Lab Weeks 10 & 11 – Gene Annotation
Helpful Tips for Finding Difficult Coding Exons (and other problems…)

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

1. Re-read pp. 87-98 in the Lab Manual
2. Re-read the Sample Annotation Problem (pp. 117-130) and the BLAST activity (pp. 99-115) and look at their answer keys on Blackboard
3. Read the Advanced Annotation Instruction Sheet (pp. 131-133)
4. **Determine whether you have a plus strand or a minus strand sequence**. If you have a minus strand sequence, click the 'reverse' button below the UCSC Genome Browser window to display the minus strand sequence going left to right (5' to 3') across the top of the browser window. Note that the base pair coordinates now DECREASE from left to right.
5. Many genes were cut in two when the genomic DNA was fragmented and packaged into fosmids for cloning. **Determine whether or not all of the gene for your isoform is physically located on the claimed fosmid.** If it is not, you do not want to spend much time looking for a CDS that you will never find on that fosmid!
6. **When you are trying to decide between or confirm intron splice sites**, change the Predicted Splice Sites setting (at bottom of Genome Browser mirror window, under 'Experimental Tracks') to dense or full to see the position and ranking of potential donor and acceptor splice sites.
7. Try these variations when doing the exon-by-exon blastx searches:
    a. **Make sure the Filter for 'Low complexity regions' box is NOT checked under the blast Algorithm parameters.**
    b. Search the smallest region of the fosmid sequence you can in your blastx searches (e.g., reduce the size of the 'haystack' when trying to find a small 'needle'). When you "get DNA" from the Genome Browser, type in the coordinates for the narrowest region of DNA in which you think your CDS resides (for example, if the gene is on the + strand of a fosmid that is 35,000 bp long and the end of the previous CDS was at 31,000, you would only have to search the region from 31,000-35,000 for the next CDS.
    c. Decreasing Word size under blast Algorithm parameters from 3 to 2 has worked for finding some small exons!
    d. Increase the Expect threshold to 1000 (or even a million, if you have greatly decreased the amount of DNA you are searching)
8. Re-read the 'Rules for Gene Annotation' on p. 96 and try some of the other methods suggested on this page, such as:
    a. Try a CLUSTAL DNA-to-DNA search at http://www.ebi.ac.uk/Tools/msa/clustalw2/
    b. Set '3-way multiz' under 'Comparative Genomics' (towards the bottom of the Genome Browser mirror window) to 'full' to look for regions of high conservation between *D. melanogaster*, *D. virilis* and your fosmid species (*D. erecta* or *D. mojavensis*).
    c. Use conservation of intron length in *D. melanogaster* (from the Gene Record Finder coordinates) to gain an idea of where an exon may begin or end in your species.
    d. Look to see if the CDS coordinates (from the Gene Record Finder) of your problem CDS overlap at all with other CDS from other isoforms for the same gene.