

**Genomic Annotation Lab Exercise**  
By Jacob Jipp and Marian Kaehler  
Luther College, Department of Biology  
Genomics Education Partnership 2010

“Genomics is a new and expanding field with an increasing impact on biological research and studies of human health. Genomic approaches can provide new insight to many long-standing biological questions. Instead of studying a single gene, biologists can now study entire genomes, or track genomic changes among related species.

"Metagenomics" is taking this approach one step further to analyze the DNA of whole populations. Genome sequencing is constantly getting cheaper, and the "\$1000 human genome" is within sight, with profound consequences for the practice of medicine ([Pettersson et al., 2009](#)). Full realization of the potential of these new developments requires a broad effort to introduce genomic approaches and bioinformatics tools into the undergraduate curriculum.” (Shaffer et al. 2010)

Specifically, genomic annotation can be divided into two types of annotation. Structural annotation is the process of identifying key genomic elements in a genome. These elements include the location and structure of genes, ORFs and their localization, coding regions, and the location of regulatory motifs. Functional annotation consists of attaching qualitative information to the genome. This can involve the function of the gene, the regulation of the gene, and its involvement in other processes.

Structural annotation of a genome can be a useful tool in comparative genomics. Through the usage of tools such as BLAST (Basic Local Alignment Search Tool), sequence similarity between two genomes can be identified. BLAST is a computer algorithm that enables one to search a database of sequences for similarity to a query sequence. A variety of queries can be used which enables sequence similarity to be identified at the protein and nucleotide level. Based on this sequence similarity, speculations can be made as to the homology of two genes. Evolutionarily speaking, these similarities can be interpreted as divergent evolution from a common ancestor.

While gene predictors have been developed, they are not robust enough to be used as the only source in eukaryotic genomic annotation. There is less information available about eukaryotic genes than prokaryotic genes. Also, these predictors struggle with gene prediction due to splice site recognition and the presence of different isoforms – proteins derived from alternatively spliced transcripts of the same gene in a species.

In this lab you will be asked to annotate a gene from *Drosophila erecta*, a fruitfly species that diverged from *D. melanogaster*, an extensively studied model organism in biology. The *D. melanogaster* genome has been thoroughly sequenced and researched. By utilizing tools such as BLAST and leveraging the large amount of research already performed on *D. melanogaster*, you will be able to identify the gene in *D. melanogaster* that has the greatest sequence similarity to the gene you are annotating. Then, knowing the predicted protein sequence of that gene and proceeding exon-by-exon, you will be able use BLAST to help you identify the exact coordinates of each exon. At the conclusion of this lab you will have structurally annotated a gene for *D. erecta*. Based on the extent of its sequence homology, you may be able to predict the protein’s function as well.

## Annotation Worksheet

Work in pairs. One worksheet per group is to be turned in at the end of the second lab period.

### Part 1: Computer Set-Up

1. Open Internet Explorer/Mozilla Firefox

Four different windows/tabs will be utilized for this lab. Each website will have a specific function throughout the course of this lab. It is important to use the website that is directed for each step.

2. Open the following four websites (either in separate windows or in separate tabs, whichever is most comfortable):

*Throughout this worksheet, these sites will be referred to by the names in quotes.*

- “Gander Genome Browser” = [gander.wustl.edu](http://gander.wustl.edu)  
-Similar to UCSC genome browser. The UCSC genome browser is an up-to-date source for genome sequence data integrated with a large collection of aligned annotations. It has downloadable files and is designed to be easy-to-use.
  
- “Flybase” = [www.flybase.org](http://www.flybase.org)  
-A comprehensive site containing genetic information and biological information about *Drosophila melanogaster* and related species. Flybase is carried out by a consortium of researchers at Harvard University, Indiana University, and Cambridge University in the UK.
  
- “Ensembl” = [www.ensembl.org](http://www.ensembl.org)  
-A joint scientific project between European Bioinformatics Institute and Wellcome Trust Sanger Institute. It was launched in 1999 and allows users to retrieve genetic information about a species.
  
- “[NCBI] BLAST” = [www.ncbi.nlm.nih.gov/blast](http://www.ncbi.nlm.nih.gov/blast)  
-A website designed to compare two different genomic sequences. A variety of inputs (including nucleotide, translated nucleotide, and protein) can be used to search an extensive database of genomes in order to find similar sequences. Sequence similarities are given as an output based upon the level of similarity.

3. Use a few minutes to search the various sites. Try to become familiar with the different attributes of the sites and the differences between them. Also scan each site for information on the different BLAST methods. Complete the charts below.

TABLE 1. Characterization of various genome databases

	<b>Wash U-modified UCSC Genome Browser</b>	<b>Flybase</b>	<b>Ensembl</b>	<b>NCBI</b>
Can sites use amino acid sequence (ie. AYWP), nucleotide sequence (ie. AGTCG), or both?				
How diverse are the sequences that have genomic data represented?				
Does this site offer comparative capabilities?				
Can individual genes be used as an input to search?				

TABLE 2. Types of BLASTing

<u>BLAST</u>	<u>Query</u>	<u>Database Searched</u>
<b>blastn</b>		
<b>blastp</b>		
<b>blastx</b>		
<b>tblastn</b>		
<b>tblastx</b>		

## **Part 2: Initial Fosmid Observation**

1. Using the Gander Genome Browser site, load the particular fosmid from *D. erecta* which you will be analyzing (you will be told the necessary information about the fosmid in lab).

2. If you scroll down you will notice a variety of different settings which allow you to view a variety of different “tracks.” (These are simply different features of the genome and information from different databases) Click “hide all.” Next, change the settings so they match those below:

GEN-SCAN = pack

Repeat Masker = dense

Click Refresh and you should be able to see the gene predictions. The gene predictions are indicated by the colored bars: The thick, solid bars are the exons and the thin lines are the introns.

3. Now that the gene predictions are shown, you should be able to identify how many genes are present in the fosmid.

How many genes are present in the fosmid? \_\_\_\_\_

Are the genes oriented in the same direction? \_\_\_\_\_

How can you tell? \_\_\_\_\_

\_\_\_\_\_

4. If you look at the bottom of the screen, you can see the repeat masker areas. Approximately (a simple estimate is fine) how much of the fosmid sequence appears to be repeats? \_\_\_\_\_

Where in the fosmid are these repeat sequences primarily located? \_\_\_\_\_

\_\_\_\_\_

5. Click on the gene that you have been assigned to annotate. When you click on the gene, you will be able to obtain the predicted protein sequence.

6. Copy the predicted protein sequence (your query sequence) of your particular *D. erecta* gene from the Gander Genome Browser and perform a blastp search against the *D. melanogaster* annotated proteins (AA) on the Flybase website (select the “GenBank protein sequences” from the database drop-down menu). A list of gene hits should come up as the result. A small E-value for gene similarity indicates a strong similarity between the two sequences.

7. Based on the color-coded chart (Note: this chart may not appear; if not, look at the table contains scores for gene similarity) that appears, determine which gene seems to have the same predicted protein sequence as your gene. Select the protein with the highest match score. Scroll down (if you click on the gene in the chart or click on the ‘score’ on the table, it should automatically transfer you to the appropriate result on the page) and examine the alignment.

What is the % of identical matches? \_\_\_\_\_

What is the % of positive matches? \_\_\_\_\_

How many gaps are in the alignment? \_\_\_\_\_

Briefly explain what these three alignment terms (% identical, % positive, gaps) mean:

---

---

What is the name of your gene? \_\_\_\_\_

### **Part 3: Gene Structure Research**

1. For annotation that will be submitted back to GEP, use the gene record finder at "<http://gander.wustl.edu/~wilson/dmelgenerecord/index.html>" (available from the GEP website under the Projects menu.) [Alternatively, if this lab will be used for demonstration purposes only, you can use Ensembl, though this site is somewhat outdated. Scroll the Search drop-down window and select Fruitfly]

2. Type in the name of your gene and hit “Find Record”.

3. Click the FlyBase protein coding gene link for your particular gene

Gene: **mav (FBgn0039914)**

maverick

Location [Chromosome 4: 643,842-646,656](#) reverse strand.

Transcripts There are 3 transcripts in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
mav-RA	<a href="#">FBtr0089103</a>	<a href="#">FBpp0088172</a>	protein_coding
mav-RB	<a href="#">FBtr0100502</a>	<a href="#">FBpp0099940</a>	protein_coding
mav-RC	<a href="#">FBtr0290233</a>	<a href="#">FBpp0288672</a>	protein_coding

4. Click the top “Transcript ID” link on the results page

(The above figure shows the appropriate link to click for the example gene, mav-RA. mav-RB and mav-RC are different isoforms for the same gene.)

Click the Transcript IDs for each of the isoforms (if applicable) and analyze the chart that appears.

Based on the statistics below each chart, what is the approximate percentage of transcript that is translated in each isoform? \_\_\_\_\_

\_\_\_\_\_

5. On the left hand side you will notice a variety of options to analyze the gene. Under the Sequence drop-down, there is a link titled “Protein.” Click this link.

6. The protein sequence for your gene will appear. It will appear as a mixture of blue and black font. Any point when the font changes color represents the beginning of a new exon. For example, if the protein sequence is initially black, then changes to blue, then returns to black, this indicates that there are three exons in the gene. Any residue that appears red represents a codon that is interrupted by the exon splice site. During individual exon annotation the red residue must be included in *both* exons that code for it.

Based on the color coding of exons, how many exons does your gene have? \_\_\_\_\_

At which (if any) exon splice sites are codons interrupted? \_\_\_\_\_

**Discussion Question 1: What are isoforms? Evolutionarily speaking, how do you think they came to be?**

## Part 4: Annotation

**\*\*For each exon you will need to complete the following steps:\*\***

You now have the gene ID and protein sequence from *D. melanogaster*. You must locate and evaluate how “true” the sequence is to the *D. erecta* genome. Essentially, you will need to compare, exon by exon, the closest match of the *D. erecta* DNA to the known exon sequences in *D. melanogaster*.

1. First, highlight ONLY the first exon that is shown in the Ensembl results page, being sure not to include any of the amino acids from different exons (i.e. if you are highlighting a “black” exon, be sure that you do not highlight any amino acids from the neighboring “blue” exons; be sure to include the red amino acid in both exons as it is encoded by both exons).

2. Once the exon has been highlighted, copy this sequence.

3. In the NCBI BLAST tab, under the Basic BLAST heading, click “tblastn.”

What are you comparing in tblastn searches? \_\_\_\_\_

4. Paste the protein sequence for the particular exon into the “query sequence.”

5. Check the box that indicates that you want to “align two or more sequences.”

6. Open the Gander Genome Browser tab and return to the fosmid ID page. **Make sure the arrow at the left of the cosmid is consistent with the direction of transcription for your gene.** Hit the 10X zoom out several times until you are sure you are looking at the entire cosmid sequence. Click the “DNA” link at the top of the page. Next, click “get DNA.” This will give you the entire DNA sequence for your fosmid.

7. Copy this DNA sequence (Clicking Ctrl (or ⌘) +A will highlight the entire region. When the entire region is highlighted, clicking Ctrl (or ⌘) +C will copy the sequence) into the “subject sequence” box in the NCBI BLAST tab. BLAST these two sequences against one another. This BLAST will allow you to determine the location of the particular exon location on the entire fosmid. The reading frame will indicate which direction the exon lies on the fosmid. A -1, -2, or -3 indicates it reads R→L and +1, +2, +3 indicates it reads L→R.

8. Return to the Gander Genome Browser. Identify the indicated beginning position of the protein sequence on the fosmid. On the Genome browser, zoom in to the particular location. Make sure the Base Position option is “Full.”

Directions for zooming: Click on any number on the base coordinate. This will zoom in 3X and center on that location. Continue to click on the coordinates until you feel that you are zoomed in sufficiently. Eventually, you will notice the individual nucleotides. To the left of these is an arrow. Verify that this arrow is pointed in the correct orientation in regards to the direction in which the gene is read. Clicking on an individual nucleotide will toggle on and off the three different reading frames for translation.

Within the first translated exon of a gene, you should notice a green block at the indicated “start” location. This green box represents a start codon for the gene, and the end of the 5’ untranslated region (5’UTR) of the mRNA. On the table located at the end of this lab, record the base pair location at which the start codon is located. (Note: we are only identifying translated exon sequences within the mRNA using this annotation strategy. 5’ and 3’ noncoding sequences are not annotated.)

9. Travel along the sequence in the appropriate direction and make sure you do not see any “red” amino acid boxes, which indicate stop codons. (Note: red boxes in a different row are ok; these are in a different reading frame and do not impact translation.)

**Discussion Question 2: Based on what you see as you scan the exon, what are the potential results of a frame shift mutation? What are the potential consequences of these results?**

10. Now travel to the indicated ending coordinate of the exon. Verify that you have identified the beginning of the intron. The beginning of the intron sequence starts with “GT” (In some rare cases, an intron will begin with “GC”). Record the base pair location where the exon ends.

At the end of the exon, note whether the coding sequence ends within or at the end of a codon. If the exon ends in the middle of the codon, record on the table how many nucleotides in the codon are coded for by the exon. (The reading frame for the next exon



will have to “match” to ensure that the coding sequence is sustained; whatever remaining nucleotides are needed in the codon must be coded by the following exon.)

11. Continue this procedure for each exon in the gene. The reading frame for each exon may not necessarily be the same. This is okay as long as an intact codon is accounted for at the exon-exon junctions. Remember that after the initial exon, each subsequent exon need *not* begin with a start codon. Instead, the end of the preceding intron will be indicated with “AG.” Within the final exon, you should see a red box indicating a stop codon. The stop codon is necessary to stop translation, and marks the beginning of the 3’ untranslated region (3’UTR) of the mRNA.

TABLE 3. ANNOTATION DATA FOR GENE \_\_\_\_\_

<u>Exon Number</u>	<u>Start Coordinate</u>	<u>Stop Coordinate</u>	<u>Bases needed to complete codon</u>
One	_____	_____	_____
Two	_____	_____	_____
Three	_____	_____	_____
Four	_____	_____	_____
Five	_____	_____	_____
Six	_____	_____	_____
Seven	_____	_____	_____
Eight	_____	_____	_____
	_____	_____	_____
	_____	_____	_____
	_____	_____	_____

**Discussion Question 3: How do the exon/intron boundaries correlate with codon reading frames?**

**Discussion Question 4: Following the stop codon, continue to scan downstream. What do you notice? Explain the evolutionary advantage of this phenomenon.**