GEP Annotation Report

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.

Student name(s): _____
Student email(s): _____
Faculty advisor(s): _____
College/university: _____

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 _____

Instructions for project with no genes

If you believe that the project does not contain any genes, please provide the following evidence to support your conclusion:

1. Perform a NCBI BLASTX search of the entire contig sequence against the "non-redundant protein sequences (*nr*)" database. Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the *nr* database as to why they do not correspond to real genes in the project.

2. For each Genscan prediction, perform a NCBI BLASTP search of the predicted amino acid sequence against the *nr* protein database using the strategy described above.

3. Examine the gene expression tracks (*e.g.*, RNA-Seq) for evidence of transcribed regions that do not correspond to alignments to known *D. melanogaster* proteins. Perform a NCBI BLASTX search against the *nr* protein database using these genomic regions to determine if they show sequence similarity to known or predicted proteins in the *nr* database.

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.

Gene report form

Gene name (*e.g., D. biarmipes eyeless*): _____
Gene symbol (*e.g., dbia_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 for all 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:

_____

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 <u>ALL</u> isoforms**, irrespective of whether they have identical coding sequences as other isoforms.

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 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 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 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 the 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: _____

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.

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. 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 an Isoform TSS report (through page 13) for each unique TSS listed in the table above. If the gene has more than one unique TSS, copy and paste this form to create as many copies as needed.

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* (see table below):
(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**

1. 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:**

2. Turn on RNA-Seq evidence tracks

If the TSS annotation is supported by RNA-Seq read coverage or splice junction predictions (*e.g.*, TopHat, regtools), **paste a Genome Browser screenshot of the region surrounding the putative TSS (±300bp) with the following evidence tracks:**

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

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

3. 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:**

4. Define the TSS search region(s)

Note:
If the blastn alignment to the initial transcribed exon satisfies the criteria listed on page 8 of Module TSS4 (i.e. a long match with low E-value, requires extrapolation of less than 150bp to the estimated TSS position, alignment is in concordance with other evidence tracks), then you can define the TSS search region as +/- 300 bp from the initial 5' nucleotide. For example, if the estimated TSS position is located at position 1500, then the narrow TSS search region would be placed at 1200-1800.

If you cannot estimate the TSS position based on the blastn alignment to the initial transcribed exon, then you can define the TSS search region(s) based on the experimental data (*e.g.*, RNA-Seq, RNA PolII ChIP-Seq) and the conservation track for the target species. 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**" as the wide search region and "**1500-2000**" as the narrow search region.

Coordinates of the narrow TSS search region:

_____

Coordinates of the wide TSS search region:
(Enter "NA" if the narrow TSS search region is defined based on the blastn alignment to the initial transcribed exon. Enter "Insufficient evidence" if a wide search region cannot be defined based on the available evidence)

_____

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*):

_____

_____

_____

5. 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 (and not a single coordinate), you should report all motif instances within the narrow TSS search region. If you did not report a narrow TSS search region due to insufficient evidence, report the motif instances in the wide TSS search region.

Coordinates of the motif search region

Your project (*e.g.*, contig10:1500-2000): _____

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.)

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		

6. Summarize all of the evidence that supports the TSS annotation postulated above.

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.)

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

If so, indicate whether the evidence listed below support the TSS position. _____

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 splice junctions	☐	☐	☐
Core promoter motifs	☐	☐	☐
Sequence conservation with other *Drosophila* species (*e.g.*, "Conservation" track on the Genome Browser)	☐	☐	☐
Gnomon, N-SCAN, and Augustus TSS predictions	☐	☐	☐
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:

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

In order to identify TSS for partial genes 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 boxes 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:

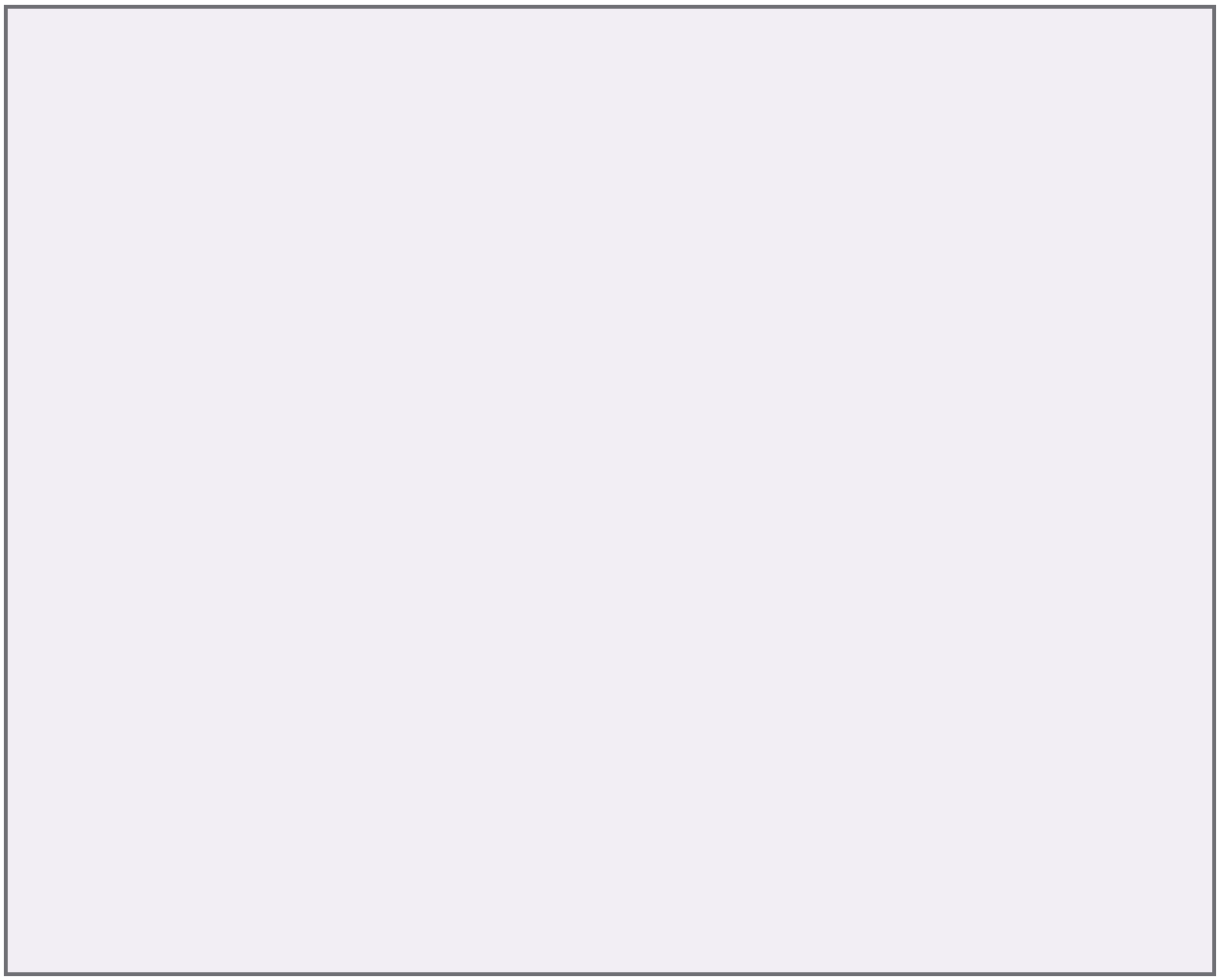
_____

Screenshot of the blastn search results:

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

_____

Screenshot of the blastn search results:



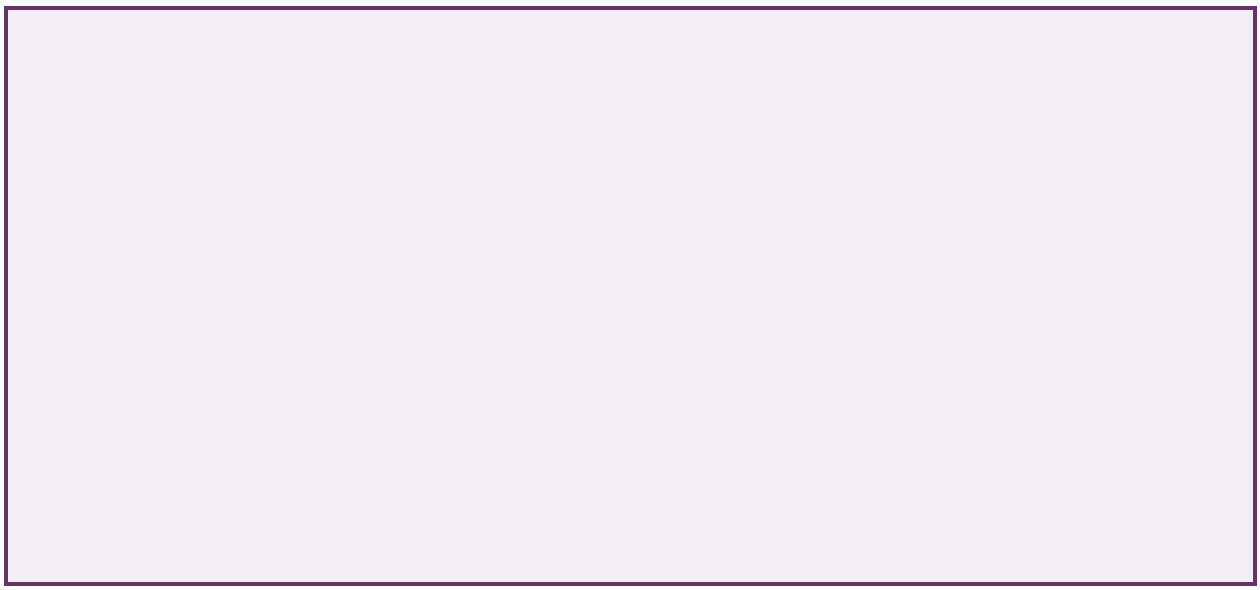
Preparing the project for submission

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.

The Annotation Files Merger also allows you to view all the gene models in the combined GFF file within the Genome Browser. Please refer to the Annotation Files Merger User Guide for instructions on how to view the combined GFF file on the Genome Browser (you can find the user guide under "Help" ➔ "Documentations" ➔ "Web Framework" on the GEP website at http://gep.wustl.edu).

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 (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 into the box below.**



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

For each region of the project with gene predictions that do not overlap with your gene annotations, perform a NCBI BLASTP search using the predicted amino acid sequence against the "non-redundant protein sequences (*nr*)" database. **Paste a screenshot of the search results into the box below.** Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the *nr* database and why these hits do not correspond to real genes in your project.