Annotation Practice Activity

[Based on materials from the GEP Summer 2010 Workshop]

Special thanks to Chris Shaffer for document review

Parts A-G

Introduction:

A genome is the total genetic content of an organism. In order to study a genome, DNA is isolated from a convenient tissue source, digested with a battery of restriction enzymes or physically broken into 1-2 kb DNA fragments and cloned into appropriate vectors. The total of all the cloned fragments is called a genomic library. Small regions of the genome with many overlapping fragments are called "contigs," which derives from a "contig"uous set of overlapping DNA sequences that can be ordered into a complete sequence.

This activity will use contig 36 of the D. erecta genomic library to gain experience in the annotation process of identifying gene(s) within a contig sequence.

Procedures:

Part I: These activities will provide practice in using Gene Record Finder, NCBI BLAST, and the UCSC Genome Browser.

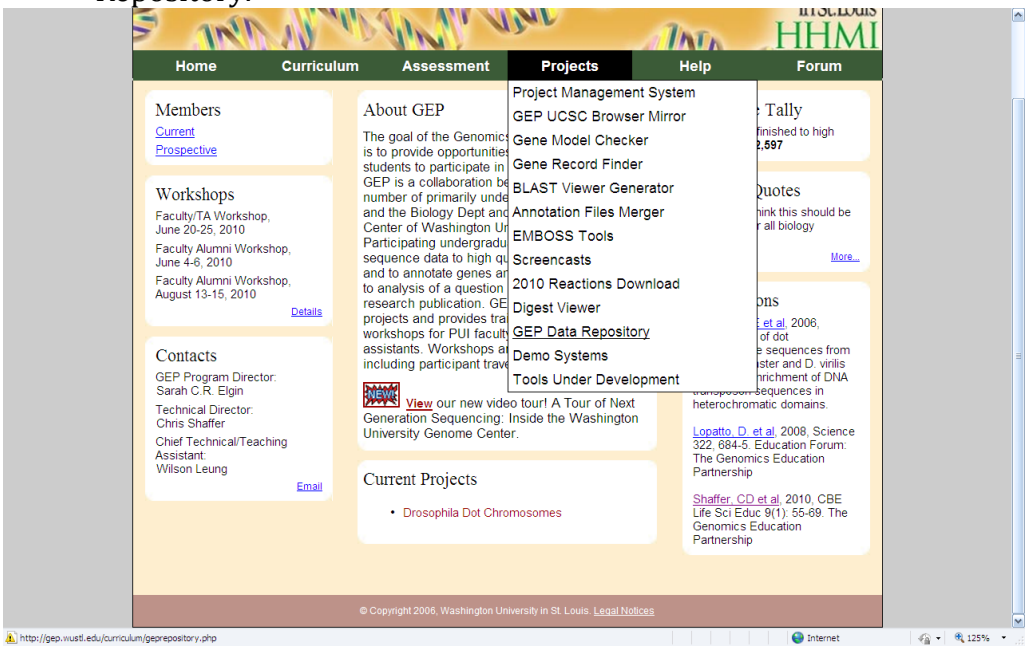
A. Open the following websites:

http://gep.wustl.edu/

http://gander.wustl.edu/

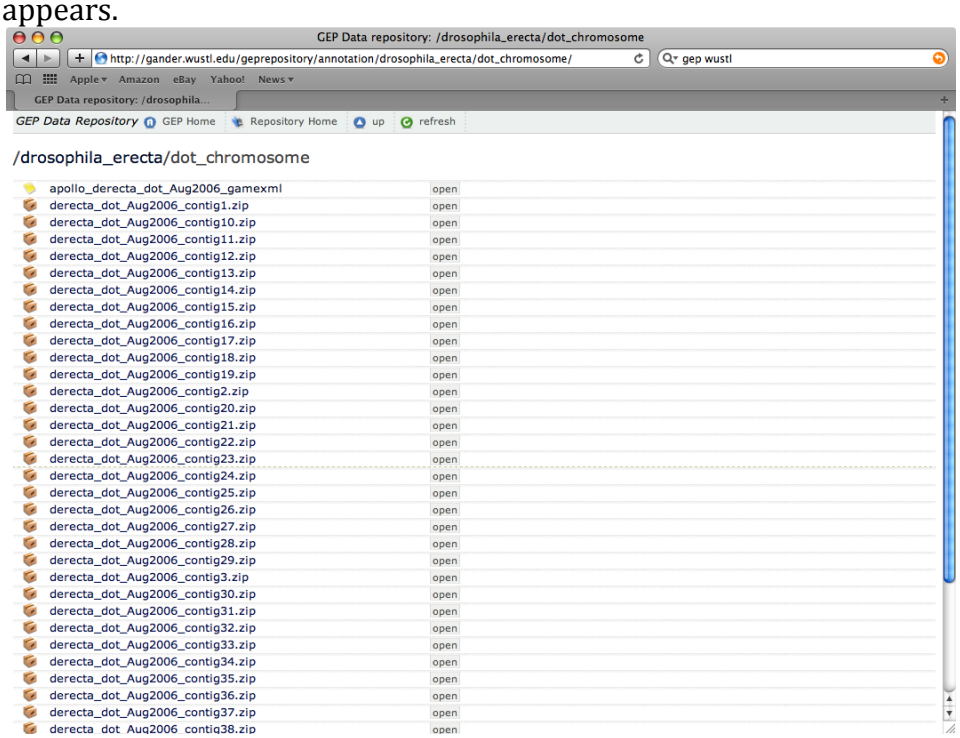
http://blast.ncbi.nlm.nih.gov/Blast.cgi

B. Go to http://gep.wustl.edu/ and select Projects and then select GEP Data Repository.



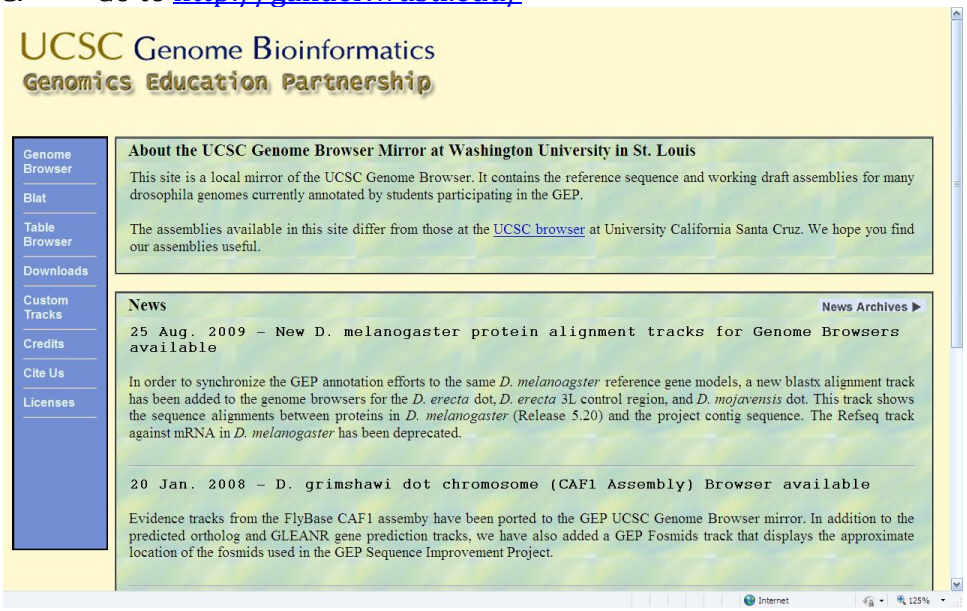
Click on the GEP Data Repository; the following window appears:

Select directory: drosophila_erecta/dot_chromosome. The following window appears.



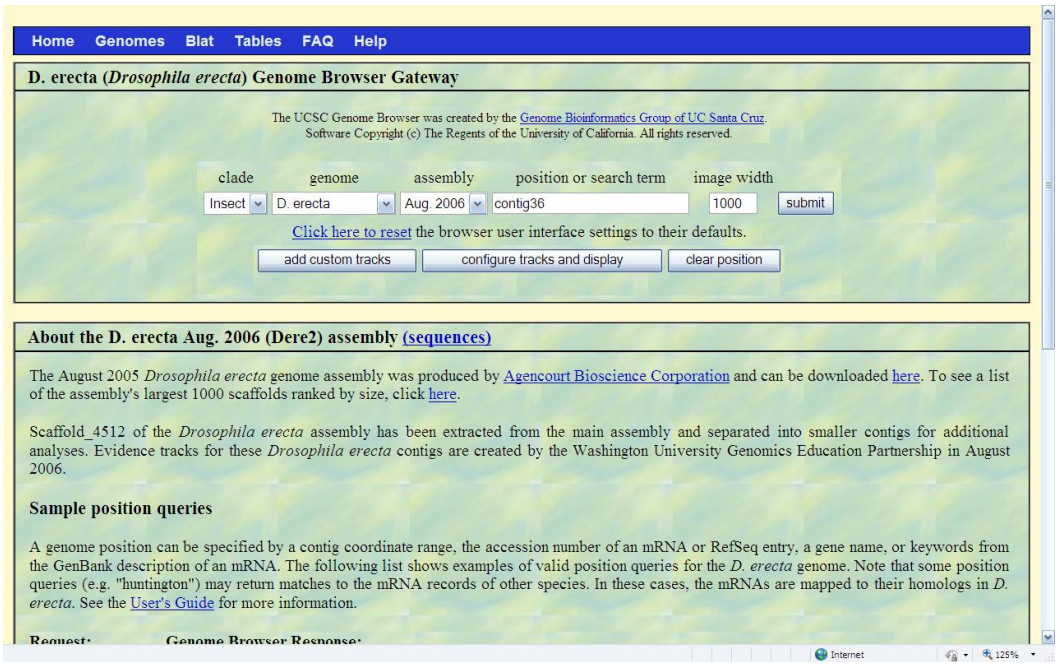
Select "derecta contig 36", which downloads file to desk top. Right click and select "Extract All"

C. Go to http://gander.wustl.edu/

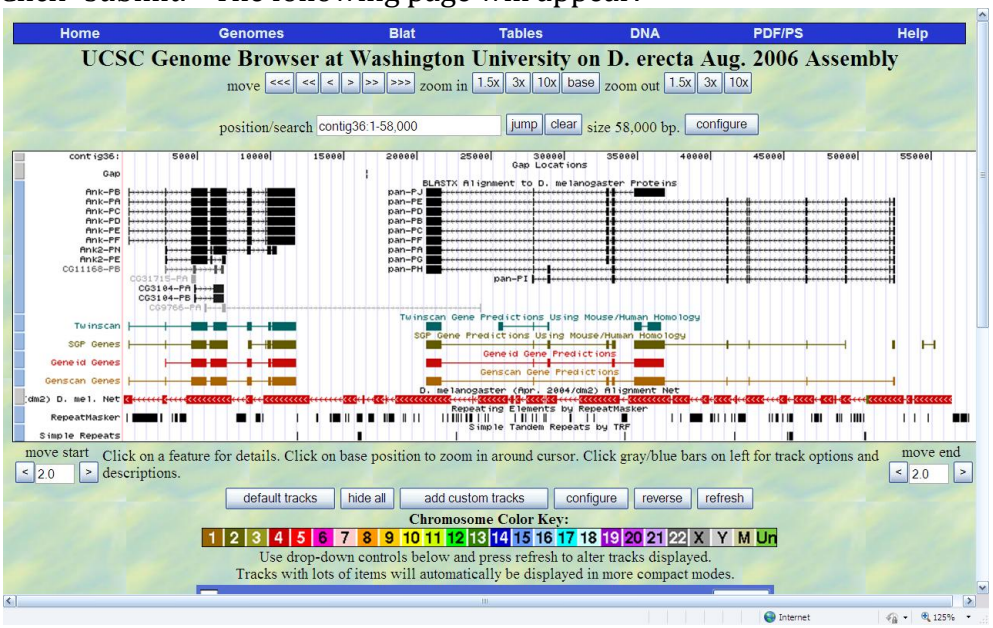


Click on Genome browser on the left side of the screen.

In the genome pull-down window, select D. erecta; in the assembly pull down window select August 2006 assembly; enter contig 36 in the “position or search window.

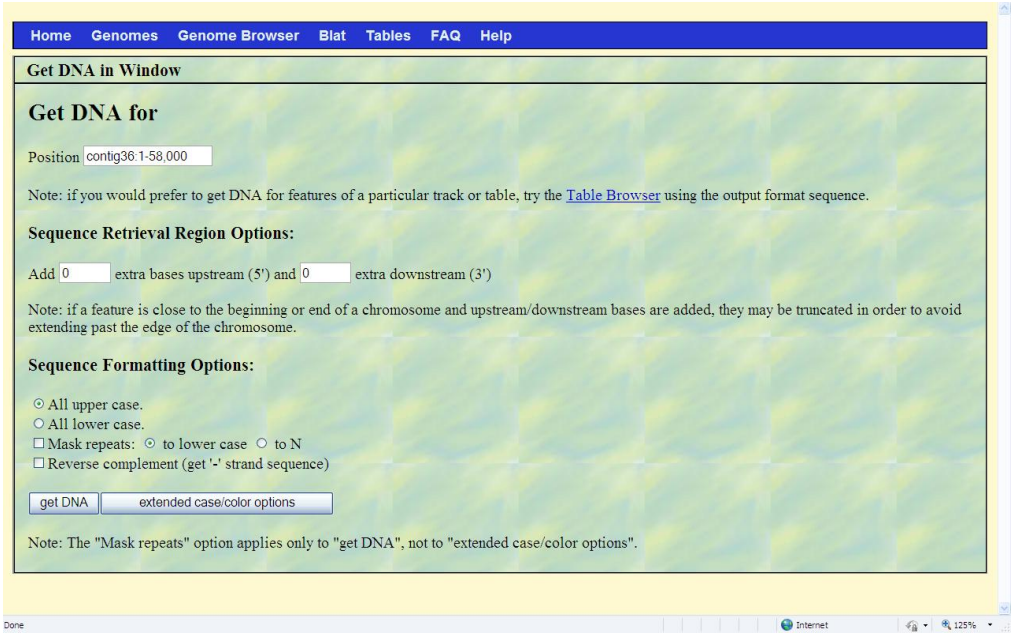


Click “submit.” The following page will appear:

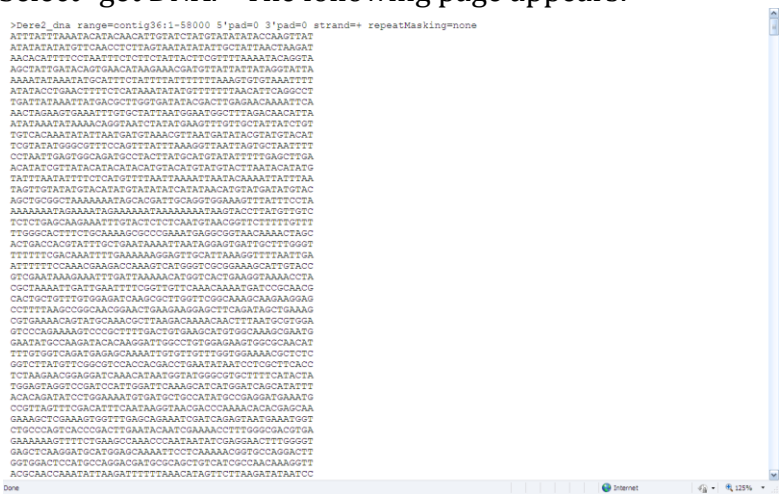


The Genome Brower window will display one or more gene models that prediction programs have identified in the contig. (Note: your image may not look exactly like this depending on your settings. Feel free to experiment with the menu’s found below the browser image to control what is seen in the image and exactly how much detail you are shown).

Select DNA at the right top of the page. The following page appears:



Select "get DNA." The following page appears:



Copy the entire sequence [it is very long—56,000 NT]; this will be pasted into a blastx window later.

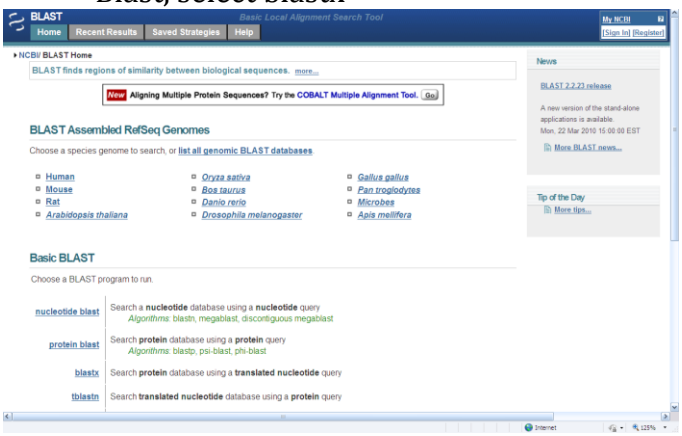
Brief Notes:

- The Genome browser will find closely related D. melanogaster genes to sequences contained in contig 36.
- Twinscan, SGP, Gene ID Genes and Genscan Genes are different computer programs that create gene models.
- Each region of the image shows in graphical form the results of some computer program that was run on the Contig36 sequence. The black boxes in the "BlastX alignment" section of the image show regions that are very similar to *Drosophila melanogaster* proteins. The exact proteins in this case are Ank and pan. Below this are the results of the gene prediction programs.

The red boxes show regions of high similarity between the DNA sequence of contig 36 and the genome of *Drosophila melanogaster*.

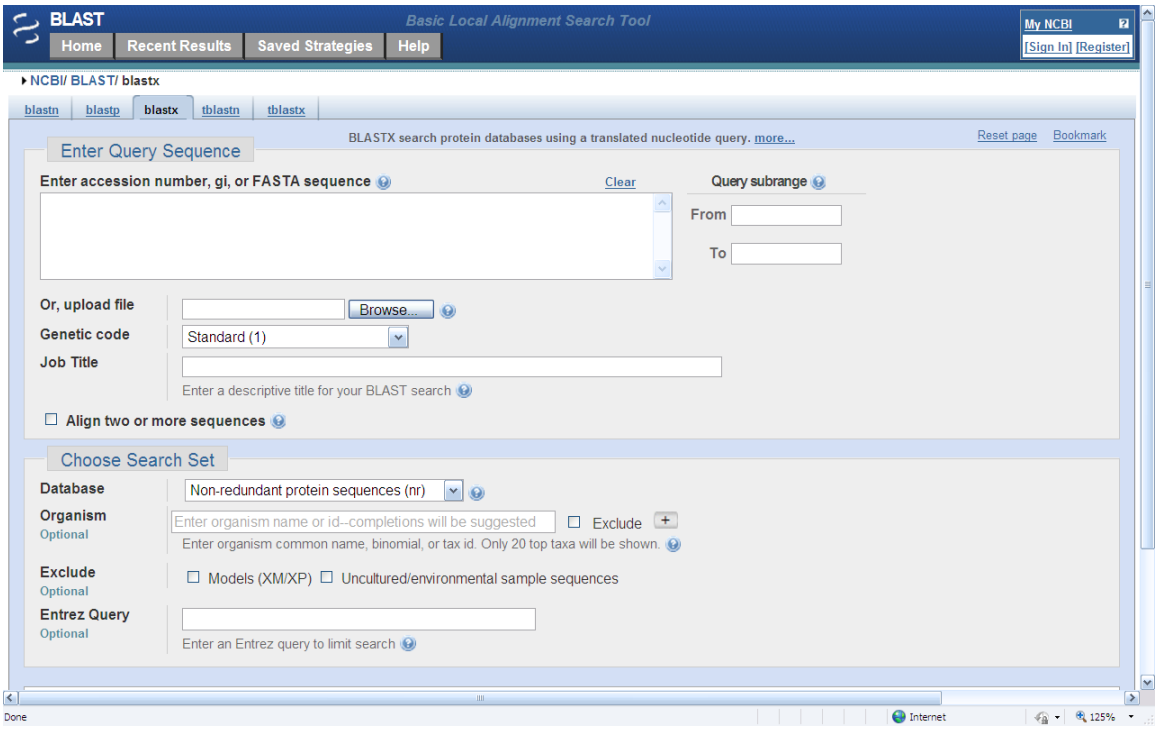
- Please note that some genes are in the reverse orientation!
- Since D. erecta is evolutionarily close to the very well annotated D. melanogaster, we will rely on the the coding sequence of D. melanogaster to help us annotate genes in D. erecta.
- Select the "pan" gene for annotation. To find the parts of contig 36 that match the D. melanogaster pan gene, use Gene Record Finder to get the entire coding sequences of each exon of the pan gene, and these will be used to annotate the pan gene and its isoforms in D. erecta.

D. Open NCBI BLAST [http://blast.ncbi.nlm.nih.gov/Blast.cgi] and under Basic Blast, select blastx



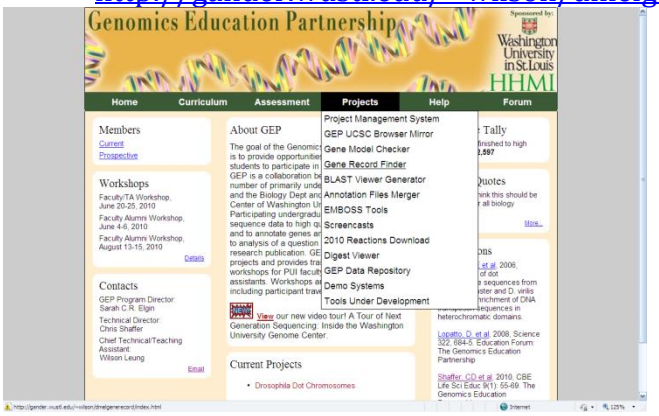
Select the box: align two or more sequences.

In the Enter Query Sequence, paste in the DNA sequence of contig36 copied from the UCSC Genome Browser.

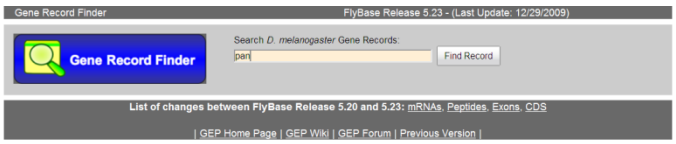


At bottom of the page, click Algorithms and under Filters and Masking, turn off low complexity filter [this give more evidence to work with]. Note: The default threshold is 10, which means the chance you would get a match to sequence by random; if blastx gives too many results, then go back and change threshold to a larger number. **Do not** blast yet!

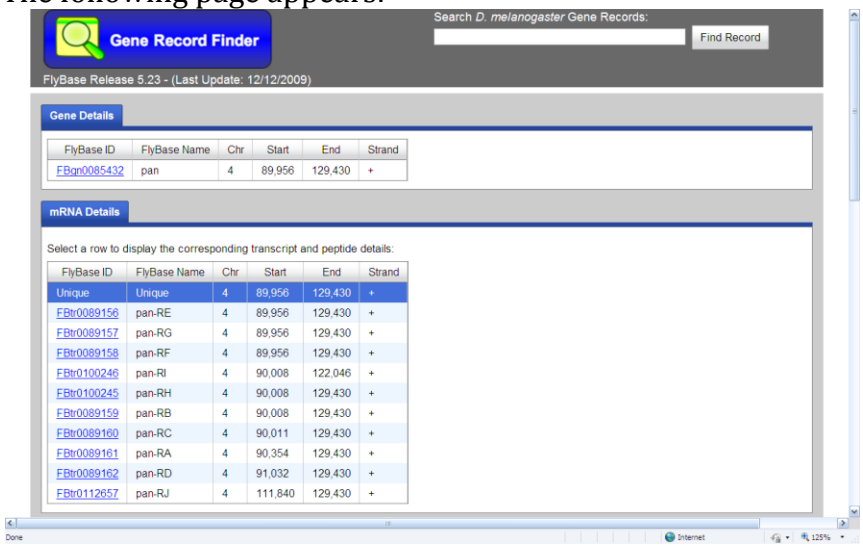
E. In a separate window, return to the GEP Home Page. Use the Project pull-down window and select "Gene Record Finder" [or go to: http://gander.wustl.edu/~wilson/dmelgenerecord/index.html].



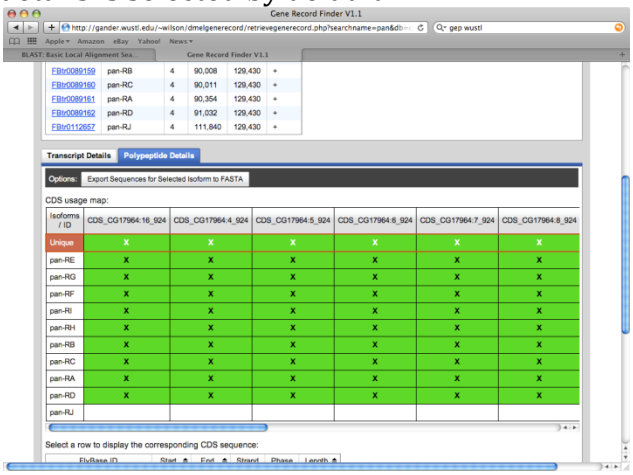
Type "pan" in the window and click "Find Record."



The following page appears:



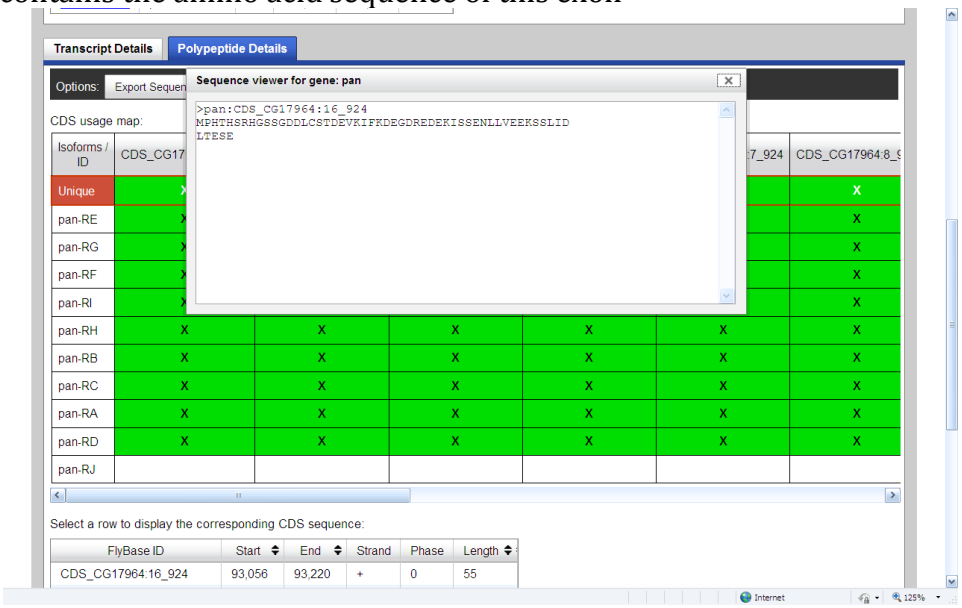
Scroll down the page to transcript details/polypeptide details. The polypeptide details is selected by default.



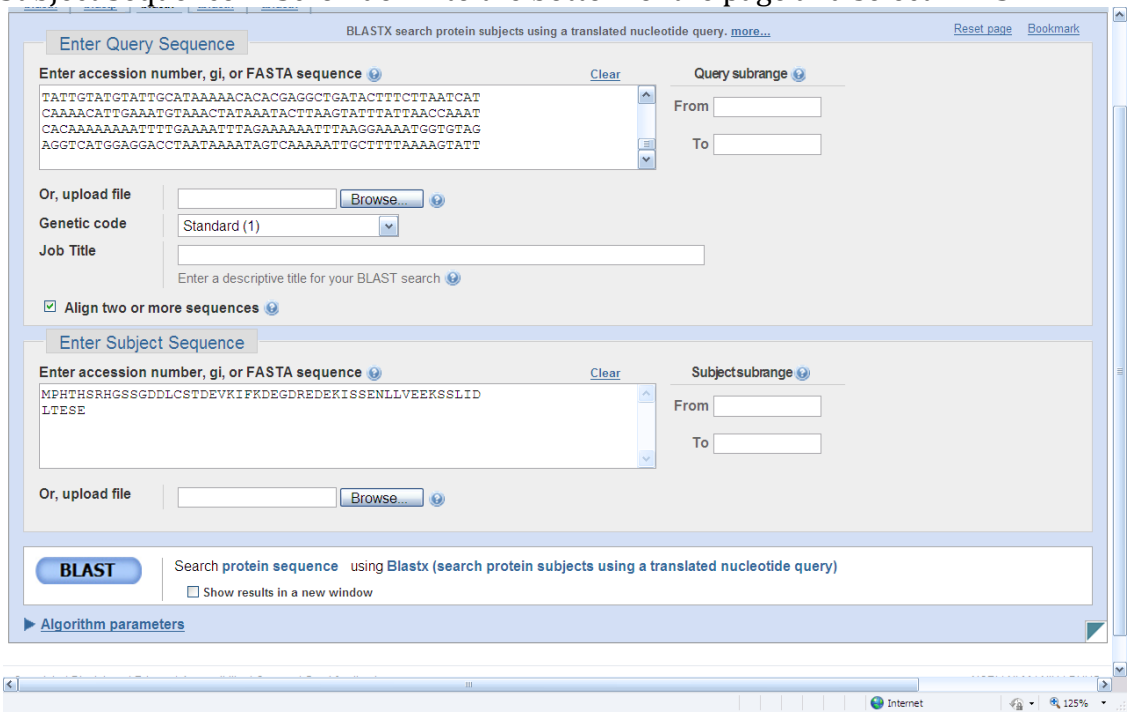
Isoforms are alternately spliced genes.

Scroll across the page to see which exons vary and which ones are similar among the isoforms.

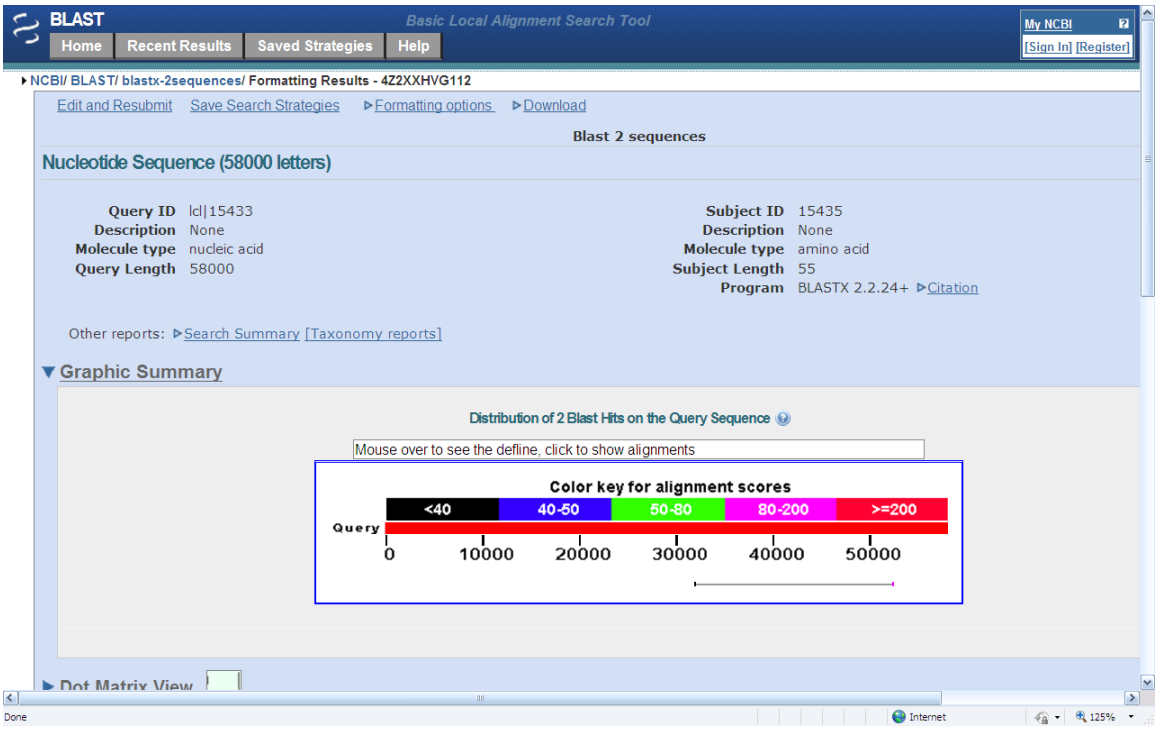
Under the CDS usage map, select the first grey box: a in a pop-up box appears that contains the amino acid sequence of this exon



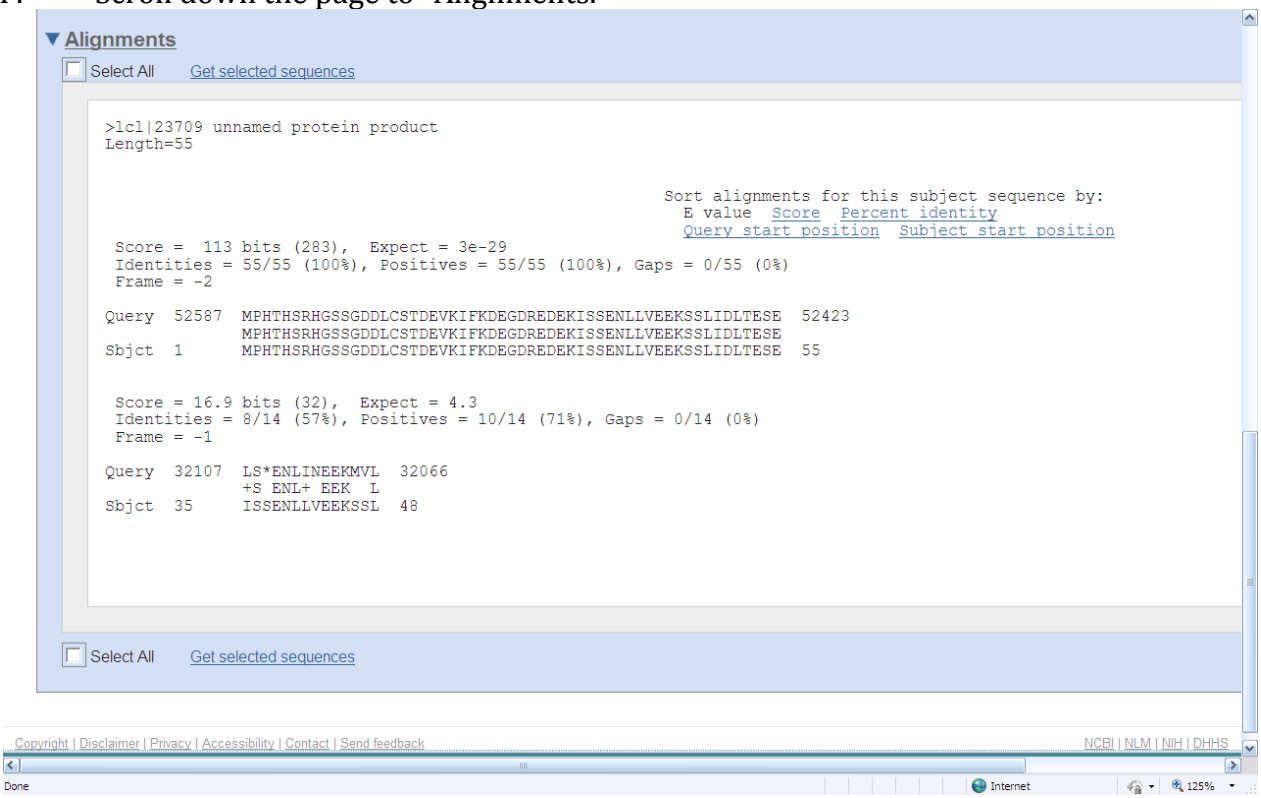
Copy this amino acid sequence and paste into the blastx window in the "Enter Subject Sequence." Scroll down to the bottom of the page and select BLAST.



Select BLAST. The following window appears:



F. Scroll down the page to "Alignments."



Brief Notes:

- The first alignment block (starts at base position 52587 in your query i.e. contig36) shows an E value is 3e-29, which means there is a very low chance of getting this match by chance alone. The second query data shows an E value of 4.3 which means these data are useless.
- The number of amino acids copied was 55. A match of 55/55 is excellent for a small exon.
- Note that the reading frame is -2, which indicates the pan gene is in the reverse orientation (i.e., it is on complementary strand to the 5'->3' strand].

- Below is an enlarged view of Query 52587

```
Query 52587  MPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDL
TESEMPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE 52423
Sbjct  1   MPHTHSRHGSSGDDLCSTDEVKIFKDEGDREDEKISSENLLVEEKSSLIDLTESE  55
```

- Query indentifies the nucleotide number of contig 36 where this alignment to exon 1 begins (52587) and where it ends (52423]. Given the high level of similarity and the fact that there is really nothing better, we have found the probable position of exon 1 in our D erecta contig36.

- Sbjct 1 is the amino acid sequence of exon 1 which you put in the bottom box.

[[Category:Faculty Resources]]