Lab Week 10: Introduction to the Genomics Education Partnership and Collaborative Genomics Research in *Drosophila*

1. Introduction

Overview of the Genomics Education Partnership

The Genomics Education Partnership (GEP) is a national, collaborative, scientific investigation that uses comparative genomics to engage students in genomics research within regular academic-year laboratory courses (http://gep.wustl.edu). Using a versatile curriculum that has been adapted to many different class settings in over 90 primarily undergraduate institutions, GEP undergraduates undertake wet-lab projects to improve draft genomic sequences and/or computer analyses of the data (including annotation of genes, assessment of repeats, exploration of evolutionary questions, etc.).

At present, the research problem entails generating finished sequence from the fourth (dot) chromosome (Figure 1.1) of various species of *Drosophila*, annotating these sequences, and making comparisons among species and to control DNA sequences from the third chromosome to discern patterns of genome organization related to the control of gene expression. GEP undergraduates have improved over 2.4 million bases of draft genomic sequence from several species of the fruit fly *Drosophila* and have produced hundreds of gene models using evidence-based manual annotation.

![Figure 1.1](image.png)

*Figure 1.1. Drosophila melanogaster* chromosomes. The small fourth (dot) chromosomes are at the red arrows (from Painter, 1934; used with permission from Oxford University Press).
The scientific interest is based on observations that the dot chromosome shows a mixture of heterochromatic and euchromatic properties. The chromosome stains intensely with fluorescent DNA-binding dyes, has a high density of repetitious sequences, is late replicating, and exhibits very low meiotic recombination, all properties of heterochromatin. At the same time, the distal 1.2-million base pair (Mb) region of the dot chromosome is amplified in polytene chromosomes, a property of euchromatin, and codes for ~80 genes, a gene density similar to that found in the euchromatic arms. Actively-transcribed regions of DNA are typically associated with euchromatic domains, while transcriptionally-silent regions of DNA are generally associated with heterochromatic domains (see pp. 215-216 and 357-359 in the Life, 9th edition textbook). Thus, understanding chromosome organization and chromatin effects on dot chromosome gene expression requires careful analysis, not just of the genes present but also of the type and distribution of repetitive elements. While this amalgamation of heterochromatic and euchromatic properties of the dot chromosome is unusual in Drosophila, the composition of the dot chromosome (30% repetitious DNA) is actually similar to euchromatic regions in a mammalian genome. Hence, the dot chromosome is an ideal candidate for studying the effects of chromatin packaging on regulating gene expression, and the results from comparative genomics can provide many insights into these questions.

DNA sequence improvement is usually done in the context of an entire semester course or independent research project. However, some publicly-available DNA assemblies are of sufficient quality that no additional improvement is needed. In addition, fosmid DNA that is successfully improved by GEP students is then deposited into the GEP’s Project Management System for future annotation. Thus, all members of the Genomics Education Partnership have the opportunity to contribute to the completion of an annotation research project, even if they do not do the sequence improvement themselves.

An additional goal of the GEP is for students to gain a more sophisticated understanding of eukaryotic gene structure and mRNA processing than is often achieved in lectures. This will be the focus of this week’s Biology 191 lab. Pages 2-13 provide a detailed introduction to the topic of gene annotation and should be read carefully. During lab, you will then work through A Sample Annotation Problem to learn the basics of gene annotation, by mapping the protein-coding regions of a relatively-uncomplicated gene that has just three exons and two introns. In the process, you will be introduced to a variety of on-line bioinformatics databases. Students who go on in biology and/or medicine will most likely work with these databases again, so understanding how to make use of them now will serve you well in the future.

What is Gene Annotation?

The first complete genome sequences of a bacterium (Haemophilus influenza) and a eukaryote (Saccharomyces cerevisiae) were published in consecutive years in the mid-1990’s (Fleischman et al., 1995; Goffeau et al., 1996). Since that time, hundreds of organisms have had their genomes sequenced. However, while continually-updated technologies are rapidly increasing the number of sequenced genomes, ascribing functions to various parts of the genome is nowhere near as automated. In the process of annotation, decisions are made as to which particular DNA sequences within a genome contain biological information. This involves identifying features in the DNA sequences and, in many cases, determining which of these features are likely to be genes (Walter and Wilkerson, 2006).
The expression of protein-coding genes in cells employs mRNA intermediates. Gene annotation in complex eukaryotes is complicated by the fact that the coding regions of most protein-coding genes are interrupted, consisting of coding **exons** and non-coding **introns**. Introns are often much longer than the exons in many genes, and the introns 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 1.2). Human annotators combine these (often contradictory) computational predictions with DNA sequence alignment and mRNA expression data to create the best-supported gene model (Walter and Wilkerson, 2006).

![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 1.2.** 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 members 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 for the **orthologous** sequences follows an overview of the organization of the gene annotation lab.

**Overview of the Biology 191 Gene Annotation Lab**

In this lab, you will annotate one specific **feature** (putative gene) of a 45,000-bp **fosmid** of DNA from *Drosophila biarmipes*, a close relative of the widely-used model organism, *Drosophila melanogaster*. Genomic sequences from closely-related species are generally straightforward to annotate. Your lab instructor will first provide an overview of the preliminary analyses of the *Drosophila biarmipes* fosmid, in which the number of features in the fosmid, the number of **isoforms** for each feature, and the relative complexity of the different features was pre-determined (by scientists at Washington University in St. Louis) using various gene predictors. You will then work through **A Sample Annotation Problem**, filling in answers to questions in the worksheet and filling out an Annotation Report as you go along. This exercise will reinforce the concept of interrupted genes in eukaryotes, which consist of coding exons and non-coding introns, and guide you through the basic annotation procedure of mapping the protein-coding regions of a gene. You will also be introduced to the GEP’s Gene Model Checker in this activity, so you know how to assess the quality of your newly-generated gene model.
Once you have passed (to your TA’s or instructor’s satisfaction!) all parts of the Gene Model Checker, you and your partner should then complete and submit (to Moodle) an Annotation Report of the mapped gene. If all goes well, you will complete most, if not all of the Annotation Report during your lab meeting time. Students who do not finish the Annotation Report during lab have up to one week from lab to submit the report. If time allows, we will talk a bit at the end of lab about the selected gene and what it encodes (if it is known), the challenges you faced with the annotation, (e.g., any problems or ambiguous spots that you encountered) and anything of interest that you learned about the gene by annotating it.

Basic Biology

Before we consider types of annotation evidence in detail, we will review a few details of basic molecular biology, to guide you in your generation of a gene model. While it is impossible in a short tutorial to cover all the relevant basic biology (you should already know about transcription and translation), there are a few specific details that should be discussed.

![Gene Model Structure](image)

**Figure 1.3.** Eukaryotic gene structure. The donor (GT) splice site and the acceptor (AG) splice site are the first two and last two nucleotides, respectively, of the DNA sequence of each intron (in red above) (Ben-Hur et al., 2008). (Figure used with permission from S. Sonnenburg, G. Rätsch & B. Schölkopf)

**Introns.** Unlike bacteria, many genes in eukaryotes have introns (Figure 1.3). These sequences are removed from the primary RNA transcript based on sequences found within the intron. In cells, splicesomes recognize DNA consensus sequences to remove the introns and join the exons together, creating an mRNA with a single open reading frame. In gene annotation, it is likewise critical to find the precise beginning and end of each exon, so that the resulting gene model also encodes a full-length polypeptide. The sequence at the beginning (5' end) of the intron is called the **donor site**, while the 3' end of the intron is called the **acceptor site.** Most eukaryotic introns begin with the nucleotide sequence 'GT' and end with the sequence 'AG' (the
GT-AG rule) in the genomic DNA. However, the larger consensus sequences that define intron donor and acceptor sites have a lot of tolerance for mismatches and can evolve quite quickly.

For the purposes of your analysis, you can identify putative intron/exon boundaries in three ways. First, you can assume that any de novo gene prediction program will predict donor and acceptor sites consistent with the basic biology of splice site sequence composition, thus any gene model generated by a de novo gene prediction site can be used to help you pick splice sites. Second, a computer program designed to find and score putative donor and acceptor sites has been run on all GEP fosmid sequences. The results of this analysis are displayed in the GEP mirror of the UCSC Genome Browser for your sequence. To simplify the results, the potential sites predicted using this program have been split into high, medium, and low quality; sites of any quality can be used as part of your gene model. Finally and probably the least reliable way to find donor and acceptor sites is to look at the sequence by eye and scan for the base sequences known to be used by the splicing mechanism. While searching by eye is the lowest quality of evidence for the prediction of an intron/exon boundary, it is often the only evidence for a given splice site. In this case, any GT before a stop codon can be considered as a potential intron donor site, while any AG is a possible acceptor site. In all cases, conservation of exon length between the reference Drosophila melanogaster sequence and the sequence being annotated should be a guiding factor in choosing between possible intron splice sites (see Conservation section below).

In rare cases (in the range of about 1 in 100), a “non-canonical” donor site with the sequence GC (instead of the canonical GT) has been found in Drosophila melanogaster. Non-canonical donor sites will never be used in gene models generated by most de novo gene predictors; however, evidence collected so far in annotation of Drosophila virilis suggest that these GC donor sites are also used in a few genes in this species. Of the non-canonical sites found in Drosophila virilis, about 50% of the time the same non-canonical site is used in Drosophila melanogaster.

Finally, it should be noted that examination of a large number of introns in Drosophila melanogaster establishes the minimum size of an intron as 43 bases (Guo et al., 1993; Talerico and Berget, 1994). It is reasonable to assume that this limit also exists in the other Drosophila species. Thus, any gene model that predicts the presence of an intron smaller than 42 bases is highly unusual and would require expression data (e.g., EST, RNA-seq) to support the hypothesis.

mRNA structure. Once the start codon, the stop codon, and all the intron/exon boundaries have been identified, you will use the GEP’s on-line Gene Model Checker to create the final coding sequence of the mRNA as well as the predicted amino acid sequence of the encoded protein. Remember that in eukaryotes, each processed, mature mRNA contains a single open reading frame that extends from the start codon all the way through all the internal exons and ends with a stop codon. Your gene model should likewise produce a putative message that contains a single long open reading frame with no internal stop codons. If the Gene Model Checker report shows that your gene model has internal stop codons, you should double check and adjust your intron/exon boundaries until no internal stop codons are found.
2. Conceptual Guide to Gene Annotation (warning - this is slow going!)

The basic idea when attempting to create a gene model for any feature is to determine a series of base pair coordinates that describes the structure of the gene. In this lab, you will map only the protein-coding regions of one mRNA molecule. The coordinates assigned in this case then describe the base position of the beginning and end of each coding sequence, which together make up most of the exon sequences in the final (mature) mRNA. Your gene model must be consistent with what is known about the basic biology of transcription, mRNA splicing, and translation. For example, since it is known that RNA polymerase does not hop back and forth between the two strands of a double-stranded DNA molecule, a gene model typically does not include sequences from both strands (see McManus et al., 2010 for exceptions). Your gene model must start on one strand, continue down the length of that strand and end on the same strand. Your gene model must also include the base position of the start codon for translation, the position of the beginning and ending of each coding region and the position of the stop codon of translation.

Your job as a gene annotator is to learn as much as you can and apply new skills and knowledge to come up with your best gene model for the selected feature from *D. biarmipes*. Your first step in annotation will be to collect and consider all the evidence you can gather about the sequence you are annotating. Once you have gathered the available evidence, each piece of evidence should be weighed against all the other evidence and used to make your gene model. The goal is to make the best gene model you can that integrates all the evidence in a way that maximizes the use of high-quality evidence, avoids internal conflicts and only uses low-quality evidence when no higher-quality evidence can be found.

The types of evidence fall into three basic categories: expression, conservation and computation. **Expression** evidence is derived from sequencing mRNA that has been isolated from the organism of choice (*D. biarmipes* in this case). The RNA sequences are then typically mapped back to the genomic DNA to provide evidence of transcription. The use of expression data (**expressed sequence tags** [ESTs], mRNA sequences, or RNA-sequencing data) is the best way to identify the coordinates of a full-length, mature mRNA transcript, including the 5' and 3' **untranslated regions** (UTRs). We do not have time to map the coordinates of the UTRs in this lab. However, if interested, there are two activities (**Exercise #2: Using mRNA and EST Evidence in Annotation** and **Browser-Based Annotation and RNA-Seq Data**) that describe how to use expression data in gene annotation on the GEP web site at: [http://gep.wustl.edu/curriculum/course_materials_WU/annotation/annotation_exercises](http://gep.wustl.edu/curriculum/course_materials_WU/annotation/annotation_exercises).

**Conservation** defines those types of evidence that rely on the assumption that the new species being annotated had a common ancestor with *Drosophila melanogaster*. Since you are not using expression data in this lab, **conservation will be your most important evidence in constructing a gene model**. Based on the principal of Occam’s razor, which declares that the best explanation for anything is the one with the fewest assumptions, the best gene model in a new species is usually the model that assumes the fewest differences between it and the gene model in *Drosophila melanogaster* (i.e. the one that has the most similarity).

The third general type of evidence is **computational**. Many computer programs have been created that attempt to recognize various features in DNA sequence. Several of these
programs have already been run on the sequences you will be annotating, and the results are available for viewing on various genome browsers. These programs are designed to identify evidence for a wide variety of biological features including genes, repeats, or various other features (e.g., intron/exon boundaries). Each of these programs has been optimized for its given purpose and as such, provides at least a hint as to a possible biological function of any given sequence. Without expression and conservation evidence, computational analysis is usually the only evidence you can fall back on to create your gene models.

Finally, if expression, conservation, and computational evidence fail to provide enough evidence for a given gene model, a few simple rules can be used to assist in creating a gene model. These rules are based on philosophical consideration of how best to “get things wrong” and are discussed in the Summary and outline that follows. Additional documents that describe techniques for tackling annotation projects with a range of difficulties are available on the GEP web site at http://gep.wustl.edu/curriculum/course_materials_WU/annotation/specific_issues.

Expression

Evidence of expression from cDNA/EST/RNA-Seq tracks can provide strong evidence for the locations of intron/exon boundaries. However, there are important factors to consider when evaluating expression data. First, some genes are expressed at very low levels or only for very short periods of time. As such, the lack of expression data for any gene should not be considered strong evidence in favor of discounting the presence of a gene. No matter how much expression data are collected, there will always be real genes that lack expression evidence. Another issue to be aware of is the high level of noise found in RNA-seq data. For example, we have observed a non-trivial amount of variation in the apparent splice site evidence in RNA-seq data that have been mapped across introns (i.e., TopHat track). In situations like this where there is a lot of variation in the evidence for the exact position of a splice donor/acceptor site, we will follow a “majority rules” approach.

Conservation

Conservation can take many forms and all of the following should be considered when generating a gene model. They are presented here in order of importance with the most important first:

Conservation of primary amino acid sequence. This is certainly one of the most important forms of evidence to guide you in construction of your gene models. It is reasonable to assume that for almost any gene found in the various Drosophila species, the encoded protein is serving a very similar if not identical function in the different species (i.e. the proteins are serving as functional orthologs). As such, one would expect the amino acid sequences to be very similar. Your use of programs that search for similarity to identify regions of similar amino acid sequence will be the foundation upon which you will build your gene models. Conservation of this type is found using computer programs like BLAST and Clustal. When conservation between the two species is very high, the identification of intron/exon boundaries is easy since the boundary will be very close to the end of the alignment. As the extent of amino acid conservation decreases, the identification of intron/exon boundaries will need to rely on other evidence discussed below. See the Advanced Annotation Instruction sheet for more on Clustal analysis and what to do if the default BLAST search fails.
Conservation of gene structure. The creation or removal of an intron in orthologous genes is a very rare event, even over evolutionary time scales. This means that the best gene model in *Drosophila biarmipes* will almost always have the same basic structure (i.e., the same number of introns and exons) as the gene model in *Drosophila melanogaster*. This rule however is not absolute; sometimes the only gene model that fits most of the evidence has a new or missing intron, so if you can find no way to construct a gene model that maintains the same number of exons, go with a gene model that keeps the total number of exons as close to *Drosophila melanogaster* as possible.

Conservation of exon length. In a surprisingly-large number of cases, we find exons that have a very similar length even when there is no detectable conservation of the encoded amino acids found near the intron/exon boundary. This has happened enough that we can come to consider more carefully any putative donor or acceptor sites that conserve exon length. For example, consider the following alignment in which BLASTX was used to find similarity between a piece of *Drosophila virilis* genomic DNA sequence and the sequence of a 45 amino acid long exon from *Drosophila melanogaster*.

Exon sequence:

```
1    CGSVVPSADYAYSPAYTQYGTTYGSYGTSSGLIYPAS
41   GPIT
```

BLAST alignment:

```
Query: 14253 CGSVVPSADYAYSPAYTQYGTTYGSYGTSSGLI 14357
C SVVP +DYAY+PAYTQYGG YGSY YGT SGLI
Sbjct: 1 CSSVVPSADYAYSPAYTQYGTTYGSYGTSSGLI 35
```

In this case, we can see that the alignment starts out well (amino acid 1 of the exon is aligning to some sequence in the genome) but ends before the end of the exon, we are missing the last 10 amino acids (remember that BLAST only gives a local alignment; that is, it does not report sequences that do not have significant similarity). In cases like this, we would concentrate our search for donor sites around base 14387 (30 bases or 10 codons down from the end of the alignment). While any donor downstream of 14357 (the end of the above alignment) would be potential candidate, donor sites found near 14387 would be strong candidates for use in the final gene model, especially if they have a high score on the donor site detection algorithm (see below).

Computational evidence

While conservation is in most cases the best evidence for constructing your gene model, you will not always have sufficient similarity to construct a viable gene model. There are also cases in which conservation will give support for several different gene models with no way to pick among the consistent models. In these cases, computational evidence is your next best source. The best approach is to rely on conservation as much as possible and adjust your models based on the computational evidence. There are two main sources for information you will want to consider as you try and determine the best gene model, splice site prediction programs and *ab initio* gene finders.
Splice site prediction program. Several of the information tracks available to you on the UCSC Genome Browser show the results of the splice site prediction program GeneSplicer. The output of this program tags potential splice donor and acceptor sites and gives them a score of between -10 and +10. In order to simplify the output, we have classified those sites with scores above 7 as high quality, scores between 0 and 7 as medium quality and scores between -10 and 0 as low quality. In general, this information can be used to help you pick donor/acceptor sites when there is no conservation. For the purposes of the GEP project, you should always pick a donor/acceptor site that maintains the open reading frame and maximizes conserved amino acids; however when there is little or no conservation or there are two or more possible donor/acceptor sites very close together, sites tagged by GeneSplicer are better candidates than sites that are not tagged, and in general, the higher the score the better.

ab initio gene prediction algorithms. The creation and optimization of ab initio gene finders is an active field of study and, as such, many different programs are available to create gene prediction sets. Many of these have been run on the section of DNA that you will be working on. The results of these analyses are available on the Genome Browser for your section of DNA. While each program has its strengths and weaknesses, for the purposes of gene model creation (selection of intron/exon boundaries) they should be considered of equal quality. The most common usage of the information created here is a majority rules/vote system. Failing any evidence from basic biology, conservation or other algorithms, the splice site that was picked by the most different programs would be picked as the donor/acceptor site.

Last and certainly least

It is certainly possible that you may run across situations where you will have ambiguous evidence and must choose between a small number of consistent choices with no evidence to help you decide (this is often the case when using the conservation of exon length rule). In these cases, when all else fails, the policy is to go with the choice that creates the largest protein. The reason for this is that it is better to add a few extra amino acids to a protein than to have a few amino acids missing. This is because if the amino acids are missing there is no way to find them in a BLAST search, but BLAST is fairly tolerant of having a few extra amino acids tucked inside an alignment. Thus it is best to err by including extra amino acids rather than missing amino acids.

It is also possible that you will run across situations where there may be only very weak evidence for one gene model over another, yet the weaker model gives a longer protein. To balance these decisions, the GEP has set a policy for the use of the computational donor/acceptor sites when picking your gene model. In general, when picking among a group of consistent intron/exon boundaries, choose the longest exon that has a boundary no more than one step (low, medium, high) worse than a boundary that creates a shorter exon. In other words, when two choices differ by two steps, go with the higher valued boundary (longer unlabelled vs. shorter medium scoring, pick the medium candidate; longer low scoring vs. shorter high scoring, pick the high-scoring candidate). Alternatively, when two choices differ only by one step (unlabelled vs. low, low vs. medium, and medium vs. high) pick the boundary that gives the longer protein.
Summary

The following is a list of the important rules for annotation based on the above discussion. All models should follow these rules as much as possible. The rules are listed in the order of importance: the best model will follow a rule higher on the list at the expense of a lower rule. Most models will not follow all rules; it is your job as an annotator to create the best model in spite of this. For example, you may need to decide between a model that follows many less important rules but breaks a single more important rule and a model that follows a more important rule over the less important rules. This balancing act is where human ability far exceeds computers. Rules are ranked into four classes:

1. **Inviolate rules** – rules for which no counter examples have ever been seen. Clear and convincing wet bench experimental evidence would be required to convince scientists that this rule should not apply to your model. Since no wet bench work is being done for these projects, they should never be broken.

2. **Important rules** – rules for which exceptions are only rarely seen. You may choose to make a model that does not follow this rule but you must note in your annotation report that this rule was not followed and document why you have decided not to follow this rule.

3. **Basic rules** – these are rules or observations that are seen more often than not but are also not followed in a significant number of models. You should make models that follow these rules if you can but be careful not to ignore more important rules just to follow these rules. You do not need to document that you did not follow rules of this type.

4. **Tie-breaking rules** – rules to help make models when all the more important rules do not help. You may wish to note the use of these rules in your annotation report to help those reviewing your annotations understand why you picked the model you did.

Refer frequently to the one page set of rules on the next page while you annotate and create your gene model.
Rules for Gene Annotation

Inviolate rules (In Basic Biology):
1. CDS of gene must begin with ATG and end with a stop codon; no internal, in-frame stop codons.
2. Exons are found in order along the source DNA.
3. The last two bases of an intron sequence must be AG.
4. Intron sequences should be at least 42 nt.

Important rules:
In Basic Biology:
5. An intron sequence should begin with GT (GC is the rare exception).
   a. GC should be used when use of GT sites breaks important rules
6. Use data in RNA-Seq tracks (both mapped reads and TopHat splice-site junctions) if available.

In Conservation:
7. Conserved amino acids identified by single exon BLAST shown in high quality alignments (i.e. high % identity and properly placed) should be included in exons.
8. The number of exons between informant and new species should be conserved.
9. The organization of exons to generate the various D. melanogaster isoforms should be conserved.
   a. Some genes have alternate splice sites for a particular exon that is unique to that isoform. You must find these alternative splice sites and create a gene model for every isoform.

Basic rules:
In Conservation:
10. Identification of conservation should be done in the following order, (based on speed, sensitivity, ease of use, and specificity).
   a. Protein-DNA BLASTx or tBLASTn with increasing expect thresholds
   b. DNA-DNA CLUSTALW alignment
   c. DNA-DNA BLASTn using very large E-score cutoff values (e.g. 10^{10})
11. Failing identification of exons by expression or conservation alignments as above, the highly-conserved regions identified in the “Comparative Genomics” (multiz) tracks should be checked.
12. Attempt to conserve exon length even if the specific amino acids are not conserved.
13. If it is difficult to identify exons based on the above, repeat the whole process using the gene models generated by the GEP in a close neighbor species as your assigned species instead of the model found in Drosophila melanogaster.

In Computation:
14. Exons that cannot be found by any type of conservation may be identified using predictions from ab initio gene finders

Tie-Breaking rules (In Basic Biology):
15. Longer exons are better; include more amino acids in the exons between the start and stop codons.
Glossary of Terms

*ab initio gene finding*: process whereby genomic DNA sequence alone is systematically searched for certain tell-tale 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. Example: transcription factors that recognize particular patterns in the promoters of the genes they regulate

**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 hybrid plasmid 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) 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. One must sort through six possible reading frames for any putative protein-encoding genomic DNA fragment: three on one strand of DNA heading in the 3’ direction (each shifted by one base with respect to the previous reading frame) and three on the other strand of DNA heading in the opposite direction (Figure 2.1).

Figure 2.1. There are six possible reading frames for any region of DNA. The colored block arrows mark the codons in each of the six reading frames. The different colors of the block arrows represent different amino acids encoded by the various codons. Note that the amino acid sequence for each of the reading frames is different. Note also the stop codon in reading frame +1.

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

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
3. A Sample Annotation Problem

Introduction

This worksheet will guide you through a series of steps that have been found to work well for annotation of species closely related to *Drosophila melanogaster*. It provides a technique that can also be the foundation of annotation in other more divergent species, but in those cases other special techniques will probably be needed. The gene you will study is from the fourth (dot) chromosome of *Drosophila biarmipes*, a species in the *melanogaster* group and *suzuki* subgroup of fruit flies. A diagram that summarizes the specific steps you will follow to annotate the *D. biarmipes* gene is below (Figure 3.1).

![Diagram of gene annotation steps](image)

**Figure 3.1.** Summary of steps in gene annotation using on-line resources

The GEP annotation strategy makes extensive use of a Washington University mirror of the University of California at Santa Cruz (UCSC) Genome Browser (including custom ‘evidence tracks’ for GEP projects), FlyBase and the National Center for Biotechnology (NCBI) BLAST Web server, all publicly-accessible Internet sites. Links to these Web sites are below.
Resources:

Important Note:
Mozilla Firefox is the Internet browser of choice in the teaching labs at Amherst College. However, other Internet browsers work equally well with all of the above databases and on-line search engines. It works best to open each of the above Web sites in a separate window (NOT tabs), so that you can have them all visible at one time on your monitor. You will also need to open the Annotation Report Word file, into which you will paste your results.

While this worksheet will do some click-by-click guidance, the NCBI BLAST and UCSC Genome Browser Mirror web pages can be daunting. For those users completely unfamiliar with these sites, there are online tutorials and user guides for the Genome Browser Mirror at http://genome.ucsc.edu/training.html. BLAST training is available online at the NCBI web page and documentation for GEP programs is available on the GEP Web site (under Help -> Documentations -> Web Framework, then click on or scroll to the ‘Guide to Other GEP Tools’ section and choose the appropriate guide).

Note: As you work through this exercise on a computer, the screenshots you get may not exactly match the images shown in the figures since these Internet sites are frequently updated. Therefore, you should work from the live Web windows as much as possible.

Identifying the Ortholog

The first step in annotating a potential feature is to identify the D. melanogaster ortholog. We can start by examining the UCSC Genome Browser on the GEP website. Go to http://gep.wustl.edu, select the Projects drop down menu, then ‘Annotation Resources’ and the ‘GEP UCSC Browser Mirror.’ Click on ‘Genome Browser’ in the left-hand column of the next window (Figure 3.2).
**Figure 3.2.** GEP UCSC Genome Browser mirror. Note that the fosmid name and/or number in the above window (in the ‘position or search term’ box) changes depending upon which fosmid is selected for study.

As in the above figure, select “D. biarmipes” genome, “Aug. 2012 (GEP / Dot)” assembly, enter contig14 in the position box and click “submit.” Do not be scared by the next window that appears 😅. To make the display a little less complicated, scroll to the bottom half of the window to close up some of the data tracks (Figure 3.3a). Adjust the settings at each of the red arrows to ‘hide’ and make sure that all the other settings are as pictured below. Then, click any of the ‘refresh’ buttons.

**Figure 3.3a.** The GEP UCSC Genome Browser view for contig14 from *Drosophila biarmipes*, lower control section of window.

The top half of the UCSC Genome Browser Mirror should now look as in Figure 3.3b (on the next page). Each colored, horizontal line (with boxes) in the white window represents a different “track” generated by a different sequence analysis software program. The names (above and to the left) of several different gene predictor tracks are repeated in the second set of display control bars in the bottom half of the Browser window (Genes and Gene Prediction Tracks in Figure 3.3a). When ‘dense’ is selected below the name of each track, the track in the Genome Browser mirror (Figure 3.3b) simply indicates that the software analysis has been run. However, to see the specific data that were generated by that software program, you need to expand the
track (by selecting ‘pack’ or ‘full’) to change how or what data are displayed. To find out more about the data in any given track, scroll down to the settings part of the window and click on the title of the track.

Figure 3.3b. The GEP UCSC Genome Browser view for contig14 in *Drosophila biarmipes*, upper, data display part of window.

*The “BLASTX alignment to D. melanogaster Proteins” tracks*

NCBI’s Basic Local Alignment Search Tool (BLAST) is a program that detects sequence similarity between a Query sequence and Subject sequences within a database. The ability to detect sequence homology allows us to identify putative genes in a novel sequence (e.g., the *D. biarmipes* fosmid). It also allows us to determine if a putative gene in the fosmid or encoded protein is related to other known genes or proteins.

BLAST is popular because it can quickly identify regions of local similarity between two sequences. The greater the similarity between two sequences, the higher the raw score (S). The S value for an alignment is calculated by summing the scores for each aligned position and subtracting penalties for gaps (Figure 3.4, a DNA:DNA alignment is shown in this example). More importantly, BLAST uses a robust statistical framework to determine if the alignment between two sequences is statistically significant. BLAST calculates Expect (E) values, which are probabilities for the number of different alignments with scores equivalent to or better than S that are expected to occur in a database search by chance alone. The lower the E value, the more significant the Score. Note that an E value of zero indicates that there is no chance that one would get such a high score by alignment of any other two sequences.
Figure 3.4. Calculation of S values (Scores) for BLAST alignments, from NCBI BLAST Glossary http://www.ncbi.nlm.nih.gov/Education/BLASTinfo/glossary2.html

The “BLASTX alignment to \textit{D. melanogaster} Proteins” tracks in Figure 3.3b show the location of all the BLAST alignments that resulted from using the translated genomic DNA sequence (contig14 from the \textit{D. biarmipes} fourth chromosome) to search the database of all known \textit{D. melanogaster} proteins. It demarcates regions in the \textit{D. biarmipes} fosmid that, when translated, have sequence similarity to proteins in \textit{D. melanogaster}. For these tracks, there are a few things to note.

1. As its name suggests, BLAST is designed to identify local regions of sequence similarity. This means that BLAST may report multiple discrete regions of sequence similarity between a query sequence and a subject sequence in a database. The red or brown colored boxes in Fig. 3.3b represent regions of sequence similarity between the translated \textit{D. biarmipes} sequence and some known \textit{D. melanogaster} proteins. The longer the block, the longer the region of sequence similarity.

2. If a cDNA sequence (generated from a spliced, mature mRNA sequence with more than one exon) is aligned to a genomic sequence, we would expect to see multiple alignment blocks that correspond to discrete exons of putative protein-coding genes; regions of the genomic sequence without significant alignment that fall between these exons/alignment blocks are introns (see Figure 1.3 above and Figure 14.9 in the \textit{Life} text, 9\textsuperscript{th} edition). In the BLASTX tracks (Fig. 3.3b), the thin lines between the alignment blocks thus represent likely intron sequences in the \textit{D. biarmipes} genes, which are not present in mature mRNA molecules.

3. These alignment tracks often show directionality (e.g., which DNA strand contains the coding information for the protein), if the arrowheads point to either the right or the left. Be aware that the arrows showing directionality can be unreliable, especially in cases where the matches are DNA to DNA. If directionality is important, it should be confirmed by other methods.

4. Note that the alignment blocks only show regions of high local similarity. Thus, these tracks will only indicate an entire exon if conservation extends across the whole exon. The extent of the alignment will mostly depend on how closely related the species is to \textit{D. melanogaster}. Therefore, it is often not possible to make firm conclusions about gene structure (number and placement of exons) based on these tracks, since exons with little
or no conservation may be missed while large exons with multiple conserved domains may be broken into multiple smaller alignment blocks.

5. If multiple mRNA isoforms exist for a single BLASTX feature, this is represented by a unique letter at the end of the isoform names (-PA, -PB, etc.).

From the BLASTX alignment tracks in Figure 3.3b, it appears that this region of the *D. biarmipes* genome has two features that show sequence similarity to putative *D. melanogaster* protein-coding genes: *CG31999* and a gene from the *yellow* family, respectively. For this lab, we will focus on the gene from the *yellow* family. Remember, however, that the coding regions for only one of these genes can be present in this region of the genome, since multiple genes cannot overlap at the exact same positions on a DNA molecule. So, *how do you decide which member of the yellow gene family is the best one to study?* To make this decision, answer Questions 1-4 below.

**Question 1**

The different colors of the alignments blocks for the different *yellow* family genes provide one hint. Click on the ‘D mel Proteins’ link in the Genes and Gene Prediction Track area of the UCSC Genome Browser to find a chart of what the different colors represent.

*Which gene has the highest (bit) scores for its alignment blocks?*

[Note that the lengths of the alignment blocks in the Genome Browser window are also longer for this gene than any of the other *yellow* family genes.]

**Question 2**

Type the name of the gene (case sensitive and without the -PA) from Question 1 into the [http://flybase.org/](http://flybase.org/) ‘Jump to Gene’ box in the right corner of the menu bar. Then, click ‘Go.’

*Which chromosome is this gene on in *D. melanogaster*?*

*Why does the fact that the *D. melanogaster* gene is on this particular chromosome (and not on a different one) strengthen the case for this gene being the best candidate for the *D. biarmipes* gene?* [Hint: none of the other yellow family genes are on this chromosome in *D. melanogaster.*]

Click the box next to Gene Model & Products in the lower left of the FlyBase window to view a zoomed-in diagram of the exons and coding sequences (CDS, in purple) for the *yellow-h* gene in *D. melanogaster*. How many CDS are there?

*How does this compare to the number of alignment blocks in the BLASTX track for the *yellow-h* gene in the Genome Browser window of the *D. biarmipes* fosmid (Fig. 3.3b)?*
The Gene Predictor tracks

Although you may now be fairly confident that yellow-h is the best candidate gene for this region of the D. biarmipes fosmid, you should strengthen this conclusion in a couple different ways. First, return to the ‘D. mel Proteins Track Settings’ page and set the ‘Filter score range: min:’ to 500 and click ‘submit.’ This removes all the other yellow family genes from the Browser window. Now, examine the five other Gene Prediction tracks in the Genome Browser window in more detail: select ‘pack’ under the names of each of the gene predictors (Genscan, Geneid, N-Scan, etc.) and click one of the ‘refresh’ buttons. Note that the ‘hits’ for each gene predictor are now numbered in the Browser window; if there is more than one separate feature identified by a particular gene predictor, they are numbered in the Genome Browser window in the physical order, from left to right, that they appear along the DNA sequence of the fosmid. Note the differences between the results of the various gene predictors!

Question 3
Which gene predictor do you think matches the best to the BLASTX output track for the yellow-h gene and why?

How does the alignment for this gene predictor differ from the BLASTX output for the entire fosmid?

You will now use sequence information from one of the gene predictor tracks to confirm that the yellow-h gene is the best candidate gene in this region of the D. biarmipes genome, by doing a BLAST search. It is quicker to BLAST using FlyBase instead of NCBI, since you are only going to search the known proteins database from Drosophila melanogaster, as follows:

1. Obtain the predicted amino acid sequence of gid_contig14_1 in the Geneid Gene Predictions track by clicking directly on any of the predicted exons for this contig (bright red in color).
2. Click on the ‘Predicted Protein’ sequence link in the next window and copy the amino acid sequence.
4. Paste the copied Predicted Protein sequence from the Genome Browser into the Sequence box and click the BLAST button.

Once the BLAST search is complete, a new web page appears with the BLAST Report. The BLAST output begins with a description of the version of BLAST used (blastp 2.2.18 in this
case), a literature reference and some details on the query sequence used in the search. The rest of the default BLAST report consists of three main sections: a graphic summary, a list of significant BLAST hits, and the corresponding amino acid alignments. We will go through each of these sections in order to help you interpret the blastp output.

a. Graphic Summary

The Graphic Summary (Figure 3.5) shows alignments (as colored bars) of the top 25 database matches with significant sequence homology to our Query sequence (line under the colored Score Key). Once again, the color of the bars corresponds to the score (S) of the alignment, with red representing the highest alignment scores. The default listing for the hits is by score, since it is generally the case that the higher the alignment score, the more significant the hit. If you click on a colored alignment bar, you will jump to the actual DNA alignment associated with that BLAST hit (see c. below). Note that since the Geneid track has a single (fused) alignment track including both features in the D. biarmipes fosmid, the CG31999 gene is included at the top of the list of hits. You can ignore the CG-31999 hit for this analysis.

<table>
<thead>
<tr>
<th>Query</th>
<th>Description</th>
<th>Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>gid_contig4_1</td>
<td></td>
<td>1205</td>
</tr>
</tbody>
</table>

Figure 3.5. A graphical overview of the first 14 blastp D. melanogaster hits to the D. biarmipes Query sequence
b. List of Significant BLAST Hits

Scrolling further down the BLAST Report window, you will find a BLAST Hit Summary table with the top 25 sequences in the *D. melanogaster* protein database, sorted according to the Score and E-value (Figure 3.6).

| BLAST Hit Summary |
|-------------------|----------------|
| Description       | Species | Score | E value |
| CG319599-PA        | Dmel   | 927.161 | 0     |
| yellow-h-PA        | Dmel   | 758.444 | 0     |
| yellow-a-PA        | Dmel   | 299.286 | 1.52705e-80 |
| yellow-b-PA        | Dmel   | 267.73 | 4.59782e-77 |
| y-PA               | Dmel   | 286.199 | 1.06792e-76 |
| yellow-f-PA        | Dmel   | 238.039 | 3.81151e-62 |
| yellow-f-PB        | Dmel   | 228.794 | 2.09205e-59 |
| yellow-f-PC        | Dmel   | 228.024 | 3.94429e-59 |
| yellow-f-PA        | Dmel   | 227.539 | 5.73885e-59 |
| yellow-PA          | Dmel   | 202.216 | 1.97794e-51 |
| yellow-f-PA        | Dmel   | 196.223 | 5.37973e-50 |
| yellow-c-PA        | Dmel   | 186.037 | 1.47910e-46 |
| yellow-d-PA        | Dmel   | 162.57  | 1.69091e-45 |
| yellow-e-PA        | Dmel   | 155.221 | 3.22210e-37 |

**Figure 3.6.** Scores and E-values for the first 14 blastp *D. melanogaster* hits to the *D. biarmipes* Query sequence.

**Question 4**

*How does the Score and E value of the *D. melanogaster* yellow-h gene alignment to the *D. biarmipes* fosmid compare to the values for the other members of the yellow gene family?*

c. List of Alignments

Following the table of BLAST hits is a section showing all of the alignment blocks for each BLAST hit. To jump right to the alignment for the protein encoded by the yellow-h gene, click on its Score in the BLAST Hit Summary table. The sequence alignment (Figure 3.7) shows you how well the Query sequence matches with the Subject sequence in the database. Since you will rely heavily on sequence alignments in your annotation efforts, examine the alignment to the *Drosophila melanogaster* yellow-h amino acid sequence more closely.

Different BLAST hits (Subject sequences in the database) are separated by definition lines that begin with a ‘>’ character, which is followed by lots of database information for that particular Subject sequence, including its FlyBase ID number, the name (or symbol, if unnamed) of the encoded protein and the length (in amino acids) of the Subject sequence. Each alignment block demarcates a local region of similarity between the Query sequence and the Subject sequence identified by the definition lines.
Figure 3.7. Alignment of the Predicted Protein sequence from *D. biarmipes* (Query) to the yellow-*h* protein product from *D. melanogaster* (Subject).

**What about the alignments themselves?** Each alignment block begins with a summary that includes the BLAST Score and Expect value (the statistical significance of the alignment), sequence **Identities** (number of identical amino acids between the Query and the Subject sequence), sequence **Positives** (amino acids that are either [1] identical or [2] mismatched but have side chains with similar physical or chemical properties, such as leucine and isoleucine) and the number of **Gaps** in the alignment. Gaps are places where the BLAST program inserts spaces to maximize the score in other parts of the alignment; gaps generally represent places where one gene sequence has diverged away from the other over evolutionary time by adding or deleting codons.
The alignment itself consists of three lines: the top Query sequence, a middle Matching sequence, and the bottom Subject sequence (Figure 3.7). The Matching sequence consists of a combination of letters (for identical amino acids in the two sequences), + characters, and empty spaces. The ‘+’ character shows the location of the mismatched Positives (see above). Places in the middle Matching line that are blank are locations where both species have amino acids in those locations, but the side chains of the mismatched amino acids are chemically or physically dissimilar. Note that these dissimilar mismatches still contribute a positive value to the overall alignment score. Finally, the ‘-’ characters in either the Query or the Subject sequence denote gaps in the alignment.

Now, click on the alignment bars or scores for some of the other yellow family genes.

Note how much better and longer the alignment is with the yellow-h gene than with the other yellow family members. Thus, we will proceed with the assumption that this region of the fosmid does indeed contain the D. biarmipes ortholog for the yellow-h gene of Drosophila melanogaster and not any of the other yellow family genes identified by the BLAST searches. ☺

**Coding DNA Sequence (CDS) by CDS Searches**

While it is true that the Geneid prediction shows significant matches to the Drosophila melanogaster yellow-h protein (Figures 3.5-3.7), this does not mean that the prediction is totally “correct”. In fact, published accuracy rates for most *ab initio* gene prediction algorithms are in the range of 20-30%. Common errors generated by Geneid and other *ab initio* gene prediction algorithms are skipped exons and errors involving the ends of the gene (split genes, fused two genes, as in this case, etc.). It is therefore very possible that the Geneid prediction is actually wrong (i.e. not perfect). Without detailed analysis of the alignment between contig14_1 and the yellow-h gene, we have no way of knowing. For now, all we know is that BLAST has aligned at least some of contig14_1 with at least some of the yellow-h gene from Drosophila melanogaster, which has contributed to a good E value for the sum of all the alignment blocks to the fosmid.

The next step is to use BLAST searches to find the best matches to the individual D. melanogaster yellow-h coding sequences (CDS), as these matches will be the best evidence we can gather as to the structure of this gene in D. biarmipes. In order to do CDS-by-CDS searches, we need the sequence of each CDS. This information is most easily obtained from the GEP’s Gene Record Finder. Go to the ‘Projects’ drop down menu on the [http://gep.wustl.edu](http://gep.wustl.edu) website, then select ‘Annotation Resources’ to get to the Gene Record Finder. Enter ‘yellow-h’ into the search box (case sensitive) to obtain the information on this gene. As you type, note that the name of the gene appears in a drop-down bar below the search box, with the number of mRNA isoforms (only 1) and the total number of exons (3) and CDS (3) for this gene. Now, click the “Find Record” button. [Note: a pdf of The Gene Record Finder User Guide is posted on the course Moodle site in the Lab 10 topic box.]

There are three regions of information in the next window that appears: Gene Details, mRNA Details and Transcript Details/Polypeptide Details. Note that the name of the isoform has changed from yellow-h-PA (in the previous database windows) to yellow-h-RA. The ‘R’ in the Gene Record Finder has simply been substituted for the ‘P’ in these other databases, and you can ignore this difference. The term CDS refers to DNA sequences that are transcribed, present in a mature mRNA molecule and code for amino acids. In most cases, the term CDS can be used
synonymously with ‘exon.’ The exceptions are those exons that include 5’- or 3’-untranslated regions, which do not code for amino acids. Also, due to alternative splicing, not all exons of a gene may appear in a particular mature mRNA. Hence, the windows for Transcript Details and Polypeptide Details reflect these situations and are not the same, with the numbered CDS referring to the specific DNA sequences that are translated in a particular mRNA molecule.

If you click on a particular row of the Polypeptide Details table, the amino acid sequence encoded by the specific CDS appears in a new pop-up window (Figure 3.8). Make sure you record the number (under Length) of amino acid residues in the Drosophila melanogaster isoform encoded by each CDS.

![Figure 3.8](image)

**Figure 3.8.** Amino acid sequence of CDS #1_1556_0 (the first CDS) for the yellow-h gene from *Drosophila melanogaster*.

To map each putative CDS onto the fosmid DNA from *D. biarmipes*, you will do searches that use the BLAST algorithm to compare two sequences to each other (bl2seq). In this case, you will do a blastx search to compare the entire translated *D. biarmipes* fosmid to the amino acid sequence of each *D. melanogaster* CDS. Begin by comparing the protein sequence of the first coding exon of yellow-h (CDS #1_1556_0) to the entire fosmid DNA sequence, as follows:

2. On the next page that appears, click the ‘Align two or more sequences’ box, which adds an “Enter Subject Sequence” box to the BLAST window.
3. Return to the UCSC Genome Browser page for *D. biarmipes* and obtain the contig14 fosmid DNA sequence by clicking ‘DNA’ in the upper menu bar (make sure contig14:1-45,000 appears in the position/search box). Click on the ‘get DNA’ button in the next window.
4. Copy and paste the fosmid DNA sequence (45,000 bp long) into the ‘Enter Query Sequence’ box of the Blast window (it is okay to include the file header when you copy and paste the DNA sequence).

5. Copy the amino acid sequence of CDS #1_1556_0 from the pop-up window of the ‘Polypeptide Details’ section of the Gene Record Finder window, and paste the CDS sequence into the ‘Enter Subject Sequence’ box.

6. Check the ‘Show results in a new window’ box next the BLAST button (which will enable you to reuse the blastx search page for each of the next two CDS searches).

7. Click on the ‘Algorithm parameters’ link, and be sure that the ‘Low complexity regions’ and two mask boxes are UN-checked.

8. You are now ready to click the BLAST button! It will take a few seconds for a new window with the blast alignment(s) to appear.

The organization of NCBI’s BLAST Report page is similar to FlyBase’s, with a Graphic Summary, followed by a table of significant hits, and then one or more specific alignments of the first CDS (Sbjct Sequence) to the translated fosmid sequence (Query Sequence). Note that in this case, only the first alignment produces a statistically-significant E value; the five other alignments below this one are from other regions of the fosmid and are so weak that they can be ignored. **Copy the best (first) alignment and paste it into the appropriate section of the Annotation Report** (fillable Word file on Moodle course site in Lab Week 10 topic box). When copying alignments, be sure to include the Score, etc. header information and shrink the margins and/or font to keep the sequences in alignment across the width of the page).

The alignment should look as below (Figure 3.9):

![Figure 3.9](attachment:image.png)

**Figure 3.9.** Alignment of the first CDS from the *D. melanogaster yellow-h* gene (Sbjct) to the translated *D. biarmipes* fosmid (Query).

One important point to understand about blastx alignments is that the numbering of the Query sequence reflects the nucleotide position in the *D. biarmipes* fosmid, while the numbers of the Sbjct sequence indicate amino acid positions in the CDS from *D. melanogaster*. By using the whole fosmid in our search, we can read the DNA base coordinates of the orthologous *D.*
biarmipes CDS directly from the alignment. Thus, the alignment for this first CDS begins at base 41,412 of the fosmid with the codon for methionine (remember your Basic Biology rules!) and ends at base 41,197. Note that the beginning of the coding region of this gene is not especially well conserved between D. melanogaster and D. biarmipes. It is also important to check whether or not the entire 78 amino acids of the D. melanogaster CDS (Sbjct line) aligns to this region of the D. biarmipes fosmid (that’s a yes in this case!).

blastx translates the input fosmid sequence in all six reading frames (Figure 2.1), and compares each translated sequence with the amino acid sequence of the CDS. Thus, there is an additional field ‘Frame’ right above the start of the alignment that indicates which frame was translated to produce the specific alignment (Figure 3.9). The frame can either be + or – and it corresponds to the relative orientation of the fosmid sequence compared to the CDS (e.g., which DNA strand contains the CDS information). For each orientation, the frame has a value that ranges from 1 to 3, which reflects the base at which the reading frame began relative to the 5’ end of the coding DNA strand. Together, the relative orientation and the frame represent all six reading frames.

Question 5
What reading frame was the fosmid translated in to best align with the first CDS?

What can you conclude from the sign of a reading frame (+ or -) about the relative magnitude of the base pair coordinates for the beginning versus the end of a mapped alignment block in the fosmid DNA?

Do the base coordinates of the beginning and end, respectively, of the alignment to the first CDS support your conclusion? Explain.

Question 6.
Repeat the same blastx searches with the second and third CDS (#2_1556_1 and #3_1556_0); copy and paste the best alignments for each CDS into the Annotation Report. Highlight the DNA base coordinates of the beginning and end of each alignment. Also highlight the frame number that was translated to generate the amino acid sequence for the alignment.

Annotating CDS Boundaries

Since the blastx alignments show conserved amino acid residues between D. melanogaster and D. biarmipes, extra nucleotides in the exon after the last or before the first complete codon of that exon may not be apparent in these alignments (if, for instance, a triplet codon has been split by an exon-intron boundary). Thus, to carefully annotate a protein-coding gene, we must find the exact mRNA splice positions that would create a processed mRNA that links these exons together to create a continuous coding block (e.g., an open reading frame).
Use the UCSC Genome Browser to navigate to the beginning of the first CDS and confirm the coordinates of the start codon for the yellow-h gene in the D. biarmipes fosmid. From the alignment in Figure 3.9, we believe the start codon begins at base 41,412. To jump directly to this region, enter the coordinates “contig14:41,400-41,450” in the ‘position/search’ box, then click ‘jump.’ Next, make sure ‘full’ is selected under the ‘Base Position’ link in the ‘Mapping and Sequencing Tracks’ area and ‘pack’ is selected under ‘D. mel Proteins’ in the box below that. Don’t forget to click a ‘refresh’ button.

Since the first alignment block was in Frame -1, it is easier analyze the fosmid by reversing the DNA sequence of the fosmid (to display the three frames going from left to right in this orientation). To do so, click the ‘reverse’ button under the white window of the Browser. The new window is shown in Figure 3.10 (you may need to zoom in a bit to clearly see the nucleotide bases). Note that the numbering of the nucleotides in the fosmid now decreases as you read from left to right.

Three reading frames for the - strand: 1st (top), 2nd (middle) and 3rd (bottom) row. Each gray box represents the amino acid encoded by the above three nucleotides in the DNA sequence.

Figure 3.10. Close-up look at the region around the beginning of the first CDS alignment

The above Genome Browser window confirms that the first CDS of the D. biarmipes yellow-h gene begins at bp 41,412 of the fosmid, as a codon for methionine (green start codon) in the first row of amino acids (translation Frame -1) begins at this position. The BLASTX alignment block to the first CDS of the yellow-h gene also begins at this position and extends towards the right side of the window. Note that NONE of the seven gene predictors show any evidence for an exon in this region of the fosmid!

Use a similar method as above to now jump to the predicted end of the first CDS (at bp 41,197). To precisely map the end of this CDS, you must examine this region for potential intron donor splice sites (see Fig. 1.3), since you are at the beginning of the first intron (Figure 3.11). The blastx alignment in Figure 3.9 indicates that you should look for the amino acid sequence ‘SILN’ in the first row of gray boxes (reading Frame -1) to find the end of the first CDS. Note that six of the seven Gene Predictors detected an exon beginning somewhere in the middle of the
first BLASTX alignment block for the *yellow-h* gene, and they are now very helpful in directing your eye to the possible end of the exon. Another helpful set of data to add to the Browser window at this time are the Predicted Splice Sites (see explanation of this track in the Computational evidence section of the Student Outline). Add these data by selecting ‘dense’ below the corresponding link in the Gene and Gene Prediction Tracks control area. Color-coded bars then appear under putative intron splice site consensus sequences (red arrow in Figure 3.11).

**Figure 3.11.** Close-up look at the region around the end of the first CDS/exon of the *yellow-h* gene.

The complementary bars for the gene predictors and the predicted intron donor splice site readily identify the ‘GT’ at base position 41,194 and 41,193 (circled in blue above) as the first two bases of the first intron. Note that this is NOT the same position as called by the BLASTX alignment (which extends the exon by one base to include the G). This is a great example for why the blastx alignments of the individual CDS you did earlier do not necessarily give you the most accurate coordinates for the end (or beginning) of exons/CDS.

An additional complexity that arises from the above coordinates is that by ending at base position 41,195 (e.g., one base BEFORE the intron donor GT site), the first exon ends in the middle of a codon. The last complete codon in the first exon is for asparagine (N in the first row of gray boxes). We can see by careful inspection that a cleavage after base 41,195 will leave two bases (GG) between the end of the last complete codon (AAT) in frame -1 and the end of the exon. We use the term “phase” to describe these remaining bases; in this case, the end of the first exon is said to be in phase 2 because two bases are present after the last complete codon and before the exon ends. To make a complete mRNA, we must find an acceptor site at the other end of this intron, such that these two bases at the end of the first exon will join with one other base at the beginning of the next exon to make a complete three-base codon (Figure 3.12). For now, we simply note that the best donor site in this region is in phase 2.

*Add the base pair coordinate and phase of the mapped 3’end of the first CDS to the alignment information from Question 6.*
When you are deciding where to splice adjacent exons together, the acceptor site that you select for an exon must have a phase that is compatible with the phase of the donor site for the previous exon, e.g., phase 0 acceptor with phase 0 donor, phase 2 acceptor with phase 1 donor, or phase 1 acceptor with phase 2 donor (as in Figure 3.12). If the acceptor site has a phase that is incompatible with the donor, the extra bases will create a frame shift. The downstream CDS will then be translated in an incorrect reading frame, creating an abnormal polypeptide.

Use the Genome Browser to navigate to the region where you suspect the second exon begins, based on the alignment you found for Question 6. Remember that intron acceptor sites have the invariant sequence ‘AG’ just before the first base in the following exon.

**Question 7.**
*There are two potential AG intron acceptor sites right next to each other near the beginning of exon 2. What are the base pair coordinates of the AG dinucleotide you think is the best one to use for your gene model and what is the evidence in support of the one you chose?*

Record the base coordinate and phase for the beginning of the second exon as deduced from the above analysis.
Question 8
Use the results of the alignments of the second and third exons from Question 6 to precisely map the 3’ end of the second exon and the beginning and end of the third exon. Record the base pair coordinates for the beginning and end of all three exons in the table below. You should also record the base pair coordinates of the stop codon (red block with a * in the middle), which should immediately follow the end of exon 3.

| Table 1. Coordinates of yellow-h CDS in contig14 of Drosophila biarmipes |
|---------------------------------|------------------|
| Base Pair Coordinates           |                  |
| CDS 1                           |                  |
| CDS 2                           |                  |
| CDS 3                           |                  |
| Stop codon                      |                  |

Checking your Gene Model with The Gene Model Checker

Overview
The next step in the process of making a gene model is to check whether or not the above coordinates for the three D. biarmipes yellow-h CDS code for a full-length polypeptide chain when spliced together. You will use the GEP’s Gene Model Checker to do so. Detailed instructions for using this software can be found in The Gene Model Checker User Guide, a pdf of which is posted on the course Moodle site.

Using the Gene Model Checker
To confirm the accuracy of a gene model using the Gene Model Checker, proceed as follows:
2. Download onto the desktop a copy of the fasta sequence file of the entire fosmid (D. biarmipes dot, contig14) from the course Moodle site (in the Lab Week 10 topic box).
3. Upload this sequence file into the first box on the left of the Gene Model Checker window.
4. Type in the name of the D. melanogaster ortholog (yellow-h-PA, case-sensitive) in the second box. Watch for this gene to appear in the drop-down menu box as you type.
5. 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: # - #, # - #, # - #
6. 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 in # - # format in the box lower down in the Gene Model Checker window.
7. You did not annotate the untranslated regions, so click the circle in from of ‘No.’
8. The information for this gene is on the (-) strand, so click the circle in front of ‘Minus’ following ‘Orientation of Gene Relative to Query Sequence.’
9. We think all the gene’s coding sequences are accounted for, so click the circle in front of ‘Complete’ in the next row.
10. Use the drop-down menu to select the Project Group (D. biarmipes Dot) and type in the Project Name (contig14) for this fosmid.
11. **Red boxes will appear around any entries that the Gene Model Checker deems incorrectly entered. Fix these before going on.**
12. Click on the ‘Verify Gene Model’ box at the bottom of the window.
13. A summary of how your model did now appears on right side of the Gene Model Checker window. Did your gene model pass on all counts?
14. 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 sequence. A simple misreading or mis-typing of intron-exon boundary coordinates is responsible for many model failures.
15. If your problem was “Fail with premature stop codons”, click on the Peptide Sequence tab in the upper-right menu bar next to “Checklist.” The symbol * in the peptide sequence that appears next will show you where these premature stop codons are. You can then use this information to find the coordinates of the problematic CDS using your saved blastx alignments.
16. **If you passed the Gene Model Checker on the first try, intentionally change some of the coordinates to see what a failure would look like!**
17. Re-check your work and the typing of the coordinates until you pass ALL parts of the Gene Model Checker. Keep the Gene Model Checker window open, as you will come back to it often while completing the Annotation Report.

*Complete the Annotation Report by following the directions and answering the questions in the report. Upload the completed Annotation Report (one per pair) to the course Moodle site by NO LATER than the start of next week’s lab.*
Acknowledgements

The GEP is supported by Howard Hughes Medical Institute grant #52005780 to Sarah C.R. Elgin at Washington University in St. Louis, Missouri.

Literature Cited


