Efficient Annotation of Genes with multiple variants

The process of annotating genes with multiple splice variants can be greatly streamlined and simplified. Since many of the exons are shared between variants, an effective approach is to determine how many unique variants are associated with a gene and the particular exons associated with each variants. The unique exons can be discovered in *D. erecta*, compiled and then assembled to complete the annotation for each variant. The steps to achieve this are indicated below. Note that it is advisable that students should have prior exposure to annotation or work as teams to make this approach feasible.

Annotation Summary for coding exons

1. Get predicted GenScan contig and protein from the goose genome browser.
2. Run this predicted protein vs. Flybase BLAST.
3. Get best match identity and note how much of your protein matched.
4. Find this annotated protein in Ensembl.org
5. Note the number of variants, number of exons and number of peptide residues in your Word documentation.
6. Pull the protein sequence for each variant individually and copy to your Word document.
7. You can now begin to isolate each unique variant. There are a number of methods to do so.
    a) One approach (**See Multiple variant annotation for example)** is to list each exon separately for each variant is MS Word (remember to include a split codon with each exon). Starting with Exon 1, Variant A, use the Find Command to search for identical sequences. Highlight any finds. Now proceed to exon 2, variant A and repeat the process. When you have completely tested Variant A, proceed to Variant B and carry out the same procedure. Note that any highlighted exons have already been discovered.
        a. With each variant, it becomes clear that most exons are shared. At the end of the process you have documented all unique exons and are ready to begin annotating them.
        b. Collect all unique variants in a table and gather their coordinates for annotation. A final summary table will allow assembly of the annotation information for each variant.
        c. Create a 2nd table listing the coordinates for each variant. These can later be entered for .gff file creation.

8. EXON Discovery and Documentation **(See annotation cheat sheet for example).** Students should carry out normal documentation procedure for each exon, including screen capture, identification of exon boundaries via splice sites, documentation of phasing of split exons and any unusual properties exhibited by the exons. Details are found in the annotation document. Highlighted steps include:

9. BLAST 2 seq (NCBI)- Using tblastn, search for each Dm exon individually against the entire contig DNA. Be certain to use the whole contig, (make sure you do not have the reverse complement); otherwise the numbering will be incorrect.

10. Copy the BLAST outcome making certain to include the lines of documentation above the actual sequence match. Paste into your word document.

11. Using the coordinates from the BLAST search, examine the sequences in the Goose Genome Browser. Precisely determine the ends of the exons and introns by finding and identifying the splice junctions. Note also: the reading frame, the phase of the last codon in the exon: (0) = codon complete; (1) or (2) means that there are 1 or 2 nucleotides remaining. Record this data for both ends of the exons.

12. Capture an image of the genome browser for the beginning and end of each exon. Note the stop/start codon coordinates under each capture. Also, note the supporting evidence (RefSeq, Genscan splice predictors, etc.)

13. Here are the parameters that you should discover for each variant:

a) location of start codon and stop codon, GT splice donor at the end of an exon, AG splice acceptor at the beginning of an exon, the precise locations at end of an exon. I recommend the following formulation of this information for later validation.

Coding Exon #	AG splice acceptor	Beginning of exon (Phase)	End of exon (Phase)	GT splice donor
1	NA	4346 (0)	4467 (1)	4468-69
2	4561-62	4563 (2)	4689 (0)	4690-91
3	4719-20	4721 (0)	4805 (2)	4806-07
4	4911-12	4913 (1)	5265 (E)	5266-8(stop)

Note that the phase of the end of a preceding exon and the phase of the beginning of a following exon should produce a complete codon. Therefore, either both are in phase 0 OR the phases add to 3.

Validation format: 4346-4467(1), 4563(2)-4689(0), 4721(0)-4805(2), 4913(1)-5265
(Note: place smaller number before larger in the case of annotating in the negative frame)

Use the ORF Translator at www.dnalc.org/bioinformatics or suitable program to build the protein. If stop codons are shown anywhere other than the last codon, check the beginnings and ends of each exon. It may be advisable with challenging variants to add one exon at a time and look for a continuous ORF.

12) Carefully examine the De DNA for each exon.

a) make certain that you find a start and stop codon
b) make sure that the phases are in frame for each exon.

Most student problems stem from incorrect location of the borders of an exon, which must be precise and in frame. Occasionally, you will have to hypothesize a different location for an exon border. Please note and justify any such anomalies which can include but are not limited to:

a) rare exon donor/acceptor sites – e.g. GC (donor), AC (acceptor). Remember that 99% of all exons are bordered by the traditional GT. . .AG borders. These rare sites are only invoked when traditional borders are not found, they are in frame and they are reasonable.
b) 'lost' start or stop codons. In rare instances, a protein may not begin with a canonical ATG but an alternative. You should be very persistent and diligent in searching for it. There are several explanations possible:
    a. indeed, there is a different amino acid (not met) beginning the protein.
    b. The start/stop codon is located in a hybrid non-coding/coding exon.
    c. The variant does not really exist in De and this is a pseudogene. This should not be evoked until an exhaustive search for anything missing has been done.

Useful tips in looking for

A) Problematic exons

1)Exons are very short

a. Raise the (E) value in tblastn from 10 to 100 to 1000 etc. progressively until you location a probable match. Since you have other surrounding exons, you have an idea of the approximate location of the short exon. Usually, you can locate the missing exon in this manner.
b. If the short coding exon is at the beginning or end of the gene, it may be part of a non-coding exon (hybrid). It may be instructive to locate it by using blastn against the contig and finding the hybrid borders. This will often reveal not only the missing aa codons but the location of the non-coding regions.

2) Cannot find start/stop

a. sometimes the start/stop can be isolated in a non-coding exon. Again, you can use the non-coding exons and blastn to explore whether this possibility has occurred.

3) Exons discovered have regions within them which are not well-matched. Although, in general, most exon sizes are somewhat conserved between Dm and De, occasionally, exons can be split or fused. This is very challenging and will require a careful search for mismatching, splice sites, etc. Please carefully document any such event.

4) Genscan has fused/cut off exons. This happens with some regularity if Genscan does not pick up start/stop codons, normal GT/AG boundaries, etc. Again, ust the Dm ortholog as your guide.

Annotation of UTR's for multiple variants

Efficient annotation of UTRs in a multi-variant gene can also be achieved in a similar manner. The annotation of UTRs is described in the **Annotation of UTRs** document. Again, a preliminary scan to discover how many unique UTRS are associated with a Dm gene can save a good deal of time and effort in annotation.

A good starting point is to gather the chromosome locations of the Dm UTRs in a table for all of the variants (see below). It is quite easy to distinguish common exons for the variants. Again, we now only need to annotate 3 forms of exon 1, 2 forms of hybrid exon 2, and 2 forms of exon 14.

Flybase gene: CG32018 (Zyx102EF)
Non-coding exons

Variant	**A**	**B**	**C**	**D**	**E**	**F**	**G**
Exon 1	1,080,991 1,081,166	1,080,991 1,081,166	1,081,077 1,081,103	1,081,077 1,081,103	1,081,023 1,081,136	1,081,023 1,081,136	1,081,023 1,081,136
Intron	1,080,885 1,080,990	1,080,904 1,080,990	1,080,904 1,081,076	1,080,885 1,081,076	1,080,904 1,081,022	1,081,023 1,081,136	1,080,087 1,081,022
Exon 2	1,080,663 1,080,884	1,080,663 1,080,903	1,080,663 1,080,903	1,080,663 1,080,884	1,080,663 1,080,903	1,080,663 1,080,884	1,079,669 1,080,086
Exon 14	1,077,606 1,077,822	1,077,608 1,077,822	1,077,608 1,077,822	1,077,606 1,077,822	1,077,608 1,077,822	1,077,606 1,077,822	1,077,608 1,077,822

Exon 1- A, B
Exon 1- C, D
Exon 1- E, F, G
Exon 2- A,B,C,D,E,F
Exon 2- G
Exon 14 – A, D, F
Exon 14- B, C, E, G

The annotation data can be assembled in table for each variant and linked to the coding exon annotation. (See **Annotation V. All exons** below). This document now contains an organized, compact gene model for .gff creation and the completion of the annotation project.

Please direct any questions or concerns to:
Dr. Gary Kuleck (gkuleck @lmu.edu) 310-338-7496
or
Nicole Yu (nyu1@lmu.edu)