Annotation Practice Activity

[Based on materials from the GEP Summer 2010 Workshop]

Special thanks to Chris Shaffer for document review

Parts H-L

Introduction:

The typical structure of a eurkaryotic gene consists of a promoter region and an open reading frame (ORF). Features of an ORF are: (1)the presence of a start codon, AUG; (2) a sequence of codons that results in a series of amino acid sequences in a putative polypeptide, and (3) a termination codon (UAG, UAA, UGA). The genomic sequence of a gene contains both the exons that give rise to an ORF and introns, intervening sequences between exons. During the splicing process of the initial mRNA transcript, the introns are removed by splicesome, a large, RNA-protein complex. Introns have canonical two nucleotide sequence at the 5' and 3' end of the intronic sequence that signal the splice sites recognized by the spliceosome. [http://www.imgt.org/textes/IMGTeducation/Aide-memoire/_UK/splicing/]

The 5' sequence is called the donor site and the 3' sequence is called the acceptor site. A critical factor in annotation of removing introns and joining exons is that the donor and the acceptor of an intron, between two exons to be spliced, must belong to the same splicing frame; otherwise, a frameshift mutation will occur.

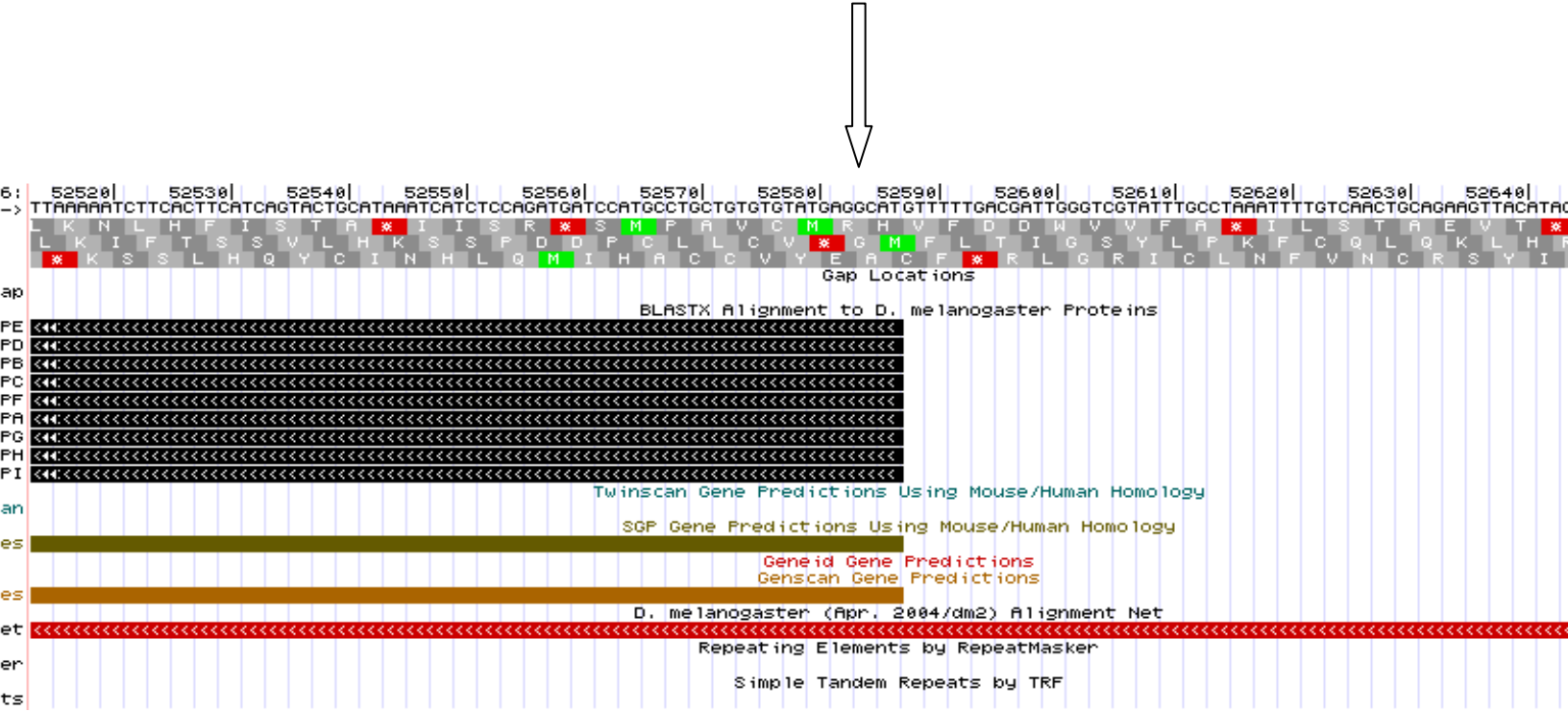
The UCSC Genome Browser has a track that detects donor/acceptor sites; in Drosophilia, the donor sequence is usually starts "GT" and the acceptor sequence is usually ends with "AG.". However, visual determination is usually needed because of sequence variation and the requirement to maintain the reading frame in the annotated sequence.

The coordinates [physical location of donor/acceptor sites in gene model] and frame are obtained from the blastx results.

From activity G, the blastx showed that the reading frame was -2 and the coordinates/frame of exon 1 are:

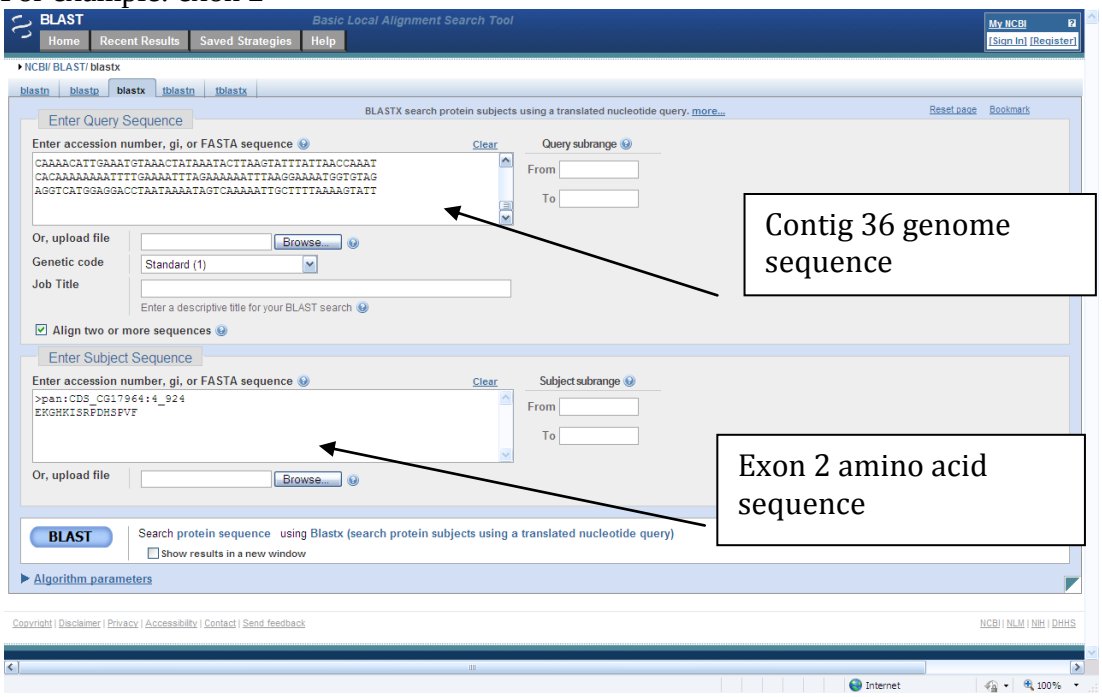
```
Query 52587  MPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDL
TESEMPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE 52423
```

Query indentifies the nucleotide number of contig 36 where exon 1 begins is NT 52587 and ends at NT 52423. These physical locations are shown in the UCSC Genome Browser [see below for example].

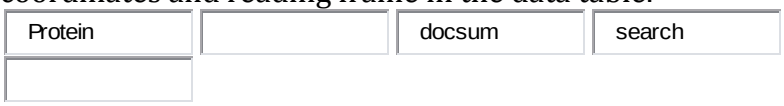


The coordinates for exon 1 of the pan gene were determined in Activity F. The coordinates and reading frames of all the exons need to be recorded in order to facilitate annotation via the UCSC Genome Browser. Therefore, a table of coordinates and frames is needed to keep track of the overall reading frame; see Data Page pan unique Coordinates. NOTE: The pan gene in Contig 36 is in the reverse direction ("complementary" or "minus" strand) and therefore the reading frames will be negative.

H. Go to Gene Record Finder (via GEP or directly: http://gander.wustl.edu/~wilson/dmelgenerecord/index.html) and enter "pan" in the window. Copy the polypeptide sequence of exon 2, and use it to do a blastx search as in Activity D. For example: exon 2



Select BLAST; on the results page, scroll down to Alignments. Note that the E value is poor for exon 2, but the length of the search was only 15 amino acids. Record the start, end coordinates and reading frame in the data table.




```
>lcl|22575 pan:CDS_CG17964:4_924
Length=15


                                                                Sort alignments for
this subject sequence by:

                                                                  E value  Score

Percent identity

                                                                  Query start

position  Subject start position
 Score = 32.7 bits (73),  Expect = 5e-05
 Identities = 13/15 (86%), Positives = 14/15 (93%), Gaps = 0/15 (0%)
 Frame = -2

Query  52182  EKGQKIARPDHSPVF  52138
              EKG KI+RPDHSPVF
Sbjct  1      EKGHKISRPDHSPVF  15
```

Repeat the above BLAST steps with all remaining exons in the unique isoform and record the information in the table. On a separate data page, save the polypeptide sequence of each exon.

Notes: (1) At the left top of the Blast results page click "Edit and Resubmit", then clear subject data and enter sequence of the next exon. This way, you won't have to enter the entire Contig 36 sequence each time you do a blastx. (2) Select the alignment with the smallest E value [it should be the first set of coordinates listed in the blast results page].

I. Go to UCSC from GEP webpage [or http://gander.wustl.edu/cgi-bin/hgGateway]

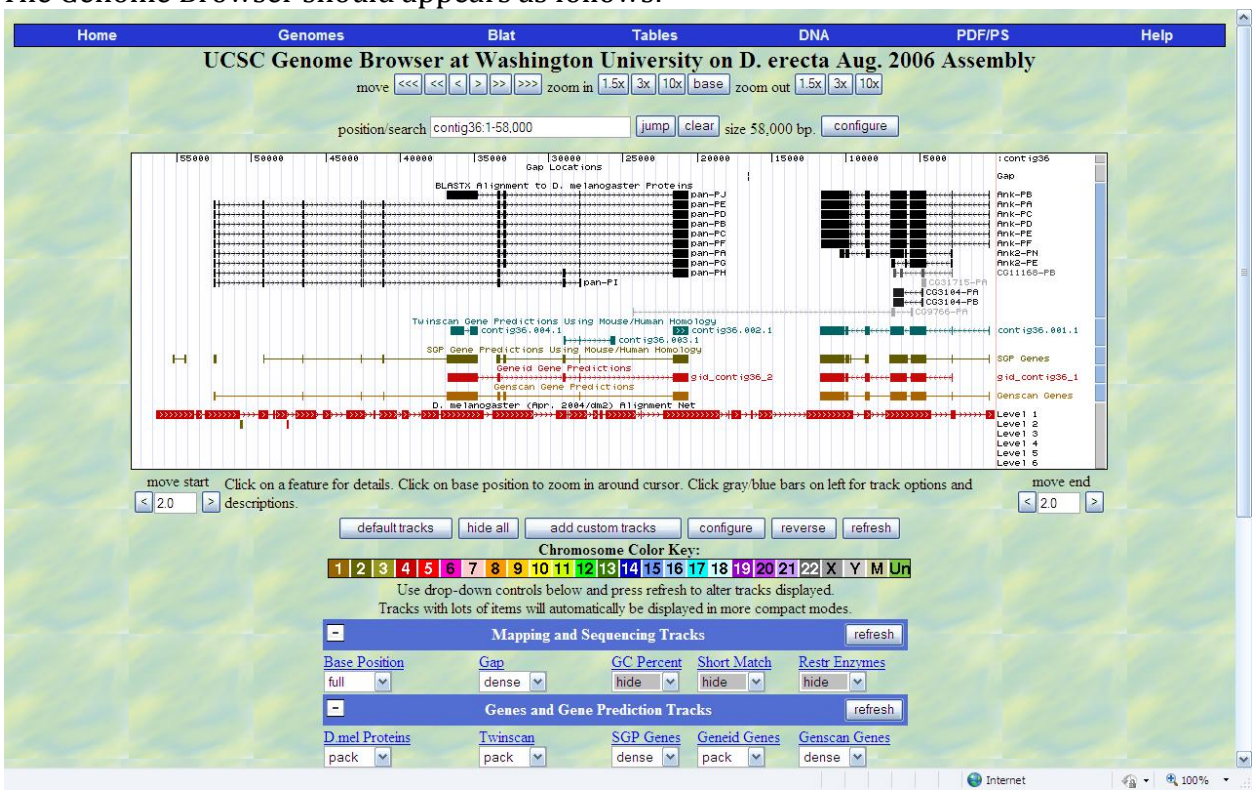
1. Since the pan gene is in the reverse orientation, select "reverse" tab that is underneath the window. This reverses the view so that the gene will appear to be running from left to right.

2. Under the "Tracks" settings, make the following changes (some may be set by default).

- Mapping and Sequencing Tracks change Base Position to full.
- Genes and Gene Prediction Tracks change D.mel Proteins to pack.
- Make sure all the gene prediction tracks (Genscan, Geneid, Twinscan, etc.) are set to either dense or pack.
- mRNA and EST Tracks, change Other ESTs to hide
- Comparative Genomics change D. mel. Chain to hide and (dm2)D.mel.Net to full
- Variation and Repeats change RepeatMasker to hide and Simple Repeats to hide.
- Experimental Tracks, change Predicted Splice sites to hide.

3. Click the refresh button at the bottom of the page.

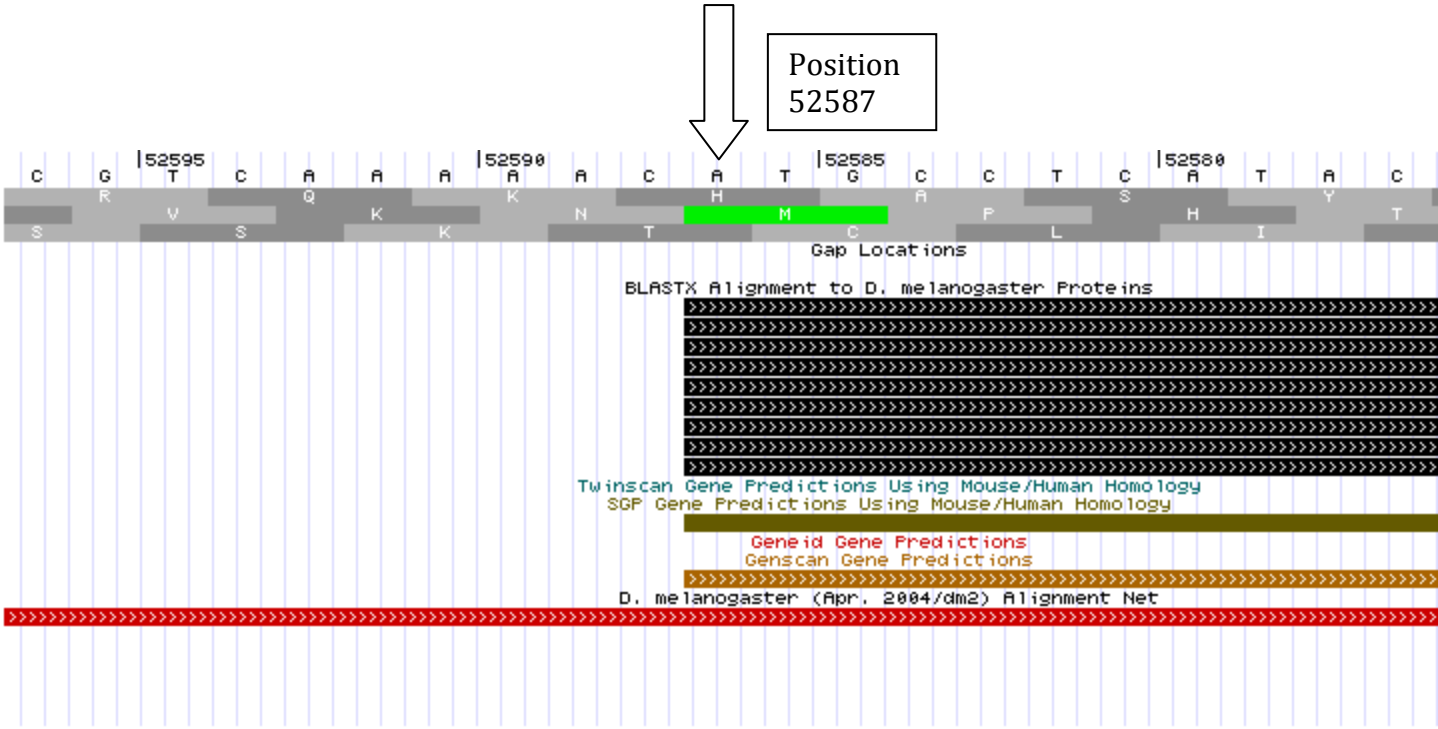
The Genome Browser should appears as follows:



There are various ways to enlarge the screen so that you can see individual nucleotides in the sequence. You can move to an approximate location by entering the beginning coordinate in the position/search window for the exon you are examining.

Place the cursor underneath the nucleotide position number of interest and click. This will center on the position and enlarge the image; repeat until you can see the nucleotide and amino acid sequence. You will need to click the numbers several times to get an image that you can easily read the numbers, as in below:

position/search contig36:52,572-52,597 jump size 26 bp. configure

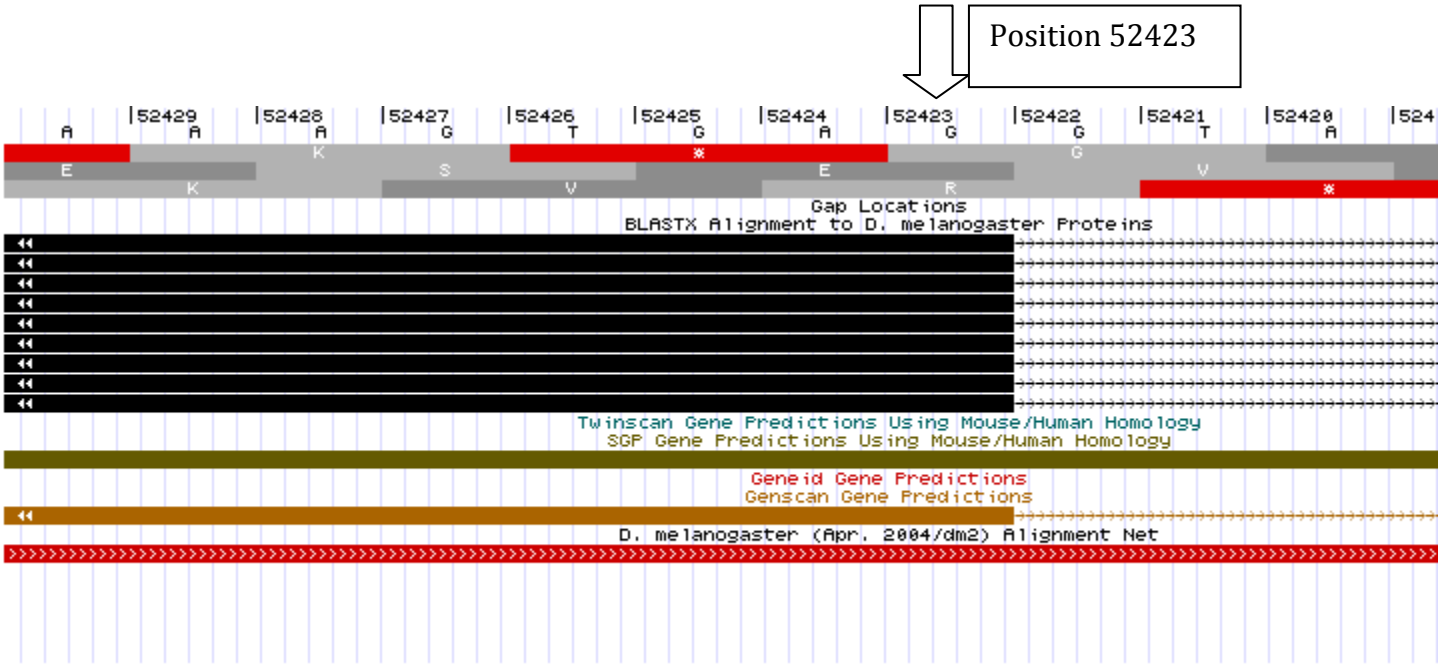


If you zoom out too much, use the zoom feature to return to a smaller image. Use the <<< move on the left or the right tab to get to the beginning/end of the gene OR enter the coordinates in the position/search window.

J. The first exon will begin with an ATG (green) and the last exon will end with a stop codon (red). In the window above, the "A" in ATG is at NT 52587 (see arrow); below the NT are the three reading frames. The green M indicates the start codon, Methionine. It is in the second row, or frame 2, as previously determined by the blastx search. (Remember, since selecting the "reverse" the minus sign is negated.)

Use the "move end" or zoom out tabs, or use the "jump" option to get to the end of exon 1.

position/search contig36:52,417-52,430 jump size 14 bp. configure



K. The last base in exon 1 is a G at position 52423 (see arrow). In the middle row of amino acids (frame 2) notice that the codon for E includes this G as the third NT of the codon E. Since the end of the exon includes a full codon, the end of exon 1 is in phase 0.

The phase is the number of bases "left over" after the last full codon in the exon. Thus, to prevent a frame shift, the phases at the end of exon and the beginning of the next must add up to three. Sometimes there are multiple possible splice sites so you will need to determine the most likely correct one based on maintaining the phase.

Phase 1: one of 3 NT left over
Phase 2: two of 3 NT left over
Phase 0: no left over NT, i.e., complete codon.

Note that at positions 52422 and 52421 the bases are GT, and this is the beginning or donor sequence, of an intron.

In Drosophilia, the beginning, or donor sequence, of an intron is GT and the end, or acceptor sequence is AG.

The donor/acceptor sequences vary among organisms.

Note: Remember that the sequence is in "reverse" so as you move from left to right, the position numbers decrease.

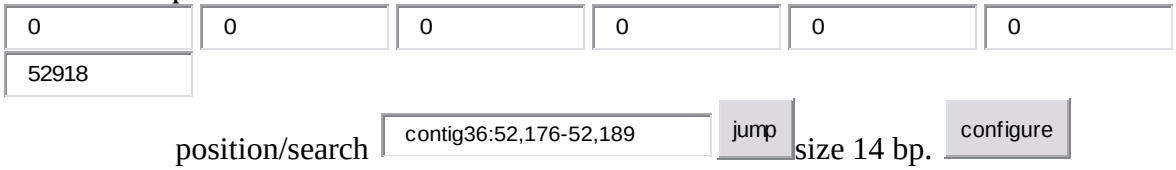
L . Identifying beginning and ends of exons or Keeping phases in phase: This is the fun part!

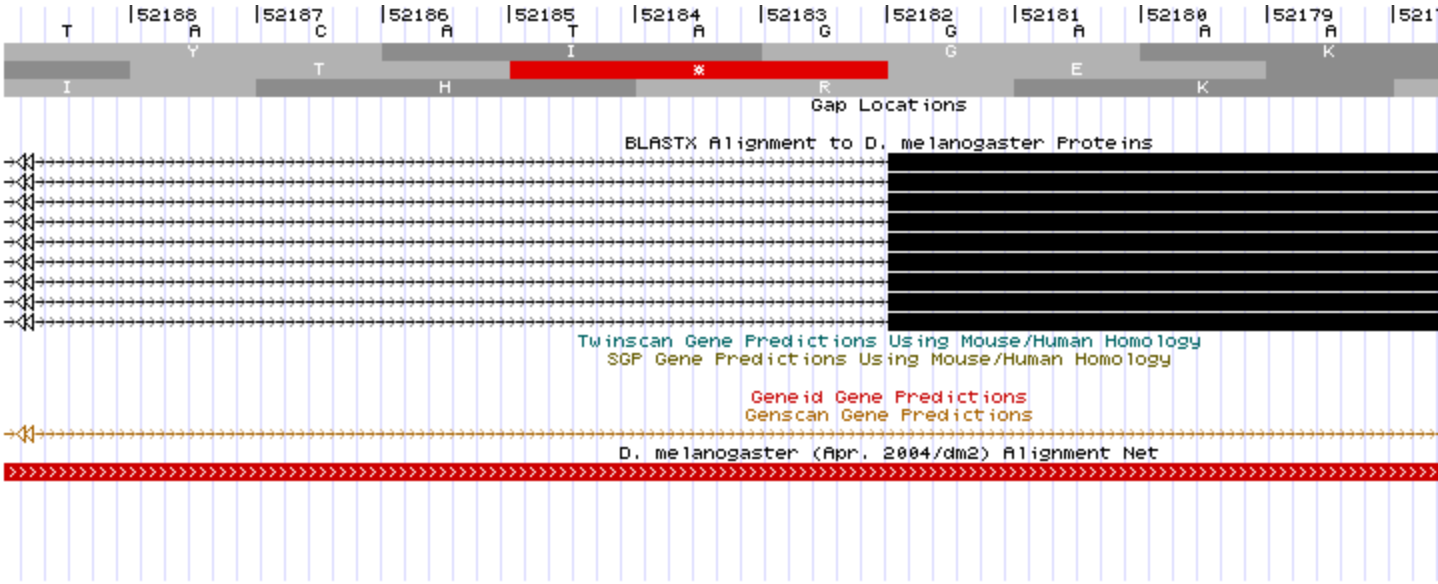
NOTE: The following coordinates are for the pan-PE isoform!

From part H, it was determined that the beginning position of exon 1 is 52587 and the end position is 52423. Exon 1 ended in reading frame 2 (middle row). Record this information, plus the beginning and end phases in a Exon Phase Score Box.

Once you have found the beginning and end location of exon 1, copy/paste the Genome Browser Windows and keep them as part of the data record.

Move the sequence in the Genome Brower to the start of the exon 2.





The second exon begins at position 52182, which is a G. This is the first of the 3 NT needed for a codon (amino acid E, codon GGA), therefore this is in phase 0, i.e., there are no left over NT needed to make a codon. The phases at the end of an exon and at the beginning of an exon need to be the same; in this case, exon 1 ended in phase 0 and exon 2 ended in phase 0.

Note: if the phases are not both 0, then the phase of the end of last exon plus the phase of the beginning of the next exon must add to 3, i.e., 3 NT in a codon.

Also note that the preceding two NT (positions 52184 and 52183) are AG, which is the acceptor sequence (end) of an intron.

M. Repeat steps J-K to determine the end of exon 2 and the beginning and end positions of the remaining exons in the pan PE gene. The last codon of the last exon must be a stop codon (TAG, TAA, TGA).

Congratulations! You have annotated your first gene.

Exon Phase Score Box

Exon Number	AG nucleotide positions	Beginning Nucleotide Position of Exon (Phase)	Ending Nucleotide Position of Exon (phase)	GT Nucleotide Positions
Exon 1		**Note: First codon must be ATG**		
Exon 3				
Exon etc				
Last exon		**Note: Must end with UAG, UAA, OR UGA**		

[[Category:Faculty Resources]]