Last Update: 08/14/2012

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: ______________________________
Student email: _____________________________
Faculty Advisor: ___________________________
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 genes and all isoforms or is it a partial report? _______________
If this is a partial report because different students are working on different regions of this sequence, please report 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 conclusions:

1. Perform a BLASTX search of the entire contig sequence against the non-redundant (*nr*) protein 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 BLASTP search using the predicted amino acid sequence against the protein database (*nr*) using the strategy described above.

3. Examine the gene expression tracks (e.g. cDNA/EST/RNA-Seq) for evidence of transcribed regions that do not correspond to alignments to known *D. melanogaster* proteins. Perform a BLASTX search against the *nr* database using these genomic regions to determine if the region is similar to any 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. Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms.

Gene report form

Gene name (i.e. *D. mojavensis eyeless*): _______________
Gene symbol (i.e. dmoj_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:
If you are annotating untranslated regions then all isoforms are unique (by definition)

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

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.

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 (i.e. dmoj_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: ______________

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 below**:

2. View the gene model on the Genome Browser

Using the custom track feature from the Gene Model Checker (see page 10 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). 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. Include the following evidence tracks in the screenshot if they are available.

1. A sequence alignment track (D. mel Protein 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, 3-way, 5-way or 7-way multiz)

Paste the screenshot of your gene model as shown on the Genome Browser 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 use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Copy and paste the alignment below**:

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

Paste a copy 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: Large vertical and horizontal gap near exon boundaries in the dot plot often indicates 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.

Preparing the project for submission

For each project, you should prepare the project GFF, transcripts 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 detail 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.

Have you annotated all the genes?

For each region of the project with gene predictions that do not overlap with putative orthologs identified in the BLASTX track, perform a BLASTP search using the predicted amino acid sequence against the non-redundant protein database (*nr*). **Provide a screenshot of the search results.** Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the *nr* database and why you believe these hits do not correspond to real genes in your project.