index (SI) for the RAMPAGE peak

\_\_\_\_\_

Promoter shape in *D. melanogaster*  
(Peaked:  $SI > -1$ ; Broad:  $SI \leq -1$ )

\_\_\_\_\_

### If the promoter of the isoform does not overlap with a RAMPAGE peak:

Use the read distributions in the “**Combined modENCODE CAGE TSS**” track (available under the “Expression and Regulation” section) to characterize the shape of the promoter.

Shape of core promoter in *D. melanogaster*:  
(Peaked / Intermediate / Broad / Insufficient Evidence)

\_\_\_\_\_

**1. Turn on RAMPAGE evidence tracks  
(Only applies to projects with these tracks)**

Examine the peaks in the “**Combined CSHL RAMPAGE TSS**” track (available under the “Expression and Regulation” section of the Genome Browser for your project).

Coordinates of the TSS position based on position with the highest RAMPAGE read density

---

Coordinates of the narrow TSS search region based on RAMPAGE peaks

---

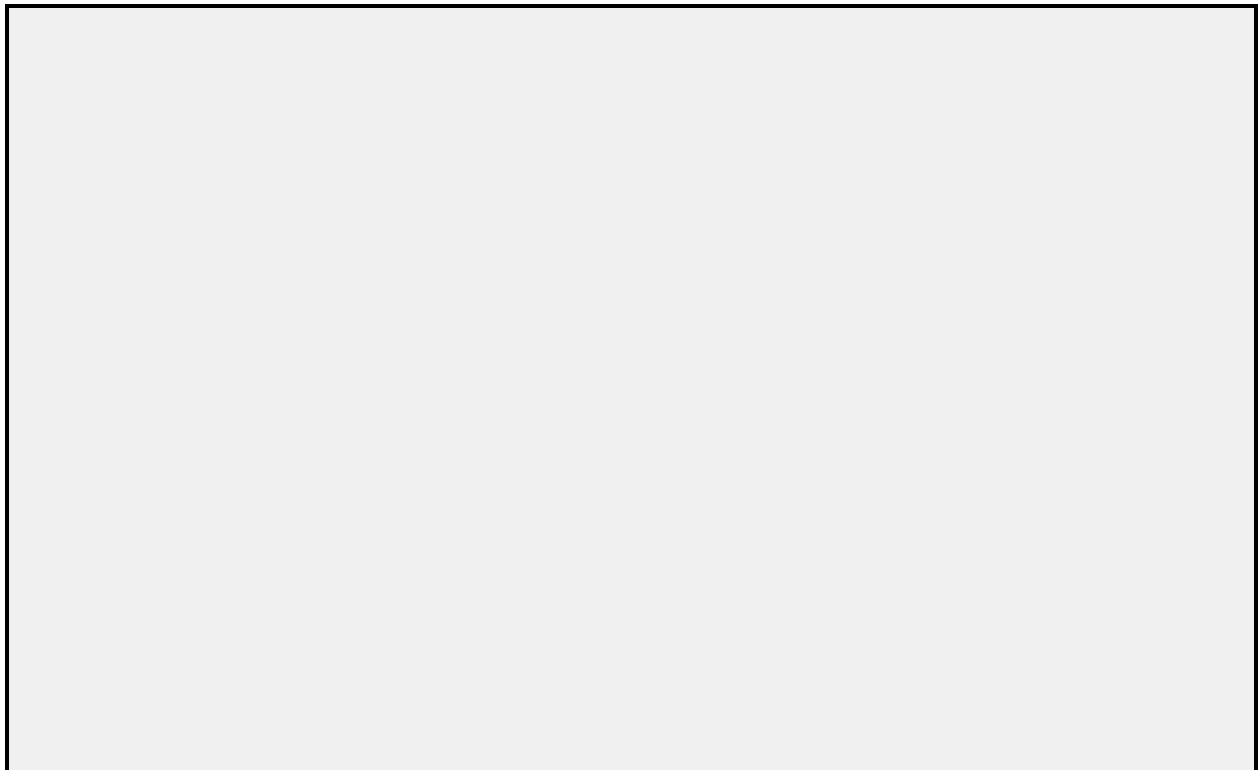
Shape Index (SI) for the RAMPAGE peak

---

Promoter shape  
(Peaked:  $SI > -1$ ; Broad:  $SI \leq -1$ )

---

If the TSS position and narrow TSS search region are supported by RAMPAGE data, **paste a Genome Browser screenshot of the region surrounding the putative TSS ( $\pm 300\text{bp}$ ) showing the “Combined CSHL RAMPAGE TSS” evidence track:**

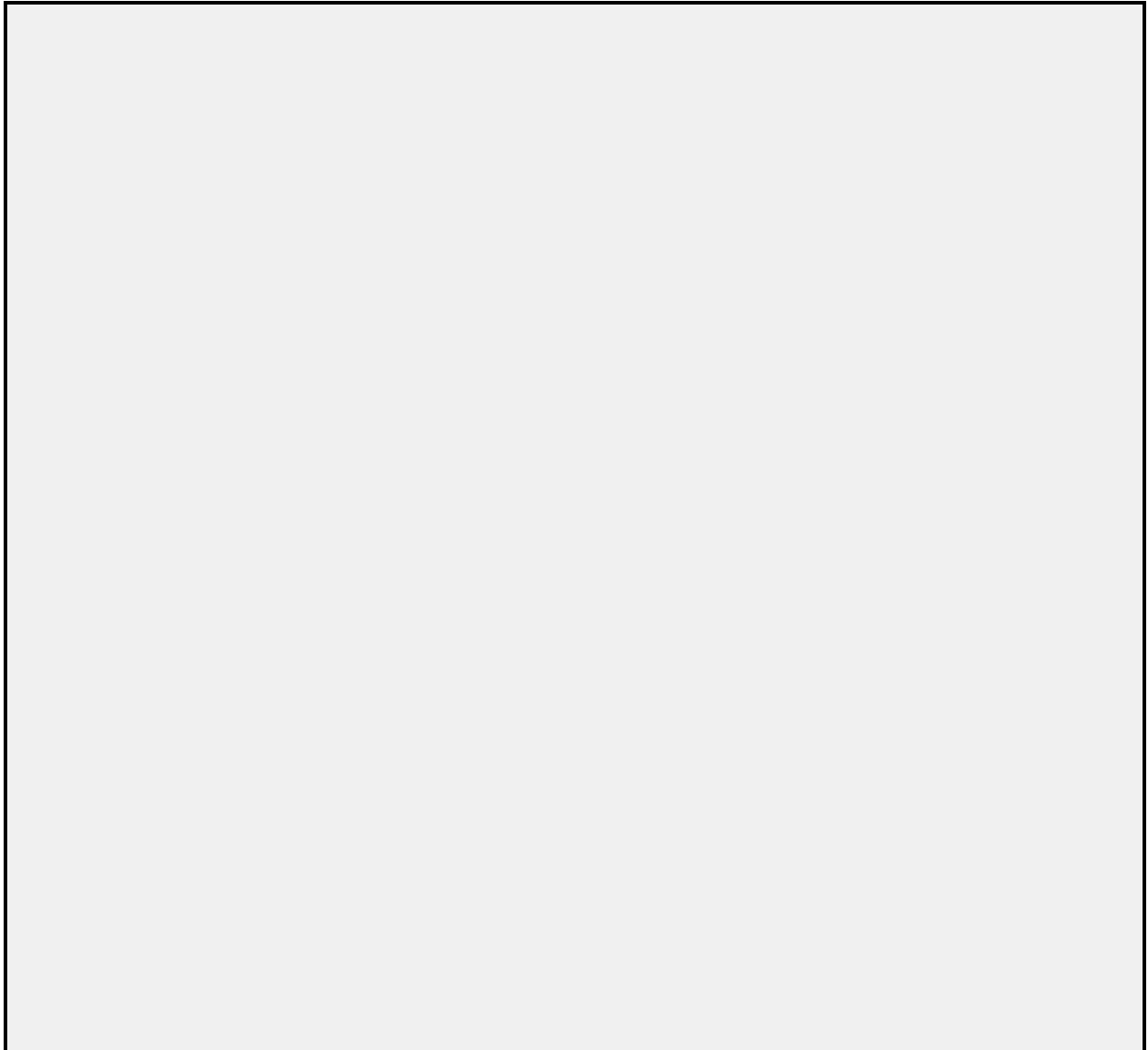


**2. Turn on ATAC-Seq evidence track  
(Only applies to projects with these tracks)**

If the wide TSS search region is supported by ATAC-Seq data, **paste a Genome Browser screenshot of the region surrounding the putative TSS ( $\pm 300\text{bp}$ ) showing the “Eye Discs ATAC-Seq” evidence track:**

Coordinates of the wide TSS search region based on ATAC-Seq peaks

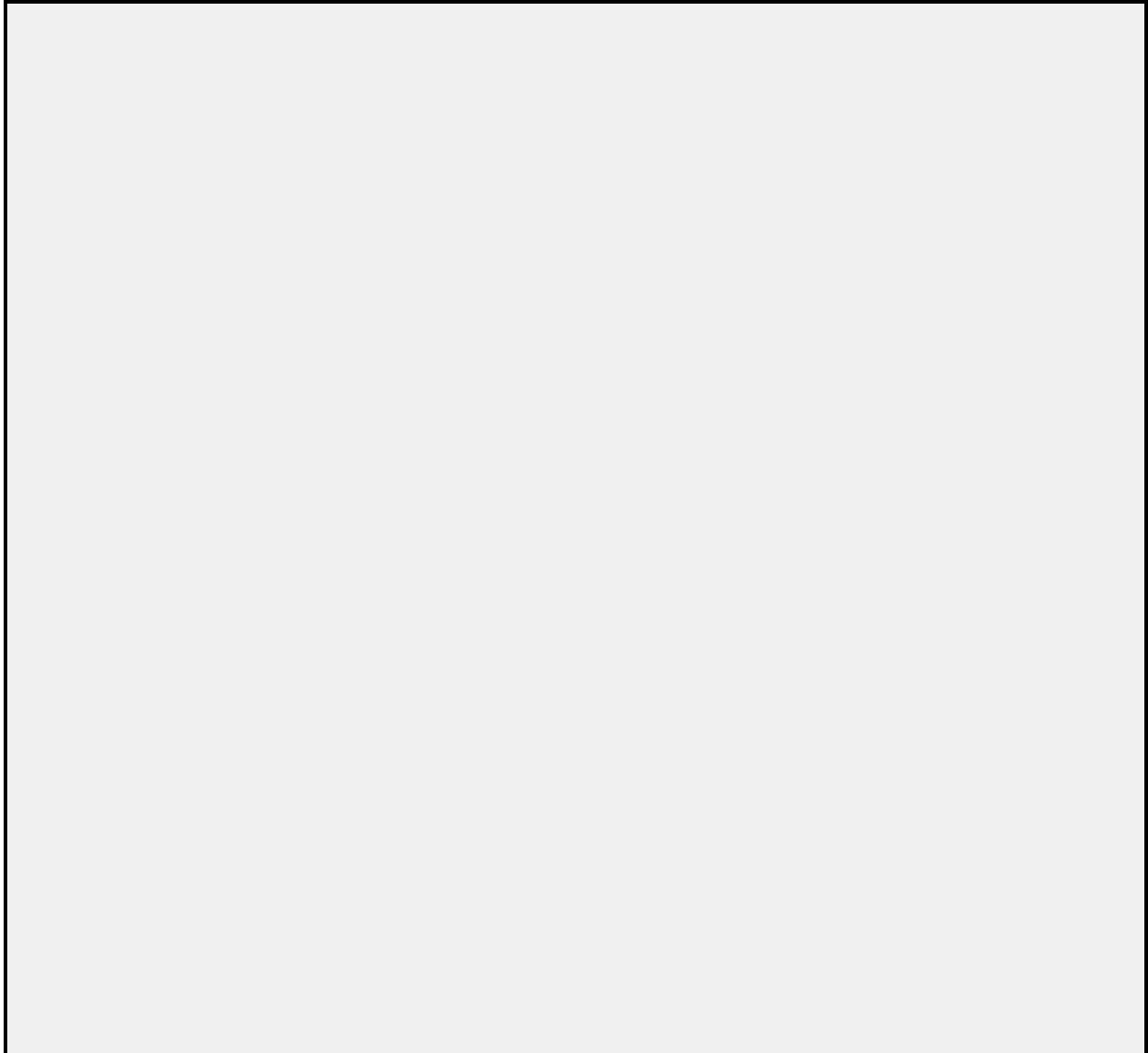
---



### 3. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (e.g., regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS ( $\pm 300\text{bp}$ ) showing the following evidence tracks:**

1. RNA-Seq Coverage or RNA-Seq Alignment Summary
2. Combined Splice Junctions or RNA-Seq TopHat



If the RNA-Seq evidence tracks indicate the TSS position, list it here: \_\_\_\_\_

If the RNA-Seq evidence tracks indicate a TSS search region, list it here: \_\_\_\_\_



#### 4. Annotate the first transcribed exon

Coordinates of the first transcribed exon based on *blastn* alignment:

---

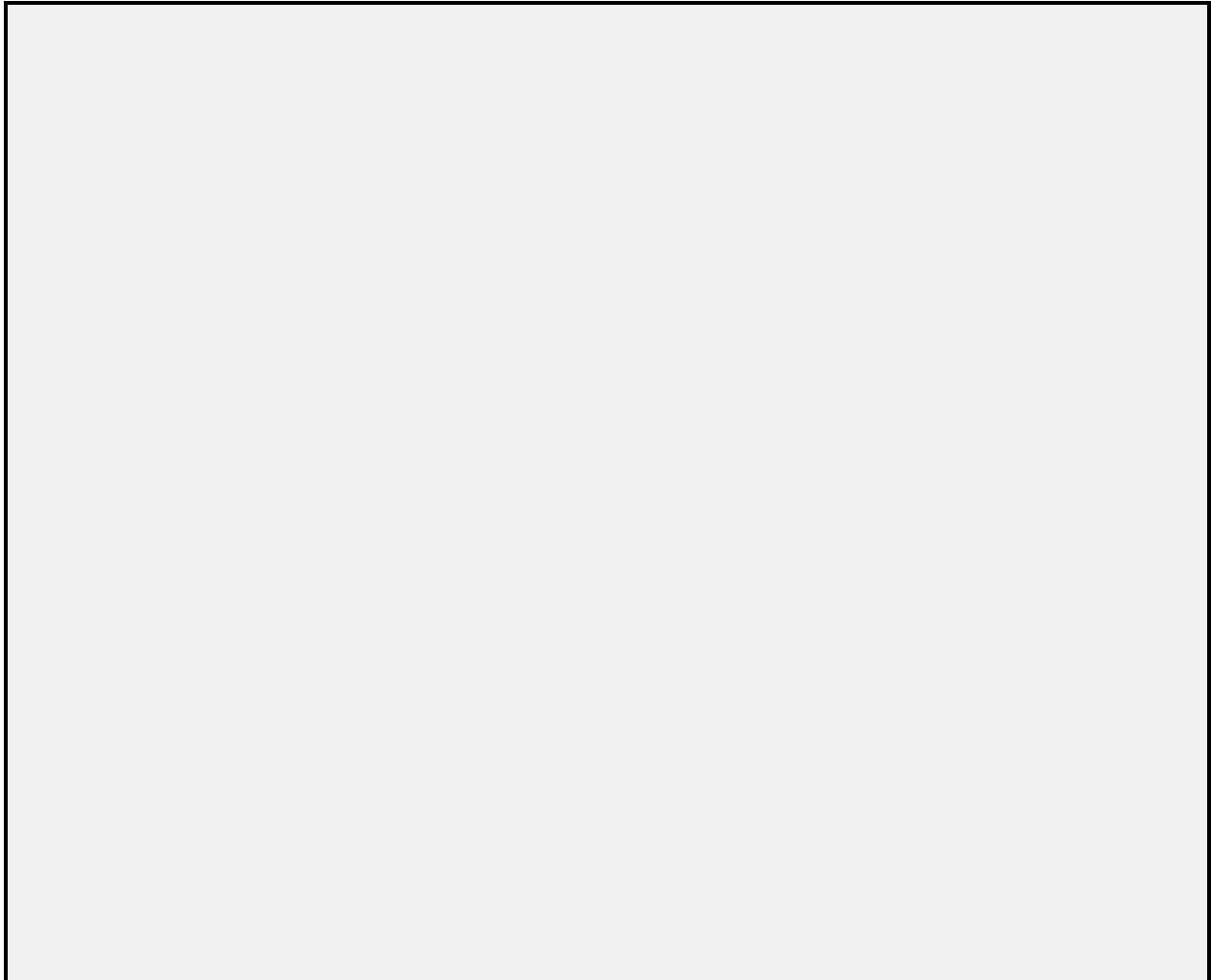
Does the *blastn* alignment cover the entire *D. melanogaster* first transcribed exon?

---

If not, specify the parts of the *D. melanogaster* exon that are missing from the *blastn* alignment.

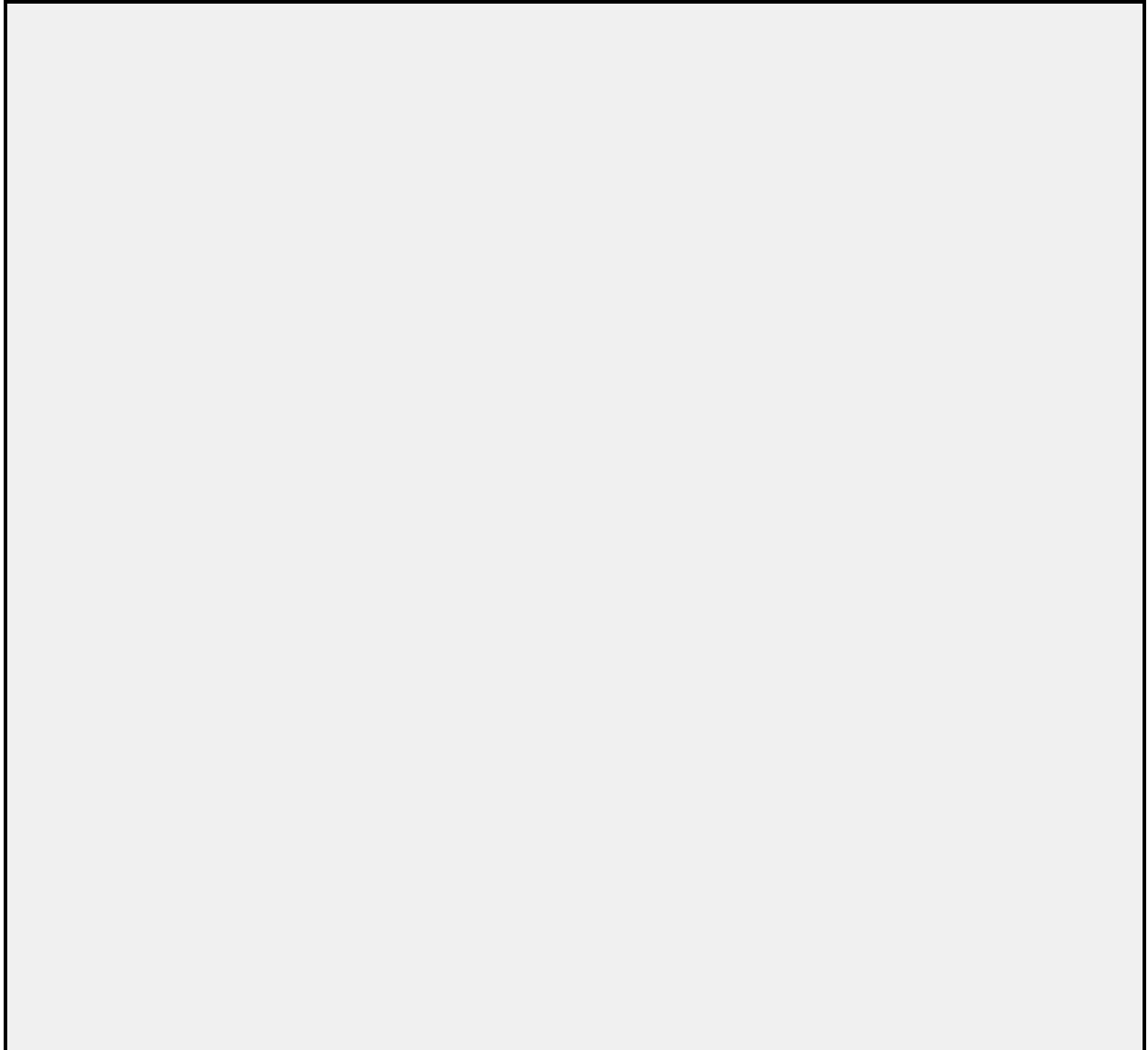
---

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:**



**5. Turn on comparative genomics tracks**

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



**6. Summarize the evidence that supports the TSS annotation postulated above**

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

Based on RAMPAGE data(if applicable): \_\_\_\_\_

Based on ATAC-Seq data (if applicable): \_\_\_\_\_

Based on RNA-Seq data: \_\_\_\_\_

Based on *blastn* alignment: \_\_\_\_\_

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.)

Were you able to define a TSS position based on the available evidence? \_\_\_\_\_

If so, indicate in the table below the evidence that supports this TSS position

If not, were you able to define a TSS search region? \_\_\_\_\_

If so, indicate in the table below the evidence that supports the TSS search region(s)

For each evidence type, enter an "X" in the cell to indicate whether the line of evidence supports, refutes, or neither supports nor refutes the TSS annotation:

Evidence type	Support	Refute	Neither
RAMPAGE peaks and read density			
ATAC-Seq peaks and log likelihood enrichment profile			
RNA-Seq coverage and splice junctions			
<i>blastn</i> alignment of the initial exon from <i>D. melanogaster</i>			
Sequence conservation with other <i>Drosophila</i> species (e.g., “Conservation” track on the Genome Browser)			
Other (please specify) (e.g., RefSeq Genes, N-SCAN PASA-EST, Augustus TSS predictions; histone modifications CHIP-Seq data).			

**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 the annotated TSS; check “Refute.” In contrast, the lack of RNA-Seq read coverage is a negative result that neither supports nor refutes the TSS annotation; check “Neither.”

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

## 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. 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 below).

Gene name (e.g., *D. ananassae eyeless*): \_\_\_\_\_

Gene symbol (e.g., *dana\_ey*): \_\_\_\_\_

FlyBase release (e.g., 6.52): \_\_\_\_\_

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

Names of the isoforms that have been removed from the current FlyBase release:

\_\_\_\_\_

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

Provide the evidence (text and figures) which support the hypothesis that these isoforms are absent in this species (e.g., changes in canonical splice sites, gene structure, etc.):

Note: For each revised gene model listed above, you will also submit gene model files which describe your isoform as a DNA sequence (FASTA), a peptide sequence (PEP), and as a collection of exon coordinates that can be visualized on the GEP UCSC Genome Browser (GFF). You should use the [Gene Model Checker](#) to **create these files for ALL revised gene models**. (If the coding sequence of the revised gene model is identical to that of another isoform, you should create the gene model files for each isoform.) 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.

All of the coordinates reported in this section should be relative to the coordinates of the original project 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 includes a CDS alignment with frame shifts or in-frame stop codons, and RNA-Seq reads with discrepant alignments compared to the project sequence.

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

Use the [Sequence Updater](#) to 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 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 symbol (e.g., dana\_ey-PA): \_\_\_\_\_

Names of any additional isoforms with identical coding sequences:

---

Is the 5' end of this isoform missing from the end of the project? \_\_\_\_\_

If so, how many putative 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 putative exons are missing from the 3' end: \_\_\_\_\_

(Define "putative exons" based on the exons present in the *D. melanogaster* ortholog)

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

## 1. Gene Model Checker checklist

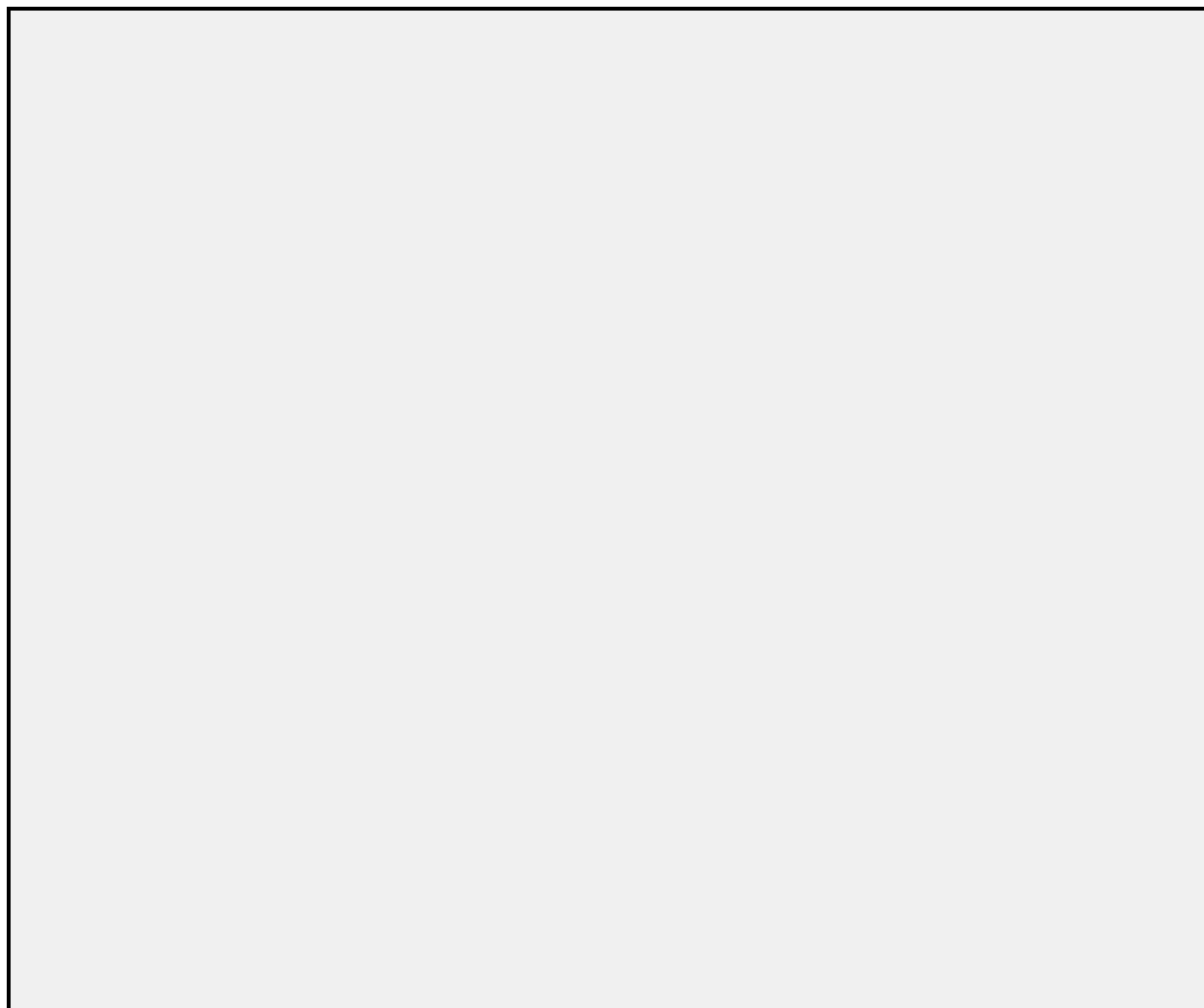
Coordinates of your final gene model for this isoform:

\_\_\_\_\_

Stop codon coordinates: \_\_\_\_\_

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:** This screenshot should show the “**Configure Gene Model**” panel with the exon coordinates and the “**Checklist**” panel with all the checklist items (i.e., from the criteria “Check for Start Codon” to “Number of coding exons matched ortholog”). If necessary, include multiple screenshots of the “Checklist” panel to capture all the checklist items.

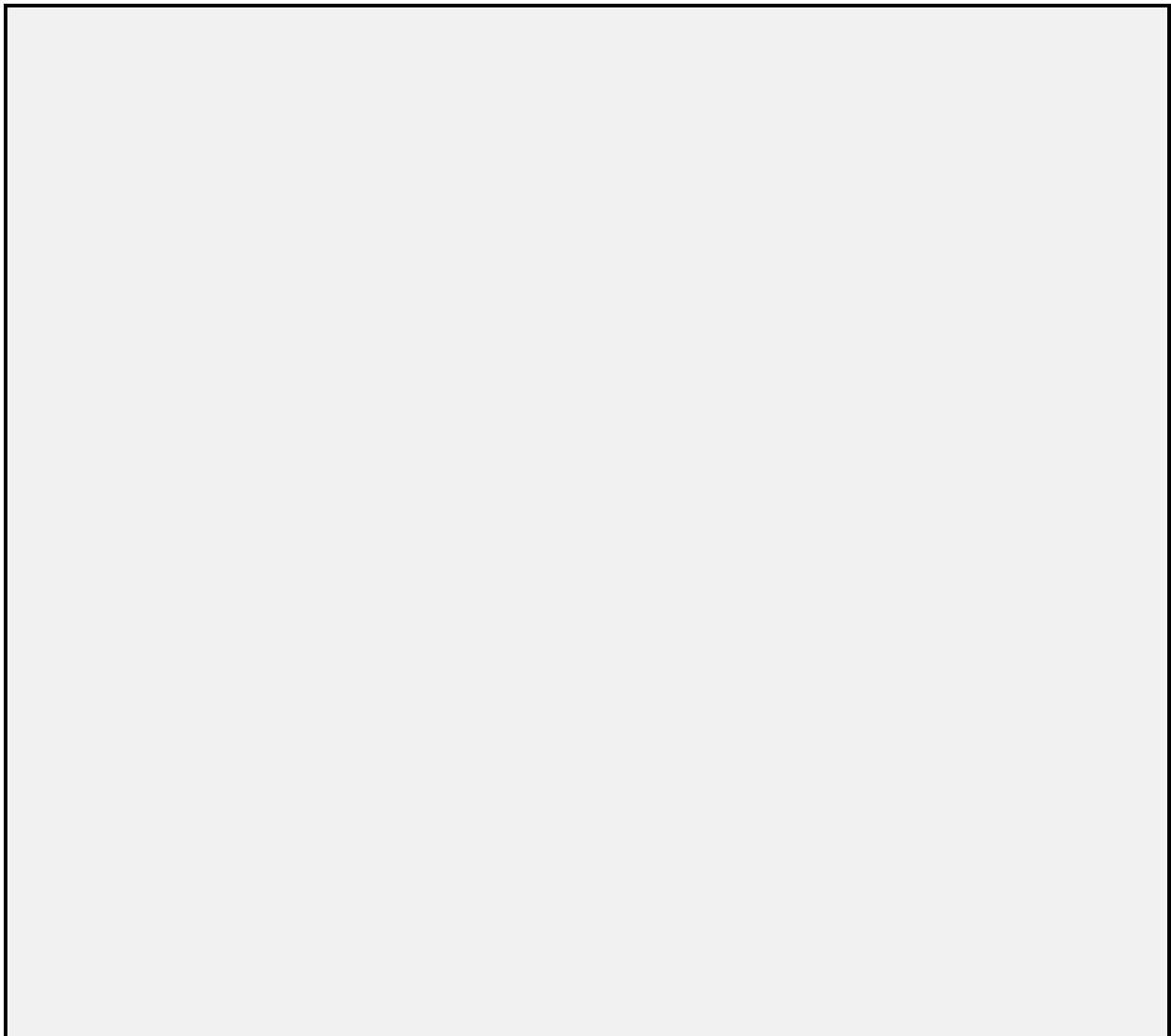


## 2. View the gene model on the Genome Browser

Click on the magnifying glass icon under the “Checklist” tab of the [Gene Model Checker](#) to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

1. A sequence alignment track (e.g., D. mel Proteins)
2. At least one gene prediction track (e.g., Genscan, N-SCAN PASA-EST, Augustus)
3. At least one RNA-Seq track (e.g., RNA-Seq Coverage)
4. A comparative genomics track (e.g., D. mel. Net Alignment, Conservation)

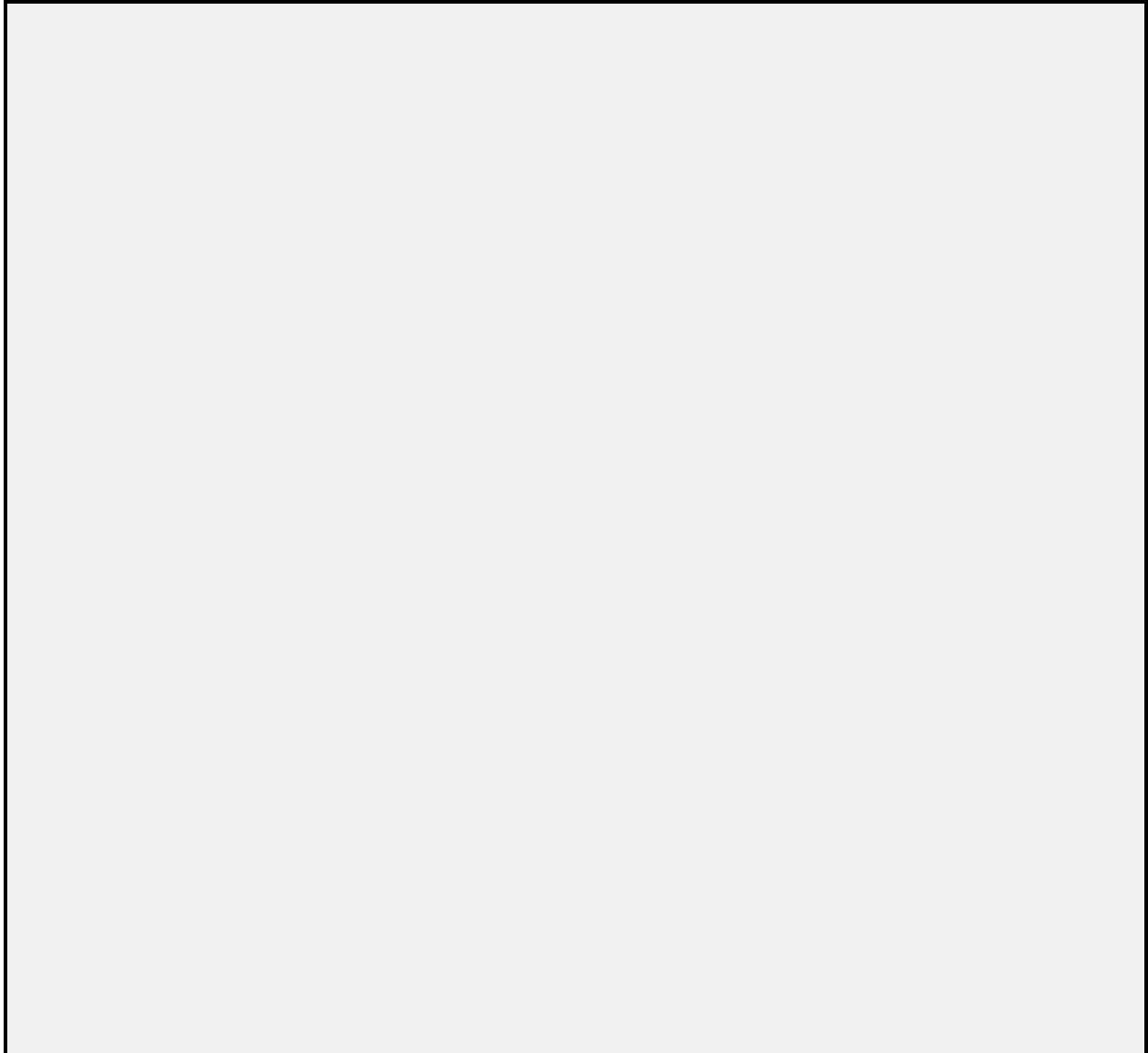
**Paste a screenshot of your gene model as shown on the GEP UCSC Genome Browser into the box below:**





### 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 at the NCBI BLAST website. **Paste a screenshot of the protein alignment into the box below:**

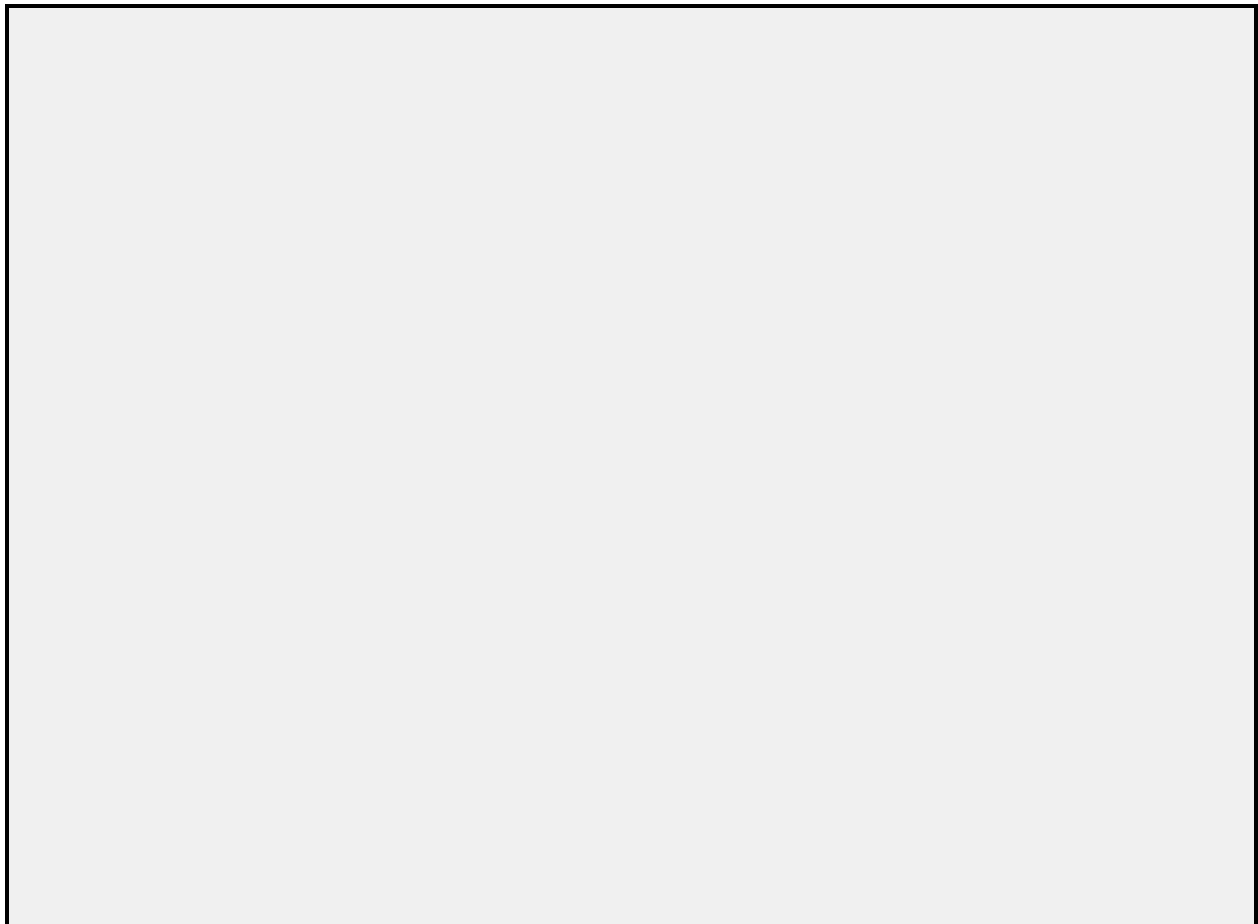


#### 4. Dot plot between the submitted model and the *D. melanogaster* ortholog

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

Provide an explanation for any anomalies on the dot plot (e.g., large gaps, which would appear as kinks in the diagonal line; regions with no sequence similarity; indications of significant insertions or deletions).

**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.



Thank you for your submission, and congratulations on completing your analysis of this region of this genome. Our planned GEP meta-analysis of the genes and genomes in this study depends on the high quality annotations accomplished by GEP students.