Tutorial 1. Using UCSC Genome Browser to explore DNA sequence and to generate a gene model.

Learning Objectives:

Students should be able to

1. Visualize genes and gene predictions using UCSC genome browser.
2. Distinguish transcript sequence and coding DNA sequence (CDS).
3. Predict location of a CDS (one coding exon) and test their prediction.
4. Retrieve DNA sequence from the UCSC genome browser.
5. Translate DNA sequence using ExPASy Translate tool.
6. Compare predicted protein to a protein database using blastp.

Goal: Use UCSC Genome Browser to find the *D. erecta* CG11077 gene and predict the location of its coding DNA sequence.

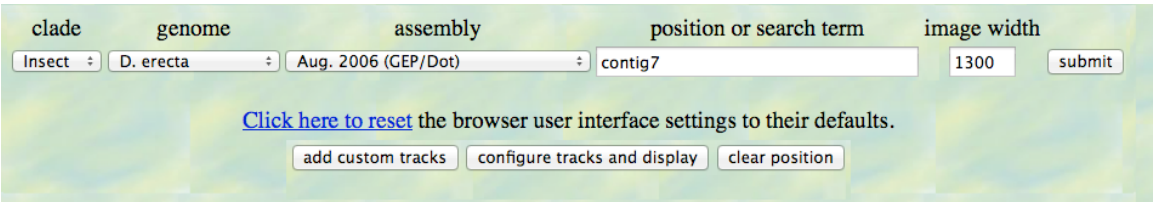
Introduction

Genome annotation includes (among other things) finding genes and describing their structure. We will construct a putative gene model in *Drosophila erecta* based on high quality manual annotations available from a closely-related species (*D. melanogaster*). The first step is to visualize genomic sequence in the context of the expression data, sequence alignment, and computational predictions available. The UCSC Genome Browser (http://genome.ucsc.edu/) provides a convenient way to visualize the DNA sequence and other important information about the genome.

For annotation and practice, we will use a custom version of the UCSC Genome Browser hosted at Washington University in St. Louis (http://gander.wustl.edu). At this site, genomic sequences from several species of fruit flies have been partitioned into overlapping segments (contigs or fosmids) that correspond to the different annotation projects.

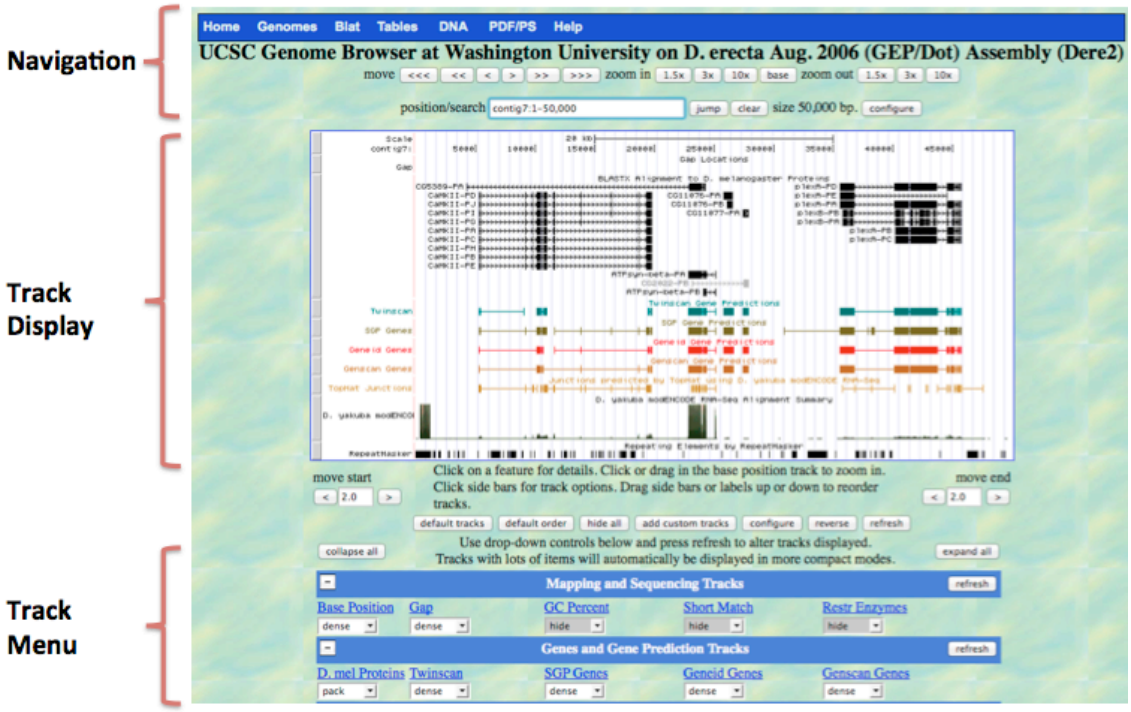
1.1 How do I find a specific contig and a gene of interest?

Open the browser: go to http://gander.wustl.edu/
Click on "Genome Browser" from the blue menu on the left of the screen



From the pull down menu on this page choose:
genome : *D. erecta*
assembly: Aug 2006 (GEP/Dot)
position: contig7
click "submit" button

You will see a very busy intimidating page, which can be divided into 3 parts:



TRACK DISPLAY WINDOW

The top of the track display window starts with a scale and a Base Position track that indicates the current base position (i.e. nucleotide number). This is similar to an X-axis with coordinates (base position). The various colorful lines and boxes below the sequence are evidence tracks and experimental data that have been mapped to the genomic location indicated by the base position track.

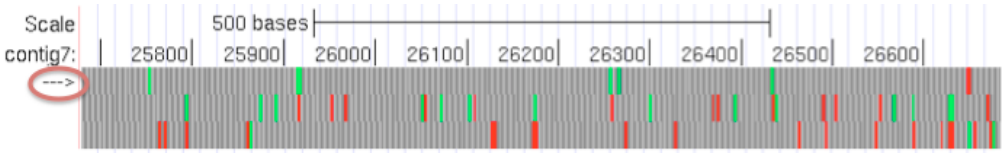
You can change what you see in this display window using two types of controls: **navigation controls** and **track menu**.

NAVIGATION CONTROLS

Above the track display window, navigation controls allow you to move along the chromosome and zoom in and out. Buttons next to "move" allow you to move left or right along the sequence in small (>) or large (>>>) steps.

Use the navigation controls to display only the gene CG11077 in the track display window (no other genes).

Now, zoom all the way to the "base" level (keep zooming in by 3X or 10X steps, or just use the "base" button). (Make sure that the Base Position track is set to "full.") Notice the arrow in the top left corner of the display window (just under words "base position" – if you click on the arrow, you will change which DNA strand is displayed (forward or reverse).



TRACK MENU
Below the window, there is a menu that allows you to choose what type of information you want to view. (The lines of information are referred to as "tracks".)

Tracks are grouped into categories: Mapping and Sequencing tracks, Gene and Gene prediction tracks, etc…. If you want to see the description of a track – click on the name of the track (blue, underlined).

You can control how compactly the information for each track is presented or whether it is hidden. Setting "full" – displays each feature on a new line. "Dense" displays all the features of the track on a single line, irrespective of whether they overlap with each other. After modifying the track display options, update the track display image by clicking on the **"refresh" button.**

1.2 How do I predict the location of the CDS?

Changing the display: Start by hitting "hide all", then "refresh." Your display now is not very informative!

Next, change the track settings to the following (from top to bottom on the track menu):
Under "Mapping and Sequencing Tracks", change Base Position to full
Under "Genes and Gene Prediction Tracks", change D. mel Proteins to full, and the remaining tracks in this section (e.g. Twinscan, Genscan Genes) to dense. Click the "refresh" button.

Sequence tracks: At the top of the Track Display you will see a nucleotide sequence, and immediately below, a translation in three reading frames in the orientation indicated by the arrow at the top left corner. Because the arrow is pointing from left to right, the translations correspond to the three reading frames in the positive strand. We will refer to these as +1 (top), +2 (middle) and +3 (bottom). You can see the other three reading frames (-1, -2, -3) by clicking on the arrow at the top left of display window. Methionines (start codons) are highlighted in green, and stop codons are in red.

Gene and Gene Prediction tracks: Below the base position track, the "D. mel Proteins" track shows the regions of the contig that have sequence similarity to proteins found in *D. melanogaster* proteins. Zoom out 10X. The rectangles show matches between amino acid sequences encoded by *D. erecta* and proteins from *D. melanogaster*. Gene prediction tracks below show connected rectangles (exons) where protein coding genes have been predicted in the *D. erecta* DNA sequence by the various gene predictors (e.g. Twinscan and Genscan) . Note the differences in gene models predicted by the different programs.

In the gene prediction tracks, coding exons are represented by rectangles connected with horizontal lines representing introns. In full display mode, arrowheads on the connecting intron lines indicate the predicted direction of transcription.

Gene Annotation: When we annotate a gene, we will use gene prediction tracks, D. mel Proteins track and other evidence tracks to define the precise location of the coding DNA sequences (CDS) within the genomic sequence. Note that "gene" is not synonymous with transcript or CDS. **Gene** includes regulatory regions of the genome in addition to various transcripts (isoforms) that are derived from alternative splicing of the gene. **Transcript** includes the CDS and the 5' and 3' **untranslated regions (UTR).** However, the non-coding regions of a gene are more difficult to define and we are not going to annotate them at this time.

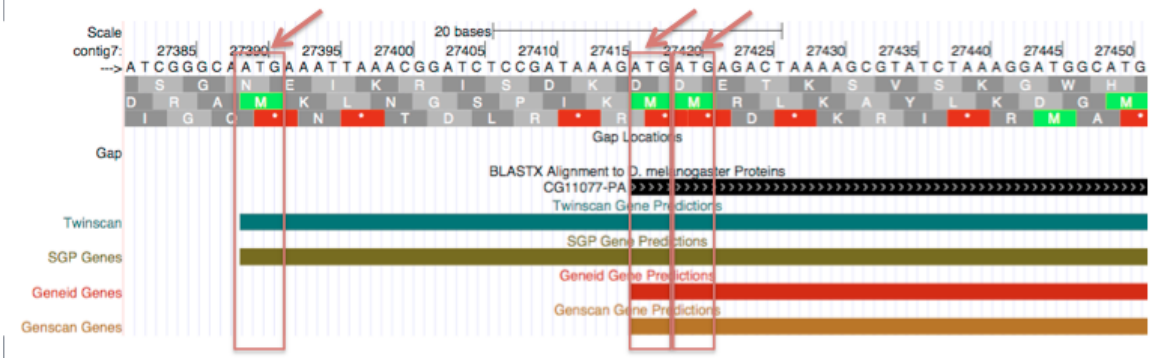
Use the navigation features of the genome browser to identify the location of the CG11077 CDS
Zoom in and out on both strands to predict the location of the *CG11077* CDS:

Reading frame _____ (e.g. +1, -1)

Location of start codon _____ (specify nucleotide position)

Location of the last coding nt_____ Stop codon_____(range)

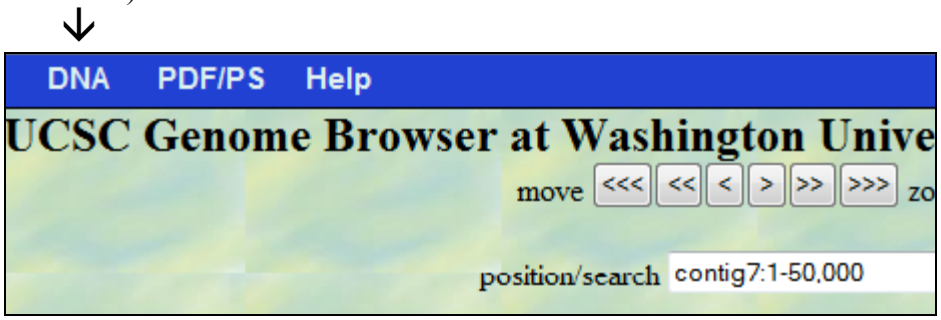
Notice that there are three possible start codons to consider:



When we construct a gene model, we are actually testing a hypothesis – the hypothesis that the *D. erecta* genome has a gene homologous to one found in the *D. melanogaster* genome. Hence we need to collect evidence that supports the conclusion that the exon coordinates are correct. To do this, we will retrieve predicted coding DNA sequence, translate it, and compare it to *D. melanogaster* protein sequence. It is usually a good strategy to start with the largest plausible exon, but be aware that this model may require modification!

1.3. How do I retrieve DNA sequence using the UCSC Genome Browser.

On the Blue menu across the top of the page (just above the words "UCSC Genome Browser") click on "DNA"



You can choose the entire contig or just a portion.
Type in the location/coordinate of your start codon, and of the last nucleotide of the stop codon from section 1.2. Click on "get DNA" to see the sequence in FASTA format. (Details on the FASTA format are found at: http://blast.ncbi.nlm.nih.gov/blastcgihelp.shtml .) Copy and paste this sequence into a **text file -** this is your predicted CDS.

1.4. How do I translate CDS?

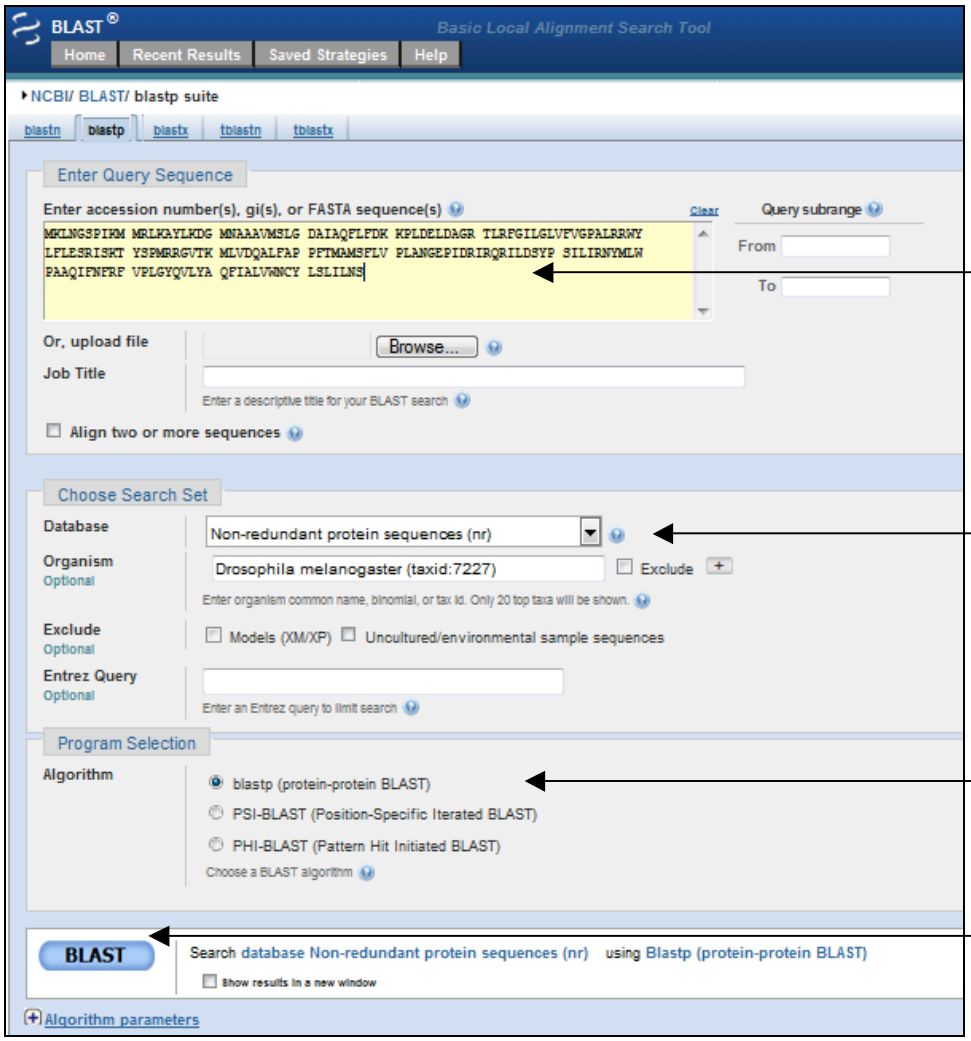
Go to ExPASy website and find the Translate tool (http://www.ebi.ac.uk/Tools/emboss/transeq; an alternative translator can be found at http://www.fr33.net/translator.php)

Paste your DNA sequence into the window, change to "Compact" output format, and translate.
Copy and paste your predicted protein sequence into a text file.

1.5. How do I check whether I chose the best Start codon?

Let's compare your predicted protein sequence to other proteins using **protein BLAST** at NCBI (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi)

Click on the "blastp" tab. Paste your predicted protein into the Query window; choose "Non-redundant protein sequences (nr)" under the "Database" field; then enter "Drosophila melanogaster (taxid: 7227)" under the "Organism" field and click on the BLAST button (see arrows on the screen shot below).



Examine the BLAST output and answer the following questions.
Note that the BLAST help pages can be found at the top navigation bar of the BLAST output page or through the NCBI Bookshelf at
http://www.ncbi.nlm.nih.gov/books/NBK21097/.

How long is your query sequence? _____
(Hint: find the number of aa at the top of the output page)

Is the top *D. melanogaster* match *CG11077*? YES or NO

Does the first amino acid of your protein (query) match to the first amino acid of the *D. melanogaster* protein? YES or NO
If no, modify your coordinates to create a gene model that matches the *D. melanogaster* protein as closely as possible. If yes, record the coordinates for your gene model below.

Conclusion: *D. erecta* gene model for *CG11077* on contig 7: strand: + or -
Coordinates of the CDS _____(range) stop codon_____(range).

Tutorial 2. Creating *D. erecta* gene models for multi-exon genes.

Learning Objectives:

Students should be able to
1. Find a gene model for *D. melanogaster* genes using **Gene Record Finder**.
2. Use **Blast 2 Sequences** to map exon locations
3. Use the **UCSC Genome Browser** to refine exon coordinates.
4. Use the **Gene Model Checker** to verify their gene models and refine coordinates.
5. Recognize donor and acceptor splice sites in a sequence.
6. Explain the meaning of the following terms: reading frame, intron, exon, splice site donor, splice site acceptor, 5' UTR, 3'UTR, start and stop codons, transcript, CDS gene.

Goal: create gene models for *CG11360* and *Slip1* on contig 18.

Procedure:
1. On the **UCSC Genome Browser Mirror**:
   Download contig18 sequence and save it as a text file.
   Find the gene we are going to annotate.
2. Find the gene model for *D. melanogaster* using the **Gene Record Finder**.
3. Use **BLAST** to map the approximate exons locations on the contig.
4/5. Use the UCSC **Genome Browser** to refine exon coordinates.
4/5. Use the **Gene Model Checker** to verify your gene model.

In this exercise we are going to use the four web-based programs listed below; open four tabs in your web browser for each of them:

The first three sites can he found on the Genomics Education Partnership website http://gep.wustl.edu/ under Projects -> Annotation Resources

1. UCSC Genome Browser
   GEP UCSC Genome Browser Mirror
   http://gander.wustl.edu/
   We learned how to use this browser in Tutorial 1.
2. Gene Record Finder
   http://gander.wustl.edu/~wilson/dmelgenerecord/index.html
3. Gene Model Checker
   http://gander.wustl.edu/~wilson/genechecker/index.html
4. NCBI BLAST
   http://www.ncbi.nlm.nih.gov/blast/

Step1. On the GEP UCSC Genome Browser Mirror:

A. Download contig18 sequence and save it as a text file.
B. Find the gene we are going to annotate.

1A. Download contig18 sequence in FASTA format.

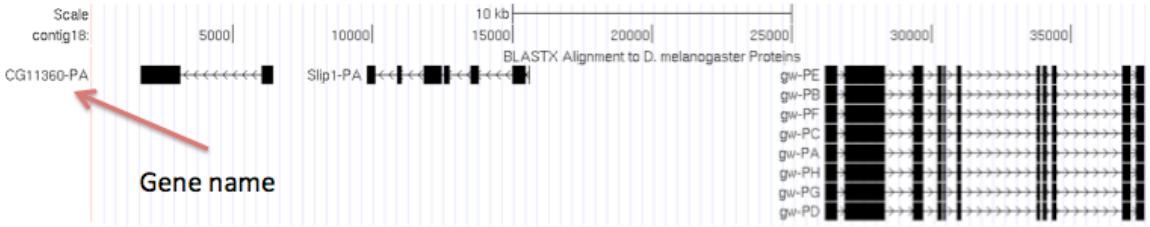
Find contig18 on the *D. erecta* Dot chromosome (Aug 2006 assembly) following the steps described in tutorial 1 (section 1.1). Retrieve the entire contig18 DNA sequence (section 1.3).

Copy all (control A on a PC, include the FASTA header that describes your sequence ">Dere2…") and paste the entire contig sequence into a text file.
(If you are using MS Word, make sure the document is saved as a text file not a doc or docx file. Alternatively use Notepad or WordPad from Programs→ Accessories on a PC)
Save this sequence file as "contig18.txt"

1B. We need to find the FlyBase gene symbol for the homologous gene in *D. melanogaster*

We will use this information in step 2 (Gene Record Finder) and step 5 (Gene Model Checker).

In the *D.mel* Proteins track (Black) in the GEP UCSC Genome Browser mirror you can find gene names, as they appear in NCBI Gene. These names often, but not always, match to FlyBase gene symbols.
If you click on the gene name (**CG11360** below)- you will get some additional information, including a BLAST hit summary with the subject name and FlyBase ID.



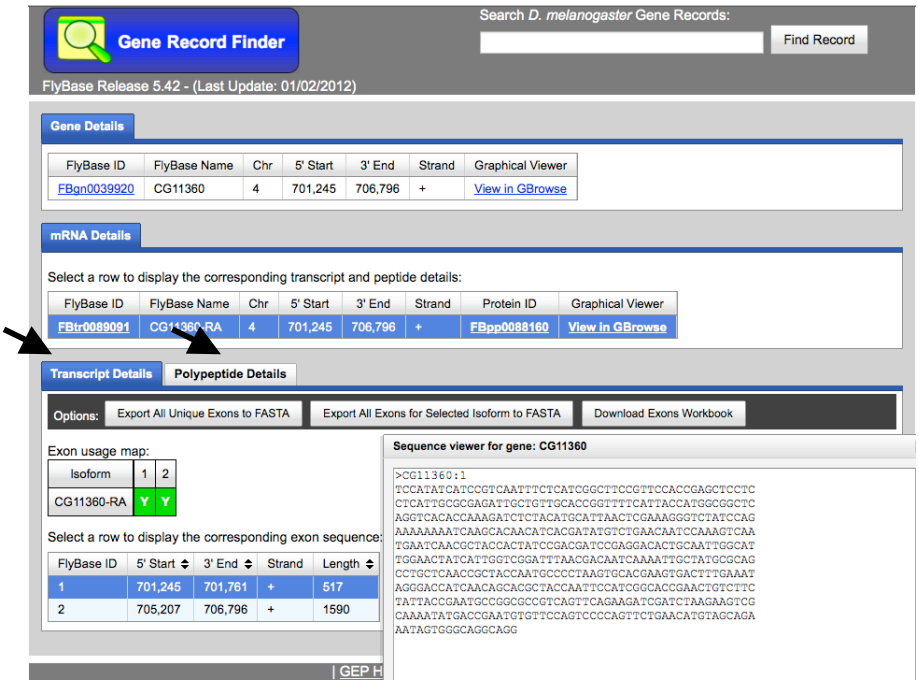
Can you tell whether *CG11360* is on the top or the bottom strand?
Look at the arrows on Genscan or another gene predictor (when it is displayed as "full"). Genes on the top strand will have the >>> symbol, those on the bottom strand, <<<<, indicating the direction of transcription.

Step 2. Find Gene Model for *D. melanogaster* using **Gene Record Finder**.

Gene Record Finder is a program that retrieves information about *D. melanogaster* gene structure from FlyBase. Most genes in *D melanogaster* have been carefully annotated,

and our starting hypothesis is that there is an orthologous gene in *D erecta* that will have a similar pattern of exons. (However, this might prove to be incorrect!) Use the name of the gene found in the UCSC Genome Browser.

Type the gene name in the search window (note that gene names in Drosophila are case sensitive: cg11360 will not work!). Click "Find Record" button to display information about the gene structure in *Drosophila melanogaster*:



In the screenshot above, the 'Transcript Details' are highlighted (see arrows). In order to see 'Polypeptide Details,' click on the appropriate tab.

All exons in the gene are numbered. The ***exon usage map*** shows the order of exons in the transcript or polypeptide (from 5' to 3'). The Polypeptide tab shows only coding exons (exons that code for amino acids), while the transcript tab shows all exons. The ***5' start*** and ***3' end*** columns in the table show the coordinates of exons on the *D. melanogaster* chromosome. Clicking on each exon in the exon table, reveals the sequence that corresponds to the selected exon.

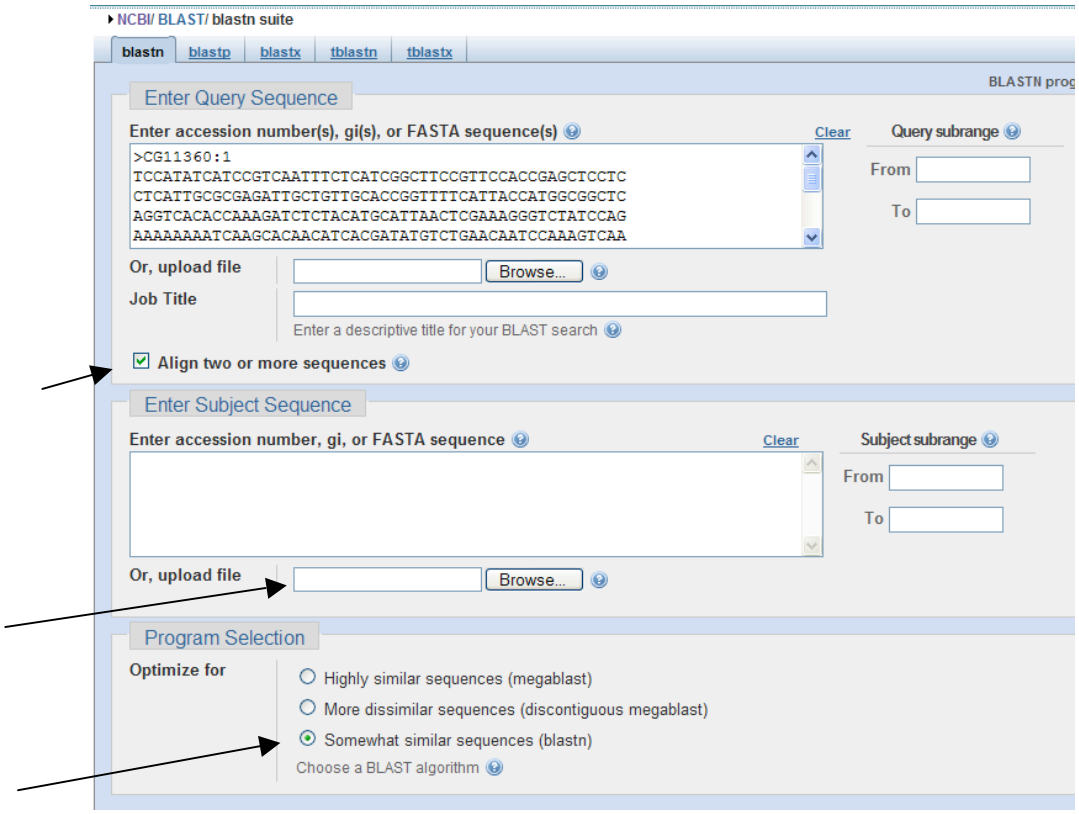
In order to generate a gene model, we will use both polypeptide and transcript details. In this simple example, there is only one transcript and one protein isoform and all exons are coding. For many genes, the situation is more complicated and this program helps organize and visualize information about the transcript and the CDS as it exists in *D melanogaster*. Note the first exon, CG11360:1, does not start with the ATG. It contains an UnTranslated Region (5' UTR).

For the next step, we will copy the sequence of each exon from *D. melanogaster CG11360* and align it to the *D. erecta* contig18 sequence.

3. Use **BLAST** to map the exon locations.

http://www.ncbi.nlm.nih.gov/blast/

Copy the nucleotide sequence of exon 1 from the *D. melanogaster CG11360* gene (which you located in the Gene Record Finder) and paste it into Query Sequence window 1. Check the "Align two or more sequences" box and upload your contig18 text file as the subject. Change the alignment to "Somewhat similar sequences (blastn)". Hit the "BLAST" button at the bottom of the window.



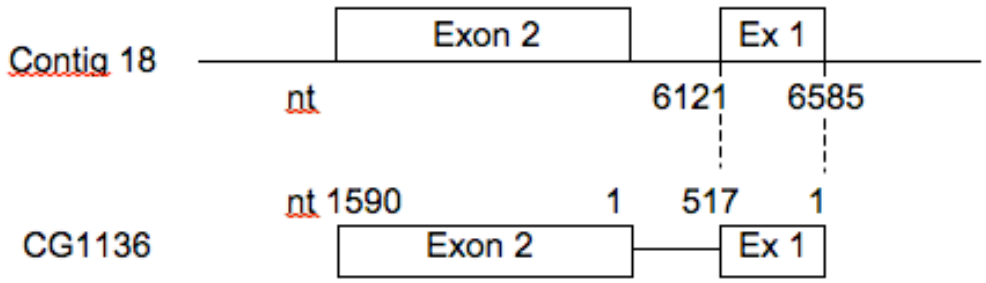
Alternative: you can paste the entire contig18 sequence (from the *D. erecta* genome browser) into the Subject Sequence window (second window).

The alignment (screenshot on next page) shows the relationship between exon 1 of *CG11360* in *D. melanogaster* and the genomic sequence from *D erecta* in our contig. Copy the coordinates from the beginning and the end of the alignment into the table below. Did we find the match to the first and last nucleotide of the *D. melanogaster* exon? Repeat the blastn search for the second exon.

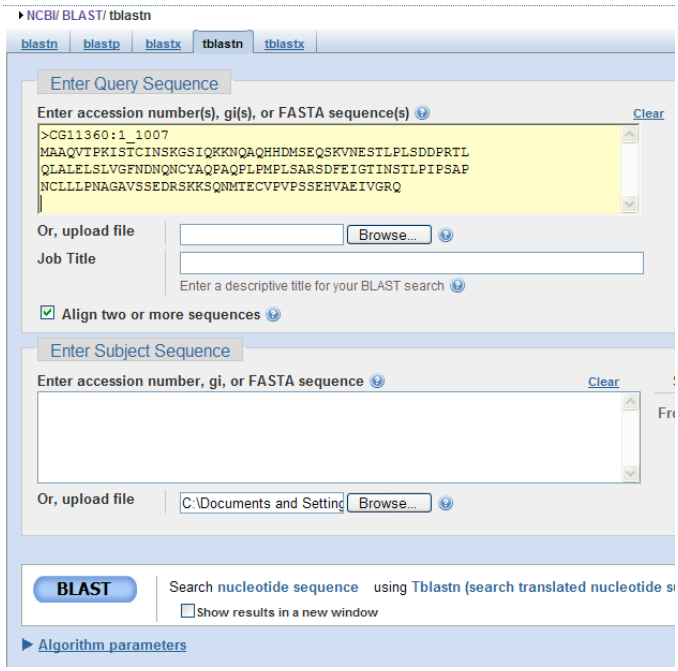
Exon #	Beginning position	End position	Notes (note any gaps in alignment or incomplete exons)
1	6585	6021	
2			

```
>lcl|12791 Dere2_dna range=contig18:1-40000 5'pad=0 3'pad=0
Length=40000

 Score =  666 bits (738),  Expect = 0.0
 Identities = 480/565 (85%), Gaps = 48/565 (8%)
 Strand=Plus/Minus

Query  1     TCCATATCATCCGTCAATTTCTCATCGGCTTCCGTTCCACCGAGCTCCTCCTCATTGCGC  60
             ||||| |||||||||||  || ||||||||||||||   |||||||  ||||||||||||||
Sbjct  6585  TCCATCTCATCCGTCAACTTGTCATCGGCTTCCGTGTCACCGAGGTCCTCCTCATTGCGA  6526

Query  61    GAGATTGCTGTTGCACCGGTTTTCATTAC------------------------------C  90
             |||| ||||||||||||||||| ||| |                              |
Sbjct  6525  GAGAGTGCTGTTGCACCGGTTTTAATTTCGCTTCAAAAAGGGGATAAGAGTGCACTTGCC  6466

Query  91    ATGGCGGCTCAGGTCACACCAAAGATCTCTACATGCATTAACTCGAAAGGGTCTATCCAG  150
             ||| ||||||| |||||||| ||||||||||   ||||||||||||||||||||||||||
Sbjct  6465  ATGTCGGCTCAAGTCACACCGAAGATCTCTACTAACATTAACTCGAAAGGGTCTATCCAG  6406

Query  151   aaaaaaaaTCAAGCACAACATCACGATATGTCTGAACAATCCAAAGTCAATGAATCAACG  210
             |||| |||   |||||| ||||||||||||||||||||||||||| ||||||||||||||
Sbjct  6405  AAAACAAACACAGCACACCATCACGATATGTCTGAACAATCCAAAGGCAATGAATCAACG  6346

Query  211   CTACCACTATCCGACGATCCGAGGACACTGCAATTGGCATTGGAACTATCATTGGTCGGA  270
             |||||||||| || |||||||| ||||||||| |||| ||||||||||||||||||||||
Sbjct  6345  CTACCACTATCAGAAGATCCGAGAACACTGCAACTGGCGTTGGAACTATCATTGGTCGGA  6286

Query  271   TTTAACGACAATCAAAATTGCTATGCGCAGCCTGCTCAACCGCTACCAATGCCCCTAAGT  330
             |||| ||||||||||||||||||||||||| ||||||||||||||||||||||||||| ||
Sbjct  6285  TTTACCGACAATCAAAATTGCTATGCGCAACCTGCTCAACCGCTACCAATGCCCCTATGT  6226

Query  331   GCACGAAGTGACTTTGAAATAGGGACCAT-CA-----------------ACAGCACGCTA  372
             |||| |||||||||||||| |||||||| ||                 ||||||||||
Sbjct  6225  GCACAAAGTGACTTTGAAATTGGGACCATGCACTCATTTTCGAAAGCAGACAGCACGCTA  6166

Query  373   CCAATTCCATCGGCACCGAACTGTCTTCTATTACCGAATGCCGGCGCCGTCAGTTCAGAA  432
             |||||||||||||||||||||||| || |||||||||| ||||||||||||||||||||||
Sbjct  6165  CCAATTCCATCGGCACCGAACTGCCTGCTATTACCGACTGCCGGCGCCGTCAGTTCAGAA  6106

Query  433   GATCGATCTAAGAAGTCGCAAAATATGACCGAATGTGTTCCAGTCCCCAGTTCTGAACAT  492
             |||||||||||||||| ||||||||||||||||||||||||||||||||||||||||||||
Sbjct  6105  GATCGATCTAAGAAGTCTCAAAATATGACCGAATGTGTTCCAGTCCCCAGTTCTGAACAT  6046

Query  493   GTAGCAGAAATAGTGGGCAGGCAGG  517
             ||||||||||||||||||| ||||
Sbjct  6045  GTAGCAGAAATAGTGGGCAGACAGG  6021
```

The figure below shows a sketch of how the *D. melanogaster* exons align to our contig. Add your coordinates for the second exon.
Do these represent transcript or CDS coordinates?



Next, let's try using "Polypeptide Details" to locate the coding regions of *CG11360* in our contig. Go back to the Gene Record Finder and copy the amino acid sequence of the first exon. We want to align protein sequence to DNA sequence, so we need to use tblastn, with the amino acid sequence of the coding exon as the query and contig genomic sequence as the subject. http://www.ncbi.nlm.nih.gov/blast/Blast.cgi



Your alignment result should looks like this:

```
>lcl|34757 Dere2_dna range=contig18:1-40000 5'pad=0 3'pad=0
Length=40000

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position
 Score =  257 bits (656),  Expect = 5e-72, Method: Compositional matrix adjust.
 Identities = 129/148 (87%), Positives = 134/148 (91%), Gaps = 6/148 (4%)
 Frame = -2

Query  1     MAAQVTPKISTCINSKGSIQKKNQAQHHDMSEQSKVNESTLPLSDDPRTLQLALELSLVG  60
             M+AQVTPKIST INSKGSIQK N A HHDMSEQSK NESTLPLS+DPRTLQLALELSLVG
Sbjct  6465  MSAQVTPKISTNINSKGSIQKTNTAHHHDMSEQSKGNESTLPLSEDPRTLQLALELSLVG  6286

Query  61    FNDNQNCYAQPAQPLPMPLSARSDFEIGTI------NSTLPIPSAPNCLLLPNAGAVSSE  114
             F DNQNCYAQPAQPLPMPL A+SDFEIGT+      +STLPIPSAPNCLLLP AGAVSSE
Sbjct  6285  FTDNQNCYAQPAQPLPMPLCAQSDFEIGTMHSFSKADSTLPIPSAPNCLLLPTAGAVSSE  6106

Query  115   DRSKKSQNMTECVPVPSSEHVAEIVGRQ  142
             DRSKKSQNMTECVPVPSSEHVAEIVGRQ
Sbjct  6105  DRSKKSQNMTECVPVPSSEHVAEIVGRQ  6022
```

Why are the coordinates reported by tblastn different from the ones reported by blastn? Re-draw your *CG11360* gene model to incorporate this new information.

4. Use the **Genome Browser** to refine your exon coordinates.

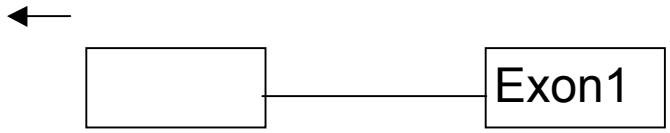
Zoom in on the gene in the Genome Browser and visually check whether the exon boundaries identified by our BLAST results are adjacent to canonical donor and acceptor splice sites. Is an open reading frame maintained following the splice?

Background on exon splicing:
In a mature mRNA, introns are removed by splicing and exons are joined together. In most cases, we can recognize introns because they usually start with a GT (donor splice site) and end with a AG (acceptor splice site), although there are occasionally exceptions to this rule.

For more information on alternative splicing, see **http://web-books.com/MoBio/Free/Ch5A4.htm**

Gene on the top strand:

Gene on the bottom strand:



Our gene is on the bottom strand. Above, sketch where GT and AG will be for the gene on the bottom strand.

Hint: The transcript/mRNA is always synthesized in the 5' to 3' direction regardless of which strand is read. We will still expect to find GT at the beginning and AG at the end of the introns.

Check your exon-intron boundaries and open reading frame to make sure your gene model has the correct donor and acceptor splice site. If not, look for correct splice sites near the predicted exon coordinates derived by Genscan or Twinscan.

5. Use **Gene Model Checker** to test your predicted gene model.

Once you have verified the exon coordinates and found the start and stop codons, you are ready to verify your model using the "Gene Model Checker." You will need your contig text file, coordinates of each exon (without a stop) and the stop codon coordinates to run the checker.

If you gene model passes the "Checker," does it mean that your model is correct?

Reflect back on the two types of BLAST you have done during annotation:

What are the advantages and disadvantages of using nucleotide sequences and amino acid sequences for finding exons?

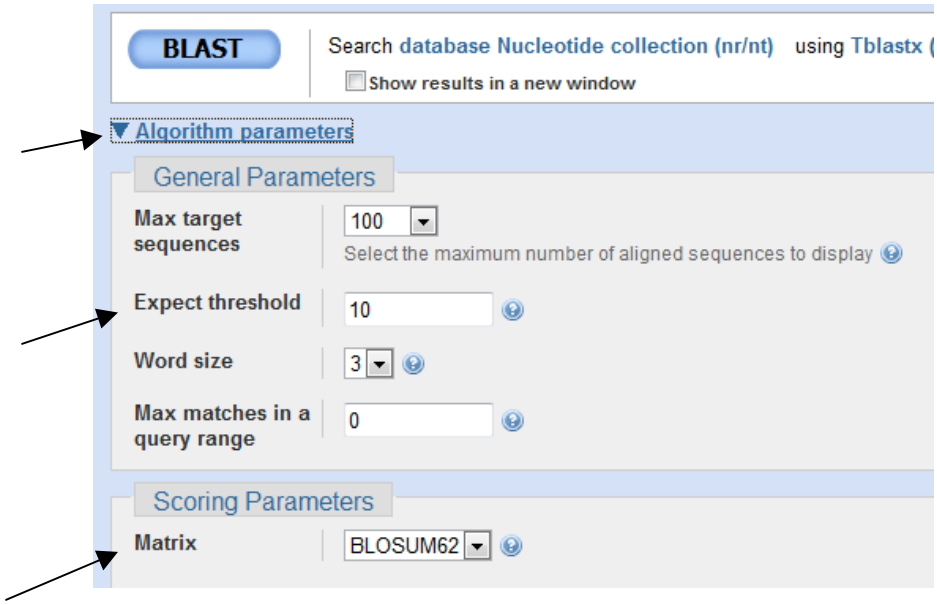
After you finish annotation *CG11360*, repeat the procedure for *Slip1* and fill out table below.

Predicted exon coordinates:

Exon #	Beginning	End	Start/stop	Notes
1	15741	15591		
2				
3				
4				
5				
6				

What should you do if you do not get an exon aligned to the contig?

If the exon is very short, you may have come up with nothing in your initial alignment. You will need to change the BLAST parameters to make the search more sensitive: below the "BLAST" button, expand the "algorithm parameters" section. Try increasing the "Expect threshold" value from 10 to 1000 or higher, or changing "matrix" to something with a lower number. Also consider adjusting the word size to a lower value. You may have to use blastx to align the amino acid sequence of the exon to the contig.



If you gene model passed the "Checker," does that mean your model is correct?

What evidence can you use to support the notion that your model is correct? (In some cases we now have data on RNA transcripts, which is very helpful!)

Summary:

Comparative genome annotation relies on the availability of a well-annotated reference genome (*D. melanogaster* in our case) and uses multiple sources of data to come up with a defensible gene model (one that is best supported by the available evidence).

These tutorials introduced you to a wide range of tools used to annotate eukaryotic genomes. Some of these tools are widely used for many applications (BLAST, UCSC Genome Browser) and some of these tools were developed specifically for our project (Gene Model Checker and Gene Record Finder). The procedure summarized below includes alternatives to the custom tools and should be applicable to annotation of other genomes for which a reference genome is available. Please, refer to the "Annotation Instruction Sheet" on the GEP website for a more detailed discussion and summary of annotation guidelines used by the GEP to annotate the genes in various Drosophila species.

Procedure for comparative *D. erecta* genome annotation

For each feature of interest:

1. Identify the likely ortholog* in *D. melanogaster* (D.m.)
   Use the *D. mel proteins* track in the UCSC genome browser for *D. erecta* (D.e.), or Blast the protein sequence predicted by GeneScan against the D.m. genome.
2. Find the gene model of the D.m. ortholog and identify all exons
   Use the Gene Record Finder, and/or
   Use FlyBase or Ensembl.
3. Find the approximate locations of exons on the contig
   Use bl2seq
4. Generate a gene model/refine coordinates:
   Find start and stop codons, and find donor and acceptor splice sites that link exons together using information from different sources: exon location, reading frame, gene predictions and others.
5. Check whether your gene model is consistent with basic biology
   Use the Gene Model Checker or the
   Translate tool (ExPASy) and blastp

* Orthologs are genes in different species that evolved from a common ancestral gene.