

What is BLAST?

- ♦ BLAST stands for
Basic **L**ocal **A**lignment **S**earch **T**ool
- ♦ Why is BLAST popular?
 - Good balance of sensitivity and speed
 - Reliability
 - Flexibility

Where Can I run BLAST?

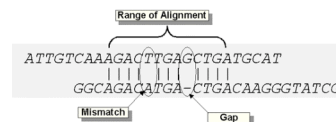
- ♦ We will use two sites:
 - NCBI (www.ncbi.nih.gov)
databases updated constantly (daily); very slow at times
 - FlyBase (<http://flybase.org/blast/>)
databases for many *Drosophila* species

BLAST output

- Graphical overview, showing alignment blocks as bars = regions of sequence similarity between Query (top) and database sequences
- List of Sequences with scores (see next slide)
 - Raw score, higher is better (length dependent)
 - Expect value, smaller is better
(length and database size independent)
- List of alignments

Calculating alignment scores

The raw score S for an alignment is calculated by summing the scores for each aligned position and the scores for gaps. In this figure, a DNA alignment is shown.



$$S = \sum (\text{identities, mismatches}) - \sum (\text{gap penalties})$$

$$\text{Score} = \text{Max}(S)$$

E value (Expectation value). The number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance. The lower the E value, the more significant the score.

The Databases

- ♦ Genbank nr/nt (protein and nucleotide versions)
 - Non-redundant large databases (compile & remove duplicates)
 - Anyone can submit, you can call your sequence anything
 - Quality low; names can be meaningless
- ♦ EST (Expressed Sequence Tags) databases
 - Short single reads of cDNA clones
 - Other short single reads
 - High error rates
- ♦ Swissprot
 - Curated from literature
 - REAL proteins; REAL functions; small;
- ♦ Genomic Databases
 - Human, Mouse, *Drosophila*, *Arabidopsis*, etc
 - NCBI, species-specific web pages

BLAST Protocols

- ♦ The most common BLAST search includes **five protocols**:

Program	Database	Query
BLASTN	Nucleotide	Nucleotide
BLASTP	Protein	Protein
BLASTX	Protein	Nt. → Protein
TBLASTN	Nt. → Protein	Protein
TBLASTX	Nt. → Protein	Nt. → Protein

BLASTN

- ♦ BLASTN
 - The query is a nucleotide sequence.
 - The database is a nucleotide database
 - No conversion is done on the query or database
- ♦ DNA :: DNA homology
 - Mapping oligos to a genome
 - Cross-species sequence exploration
 - Annotating genomic DNA with ESTs

BLASTP

- ♦ BLASTP
 - The query is an amino acid sequence
 - The database is an amino acid database
 - No conversion is done on the query or database
- ♦ Protein :: Protein homology
 - Protein function exploration
 - Novel gene → makes parameters more sensitive

BLASTX

- ♦ BLASTX
 - The query is a nucleotide sequence
 - The database is an amino acid database
 - All six reading frames are translated on the query and used to search the database
- ♦ Coding nucleotide seq :: Protein homology
 - Gene finding in genomic DNA
 - Annotating ESTs (and Shotgun Sequence)

TBLASTN

- ♦ TBLASTN
 - The query is an amino acid sequence
 - The database is a nucleotide database
 - All six frames are translated in the database and searched with the protein sequence
- ♦ Protein :: Coding Nucleotide DB homology
 - Mapping a protein to a genome
 - Mining ESTs (Shotgun DNA) for protein similarities

TBLASTX

- ♦ TBLASTX
 - The query is a nucleotide sequence
 - The database is a nucleotide database
 - All six frames are translated on the query and on the database
- ♦ Coding :: Coding homology
 - For searching distantly-related species
 - Sensitive but expensive

Appendix D: Completed Annotation Report for the *Spinophilin* G Isoform of *Drosophila erecta*

Annotation report

Student Name: xxxxxxxx & xxxxxxxxxx
Student E-mail: xxxxxxxx@amherst.edu & xxxxxxxx@amherst.edu
Faculty Advisor: Dr. Julie Emerson
College/University: Amherst College

Project name: derecta_2nd3Lcontrol_Nov2011_fosmid52
Project species: *Drosophila erecta*
Date of submission: Dec. 7, 2011
Size of project in base pairs: 28069
Number of genes in project: 1
Complete Report

For each gene complete the following Gene Report Form (copy and paste to create as many copies as needed, be sure to create enough isoform reports within your gene form for all isoforms):

=====Gene Report Form=====

Gene name: D erecta Spinophilin
Gene symbol: dere_Spn
Approximate location in project (from start codon to stop codon): 37067-8998
Number of isoforms in *D. melanogaster*: 10
Number of unique isoforms based on coding sequence: 8
List names of unique isoforms (i.e. PA, PC etc): PB/PC/PH, PI, PJ, PG, PF, PD, PE, PK
Number of unique isoforms found in this project: 1

=====Isoform report=====

For each (protein) isoform complete the following (copy and paste to create as many copies as needed):

Gene-isoform name: dere_Spn-PG
Is the 5' end of this isoform missing off the end of project: no
Is the 3' end of this isoform missing off the end of the project: no

Enter the coordinates of your final gene model for this isoform into the gene model checker and paste a screen shot of the results below:

The screenshot shows the Gene Model Checker web application in a Mozilla Firefox browser. The 'Configure Gene Model' section on the left contains the following details:

- Model Details:**
 - Fosmid Sequence File: - Ortholog in *D. melanogaster*:
 - Coding Exon Coordinates:
 - Annotated Untranslated Regions? ☐ Yes ☒ No
 - Orientation of Gene Relative to Query Sequence: ☐ Plus ☒ Minus
 - Completeness of Gene Model Translation: ☒ Complete ☐ Partial
 - Stop Codon Coordinates:
- Project Details:**
 - Project Group:
 - Project Name:

At the bottom of the configuration section are buttons for 'Verify Gene Model' and 'Reset Form'.

The 'Checklist' section on the right displays a table of checks:

View	Criteria	Status	Message
<input checked="" type="checkbox"/>	Check for Start Codon	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 1	Pass	Already checked for Start Codon
<input checked="" type="checkbox"/>	Donor for CDS 1	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 2	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 2	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 3	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 3	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 4	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 4	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 5	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 5	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 6	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 6	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 7	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 7	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 8	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 8	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 9	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 9	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 10	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 10	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 11	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 11	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 12	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 12	Pass	
<input checked="" type="checkbox"/>	Acceptor for CDS 13	Pass	
<input checked="" type="checkbox"/>	Donor for CDS 13	Pass	Already checked for Stop Codon
<input checked="" type="checkbox"/>	Check for Stop Codon	Pass	
<input checked="" type="checkbox"/>	Additional Checks	Pass	
<input checked="" type="checkbox"/>	Number of coding exons matched <i>D. melanogaster</i> ortholog	Pass	

At the bottom left, there is a link for 'External Links: [Old Gene Checker](#)'. The browser's status bar at the bottom right shows 'zotero'.

Using the custom track feature from the Gene Model Checker, capture a screen shot of your gene model shown on the browser for your project; zoom in on only your isoform. **If available**, also show these tracks: other ref seq; all relevant Gene Prediction Tracks, 3-way and 5-way multi Z). If you need help, see lab instructor and/or read the bottom of page 9 in The Gene Model Checker User Guide, on the “Documentations” page under the Help menu at gcp.wustl.edu. (Type comments about your model below the screen shot.):

Query	241	TTRTHSDLNRCDIIRTVPGTGLIMDSEKVAKPAMEPPQPPPNASPNPPMRAQAPPEIKPR	300
Sbjct	240	TTRTHSDLNRCDIIRTVPGTGLIMDSEKVAKPAMEPPQP PNASPNPPMRAQAPPEIKPR	299
Query	301	SGKIGSPVKSPPLPPIPAVKPKNVSPVKFNPDRLRQSPTKTADNSPPPPPAKSAAVLQRS	360
Sbjct	300	SGKIGSPVKSPPLPPIPAVKPKNVSPVKYNPDRLRQSPTKTADNSPPPPPAKSAAVLQRS	359
Query	361	LMQEQQELLRNNSCDQGVAPIPEKPRKKSVDLIEDTLPLTNCSTPSSCASPTSSYLMQPA	420
Sbjct	360	LMQEQQELLRNNSCDQGVAPIPEKPRKKSVDLIEDTLPLTNCSTPSSCASPTSSYLMQPA	419
Query	421	KRGSLDGGSGNGQYPGNLSGSTNSATSGSPVASASSGPSSPVHTEDEKQENESTEKSEM	480
Sbjct	420	KRGSLDGGSGNGQYPGNLSGSTNSAASGPVASASSGPSSPVHTEDEKQENESTEKSEL	479
Query	481	EYYHGGNYNSVPRRRRSENEGRKSVDESSPSANNSQQQQQHSIPGSAAGSPQRVANKRSS	540
Sbjct	480	EYYHGGNYNSVPRRRRSENEGRKSVDESSPSANNSQQQQQHSIPGSATGSPQRVANKRSS	539
Query	541	ITVNMPAAGLGQRPPSIISTTSQDEGGFNESAPELKAKLQPAYDQTEEQPHSLNYVDVGY	600
Sbjct	540	ITVNMPAAGLGQRPPSIISTTSQDEGGFNESAPELKAKLQPAYDQTEEQPHSLNYVDVGY	599
Query	601	RLNPDGSESREVYGSEAELYDTAKVTDQMQRKFHGANGFGQESSTVYAI IKPDVQESQPVA	660
Sbjct	600	RLNPDGSESREVYGSEAELYDTAKVTDQMQRKFHGANGFGQESSTVYAI IKPDMQESQPVA	659
Query	661	PSRSVLIQSPNSSSVESGSPHRSYSSPPVGVVSPIRRNSNQDQSVGGGG--SAKTTTP	718
Sbjct	660	P+R VL+QSP SSSVEGSPLHRGSY+SPPVGVVSPIRRNSNQDQ VGGGG SAK+TP	719
Query	719	QCSPARSALVKGIAPIASIDAHEEEELDLVEEDEHLAVEYVEVLELQQDEEEEEEAPVLPE	778
Sbjct	720	PCSPARSAMVKGIAPIASIDAHEEEELDLVEEDEHLAVEYVEVLELQQDDDEEEAPVLPE	779
Query	779	RRAPAQGSLELQDLEYADTSAGEDEEDI INHLKDGDLVDELIDDVVDEVIKVVHNHSA	838
Sbjct	780	RRAPAQGSLELQDLEYADTSAGEDEEDI INHLKDGDLVDELIDDVVDEVIKVVHNHSA	839
Query	839	TAPSIQAATPAAAI PREDSLPDDMTAAEAERLLSSRQQSLLSDEQAKEVEQILNAAPSVG	898
Sbjct	840	TAP IQAA PAAAI PR DSLPDDMTAAEAERLLSSRQQSLLSDEQAKEVEQILNAAPSVG	899
Query	899	VAVATVVATATSPTS IKNLIEDLPGQSAVAASAANGEQDIQIAAVPAIVEEDEDEEEDFP	958
Sbjct	900	VAVATVVATATSPTS IKNLIEDLPGQ+AVAASAANGEQDIQIAAVPAIVEEDEDEEEEF	959
Query	959	EDDEED-HARADFDANGGDADGSDSDDVEAVDIVGYGHASTALNATFVKADSTETTTTTT	1017
Sbjct	960	++D+E HARADFDANGGDADGSDSDDVEAVDIVGYGHASTALNATFVKADSTETTTTTT	1019
Query	1018	PSTATTATTRHDDDEPEWLRDVLEAPKRSLENLLITSATSSRAPGQREELNGYDLHEKH	1077
Sbjct	1020	PSTATTATTRHDDDEPEWLRDVLEAPKRSLENLLITSATSSRA GQREELNGYDL EKH	1079

Query	1078	SDLNQTYITGGESLHESIVSVESTQSDATLNQTTTIDDSIISSKHNSTYSLADAEQATSS	1137
Sbjct	1080	SDLNQTY+TGGESLHESIVSVESTQSDATLNQTTTIDDSIISSKHNSTYSLADAEQAT+S	1139
Query	1138	TVLSTGVTELDSDSQQYYIPEYPPVRSKEVLVEAGVHYFEDGNFWMEVPGLLDFDDDDCSYP	1197
Sbjct	1140	TVLSTGVTELDSDSQQYYIPEYPPVRSKEVLVEAGVHYFEDGNFWMEVPGLLDFDDDDCSYP	1199
Query	1198	PITVRKNPKVRFSSGPIHVVSTFVSNDYDRNEDVDPVAASAEYELEKRVEKMHVFPVEL	1257
Sbjct	1200	PITVRKNPKVRFSSGPIHVVSTFVSNDYDRNEDVDPVAASAEYELEKRVEKMHVFPVEL	1259
Query	1258	MKGPEGLGLSIIIGMGVGADAGLEKLGIFVKTTITDNGAAARDGRIQVNDQIIIEVDGKSLVG	1317
Sbjct	1260	MKGPEGLGLSIIIGMGVGADAGLEKLGIFVKTTITDNGAAARDGRIQVNDQIIIEVDGKSLVG	1319
Query	1318	VTQAYAASVLRNTSGLVKFQIGRERDPENSEVAQLIRLSLQADREKEERLKRQQEEYLRR	1377
Sbjct	1320	VTQAYAASVLRNTSGLVKFQIGRERDPENSEVAQLIRLSLQADREKEERLKRQQEEYLRR	1379
Query	1378	TLDYSEDSTQPVSANSSVCEGPSSPVQVEHPMEVEATHSQEVESLKRLLQESMGCLVKE	1437
Sbjct	1380	TLDYSEDSTQPVSANSSVCEGPSSPVQVEHPMEVEATHSQEVESLKRLLQESMGCLVKE	1439
Query	1438	EIIQNLKRKLVKLETTGNENELLSERLRQSERELGNIRKEAANLQNMLQQSQGYMALDK	1497
Sbjct	1440	EIIQNLKRKLVKLETTGNENELLSERLRQSERELGNIRKEAANLQNMLQQSQGYMALDK	1499
Query	1498	KYNKAKRLVREYQQRELDMCHREEFYQQLLQEKDTEYNALVKKLKDRVINLEHELQETQR	1557
Sbjct	1500	KYNKAKRLVREYQQRELDMCHREEFYQQLLQEKDTEYNALVKKLKDRVINLEHELQETQR	1559
Query	1558	KAGFPVGLPYDSATLKLTPQMMRKTPPKPLFHKLETELSLTEISDLSPDGDGVKTATVER	1617
Sbjct	1560	KAGFPVGLPYDSATLKLTPQMMRKTPPKPLFHKLETELSLTEISDLSPDGDGVKTATVER	1619
Query	1618	KVPVKDELDAAVPQHELLDNSINKTKIDLANRQLPSANGNSSTSNGAAVDLGQLSNGNLL	1677
Sbjct	1620	KVPVKDELDAAVPQHELLDNS+NKTKIDLANRQLPSANGNSSTSNGAADLGQLSNGNLL	1679
Query	1678	KRSRSNSRSSDCTLDDTDEEEERESEALNLAGAPVAHETISLSNGNSHLLANVNNLLQHH	1737
Sbjct	1680	KRSRSNSRSSDCTLDDTDEEEERESEALNLAG PV HETISLSNGNSHLLANVNNLLQHH	1739
Query	1738	PPAMATVIATPSNGHLGTTTTPILLNSTSSASSSSSNQSTAREAQINQLYAQVHKDPSKQQ	1797
Sbjct	1740	PPAMA+V+ATPSNGHLGTTTTPILLNSTSSASSSSSNQSTAREAQINQLYAQVHKDPSKQQ	1799
Query	1798	HQQQQQQQQQAQAVTTSIPSIFKNALGSPADNGLNDFHRGSMTTFGTGPATSSNRDLNSS	1857
Sbjct	1800	-HQQQQQQQQQAQAVTTSIPSIFKNALGSPADNGLNDFHRGSMTTFGTGPATSSNRDLNSS	1858
Query	1858	YDSILGSNDKLAENDPAESWMPSSRRRVAPNGSKVPLPGSSFTDQLNQALSDRERRLGDG	1917
Sbjct	1859	YDSILGSNDKLAENDPAESWMPSSRRRVAPNGSKVPLPGSSFTDQLNQALSDRERRLGDG	1918

Query	1918	SSRHSSDDYTEINKSQSAAAINCKTLN	IRQAVNEAQPKVPWQQQHHQQIQQQPSAHTT	1977
Sbjct	1919	SSRHSSDDYTEINKSQSAAAINCKTLN	IRQAVNEAQPKVPWQQQHHQQIQQQPSAHTT	1978
Query	1978	GPPSPTSMSSGCSSPGYSPSRTL	DLGSSSSFS	2037
Sbjct	1979	GPPSPTSMSSGCSSPGYSPSRTL	DLGSSSSFS	2038
Query	2038	LMGIELERYIPVFKENNVEGGALL	TLDSKDFKTLGICGDDKHRLKKRLKDLKANIEKERK	2097
Sbjct	2039	LMGIELERYIPVFKENNVEGGALL	TLDSKDFKTLG+CGDDKHRLKKRLKDLKANIEKERK	2098
Query	2098	DM	2099	
Sbjct	2099	DM	2100	

As can be seen from the excellent correlation between our predicted protein in *D. erecta* against the equivalent protein in *D. melanogaster* in the Blastp, our prediction is most likely very accurate.

Appendix E

Which gene predictor matches the best to the BLASTX output?

Question 2

Which chromosome is it on in D. melanogaster? 4

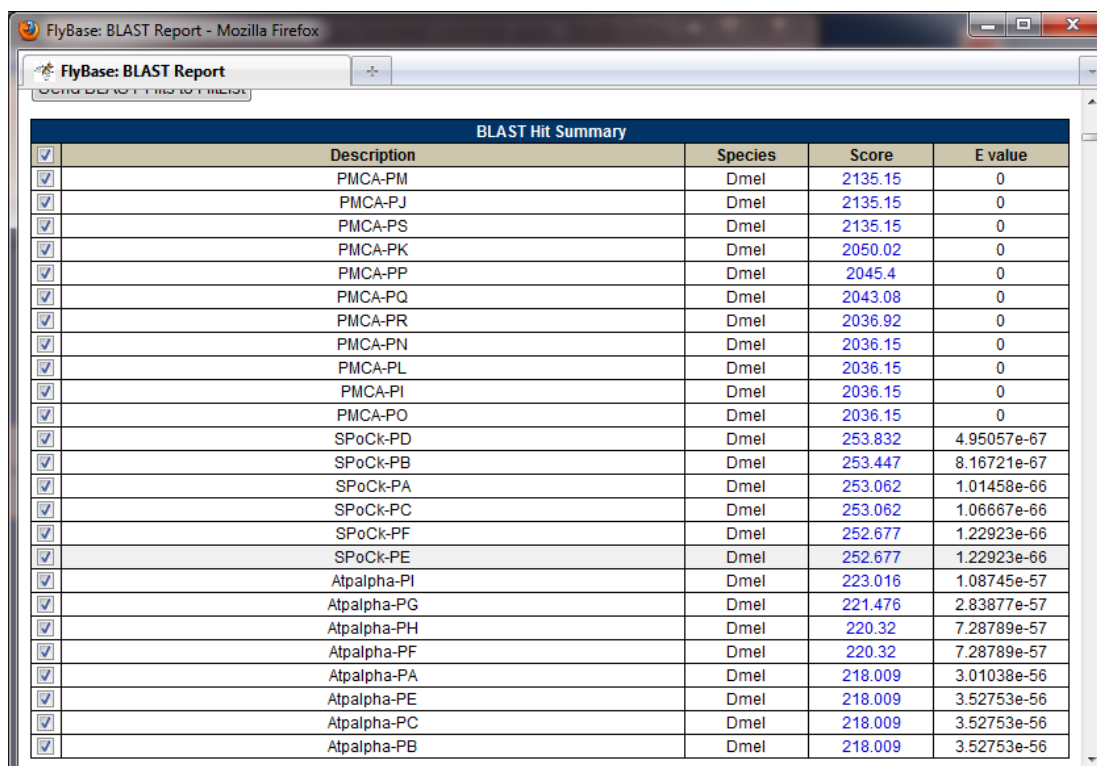
Why does the fact that the *D. melanogaster* gene is on this particular chromosome (and not on a different one) strengthen the case for this gene being an ortholog to the *D. grimshawi* gene?

Twenty-seven of 28 genes are shared between *D. melanogaster* and *D. virilis* in a region of the fourth (dot) chromosome that is shown in the introductory Power Point slides (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. The fosmid we are analyzing for this exercise contains dot chromosome DNA from *D. grimshawi*. 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*. Thus, the fact that the PMCA gene is on the 4th chromosome in *D. melanogaster* is additional evidence that this region of the *D. grimshawi* dot chromosome contains the PMCA ortholog.

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 high S scores by alignment of any two random sequences).



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

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 in *D. melanogaster*, which further decreases our confidence that they are present in this fosmid (which contains DNA from the 4th chromosome of *D. grimshawi*). Thus, the evidence indicates that it is the PMCA ortholog that is found in this region of *D. grimshawi* DNA.

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 Figure 4 on p. 27 of the Sample Annotation Problem and the table below, one can see that all but one (the P isoform) of the protein isoforms have CDS #22_... through #8_... in sequential order, with CDS #22_... being the first (5'-end) coding exon of the gene). However, as seen below, the other end of the protein varies across several isoforms.

CDS usage map:

Isoform	1	13_1157_0	12_1157_2	11_1157_0	10_1157_2	9_1157_0	8_1157_1	7_1157_0	6_1157_0	5_1156_0	3_1157_0	1_1157_0
PMCA-RO	Y	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RQ	Y	Y	Y	Y	Y	Y	Y	Y				Y
PMCA-RK	Y	Y	Y	Y	Y	Y	Y					Y
PMCA-RJ	Y	Y	Y	Y	Y	Y	Y			Y		
PMCA-RI	Y	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RR	Y	Y	Y	Y	Y	Y	Y	Y			Y	Y
PMCA-RP	Y	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RL	Y	Y	Y	Y	Y	Y	Y				Y	Y
PMCA-RS	Y	Y	Y	Y	Y	Y	Y			Y		
PMCA-RM	Y	Y	Y	Y	Y	Y	Y			Y		
PMCA-RN	Y	Y	Y	Y	Y	Y	Y		Y			

Instructor's Note*: The isoform numbers after the first underscore in the above screenshot and in the text and Figures 4 and 5 of the Sample Annotation Problem change concomitant with updates to FlyBase and the Gene Record Finder. These differences can be ignored. (The above screenshot was taken on May 29, 2012.)

Which mRNA isoforms code for identical protein isoforms?

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

What might the different protein isoforms tell you about the (minimum) number of stop codons that are used in the expression of this gene?

Isoforms O, Q, K, I, R, P and L include the 3'-most exon (#1_...), so there must be a stop codon somewhere in this exon. However, there must also be a stop codon in exon #5_..., since this is the final exon of isoforms J, S and M; isoform N ends with exon #6_..., which therefore must also have a stop codon. So, the minimum number of stop codons is three.

Given the above answer, why do you think there is not a protein isoform that includes all of the CDS?

All isoforms that include the final exon (#1_...) also omit CDS #6_... and #5_.... Since we know from the above answer that exons #6_... and #5_... have stop codons in them, the only way that a protein can be made that includes more 3' coding sequences (e.g., #3_... and/or #1_...) is to omit (by alternative splicing) exons #6_... and #5_....

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 and shrink the margins and/or font to keep the sequences in alignment). 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	LSGSKADEEHRRETFGSNVIPPCKPKTFLTLVWEALQDVTLLIILEVAALVSLGLSFYKPA	3436
		LSGSKADEEHRRETFGSNVIPPCKPKTFLTLVWEALQDVTLLIILEVAALVSLGLSFYKPA	
Sbjct	1	LSGSKADEEHRRETFGSNVIPPCKPKTFLTLVWEALQDVTLLIILEVAALVSLGLSFYKPA	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

Alignment on next page:

```

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

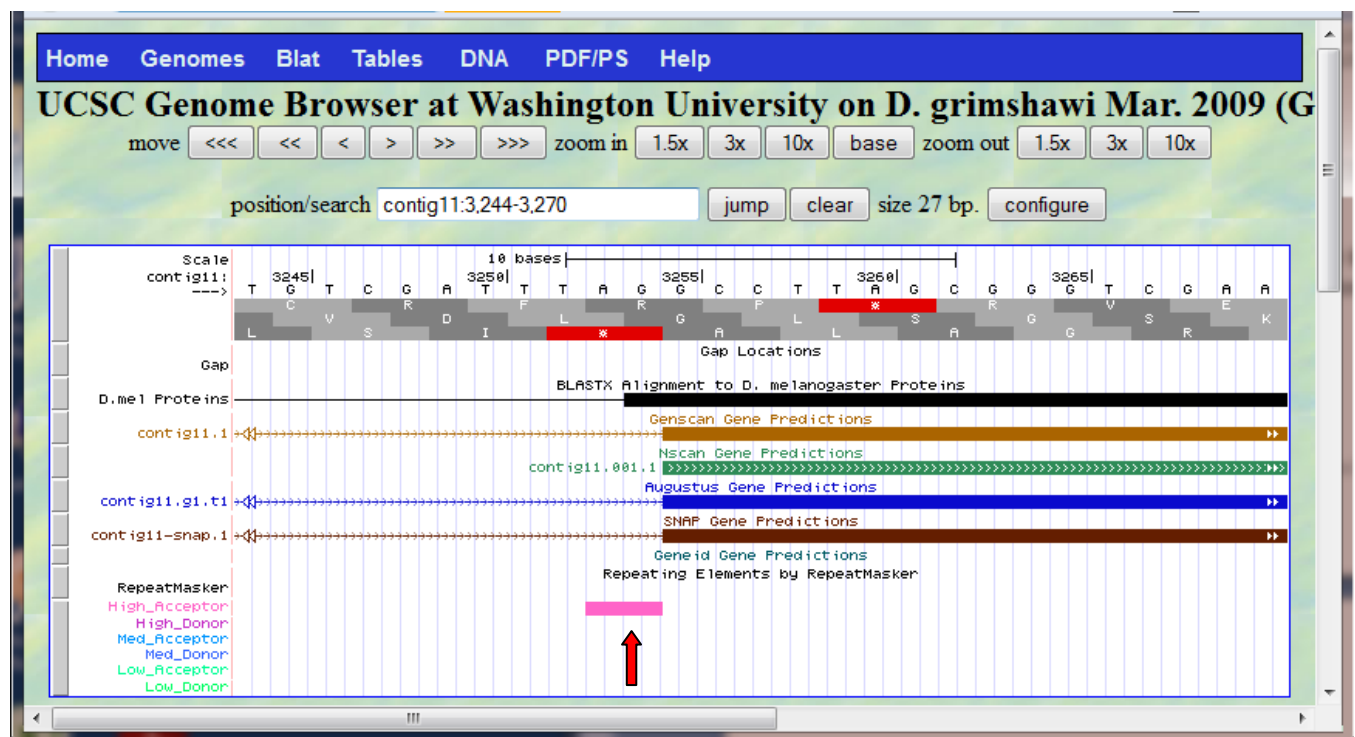
Query  4153  CQISVGDILVGDIAQIKYGDLLPADGCLIQSNDLK  4257
          CQISVGDILVGDIAQ+KYGDLLPADGCLIQSNDLK
Sbjct   61   CQISVGDILVGDIAQVKYGDLLPADGCLIQSNDLK  95

```

Question 6

Look around the region where the alignment to CDS 21_...* (the second exon) begins. How many acceptor sites can you find? **See Instructor's Note on page 3.*

The UCSC Genome Browser window for this area of the fosmid is shown below. The second exon begins at around base 3257. There are two acceptor sites in this region: the “AG” at 3253-4 and the “AG” at 3260-1.



Considering the frame of the conserved amino acids you found in question 5, what is the phase of each putative acceptor site you find?

From the exon by exon blastx search results on page 4, we know that the reading frame with the conserved amino acid sequence for exon 21 is Frame +2 of the fosmid. The Frame +2 amino acid sequence is the second row of light and dark gray boxes in the above screenshot. Note that in Frame +2, the 3253-4 acceptor results in the exon beginning two bases (GC) before the first complete codon (for the L amino acid); 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).

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?

Since the intron donor sequence after the end of the first exon creates a phase 1 exon (see page 31 of the Sample Annotation Problem), 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.

Itemize what other evidence you could consider if you have two or more possible donor/acceptor pairs.

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 (see red arrow in previous screen shot).
3. For any combinations that are indistinguishable by the above criterion, by convention, the pair that creates the longest protein should be picked.

Finally, record the base coordinates for the first exon and the beginning of the second exon as deduced from your complete analysis.

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, e.g. the first base after the ‘AG’ that ends the intron).

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.