Lab Weeks 10 & 11 - The How to Guide for Gene Annotation in *Drosophila:* Steps to Complete and Submit Projects

(adopted from the Genomics Education Partnership - http://gep.wustl.edu/)

Overview

The DNA to be annotated is present in **fosmids**, which are cloning vectors that contain approximately 40,000-60,000 bp of DNA sequence from various *Drosophila* species. Each student will work with a partner to annotate a gene, or one specific isoform of a gene that encodes multiple protein isoforms. This year, all students will annotate genes from the long arm of the third chromosome (3L) from *Drosophila erecta* or *Drosophila mojavensis*. These sequences are being used as controls to which the dot (fourth) chromosome data will be compared. We anticipate that annotation of genes on this chromosome will yield interesting results as well. Also note that *D. mojavensis* is more distantly related to *D. melanogaster* then is *D. erecta* (see phylogeny at http://flybase.org/static_pages/species/sequenced_species.html), which may make annotating sequences from *D. mojavensis* a bit more challenging!

Steps 1 and 2 – Finding and Confirming the *Drosophila melanogaster* Ortholog

As a group, we will examine each fosmid selected for your lab section. The researchers at Washington University have pre-screened each fosmid to identify one (or more) candidate *D. melanogaster* ortholog(s) to each gene feature. The UCSC Genome Browser window for each fosmid shows the results of that screen. ***Refer to pp. 117-120 of "A Sample Annotation Problem" for more detail about the information contained in the Genome Browser window***. In brief, the tracks of the UCSC Genome Browser window show all the database and gene predictor searches that were done on each fosmid. By selecting "full" for the black "D. mel Proteins" track, one can see the detailed results of the BLASTX search that looked for sequence similarity between the translated fosmid sequence and all known *Drosophila melanogaster* proteins (the amino acid sequences of which were obtained by translating the mRNA sequences for each isoform). You should also compare the features identified by the different gene predictors (e.g., Genscan, Nscan, etc.) in the UCSC Browser window to the "D. mel Proteins" track to gain additional support for particular gene orthologs.

Confirm that your selected gene is the best candidate ortholog from *Drosophila melanogaster* for your particular fosmid feature by doing a blastp search of the Predicted Protein sequence of the gene against all known *D. melanogaster* proteins using FlyBase (see the middle of p. 121 in the lab manual).

Step 3 – Discover Something about Your Selected Gene

Go to Flybase.org and do a QuickSearch of your selected feature to find out its Gene Symbol and Gene Name (if known) in *Drosophila melanogaster*. Select 'genes' in the menu for Data Class. Do not include the hyphen and the (right) isoform letters when typing your gene symbol (case-sensitive) into FlyBase; look for (and select) its symbol in the drop-down box as you type. In the results window that next appears, you should be able to confirm the chromosomal location of the gene in *D. melanogaster*. Does it match the chromosomal source of the fosmid DNA for your gene (e.g., is it also on the 3L chromosome in *D. melanogaster*)?

Next, click on the small box to the left of 'Gene Model & Products' in the blue area under the diagram to see an expanded view of the gene and all of its mRNA (Transcript) and protein (CDS) isoforms. We find that spending a few minutes now looking over the Transcript and CDS models on

FlyBase helps considerably in making sense of the Gene Record Finder charts. Note that the orange boxes in each of the Transcripts are the coding exons, whereas the gray boxes at the ends represent exon sequences that are untranslated (e.g., they are the 5' and 3' untranslated regions [UTRs], respectively). Finally, click on the 'Summary Information' box and/or follow other links in FlyBase to find out what is known (if anything) about the function of the protein encoded by the *Drosophila melanogaster* ortholog to your gene.

Now you are ready to begin the detailed gene annotation of your selected feature. If successful, you will be the first or second group (Washington University requires two independent analyses of each fosmid) to completely annotate this region of this Drosophila genome. ***How exciting is that!!***

Step 4 – Investigating the Coding Sequences (CDS)

First, open up a Web browser window and a Word document. ***Each time that you go to a different Web site, open a separate browser window (NOT tab), as you will need to have two or more of the browser windows open and visible at the same time.*** Now, map the position in the *D. spp* fosmid of each coding exon, as follows (***see pp. 111-112 and 121-125 in the lab manual for more detailed directions***):

a. Use the Gene Record Finder (http://gep.wustl.edu/ -> Projects -> Annotation Resources ->) to obtain the amino acid sequences of all the *D. melanogaster* coding sequences (CDS) found in your particular gene (or isoform, if the gene encodes multiple isoforms).

b. Highlight your particular isoform by clicking on the corresponding row in the CDS usage map, then ***paste a screen shot (Ctrl-Alt-Print Screen) into the Word document of the table below*** in the 'Polypeptide Details' window, which shows the lengths of all the CDS for your particular protein isoform**.**

c. Obtain the amino acid sequence of each exon by clicking on its row, or obtain all of the CDS sequences at one time by clicking on the 'Export All CDS's for Selected Isoform to FASTA' tab next to Options. Align each *D. melanogaster* CDS to the *D. mojavensis* or *D. erecta* fosmid sequence by doing blastx searches at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi). Search one by one, examine amino acid alignment, note base coordinates of the beginning and end of each coding sequence, and note frame. *Note*: frames for all exons must be all – OR all + for any gene. The frame *number* can vary, though (1, 2 or 3), from CDS to CDS.

d. ***Pay careful attention to the lengths of the alignment blocks*** as compared to the length of the *D. melanogaster* CDS, to make sure that you have found the full-length CDS in your fosmid. Sometimes CDS lengths do vary between species, though, so do speak with your instructor to see if this might be the case for any alignment blocks of unusual lengths.

e. ***Copy and paste all the information (including the frame) for the best blastx alignment to each D. melanogaster CDS into the same Word document as above.*** Be sure to title each alignment with the FlyBase ID number of each from the CDS table of the Gene Record Finder. You will likely need to set the page margins to 0.5 inch or reduce the font size to 9, so the format of the Word document will look like the BLAST results window (e.g., 60 amino acids per row, with the number at the right on the same line).

Step 5 – Creating a Gene Model by Careful Mapping of the Exon-Intron Boundaries

Confirm the exact base coordinates of the start and stop codons and the end and beginning of each coding exon in between to create a gene model, as follows (***see pp. 126-127 of "A Sample Annotation Problem"***):

a. Keep the Word document you have already created open on one side of your monitor. Then, re-open the GEP UCSC Genome Browser (http://gep.wustl.edu/ -> Projects -> Annotation Resources ->) and use the Gene Prediction Tracks to line up the beginning and end of each alignment block from Step 4 with the fosmid nucleotide and amino acid sequences (in 3 different frames).
b. Use alignment blocks in the gene predictors, frames, conserved amino acid sequences in your *D. spp* from the blastx alignments and intron splice site consensus sequences to map the exon-intron boundaries.
c. **If the coding information for your feature is on the minus (–) strand of DNA**, click the 'reverse' button (below the tracks window) to map the exon-intron boundaries. This flips the fosmid DNA sequence so that the minus strand now goes from left to right in the 5' to 3' direction. **Note** that the numbering of the base pair coordinates of the minus strand also *decreases* in sequence from left to right after 'reverse' is clicked.
d. **Add** the coordinates of the beginning and end of each CDS, the phase of each exon-intron and intron-exon junction, and the coordinates of the stop codon to the Word document from Step 4. **Save this Word document to your folder on the U drive or a thumb drive.**

Step 6 – Checking Your Gene Model, Part I

Confirm the accuracy of your model using the Gene Model Checker, as follows (for detailed instructions, ***see pp. 128-130 in the lab manual and/or the on-line GEP Gene Model Checker User Guide at*** *http://gep.wustl.edu/help/documentations/index#gene_model_checker_guide).*

a. Select 'Gene Model Checker' from the Projects -> Annotation Resources -> drop-down menu at http://gep.wustl.edu/.
b. Right-click on and download a copy of the fasta sequence file for your particular fosmid on to the desktop from the '2011 Drosophila Gene Annotation Labs' folder on Blackboard. Upload this sequence file into the first box on the left of the Gene Model Checker window.
c. Type in the name (case-sensitive) of the *D. melanogaster* isoform that is the ortholog of your gene in the second box – watch for this gene to appear in the drop-down menu box as you type.
d. Enter the base number of the beginning and end of each coding sequence, in the order they appear in the in the table above, in the following format: # - #, # - #, # - #, etc.
e. ***Do not include the stop codon in the last CDS since the stop codon does not code for an amino acid***. Instead, enter the stop codon coordinates in # - # format in the box lower down in the Gene Model Checker window.
f. If the information for your gene is on the (-) strand of the fosmid, click the circle in front of 'Minus' following 'Orientation of Gene Relative to Query Sequence.'
g. If your gene was cut off by one end of the fosmid and not all the gene's coding sequences were present, click the circle in front of 'Partial Translation' in the next row. You should also check 'Partial Translation' if you were unable to map all the CDS in the ortholog to the fosmid sequence
h. Use the drop-down menu to select the species and assembly date for your fosmid (***see GEP Project Management System handout for this information***).
i. Type in the name/number of your contig or fosmid, e.g., contig32, fosmid35, etc.

j. ***Note: Steps h and i must have compatible information from the Project Management System handout. Red boxes will appear around any entries that the Gene Model Checker deems incorrectly entered. Fix these before going on.***
k. Click on the 'Verify Gene Model' box at the bottom of the window.
l. A summary of how your model did now appears on right side of the Gene Model Checker window. If there are failed parts of the Gene Model Checker, click on the small + box to the left of the failed part to get more information on the problematic sequence. A simple misreading or mis-typing of intron-exon boundary coordinates is responsible for many model failures.
m. If your problem was "Fail with premature stop codons', click on the Peptide Sequence tab in the upper-right menu bar next to "Checklist.' The symbol * in the peptide sequence that appears next will show you where these premature stop codons are. You can then use this information to find the coordinates of the problematic CDS using your saved blastx alignments. Start with the most 5' premature stop codon, since if you fix this one, all the rest may disappear!
n. Re-check your work and the typing of the coordinates until you pass **ALL** parts of the Gene Model Checker. **Note**: if necessary, you may elect to use a non-canonical GC intron donor site, in which case you are allowed to fail this part of the Gene Model Checker (as long as you then explain your rationale for using a GC splice site in the Annotation Report).
o. Click on the 'Peptide Sequence' tab to see the amino acid sequence created by your gene model. You will next compare this sequence to the *D. melanogaster* protein sequence for a final check of your gene model.

Step 7 - Checking Your Gene Model, Part II

It is important to now check how well the amino acid sequence predicted by your gene model aligns with the sequence of the putative *D. melanogaster* ortholog. You will do this by performing a blast2seq (blast, protein-protein) alignment that compares the entire amino acid sequence of the *D. melanogaster* ortholog to a translation of the coding sequence of the *D. mojavensis* or *D. erecta* gene/isoform predicted by your gene coordinates model. Proceed as follows:

a. Go to NCBI's Blast Home page (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and open a protein blast (blast) search window.
b. Enter the name of the gene/isoform that you annotated in the 'Job Title' box and check the 'Align two or more sequences' box.
c. Click on the 'Peptide Sequence' tab of the Gene Model Checker to obtain the amino acid sequence of the entire polypeptide predicted by your gene model. Copy and paste this sequence into the 'Query Sequence' box of the blast window.
d. Go to Flybase.org to obtain the complete amino acid sequence of the *D. melanogaster* ortholog, as follows.
    I. Select 'polypeptides' in the Data Class box of the Quick Search window.
    ii. Type the gene name in the 'Enter text' box (including hyphen and isoform information, which should appear in the pop-up window as you start typing).
    iii. Click on the 'Sequence' link in the lower-right corner of the Search results window. Copy and paste this amino acid sequence into the 'Subject Sequence' box in the blast search window.
    iv. Make sure the Low complexity regions box is unchecked in the Algorithm parameters. BLAST away!

Examine this alignment to assess whether the character of your predicted protein looks right. For instance, does the overall length of your predicted *D. erecta* or *D. mojavensis* (Query) sequence match well to the *D. melanogaster* ortholog? Are there large gaps (-------) in the *D. melanogaster* (Subject) Sequence in the interior of the alignment? If so, these perhaps could be regions where your *D. erecta* gene model has an additional intron, in which case you should search through this region of the fosmid in the UCSC Browser for additional intron splice sites. Conversely, are there large gaps (------) in the *D. grimshawi*, *D. erecta* or *D. mojavensis* sequence? This would mean that the predicted *D. spp.* protein is missing amino acids that are present in *D. melanogaster*. This could be hard to reconcile if it is a protein that is functional and has been well characterized in *D. melanogaster*. So, you should search these regions of the fosmid on the UCSC Browser to see if you missed coding sequences and truncated the protein.

If this analysis causes you to change your gene model, enter your new coordinates into the Gene Model Checker and re-do the blast2seq alignment. ***Copy and paste a screen shot of the Gene Model Checker window (with all passes!) and the final blastp alignment into the appropriate spaces in the Annotation Report form (see Step 10 below). A brief discussion of ideas from the above paragraph should follow the blastp alignment.***

Step 8 – Making a Diagram of Your Gene Model

You will now create an image of your gene model, using the coordinates you typed into the Gene Model Checker, as follows.

a. Click on the small magnifying glass to the left of any of the CDS entries in the right-hand window of the Gene Model Checker. This will open up a new UCSC Browser Window, which now has a red 'Custom Track".
b. Enter the beginning and ending base coordinates, respectively, of the start and stop codon (plus ~50-100 bases on either end) into the 'position/search' box to zoom in to the entire gene feature that you annotated.
c. Set the black D. mel Proteins track to 'pack' and make sure that the Gene Prediction tracks (set to dense) that best support your model are visible. If available, make these tracks visible as well: other ref seq and 3-way multiZ or 5-way multiZ).
d. ***Take a screen shot of the Browser window and paste it into the Annotation report form. Below this screen shot, discuss how well the different gene predictor tracks support your model.*** If there are any discrepancies between the different tracks, discuss which discrepancies you think support or do not support your particular gene model and whether or not you consider your model more accurate than some (or all!) of the gene predictors.

Step 9 – Getting the Data Together

Next, you will create three data files for your gene model, which will be combined with others from that fosmid for final submission to Washington University, as follows.

a. Click on the 'Downloads' tab in the upper-right corner of the Gene Model Checker window.
b. Then, right-click on each of the three links that appear (GFF File, Transcript Sequence File and Peptide Sequence File) to save each particular file to your desktop.
c. ***Finally, save a copy of each of these files into a folder on the U drive or a thumb drive.*** The name of the folder should include your last name and the name of the gene/ isoform that you annotated**.**

Step 10 – Finishing the Annotation Report Form

Complete an Annotation Report for your isoform/gene as follows.

a. Download the Annotation Report Word file from the '2011 Drosophila Gene Annotation Labs' folder on Blackboard and rename it, including your last names and the name of the gene/isoform that you annotated.
b. Type information for your gene/isoform into the appropriate spaces in the Word version of the Annotation report. Precede the gene name by D. erecta or D. mojavensis and the Gene symbol (and Gene-isoform name) by dere_ or dmoj_ in the report (name and symbol from Step 3 above).
c. Fill in the Gene Record Finder number of the missing CDS for any gene/isoform that is cut off by one end of the fosmid and is missing one or more CDS at the beginning or end of the isoform.

Each student team should submit their individual Annotation Report file, the three data files from Step 9 and the Word file (from Steps 4 and 5 above) for instructor review and grading. Submit these files in a single folder, which includes your names and the gene name in its title, and put this folder in the appropriate fosmid folder on the lab thumb drive and on your U drive. Be prepared to talk briefly to the rest of the lab section (over pizza the last week of lab!) about any interesting characteristics and/or challenging aspects of annotating your particular gene feature.

Finally, Dr. Emerson will compile all the individual Annotation Reports into one final report for each fosmid that we claimed this semester. She will then log into the GEP's Project Management System (under the 'Projects' tab) to upload the final Annotation Report and the three composite data files for each fosmid, thereby sending these data along to scientists at Washington University in St. Louis!