# Lab Weeks 8-11: Introduction to the GEP Collaborative *Drosophila* Annotation Research Project

**[adopted from the Genomics Education Partnership (GEP) - http://gep.wustl.edu/]**

## Overview of GEP Research Goals

The Genomics Education Partnership is a national, collaborative, scientific investigation of a problem in genomics, involving wet-lab generation of a large data set (e.g., **finishing** a genomic sequence) and 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) 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.

The scientific interest is based on observations that the dot chromosome is largely **heterochromatic** in some species, but largely **euchromatic** in others. Actively-transcribed regions of DNA are typically thought to be euchromatic, with transcriptionally-silent regions of DNA often assuming a heterochromatic state (see pp. 187-188 and pp. 322-323 in the *Life* text). However, expressed genes on the dot chromosome in *Drosophila* appear to reside in euchromatic regions in some species and heterochromatic regions in others. Thus, understanding chromatin effects on dot chromosome gene expression requires careful analysis not just of the genes present, but of the type and distribution of repetitive elements. Since the dot chromosome is a fraction of the size of the other chromosomes (containing just over 80 genes in *Drosophila melanogaster*), it is a simpler system with which to explore chromatin effects on gene expression.

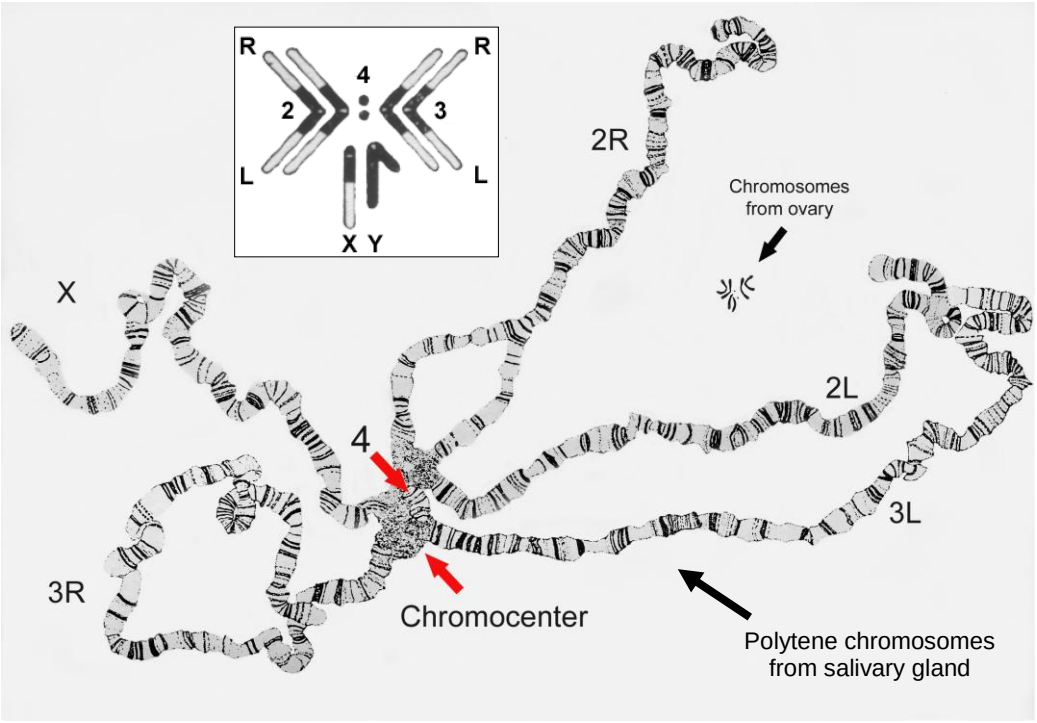



Figure 1. *Drosophila melanogaster* chromosomes

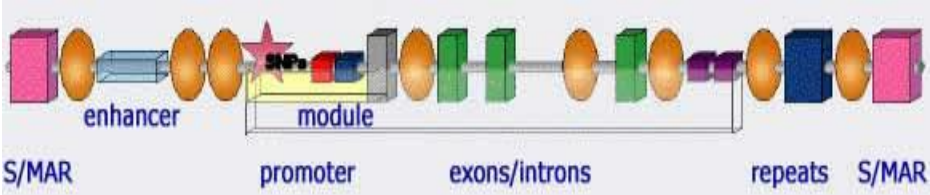
Finishing is usually done in the context of a whole semester course or independent research project. However, members of the Genomics Education Partnership can participate in an annotation project without having to be involved in finishing a sequence, as some publically-posted sequence data are of sufficient quality that no further finishing is needed. Thus, students in the Biology 191 course will have the opportunity to contribute to the completion of an original annotation project!

## What is Gene Annotation?

*(This section modified from the Annotation for Amateurs Website: http://www.plantgdb.org/tutorial/annotatemodule/ and Wikipedia: http://en.wikipedia.org/wiki/Genome_project )*

Many organisms have had their entire genome sequenced; however this is not the end of a genome project. **Annotation** is the process by which pertinent biological information is attached to these raw DNA sequences. This involves identifying **features** in the DNA and determining which of these features can be called **genes**. 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** (which are often much longer than the exons). Thus, gene annotation must include careful mapping of all the exon-intron boundaries, to create a gene model that results in the translation of a full-length polypeptide chain.

The diagram below represents a tiny fragment of DNA, a single hypothetical gene. Notice that there are various parts within the gene. Some of these parts will code for a protein (the exons), others contain regulatory information (promoters and enhancers), and some will not be translated (introns).




One important aspect of annotation is identifying which parts of a genome are transcribed and processed into a mature mRNA molecule, as this is the intermediate stage between a gene and the protein that the gene encodes. Obviously, computer programs are essential to this process, as they can scan newly-sequenced DNA regions for **open reading frames** and **consensus sequences** of promoters, enhancers, intron splice sites, etc. However, human brains are often required to evaluate computer-generated gene models, which often do not agree with each other when generated by different programs.

## Overview of Gene Annotation Labs

In these labs, students will work in pairs to annotate one or two specific features of a 40-60 kb **fosmid** of DNA from a species of Drosophila. The basic GEP annotation strategy makes extensive use of the University of California at Santa Cruz (UCSC) Genome Browser (including custom 'tracks' for our own projects), NCBI Blast and FlyBase, all publically-accessible Internet sites.

You will first complete an on-line activity that introduces you to the National Center for Biotechnology Information (NCBI) BLAST search engine and the GEP's own Gene Record Finder Web site, as you examine a DNA sequence from *Drosophila yakuba*, a close relative of the widely-used model organism, *Drosophila melanogaster*.

Genomic sequences from species such as *Drosophila mojavensis* or *Drosophila grimshawi*, which are more distantly related to *Drosophila melanogaster*, are generally more difficult to annotate. However, you will next work with a highly-conserved gene from *Drosophila grimshawi* by working through "A Sample Annotation Problem", filling in answers to questions in the worksheet 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 exon/intron boundaries.

The next scheduled lab(s) will be devoted to work on your individual annotation project, with help from the TA's and the lab instructor. We will first provide an overview of the preliminary analyses of the claimed *Drosophila* fosmids from the GEP, in which the number of **features** (putative genes) in the fosmid(s), the number of **isoforms** for each feature, and the relative complexity of the different features was pre-determined using various gene predictor software platforms. Specific fosmids from *Drosophila erecta* or *Drosophila mojavensis* were selected for annotation based on these and additional criteria. You and your lab partner will then (by random draw!) select one or two features and/or isoforms from the pre-selected fosmids to annotate for your individual project.

You will also be introduced to the Gene Model Checker in the second activity, so you know how to check your work. Once you have passed (*to your instructor's satisfaction!)* all parts of the Gene Model Checker and the final BLAST alignment, you will then complete an Annotation Report for your mapped feature. Students who annotated different isoforms of the same gene should meet to create one group gene report for that feature.

The Annotation Reports should be completed by the end of the final lab. If time allows, each pair of students (or group if the feature has multiple isoforms) will give a few-minute overview of your feature to the rest of the lab section (with Power Point slides if you wish), talking a little bit about the gene and what it encodes (if it's 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.

## Conceptual Guide to Gene Annotation (*warning: this is heavy going!)*

The following guide is intended to help you think about the various types of evidence you should consider as you attempt to annotate genes in *D. mojavensis.* The same considerations will also apply to species that are closer to *Drosophila melanogaster*, such as *Drosophila erecta*. However, in many of these species the level of conservation will be high enough that some of the other forms of evidence will rarely need to be considered. Your job as a gene annotation researcher is to learn as much as you can within the annotation labs time frame and apply these skills and knowledge to come up with your best gene model.

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 cases where evidence of

expression (**expressed sequence tags** [ESTs] or mRNA sequences) is available, one may be able to identify the coordinates of the full-length transcript including the **5' and 3' untranslated regions**; otherwise one must focus on the protein-coding domains. The coordinates assigned in these cases would describe the base position of the beginning and end of each piece of coding sequence that together make up most of the **exons** 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 cannot include sequences from both strands. It must start on one strand, continue down the length of that strand and end on the same strand. **At the minimum, your gene model must include the base position of the *start codon* for translation, the position of the beginning and ending of each coding exon and the position of the *stop codon* of translation**. For species sufficiently close to *Drosophila melanogaster* (e.g., *Drosophila erecta*), it might also be possible to identify the 5' and 3' untranlsated regions of the mature mRNA by sequence similarity to *D. melanogaster* **cDNA** sequences. It may also be possible, as stated above, to identify these regions if there is mRNA sequence data available for your feature. More information about how to use the mRNA sequence evidence will be provided to you later.

Your first step in annotation will be to collect and consider all the information or 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 used fall into two basic categories, conservation and computation. **Conservation** defines those types of evidence that rely on the assumption that the new species being annotated had a common ancestor with *Drosophila melanogaster*. 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 model for the explanation of anything is the one with the fewest assumptions, the best gene model would be the model that assumes the fewest mutations (i.e. is the one that has the most similarity to the *Drosophila melanogaster* gene model).

The second general type of evidence is **computational**. Many computer programs have been written that attempt to recognize various features in DNA sequence. Several of these programs have already been run on the fly genomes and the results are available for viewing on various genome browsers. These programs are designed to do a variety of functions including gene prediction, recognition of repeats of various types, or identification of other features (e.g. intron/exon boundaries). Each of these programs has been worked on and optimized for its given purpose and as such provides at least a hint as to a possible biological function of any given sequence. Barring sequence conservation, this type of evidence is usually the only evidence one can fall back on to create gene models from newly sequenced genomes.

Finally, if neither conservation nor computational evidence can be found, 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 Advanced Annotation Instruction sheet.

## Basic Biology

Before we consider types of annotation evidence in more 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.

**Introns.** Unlike bacteria, many genes in eukaryotes have introns (Figure 2). These sequences are removed from the primary RNA transcript based on sequences found within the intron. The sequence at the beginning (5' end) of the intron is called the **donor site**, while the 3' end 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.

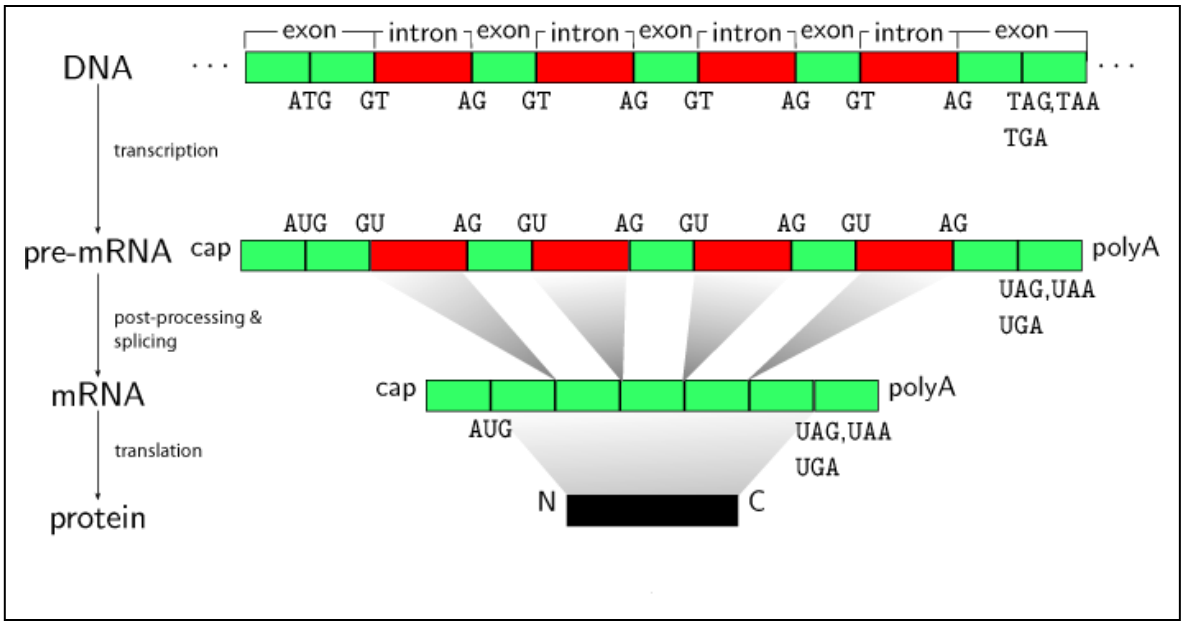



Figure 2. Eukaryotic gene structure. The donor (GT) splice site and the acceptor (AG) splice site are the first two and last two nucleotides, respectively, of the DNA sequence of each intron (in red above). (Figure from http://svmcompbio.tuebingen.mpg.de/splicing.html)

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 sequences. The results of this analysis are displayed in the GEP 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 can be considered as a potential intron donor site, while any AG is a possible acceptor site.

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 detected in *Drosophila melanogaster*. Non-canonical donor sites will never be used in gene models generated by the *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* put the minimal size of an intron at 43 bases (see Guo M et al. 1993. Mol Cell Biol 13:1104 and Talerico M and Berget SM. 1994. Mol Cell Biol 14:3434). 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 very likely incorrect and a different gene model that does not have such a small intron should be made.*

**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 predict the final coding sequence of the mRNA as well as the predicted amino acid sequence of the encoded protein. Remember that in eukaryotes, each 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.*

## 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 the most important form of evidence that will 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 can be easy since the boundary will be very close to the end of the alignment. As the extent of amino acid conservation goes down, the identification of intron/exon boundaries will need to rely on other evidence as 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 mojavensis* or *Drosophila erecta* will almost always have the same basic structure

(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 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 surprising number of cases, we have found 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    CGSVVPSADYAYSPAYTQYGGTYGSYSYGTSSGLIYNPAS
 41    GPITT
```

BLAST alignment:

```
Query: 14253 CGSVVPSADYAYSPAYTQYGGTYGSYSYGTSSGLI 14357
              C SVVP +DYAY+PAYTQYGG YGSY YGT SGLI
Sbjct:     1 CSSVVPGSDYAYNPAYTQYGGAYGSYGYGTGSGLI 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 genome browser show the results of the splice site prediction program GeneSplicer. The output of this program tags potential spice 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 that have been tagged by GeneSplicer are better picks 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 on the side of extra and not on the side 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. Said another way, when two choices differ by two steps, go with the higher valued boundary (longer unlabelled vs. shorter medium scoring, pick the medium; longer low scoring vs. shorter high scoring, pick the high scoring). 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 choose 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-summary 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 using CLUSTALW
    c. DNA-DNA BLASTn using very large E-score cutoff values (e.g. $10^{10}$) *(see Advanced Annotation Instruction sheet for b. and c.)*
11. Failing identification of conservation by BLAST or CLUSTALW, check for EST evidence (BLASTn to EST database at NCBI –restrict to proper species).
12. Failing identification of exons by EST evidence, the highly conserved regions identified in the "Comparative Genomics" (multiz) tracks should be checked.
13. Attempt to conserve exon length even if the specific amino acids are not conserved.
14. If it is difficult to identify exons based on sequence conservation or computational results, repeat the whole process using the gene models already called by GEP in *D. virilis* as the reference model instead of *Drosophila melanogaster*.

**In Computation**

15. Exons that cannot be found by any type of conservation may be identified using predictions from *ab intio* gene finders

***Tie-Breaking rules:***

**In Basic Biology**

16. 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 extrinsic 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 regions/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 (*see diagram on next page*).
***Orthologs***: related genes in <u>different</u> species, which are derived from the same gene in a common ancestral species
***Paralogs***: related genes <u>within</u> a species, which have arisen by a duplication event

**Introns:** gene sequences that are transcribed into RNA but are removed during splicing (see http://en.wikipedia.org/wiki/Intron for more on introns)

**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):** the part of an mRNA molecule that potentially codes for a polypeptide. ORFs are located between the start codon (AUG) and a stop codon (see table)

for translation. One must sort through six possible reading frames for any putative protein-encoding genomic DNA fragment: 3 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 stand of DNA heading in the opposite direction.

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

**3' Untranslated Region (UTR):** mRNA sequence between the Stop codon for translation and the poly A tail

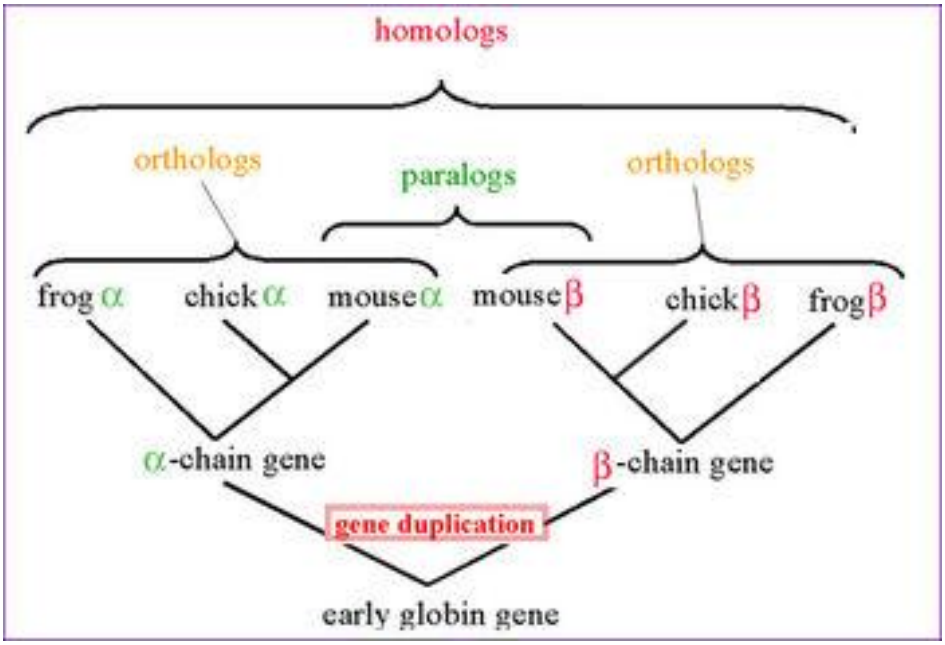



**Image** from http://schaechter.asmblog.org/schaechter/2009/03/of-terms-in-biology-ortholog.html

**Codon Chart** (codons 5'-> 3')

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| GCA<br>GCC<br>GCG<br>GCU | AGA<br>AGG<br>CGA<br>CGC<br>CGG<br>CGU | GAC<br>GAU | AAC<br>AAU | UGC<br>UGU | GAA<br>GAG | CAA<br>CAG | GGA<br>GGC<br>GGG<br>GGU | CAC<br>CAU | AUA<br>AUC<br>AUU | |
| Ala | Arg | Asp | Asn | Cys | Glu | Gln | Gly | His | Ile | |
| A | R | D | N | C | E | Q | G | H | I | |
| UUA<br>UUG<br>CUA<br>CUC<br>CUG<br>CUU | AAA<br>AAG | AUG | UUC<br>UUU | CCA<br>CCC<br>CCG<br>CCU | AGC<br>AGU<br>UCA<br>UCC<br>UCG<br>UCU | ACA<br>ACC<br>ACG<br>ACU | UGG | UAC<br>UAU | GUA<br>GUC<br>GUG<br>GUU | UAA<br>UAG<br>UGA |
| Leu | Lys | Met | Phe | Pro | Ser | Thr | Trp | Tyr | Val | stop |
| L | K | M | F | P | S | T | W | Y | V | |