Lab Week 9 - A Sample Annotation Problem - Answers

Adapted by Chris Shaffer from a worksheet by Varun Sundaram, Bio4342 Class of 2009

Question 1

Which gene predictor matches the best to the BLASTX output?

The Nscan Gene Predictor matches up best to the BLASTX output, showing three features that line up well to the three different features (putative protein-coding genes) in the BLASTX track.

The PMCA gene. If one clicks on any of the alignment blocks in the BLASTX track for each of the features, a BLAST Summary Viewer window appears with details of the alignments, including the E values. Gene CG42314 appears to match much better (larger, more numerous blocks and lower E values) than the others and is thus our best candidate for the putative ortholog.

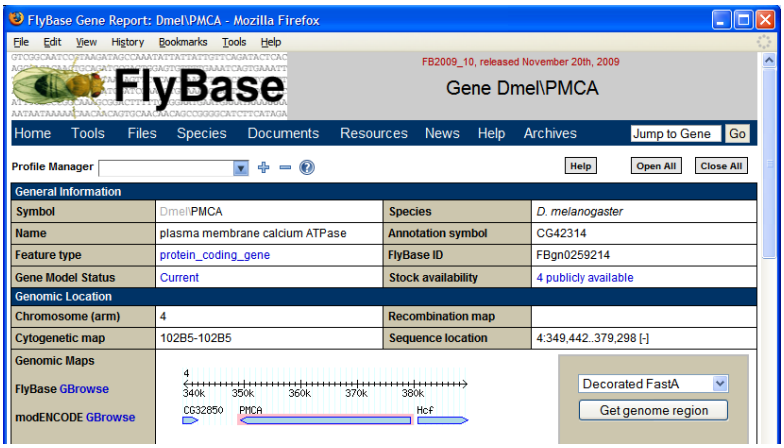
Question 2

What is the symbol of the D. melanogaster protein-coding gene in the BLASTX track that appears to match well to the left-most D. grimshawi feature, predicted by Genscan or Nscan?

PMCA

Type this gene symbol (case sensitive) into the flybase.org Quick Search box. What is the name of this gene?

Plasma membrane calcium ATPase



Which chromosome is it on in D. melanogaster? 4

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 an orthologue to the D. grimshawi gene?

In the region of the fourth (dot) chromosome from *Drosophila virilis* shown in the Power Point slides (in the lab manual from last week), 27/28 genes are shared between *D. melanogaster* and *D. virilis* (figure from 2006 *Genome Biology* paper). This shows that there has been little

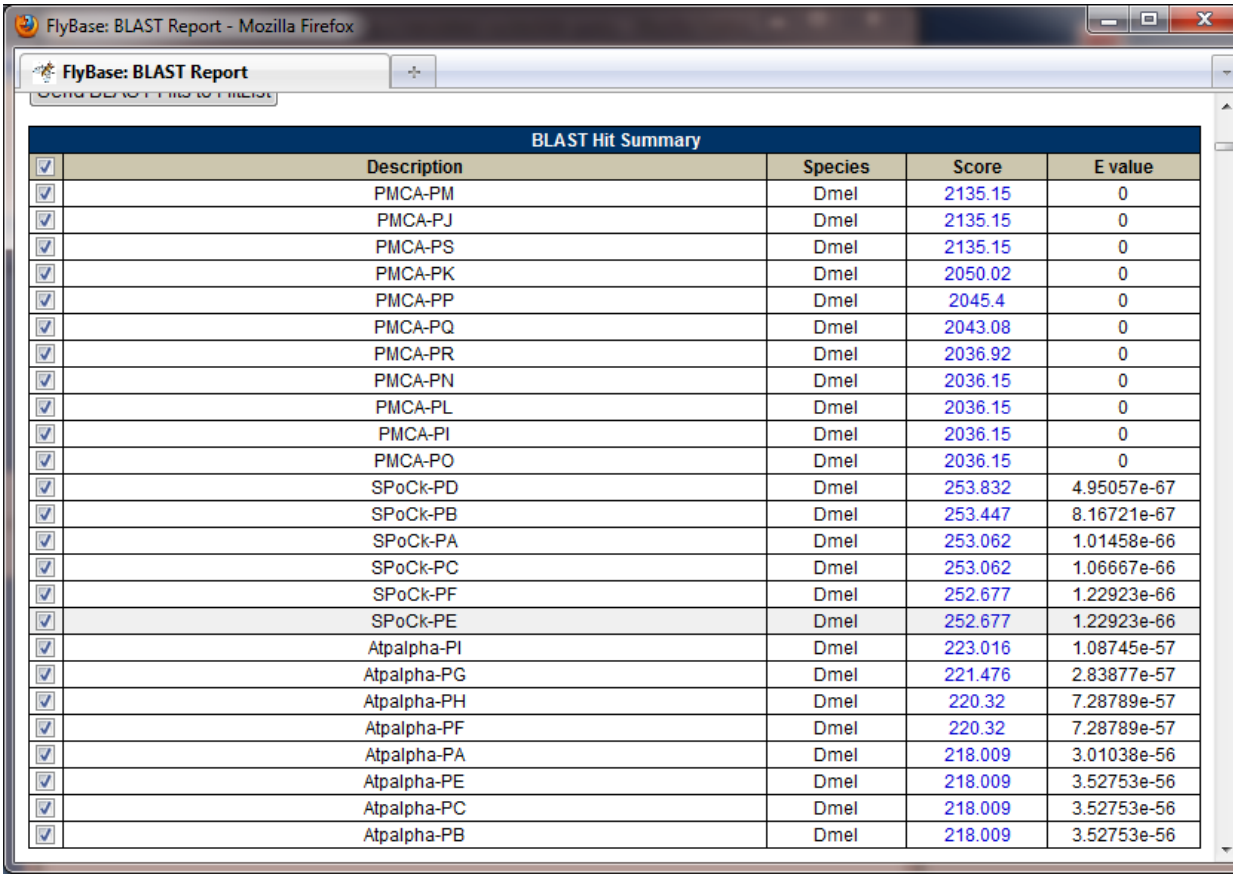
movement of genes off of the 4th chromosome since the two species diverged. Since the evolutionary distance between *D. grimshawi* and *D. melanogaster* is similar to that between *D. virilis* and *D. melanogaster* (see the 12 *Drosophila* genomes phylogenetic tree), we would expect that there would be similar levels of gene conservation on the 4th chromosome of *D. grimshawi* and *D. melanogaster.* The fosmid contains dot chromosome DNA from *D. grimshawi.* Thus, the fact that the PMCA gene is on the 4th chromosome in *D. melanogaster* is good evidence that this region of the *D. grimshawi* dot chromosome contains the PMCA orthologue in *D. grimshawi*.

Question 3

Examine the list of the top 25 hits. How do the Scores and E-values of the PMCA isoforms compare to the other hits?

The results (see screen shot below) show hits to various isoforms of PMCA with Scores of over 2000 and E-values of 0 (which indicates that there is zero percent probability that we could have gotten those S scores by alignment of any two random sequences). The next best hits are to isoforms of the SPoCk and Atpalpha genes, with Scores about 10-fold lower and E-values ranging from 10^{-67} to 10^{-52}. Looking at the alignments, one sees the reason for these lower scores and higher E-values, as the aligned sequences are much shorter and much less similar. Turns out that the SPoCk and Atpalpha genes are also NOT on the dot chromosome, which further decreases our confidence that they are orthologs to genes in this fosmid. Thus, the best evidence indicates that it is the PMCA ortholog that is found in this region of *D. grimshawi* DNA.

blastp hit summary:



BLAST Hit Summary

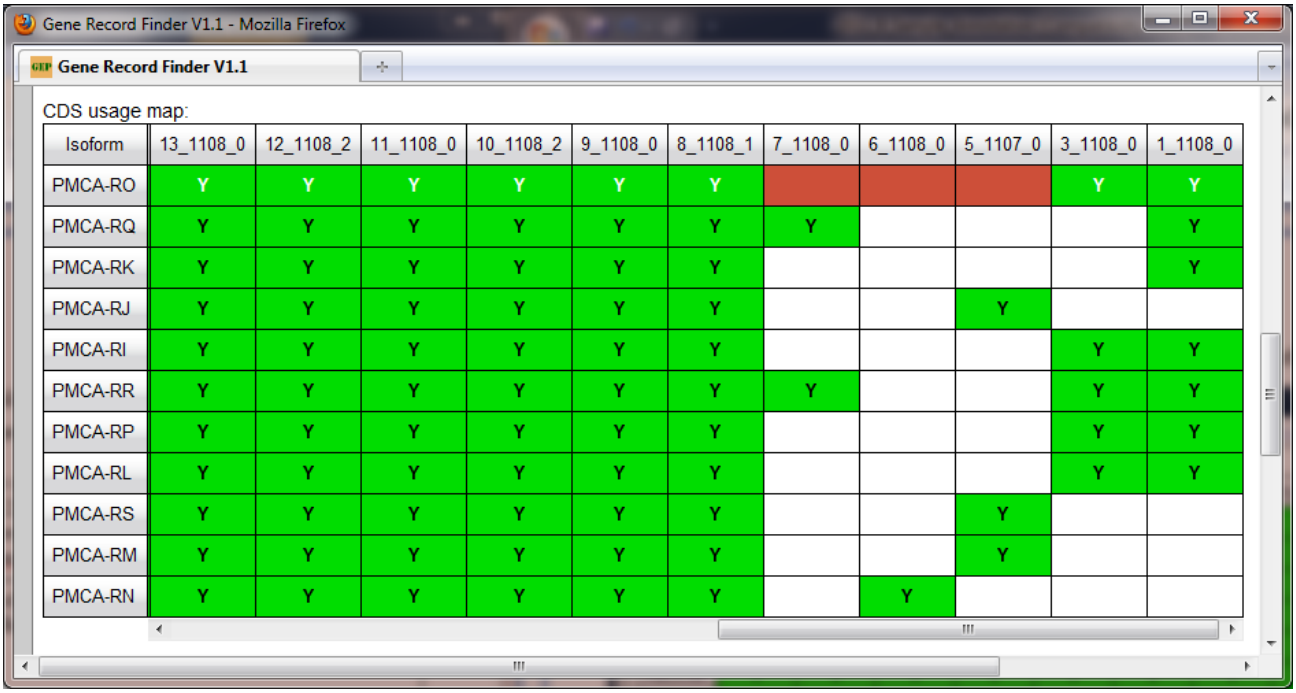
Description	Species	Score	E value
PMCA-PM	Dmel	2135.15	0
PMCA-PJ	Dmel	2135.15	0
PMCA-PS	Dmel	2135.15	0
PMCA-PK	Dmel	2050.02	0
PMCA-PP	Dmel	2045.4	0
PMCA-PQ	Dmel	2043.08	0
PMCA-PR	Dmel	2036.92	0
PMCA-PN	Dmel	2036.15	0
PMCA-PL	Dmel	2036.15	0
PMCA-PI	Dmel	2036.15	0
PMCA-PO	Dmel	2036.15	0
SPoCk-PD	Dmel	253.832	4.95057e-67
SPoCk-PB	Dmel	253.447	8.16721e-67
SPoCk-PA	Dmel	253.062	1.01458e-66
SPoCk-PC	Dmel	253.062	1.06667e-66
SPoCk-PF	Dmel	252.677	1.22923e-66
SPoCk-PE	Dmel	252.677	1.22923e-66
Atpalpha-PI	Dmel	223.016	1.08745e-57
Atpalpha-PG	Dmel	221.476	2.83877e-57
Atpalpha-PH	Dmel	220.32	7.28789e-57
Atpalpha-PF	Dmel	220.32	7.28789e-57
Atpalpha-PA	Dmel	218.009	3.01038e-56
Atpalpha-PE	Dmel	218.009	3.52753e-56
Atpalpha-PC	Dmel	218.009	3.52753e-56
Atpalpha-PB	Dmel	218.009	3.52753e-56

Question 4

Hint☺: Examine both sides of the Polypeptide Details window to answer these questions.

How many different protein isoforms exist for this gene?

There are 11 different mRNA isoforms, but only seven different protein isoforms. From the chart on p. 123, one can see that all but one (the P isoform) of the protein isoforms have CDS 22 through 14 in sequential order (with CDS 22 being the first coding exon in the gene). However, from the table below, one sees that the other end of the protein varies across several isoforms.



Isoform	13_1108_0	12_1108_2	11_1108_0	10_1108_2	9_1108_0	8_1108_1	7_1108_0	6_1108_0	5_1107_0	3_1108_0	1_1108_0
PMCA-RO	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RQ	Y	Y	Y	Y	Y	Y	Y				Y
PMCA-RK	Y	Y	Y	Y	Y	Y					Y
PMCA-RJ	Y	Y	Y	Y	Y	Y			Y		
PMCA-RI	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RR	Y	Y	Y	Y	Y	Y	Y			Y	Y
PMCA-RP	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RL	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RS	Y	Y	Y	Y	Y	Y			Y		
PMCA-RM	Y	Y	Y	Y	Y	Y			Y		
PMCA-RN	Y	Y	Y	Y	Y	Y		Y			

Which mRNA isoforms code for identical protein isoforms?

mRNA isoforms O, I and L code for identical proteins; mRNA isoforms J, S and M code for identical proteins. mRNA isoforms Q, K, R, P, and N code for unique proteins.

How do the different protein isoforms vary with respect to coding sequence (CDS) usage?

Isoform P skips CDS 18 and also CDS 7 through 5. See table above for differences between other isoforms. The R protein isoform is the longest (uses all but CDS 6 and 5), and NONE of the protein isoforms uses every one of the CDS of the PMCA gene.

Question 5.
Repeat the same blastx searches with the next two CDS's (#21 and 20); copy and paste the best alignments into a Word document (when copying alignments be sure to include the Score, etc. header information). What are the DNA base coordinates of the beginning and end of each alignment? What frame was translated to generate the amino acid sequence for each alignment?

Answers highlighted in red below

Second exon (CDS 21):

```
Length=64

 Score =  129 bits (324),  Expect = 4e-34
 Identities = 64/64 (100%), Positives = 64/64 (100%), Gaps = 0/64 (0%)
 Frame = +2

Query  3257  LSGSKADEEHRRETFGSNVIPPKPPKTFLTLVWEALQDVTLIILEVAALVSLGLSFYKPA  3436
             LSGSKADEEHRRETFGSNVIPPKPPKTFLTLVWEALQDVTLIILEVAALVSLGLSFYKPA
Sbjct  1     LSGSKADEEHRRETFGSNVIPPKPPKTFLTLVWEALQDVTLIILEVAALVSLGLSFYKPA  60

Query  3437  DEDA  3448
             DEDA
Sbjct  61    DEDA  64
```

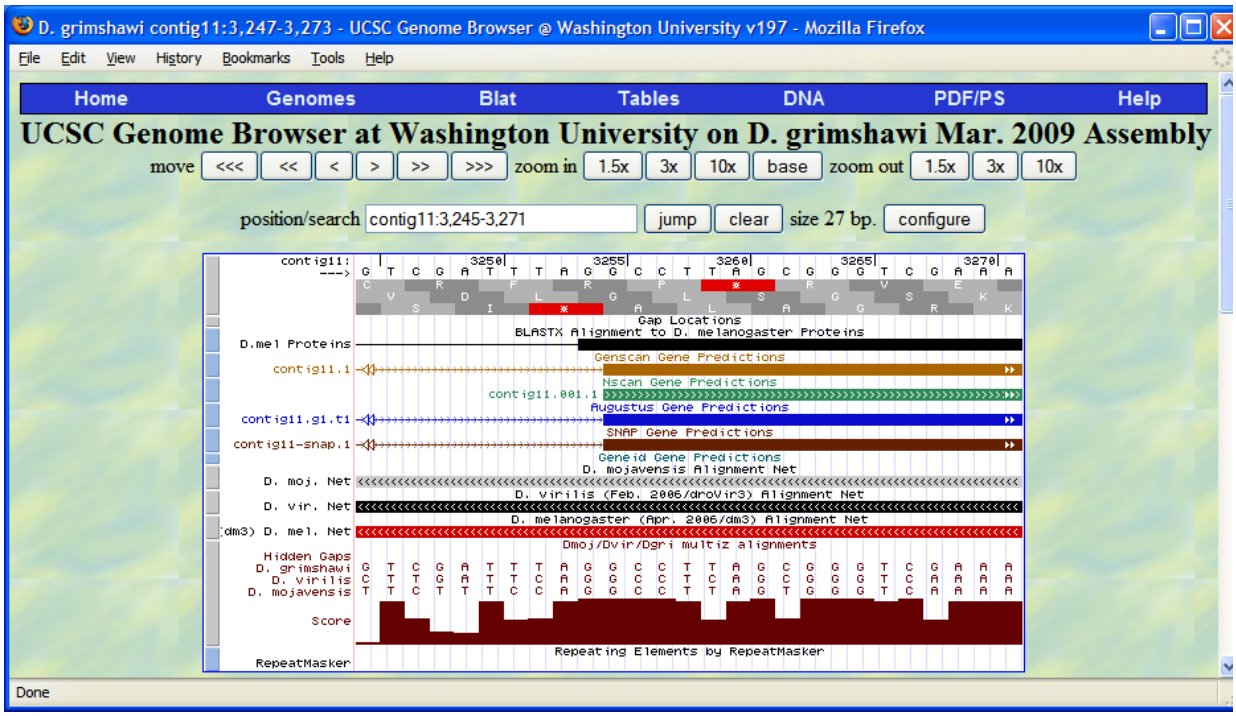
Third exon (CDS 20):

```
Length=95

 Score =  189 bits (479),  Expect = 9e-52
 Identities = 92/95 (96%), Positives = 95/95 (100%), Gaps = 0/95 (0%)
 Frame = +1

Query  3973  LLQEEDEHHGWIEGLAILISVIVVVIVTAFNDYSKERQFRGLQNRIEGEHKFSVIRGGEV  4152
             +LQEE+EHHGWIEGLAILISVIVVVIVTAFNDYSKERQFRGLQNRIEGEHKFSVIRGGEV
Sbjct  1     VLQEEEEHHGWIEGLAILISVIVVVIVTAFNDYSKERQFRGLQNRIEGEHKFSVIRGGEV  60

Query  4153  CQISVGDILVGDIAQIKYGDLLPADGCLIQSNDLK  4257
             CQISVGDILVGDIAQ+KYGDLLPADGCLIQSNDLK
Sbjct  61    CQISVGDILVGDIAQVKYGDLLPADGCLIQSNDLK  95
```

Question 6

Look around the region where the alignment to CDS 21_1108_2 (the second exon) begins. How many acceptor sites can you find? Considering the frame of the conserved amino acids you found in question 5, what is the phase of each putative acceptor site you find? Using just phase information, which if any of these acceptor sites is/are usable to maintain the proper translation frame throughout the first two exons? Itemize what other evidence you could consider if you have two or more possible donor/acceptor pairs. Finally, record the base coordinates for exon 1 and the beginning of exon 2 based on your complete analysis.

The alignment begins at around base 3257. This region is shown here:



There are two acceptor sites in this region: the "AG" at 3253-4 and the "AG" at 3260-1. From the exon by exon blastx search above, we know that the frame with the conserved amino acids for the 19_945 exon is frame +2 (the second row of light and dark gray boxes). Note that in frame +2, the 3253-4 acceptor results in the exon beginning two bases (GC) before the first complete codon (the L); this we denote "phase 2." The acceptor at 3260-1 has one base (C) before the first complete codon (the codon for the G) in frame +2 (phase 1). Since the intron donor sequence after the end of the first exon creates a phase 1 exon, we must find an acceptor that results in a phase 2 start to the second exon. Thus, the best acceptor is the 3253-4 "AG". Note that the other acceptor is not only out of phase, but, if used, would cause one of the conserved amino acids (the L) to be omitted from the protein.

When two or more intron donor/acceptor pairs are found, the following should be considered:

1. The pair that maximizes the inclusion of conserved amino acids would be strongly favored.
2. Pairs that have more bioinformatic support would be favored. This includes co-incidence with gene predictors (the more the better) and higher scoring predicted splice sites.

3. Finally, for any combinations that are indistinguishable by the above criterion, by convention, the pair that creates the longest protein should be picked.

The final result has coordinates of the first exon as 3035 – 3191 (with phase 1).
The start of the second exon is phase 2 and mapped to 3255 (the first base in the exon).

Question 7
Use the results of the alignment of the second and third exons in question 5 to locate the 3' end of the second exon and the beginning (5' end) of the third exon.

Following the general procedure above, the end of the second exon (CDS 21) is found at 3449 and is phase 1, and the beginning of the third exon (CDS 20) is found at 3971 (with phase 2).