An introduction to the complete GEP gene annotation process using a simple example from *Drosophila erecta*

Ken Saville\(^1\) and Gerard McNeil\(^2\)

1. Biology Department, Albion College, 611 E. Porter, Albion Mi. 49224 (ksaville@albion.edu)
2. City University of New York, Jamaica, NY 11451 (gmcneil@york.cuny.edu)

Acknowledgements: Much of this material was modified from resources available through the Genomics Education Partnership (GEP), available at [http://gep.wustl.edu](http://gep.wustl.edu). In particular, the material presented at the 2012 ABLE workshop by Emerson *et al.*, was used as a starting point.

Overview

The Genomics Education Partnership (GEP) is a national, collaborative, scientific investigation of a problem in genomics, involving wet-lab generation of a large data set (e.g., sequence improvement of genomic DNA) and computer analyses of the data (including annotation of genes, assessment of repeats, exploration of evolutionary questions, etc.). Overall, the goal of annotation is to create gene models for all the genes in a genome. The specific goal of the GEP is to annotate the genomes of several *Drosophila* species, using the genome of *D. melanogaster* as a reference genome. In particular, the GEP is focused on genomic regions in other species that correspond to chromosome four of *D. melanogaster*. Chromosome four is also referred to as the dot chromosome and the Muller F element (Figure 1). Thus, the current research problem entails generating finished sequence from the fourth (dot) chromosome of various species of *Drosophila*, annotating these sequences, and making comparisons among species to discern patterns of genome organization related to the control of gene expression. Here we will concern ourselves only with the process of gene annotation.

A brief description of the scientific background of the current research problem

The scientific interest is based on observations that the dot chromosome shows a mixture of heterochromatic and euchromatic properties. The dot chromosome is like heterochromatin in that it stains intensely with fluorescent DNA binding dyes, has a high density of repetitious sequences, is late replicating, and exhibits very low meiotic recombination. At the same time, the distal region (i.e. the region not near the centromere) of the dot chromosome is amplified in polytene chromosomes, which is a property of euchromatin, and codes for ~80 genes, a gene density similar to that found in the euchromatic arms. In most genomes, actively-transcribed regions of DNA are typically associated with euchromatic domains, while transcriptionally-silent regions of DNA are generally associated with heterochromatic domains. The control of gene expression is important to many areas of cellular and molecular biology, including the understanding of many human diseases. An understanding of chromosome organization and chromatin effects on dot chromosome gene expression in *Drosophila* will thus shed light on the mechanisms of gene regulation in general. An understanding of these chromatin effects requires careful analysis, not just of the genes present but ultimately of the type and distribution of repetitive elements and non-protein coding regions of the genome. However, we will concern ourselves here primarily with the coding regions.
What is Gene Annotation?

Gene annotation refers to the process of ascribing functions to various parts of a newly sequenced genome. While continually updated and highly automated technologies can generate a vast amount of sequence data and leads to a rapidly increasing number of sequenced genomes, the process of gene annotation is nowhere near as automated and requires careful analysis by trained human annotators (soon to be you!). In the process of annotation, decisions are made as to which subset of DNA sequences within a large genomic sequence contains biological information. This involves identifying features in the DNA sequences and, in many cases, determining which of these features are likely to be genes. As you know, the expression of protein-coding genes in cells employs mRNA intermediates and the coding regions of most eukaryotic protein-coding genes are interrupted, consisting of coding exons and non-coding introns. In fact, introns are often much longer than the exons of many genes, and they must be spliced out of the initial RNA transcript. Thus, gene annotation includes careful mapping of all the exon-intron boundaries, to create a gene model that results in the translation of a full-length polypeptide chain. Gene annotation is facilitated by computer programs, which scan the newly-sequenced DNA for intron splice sites and open reading frames; the programs may also search for consensus sequences of known gene regulatory regions (such as
**promoters** and enhancers), which dictate when and where a gene is transcribed into RNA (Figure 2). Human annotators combine these (often contradictory) computational predictions with sequence alignment and gene expression data (e.g. RNA-seq data) to create the best-supported gene model.

![Diagram of a single, hypothetical eukaryotic gene. Some of the rectangles code for amino acid sequences of a protein (the exons), others contain regulatory information (promoters and enhancers), and some areas are transcribed but not translated (introns between the exons).](image)

**Figure 2.** Diagram of a single, hypothetical eukaryotic gene. Some of the rectangles code for amino acid sequences of a protein (the exons), others contain regulatory information (promoters and enhancers), and some areas are transcribed but not translated (introns between the exons).

GEP students will use a comparative genomics approach to gene annotation, comparing new genomic DNA sequences (and mRNA sequences, if available) from various *Drosophila* species to known reference DNA and mRNA sequences from *Drosophila melanogaster*. A discussion of the various rules and criteria that are used to generate gene models can be found in the [Annotation instruction sheet](#) on the GEP website. A copy of this document will be provided to you as a reference for future annotation projects. The following glossary of terms is also provided for your reference.

**Glossary of terms commonly used in the process of gene annotation**

**ab initio gene finding**: process whereby genomic DNA sequence alone is systematically searched for certain telltale signs of protein-coding genes (as opposed to experimental evidence in the form of an identified mRNA or protein molecule encoded by that DNA sequence

**Chromatin**: the DNA-protein complex found in eukaryotic chromosomes
- **Euchromatin**: chromatin that is diffuse and non-staining during interphase; may be transcribed.
- **Heterochromatin**: chromatin that retains its tight packaging during interphase; often not transcribed

**Coding DNA Sequences (CDS)**: the subset of exon sequences that are translated into protein

**cDNA (complementary DNA)**: a DNA copy of an mRNA molecule, manufactured using the enzyme reverse transcriptase

**Consensus sequence**: the most common nucleotide (or amino acid) at a particular position after multiple, related sequences are aligned and similar functional sequence motifs are found.

**Contig**: A region of contiguous sequence containing genes to be annotated.
**Exons:** gene sequences that are transcribed into RNA and are present in the mature (spliced) mRNA molecule

**Expressed Sequence Tag (EST):** partial, single (e.g., one shot) sequence read of a cDNA molecule. The sequence is of relatively-low quality, usually 500 to 800 nucleotides in length.

**Feature:** any region of defined structure/sequence in a genomic fragment of DNA. Features would include genes, pseudogenes and repetitive elements. Most people are interested in identifying the protein-encoding genes.

**Fosmid:** a cloning vector, which has been manufactured to accept DNA inserts of ~40,000 base pairs (bp) [normal plasmids are able to carry only 1-20 kb]; usually propagated in *E. coli*, which can each only contain one fosmid

**Homologous genes:** genes that have similar sequences because they are evolutionarily related; there are two different types of homology

- **Orthologs:** related genes in different species, which are derived from the same gene in a common ancestral species
- **Paralogs:** related genes within a species, which have arisen by a duplication event

**Introns:** gene sequences that are transcribed into RNA but are removed during splicing

**Isoform:** any of several different forms of the same protein. Isoforms may be produced from different alleles of the same gene, from the same gene by alternative splicing, or may come from closely-related genes.

**Open Reading Frame (ORF):** a segment of the genome that potentially codes for a polypeptide chain (or part of a polypeptide chain). Also the part of an mRNA molecule that potentially codes for a polypeptide. ORFs are located between the start codon (ATG; AUG in the mRNA) and a stop codon (TAA, TAG, TGA; UAA, UAG, UGA in the mRNA) for translation. There are six possible reading frames for any potentially protein-encoding genomic DNA fragment: three on one strand and three on the other strand of DNA heading in the opposite direction (Figure 3).
**Figure 3.** The six possible reading frames for any region of DNA.

**Polytene chromosomes:** giant chromosomes that form when multiple rounds of DNA replication occur and the sister chromatids remain attached to and aligned with each other.

**Splice donor:** The sequence at the beginning of an intron, also called the 5' splice site. This typically has the sequence GT in DNA and GU in RNA.

**Splice acceptor:** The sequence at the end of an intron, also called the 3' splice site. This typically has the sequence AG in DNA and in RNA.

**5' Untranslated Region (5' UTR):** mRNA sequence between the 5' cap and the start codon for translation.

**3' Untranslated Region (3' UTR):** mRNA sequence between the stop codon for translation and the poly A tail.
A sample annotation project of a simple gene for *D. erecta*

The gene selected for annotation in this exercise is from the *Drosophila erecta* genome, and, therefore, is relatively easy to annotate because *D. erecta* is evolutionarily very closely related to *D. melanogaster*. That is, the sequence will be quite similar to that of *D. melanogaster*, so the features of the gene, such as the start codon, the intron/exon boundaries and the stop codon, should be easily recognizable. Also, this gene is well characterized in *D. melanogaster* and thus has an easily identifiable name and function (as opposed to a 'CG' gene that has been identified based simply by computer-based predictions). Keep in mind that other genes from *D. erecta* and especially genes from more distantly related species will present additional problems for annotation. Guidance for dealing with these more advanced problems can be found in other materials available on the GEP website. However, the objective here is to familiarize the student with the various tools (listed above) and basic strategies involved in the overall process of determining a gene model by annotation, checking the gene model, and submitting the result to GEP. This process can be summarized as follows, and this is the basic sequence of events covered in this document.

1. Identify and get an overview of the region to be identified in the UCSC Genome Browser mirror. The overall region is typically referred to as a fosmid or a contig, depending on the exact region. The region will consist of approximately 40,000-50,000 base pairs of DNA.
2. Download the complete DNA sequence of the fosmid or contig. This sequence will be saved as a text file and used for various BLAST searches as the region is annotated.
3. Identify the specific *D. erecta* feature to be analyzed. Ultimately, every feature from the region will be annotated, but we need to start with one.
4. Identify the ortholog in *D. melanogaster*.
5. Find the gene record of the ortholog from *D. melanogaster* using Gene Record Finder. In most cases, several versions, or isoforms, will be found for the gene. We will need to annotate all isoforms for each gene.
6. Compare the exon sequences from Gene Record Finder to the downloaded fosmid (or contig) DNA sequence. This is the primary method of mapping specific features of the gene (start codon, intron/exon boundaries and stop codon). This process results in a gene model consisting of a series of coordinates that correspond to the various parts of the gene model.
7. Use the coordinates determined in step 6 to verify the gene model using the program called Gene Model Checker.
8. Once the gene model is accepted by Gene Model Checker, various additional checks need to be made. These include:
   1. Comparing the peptide sequence predicted by your gene model to the *D. melanogaster* peptide sequence using BLAST.
   2. Generating a 'dot plot' of the above comparison. This is a graphical representation of the alignment and is useful for identifying trouble spots.
   3. Generating files needed for the final GEP submission. These files include a GFF file, transcript file and a peptide file. These files are generated by the Gene Model Checker and need to be downloaded and submitted to GEP.
9. Complete the GEP report. This document needs to be filled out for each isoform annotated and includes all of the information described above.
10. Submit the project report using the GEP project management system.
These steps are summarized graphically in the following Figure.

**Figure 4.** Summary of steps involved in gene annotation using GEP annotation tools Reproduced from Emerson et al. 2012.

**Web based annotation tools**

The four major web based tools we will use are listed below. You should open each site in a separate tab in your web browser and try to keep them open as you complete this tutorial.

1. The GEP version of the UCSC genome browser: [http://gander.wustl.edu/](http://gander.wustl.edu/)
Becoming familiar with the UCSC genome browser

The UCSC (University of California-Santa Cruz) browser is a centralized resource for analyzing sequences from a variety of genomes, including the human genome. We are using the GEP version of the UCSC browser, which organizes the specific *Drosophila* genome sequences we need for easier access. The GEP version can be accessed directly at http://gander.wustl.edu, but we will access it through the GEP website.

We will first explore the various features of the genome browser, then focus on one gene and complete the annotation and submission process for that gene. The gene we will use as an example is contained in contig38 of the *Drosophila erecta* dot chromosome. This can be found through the GEP website by navigating to the following page: From the GEP home page (http://gep.wustl.edu), click on Projects, then Annotation Resources, then GEP UCSC Genome Browser Mirror. This should bring you to the genome browser front page. From here, click on the Genome Browser link (in the left hand column) to access the Genome Browser gateway page.

Once at the genome browser, navigate to contig38 of the *D. erecta* Aug. 2006 (GEP/Dot) assembly by configuring the form as shown in the screenshot below:

Click **submit**.

This will bring you to a screen that represents contig38. The exact appearance of this screen may vary depending on which features are selected. An example of this screen is shown below. First notice the 'position/search' box. It should read 'contig38: 1-53,840'. This tells us that the complete sequence for this region is 53,840 base pairs. To make sure we're all starting at the same place, scroll down a bit and click on 'hide all'.
Once you select 'hide all', the screen should look like that shown below. You'll probably need to scroll up and down to see everything. The available evidence tracks are listed below the genome browser graphics.

This UCSC genome browser page shows several categories of information pertaining to this genomic region. Within each category are several 'tracks' of information. We can select which tracks we want to display using the various dropdown menus. The display options for each track are: hide, dense, squish, pack, and full. These terms represent various amounts of information to
display from the least (hide) to the most (full). You will need to play around with these to get an idea of what will be displayed in each case. As we continue through this walkthrough we will tell you which tracks are most important for the task at hand. Note - each time you change the settings of the drop down menus you need to click 'refresh' in order for those changes to be appear in the genome browser.

Here we give a general description of the various tracks. You can also click on any of the hyperlinks above the display option drop down boxes to get a more detailed description of each track. Try this now. Click on the hyperlink (the blue underlined text) for 'Base Position' and read the description of the information contained in this track.

The various information tracks are:

**Mapping and Sequencing Tracks**

For this one we will mainly use the 'Base Position' track. We will usually keep this set to full. You can also zoom to the base level to see the actual DNA sequence and the three translated reading frames. To do this, scroll to the top of the page, find where it says 'zoom 1.5X, 3X, 10X, and base'. Click on 'base'. The display screen should now look like this:

Now you can see the DNA sequence (starting with CGCTTTTA) and three possible reading frames translated from this sequence. The green M's represent start codons and the red asterisks represent stop codons. We'll be discussing reading frames more as we go.

Also notice that by zooming to the base level we have changed the amount of the sequence represented. Notice that in the example shown above (your specific numbers may vary) the position/search window now reads contig38:26,846-26,996. We need to reset this to the full contig before we proceed. To do this, either click zoom out 10X several times until it reads 1-53,840, or simply type contig38 (don't type the colon) into the “position/search” box and hit jump. That should bring you back to this screen:
Genes and Gene Prediction Tracks:

This category contains several tracks, each of which represents a different method for identifying interesting features in this *D. erecta* contig:

These tracks are:

**D. mel proteins** - This track shows potential genes based on a BLASTX comparison of the *D. erecta* contig38 to the collection of known *D. melanogaster* proteins. Recall that BLASTX translates a given DNA sequence, in this case contig38, into six possible reading frames and then compares each translated sequence to known proteins contained in a database (in this case the *D. mel* protein database.) Only three possible reading frames are shown. These are based on the DNA sequence being transcribed and translated from left to right and are referred to as the +1, +2, and +3 reading frames. It's possible that the DNA sequence is actually transcribed and translated from right to left, these possible reading frames are called -1, -2, and -3. If we need to we can reverse complement the contig DNA sequence by clicking the reverse arrow (to the left of the “Base Position” track), but we'll deal with that when the time comes.

Change the D. mel protein track to the various levels (hide, dense, squish, etc.) to see the various amounts of information that are displayed at each display mode. Each time you change the menu, click on the “refresh” button to update the graphical display. Now, set the menu to 'pack' and click refresh. Before moving on, make sure your screen looks like this:
Ok, now we're getting some information. Each set of black or gray bars represents a BLASTX 'hit'. In other words each set of bars represents a protein from *D. melanogaster* that shares sequence similarity with a predicted protein from contig38 as determined using the BLASTX program. This information is very useful since the *D. melanogaster* genome is very well characterized and annotated in detail, therefore we can identify important regions (e.g. genes) in other Drosophila genomes by comparing their genomic sequences to the *D. melanogaster* genome. The black bars represent highly similar BLASTX hits, and the gray bars represent proteins with lower degree of similarity.

The names of the *D. mel* genes are shown to the left of the BLASTX hits. For example, the hits to the far left are labeled RpS3A-PA, RpS3A-PF, RpS3A-PB. RpS3A is the official FlyBase symbol for this gene. The PA, PF, and PB suffixes designate the three different isoforms of this gene. Recall that isoforms are different transcribed products of the same gene. For example two mRNAs that are generated by alternative splicing would result in two isoforms of that gene.

Notice that the sequence to the right matches several different genes (JYalpha-PB, CG40625-PB, etc.). Before we can construct a gene model, we need to first correctly identify the likely ortholog contained in our sequence. We can use various metrics such as E-value, percent identity, and synteny to help identify the correct ortholog. However, we will focus on a simple case in this exercise where the ortholog assignment is unambiguous.

*Based on what you see so far, how many genes, and how many isoforms of each, are contained within this sequence? What are the names of these genes and isoforms?*

**Twinscan, SGP genes, Geneid genes, and Genscan genes:** These tracks all represent results from different computational gene predictors. These types of programs are sometimes referred to as *ab initio* (*ab initio* means 'from the beginning') gene predictors, and essentially predict gene features by searching the DNA sequence for signals (codon bias, base compositions, start codons, potential splice sites, etc.) consistent with the coding regions of a gene. The exact methods of prediction are beyond the scope of this tutorial, but you should consider these tracks as potential evidence to support or refute your final gene model. We will focus on the Genscan gene predictor for this example because we have found Genscan has high sensitivity in detecting coding regions within a genomic sequence. It is also important that these predictions are mostly independent from the BLASTX hits. As we accumulate evidence for a gene model, having evidence from multiple independent sources strengthens our case. To see a couple of gene predictor results, we will set the Genscan and Twinscan tracks to 'pack', to get this screen:
The two types of gene predictions appear at the bottom of the screen. The thick bars represent exons and the thinner lines represent introns. Note that both gene predictors identified roughly the same set of coding regions. However, the Genscan program predicts three genes called contig38.1, contig 38.2, and contig 38.3 and Twinscan predicts four genes. Also notice that the Genscan program has combined two genes identified by BLASTX (Rp53A and ci) into one predicted gene called contig38.1. Although we don't yet know which of these scenarios is correct, gene predictors often make these kinds of mistakes. Part of your job as an annotator is to resolve these types of issues and provide evidence to support a final gene model.

The 'Predicted Splice Site' track displays splice site predictions generated by the program GeneSplicer. It is often useful in deciding between potential splice sites when generating a gene model. We will use this later in this tutorial as we annotate one of the genes contained in this contig.

**mRNA and EST tracks:** This track displays alignments of expressed sequence tags (ESTs) from species other than *D. erecta* (the reason other species are used is because the data isn't available for *D. erecta*). ESTs are single-read sequences derived from cDNA copies of mRNAs, and therefore usually represent fragments of transcribed genes. So ESTs are evidence that a particular region of the genome is transcribed into RNA, which is evidence of a gene in that region. Gene expression data such as these are useful for resolving conflicting evidence from other sources; however, we will postpone further discussion of this until we encounter more complex cases.

**RNA seq tracks:** Tracks in this section were created by mapping *D. yakuba* mRNA-Seq reads (generated by the modENCODE project) against the *D. erecta* contig sequences. RNA-seq refers to large-scale next generation sequencing (NGS) methods to sequence all expressed RNAs in a RNA library. This is another set of data reflecting gene expression and so can be useful when trying to identify specific genes and specific exons of genes. Again, we will postpone further discussion of this until we encounter more complex cases.
Comparative genomics: D. mel Net: This track shows the best alignment between the D. melanogaster genome assembly and the D. erecta contig. It is useful for finding orthologous regions and for studying genome rearrangement. While we won't be using these directly in our annotation process, we will include this track in our final screenshot for our final report for further analysis by the GEP.

A discussion of the tracks in the "Variation and Repeats" section is beyond the scope of this tutorial.

The process of gene annotation

For the purposes of this tutorial, we will describe the annotation of a gene from Drosophila erecta. Again, it is important to realize that we have chosen a simple example in order to familiarize the student with the steps involved from the beginning to the end of the annotation for one gene. Actual annotation projects, particularly of genes from more distantly related species will present more complicated problems that will require additional approaches to resolve. There is a wealth of information available on the GEP website to help address some of these advanced problems. The goal of this tutorial is for students to go through the entire annotation process using a relatively simple example in order to get an overview of the tools and procedures involved.

Although there are many different ways to annotate a gene, we will start with one particular approach and modify it as necessary when particular issues arise. In brief, this approach, as applied to D. erecta genes is to

1. Identify and get an overview of the region to be identified in the UCSC Genome Browser Mirror.
2. Download the complete DNA sequence of the fosmid or contig.
3. Identify the specific D. erecta gene to be analyzed. Ultimately, every gene in the contig will be annotated, but we need to start with one.
4. Identify the 'ortholog' in D. melanogaster.
5. Find the 'Gene Record' of the ortholog from D. melanogaster using 'Gene Record Finder'. In most cases, several versions, or isoforms, will be found for the gene. We will need to annotate all isoforms for each gene.
6. Compare the exon sequences from Gene Record Finder to the downloaded fosmid (or contig) DNA sequence.
7. Use the coordinates determined in step 6 to verify the gene model using the program called Gene Model Checker.
8. Once the gene model is accepted by Gene Model Checker, a number of additional checks will need to be made. These include:
   1. Comparing the peptide sequence predicted by your gene model to the D. melanogaster peptide sequence using BLAST.
   2. Generating a 'dot plot' of the above comparison. This is a graphical representation of the alignment and is useful for identifying trouble spots (e.g. incorrect splice sites).
3. Generating files needed for the final GEP submission. These files include GFF file, a transcript file and a peptide file. These files are generated by the Gene Model Checker and need to be downloaded and submitted to GEP.
9. Complete the GEP report-ugh!
10. Submit the report using the GEP project management system - yay!

**Sample Annotation**

Once again, open each of the following web pages in a different tab in your browser

1. The GEP version of the UCSC genome browser:  [http://gander.wustl.edu/](http://gander.wustl.edu/)

1. *Inspect the various evidence tracks of the UCSC browser to identify potential genes in a particular contig of the D. erecta genome.*

We will annotate one of the genes contained in contig38. So navigate to the appropriate region of the UCSC browser. Forgot already? Check your notes. You should be able to get to the following screen.

![Image of UCSC Genome Browser](image_url)

Click **submit**.

Adjust the various evidence tracks so that the screen appears as below
We will annotate the gene identified in the BLASTX search as *ci*. Because only one gene has been identified in the automated BLASTX search we will assume this is the ortholog. In many cases, additional steps (e.g. a FlyBase BLASTP search of the Genscan gene prediction against the collection of *D. melanogaster* proteins) will be required to verify the ortholog assignment. For example, there are several hits to the rightmost gene in this fosmid and some investigation would be required to identify the correct ortholog. From initial inspection of the BLASTX hits for *ci*, it appears that there are three isoforms of this gene. We will need to annotate each isoform in detail.

2. Download the complete DNA sequence of the fosmid or contig. We will download the entire contig 38 sequence from the UCSC browser page shown above. **IMPORTANT:** Make sure that the position/search window includes the entire DNA sequence (in this case bases 1-53,840).

Click on the 'DNA' link at the top of the page to get the following screen.
Click on 'get DNA'. This will lead to a screen with the entire 53,840 DNA sequence. Click inside the text box and select the entire sequence (Ctrl-A on MS Windows, ⌘-A on Mac OS X). Copy the entire sequence onto the clipboard (Ctrl-C on MS Windows, ⌘-C on Mac OS X), including the first line that reads >Dere2_dna range=contig38:1-53840 5'pad=0 3'pad=0 strand=+ repeatMasking=none.

Open a Text Editor (Notepad or Wordpad on MS Windows, TextEdit on Mac OS X). Paste the entire DNA sequence into the plain text document and save it as contig38sequence.txt.

3. **Identify the specific D. erecta gene to be analyzed.** Ultimately, every gene from the region will be annotated, but we need to start with one. We will start with the BLASTX hit to the ci gene. It is relatively small and simple, but also has a few isoforms. It also has an interesting name and a well-characterized function.

4. **Identify the ortholog in D. melanogaster.** The BLASTX hit shown in the browser is probably 'correct' in identifying the D. mel ortholog. However, we will independently confirm this by investigating one of the gene predictions. In this case, we will use the Twinscan prediction because it doesn't have the complication of being fused with another gene. The basic strategy here is to get the protein sequence predicted by Twinscan and use it in a blastp search of *D. melanogaster* proteins. To get the protein sequence click on the Twinscan prediction (contig38.002.1):

   Twinscan Gene Predictions (contig38.002.1)
   
   **Position:** contig38:15063-19519  
   **Genomic Size:** 4457  
   **Strand:** +

   **Links to sequence:**

   - Predicted Protein
   - Predicted mRNA from genomic sequences
   - Genomic Sequence from assembly

   This brings up the following sequence, which is the protein sequence of the predicted Twinscan gene. This is in FASTA format. The > symbol designates this first line of the file as the name of the sequence. Select and copy this entire sequence, including the first line.
>contig38.002.1 length=1327

MYLFFYRYSIELQASRRAAAAVAAATVLPGSPCINQLHPTDVSSSVTVP
SIIQTVGPSDSIIPQFICENTLANATGQHHQFPQHVHNLNVGQP
HVFHDFHPAYRIFGPSQLYSQRNSNSSHFDFHFSVDGNPRFPFGSSIR
ASISRKRALSSPSFDimonialfnagSSSTAGSYGH
ISANALPMHSV/HSTLRLQIQAIHLRASAGLNPMTSQVAASGFSIHM
SASAACLTVNDVHPNSSDSSSHITTSSTLKNLDDGKGKKGHDVTFQPS
STSAGAQAQVDEASSHLSLDCYNVNNIKSIPGIDIKVSTRLDEYINC
TASTPSNEYDCANDTTIIDKDepDIEIETNCHWRSCCIEFITQDELVKI
NNDHITQNKAFVCRWDECTRGEKFQKMAVYLMVHRMRTGKEKPHCRA
YSLLENKTHLRSDRTGKEQPTCEYVGCSKAFNASDRAKHQNRTHSNEKPY
ICKAPGCTKYTDPSSLRKHKTIVHAEFVYANKKHKLGLLDDVNSRQLRDR
NSHRSHNLQEHNFSSPCSEDSHMKIGLTSSSPISIDISSSSHQQLYN
GVRASDSLLTYSFPDVAENILDLLDGWNCDDDVDVADLPILRACMVIGSG
HASASTIGGAVLARQRFRSLQRTKGINSTLMLCNPESNHTIGISELNQ
RITELKMEPTGTAEGIKIPMPTNATIGGPPEELLQNOGTRNTVNLQGQIS
TASGVSQGQRFRDSQNASTYYYGMSQSRSSQSSQVSSIPTMRSPCTCT
TTTASFYDPISSPGCSSRSSQMSNSANCYAFSTSTGLPIIINKDSNNTNAF
INKNPLGVNSVGGNSLPPPSHLATNKLRLQGKDENCYHNTSGR
FCIPSCMHLNKNSNPVQGEFEDVIANNLRRQTEVFNLLNDSTTNI
PRLSTTPNSDFITVGTNTNIASSINKDSLRELCETVPIKADMAMTSQHP
NERINLDEVEEILPDMQLYSLVKEQNHTKEHQTMEGSSYETLT
SNHREQSNIYNSKQILAPPSVDIOQPNNTIQDFPWMATIGGSFSQR
QSTTVNVPHGAKCGSPHHQCEKIINTIDIKQTPLPPAYQRTEKIP
NFQIIIDSMSTSLPELVNNVSITYQNETENIFEVHRDHDNEIQCQIGISQ
MSPTNLNNEQGISATQPISTIKLSFSETQKIVCDQTQNNSSVMHLD
YTQTLVEYVQSCQNWMTTSTNVTPRMPAPGGMQVNTTLPDVSSTHYP
GHTRMVIDMTTTSSLELSEENYLYMMQ

Navigate to FlyBase at [flybase.org](http://flybase.org)

Click on BLAST

Use the protein sequence predicted by Twinscan to search the *D. melanogaster* protein database using BLASTP.
This should result in multiple BLASTP hits (just the top 3 are shown in the screenshot below). Look familiar?

Scroll down to see the scores and E values for these alignments. Note that the E values for the first three hits are 0, which strongly suggests that this gene (ci) is the *D. melanogaster* ortholog of the *D. erecta* gene we are trying to annotate.

We'll return to FlyBase later. To get more information about the *D. melanogaster* ci gene, we will use the Gene Record Finder available at the GEP web site. To find this resource navigate to the GEP site ➔ Projects ➔ Annotation Resources ➔ Gene Record Finder. It should look like this:
Type “ci” into the search box. Note that because gene names in Drosophila are case-sensitive, the search term for the Gene Record Finder is also case-sensitive. Remember this when you search additional genes. The Gene Record for the ci gene is shown below.

To learn more about this gene click on the link below FlyBase ID. Using the FlyBase Gene Record for *ci*, you can easily determine the name of the gene as well as a lot of additional information (expression patterns, mutants, literature references, etc.) associated with this gene.
From the FlyBase Gene Report for *ci*, we see the full name of this gene is *cubitus interruptus* and that it is located on chromosome 4. We can also see that *RpS3A* and *plexB* are in the same region, which is consistent with the genomic region of *D. erecta* we saw previously in the GEP UCSC Genome Browser. In addition to the protein coding genes there are two features surrounding the *ci* gene that have a CR prefix (CR44440 and CR43957). Features with the CR prefix usually denote a non-protein coding gene (such as lncRNAs). This may require further investigation, but we'll ignore it for now.

*Click on summary information. Briefly, what is the function of the protein encoded by this gene?*

We can also learn something about the structure of the gene by clicking on GBrowse, either from the FlyBase page, or from Gene Record Finder, which yields this screen.
The GBrowse results show us the organization of translated and untranslated regions of the mRNA, which may come in handy if we need to find a small exon later. For now we will return to the Gene Record Finder page to start annotating each exon.

Construct gene model by comparing each *D. melanogaster* coding exons in the ci-RB isoform from the Gene Record Finder to the contig38 DNA sequence.

Reopen the tab for the Gene Record Finder. Type ci into the search window to get the following:

The critical information from Gene Record Finder is the organization of the exons in the *D. melanogaster* gene (circled). In this view, the 'polypeptide details' tab is selected. Note there are three isoforms ci-RA, ci-RB and ci-RC. Because each of these is different in the polypeptide view, each isoform codes for a different protein sequence. We will need to annotate each of these.
You should click on the transcript details to see the isoforms at the RNA level. In some cases there will be more unique isoforms in the transcript details section. These differ at the RNA level, but don't affect the protein sequence. For example, two RNA isoforms might differ in the 5'-UTR, but be the same for the coding exons. We will need to report these isoforms, but they don't require any further analysis.

We will begin by annotating isoform ci-RB, which has 5 exons. Some of these exons are shared with other isoforms so they will need to be annotated only once, but incorporated into the model for each isoform. Exons that are unique to the other isoforms will be annotated separately.

Each exon is designated with a number, listed in the top row of the table (7_1634_0; 6_1634_1, etc.). While these numbers are rather awkward to work with, it is best to keep track of them as we proceed.

Let's start with the first exon of isoform ci-RB. The number for this initial coding exon is 6_1629_0. Select that number in the top row. A window displaying the amino acid sequence of that exon will pop-up, as shown below.

Our goal is to locate the corresponding sequence in the *D. erecta* contig38 sequence. To do this we will perform a BLASTX comparison of the exon sequence shown above and the entire DNA sequence of contig38. Recall that the definition of a BLASTX search is to "Search a protein database using a translated nucleotide query." In this case the exon sequence will serve as the 'protein database' and the DNA sequence from the contig38 will serve as the 'translated nucleotide query'. This DNA sequence will be provided with the annotation package. However, you can also retrieve the sequence using the "get DNA" functionality of the GEP UCSC Genome.
Browser. You should already have a file that contains the DNA sequence for contig38. If not, get it from the UCSC browser as described earlier in this document.

We are now ready to begin mapping exons from Gene Record Finder to the contig38 sequence.


Click on the “blastx” link under the “Basic BLAST” section

Click on 'align two or more sequences'

This should lead to this screen:

![BLAST Screen](image)

Note: You are reminded of the types of query and subject sequences required to perform this search at the top of the BLASTX interface: the query (top box) needs to be DNA and the subject (bottom box) needs to be the amino acid sequence. Use the Browse button under the “Enter Query Sequence” section to select your contig38sequence.txt file. For the subject, return to the Gene Record Finder tab (or reopen the page if you closed it), copy the first coding exon of ci-RB from Gene Record Finder, and paste the amino acid sequence into the subject box. The search form should look like this:
Before you click BLAST, click on Algorithm parameters to expand the section. Find the 'low complexity filter' check box and unselect it (this just helps with the search and is covered in more detail in other BLAST tutorials).

Now click BLAST.

When your BLAST results are returned, scroll down a bit to find the following alignment

The first alignment shows where the exon sequence (subject) aligns with the contig38 sequence (query). Recall that the query sequence was the entire 53,840 base pair DNA sequence. However, the alignment is to an amino acid sequence. This is because blastx translated the DNA into all 6 possible reading frames, and then compared each translated sequence to the *D. melanogaster* coding exon sequence. However, the numbers (15405 and 15455) refer to the DNA sequence coordinates in contig38. We will use these numbers to locate the exon sequence
in the contig. Also notice the first alignment has an E value of 1e-06 and, importantly, is in frame +3. The frame information will be important later. Select and copy the relevant information (boxed), and paste it into a Word document. This document will serve as your notebook as you annotate this gene. Notice that the 'relevant information' includes the name of the exon ci: 6_1629_0, the frame, and the alignment. In some cases the alignment will span several lines and you will need to copy all of the lines to capture the entire alignment.

We will now return to the GEP UCSC Genome Browser to examine this region more closely and map the specific base positions that that correspond to this exon.

![Image of Genome Browser](image1)

This should look familiar by now.

We will want to zoom in to the relevant region. Recall that the alignment spans from 15405-15455, so zoom in to this region, but include a little room on each side. One way to do this is to type the desired coordinates into the search box. Let's try contig38:15390-15475. Click on jump, which should lead to this view:

![Image of Genome Browser](image2)

Remember that the black bar represents an automated BLASTX alignment against the collection of *D. melanogaster* proteins and the brown bar is the Genscan predicted gene. Notice that the alignment correlates with isoform B and that isoforms A and C begin further upstream. The Genscan program did not predict the beginning of the B isoform. Also, notice the green M in frame +3 (the frames are numbered +1, +2, and +3 from top to bottom). This corresponds to nucleotide 15405 (notice the ATG beginning at that spot in the DNA sequence).
As we build our gene model, we will need to record the specific nucleotide numbers that correspond to the relevant parts of the gene. You should record the coordinates in your Word file or in an Excel spreadsheet, so that you can easily keep track of all the coordinates (e.g. via copy and paste). In this case, the relevant gene feature is the start codon, and the nucleotide coordinate is 15405, which corresponds to the A of the ATG. The end of the blastx alignment was at nucleotide 15455. If that's the end of the first exon, what DNA sequence should be present at the beginning of the intron? Is that sequence present?

OK, so the sequence at the beginning of an intron is a GU in the RNA, which corresponds to a GT in the DNA. There is a GT near the end of our alignment. We can zoom in on that area to get a better look at exactly which nucleotide corresponds to the end of the exon. We can type in coordinates that surround that area, or, you can click on the ruler bar at the top of the sequence and select the region of interest. This might take some trial and error, but once you get the hang of it, it will make navigation of the sequence much easier. Zooming and navigating to the area near the end of the BLASTX alignment gives us the following screen:

The nearest GT is at position 15457-15458. This is also the site selected by the Genscan model. Notice that the Genscan bar stops immediately before the GT. If we decide that this is indeed the best splice donor site to use (and because our alignment strongly suggests it at this point and it agrees with the Genscan gene prediction, we can be reasonably confident), we would select the last nucleotide of the exon as the last coordinate of this exon in our gene model. This coordinate would be 15456, which is the last nucleotide before the G of the GT splice donor site. Notice that if we had just used the blastx alignment we would have been tempted to use 15455 as the coordinate. This may seem like a trivial difference, but as you probably know, a single insertion or deletion will introduce a frame shift mutation which will likely lead to a truncated protein product. As we annotate various genes, it is important to precisely specify the exon coordinates. To verify the set of exon coordinates are correct, we will use the Gene Model Checker later to verify the gene model. We will need to correct errors reported by the Gene Model Checker before we can submit our model.

**Splice site predictor.** Although we probably don't need additional evidence to support this choice for a splice donor site, this is a good point to introduce the Predicted Splice Sites track. Sometimes, two potential splice sites might make sense and we will need to distinguish between them. One source of additional evidence for a potential splice site is the splice site predictor track. To try this out, navigate back to the UCSC genome browser and zoom into the general region surrounding the above splice site.
Scroll down to the “Gene and Gene Prediction Tracks” section, and set the display mode for the “Predicted Splice Sites” track to “dense”. Hit refresh.

Scroll back to the top of the page, and notice the appearance of a blue box. Over to the left, a key tells us that this is a 'medium donor'. This means that the splice site predictor has identified this as a potential splice donor site, with a medium level of confidence. Hence the “Predicted Splice Site” track supports the splice donor we have previously selected.

A bit of a digression to explore Reading Frames and Splicing Phases:

At this point it's worth taking a slight detour to discuss the concepts of reading frame and splicing phase. A further discussion of this can be found at the GEP web site in the document "Primer on Reading Frames and Phase." For this exon our BLASTX alignment identified the translation in the +3 reading frame in *D. erecta* matches best to the first coding exon of *ci-RB*. This is the bottom of the three frames shown in the Base Position track above. The letters in the frame refer to the single letter amino acid code and have the sequence F - H - G near the splice junction. Notice that the nucleotides corresponding to the last amino acid (G, which stands for the amino acid glycine) before the splice site are GGT (this would be GGU in the RNA, but we'll keep it GGT here). However, splicing would cut the RNA between the two G's as follows G|GT, removing the GT and the remaining G would then be joined with the nucleotides in the next exon (immediately following the splice acceptor site) to form a complete codon. Therefore, the first two nucleotides of the next exon would be used to complete the codon and this would then 'maintain the open reading frame for the rest of the exon'. Because one nucleotide is 'left dangling' at the end of the first exon it is said to be in phase 1. It will need to be matched with a
phase 2 acceptor site in the next exon in order to complete the codon and maintain the open reading frame.

At the donor splice sites, the phase can be defined as the number of nucleotides between the end of the last codon and the GT. If there's 1, then it's phase 1, if 2 it's phase 2, if 0, it’s phase 0. For the acceptor site, the phase is the number of nucleotides between the AG of the intron and the first full codon. If there is 1 nucleotide in between, then the acceptor is in phase 1, if 2, phase 2, if 0, then phase 0.

This is a bit confusing (a bit?), so let's take a look at a simplified example.

In the figure below, the splice site occurs after the first A of a codon. This means that splicing will leave an incomplete codon in the first exon. When splicing to the next exon, the first two nucleotides will need to 'complete' this codon (you complete me?). In the figure below, using the A from the first exon and the first two A's from the next exon completes the codon.

In the example above, a phase 1 exon is matched with a phase 2 exon. The phases of the donor and acceptor must always be compatible with each other such that the sum of the donor and acceptor phases must either be 0 or 3.

<table>
<thead>
<tr>
<th>If the donor phase is</th>
<th>then the acceptor phase must be</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>
Once these are matched up, the next codon will be in the proper frame. In the above example CCC is the next codon. Only one reading frame is shown for simplicity. In your annotation, each coding exon could be in a different reading frame. This is OK as long as the donor and acceptor phases of the adjacent exons are compatible with each other. Remember that the reading frame reported by the BLASTX alignment is relative to the beginning of the contig sequence. As long as you use the reading frames of the coding exons as reported by BLASTX when you determine the phases of the donor and acceptor sites, the processed mRNA (with the introns removed) will maintain a single contiguous open reading frame. OK - still confusing, but with some practice you'll get it.

So, where were we? Oh yeah, annotating exon 2 of the ci-RB isoform.

Performing a blastx alignment of the second CDS (5_1629_2) of the ci-RB isoform and the entire DNA sequence of contig38 (that is, doing the same thing for CDS 2 as we did for CDS 1) leads to the following alignment:

From this alignment it looks like the exon begins near nucleotide 15521. Notice also that the reading frame for this exon is in frame +2. Even though this is different than the +3 frame from the first exon, it's OK. We will need to choose the appropriate splice acceptor site so that the reading frame switches from +3 to +2. Zooming to the region surrounding 15521 gives us the following view. The nearest potential splice acceptor (AG) is circled.

If we were to use this AG, the first nucleotide of the next exon would be the A at position 15519. Remember, that we need to use the first two nucleotides to complete the codon using the G
remaining on the first exon. That is, because the previous exon ended in phase 1, we need to match it with a phase 2 acceptor site. If we do this, using the G from the previous exon and the AT (boxed) in the current exon will form the codon GAT, then the next codon would be CCT. Notice that this CCT corresponds to the P (for proline) in reading frame +2. This is good, because the alignment of the second exon is in the +2 frame. Also notice that the alignment begins PYV …, which is what we see at the beginning of the second reading frame. So it makes sense to use this as the splice acceptor site, and the first nucleotide coordinate of this exon is 15519. Record this in the file you are using to keep track of the developing gene model.

So far we have three coordinates: the start codon at 15405, the end of the first exon at 15456, and the beginning of the second exon at 15519. We can use these coordinates to begin constructing a gene model in the following format: 15405-15456, 15519-…… We then need to repeat this process to identify the coordinates of the remaining splice donor and acceptor sites as well as the stop codon for the ci-RB isoform.

You should now repeat the blastx comparisons for each of the remaining exons to identify and record the coordinates for all the relevant splice sites and the stop codon.

Final Jeopardy music playing while we wait for you to complete this task ..........................

OK

Use your blastx results to convince yourself that the exon coordinates are indeed correct. Note, when doing blastx alignments of the different exons, remember to turn off the 'low complexity filter'. This option can be found under the 'Algorithm Parameters' section at the bottom of the BLAST configuration page. The following numbers are the correct coordinates for this gene.

15405-15456, 15519-16156, 16231-16743, 16808-16972, 17027-19516 (19517-19519)

Next, we need to input these coordinates into the gene model checker, to see if the model is 'correct'.

Verifying the Gene Model with Gene Model Checker

To confirm the accuracy of a gene model using the Gene Model Checker, proceed as follows:
2. Upload the text file of your contig sequence file into the first box (i.e. the “Fosmid Sequence File” field ) on the left of the Gene Model Checker window.
3. Type in the name of the D. melanogaster ortholog (ci-PB) in the second box– watch for this isoform to appear in the drop-down menu box as you type.
4. Enter the base number of the beginning and end of each coding sequence, in the order they appear in the table above, in the following format: # - #, # - #, # - #, etc.
5. Do not include the stop codon in the last CDS since the stop codon does not code for an amino acid. Instead, enter the stop codon coordinates field in # - # format in the box labeled “Stop Codon Coordinates” in the Gene Model Checker window.
6. If the information for your gene were on the (-) strand, you would select ‘Minus’ following ‘Orientation of Gene Relative to Query Sequence.’
7. We think all the gene’s coding sequences are accounted for, so select ‘Complete’ in the next row.
8. Use the drop-down menu to select the Project Group (D. erecta Dot) and type in the Project Name (contig38).
9. The Gene Model Checker will validate each entry as you enter them. Red boxes will appear around any entries that failed the validation. Fix these before going on.
10. Click on the ‘Verify Gene Model’ box at the bottom of the window. A checklist will appear on the right panel.

11. If there are no issues with your gene model, all the items on the checklist will be green under the “status” column and read 'pass' (see Figure below). This just means that the coordinates you input are consistent with known splice site sequences, and there are no stop codons in the middle of your gene model. It doesn't necessarily mean that the gene model is completely correct (but it's a good sign 😊).
12. If you passed the Gene Model Checker on the first try, intentionally change some of the coordinates to see what a failed gene model would look like!

13. If there are failed parts of the Gene Model Checker, click on the small + box to the left of the failed part to get more information on the problematic feature. A simple misreading or mistyping of intron-exon boundary coordinates is responsible for many of the model failures. Identify and fix these errors before proceeding to the next step.

14. We now want to see a graphical representation of the model you created. To do this, click on one of the small magnifying glass to the left of any of the items in the checklist. Your gene model will appear as a red bar in the UCSC Genome Browser. Zoom out and center the image to get a good view of your gene model with the other evidence tracks.

Final checks and preparing information for the report

A few additional checks need to be completed and a few files need to be downloaded prior to submitting your report to the GEP Project Management System. These checks and files are
generated by clicking on the various tabs in the Gene Model Checker (circled in the figure below).

For the final report we need the information found in the Dot Plot tab and the Downloads tab.

1. Click on the Dot Plot tab to generate the following figure, then copy and paste the figure into your notebook, and ultimately into your final report. The Dot Plot is a graphical representation of the alignment between the protein sequence of your gene model (Submitted Sequence on the Y-axis) and the D. melanogaster ci-PB protein sequence (X-axis). Lines in the dot plot denote regions of similarity between the two sequences. The diagonal line (with slope of ~1) indicates that the two sequences have high degree of similarity and the total lengths of the two sequences are similar. If there were any major horizontal or vertical gaps (especially near splice boundaries), then we will need to reexamine the region to determine if there is a better splice site candidate.
2. Generate a protein alignment by clicking the hyperlink labeled 'protein alignment' at the top of the Dot Plot page, then copy and paste this alignment into your notebook and ultimately into the final report. Note you can also click on the plain text version. You may need to play around with this a bit (adjust margins, font, etc.) to get it to look presentable in your final report. For example to get the following alignment to actually line up, the entire alignment was selected and changed to the font 'courier new' font size 10. In courier fonts, all characters are the same size, making it much easier to keep the bases aligned in the report. Also, while the alignment was selected, the right indent (in the ruler bar at the top of the Microsoft Word document) was adjusted slightly to the right.

```
# Program: stretcher
# Rundate: Tue 5 Feb 2013 20:08:07
# Commandline: stretcher
# -asequence /home/wilson/public_html/genechecker/trash/130205195230.q.fasta
# -bsequence /home/wilson/public_html/genechecker/trash/130205195230.s.fasta
# -aformat pair
# -outfile /home/wilson/public_html/genechecker/trash/align_130205195230.html.txt
# Align_format: pair
# Report_file: /home/wilson/public_html/genechecker/trash/align_130205195230.html.txt

# Aligned_sequences: 2
# 1: ci-PB
# 2: Submitted_Sequence
# Matrix: BLOSUM62
# Gap_penalty: 12
# Extend_penalty: 2
# Length: 1291
# Identity: 1107/1291 (85.7%)
# Similarity: 1174/1291 (90.9%)
# Gaps: 17/1291 (1.3%)
# Score: 5701

<table>
<thead>
<tr>
<th></th>
<th>ci-PB</th>
<th>Submitted_Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1 MEQLYSLQRTNSASSFHPYVNCASAFHLAGLGLGSADFGLGSALARGLSGS</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>1 MEQLYSLQRTNSASSFHPYVNCASAFHLAGLGLGSADFGLGSG</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>51 LHNAAVAAAAGSLASTDFHFSVDGRNLGSPPGGSIRASISRKALS</td>
<td>100</td>
</tr>
<tr>
<td></td>
<td>97 LHHAAVAAAALSTDFHFSVDGRN---PPGGSIRASISRKALS</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>97 LHHAAVAAAALSTDFHFSVDGRN---PPGGSIRASISRKALS</td>
<td>96</td>
</tr>
<tr>
<td></td>
<td>101 SSSPYSDFDSINSMIRFSNPNSLATIMNGSRGRAASGSYHISATALNPMS</td>
<td>150</td>
</tr>
<tr>
<td></td>
<td>97 SSSPYSDFDSINSMIRFSNPNSLATIMNGSRGSSTASGSYHISANALNPMS</td>
<td>146</td>
</tr>
<tr>
<td></td>
<td>151 HVHSTRLQQIQAHLLRAASNLNPMTDNAASGFSISMTPSVTASSLRVND</td>
<td>200</td>
</tr>
</tbody>
</table>
```
Submitted_Seq  147 HVHSTRLQQIQAHLRASAGLNPMTSQVQAASGSF1GHMSASACLRNVN  196

**ci**-**PB**
201 VHPNLDSDHIQTITSPTV---TKDVSVQVPAAAFSLKNNLDDEAREKGGPFKD  247

Submitted_Seq  197 VHPNLDSPSIHHIITSTVRLNEDNPQIAAALSINLNLDDGGKGGKFKD  246

ci-PB  248 VVPEQPSSTSGVQQEADASSQSLDCRYNNVTGPGDPVDVKNVSRL  297

Submitted_Seq  247 VVTEQPSSTGAVQQEADASSHLSDCRYNNVNNISIPDIDKTVSTRL  296

ci-PB  298 DEYINCQISIPSNEYCANADTTDKDEPGFETNCHWRSRCIEFITQ  347

Submitted_Seq  297 DEYINCQTASTPSNEYCANADTTDKDEPGFETNCHWRSRCIEFITQ  346

ci-PB  348 DELVKKHINNDIQTNKAFVCRWEDCTRGEKPKFAQYMLVHMRRTGEK  397

Submitted_Seq  347 DELVKKHINNDIQTNKAFVCRWEDCTRGEKPKFAQYMLVHMRRTGEK  396

ci-PB  398 PHKCTFEGCFKAYSRELNKLTHLRSHTGEKFYCEYPGCSKAFNASDRA  447

Submitted_Seq  397 PHKCTFEGCFKAYSRELNKLTHLRSHTGEKFYCEYPGCSKAFNASDRA  446

ci-PB  448 KHQNRTSNEKYPICKAPGCTKRTDPSSLRHKVHQTGAEFYANKKHKG  497

Submitted_Seq  447 KHQNRTSNEKYPICKAPGCTKRTDPSSLRHKVHQTGAEFYANKKHKG  496

ci-PB  498 LPLDDVNSRLQRDNHSHRNQEHNNIDSPCSEDSHMGKILGTSSPIKS  545

Submitted_Seq  497 LPLDDVNSRLQRDNHSHRNQEHNNIDSPCSEDSHMGKILGTSSPIKS  546

ci-PB  546 ESDISSNNHLTVNGVRAEDSLLTYSPPDLAENNLDDGWCDDDDVADL  595

Submitted_Seq  545 ESDISSNNHLTVNGVRAEDSLLTYSPPDLAENNLDDGWCDDDDVADL  594

ci-PB  596 PIVLRAVMVIGINNASASTIGSGVLRQFRGRQLQTKGINSSTMLCNIP  645

Submitted_Seq  595 PIVLRAVMVIGINNASASTIGSGVLRQFRGRQLQTKGINSSTMLCNIP  644

ci-PB  646 ESRNTFGISELNRQTILEKMEPTDAGAIEKIPLPNTITGYTEDEPLQNT  695

Submitted_Seq  645 ESRNTFGISELNRQTILEKMEPTDAGAIEKIPLPNTITGYTEDEPLQNT  694

ci-PB  696 SFRNTVSNKQG--TVGSIGQCPFQRDSQNSTASYGMSQSRRSS SQSV  743

Submitted_Seq  695 SFRNTVSNKQG--TVGSIGQCPFQRDSQNSTASYGMSQSRRSS SQSV  742

ci-PB  746 SSIPMTRPNPCSNST--ASYPDIPSCSRSSQMSGANCSPTSTSGLP  792

Submitted_Seq  745 SSIPMTRPNPCSNST--ASYPDIPSCSRSSQMSGANCSPTSTSGLP  791

ci-PB  796 VLNKESNKLNACNKPNIGVQVGITYNSSLPPSSHLIATNLQRLQK  842

Submitted_Seq  795 VLNKESNKLNACNKPNIGVQVGITYNSSLPPSSHLIATNLQRLQK  841

ci-PB  843 DSEEYHNFTSGRFSVPSYMSHLHIKNNKPVGENEFDKANASRA-RQTD  889

Submitted_Seq  842 DSEEYHNFTSGRFSVPSYMSHLHIKNNKPVGENEFDKANASRA-RQTD  888
### SubmittedSeq

<table>
<thead>
<tr>
<th></th>
<th>SubmittedSeq</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>847</td>
<td>DSENCYHNFTSGRFCIPSCMHSLSLHMKNSNPVQNEFSDKVIANNTLRRQTE</td>
<td>896</td>
</tr>
<tr>
<td>890</td>
<td>PVPNILDPLTISRFSTTPHSFDINVGKTNNIASSINKDNLRLKSDLFTVS</td>
<td>939</td>
</tr>
<tr>
<td>897</td>
<td>PVPNILDLSLNPRLSTTPNSFDITVGKTNNIASSINKDNLRLKELCTV</td>
<td>946</td>
</tr>
<tr>
<td>940</td>
<td>IKADAMMTSDQHPNERINLDEVEELILPDEMLQYLANLWLKDTHNEKEHQ</td>
<td>989</td>
</tr>
<tr>
<td>947</td>
<td>IKADAMMTSDQHPNERINLDEVEELILPDEMLQYLSDKEDTHNEKEHQ</td>
<td>996</td>
</tr>
<tr>
<td>990</td>
<td>AVPVGNSVSETIASNYRESNIYYTNKQILTPGTVPTTPNTKFTVQD</td>
<td>1039</td>
</tr>
<tr>
<td>997</td>
<td>TEAMGSSVYETLTSHNYREQSNIQHQLAPANSVDVDIQPNTTN-TIQD</td>
<td>1045</td>
</tr>
<tr>
<td>1040</td>
<td>KFMTAVGSGFSQRELSTLVAPNEHGHAKCESFHHSQKYMNTDIGSKQQ</td>
<td>1089</td>
</tr>
<tr>
<td>1046</td>
<td>KFPMTAIGGSGFSQRQSSLTVVPNEHGHAKCGSFHHQCEKIINTDIKQQ</td>
<td>1095</td>
</tr>
<tr>
<td>1090</td>
<td>SALPSAHQRTKESYNYIQIDSSMTSLPLENVSIYPRNETENIFKVGHD</td>
<td>1139</td>
</tr>
<tr>
<td>1096</td>
<td>TPLPPAYQRTKPNFQIDSSMTSLPLENVSYITQNETENIFVHRD</td>
<td>1145</td>
</tr>
<tr>
<td>1140</td>
<td>HDNEIQCGIQSSQYMSPTSNLNNDDGYFSTNMQTPITTSKLHFPEPKIVC</td>
<td>1189</td>
</tr>
<tr>
<td>1146</td>
<td>HDNEIQCGIISQSMPSPTSNLNNQUQISTANIPFITISKLFSSETQKIVC</td>
<td>1195</td>
</tr>
<tr>
<td>1190</td>
<td>DTQASNTSVMLHTQRTLECYQSNCNWTNNTSTQIQSLPG-MPVNN</td>
<td>1238</td>
</tr>
<tr>
<td>1196</td>
<td>DTQTVNNSSVMHLDHTQRTLEVQSCQWMTNVTNPIQAPPGGMQVNT</td>
<td>1245</td>
</tr>
<tr>
<td>1239</td>
<td>TLFPDVSSTHTHYHTNVMINDTSSLTSLEENRYLQM</td>
<td>1279</td>
</tr>
<tr>
<td>1246</td>
<td>TLFPDVSSTHTHYHTNVMINDTSSLTSLEENRYLQM</td>
<td>1286</td>
</tr>
</tbody>
</table>

3. To access files that need to be downloaded for project submission, click on the “Downloads” tab to get the following screen.

**Right-click on the links below to save the files required for project submission:**

- **GFF File**
- **Transcript Sequence File**
- **Peptide Sequence File**

Right click on each file and save the files to a folder on the Desktop of your computer, or other convenient location (such as a jump drive).

The filenames will appear something like the following when saved: 130128204226.fasta; 130128204226.gff; 130128204226.pep. The numbers incorporate the date the files were saved into a file name. The extensions refer the file type: .fasta is the transcript file, .pep is the peptide file, and .gff contains the coordinates of your gene model from Gene Model Checker. By default, Mac OS X and MS Windows will hide known file extensions so the fasta extension might not
show up on your computer. You should rename these files to the name of the isoform, for example ci_RB.fasta.

**Project Submission**

Although we won't be submitting the practice annotation of the ci-RB isoform, for your final project, once the gene model is complete, the annotation information for your project must be submitted to the GEP using the Project Management System.

A brief overview of the submission procedure is as follows.

1. Navigate to the annotation submission report form via the GEP web site by following these links: GEP ➔ curriculum ➔ course materials ➔ Washington University ➔ GEP specific issues ➔ GEP annotation report. Download the MS Word version of the report form.

2. Fill out this form. A completed sample report for the ci-RB isoform annotated in this tutorial is included below.

3. Download and merge all the GFF, transcript and peptide files using the Annotation File Merger. (GEP ➔ Projects ➔ Annotation Resources ➔ Annotation Files Merger)

4. Once the form is filled out, and all GFF, transcript, and peptide files have been merged together, navigate to the Annotation Submission form in the Project Management System.

**GEP ➔ Projects ➔ Project Management system to this screen**

The GEP faculty will need to log in to this site, leading to the screen on the following page
Click on 'Summit Annotation Projects'. You will need to have claimed the project previously before you can submit the results. Click on your project and follow instructions for submission and upload all files as instructed.

And you're done, now that wasn't so hard was it? 😊

**ASSIGNMENT:**

Repeat the annotation process for the ci-RA or ci-RC isoform.

Turn in to your professor, by email, a completed report with all necessary figures pasted in as described above. Also download the .gff, .fasta, and .pep files and attach these along with your submission.

*A completed report for the ci-RB isoform is shown starting on the next page*
GEP Annotation Report

Note: For each gene described in this annotation report, you should also prepare the corresponding GFF, transcript and peptide sequence files as part of your submission.

Student name: ________________________________
Student email: ________________________________
Faculty Advisor: ________________________________
College/University: ________________________________

Project details
Project name: __derecta_dot_Aug2006_contig38__________________
Project species: ___D. erecta__________________
Date of submission: __2/5/2013_________
Size of project in base pairs: ___53,840_________
Number of genes in project: ________________

Does this report cover all genes and all isoforms or is it a partial report? ___Partial report___
If this is a partial report because different students are working on different regions of this sequence, please report the region of the project covered by this report:
from base ___15405 ________ to base 19519__________

Instructions for project with no genes
If you believe that the project does not contain any genes, please provide the following evidence to support your conclusions:

1. Perform a BLASTX search of the entire contig sequence against the non-redundant (nr) protein database. Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the nr database as to why they do not correspond to real genes in the project.

2. For each Genscan prediction, perform a BLASTP search using the predicted amino acid sequence against the protein database (nr) using the strategy described above.

3. Examine the gene expression tracks (e.g. cDNA/EST/RNA-Seq) for evidence of transcribed regions that do not correspond to alignments to known D. melanogaster proteins. Perform a BLASTX search against the nr database using these genomic regions to determine if the region is similar to any known or predicted proteins in the nr database.
Complete the following Gene Report Form for each gene in your project. Copy and paste the sections below to create as many copies as needed. Be sure to create enough Isoform Report Forms within your Gene Report Form for all isoforms.

Gene report form

Gene name (i.e. *D. mojavensis* eyeless): ___*D. erecta cubitus interruptus__

Gene symbol (i.e. *dmoj_ey*): ___dere_ci__

Approximate location in project (from 5’ end to 3’ end): ___15405-19519__

Number of isoforms in *D. melanogaster*: ___3___

Number of isoforms in this project: ___1___

Complete the following table for all the isoforms in this project:

*If you are annotating untranslated regions then all isoforms are unique (by definition)*

<table>
<thead>
<tr>
<th>Name of unique isoform based on coding sequence</th>
<th>List of isoforms with identical coding sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>ci-RB</td>
<td>ci-RB (in some cases isoforms will differ at the RNA level, but have the same coding sequence. List all of these here. Also upload the GFF, transcript, and pep files for these isoforms)</td>
</tr>
<tr>
<td>Other isoforms will differ at the protein level, e.g. ci-RA and ci-RC in our example. List those in separate boxes here.</td>
<td>If there are multiple isoforms coding for the same proteins for these, list them here. Also upload the GFF, transcript, and pep files for these isoforms</td>
</tr>
</tbody>
</table>

Note: For isoforms with identical coding sequence, you only need to complete the Isoform Report Form for one of these isoforms (i.e. using the name of the isoform listed in the left column of the table above). However, you should generate GFF, transcript, and peptide sequence files for ALL isoforms, irrespective of whether they have identical coding sequences as

Isoform report form

Complete this report form for each unique isoform listed in the table above (copy and paste to create as many copies of this Isoform Report Form as needed):

Gene-isoform name (i.e. *dmoj_ey*-PA): ___dere_ci-RB__

Names of the isoforms with identical coding sequences as this isoform

----------------------------------------------------------------------------------------------------------------------------------

----------------------------------------------------------------------------------------------------------------------------------
Is the 5’ end of this isoform missing from the end of project: ____no___________ If so, how many exons are missing from the 5’ end: ___________

Is the 3’ end of this isoform missing from the end of the project: ____no___________ If so, how many exons are missing from the 3’ end: ________________

1. **Gene Model Checker Checklist**

Enter the coordinates of your final gene model for this isoform into the Gene Model Checker and paste a screenshot of the checklist results below:

2. **View the gene model on the Genome Browser**

Using the custom track feature from the Gene Model Checker (see page 10 of the Gene Model Checker user guide on how to do this; you can find the guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at [http://gep.wustl.edu](http://gep.wustl.edu)). Capture a screenshot of your gene model shown on the Genome Browser for your project; zoom in so that only this isoform is in the screenshot. Include the following evidence tracks in the screenshot if they are available.

1. A sequence alignment track (D. mel Protein or Other RefSeq)
2. At least one gene prediction track (e.g. Genscan)
3. At least one RNA-Seq track (e.g. RNA-Seq Alignment Summary)
4. A comparative genomics track
   (e.g. Conservation, D. mel. Net Alignment, 3-way, 5-way or 7-way multiz)
Paste the screenshot of your gene model as shown on the Genome Browser below:

![Genome Browser Screenshot]

3. Alignment between the submitted model and the *D. melanogaster* ortholog

Show an alignment between the protein sequence for your gene model and the protein sequence from the putative *D. melanogaster* ortholog. You can use the protein alignment generated by the Gene Model Checker or you can generate a new alignment using BLAST 2 Sequences (*bl2seq*). **Copy and paste the alignment below:**

```plaintext
| ci-PB       | 1 MEQLYSLQRTNSASSFHDPYVNCASAFHLAGLGLGSADFGLGSRGGLSLGE | 50 |
| Submitted_Seq | 1 MEQLYSLQRTNSASSFHDPYVNCASAFHLAGLGLGSADFGLGSRGGLSLGE | 50 |
| ci-PB       | 51 LHNAAVAAAAGSLASTDFHFSVDGNRRLGSPRPFGGSRASRASRKLARS | 100 |
| Submitted_Seq | 51 LHNAAVAAAAGSLASTDFHFSVDGNR----PRPGGSASRASRKLARS | 96 |
| ci-PB       | 101 SSPPSSFDINSFMIRFSNLSLATIMNGSGASSAASGYSYGHSATGNPMMS | 150 |
| Submitted_Seq | 97 SSPPSSFDINSFMIRFSNLSLATIMNGSGASSAASGYSYGHSANALPNMS | 146 |
| ci-PB       | 151 HVHSTRLOIQAHLLRASAGLNPMTPOQVAASGFSIGMHSATGNPMMS | 200 |
| Submitted_Seq | 147 HVHSTRLOIQAHLLRASAGLNPMTQVAASGFSIGMHSASACLRVN | 196 |
| ci-PB       | 201 VHPNLSDSHIGITTSPTV---TKDVSOVPAASLKNKDDLAREKKGFKD | 247 |
| Submitted_Seq | 197 VHPNLSDSHIGITTSPTV---TKDVSOVPAASLKNKDDLAREKKGFKD | 246 |
| ci-PB       | 248 VVPEQPSSTSGVQAQVEADSASSQLSDRCNYNNTGFPVGDKVNSRL | 297 |
| Submitted_Seq | 247 VVPEQPSSTSGVQAQVEADSASSQLSDRCNYNNTGFPVGDKVNSRL | 297 |
| ci-PB       | 298 DEYINCGRSIPSEYDACNADTIDKDEPGFITCNHWRSCTFEITQ | 347 |
| Submitted_Seq | 297 DEYINCGRSIPSEYDACNADTIDKDEPGFITCNHWRSCTFEITQ | 346 |
| ci-PB       | 348 DELVKBHDHDTQNKKAFCVCRWEDCCTRGEKPKAQMVLVVHMRRTGEK | 397 |
| Submitted_Seq | 347 DELVKBHDHDTQNKKAFCVCRWEDCCTRGEKPKAQMVLVVHMRRTGEK | 396 |
```
4. Dot plot between the submitted model and the *D. melanogaster* ortholog

Paste a copy of the dot plot of your submitted model against the putative *D. melanogaster* ortholog (generated by the Gene Model Checker). Provide an explanation for any anomalies on the dot plot (e.g. large gaps, regions with no sequence similarity).

**Note:** Large vertical and horizontal gap near exon boundaries in the dot plot often indicates that an incorrect splice site might have been picked. Please re-examine these regions and provide a detail justification as to why you have selected this particular set of donor and acceptor sites.
Preparing the project for submission

For each project, you should prepare the project GFF, transcripts and peptide sequence files (for ALL isoforms) along with this report. You can combine the individual files generated by the Gene Model Checker into a single file using the Annotation Files Merger.

The Annotation Files Merger also allows you to view all the gene models in the combined GFF file within the Genome Browser. Please refer to the Annotation Files Merger User Guide for detail instructions on how to view the combined GFF file on the Genome Browser (you can find the user guide under “Help” -> “Documentations” -> “Web Framework” on the GEP website at http://gep.wustl.edu).

Paste a screenshot (generated by the Annotation Files Merger) with all the gene models you have annotated in this project.

For the practice annotation you are only annotating one isoforms so you don't need to merge these files. For your final project you may need to merge the files for several isoforms of a gene.

The Annotation Files Merger is found on the GEP website at

GEP ➔ Projects ➔ Annotation Resources ➔ Annotation Files Merger

Have you annotated all the genes?

For each region of the project with gene predictions that do not overlap with putative orthologs identified in the BLASTX track, perform a BLASTP search using the predicted amino acid sequence against the non-redundant protein database (nr). Provide a screenshot of the search results. Provide an explanation for any significant (E-value < 1e-5) hits to known genes in the nr database and why you believe these hits do not correspond to real genes in your project.

For example, if there’s a Genscan prediction, but no corresponding blastx hit, click on the Genscan gene model, get the predicted protein sequence and use this in a blastp search of the D. melanogaster nr protein database (the nr database, not the RefSeq database).

References:

The GEP website:  http://gep.wustl.edu