BIOL 419/519 Bioinformatics Research
Introduction to Gene Models
Paul Szauter – June 18, 2013

Introduction

In this exercise, we will use a custom installation of the UCSC Genome Browser at GEP to view a *D. ananassae* fosmid. The UCSC Genome Browser allows a large variety of different data types to be displayed against a physical map of a fosmid or contig. We will be using the UCSC Genome Browser extensively for annotation. In this exercise, we will use it to build our first gene model.

1. Download the fosmid sequence and associated files

Use the link on the **Class Projects** page to locate the 3L control fosmids from *Drosophila ananassae*. Download the folder **dananassae_3Lcontrol_Jan2013_fosmid_1475K17**. This is already on the desktop of the classroom Mac laptops.

The folder contains another folder named *src* that contains the sequence of the fosmid. It also contains a folder named *analysis* that has summaries of various kinds of analysis.

2. Finding *Hem* on fosmid 1475K17

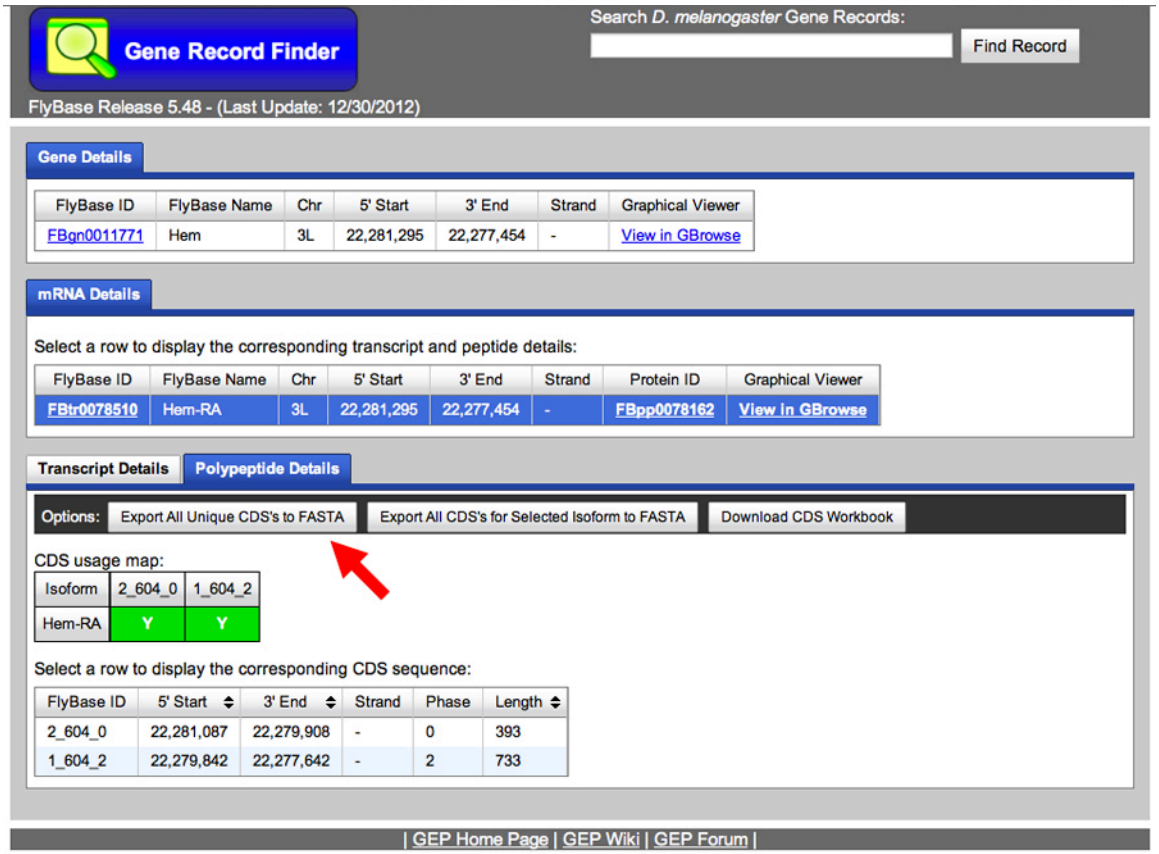
In our earlier BLASTX analysis, we found two segments of the *D. melanogaster Hem* protein that align to segments of the fosmid, as shown on the course website:

http://www.discoveryandinnovation.com/bioinformatics/results2/Szauter/szauter_report1.html#Hem

The protein sequence NP_524214.1 (*D. melanogaster Hem* protein) aligns to the six-frame translation of the fosmid in two segments, 4943-3759 in frame -1, and 3703-1520 in frame -2. This seems to indicate that the *D. melanogaster* Hem protein has at least two exons. We can use the Gene Record Finder to get the coding sequences of the *D. melanogaster* Hem protein. Navigate to the **Gene Record Finder at GEP** using the Tools page on the course website:

http://gander.wustl.edu/~wilson/dmelgenerecord/index.html

Enter `Hem` and click **Find Record**. Your screen should look like the screenshot on the next page.



This shows that the *D. melanogaster Hem* gene has a single isoform with two CDSs (two coding exons). Click the link to Export All Unique CDS's in FASTA. This gives us:

```
>Hem:2_604_0
MARPIFPNQQKIAEKLIILNDRGLGILTRIYNIKKACGDTKSKPGFLSEK
SLESSIKFIVKRFPNIDVKGLNAIVNIKAEIIKSLSLYYHTFVDLLDFKD
NVCELLTTMDACQIHLDITLNFELTKYYLDLVVTYVSLMIVLSRVEDRKA
VLGLYNAAYELQNNQADTGFPRLGQMILDYEVPLKKLAEEFIPHQRLLTS
ALRSLTSIYALRNLPADKWREMQKLSLVGNPAILLKAVRTDTMSCEYISL
EAMDRWIIFGLLLNHQMLGQYPEVNKIWLSALESSWVVALFRDEVLQIHQ
YIQATFDGIKGYSKRIGEVKEAYNTAVQKAALMHRERRKFLRTALKELAL
IMTDQPGLLGPKAIFIFIGLCLARDEILWLLRHNDNPPLLKNK


>Hem:1_604_2
KSNEDLVDRQLPELLFHMEELRALVRKYSQVMQRYYVQYLSGFDATDLNI
RMQSLQMCPEDESIIFSSLYNTAAALTVKQVEDNELFYFRPFRLDWFRLQ
TYMSVGKAALRIAEHAELARLLDSMVFHTRVVDNLDEILVETSDLSIFCF
YNKMFDDQFHMCLEFPAQNRYIIAFPLICSHFQNCTHEMCPEERHHIRER
SLSVVNIFLEEMAKEAKNIITTICDEQCTMADALLPKHCAKILSVQSARK
KKDKSKSKHFDDIRKPGDESYRKTREDLTTMDKLHMALTELCFAINYCPT
VNVWEFAFAPREYLCQNLEHRFSRDLVGMVMFNQETMEIAKPSELLASVR
AYMNVLQTVENYVHIDITRVFNNCLLQQTQALDSHGEKTIAALYNTWYSE
VLLRRVSAGNIVFSINQKAFVPISPEGWVPFNPQEFSDLNELRALAELVG
PYGIKTLNETLMWHIANQVQELKSLVSTNKEVLITLRTSFDKPEVMKEQF
KRLQDVDRVLQRMTIIGVIICFRNLVHEALVDVLDKRIPFLLSSVKDFQE
HLPGGDQIRVASEMASAAGLLCKVDPTLATTLKSKKPEFDEGEHLTACLL
MVFVAVSIPKLARNENSFYRATIDGHSNNTHCMAAAINNIFGALFTICGQ
SDMEDRMKEFLALASSSLLRLGQESDKEATRNRESIYLLLDEIVKQSPFL
TMDLLESCFPYVLIRNAYHGVYKQEQILGLAL*
```

Copy these sequences to a text file and save the file.

Use the links on the **Tools** page to navigate to BLASTX at NCBI.

Click the link to **Align two or more sequences**.

Upload the fosmid sequence as the Query sequence.

Paste the first CDS sequence in the bottom box as the Subject sequence and click BLAST.

Record the coordinates of the fosmid that align with the first CDS.

Repeat this for the second CDS.

Did you get 4973 – 3765 in frame -1 for exon 1 and 3700 – 1520 in frame -2 for exon 2?

3. *D. ananassae* fosmid 1475K17 in the UCSC Genome Browser at GEP

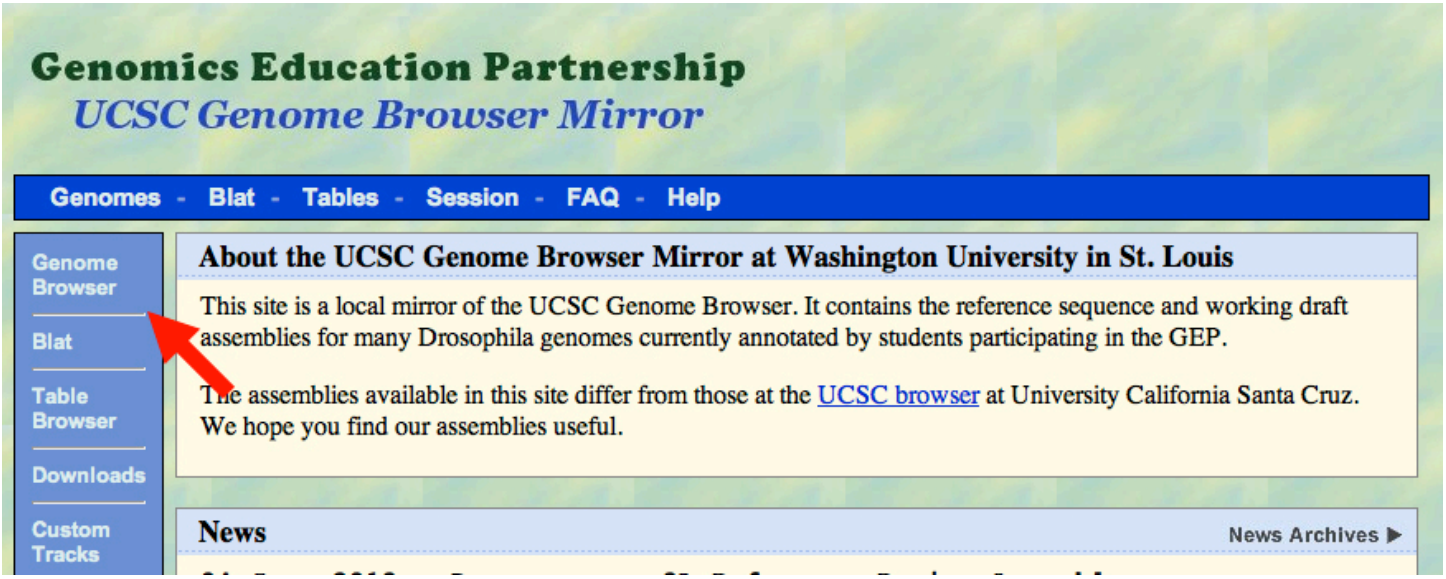
On the course website, navigate to the **Tools** page:

http://www.discoveryandinnovation.com/bioinformatics/tools.html

Click the link to the UCSC Genome Browser at GEP to go to:

http://gander.wustl.edu/

In the page that appears (shown below), click the **Genome Browser** link.



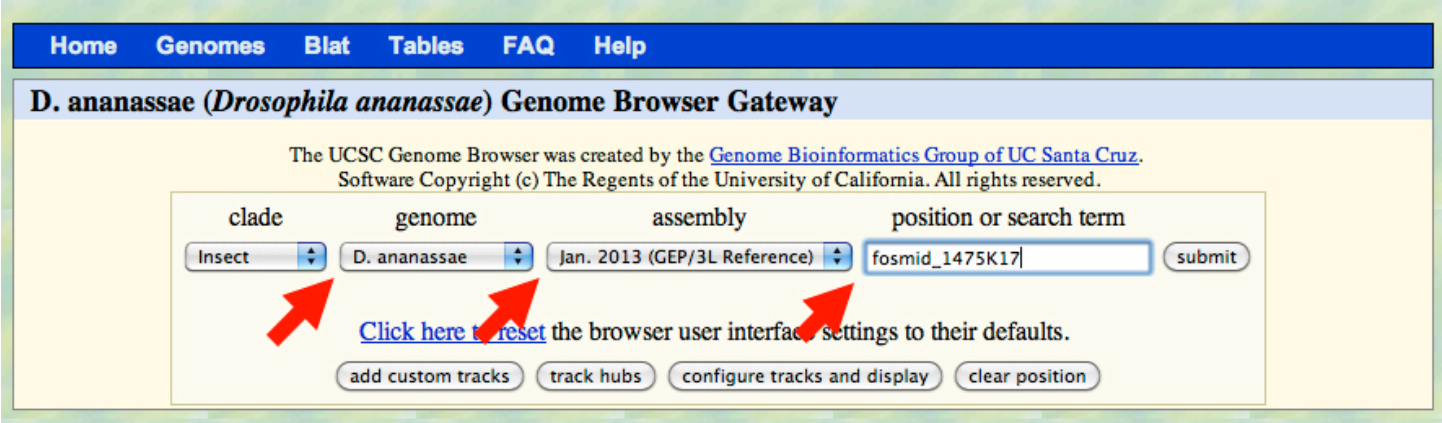
In the page that appears (shown below):

1. Select **D. ananassae** as the **Genome**.

2. Select **Jan. 2013 (GEP/3L Reference)** as the **assembly**.

3. Enter `fosmid_1475K17` as the **position or search term**.

4. Click **Submit**.



In the page that appears, there will be a visual display with many control buttons. Set the controls as described below.

1. In **Mapping and Sequencing Tracks**, set **Base Position** to full, all other tracks to **hide**.

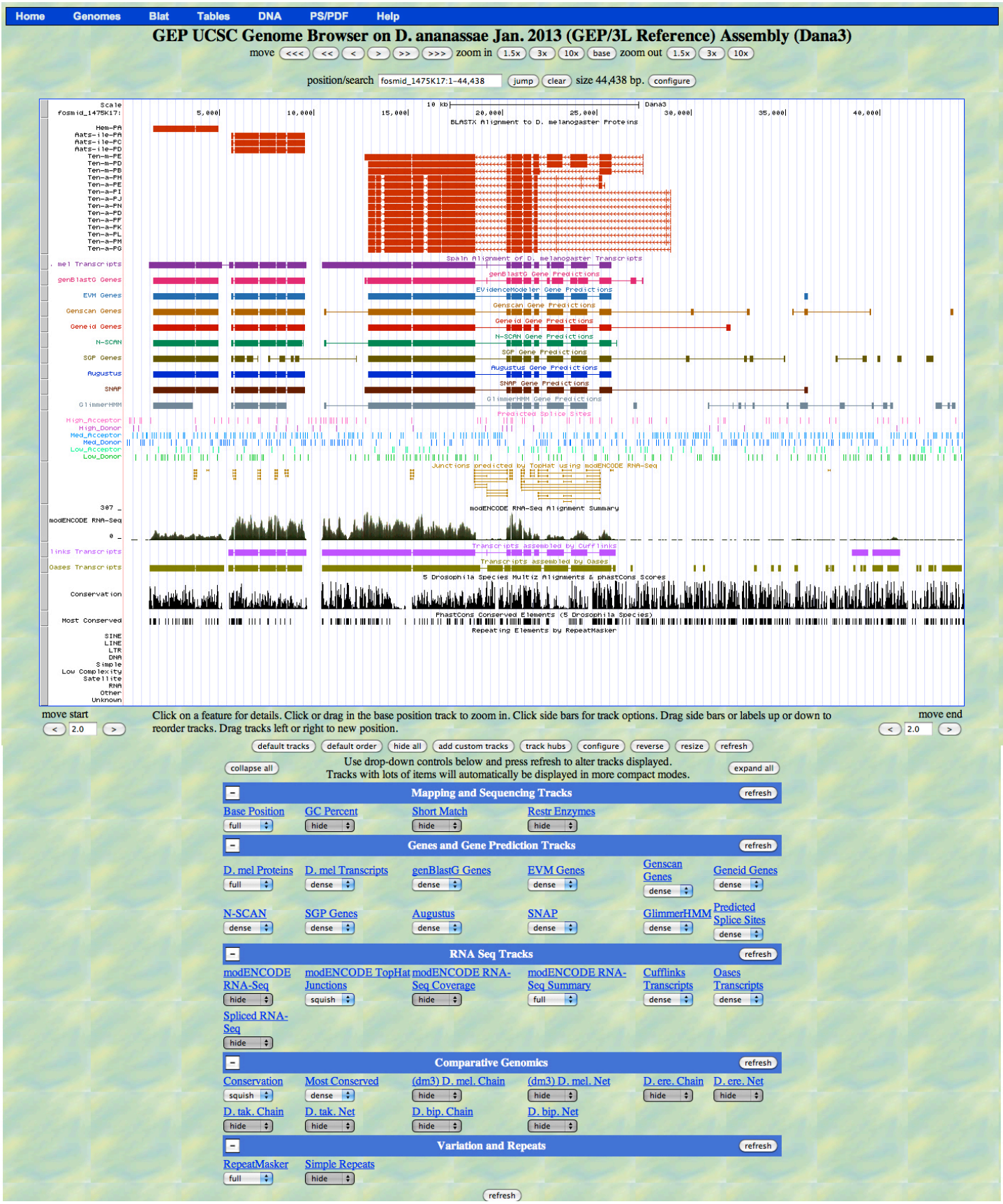
2. In **Genes and Gene Prediction Tracks**, set all to **dense**.

3. In **RNA Seq Tracks**, set **modENCODE TopHat** junctions to **squish**, **modENCODE RNA-Seq Summary** to **full**, **Cufflinks Transcripts and Oases Transcripts** to **dense**, and all others to **hide**.

4. In **Comparative Genomics**, set **Conservation** to **squish**, **Most Conserved** to **dense**, and all others to **hide**.

5. In **Variation and Repeats**, set **RepeatMasker** to **full** and **Simple Repeats** to **hide**.

Your display should now resemble the image on the next page.


Home
Genomes
Blat
Tables
DNA
PS/PDF
Help
GEP UCSC Genome Browser on D. ananassae Jan. 2013 (GEP/3L Reference) Assembly (Dana3)
move
zoom in
zoom out
position/search
fosmid_1475K17:1-44,438
jump
clear
size 44,438 bp.
configure
BLASTX Alignment to D. melanogaster Proteins
Spaln Alignment of D. melanogaster Transcripts
genBlastG Gene Predictions
EVidenceModeler Gene Predictions
Genscan Gene Predictions
Geneid Gene Predictions
N-SCAN Gene Predictions
SGP Gene Predictions
Augustus Gene Predictions
SNAP Gene Predictions
GlimmerHMM Gene Predictions
Predicted Splice Sites
Junctions predicted by TopHat using modENCODE RNA-Seq
modENCODE RNA-Seq Alignment Summary
Transcripts assembled by Cufflinks
Transcripts assembled by Oases
5 Drosophila Species Multiz Alignments & phastCons Scores
PhastCons Conserved Elements (5 Drosophila Species)
Repeating Elements by RepeatMasker
move start
move end
Click on a feature for details. Click or drag in the base position track to zoom in. Click side bars for track options. Drag side bars or labels up or down to reorder tracks. Drag tracks left or right to new position.
default tracks
default order
hide all
add custom tracks
track hubs
configure
reverse
resize
refresh
collapse all
Use drop-down controls below and press refresh to alter tracks displayed.
Tracks with lots of items will automatically be displayed in more compact modes.
expand all
Mapping and Sequencing Tracks
Base Position
GC Percent
Short Match
Restr Enzymes
Genes and Gene Prediction Tracks
D. mel Proteins
D. mel Transcripts
genBlastG Genes
EVM Genes
Genscan Genes
Geneid Genes
N-SCAN
SGP Genes
Augustus
SNAP
GlimmerHMM
Predicted Splice Sites
RNA Seq Tracks
modENCODE RNA-Seq
modENCODE TopHat Junctions
modENCODE RNA-Seq Coverage
modENCODE RNA-Seq Summary
Cufflinks Transcripts
Oases Transcripts
Spliced RNA-Seq
Comparative Genomics
Conservation
Most Conserved
(dm3) D. mel. Chain
(dm3) D. mel. Net
D. ere. Chain
D. ere. Net
D. tak. Chain
D. tak. Net
D. bip. Chain
D. bip. Net
Variation and Repeats
RepeatMasker
Simple Repeats

Use shift-drag to zoom in on the first exon of *Hem* (on the right). Use the zoom in and move buttons at the top of the display to get a view like that shown below.

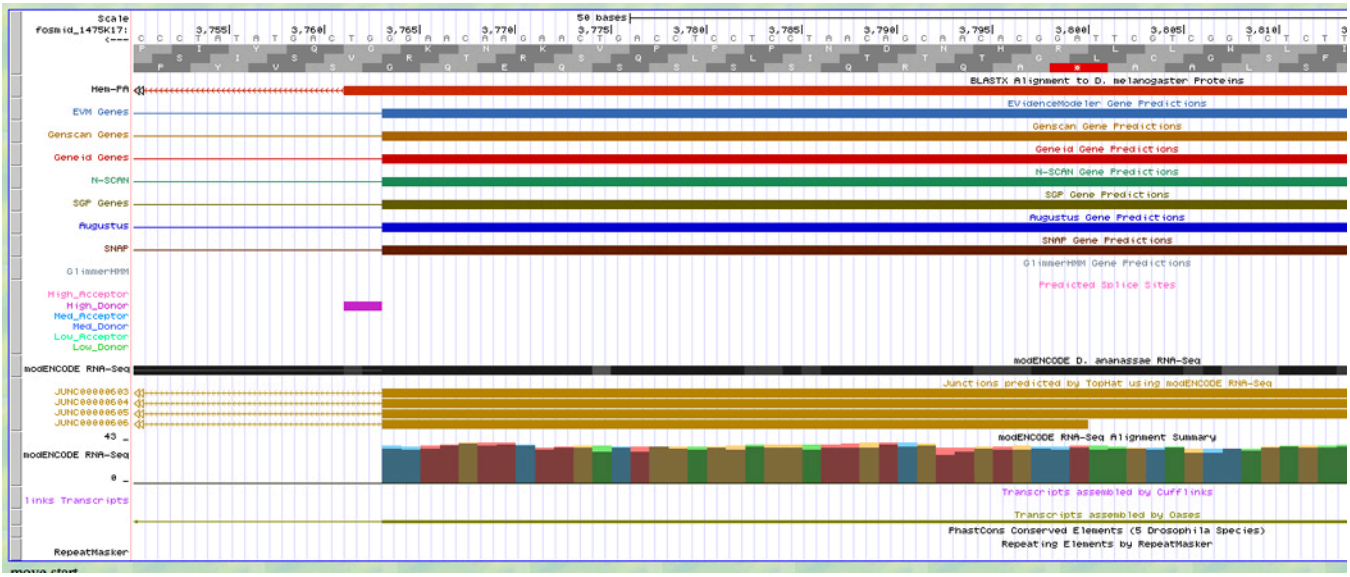
Make sure to click the arrow in the upper left part of the display so that the arrow is pointing to the left (minus strand). Notice the codon tracks below the coordinates. You can see three frames. MET (start) codons are shown in green, STOP codons are shown in red.

The top track is frame -1, the middle track is -2, and the bottom track is -3. See if you can find an open reading frame (no STOP codons) above the first exon. How can you make sure that you are correct?

Use shift and zoom to enlarge the 5' end of the first exon (on the right of the image above) until your screen matches the screenshot below.

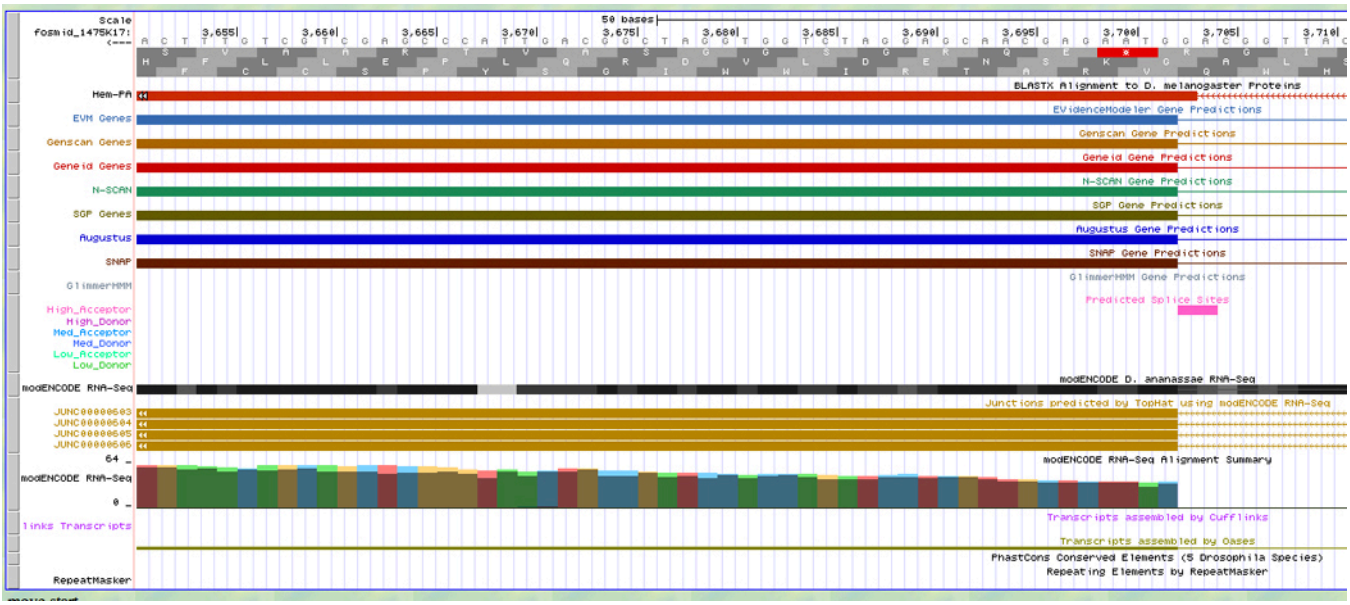
You can now read the bases as well as the amino acid sequence. Notice the green M corresponding to the ATG (backwards). You need to record the coordinate of the A in the ATG. It is 4943.

Navigate to the 3’ end of the first exon until your screen resembles that shown below.



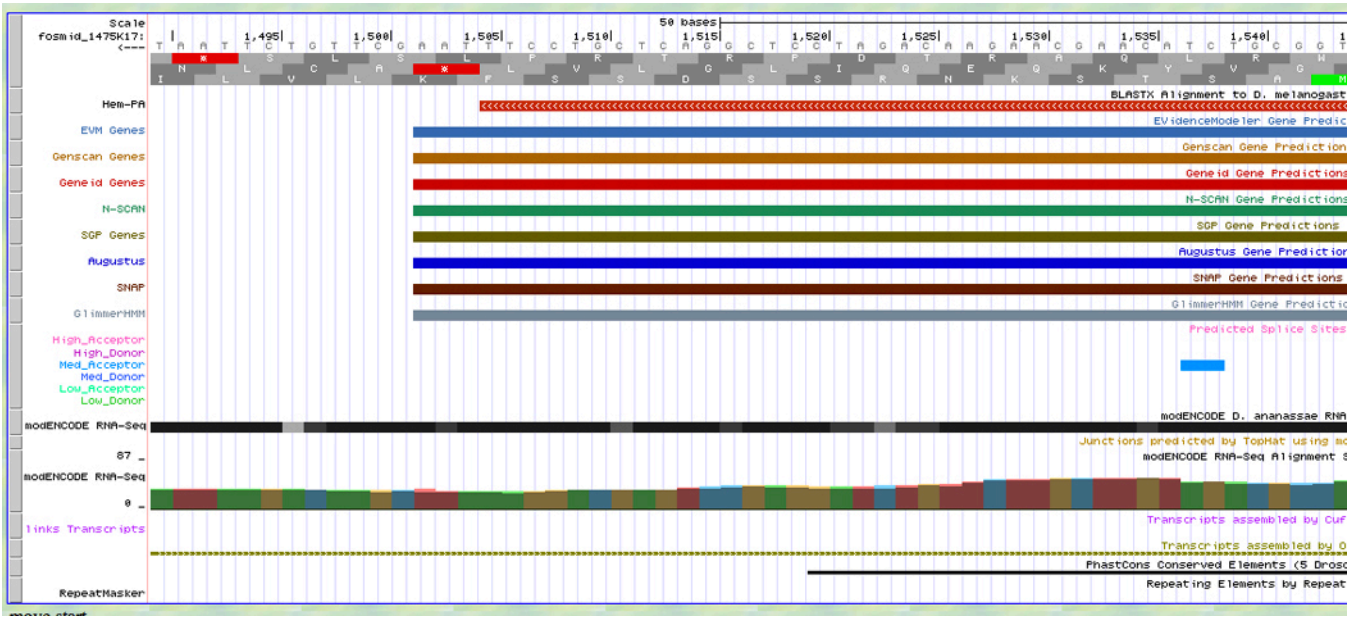
There are some additional features visible in the browser at the splice junction. First, notice that the alignment to the *D. melanogaster* Hem protein stops at the end of the exon. Below the end of the alignment, all of the gene prediction programs identify the end of the exon. There is a splice site prediction in magenta to the left of the end of the exon. The RNA-Seq data (in ochre) support the occurrence of a splice here.

Note that the bases directly above the splice site prediction are GT (backwards). This is almost always the splice donor site. The GT bases, and all the bases to the left, are the intron, spliced out of the mRNA. Record the coordinate of the last base in the first exon. It is 3764.



This is the beginning of the second exon. Note the prediction of the splice acceptor at the end of the intron. Record the coordinate of the first base of the second exon. It is 3702.

Navigate to the end of the second exon until your screen resembles the screen shown below.



Notice the red stop codon in frame -2 at the end of the alignment to the *D. melanogaster* Hem protein. Record the coordinate of the last base of the coding region (not the stop) of exon 2. It is 1505. Note that the stop is 1504-1502.

The coordinates that we have collected make up the gene model. Exon 1 is 4943-3764, exon 2 is 3702-1505, and the stop is 1504-1502. This is a complete model of the *D. ananassae* orthlog of the *D. melanogaster Hem* gene.

4. Verification of the gene model for *Hem*.

Use the **Tools** menu on the course website to navigate to the **Gene Model Checker**:

http://gander.wustl.edu/~wilson/genechecker/index.html

Upload the fosmid fasta sequence using the Browse button.

Enter the name of the *D. melanogaster* ortholog (Hem). The field will autofill to Hem-PA (one isoform).

Enter the coding exon coordinates (4943-3764, 3702-1505) in the box.

Select **No** for **Annotated Untranslated Regions?**

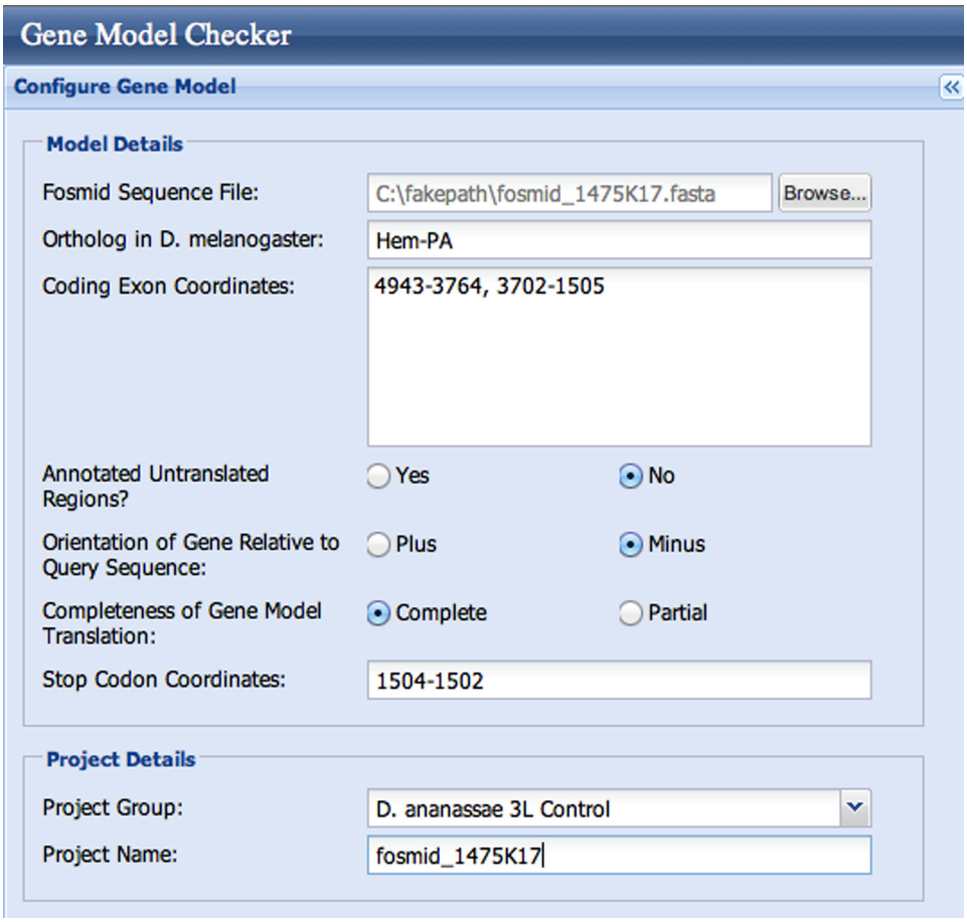
Select the **Minus** strand.

Set the gene model to **Complete**.

Enter the stop codon coordinates 1504-1502.

Select **D. ananassae 3L Control** as the **Project Group** and enter `fosmid_1475K17` as the **Project Name**.

If all has gone well, your screen will look like the screenshot on the next page.



Click the **Verify Gene Model** button. A set of automatic checks is run. A checklist appears on your screen. Try changing one or more of the settings in the Gene Model Checker (for example, of the coordinates of one of the exons) to see what happens to the checklist when you push the button.

Click the tab to see the **Dot Plot**. Click the link in the image to see the alignment. This is a good model of the *D. ananassae Hem* gene. It passes all the checks, and aligns well to the *D. melanogaster* ortholog. The dot plot looks great.