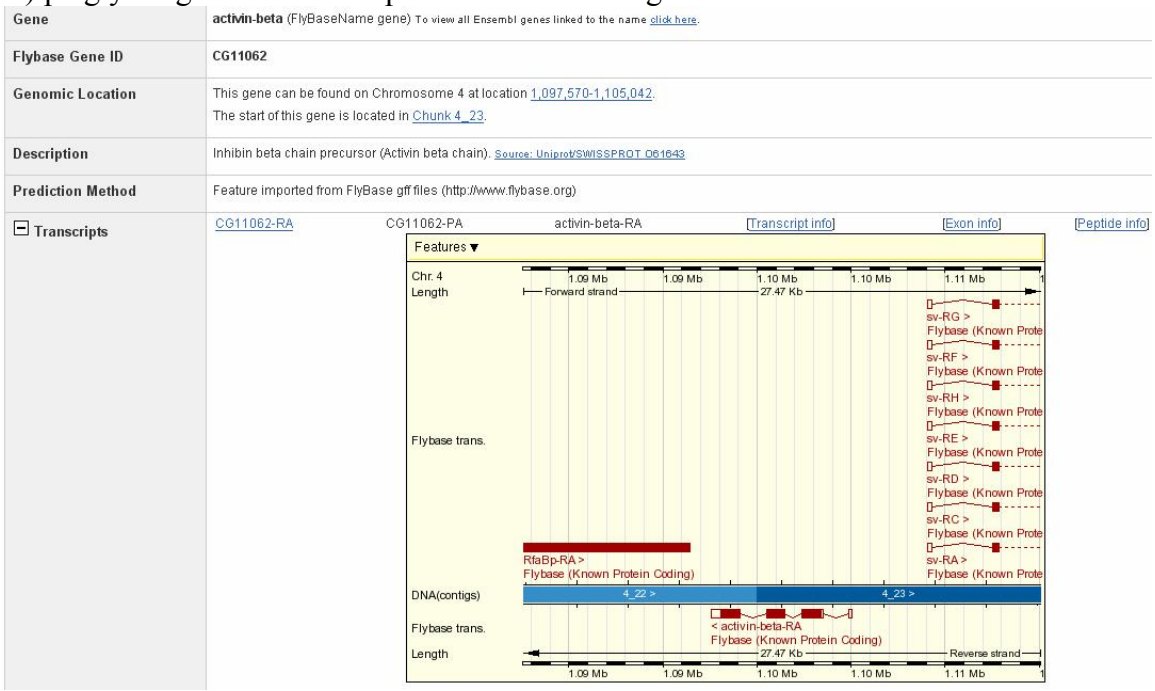
Annotating individual variant UTR regions

I. Finding 5' UTRs (CG11062, activin-beta)

Start as you would for annotating coding exons

1.) plug your gene ID into http://www.ensembl.org/index.html



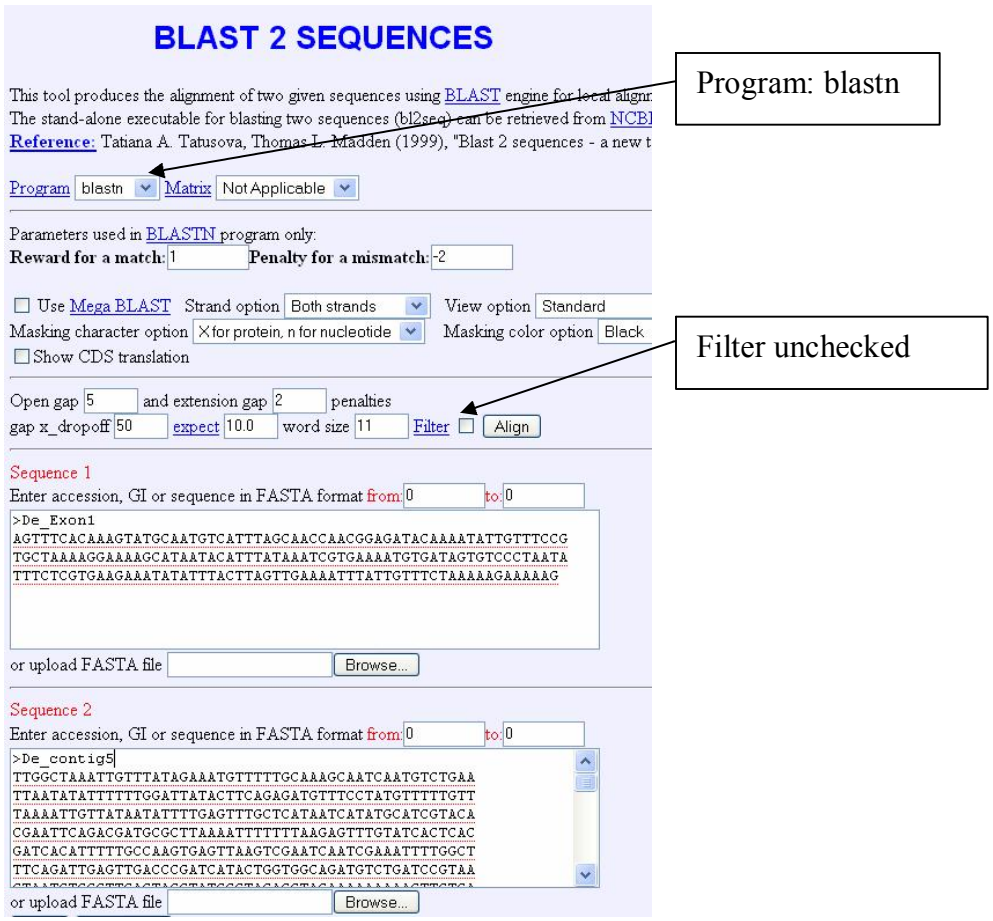
→Click on [Exon info]

2.) In the this example we are using contig 5.3 variant A, you will get a screen that will look similar to this:

No.	Exon / Intron	Chr	Strand	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence								..........aaatattgttattttcaagaaagcttttgaaaagcttaaagtgtcatctt
1	activin-beta:1:1104864:1105042	4	-1	1,104,864	1,105,042	-	-	179	AGTTTCACAAAGTATGCAATGTCATTTAGCAACCAACGGAGATACAAAATATTGTTTCCG TGCTAAAAGGAAAAGCATAATACATTTATAAATCGTGAAAATGTGATAGTGTCCCTAATA TTTCTCGTGAAGAAATATATTTACTTAGTTGAAAATTTATTGTTTCTAAAAAGAAAAAG
	Intron 1-2	4	-1	1,103,524	1,104,863			1,340	gtgatacatattgttacatacactg..........aatcgaaaatttgtattttgaacag
2	activin-beta:2:1102377:1103523	4	-1	1,102,377	1,103,523	0	1	1,147	GCAGACGAAATGAACGACAGAAAGGAAAGGAAAGGAAAAAGATTTGCAATGATTAAAAAA ATATTTACCAGAAAAATAAAATTGACAATACTTGTACCACGTTTCTAAGGGACAGTGCCT GTACTACACGCTTGCCATATATCAAAGATGCGATTTGCCTTCGATTCTAATCATTCGCAA TCGGGGGCGCCATTCAAAGGCAGCAGGTGTTTCTTTAATTGTCAATGCATCTGCTGTCGC CAAGGATGCTGCGTTGTGGTTGTAAAGTGCTGTTGCTGCTTTAACTTAAACTGCTGCAAC AGCCTTGGCTCCCGGAAGTCATTTCCACAACCCGCTGCAATGCGTAAAAAAGTTGCTGAC CTCGAAGTCCTTAGAGTATCAAGGTTTGTGGCGGTTATTTTAGTGCTGGCTCGGTGGGTT ACTGCGGTAGCGACACTCCTGACAAGCTGCATACTCCTAGACATATTTTCCGTGCCTGGC CAGTCTGGAGTTGCAGATAGAAGCCAAGCCAGCAGTAGGACAGTGCACGTCTCGGTTCCT ACCACACCTAATGAAACTCCCAGTAGCACTTCGGAGACGAAGCTAAAGTTGCTTTATGGG TATACATCGTATGACATAAATAACGACCAACAGGTAAAGTCCAACAATTTATGTAGAGTG CTTTGTAAAAGTCGCAATCGTAAACGACAGCGAAGGAGGCGACGCCGACGCAATCACAGA CGACGCAGGCACAGATATACTAAGCGACTTCATCATCTAATGCAAGATAATATGAGCGGC TTTGAGCAAAGACTTAATTTTAGCGATGCCAAATGCCAGTCTTTGGAGACAAATTACGGA

→The purple indicates UTR regions. Note that Exon 2 is a hybrid exon, containing part UTR and part coding sequence (in black).

3.) Take the DNA sequence (179 nt) from the first exon and blastn it against the entire De contig5.



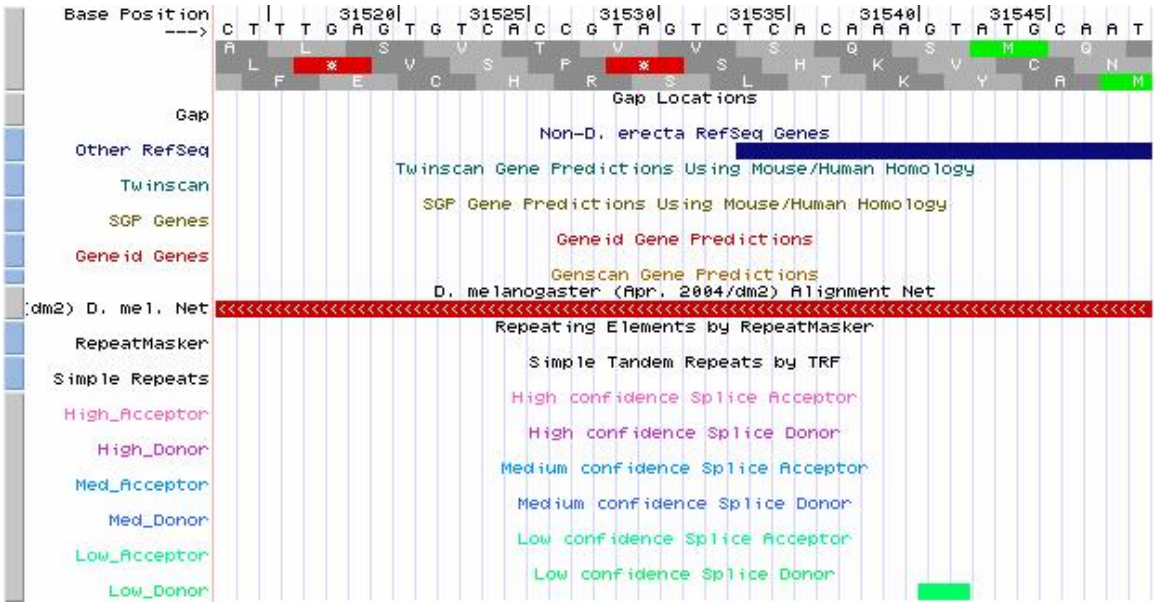
Program: blastn

Filter unchecked

→Click align

4.)Copy your result and paste it into your word document:

```
Score =  133 bits (69),  Expect = 2e-28
 Identities = 95/108 (87%), Gaps = 0/108 (0%)
 Strand=Plus/Plus

Query  1      AGTTTCACAAAGTATGCAATGTCATTTAGCAACCAACGGAGATACAAAATATTGTTTCCG  60
              ||| ||||||||||||||||||||||||||||| |||||||||||| ||||||||||
Sbjct  31530  AGTCTCACAAAGTATGCAATGTCATTTAGCAACCGACGGAGATACAATATATTGTTTCTA  31589

Query  61     TGCTAAAAGGAAAAGCATAATACATTTATAAATCGTGAAAATGTGATA  108
              |||| |||| ||||||| || || | || ||| ||||||||||||||
Sbjct  31590  TGCTGAAAGAAAAAGCAAAAAACGTGTACAAAACGTGAAAATGTGATA 31637
```



The start of exon 1, AGT is found at 31530.

→Note that this is only part of the first exon. The blastn query reveals a match up for nt #s 1-108, but the first exon is 179 nts long. This is addressed further down.

Seen below in green is where the last part matches up to in blast.
AGTTTCACAAAGTATGCAATGTCATTTAGCAACCAACGGAGATACAAAATAT
TGTTTCCGTGCTAAAAGGAAAAGCATAATACATTTATAAATCGTGAAAATGT
GATAGTGTCCCTAATATTTCTCGTGAAGAAATATATTTACTTAGTTGAAAATT
TATTGTTTCTAAAAAGAAAAAG

```
Score =  133 bits (69),  Expect = 2e-28
 Identities = 95/108 (87%), Gaps = 0/108 (0%)
 Strand=Plus/Plus

Query  1      AGTTTCACAAAGTATGCAATGTCATTTAGCAACCAACGGAGATACAAAATATTGTTTCCG  60
              ||| ||||||||||||||||||||||||||||| |||||||||||| ||||||||||
Sbjct  31530  AGTCTCACAAAGTATGCAATGTCATTTAGCAACCGACGGAGATACAATATATTGTTTCTA  31589

Query  61     TGCTAAAAGGAAAAGCATAATACATTTATAAATCGTGAAAATGTGATA  108
              |||| |||| ||||||| || || | || ||| ||||||||||||||
Sbjct  31590  TGCTGAAAGAAAAAGCAAAAAACGTGTACAAAACGTGAAAATGTGATA  31637
```

→To find the end, extend the *D. erecta* DNA sequence and compare it in ClustalW. The basic assumption is that the sequence is present for the remainder of the nts in *D. erecta.* A good start is to the DNA sequence by the missing 71 nts to 31708.

```
>Dere2_dna range=contig5:31530-31708 5'pad=0 3'pad=0 revComp=FALSE
strand=? repeatMasking=none
AGTCTCACAAAGTATGCAATGTCATTTAGCAACCGACGGAGATACAATATATTGTTTCTATGCTGAAAGAA
AAAGCAAAAAACGTGTACAAAACGTGAAAATGTGATATTTCTTACAGAGAGTGTTTTATAATATTTCTGTT
AAAGAAATCGACCGTGAACAAATAGTATTTTCATTAA
```

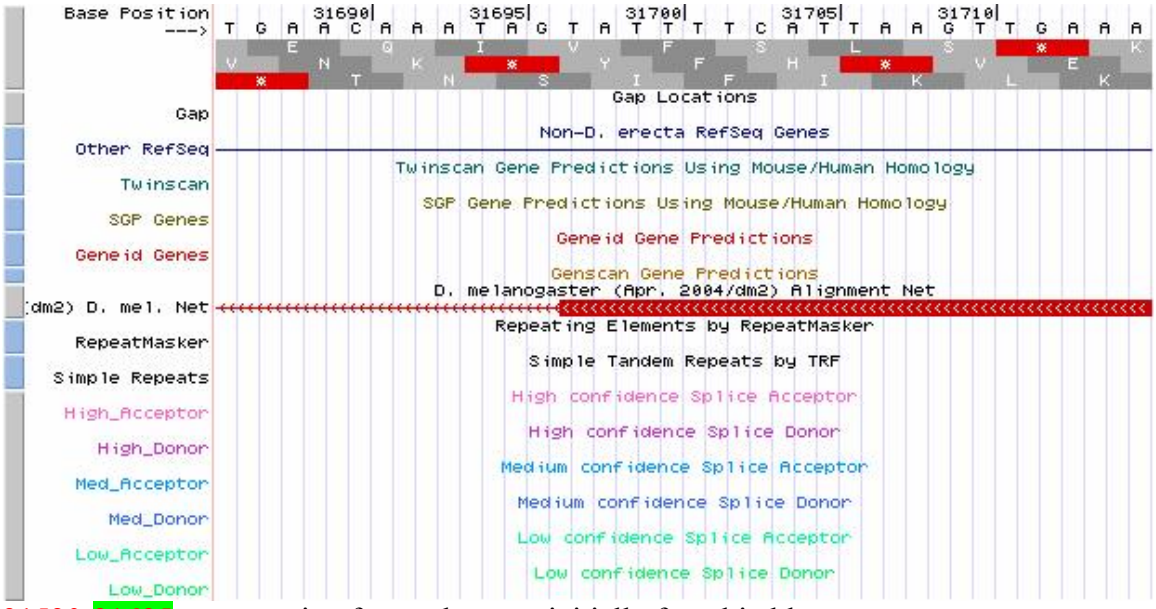
31530-31708

```
de             AGTCTCACAAAGTATGCAATGTCATTTAGCAACCGACGGAGATACAATATATTGTTTCTA 60
dm             AGTTTCACAAAGTATGCAATGTCATTTAGCAACCAACGGAGATACAAAATATTGTTTCCG 60
               *** ***************************** ************ **********
```

```
de              TGCTGAAAGAAAAAGCAAAAAACGTGTACAAAACGTGAAAATGTGATA---TTTCTTACA 117
dm              TGCTAAAAGGAAAAGCATAATACATTTATAAATCGTGAAAATGTGATAGTGTCCCTAATA 120
                **** **** ******* ** ** * ** *** ***************   *  ** * *

de              GAGAGTGTTTTATAATATTTCTGTTAAA---GAAATCGACCGTGAACAAATAGTATTTTC 174
dm              TTTCTCGTGAAGAAATATATTTACTTAGTTGAAAATTTATTGTTTCTAAAAAGAAAAAGG 179
                      **     ***** * *  * *     ****  *  **    *** ** *

de              ATTAA 179
dm              T---- 179
```

Note the GT splice junction in Dm sequence. There are 28/71 nt matches, not enough of a match for blastn to discover but indicative of some sequence preservation. Note the Dm GT splice junction (in red) is not preserved; an upstream site in De (in green) is hypothesized to end exon 1 at 31695.

→ The end is found at the end of the match for clustalW



31530-31695 an extension from what was initially found in blast.

5.) Now look for the second exon. Note that this is a hybrid exon, meaning that it contains both UTR and coding sequence.

Exon2

GCAGACGAAATGAACGACAGAAAGGAAAGGAAAGGAAAAAGATTTGCAATGATTAAAAAA
ATATTTACCAGAAAAATAAAATTGACAATACTTGTACCACGTTTCTAAGGGACAGTGCCT
GTACTACACGCTTGCCATATATCAAAGATGCGATTTGCCTTCGATTCTAATCATTCGCAA
TCGGGGGCGCCATTCAAAGGCAGCAGGTGTTTCTTTAATTGTCAATGCATCTGCTGTCGC
CAAGGATGCTGCGTTGTGGTTGTAAAGTGCTGTTGCTGCTTTAACTTAAACTGCTGCAAC
AGCCTTGGCTCCCGGAAGTCATTTCCACAACCCGCTGCAATGCGTAAAAAAGTTGCTGAC
CTCGAAGTCCTTAGAGTATCAAGGTTTGTGGCGGTTATTTTAGTGCTGGCTCGGTGGGTT
ACTGCGGTAGCGACACTCCTGACAAGCTGCATACTCCTAGACATATTTTCCGTGCCTGGC
CAGTCTGGAGTTGCAGATAGAAGCCAAGCCAGCAGTAGGACAGTGCACGTCTCGGTTCCT
ACCACACCTAATGAAACTCCCAGTAGCACTTCGGAGACGAAGCTAAAGTTGCTTTATGGG
TATACATCGTATGACATAAATAACGACCAACAGGTAAAGTCCAACAATTTATGTAGAGTG
CTTTGTAAAAGTCGCAATCGTAAACGACAGCGAAGGAGGCGACGCCGACGCAATCACAGA
CGACGCAGGCACAGATATACTAAGCGACTTCATCATCTAATGCAAGATAATATGAGCGGC
TTTGAGCAAAGACTTAATTTTAGCGATGCCAAATGCCAGTCTTTGGAGACAAATTACGGA

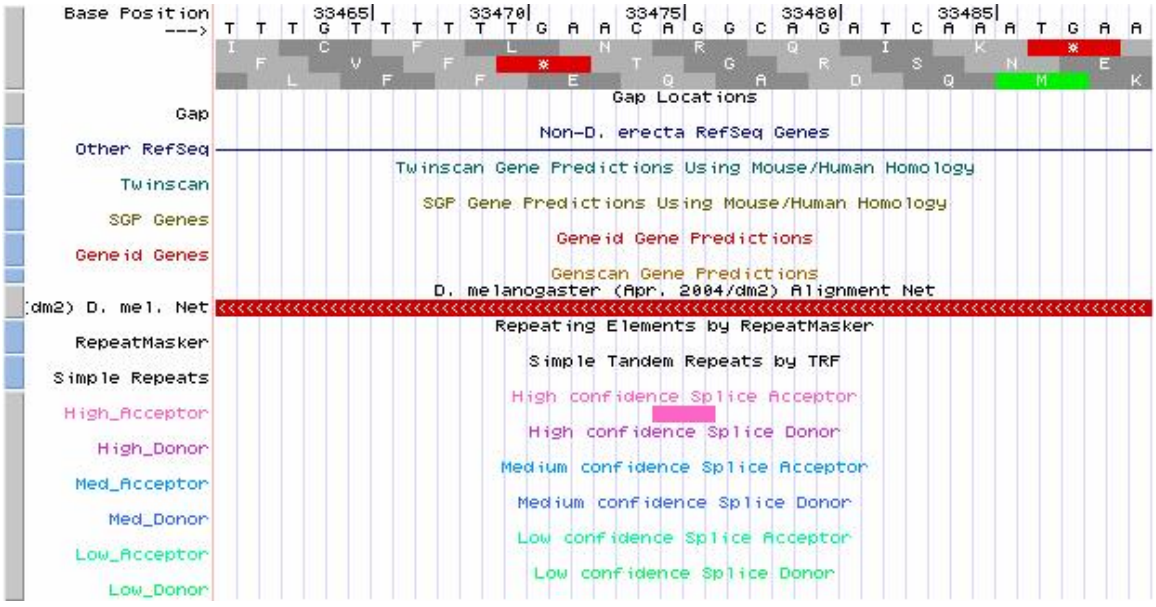
```
ACTAATTATGACTTAGTACAAGGAGGTAAACTATTTAGTCAGTCAGAGAGAAGCCTACTG
GTGTCCCCTTTGAGGGAAATTGAAGCACCTTGGCCAGCGATTCATGGTTCAATGCGTAAC
TGTTCAAAGATTAAACGCAATAGAGCCAATCTTATTTGGCTTCTAATTGGACTCGTCTGG
TTTGAAGTCAAACTTATAAATTGCAATGGGATCAGCAGTAGTAATTATTATGCTTCGAAT
TTGGAGAGTCACAAGGGCTGCACCTTGTGCCATGAAAGCGGAAAGCCCAACATATACACC
GATAAAG
```

Since we have already located the coding exon, we need to look for the upstream 180 nts of the non-coding segment of this hybrid exon. Note that beginning of the exon is not discovered by BLASTn.

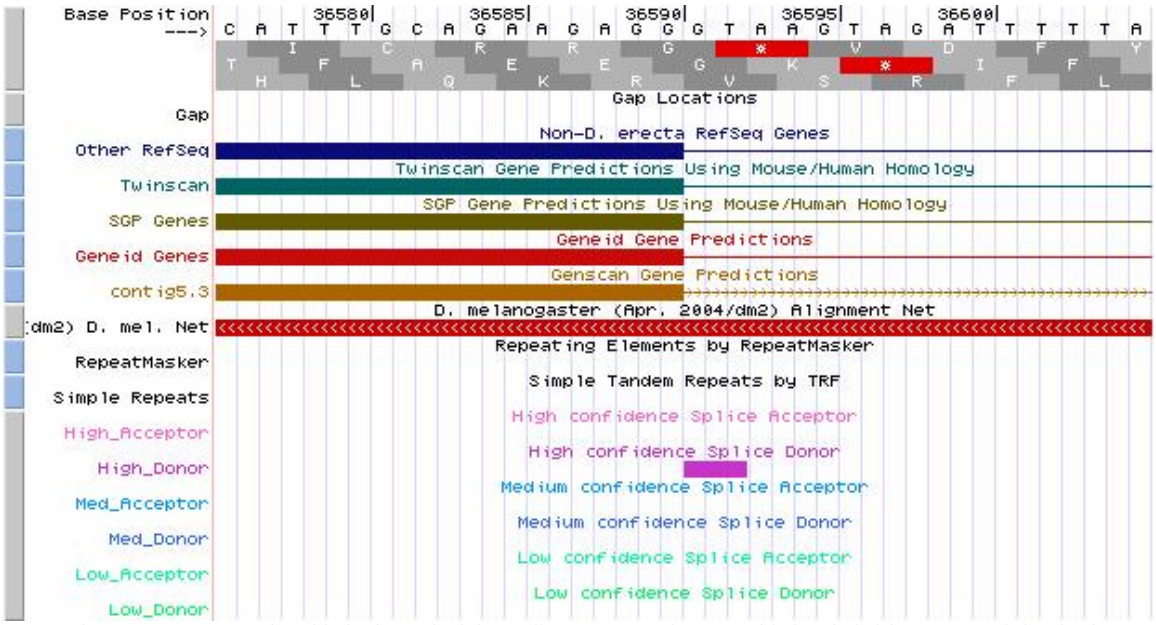
```
Score =  110 bits (57),  Expect = 2e-21
 Identities = 91/108 (84%), Gaps = 0/108 (0%)
 Strand=Plus/Plus

Query  35     GAAAAAGATTTGCAATGATTAAAAAAATATTTACCAGAAAAATAAAATTGACAATACTTG  94
              ||||||||||||||||  |   ||||| |||||||| ||||||||| || || ||||||
Sbjct  33504  GAAAAAGATTTGCAATGGATCGGAAAATGTTTACCAGGAAAATAAAAGTGCCATTACTTG  33563

Query  95     TACCACGTTTCTAAGGGACAGTGCCTGTACTACACGCTTGCCATATAT  142
              || ||| | |||||||| || |||| ||||||||||||||||| ||||
Sbjct  33564  TATCACTTCTCTAAGGGGCACTGCCAGTACTACACGCTTGCCAAATAT  33611
```

→ To find start and end sequences it is sometimes helpful to use the intronic sequences before or after the exon. Sometimes conservation provides a good match.

```
Search of intronic region: aatcgaaaatttgtattttgaacag
 Score = 39.1 bits (20),  Expect = 0.13
 Identities = 22/23 (95%), Gaps = 0/23 (0%)
 Strand=Plus/Plus

Query  3      TCGAAAATTTGTATTTTGAACAG  25
              |||||||||||| ||||||||||
Sbjct  33454  TCGAAAATTTGTTTTTTGAACAG  33476
```



start of Exon2 at 33477 with GCAGATCAAA



Ex2 End at 36590(1), note that this is the end of the coding sequence because this is a hybrid exon containing both UTR and coding sequence.

Ex2: 33476 – 36590(1)

These strategies should be used for each of the UTR regions as well as the leader and end sequences.

II. Finding 3' UTRs.

1) hybrid exons may be found by extending the De DNA sequence beyond the stop codon using the techniques and guidelines described herein.
2) Non-coding exons may also be discovered as described below.

We will find the 3'UTR for the activin-beta gene. Using blast2seq, search for the Dm sequence in the entire De contig5. Here are the results. Note that blastn discovered nearly all 510 bp.

```
Score =  233 bits (121),  Expect = 6e-58
 Identities = 162/180 (90%), Gaps = 1/180 (0%)
 Strand=Plus/Plus

Query  113    ATATACATAAAGTAGACTCAATTTTATTTTATACTTAGCTATGCTGGTGACAATATTTGT  172
              ||||| |||||||||||||||| ||||||||| ||||  |||||| ||||||||||||
Sbjct  38186  ATATAAATAAAGTAGACTCAATTATATTTTATATTTAGACATGCTGATGACAATATTTGT  38245

Query  173    ATATTTACGAACAAATCCAAATTGAGGAAGTGCCTAAATTACGTTAATGAAATATTTGTA  232
              ||||||||||||| || ||||||||||||||||||||||| |||  ||||||| ||
Sbjct  38246  ATATTTACGAACAAACCCTAATTGAGGAAGTGCCTAAATTACGTAAAT-CAATATTTATA  38304

Query  233    CATTTTAAGAATATTTCAAAAATCACTAAAATATCTTTGTACTAATAAAATTCGATATTT  292
               || |||||||||||||| | |||||| |||||||||||||||||||||| |||||
Sbjct  38305  AATATTAAGAATATTTCAAATAACACTAAGATATCTTTGTACTAATAAAATTCGCTATTT  38364

_____________________________________________________________________________
Score =  177 bits (92),  Expect = 3e-41
 Identities = 166/198 (83%), Gaps = 4/198 (2%)
 Strand=Plus/Plus

Query  313    TATGTTCAAATTAACATAAGTATAAAACGAAATGATTTAATAACCTATAACATGAGCAAA  372
```

```
            |||||||||    |||||||| ||||| ||||  ||||||| ||||||| ||||||
Sbjct  38372  TATGTTCAATAATACATAAGTACAAAACCAAATATTTTAATAGCCTATAATATGAGC---  38428

Query  373    GCGTCGCGCTTTTTTTATTGTCAAAACATTAATTTTACTAACTTGAAAAGCTTATATCAC  432
            ||  ||| ||||  |||||||| ||||||| |||||||||||||||||| || |||||||
Sbjct  38429  GCCGCGC-CTTTCGTTATTGTCTAAACATTCATTTTACTAACTTGAAAACCTAATATCAC  38487

Query  433    AGACATGTAATAAATATTTCATATTACAGTTTAATAAAGTATTAATATAAGGATTACTAT  492
            |||   ||||||  |||||||||||||| |||||||||||||||| ||||||||||||||
Sbjct  38488  AGATTAGTAATACTTATTTCATATTACACTTTAATAAAGTATTAAAATAAGGATTACTAT  38547

Query  493    ATGAAATAAAATAAATTT  510
            |||||||  ||| |||||
Sbjct  38548  ATGAAATTCAATGAATTT  38565
```

```
Score = 54.5 bits (28),  Expect = 4e-04
 Identities = 83/108 (76%), Gaps = 10/108 (9%)
 Strand=Plus/Plus

Query  3      TGCCTTACAATTTTATATTTTCCGTCCGATAGAAATAAAAAT----------ATATGTGT  52
            |||||| ||||||||||||||||||  |||| ||||| ||          ||| ||||
Sbjct  38079  TGCCTTTCAATTTTATATTTTCCGTCCTGTAGACATAAACATCTAGACATTGATAAGTGT  38138

Query  53     CCTAACTGAGCGCCAATCTCTTAACGAAATCTTTTACTTTTAAGTTAA  100
            ||||||| |||| |||  ||  |||||| ||||| |||||| ||||||
Sbjct  38139  CCTAACTTAGCGTCAAAGTCCAAACGAAGTCTTTAACTTTTTAGTTAA  38186
```

Reconstructing the 3 segments:
Dm: 3-100 , 113-292, 313-510
De: 38079-38186, 38186-38364, 38372-38565,

If we extend the De sequence upstream by 5 nts (38074 (includes potential splice site), we have an approximate starting point to initiate a ClustalW analysis to finish the annotation. We recover contig5:38074-38565 from the goose server to initiate the analysis.

```
Dere2_contig5_38063-38565      AGCTTGATCCGGTATCTGCCTTTCAATTTTATATTTTCCGTCCTGTAGAC
Dm_3'UTR                       ------------AGCTTGCCTTACAATTTTATATTTTCCGTCCGATAGAA
                                           ****** ********************  ****

Dere2_contig5_38063-38565      ATAAACATCTAGACATTGATAAGTGTCCTAACTTAGCGTCAAAGTCCAAA
Dm_3'UTR                       ATAAAAAT----------ATATGTGTCCTAACTGAGCGCCAATCTCTTAA
                               ***** **          *** *********** **** ***  **  **

Dere2_contig5_38063-38565      CGAAGTCTTTAACTTTTTAGTTAA-------------TATAAATAAAGTA
Dm_3'UTR                       CGAAATCTTTTACTTTTAAGTTAAAAGTACATAAATATATACATAAAGTA
                               **** ***** ****** ******             **** ********

Dere2_contig5_38063-38565      GACTCAATTATATTTTATATTTAGACATGCTGATGACAATATTTGTATAT
Dm_3'UTR                       GACTCAATTTTATTTTATACTTAGCTATGCTGGTGACAATATTTGTATAT
                               ********* ********* ****  ****** ****************

Dere2_contig5_38063-38565      TTACGAACAAACCCTAATTGAGGAAGTGCCTAAATTACGTAAATCAA-TA
Dm_3'UTR                       TTACGAACAAATCCAAATTGAGGAAGTGCCTAAATTACGTTAATGAAATA
                               *********** ** ************************ *** ** **

Dere2_contig5_38063-38565      TTTATAAATATTAAGAATATTTCAAATAACACTAAGATATCTTTGTACTA
Dm_3'UTR                       TTTGTACATTTTAAGAATATTTCAAAAATCACTAAAATATCTTTGTACTA
                               *** ** ** **************** * ****** **************

Dere2_contig5_38063-38565      ATAAAATTCGCTATTTC-------------ATTAATTATGTTCAATAATA
Dm_3'UTR                       ATAAAATTCGATATTTTTAATTATTACTTTATTTAATATGTTCAAATTAA
                               ********** *****              *** * *********    *

Dere2_contig5_38063-38565      CATAAGTACAAAACCAAATATTTTAATAGCCTATAATATGAGCGCCGCGC
Dm_3'UTR                       CATAAGTATAAAACGAAATGATTTAATAACCTATAACATGAGCAAAGCGT
```

```
                                ******** ***** ****  ******* ******* ******   ***

Dere2_contig5_38063-38565       C----TTTCGTTATTGTCTAAACATTCATTTTACTAACTTGAAAACCTAA
Dm_3'UTR                        CGCGCTTTTTTTATTGTCAAAACATTAATTTTACTAACTTGAAAAGCTTA
                                *    ***  ******** ******* ***************** ** *

Dere2_contig5_38063-38565       TATCACAGATTAGTAATACTTATTTCATATTACACTTTAATAAAGTATTA
Dm_3'UTR                        TATCACAGACATGTAATAAATATTTCATATTACAGTTTAATAAAGTATTA
                                *********   ******  ************** ***************

Dere2_contig5_38063-38565       AAATAAGGATTACTATATGAAATTCAATGAATTT
Dm_3'UTR                        ATATAAGGATTACTATATGAAATAAAATAAATTT
                                * ********************  *** *****
```

Commentary: Note the strong conservation of sequence similarity. The difficulty arises at the upstream part of the 3'UTR. Note that the Dm AG splice site (red) is not preserved but a high confidence AG acceptor site (green) is present at 38063. Therefore, we hypothesize that the De 3'UTR lies at 38065-38565.

III. Handling challenges of finding 5' and 3' UTRS.

1) Suppose that the techniques described do not work. Here are some alternative strategies.

a. Use sequences upstream of the 5'UTR to 'anchor' the first exon. They are often highly conserved.

b. Use sequences downstream of the 3'UTR to anchor the last exon. They are often highly conserved.

c. Intronic sequences are also sometimes preserved and could serve as a point of departure to find nearby exons. This step, however, should be done if all else fails.

IV. Summary of strategies

1) Since we working with non-coding regions, it is necessary to use the DNA sequences to find non-coding exons. Although more challenging since they will not be as strongly conserved, it should be possible to discover Dm UTRs in the De contig for a conserved gene. It is **necessary** to do the coding exons first to ensure that the gene model exists.

2) It is problematic sometimes to locate the ends of UTRs. If the splice sites are not easily uncovered, search nearby sequences to find most likely replacement sites.

3) ClustalW analysis is used to extend matches found in Blastn and explore potential end points of a UTR. Again, look for probable splice sites if the Dm sites do not match and extend accordingly.

4) Use upstream and downstream sequences contiguous with the 5' and 3' UTRs, respectively to aid in discovering exon boundaries.

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