Lab Week 10 - Answer Key to *A Sample Annotation Problem*

For this lab, we will focus on the gene from the *yellow* family…*So, how do you decide which member of the yellow gene family is the best one to study?* To make this decision, answer Questions 1-4 below.

Question 1
The different colors of the alignments blocks for the different *yellow* family genes provide one hint. Click on the 'D mel Proteins' link in the Genes and Gene Prediction Track area of the UCSC Genome Browser to find a chart of what the different colors represent.

Which gene has the highest (bit) scores for its alignment blocks? *yellow-h* (which has red alignment blocks with scores >500)

[Note that the lengths of the three alignment blocks in the Genome Browser window are also longer for this gene than any of the other *yellow* family genes.]

Question 2
Type the name of the gene (case sensitive and without the -PA) from Question 1 into the http://flybase.org/ 'Jump to Gene' box in the right corner of the menu bar. Then, click 'Go.'

Which chromosome is this gene on in D. melanogaster? 4 (the dot chromosome!)

Why does the fact that the D. melanogaster gene is on this particular chromosome (and not on a different one) strengthen the case for this gene being the best candidate for the D. biarmipes gene? [Hint: none of the other yellow family genes are on this chromosome in D. melanogaster.]

The fosmid we are analyzing for this exercise contains 4th chromosome DNA from *D. biarmipes.* A figure in the 2006 *Genome Biology* paper (see e-reserve list) shows that 27 of 28 genes are shared between *D. melanogaster* and *D. virilis* in one region of the 4th (dot) chromosome. This shows that there has been little movement of genes off of the 4th chromosome since the two species diverged. Since *D. biarmipes* is more closely-related to *D. melanogaster* than is *D. virilis* (the former two are in the same species group whereas *D. virilis* is in a different subgenus), we would expect similar (or higher) levels of gene conservation on the 4th chromosome of *D. biarmipes* and *D. melanogaster.* Thus, the fact that the *yellow-h* gene is on the 4th chromosome in *D. melanogaster* is additional evidence that this region of the *D. biarmipes* dot chromosome contains the *yellow-h* ortholog. In fact, Quick Searches in FlyBase show that none of the other members of the *yellow* gene family are on the 4th chromosome.

Click the ⊞ box next to Gene Model & Products in the lower left of the FlyBase window to view a zoomed-in diagram of the exons and coding sequences (CDS, in purple) for the yellow-h gene in D. melanogaster. How many CDS are there? 3

How does this compare to the number of alignment blocks in the BLASTX track for the yellow-h gene in the Genome Browser window of the D. biarmipes fosmid (Fig. 3.3b)? They are the same!

The Gene Predictor tracks

Although you may now be fairly confident that *yellow-h* is the best candidate gene for this region of the *D. biarmipes* fosmid, you should strengthen this conclusion in a couple different ways. First, return to the 'D. mel Proteins Track Settings' page and set the 'Filter score range: min:' to 500 and click 'submit.' This removes all the other *yellow* family genes from the Browser window. Now, examine the five other Gene Prediction tracks in the Genome Browser window in more detail: select 'pack' under the names of each of the gene predictors (Genscan, Geneid, N-Scan, etc.) and click one of the 'refresh' buttons. Note that the 'hits' for each gene predictor are now numbered in the Browser window; if there is more than one separate feature identified by a particular gene predictor, they are numbered in the Genome Browser window in the physical order, from left to right, that they appear along the DNA sequence of the fosmid. Note the differences between the results of the various gene predictors!

Question 3

Which gene predictor do you think matches the best to the BLASTX output track for the yellow-h gene and why?

Geneid, N-SCAN and Augustus match up the best with respect to the number (3) and length of blocks they share with the BLASTX alignment for the *yellow-h* gene.

How does the alignment for this gene predictor differ from the BLASTX output for the entire fosmid?

All of the gene predictors fuse the *yellow-h* gene feature to the CG31999 feature. Only the SNAP predictor splits the fosmid into two features (although the *yellow-h* feature is still fused to some of the CG31999 feature in the SNAP track).

You will now use sequence information from one of the gene predictor tracks to confirm that the *yellow-h* gene is the best candidate gene in this region of the *D. biarmipes* genome, by doing a BLAST search. It is quicker to BLAST using FlyBase instead of NCBI, since you are only going to search the known proteins database from *Drosophila melanogaster*, as follows:

1. Obtain the predicted amino acid sequence of gid_contig14_1in the Geneid Gene Predictions track by clicking directly on any of the predicted exons for this contig (bright red in color).
2. Click on the 'Predicted Protein' sequence link in the next window and copy the amino acid sequence.
3. Open the search engine at http://flybase.org/blast and select the 'Annotated Proteins (AA)' Database and the 'blastp: AA -> AA' Program.
4. Paste the copied Predicted Protein sequence from the Genome Browser into the Sequence box and click the BLAST button.

Once the BLAST search is complete, a new web page appears with the BLAST Report. The BLAST output begins with a description of the version of BLAST used (blastp 2.2.18 in this

case), a literature reference and some details on the query sequence used in the search. The rest of the default BLAST report consists of three main sections: a graphic summary, a list of significant BLAST hits, and the corresponding amino acid alignments. We will go through each of these sections in order to help you interpret the blastp output.

a. Graphic Summary

The Graphic Summary (Figure 3.5) shows alignments (as colored bars) of the top 25 database matches with significant sequence homology to our Query sequence (line under the colored Score Key). Once again, the color of the bars corresponds to the score (S) of the alignment, with red representing the highest alignment scores. The default listing for the hits is by score, since it is generally the case that the higher the alignment score, the more significant the hit. If you click on a colored alignment bar, you will jump to the actual DNA alignment associated with that BLAST hit (see c. below). Note that since the Geneid track has a single (fused) alignment track including both features in the *D. biarmipes* fosmid, the CG31999 gene is included at the top of the list of hits. You can ignore the CG-31999 hit for this analysis.

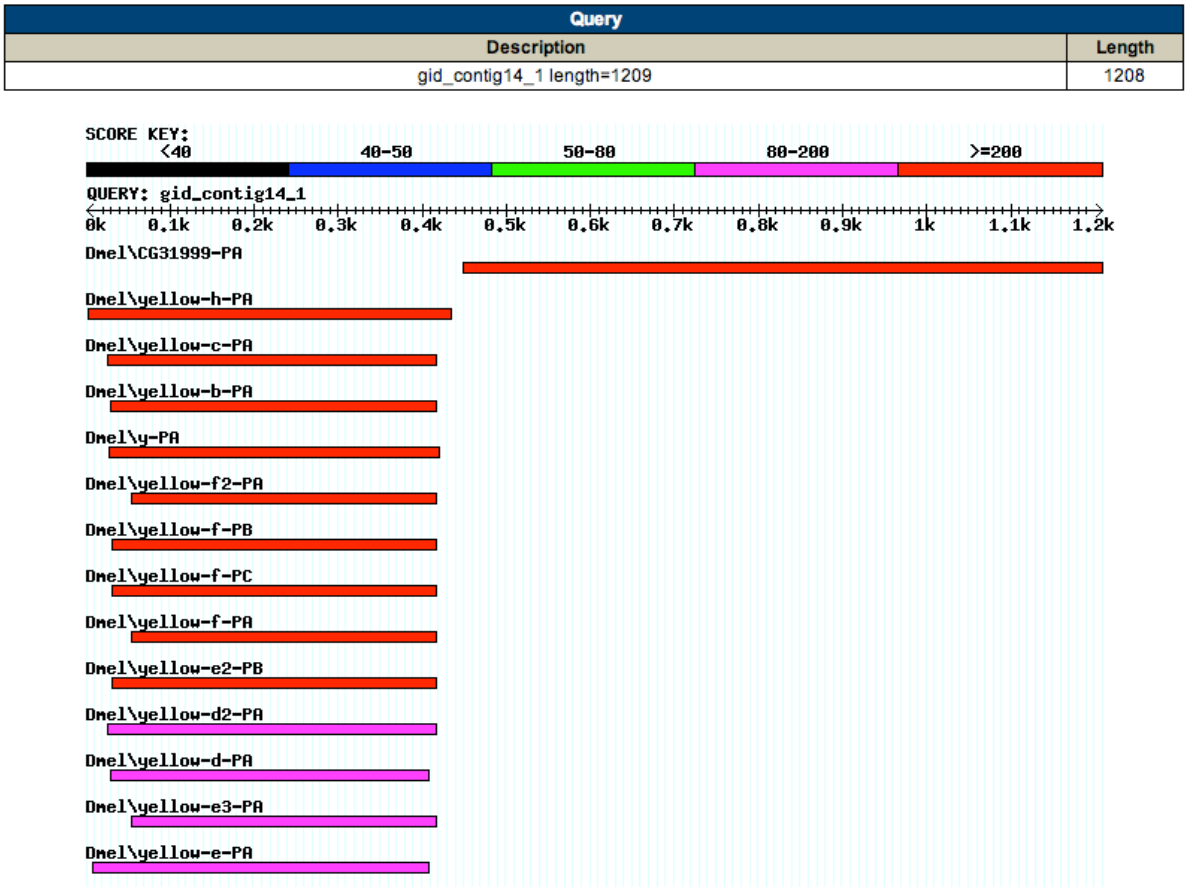

Figure 3.5. A graphical overview of the first 14 blastp *D. melanogaster* hits to the *D. biarmipes* Query sequence

b. List of Significant BLAST Hits

Scrolling further down the BLAST Report window, you will find a BLAST Hit Summary table with the top 25 sequences in the *D. melanogaster* protein database, sorted according to the Score and E-value (Figure 3.6).

BLAST Hit Summary				
☑	Description	Species	Score	E value
☑	CG31999-PA	Dmel	927.161	0
☑	yellow-h-PA	Dmel	758.444	0
☑	yellow-c-PA	Dmel	299.286	1.52705e-80
☑	yellow-b-PA	Dmel	287.73	4.59782e-77
☑	y-PA	Dmel	286.189	1.06792e-76
☑	yellow-f2-PA	Dmel	238.039	3.81151e-62
☑	yellow-f-PB	Dmel	228.794	2.09205e-59
☑	yellow-f-PC	Dmel	228.024	3.94429e-59
☑	yellow-f-PA	Dmel	227.639	5.37086e-59
☑	yellow-e2-PB	Dmel	202.216	1.97794e-51
☑	yellow-d2-PA	Dmel	196.823	8.37973e-50
☑	yellow-d-PA	Dmel	186.037	1.47916e-46
☑	yellow-e3-PA	Dmel	182.57	1.69091e-45
☑	yellow-e-PA	Dmel	155.221	3.22216e-37

Figure 3.6. Scores and E-values for the first 14 blastp *D. melanogaster* hits to the *D. biarmipes* Query sequence.

Question 4

How does the Score and E value of the D. melanogaster yellow-h gene alignment to the D. biarmipes fosmid compare to the values for the other members of the yellow gene family?

The *yellow-h* alignment to the translated fosmid produces a score of 758.4 and an E value of 0 (which indicates that there is essentially no chance of getting such a high score by randomly aligning any other two sequences). The scores of the other *yellow* genes range from 155 to 299. Although their E values are still low (10^{-37} to 10^{-80}), they are not zero as with the *yellow-h* gene.

c. List of Alignments

Following the table of BLAST hits is a section showing all of the alignment blocks for each BLAST hit. To jump right to the alignment for the protein encoded by the *yellow-h* gene, click on its Score in the BLAST Hit Summary table. The sequence alignment (Figure 3.7) shows you how well the Query sequence matches with the Subject sequence in the database. Since you will rely heavily on sequence alignments in your annotation efforts, examine the alignment to the *Drosophila melanogaster* yellow-h amino acid sequence more closely.

Different BLAST hits (Subject sequences in the database) are separated by definition lines that begin with a '>' character, which is followed by lots of database information for that particular Subject sequence, including its FlyBase ID number, the name (or symbol, if unnamed) of the encoded protein and the length (in amino acids) of the Subject sequence. Each alignment block demarcates a local region of similarity between the Query sequence and the Subject sequence identified by the definition lines.

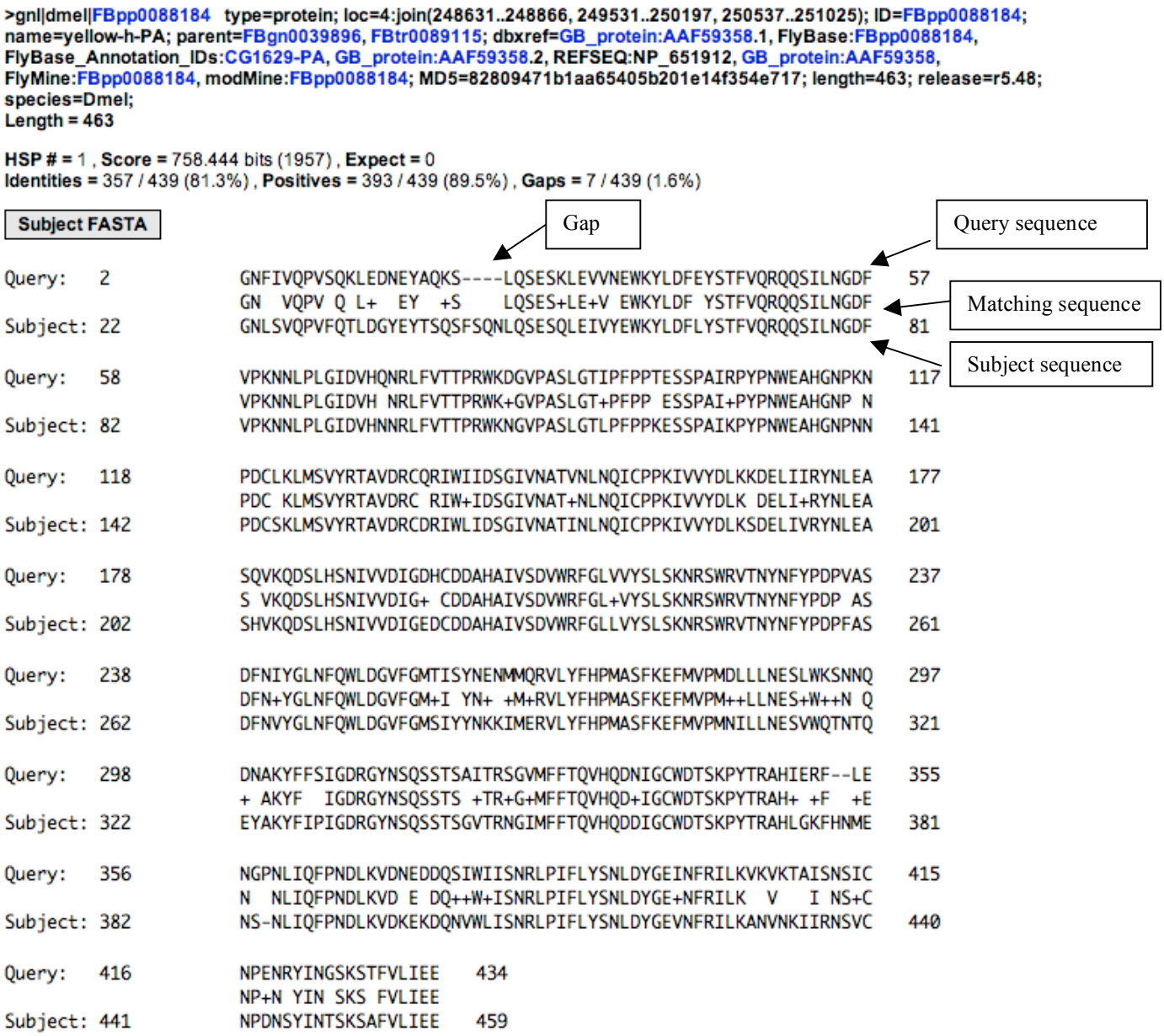

Figure 3.7. Alignment of the Predicted Protein sequence from *D. biarmipes* (Query) to the *yellow-h* protein product from *D. melanogaster* (Subject).

What about the alignments themselves? Each alignment block begins with a summary that includes the BLAST Score and Expect value (the statistical significance of the alignment), sequence **Identities** (number of identical amino acids between the Query and the Subject sequence), sequence **Positives** (amino acids that are either [1] identical or [2] mismatched but have side chains with similar physical or chemical properties, such as leucine and isoleucine) and the number of **Gaps** in the alignment. Gaps are places where the BLAST program inserts spaces to maximize the score in other parts of the alignment; gaps generally represent places where one gene sequence has diverged away from the other over evolutionary time by adding or deleting codons.

The alignment itself consists of three lines: the top Query sequence, a middle Matching sequence, and the bottom Subject sequence (Figure 3.7). The Matching sequence consists of a combination of letters (for identical amino acids in the two sequences), + characters, and empty spaces. The '+' character shows the location of the mismatched **Positives** (see above). Places in the middle Matching line that are blank are locations where both species have amino acids in those locations, but the side chains of the mismatched amino acids are chemically or physically dissimilar. Note that these dissimilar mismatches still contribute a positive value to the overall alignment score. Finally, the '-' characters in either the Query or the Subject sequence denote gaps in the alignment.

Now, click on the alignment bars or scores for some of the other yellow family genes.

Note how much better and longer the alignment is with the *yellow-h* gene than with the other *yellow* family members. Thus, we will proceed with the assumption that this region of the fosmid does indeed contain the *D. biarmipes* ortholog for the *yellow-h* gene of *Drosophila melanogaster* and not any of the other *yellow* family genes identified by the BLAST searches. ☺

Coding DNA Sequence (CDS) by CDS Searches

While it is true that the Geneid prediction shows significant matches to the *Drosophila melanogaster* yellow-h protein *(*Figures 3.5-3.7), this does not mean that the prediction is totally "correct". In fact, published accuracy rates for most *ab initio* gene prediction algorithms are in the range of 20-30%. Common errors generated by Geneid and other *ab initio* gene prediction algorithms are skipped exons and errors involving the ends of the gene (split genes, fused two genes, as in this case, etc.). It is therefore very possible that the Geneid prediction is actually wrong (i.e. not perfect). Without detailed analysis of the alignment between contig14_1 and the *yellow-h* gene, we have no way of knowing. For now, all we know is that BLAST has aligned at least some of contig14_1 with at least some of the *yellow-h* gene from *Drosophila melanogaster*, which has contributed to a good E value for the sum of all the alignment blocks to the fosmid.

The next step is to use BLAST searches to find the best matches to the individual *D. melanogaster yellow-h* coding sequences (CDS), as these matches will be the best evidence we can gather as to the structure of this gene in *D. biarmipes.* In order to do CDS-by-CDS searches, we need the sequence of each CDS. This information is most easily obtained from the GEP's Gene Record Finder. Go to the 'Projects' drop down menu on the http://gep.wustl.edu website, then select 'Annotation Resources' to get to the Gene Record Finder. Enter 'yellow-h' into the search box (case sensitive) to obtain the information on this gene. As you type, note that the name of the gene appears in a drop-down bar below the search box, with the number of mRNA isoforms (only 1) and the total number of exons (3) and CDS (3) for this gene. Now, click the "Find Record" button. [*Note*: a pdf of The Gene Record Finder User Guide is posted on the course Moodle site in the Lab 10 topic box.]

There are three regions of information in the next window that appears: Gene Details, mRNA Details and Transcript Details/Polypeptide Details. Note that the name of the isoform has changed from yellow-h-PA (in the previous database windows) to yellow-h-RA. *The 'R' in the Gene Record Finder has simply been substituted for the 'P' in these other databases, and you can ignore this difference.* The term CDS refers to DNA sequences that are transcribed, present in a mature mRNA molecule *and* code for amino acids. In most cases, the term CDS can be used

synonymously with 'exon.' The exceptions are those exons that include 5'- or 3'-untranslated regions, which do not code for amino acids. Also, due to alternative splicing, not all exons of a gene may appear in a particular mature mRNA. Hence, the windows for Transcript Details and Polypeptide Details reflect these situations and are not the same, with the numbered CDS referring to the specific DNA sequences that are translated in a particular mRNA molecule.

If you click on a particular row of the Polypeptide Details table, the amino acid sequence encoded by the specific CDS appears in a new pop-up window (Figure 3.8). *Make sure you record the number (under Length) of amino acid residues in the Drosophila melanogaster isoform encoded by each CDS.*

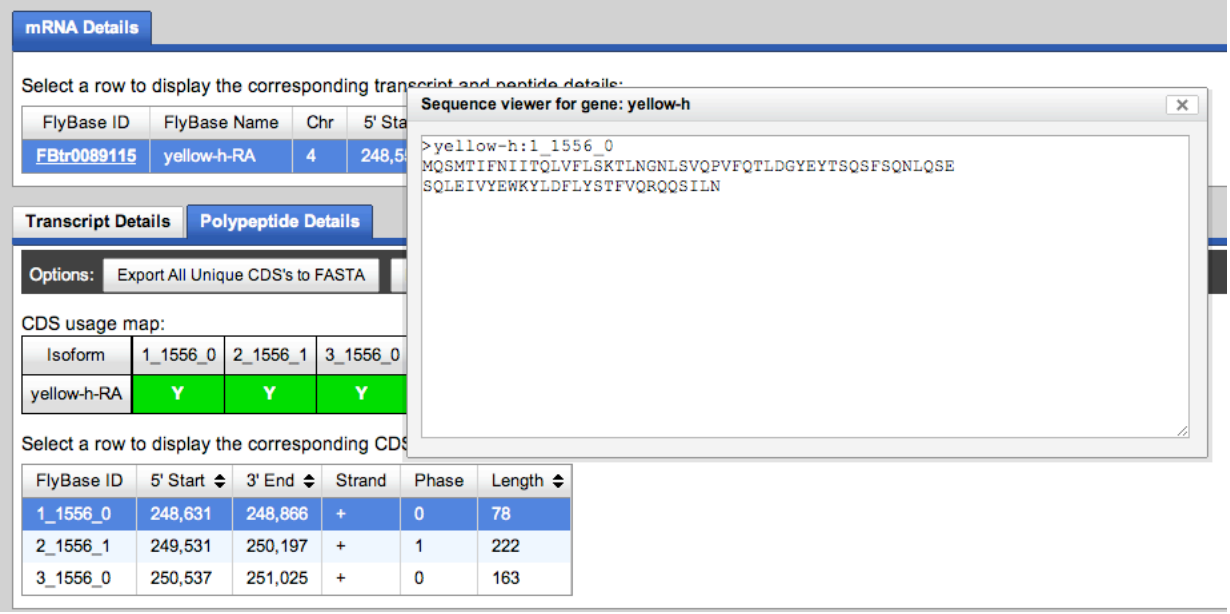

Figure 3.8. Amino acid sequence of CDS #1_1556_0 (the first CDS) for the *yellow-h* gene from *Drosophila melanogaster*.

To map each putative CDS onto the fosmid DNA from *D. biarmipes*, you will do searches that use the BLAST algorithm to compare two sequences to each other (bl2seq). In this case, you will do a blastx search to compare the entire translated *D. biarmipes* fosmid to the amino acid sequence of each *D. melanogaster* CDS. Begin by comparing the protein sequence of the first coding exon of *yellow-h* (CDS #1_1556_0) to the entire fosmid DNA sequence, as follows:

1. Go to the NCBI BLAST search site (http://blast.ncbi.nlm.nih.gov/Blast.cgi) and click on 'blastx' in the Basic BLAST section.
2. On the next page that appears, click the 'Align two or more sequences' box, which adds an "Enter Subject Sequence" box to the BLAST window.
3. Return to the UCSC Genome Browser page for *D. biarmipes* and obtain the contig14 fosmid DNA sequence by clicking 'DNA' in the upper menu bar (make sure contig14:1-45,000 appears in the position/search box). Click on the 'get DNA' button in the next window.

4. Copy and paste the fosmid DNA sequence (45,000 bp long) into the 'Enter Query Sequence' box of the Blast window (it is okay to include the file header when you copy and paste the DNA sequence).
5. Copy the amino acid sequence of CDS #1_1556_0 from the pop-up window of the 'Polypeptide Details' section of the Gene Record Finder window, and paste the CDS sequence into the 'Enter Subject Sequence' box.
6. Check the 'Show results in a new window' box next the BLAST button (which will enable you to reuse the blastx search page for each of the next two CDS searches).
7. Click on the 'Algorithm parameters' link, and be sure that the 'Low complexity regions' and two mask boxes are UN-checked.
8. You are now ready to click the BLAST button! It will take a few seconds for a new window with the blast alignment(s) to appear.

The organization of NCBI's BLAST Report page is similar to FlyBase's, with a Graphic Summary, followed by a table of significant hits, and then one or more specific alignments of the first CDS (Sbjct Sequence) to the translated fosmid sequence (Query Sequence). Note that in this case, only the first alignment produces a statistically-significant E value; the five other alignments below this one are from other regions of the fosmid and are so weak that they can be ignored. ***Copy the best (first) alignment and paste it into the appropriate section of the Annotation Report*** *(fillable Word file on Moodle course site in Lab Week 10 topic box). When copying alignments, be sure to include the Score, etc. header information and shrink the margins and/or font to keep the sequences in alignment across the width of the page)*.

The alignment should look as below (Figure 3.9):

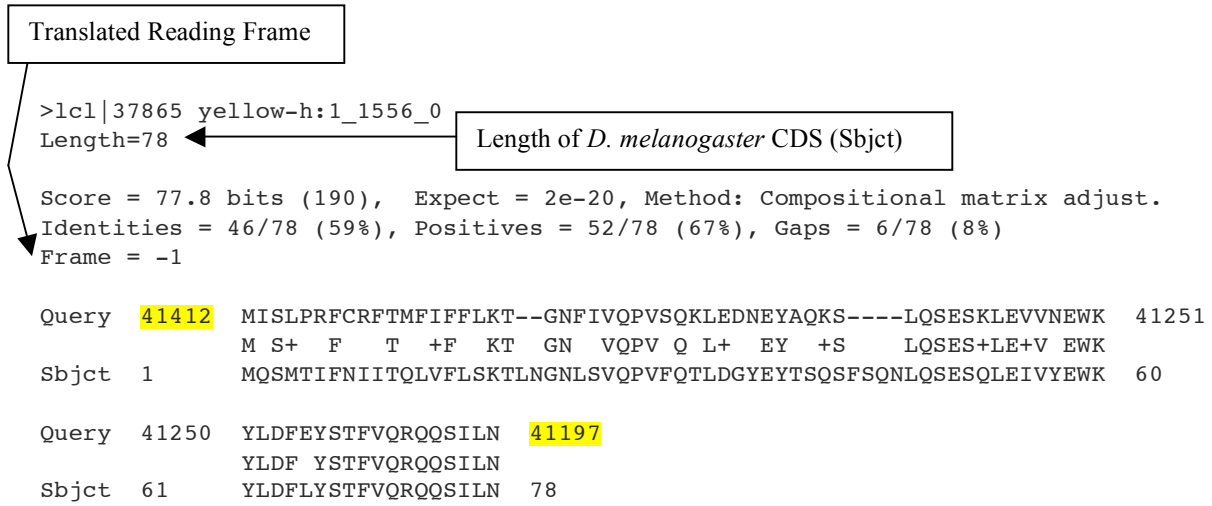

Figure 3.9. Alignment of the first CDS from the *D. melanogaster yellow-h* gene (Sbjct) to the translated *D. biarmipes* fosmid (Query).

One important point to understand about blastx alignments is that the numbering of the Query sequence reflects the nucleotide position in the *D. biarmipes* fosmid, while the numbers of the Sbjct sequence indicate amino acid positions in the CDS from *D. melanogaster*. By using the whole fosmid in our search, we can read the DNA base coordinates of the orthologous *D.*

biarmipes CDS directly from the alignment. Thus, the alignment for this first CDS begins at base 41,412 of the fosmid with the codon for methionine (remember your Basic Biology rules!) and ends at base 41,197. Note that the beginning of the coding region of this gene is not especially well conserved between *D. melanogaster* and *D. biarmipes*. It is also important to check whether or not the entire 78 amino acids of the *D. melanogaster* CDS (Sbjct line) aligns to this region of the *D. biarmipes* fosmid (that's a yes in this case!).

blastx translates the input fosmid sequence in all six reading frames (Figure 2.1), and compares each translated sequence with the amino acid sequence of the CDS. Thus, there is an additional field 'Frame' right above the start of the alignment that indicates which frame was translated to produce the specific alignment (Figure 3.9). The frame can either be + or – and it corresponds to the relative orientation of the fosmid sequence compared to the CDS (e.g., which DNA strand contains the CDS information). For each orientation, the frame has a value that ranges from 1 to 3, which reflects the base at which the reading frame began relative to the 5' end of the coding DNA strand. Together, the relative orientation and the frame represent all six reading frames.

Question 5
What reading frame was the fosmid translated in to best align with the first CDS? -1

What can you conclude from the sign of a reading frame (+ or -) about the relative magnitude of the base pair coordinates for the beginning versus the end of a mapped alignment block in the fosmid DNA?

Since it is a minus (-) frame, the "bottom" strand of the fosmid (5' to 3' going from right to left) has the information for the CDS of the *yellow-h* gene. The numbering of both strands of the DNA molecule starts at the far left in the Genome Browser with base pair #1 and increases as one proceeds to the right (to base pair #45,000 at the far right of the Browser window). Thus, as one moves in the 5' to 3' direction along the *yellow-h* gene, the numbers should decrease (since one is heading from right to left along the bottom strand of the DNA molecule).

Do the base coordinates of the beginning and end, respectively, of the alignment to the first CDS support your conclusion? Yes *Explain.* The numbers do in fact decrease, from bp #41,412 at the beginning of the CDS to bp #41,197 at the end of the CDS.

Question 6.
Repeat the same blastx searches with the second and third CDS (#2_1556_1 and #3_1556_0); copy and paste the best alignments for each CDS into the Annotation Report. Highlight the DNA base coordinates of the beginning and end of each alignment. Also highlight the frame number that was translated to generate the amino acid sequence for the alignment. See Annotation Report Key

Annotating CDS Boundaries

Since the blastx alignments show conserved amino acid residues between *D. melanogaster* and *D. biarmipes*, extra nucleotides in the exon after the last or before the first complete codon of that exon may not be apparent in these alignments (if, for instance, a triplet codon has been split by an exon-intron boundary). Thus, to carefully annotate a protein-coding gene, we must find the exact mRNA splice positions that would create a processed mRNA that links these exons together to create a continuous coding block (e.g., an open reading frame).

Use the UCSC Genome Browser to navigate to the beginning of the first CDS and confirm the coordinates of the start codon for the *yellow-h* gene in the *D. biarmipes* fosmid. From the alignment in Figure 3.9, we believe the start codon begins at base 41,412. To jump directly to this region, enter the coordinates "contig14:41,400-41,450" in the 'position/search' box, then click 'jump.' Next, make sure 'full' is selected under the 'Base Position' link in the 'Mapping and Sequencing Tracks' area and 'pack' is selected under 'D. mel Proteins' in the box below that. Don't forget to click a 'refresh' button.

Since the first alignment block was in Frame -1, it is easier analyze the fosmid by reversing the DNA sequence of the fosmid (to display the three frames going from left to right in this orientation). To do so, click the 'reverse' button under the white window of the Browser. The new window is shown in Figure 3.10 (you may need to zoom in a bit to clearly see the nucleotide bases). Note that the numbering of the nucleotides in the fosmid now decreases as you read from left to right.

Three reading frames for the - strand: 1st (top), 2nd (middle) and 3rd (bottom) row.
Each gray box represents the amino acid encoded by the above
three nucleotides in the DNA sequence.

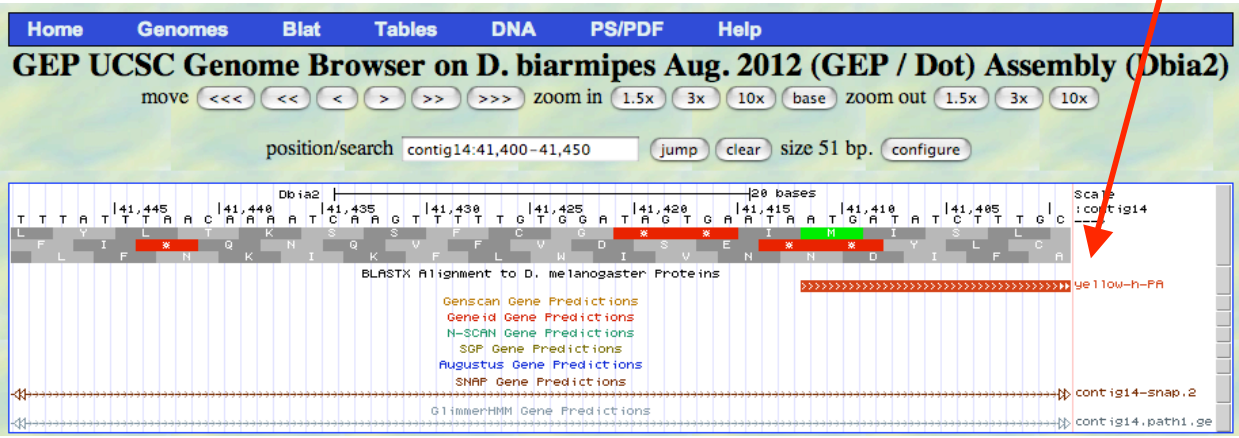

Figure 3.10. Close-up look at the region around the beginning of the first CDS alignment

The above Genome Browser window confirms that the first CDS of the *D. biarmipes yellow-h* gene begins at bp 41,412 of the fosmid, as a codon for methionine (green start codon) in the first row of amino acids (translation Frame -1) begins at this position. The BLASTX alignment block to the first CDS of the *yellow-h* gene also begins at this position and extends towards the right side of the window. Note that NONE of the seven gene predictors show any evidence for an exon in this region of the fosmid!

Use a similar method as above to now jump to the predicted end of the first CDS (at bp 41,197). To precisely map the end of this CDS, you must examine this region for potential intron donor splice sites (see Fig. 1.3), since you are at the beginning of the first intron (Figure 3.11). The blastx alignment in Figure 3.9 indicates that you should look for the amino acid sequence 'SILN' in the first row of gray boxes (reading Frame -1) to find the end of the first CDS. Note that six of the seven Gene Predictors detected an exon beginning somewhere in the middle of the first BLASTX alignment block for the *yellow-h* gene, and they are now very helpful in directing

your eye to the possible end of the exon. Another helpful set of data to add to the Browser window at this time are the Predicted Splice Sites (see explanation of this track in the **Computational evidence** section of the Student Outline). Add these data by selecting 'dense' below the corresponding link in the Gene and Gene Prediction Tracks control area. Color-coded bars then appear under putative intron splice site consensus sequences (red arrow in Figure 3.11).

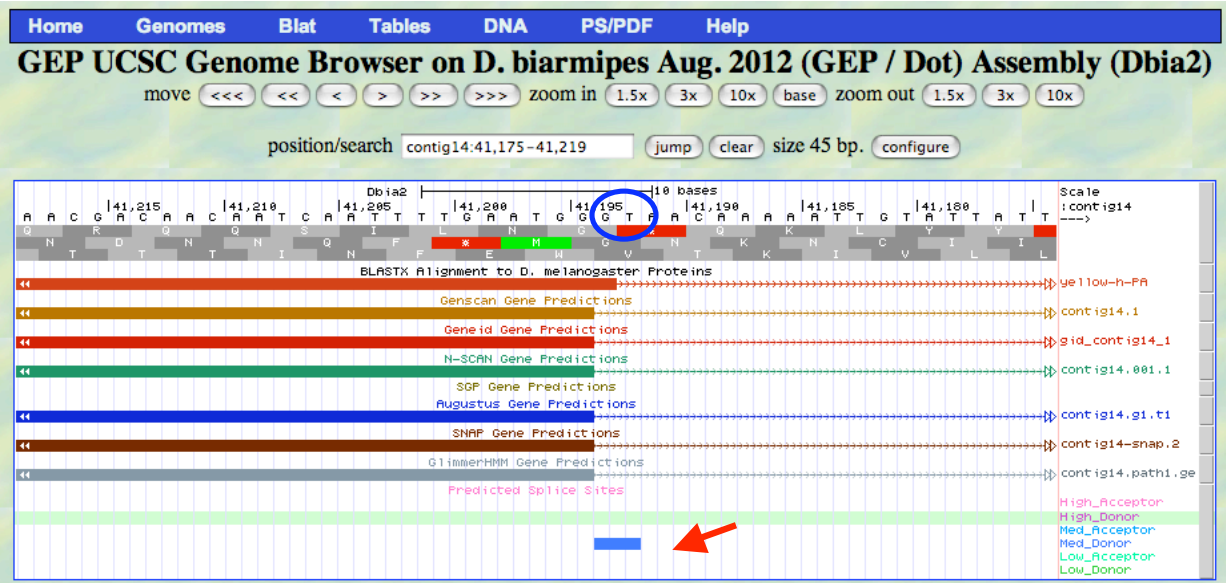

Figure 3.11. Close-up look at the region around the end of the first CDS/exon of the *yellow-h* gene.

The complementary bars for the gene predictors and the predicted intron donor splice site readily identify the 'GT' at base position 41,194 and 41,193 (circled in blue above) as the first two bases of the first intron. Note that this is NOT the same position as called by the BLASTX alignment (which extends the exon by one base to include the G). This is a great example for why the blastx alignments of the individual CDS you did earlier do not necessarily give you the most accurate coordinates for the end (or beginning) of exons/CDS.

An additional complexity that arises from the above coordinates is that by ending at base position 41,195 (e.g., one base BEFORE the intron donor GT site), the first exon ends in the middle of a codon. The last complete codon in the first exon is for asparagine (N in the first row of gray boxes). We can see by careful inspection that a cleavage after base 41,195 will leave two bases (GG) between the end of the last complete codon (AAT) in frame -1 and the end of the exon. We use the term "phase" to describe these remaining bases; in this case, the end of the first exon is said to be in phase 2 because two bases are present after the last complete codon and before the exon ends. To make a complete mRNA, we must find an acceptor site at the other end of this intron, such that these two bases at the end of the first exon will join with one other base at the beginning of the next exon to make a complete three-base codon (Figure 3.12). For now, we simply note that the best donor site in this region is in phase 2.

Add the base pair coordinate and phase of the mapped 3'end of the first CDS to the alignment information from Question 6.

See Annotation Report Key

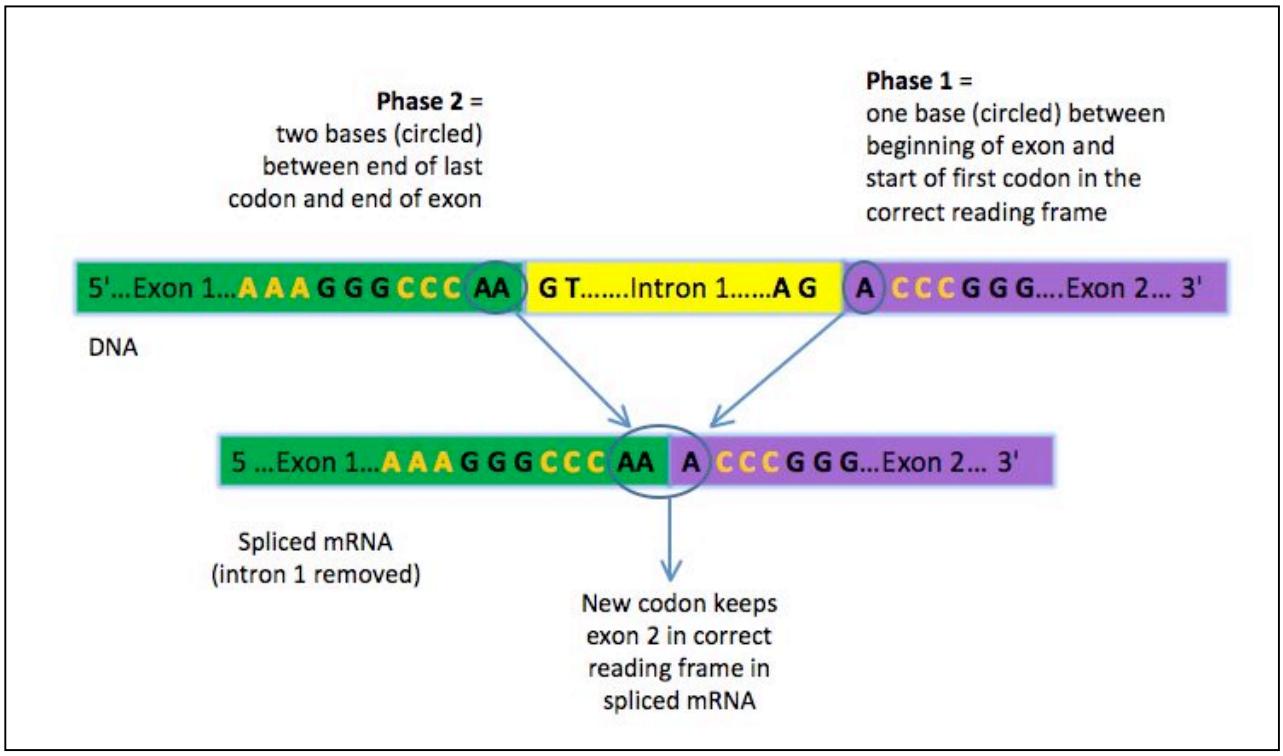

Figure 3.12. Example of compatible phases at the end and beginning of two adjoining exons

When you are deciding where to splice adjacent exons together, the acceptor site that you select for an exon must have a phase that is compatible with the phase of the donor site for the previous exon, e.g., phase 0 acceptor with phase 0 donor, phase 2 acceptor with phase 1 donor, or phase 1 acceptor with phase 2 donor (as in Figure 3.12). If the acceptor site has a phase that is incompatible with the donor, the extra bases will create a frame shift. The downstream CDS will then be translated in an incorrect reading frame, creating an abnormal polypeptide.

Use the Genome Browser to navigate to the region where you suspect the second exon begins, based on the alignment you found for Question 6. Remember that intron acceptor sites have the invariant sequence 'AG' just before the first base in the following exon.

Question 7.
There are two potential AG intron acceptor sites right next to each other near the beginning of exon 2. What are the base pair coordinates of the AG dinucleotide you think is the best one to use for your gene model and what is the evidence in support of the one you chose?

The first one at bp #33,591-33,590.

There is a high-acceptor, predicted splice site under this dinucleotide (pink bar in Genome Browser window) and all seven gene predictors start the second exon at bp #33,591. [Note that the BLASTX alignment block is once again off by one base pair at the beginning of exon 2.] Also, since the phase at the end of exon 1 is phase 2, the phase to begin exon 2 must be phase 1 (to create a triplet codon after the first intron is spliced out). The CDS blastx alignment from Question 6 indicates that the reading frame of the second exon is -1 and that the first few amino

acids of the CDS are DFVP. If the exon begins at 33,591, there will be one base (A) between the beginning of the exon and the codons/reading frame for the DFVP amino acids. So, choosing the first AG ensures that the correct reading frame is maintained after the first intron is spliced out AND conserves the overall length of the polypeptide.

If the second AG (at 33,589-33,588) was used as the splice acceptor site, the phase at the beginning of exon 2 would be phase 2. A phase 2 (from exon 1) + phase 2 situation would create a quadruplet 'codon' after the intron is spliced out, causing a frame shift during translation of the spliced mRNA and a dramatic change in the amino acid sequence downstream of the frame shift.

Record the base coordinate and phase for the beginning of the second exon as deduced from the above analysis. See Annotation Report Key

Question 8

*Use the results of the alignments of the second and third exons from Question 6 to precisely map the 3' end of the second exon and the beginning and end of the third exon. Record the base pair coordinates for the beginning and end of all three exons in the table below. You should also record the base pair coordinates of the stop codon (red block with a * in the middle), which should immediately follow the end of exon 3.*

Table 1. Coordinates of *yellow-h* CDS in contig14 of *Drosophila biarmipes*

	Base Pair Coordinates
CDS 1	41,412-41,195
CDS 2	33,589-32,923
CDS 3	32,031–31,549
Stop codon	31,548-31,546

Checking your Gene Model with The Gene Model Checker

Overview

The next step in the process of making a gene model is to check whether or not the above coordinates for the three *D. biarmipes yellow-h* CDS code for a full-length polypeptide chain when spliced together. You will use the GEP's Gene Model Checker to do so. Detailed instructions for using this software can be found in The Gene Model Checker User Guide, a pdf of which is posted on the course Moodle site.

Using the Gene Model Checker

To confirm the accuracy of a gene model using the Gene Model Checker, proceed as follows:

1. Select 'Gene Model Checker' from the Projects -> Annotation Resources drop-down menu at http://gep.wustl.edu/.

2. Download onto the desktop a copy of the fasta sequence file of the entire fosmid (*D. biarmipes* dot, contig14) from the course Moodle site (in the Lab Week 10 topic box).
3. Upload this sequence file into the first box on the left of the Gene Model Checker window.
4. Type in the name of the *D. melanogaster* ortholog (yellow-h-PA, case-sensitive) in the second box. Watch for this gene to appear in the drop-down menu box as you type.
5. Enter the base number of the beginning and end of each coding sequence, in the order they appear in the table above, in the following format: # - #, # - #, # - #
6. *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.
7. You did not annotate the untranslated regions, so click the circle in from of 'No.'
8. The information for this gene is on the (-) strand, so click the circle in front of 'Minus' following 'Orientation of Gene Relative to Query Sequence.'
9. We think all the gene's coding sequences are accounted for, so click the circle in front of 'Complete' in the next row.
10. Use the drop-down menu to select the Project Group (D. biarmipes Dot) and type in the Project Name (contig14) for this fosmid.
11. *Red boxes will appear around any entries that the Gene Model Checker deems incorrectly entered. Fix these before going on.*
12. Click on the 'Verify Gene Model' box at the bottom of the window.
13. A summary of how your model did now appears on right side of the Gene Model Checker window. Did your gene model pass on all counts?
14. 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.
15. 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.
16. *If you passed the Gene Model Checker on the first try, intentionally change some of the coordinates to see what a failure would look like!*
17. Re-check your work and the typing of the coordinates until you pass *ALL* parts of the Gene Model Checker. Keep the Gene Model Checker window open, as you will come back to it often while completing the Annotation Report.

Complete the Annotation Report by following the directions and answering the questions in the report. Upload the completed Annotation Report (one per pair) to the course Moodle site by NO LATER than the start of next week's lab.

Acknowledgements

The GEP is supported by Howard Hughes Medical Institute grant #52005780 to Sarah C.R. Elgin at Washington University in St. Louis, Missouri.

Literature Cited

Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B. and G. Rätsch. 2008. Tutorial: Support vector machines and kernels for computational biology. http://svmcompbio.tuebingen.mpg.de/splicing.html

Fleischmann, R., Adams, M., White, O., Clayton, R., Kirkness, E., Kerlavage, A., Bult, C., Tomb, J., Dougherty, B. and J. Merrick. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae Rd*. *Science* 269: 496–512

Genomics Education Partnership. 2012. http://gep.wustl.edu/

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., H. Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H. and S. G. Oliver. 1996. Life with 6000 genes. *Science* 274: 546, 563–567

Guo, M., Lo, P. C. and S. M. Mount. 1993. Species-specific signals for the splicing of a short *Drosophila* intron *in vitro*. *Molecular and Cellular Biology* 13: 1104-1118.

McManus, C. J., Duff, M. O., Eipper-Mains, J. and B. R. Graveley. 2010. Global analysis of trans-splicing in *Drosophila*. *Proceedings of the National Academy of Sciences USA* 107: 12,975-12,979.

Painter, T. S. 1934. Salivary chromosomes and the attack on the gene. *Journal of Heredity* 25: 465-476.

Talerico, M. and S. M. Berget. 1994. Intron definition in splicing of small *Drosophila* introns. *Molecular and Cellular Biology* 14: 3434-3445.

Walter, C. and M. Wilkerson. 2006. *Annotation for Amateurs* website, http://www.plantgdb.org/tutorial/annotatemodule/index.html