Annotation Instruction Sheet

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 *Drosophila virilis*, *D. mojavensis*, or *D. mojavensis*. The same considerations will also apply to species that are closer to *Drosophila melanogaster*, 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.

The basic idea when attempting to create a gene model is to determine a series of coordinates that describe the structure of the gene. In species where evidence of expression (EST or cDNA sequences) is available one may be able to identify the coordinates of the full-length transcript including 5’ and 3’ untranslated regions; otherwise we will 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 makes up the exons in the final 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, your 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. Your gene model will include the base position of the start codon, the position of the beginning and ending of each coding exon and the position of the stop codon. For species sufficiently close to *Drosophila melanogaster* (e.g. *Drosophila erecta*) it might also be possible to identify the non-coding parts of the mature mRNA by sequence similarity to *D. melanogaster* cDNA sequences. It may also be possible to identify these regions if there is a sufficiently large set of sequenced cDNA clones. However, neither of these conditions apply to the annotation of *Drosophila virilis*, *Drosophila mojavensis*, or *D. mojavensis* and as such the goal of annotation in these species will simply be to identify the coding regions of the gene without the need to identify the non-coding 5’ and 3’ untranslated ends.

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 for 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 which 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 on 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. Baring sequence conservation, this type of evidence is usually the only evidence you can fall back on to create your gene models.

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 at the end.
**Basic Biology**

Before we consider types of evidence in more detail we will discuss a few details of basic biology, these will 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. These sequences are removed from the primary RNA transcript based on sequences found within the intron. The sequence at the beginning or 5’ end of the intron is called the donor site, while the 3’ end is called the acceptor site. The consensus sequence that defines donor and acceptor sites has 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 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 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 which 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 it is possible to predict the final coding sequence of the mRNA as well as the predicted amino acid sequence of the encoded protein. Remember than 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 you 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 these should be considered when generating your 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 all genes found in the various drosophila species, the encoded proteins are serving very similar if not identical functions (i.e. they 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 as 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.

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 virilis 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:

<table>
<thead>
<tr>
<th></th>
<th>CGSVVPSADYAYSPAYTQYGGTYGYSYGTSGLIYNPAS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>GPITT</td>
</tr>
</tbody>
</table>

BLAST alignment:

Query: 14253 CGSVVPSADYAYSPAYTQYGGTYGYSYGTSGLI 14357

C SSVPP +DYAY+PAYTQYGG YGSYS YGT SGLI

Sbjct: 1 CSSVVGPSDYAYNPAYTQYGGAYGYSYGCTGGLI 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 \textit{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 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 which have been tagged by GeneSplicer are better picks than sites which are not tagged, and in general, the higher the score the better.

**ab initio gene prediction algorithms.** The creation and optimization of \textit{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.

**Special situations**

There are a few situations that deserve special comments. You may not run across these situations but the comments below can be helpful if you do.

**Conservation on one end of an intron.** Sometimes when searching for similarity between \textit{D. melanogaster} exons and your genomic region you will find one exon with a very good match to the intron/exon junction at one end of the intron and no match at the end of adjacent exon at the other end of the intron. It is still sometimes possible to find the unaligned site by using a string search instead of a BLAST similarity search. This is probably best explained by example. Consider these two alignments for the first and second exon of some hypothetical gene. In \textit{D. melanogaster} the first exon is 68 aa long and the second exon is 28 aa long:

Exon 1 sequence:

```
    1  MDINNEIENISIDINIKQAQKLKEQELKAQQYQQN
  41  YNPASGPITETTTTTVVTKDSEEET
```

And alignment to our hypothetical genomic region:

```
Query  5736  MDINNEIENISIDINIKQAQKLKEQELKAQQYQQN
       MD NN+I NIISIDINIKQAQKL+ E ++ +    +Q
Sbjct  1  MDFNNQILNIISIDINIKQAQKLQNECQSGELDLH  38
```

Exon 2 sequence:
Notice the strong alignment at the start of exon 1. This gives good evidence where exon 1 starts but given its length in *D. melanogaster* it may be difficult to find the donor site at the end of exon 1. We would certainly follow the exon length conservation rule above and look downstream about 90 bases (i.e. we are missing 30 aa or 90 bases of the end of the first alignment), but this just gives the general area where we might expect the end of the exon. Interestingly, exon 2 starts with a very strong alignment. Thus, when considering the amino acid sequence of the *D. melanogaster* protein with the amino acid sequence of the potential new gene, we have identified two conserved domains separated by a region without conservation. It would be somewhat unlikely that the 5' end of the downstream conserved domain coincides exactly with the 5' end of exon 2. If these two 5' ends do not coincide we would expect exon 1 should end with at least a few conserved amino acids. Since BLAST did not detect these amino acids (i.e. only the 5' end of exon 1 shows an alignment) it is likely that the number of conserved amino acids at the 3' end of exon 1 is very small (one or two). In these cases a search for a short DNA sequence that would code for one or two conserved amino acids next to an in frame splice junction may be fruitful. In the example above then exon 1 might end with the same 1 or 2 aa as the exon in *D. melanogaster*. To start the search we must first find the phase of the acceptor site at the beginning of exon 2. Since there is a very strong alignment that ends at the first amino acid we expect an acceptor site to be 0, 1 or 2 bases upstream of base 6259. In this example we will assume that the acceptor site at the start of exon 2 immediately precedes the codon for the glutamic acid (E). As such we would look for a DNA sequence to end exon 1 that codes for the amino acid E (glutamic acid) and then T (threonine) and then a donor site. If the acceptor site at the beginning of exon 2 had one base and then the codon for the glutamic acid we would look for a sequence to end exon 1 which was a codon for E, a codon for T, any single base and then the donor site. Since the codon table is degenerate we will need to look for a number of different sequence that could code for ET(donor). Checking the codon table we see that E has two codons, GAG and GAA, and T has four codons, ACT, ACC, ACA, ACG. If these two amino acids were conserved and we want a phase of 0 to match the acceptor of exon 2, then any of 8 different sequences could code for ET(donor):

- GAGACTGT
- GAGACCGT
- GAGACAGT
- GAGACGGT
- GAAACTGT
- GAAACCGT
- GAAACAGT
- GAAACGGT

If the potential region where these amino acids can be is small it is possible to simply search by eye. It is also possible to use BLAST to search your sequence if you change some of the parameters to specifically allow for these very short alignments. First, set the word size to a number less or equal to the length of the sequencing you are searching with. Also, be sure to turn OFF the filter and set a very large cut-off score (different implementations of BLAST calculate E-scores differently when comparing two sequences, it is best to experiment with the version of BLAST you are using to empirically determine the best cut-off score). Since there are 8 different ways to code for “ET(donor)” you would need to do 8 BLAST searches. (To ensure that the BLAST tool you are using can detect these very small alignments you may wish to do a positive control search with a sequence you know does exist within your subject to verify that it can be found).

If one of these sequences is found in the correct location (i.e. downstream of the exon 1 alignment but before exon 2 and on the correct strand) and it is in the correct frame (i.e. the same frame as the early exon 1
alignment) and is in the proper phase (i.e. links with the correct frame in exon 2) you have found pretty strong evidence for the end of exon 1 and this site should be picked.

**Very small exons.** It can be quite difficult to find very small exons. Be sure to increase the expect cut-off score until you start to see hits no matter how poor the alignment looks. To avoid sorting through lots of false alignments you can restrict the region searched. For example, if you have the upstream and downstream exons already mapped, do not search the entire region; rather use the “from:” and “to:” boxes in BLAST 2 sequences to restrict your search to the region between the two mapped exons. You can also try to make the search more sensitive by changing the word size from 3 to 2 but be aware that this may cause you to miss high quality blast hits if the sequence you are searching is large so be sure to restrict the area searched using the above technique if you reduce the word size. Remember very weak similarity, if in the right place and with usable donor and acceptor sites, is probably identifying the correct exon.

The initial exon is often a very small exon. These exons can be searched for using the short degenerate search trick described above, but will often fail. In this case, you should look for any candidate methionine codon upstream of your second exon which is also very close to an in frame medium or high quality donor site. If more that one is found in the proper region pick the one closest to the next exon.

**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 which 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.