F Element Project: Annotation 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: For each gene described in this annotation report, you should also prepare the corresponding **GFF, transcript and peptide sequence files** as part of your submission.

Complete the following Gene Report Form for each gene in your project. Copy and paste the sections below to create as many copies as needed within this report. Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms. 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). If you cannot find evidence for any protein-coding genes in the project, jump to the "Check for additional features in your project" section.

Gene Report Form

Gene name (e.g., *D. ananassae eyeless*): _____
Gene symbol (e.g., *dana_ey*): _____

Approximate location in project (from 5' end to 3' end): _____
Number of isoforms in *D. melanogaster:* _____
Number of isoforms in this project: _____

Complete the following table, including all of the isoforms in this project:

Name(s) of unique isoform(s) based on coding sequence	**List of isoforms with identical coding sequences**

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: In addition to submitting your annotation report, you will also submit gene model files which describe your isoform(s) 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). While we only require one Isoform Report form per unique coding sequence, **we also require a full set of gene model files (GFF, FASTA, and PEP) for *ALL* isoforms, even if their coding sequence is identical** to that of another isoform. See page 31 of the Gene Model Checker User Guide for details.

Consensus Sequence Errors Report Form

Complete this section if you have identified errors in the project consensus sequence that affect the annotation of the gene described above.

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:**

Isoform Report Form

Complete this report form for each unique isoform listed in the table above. Copy and paste this form to create as many copies of this Isoform Report Form as needed.

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)

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

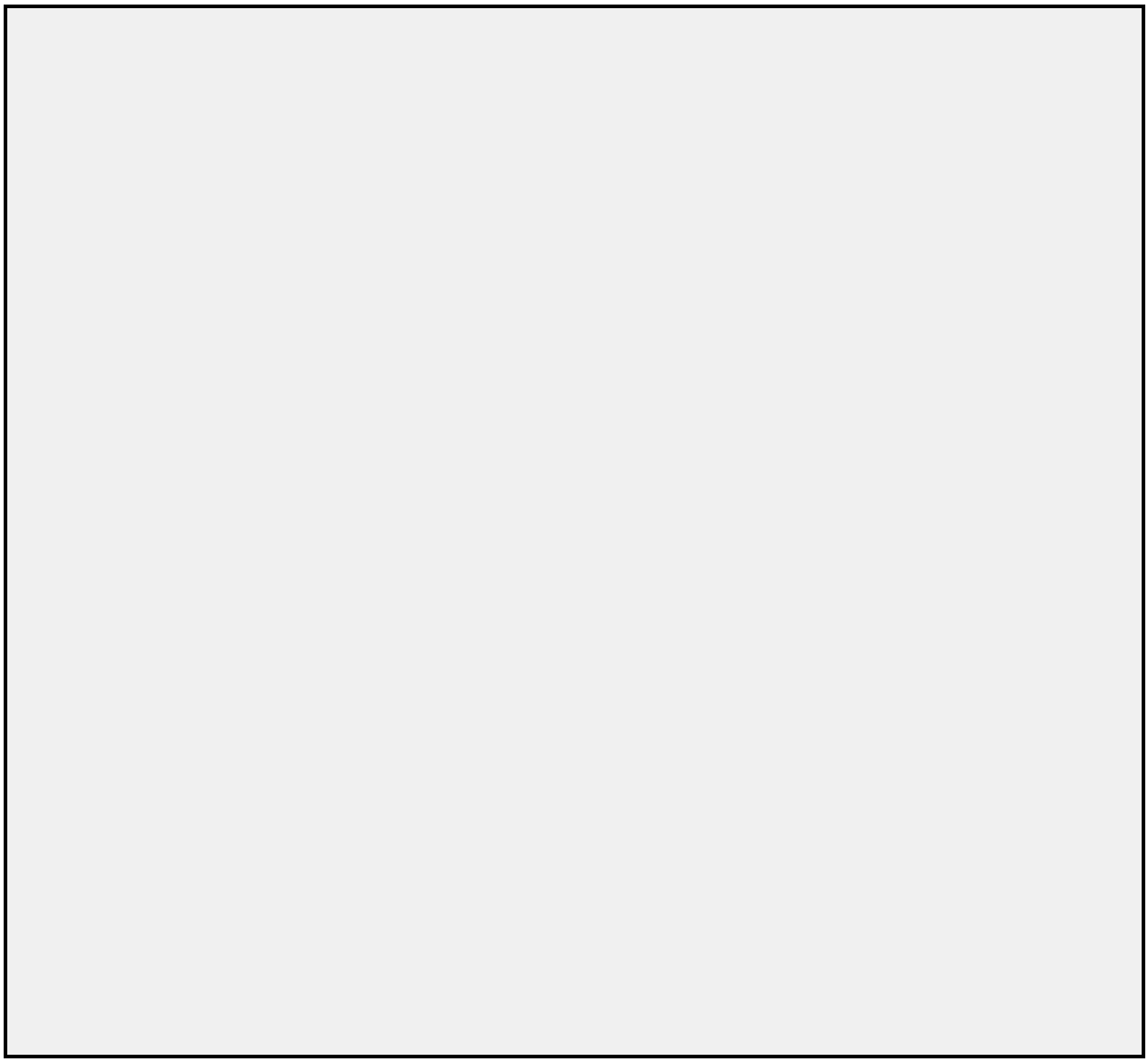
Click on the magnifying glass icon under the "Checklist" tab of the Gene Model Checker to view your gene model on the GEP UCSC Genome Browser. Zoom in so that **only this isoform is in the genome browser window, and capture a screenshot that includes the following evidence tracks if they are available:**

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.

Transcription Start Sites (TSS) Report Form (Optional)

Note: Complete this section if you have annotated the TSS for the gene above. This section is **optional** and you do not need to complete this section to submit the project.

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 18) 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 ≤ -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 ≤ -1)

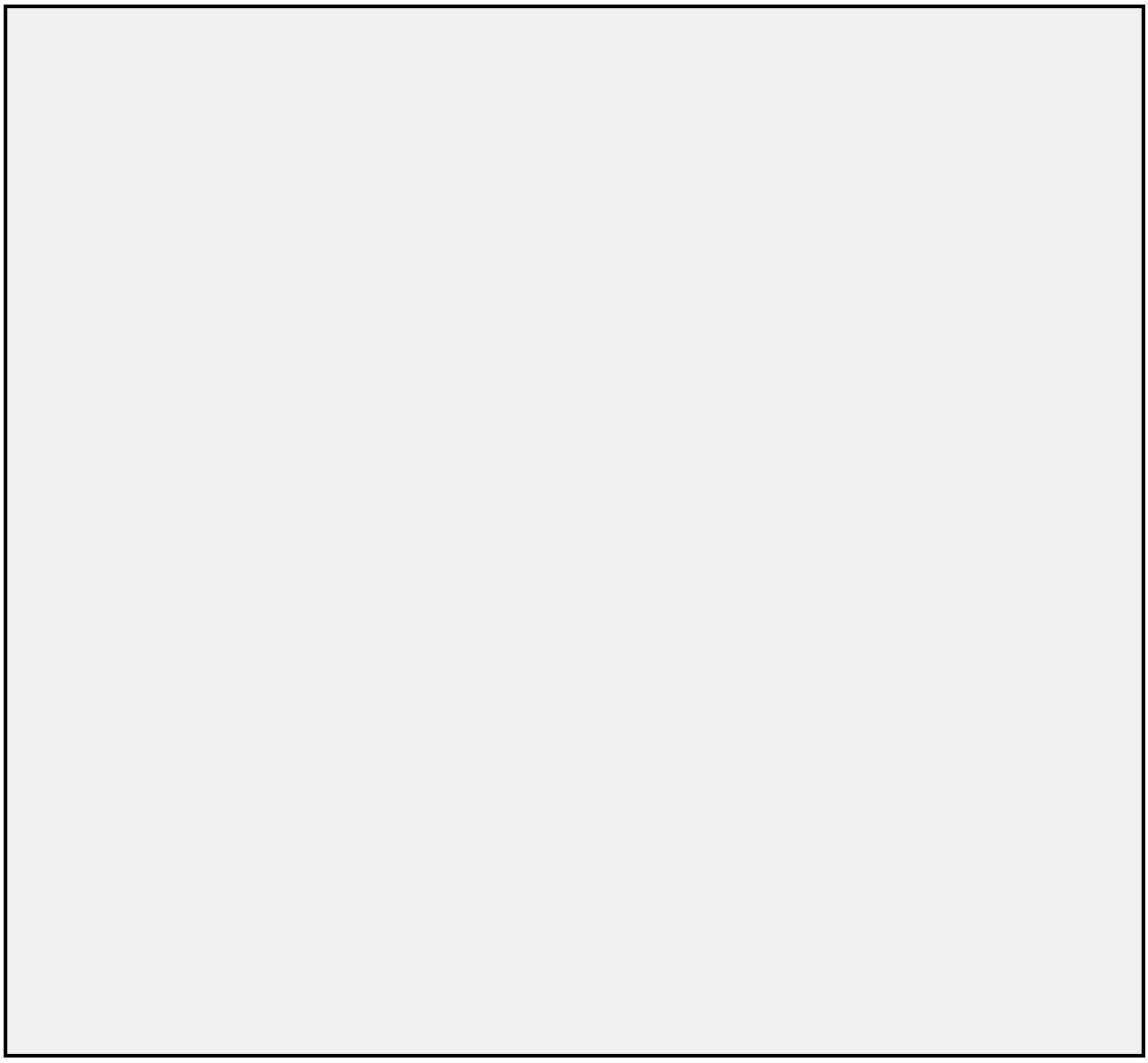
_____

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 (±300bp) 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 (±300bp) showing the "Eye Discs ATAC-Seq" evidence track:**

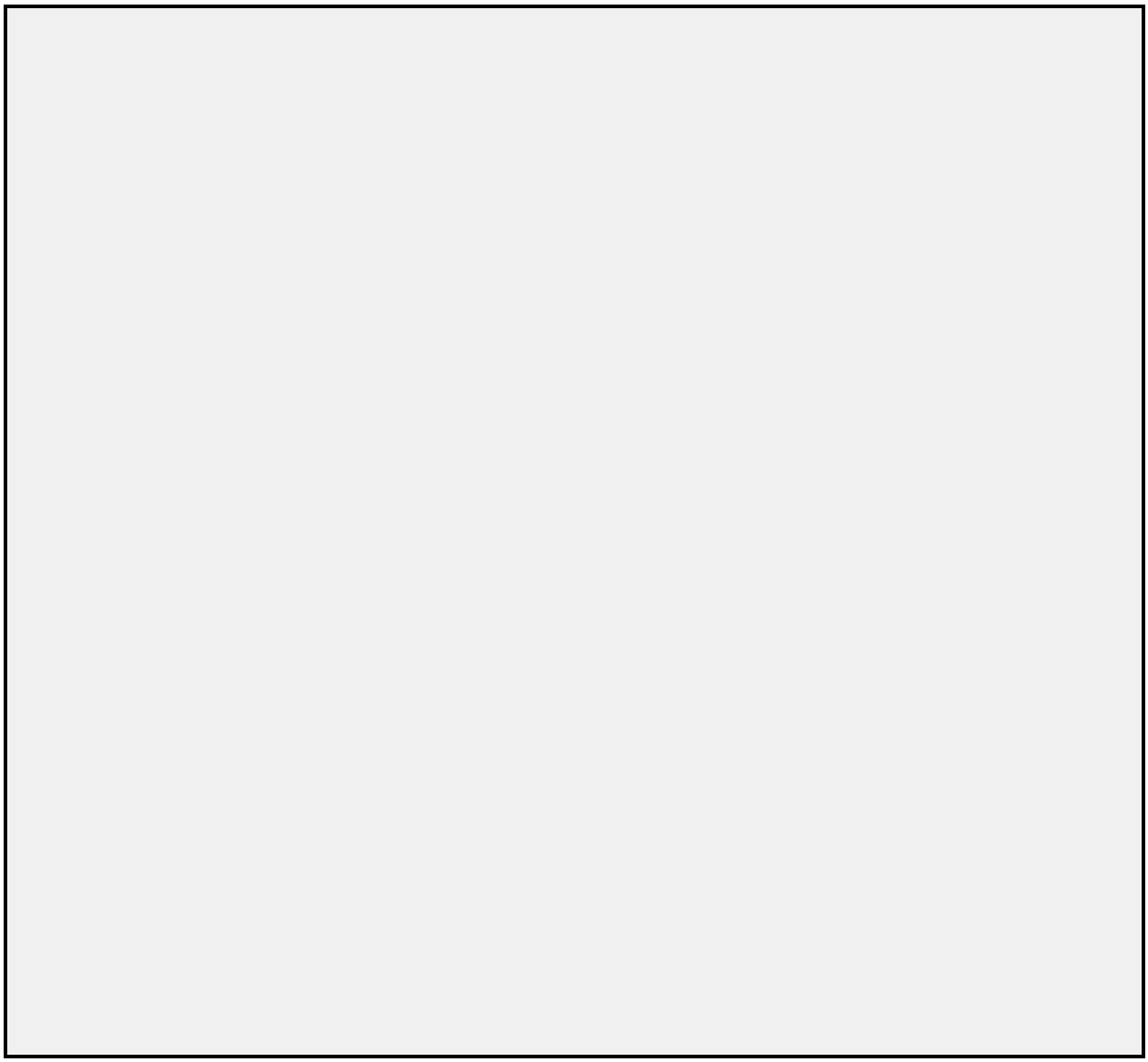
Coordinates of the wide TSS search region based on ATAC-Seq peaks



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 (±300bp) 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:

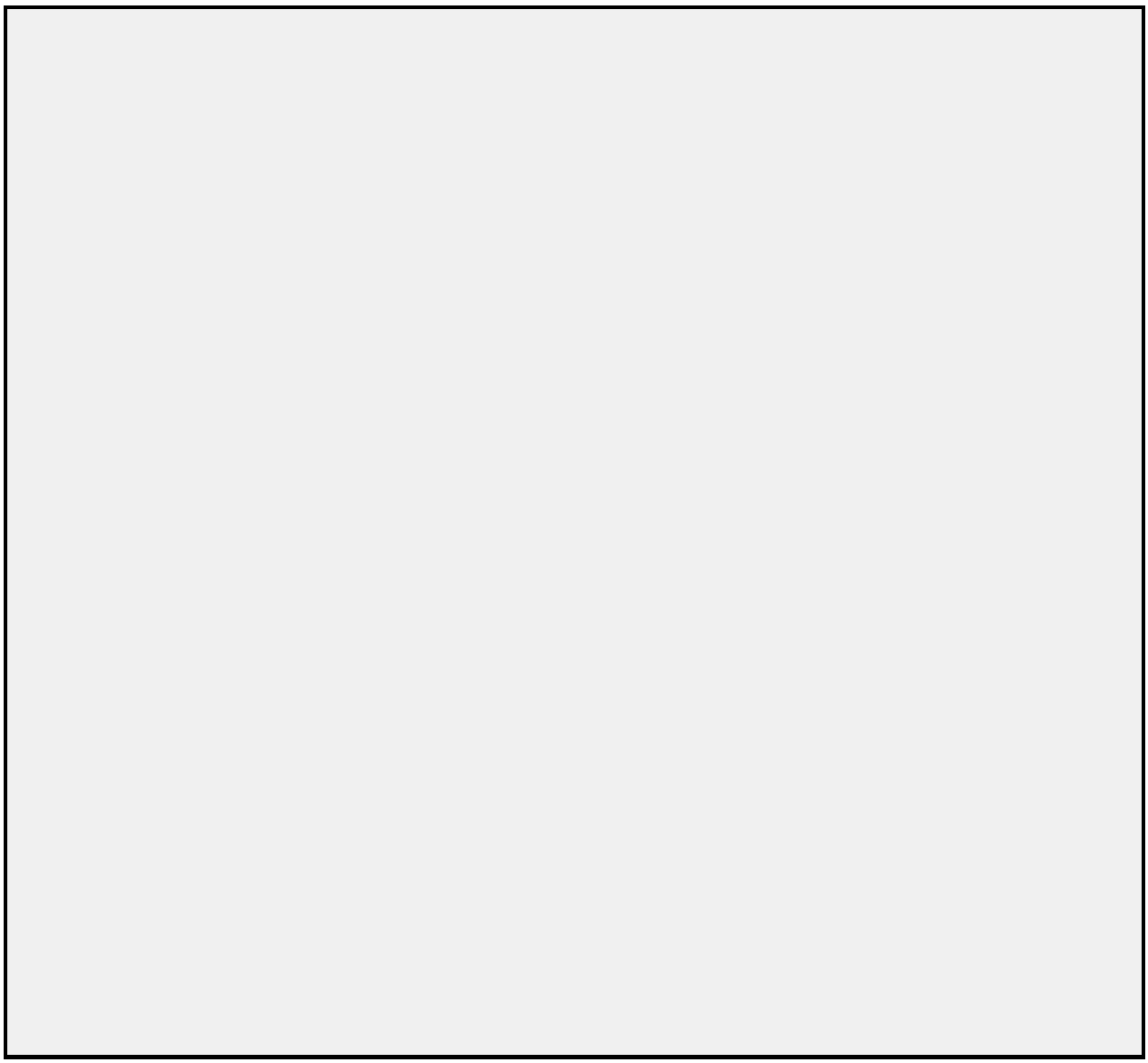
Does the *blastn* alignment cover the entire *D. melanogaster* first transcribed exon?

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			
blastn alignment of the initial exon from *D. melanogaster*			
Sequence conservation with other *Drosophila* 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:

Check for additional features in your project

For each Genscan gene prediction that does not overlap with the genes you have already annotated, perform the following analyses to determine if the feature corresponds to a protein-coding gene, pseudogene, or partial gene duplication.

1. Perform a FlyBase *blastp* search of the predicted protein sequence from Genscan against the *D. melanogaster* "**Annotated proteins**" database. Report the significant matches (E-value < 1e-5) to protein sequences in *D. melanogaster*:

2. If there are significant matches to *D. melanogaster* proteins, analyze the genomic region immediately surrounding the Genscan prediction using the exon-by-exon strategy. Report your findings:
    - If the feature is a functional protein-coding gene, construct the gene model in the target species and provide the supporting evidence for the gene model in a new Gene Report Form

    - If the feature is a pseudogene or a partial gene duplication, provide the evidence (text and figures) which support these hypotheses:
        - Evidence for a pseudogene includes in-frame stop codons, and frameshifts within coding exons
        - Changes in gene structure from a multi-exon gene in *D. melanogaster* to a single-exon gene in the target species could indicate a retrotransposed pseudogene or retrogene

3. Perform a NCBI *blastp* search of the predicted protein sequence from Genscan against the "**Reference proteins (refseq_protein)**" database. By default, this will search all the reference proteins associated with all the organisms in databases. Report the significant matches (E-value < 1e-5) to curated RefSeq gene models:
    - Protein records curated by the NCBI RefSeq database have the prefix "**NP_**"

In addition, you should examine the gene expression tracks (e.g., RNA-Seq data) for evidence of transcribed regions that do not correspond to the features you have already annotated, or transposon remnants identified by RepeatMasker. Perform an NCBI *blastx* search of these genomic regions against the **refseq_protein** database to determine if they show significant similarity (E-value < 1e-5) to curated RefSeq gene models (i.e., protein records with the prefix "**NP_**"). Report as above.

Note: The primary goal for this section of the Annotation Report is to identify genomic features that require further investigations. Detailed analysis of these genomic features might require tools and techniques beyond the standard annotation protocols. The "Annotation of Other Genomic Features within the F Element Project" presentation on the GEP website illustrates several strategies that can be used to annotate some of these features (e.g., paralogs, pseudogenes, retrogenes, and partial gene duplications).

Preparing the Project for Submission

For each project, you should prepare the project GFF, transcript, and peptide sequence files for **ALL** isoforms along with this report. You can combine the individual files generated by the Gene Model Checker into a single file using the Annotation Files Merger. Once you have combined the GFF files into a single file, click on the "**Show Track**" button to view all the gene models in the combined GFF file within the Genome Browser.

Paste a screenshot (generated by the Annotation Files Merger) with all the gene models you have annotated in this project into the box below.

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. **Paste a screenshot of the Genome Browser (generated by the Annotation Files Merger) showing the locations of all the consensus errors with respect to the original project sequence into the box below.**

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.