Lo, Louis
April 20, 2006
Annotation Report

Annotating Fosmid 14p24 of *D. Virilis* chromosome 4

Introduction

In the first half of Research Explorations in Genomics I finished a 38kb fragment of chromosome 4 in *D. virilis*. The finished sequence information is valuable to researchers because it allows them to conduct a variety of analyses including comparative sequence analysis between species. To further increase the usefulness of my sequence I will annotate important genes and repetitive elements that exist in my fragment. Doing so gives researchers important information about the existence and location of these features. Using both the sequence data and the annotation data it may also be possible to learn more about gene regulation and the functions of heterochromatic and euchromatic domains. Through my annotation analysis I was able to find four genes in my fosmid: CG32000, Ankyrin, Rhomboid-5 and CG4038 (Table 1). Figure 1 shows a final map of my fosmid while figure 2 shows a map with features predicted by Genscan.

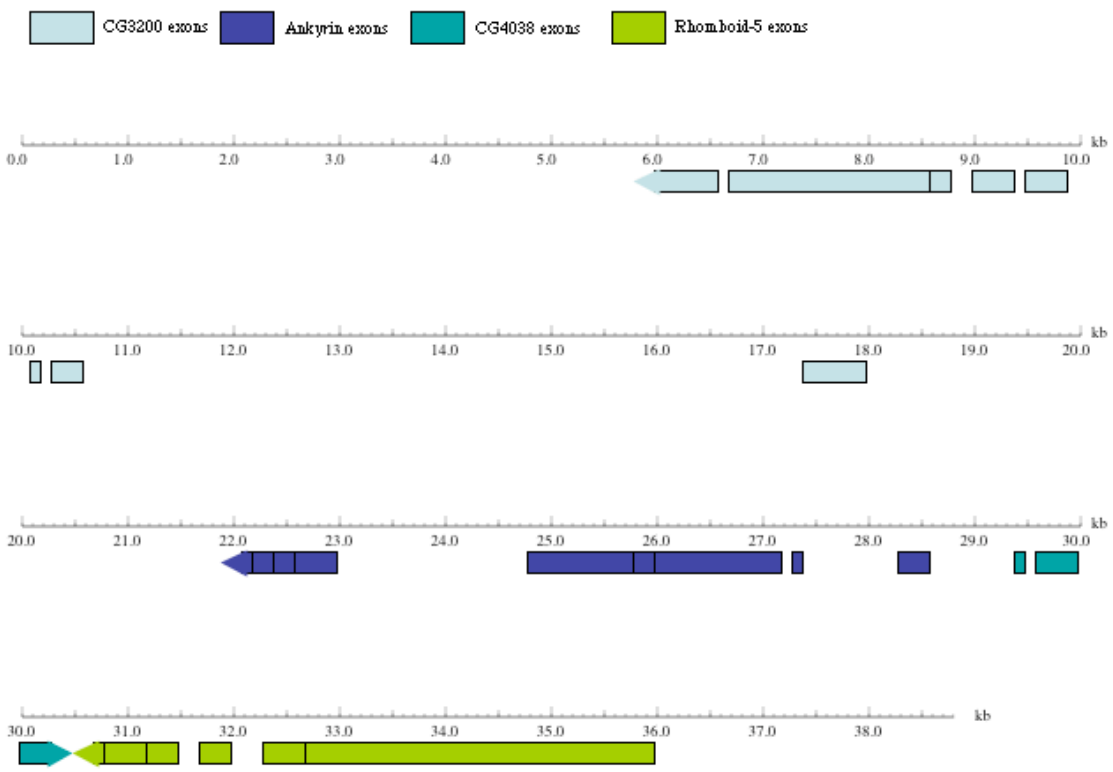

Figure 1. Fosmid 14p24 map, no repeats are shown because none were >500bps in length.

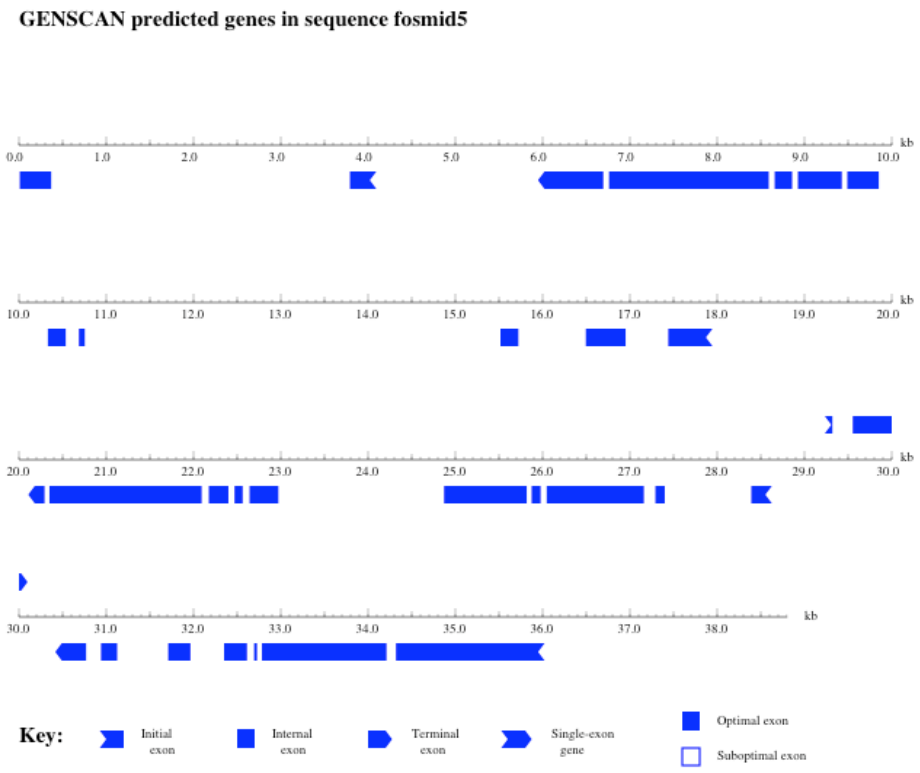

Figure 2. Genscan map output

Table 1. Summary of annotated features

Feature	Gene	Accession #
1	Mispredicted	Feature mispredicted by Genscan
2	CG3200-PG	NP_995587.1
3	Ankyrin-D	NP_787124.1
4	CG4038	NP_477043
5	Rhomboid-5	NP_995679

Genes

Feature 5

For feature 5 Genscan predicted a gene with seven exons spanning from 30347 to 35948 in reverse orientation. Using BlastP to search the non-redundant database with Genscan's putative amino acid sequence, I found matches to the rhomboid-5 gene on chromosome 2L of *D. melanogaster* (Figure 1). The integral membrane protein rhomboid-5 has dolichyl-diphospho-oligosaccharide-protein glycosyltransferase activity. It is important for signal transduction.

Figure 3. Portion of Blastp between Genscan prediction and nr database

The Blast results show that there is 64% sequence identity and an e-value of 0 between the *D. melanogaster* rhomboid-5 gene and my fosmid. Because of the evolutionary distance between *D. melanogaster* and *D. virilis* this is around what we would expect for a homolog – especially considering that there is a large region of conservation highlighted in Figure 3. With this information I decided that it was probable that there was a rhomboid-5 homolog in *D. virilis*. To confirm this I tried to locate regions in my *D. virilis* fragment that were homologous to the rhomboid-5 exons. Using the mRNA sequence of the *D. melanogaster* rhomboid-5 gene I was able to find each exon in my sequence individually. I began by converting the mRNA sequence of each exon to amino acid sequences. I than was able to use a Bl2seq with Blastx to find where each of the exons had homology to my fragment. This provided me with a finer estimate of exon/intron boundaries than Genscan does (Figure 4).

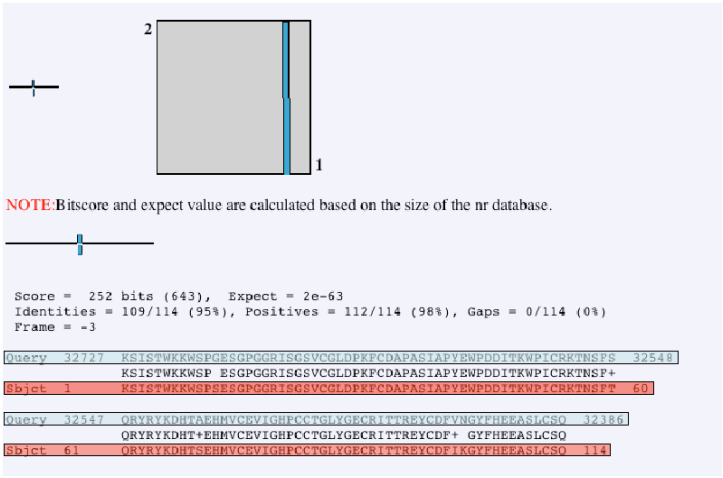

Figure 4. Bl2seq Blastx between fragment sequence (teal) and rhomboid-5 exon 2 sequence(red)

After I found exon/intron boundary estimates using Bl2seq, I then predicted exact boundaries using my estimates and the UCSC genome browser. Looking at the nucleotide sequence I found the areas where the introns began and ended. Because the homolog is in reverse orientation it's intron boundaries are marked by GA and TG (Figure 5). I used this procedure to determine all of the boundaries of the exons and introns.

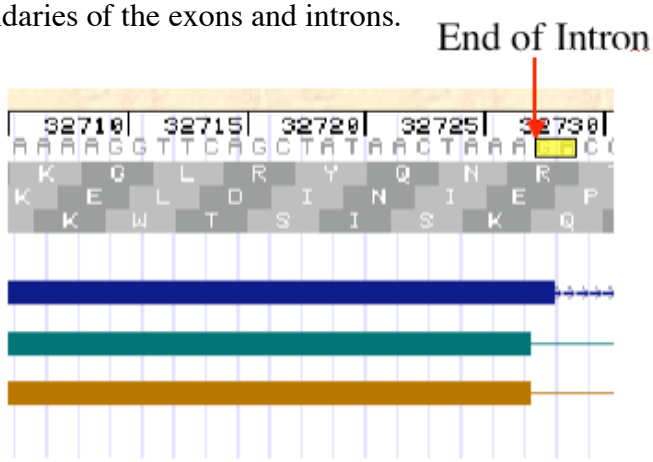

Figure 5. UCSC genome browser showing example intron boundary. Because the homolog is in reverse orientation GA marks the end of the intron

To check the accuracy of my exon boundaries I extracted the exons from my fosmid based on the boundaries I found using the UCSC genome browser. I combined all of the exons and converted them from nucleotide to amino acid sequence. If I have properly determined the boundaries the amino acid sequence should start with methionine and have a stop codon at the end. My first attempt revealed that I had not accurately determined the exon boundaries because there were many interspersed stop codons in the translated sequence (Figure 6).

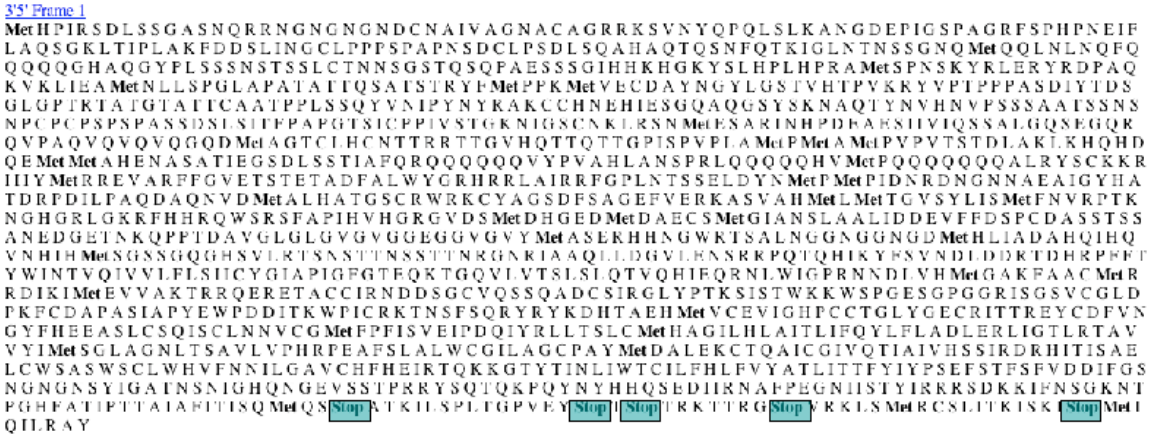

Figure 6. Initial translation – many stop codons(green) within sequence

The stop codons were all concentrated towards the end of the gene; I determined that I had predicted the last exon to be too long. I decided that the first stop codon represented the real stop codon in the gene because the others would leave stop codons within the sequence. I adjusted the exon boundary accordingly which resulted in a translation that had only one stop codon (Figure 7).

3'5' Frame 1

Met H P I R S D L S S G A S N Q R R N G N G N G N D C N A I V A G N A C A G R R K S V N Y Q P Q L S L K A N G D E P I G S P A G R F S P H P N E I F L A Q S G K L
T I P L A K F D D S L I N G C L P P P S P A P N S D C L P S D L S Q A H A Q T Q S N F Q T K I G L N T N S S G N Q Met Q Q L N L N Q F Q Q Q Q Q G H A Q G Y P L S S
S N S T S S L C T N N S G S T Q S Q P A E S S S G I H H K H G K Y S L H P L H P R A Met S P N S K Y R L E R Y R D P A Q K V K L I E A Met N L L S P G L A P A T A T
T Q S A T S T R Y F Met P P K Met V E C D A Y N G Y L G S T V H T P V K R Y V P T P P P A S D I Y T D S G L G P T R T A T G T A T T C A A T P P L S S Q Y V N I P Y
N Y R A K C C H N E H I E S G Q A Q G S Y S K N A Q T Y N V H N V P S S S A A T S S N S N P C P C P S P S P A S S D S L S I T F P A P G T S I C P P I V S T G K N I G S
C N K L R S N Met E S A R I N H P D E A E S I I V I Q S S A L G Q S E G Q R Q V P A Q V Q V Q V Q G Q D Met A G T C L H C N T T R R T T G V H Q T T Q T T G P I S P
V P L A Met P Met A Met P V P V T S T D L A K L K H Q H D Q E Met Met A H E N A S A T I E G S D L S S T I A F Q R Q Q Q Q Q Q V Y P V A H L A N S P R L Q Q Q Q
Q H V Met P Q Q Q Q Q Q A L R Y S C K K R I I I Y Met R R E V A R F F G V E T S T E T A D F A L W Y G R H R R L A I R R F G P L N T S S E L D Y N Met P Met P I D
N R D N G N N A E A I G Y H A T D R P D I L P A Q D A Q N V D Met A L H A T G S C R W R K C Y A G S D F S A G E F V E R K A S V A H Met L Met T G V S Y L I S
Met F N V R P T K N G H G R L G K R F H H R Q W S R S F A P I H V H G R G V D S Met D H G E D Met D A E C S Met G I A N S L A A L I D D E V F F D S P C D A S S T
S S A N E D G E T N K Q P P T D A V G L G L G V G V G G E G G V G V Y Met A S E R H H N G W R T S A L N G G N G G N G D Met H L I A D A H Q I H Q V N H I H
Met S G S S G Q G H S V L R T S N S T T N S S T T N R G N R I A A Q L L D G V L E N S R R P Q T Q H I K Y F S V N D L D D R T D H R P F F T Y W I N T V Q I V V L F
L S I I C Y G I A P I G F G T E Q K T G Q V L V T S L S L Q T V Q H I E Q R N L W I G P R N N D L V H Met G A K F A A C Met R R D I K I Met E V V A K T R R Q E R E
T A C C I R N D D S G C V Q S S Q A D C S I R G L Y P T K S I S T W K K W S P G E S G P G G R I S G S V C G L D P K F C D A P A S I A P Y E W P D D I T K W P I C R K
T N S F S Q R Y R Y K D H T A E H Met V C E V I G H P C C T G L Y G E C R I T T R E Y C D F V N G Y F H E E A S L C S Q I S C L N N V C G Met F P F I S V E I P D Q I
Y R L L T S L C Met H A G I L H L A I T L I F Q Y L F L A D L E R L I G T L R T A V V Y I Met S G L A G N L T S A V L V P H R P E A F S L A L W C G I L A G C P A Y
Met D A L E K C T Q A I C G I V Q T I A I V H S S I R D R H I T I S A E L C W S A S W S C L W H V F N N I L G A V C H F H E I R T Q K K G T Y T I N L I W T C I L F H L
F V Y A T L I T T F Y I Y P S E F S T F S F V D D I F G S N G N G N S Y I G A T N S N I G H Q N G E V S S T P R R Y S Q T Q K P Q Y N Y H H Q S E D I I R N A F P E G N
I I S T Y I R R R S D K K I F N S G K N T P G H F A T I P T T A I A F I T I S Q Met Q S Stop

Figure 7. Completed translation

This feature is a homolog of the rhomboid-5 gene found in *D. melanogaster*. This is supported by the low e-value found using Blastp, the individual exon similarity and the lack of stop codons in the exonic regions of *D. virilis*.

Feature 1

For feature 1 Genscan predicted a two-exon gene spanning from 9 to 5129. Using BlastP to search the non-redundant database with Genscan's putative amino acid sequence I found identity to CG33978 with an e value of 10^-43 (Figure 8).

```
>gi|85724738|ref|NP_001033801.1|  CG33978-PA [Drosophila melanogaster]
 gi|84795108|gb|ABC65827.1|  CG33978-PA [Drosophila melanogaster]
Length=4056

 Score =  178 bits (451),  Expect = 1e-43
 Identities = 89/130 (68%), Positives = 107/130 (82%), Gaps = 2/130 (1%)

Query  68   PPLGSAQAHDLTTVLLVNNDGGGRLGDSHGRFLSVRPEIGLLTSTARTFIQEGVTTEYAT  127
            P L +    D+TTVLLVNN+ G  +GD HGRFL+VRPEIG+L STARTFIQ+G+TTE+AT
Sbjct  100  PTLPAISGDDITTVLLVNNESG-HMGDFHGRFLTVRPEIGVLKSTARTFIQDGITTEFAT  158

Query  128  QVVGTTLNNGRLYAQYLKKSSRVLFENRQMSPSVVTSWVGEGDP-QTRSYLQSHNDLLDA  186
            ++VGTTLNNGRLYAQYLKKSSRVL+EN  +SPSVVTSWVGE D  QT   LQSHNDL +
Sbjct  159  KIVGTTLNNGRLYAQYLKKSSRVLYENENISPSVVTSWVGEEDSLQTPLVLQSHNDLFNI  218

Query  187  EAPDWREIDD  196
            +  +W++IDD
Sbjct  219  DDSNWQDIDD  228
```

Figure 8. BlastP results of Genscan's putative amino acid sequence against the non-redundant database

The result showed identity in only 197 amino acids; CG33978 is 4056 amino acids long and has ten exons. Because the feature is at the end of fosmid 14p24 it is possible that part of the gene was not included in my fosmid. To investigate I used Bl2seq Blastx to search for amino acid sequence of each *D. melanogaster* exon in my fosmid sequence. This is the same procedure that I used to find exon boundaries for feature 5. I was unable to find significant matches for any of the exons. If the gene was not fully included in my fosmid I would still expect exons from one end of the gene to be present in my fosmid,

but I could not find any using Bl2seq. I then used Blat to look for Genscan's putative amino acid sequence in the *D. melanogaster* genome to visualize the match obtained using Blastp (Figure 9).

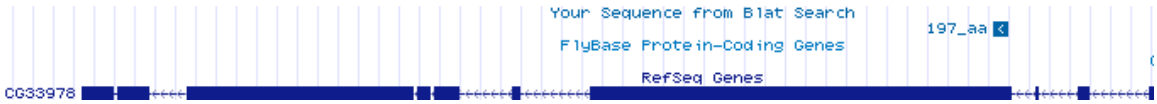

Figure 9. Blat showing region of identity match

The match was to a region in exon 4 of the *D. melanogaster* gene. Because of the e-value from the initial BlastP search I felt that it was unlikely that this 197 amino acid region could have coincidentally shared identity with the *D. melanogaster* gene, but I was unable to find any CG33978 exons in my fosmid. I then searched using nucleotide sequences from both the introns and exons of CG33978 hoping to find similarity in the untranslated regions. I used the nucleotide sequence of CG33978 from both ends of the gene to find a matching region with BlastN; this should allow me to determine which end of the gene was present in my fosmid. Again I found no significant results. The evidence suggests that feature 1 is neither a gene nor a pseudogene. Rather it is a Genscan mispredicted feature that has a small region of identity to CG33978. I am unsure why this small region has similarity to CG33978.

Feature 2

For feature 2 Genscan predicted 10 exons spanning from 5303 to 19146. Using BlastP to search the non-redundant database with Genscan's putative amino acid sequence I found matches to various isoforms of CG32000. To visualize all of the isoforms I performed a Blat search using the putative amino acid sequence with the *D. melanogaster* genome (Figure 10).

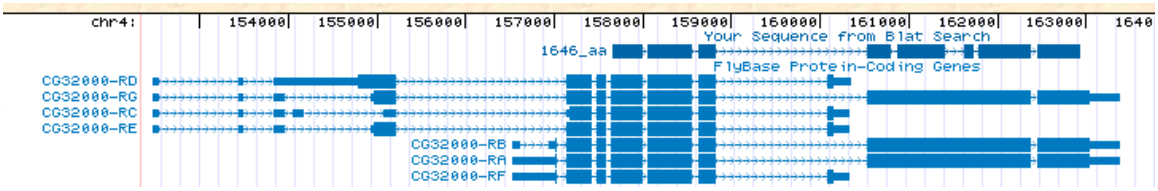

Figure 10. Blat search showing CG32000 isoforms

Analyzing the visualization I chose to use isoform RG because it was both the longest and most complete. Using these criteria to choose the isoform gives me the greatest possibility of finding all CG32000 regions within my fosmid. CG32000-RG has eight coding exons.

Using the same process as in feature 5, I determined boundaries of each individual exon of CG32000 in my fosmid using Bl2seq BlastP for estimates based on amino acid identity, and the UCSC genome browser for exact boundary positions. I was able to find boundaries in my fosmid for 6 of the 7 exons but was having trouble placing my first exon. Using Bl2seq BlastP I found that exon 1 had little identity to my fosmid (Figure 11).

```
 Score = 20.8 bits (42),  Expect = 9660060
 Identities = 14/64 (21%), Positives = 29/64 (45%), Gaps = 5/64 (7%)
 Frame = -2

Query  16708  SILSTTDQSLPTN----LNNNNTATRSSLNTPAKSTSSPPHHHRSNERHPKHKDSSSIVD  16541
               S+       +P+N    + NNNT       T A+S      ++R ++     ++ +S+++
Sbjct  8      SVQPKKSDKVPSNKIKKVENNNTLVNGCSKTSARSVPL-LKYNRPDQDGDSEENITSVLE  66

Query  16540  PAED  16529
               P  D
Sbjct  67     PNVD  70
```

Figure 11. Bl2seq BlastP of exon 1 with my fosmid

This match also does not begin with a methionine, which acts as the first amino acid of any protein. Because all of the other exons were located in the middle of my fosmid I thought that it would be unlikely that the first exon was not in my fosmid. It may have diverged too much for Bl2seq to recognize sequence similarity. To find the boundaries for the first exon I used Genscans prediction as my initial estimate. It predicted the first exon to be from 17,440 to 17,871. Using the UCSC genome browser I found that 17,871 was the first base of a start codon. I chose this to be the start codon of my protein because it starts a large open reading frame that ends around 17,430. It is unlikely that a large open reading frame of this size occurred by chance. Although I am fairly confident that Genscan had correctly predicted the location of the start codon, I had no evidence as to where the end of the exon was. The genome browser did not mark any splice acceptor sites in the entire open reading frame. Also, without EST evidence I could not find any evidence that supported a particular boundary. Because of the lack of evidence I decided to choose the boundary that created the longest exon (Figure 12).

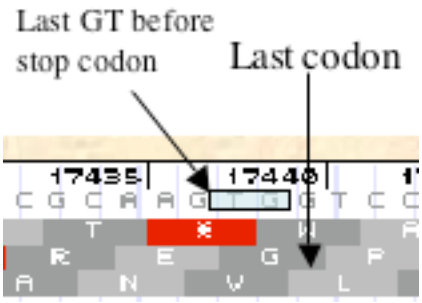

Figure 12. Exon 1 end

Although my selection of the exon end is rather arbitrary, it is my best possible guess. I had ruled out the possibility that this was a pseudogene because of the strong sequence similarity of the other seven exons in my fosmid. Because of this I was forced to find boundaries for the first exon. I decided to choose the last possible exon boundary in the open reading frame. Doing so ensured that no amino acids that are actually part of the exon were left out. I will be able to confirm the validity of my exon 1 boundaries when EST data for *D. virilis* becomes available.

After I had determined the boundaries for all eight exons I put the sequences together and translated, just like I did for feature 5. I found that the amino acid sequence started with a start codon, ended with a stop codon and had no interspersed stop codons. This along with the mapping of all exons and initial BlastP results provides good evidence that this feature is a homolog of CG32000.

Feature 4

For feature 4, Genscan predicted two exons spanning from 29049 to 30286. Using BlastP to search the non-redundant database with Genscan's putative amino acid sequence I found matches to CG4038, the "H/ACA ribonucleprotein complex subunit 1-like protein (GCR 101 snRNP)" gene on chromosome 2R of *D. melanogaster* (Figure 13). CG4038 is involved in the synthesis of rRNAs by specifying sites of uridine to pseudouridine conversion. The BlastP result was returned with an e-value of 10^-44 and had 88% identity.

```
>gi|17137002|ref|NP_477043.1|  CG4038-PA [Drosophila melanogaster]
 gi|7291241|gb|AAF46672.1|  CG4038-PA [Drosophila melanogaster]
 gi|68565897|sp|Q7KVQ0|NOLA1_DROME  H/ACA ribonucleoprotein complex subunit 1-like protein (GCR 101
snRNP)
Length=237

 Score =  181 bits (458),  Expect = 1e-44
 Identities = 84/95 (88%), Positives = 91/95 (95%), Gaps = 0/95 (0%)

Query  35   FDQGPPERVIALGNFSYACQNDLVCKVDIDDVPYFNAPIFLENKEQIGKIDEIFGTVRDY  94
            FD GPPERVI LGN+ Y+CQNDLVCKVDI DVPYFNAPIFLENKEQ+GKIDEIFGTVRDY
Sbjct  61   FDTGPPERVIPLGNYVYSCQNDLVCKVDIQDVPYFNAPIFLENKEQVGKIDEIFGTVRDY  120

Query  95   SVSIKLSDNIFANSFKPNQQLFIDPGKLLPISRFL  129
            SVSIKLSDN++ANSFKPNQ+LFIDPGKLLPI+RFL
Sbjct  121  SVSIKLSDNVYANSFKPNQKLFIDPGKLLPIARFL  155
```

Figure 13. Blastp match between Genscan prediction and nr database

The match is for only 95 amino acids while the protein in *D. melanogaster* is 237 amino acids long. Also the gene in *D. melanogaster* has three exons while Genscan only predicts two for this feature. To investigate this feature further, I ran Bl2seq with BlastP to search for each individual *D. melanogaster* exon in my fosmid. After raising the expected value to 10,000, I was able to find sequence similar to exon 1 to my fosmid but only for 13 amino acids; exon 1 in the *D. melanogaster* gene is 63 amino acids in length (Figure 14).

```
 Score = 25.0 bits (53),  Expect(2) = 3e-04
 Identities = 9/13 (69%), Positives = 10/13 (76%), Gaps = 0/13 (0%)
 Frame = +1

Query  29452  MAFGRPRGSGGRG  29490
              M FG+PRG GG G
Sbjct  1      MGFGKPRGGGGGG  13
```

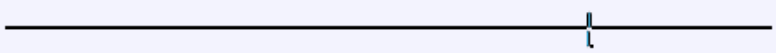

Figure 14. Bl2seq Blastx of *D. melanogaster* exon 1 versus fosmid 5

When comparing the identities for the two exon searches we see that exon 2 has much greater sequence identity between *D. melanogaster* and *D. virilis* than exon 1 does. The high level of identity for exon 2 suggests that the feature in *D. virilis* is in fact a gene, even though exon 1 exhibits such low identity. This is because the identity implies conservation – through evolutionary time we would only expect regions of DNA that were of importance to be conserved. Exon 2 would only have been conserved if the protein were functional which requires the expression of all exons. Exon 1 may not be as important for the function of the protein so its sequence may not have been so strongly conserved. The third exon was found in this same way with good identity between the species.

After I had estimates of exon locations I used the UCSC genome browser to find specific exon boundaries. When searching for these boundaries I found that the open reading frame that exon 2 was in was very much longer than the estimate based on identity to the *D. melanogaster* gene. Examining the amino acid sequence of the *D. melanogaster* gene shows G-rich regions that flank the 2nd exon (Figure 15).

MGFGKPRGGGGGGGRGFGGGGGGGGRGFGGGGGGRGGGGGRGGGGGFGRGGGGRGGGRG
AFDTGPPERVIPLGNYVYSCQNDLVCKVDIQDVPYFNAPIFLENKEQVGKIDEIFGTVR
DYSVSIKLSDNVYANSFKPNQKLFIDPGKLLPIARFLPKPPQPKGAKKAFTNNRGGGGG
GGGFGGRGGGRGGGGRGGGGGGRGGGGFRGGAGRNGGGGGGGGFNRGRGGGGGGGGGRGR
W

Figure 15. *D. melanogaster* CG4038 amino acid sequence. Colors indicate different exons

Examining the amino acid sequence predicted using the UCSC genome browser shows that the G-rich region in exon 1 of *D. virilis* is missing while the extended region of exon 2 is all G-rich. This suggests that the intron between exon 1 and exon 2 has "moved" in this gene. The G-rich regions flanking the introns might have made it easier for the intron to shift relative to the two exons. Further evidence to support intron movement comes from the UCSC genome browser. The boundaries that I identified for the end of exon 1 and the beginning of exon 2 are marked by the genome browser as a *medium confidence splice acceptor* and *high confidence splice donor* respectively (Figure 16).

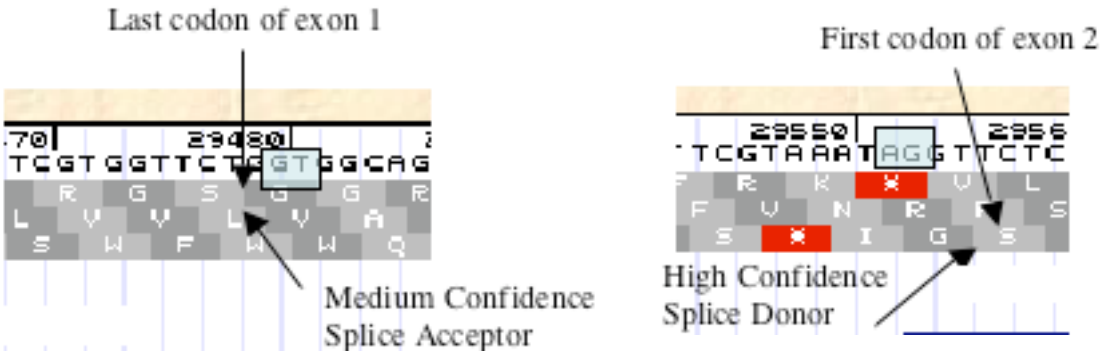

Figure 16. UCSC genome browser view of exon boundaries

As a final check for intron movement I used Bl2seq BlastP to compare the *D. virilis* exons with the *D. melanogaster* gene. I was able to find good amino acid sequence similarity between the two. With the presence of G-rich regions flanking the first intron

in *D. melanogaster*, strong confidence in splice sites proposed for *D. virilis* and sequence similarity between *D. melanogaster* and *D. virilis* exons there is strong evidence to support my hypothesis of intron movement (Figure 17).

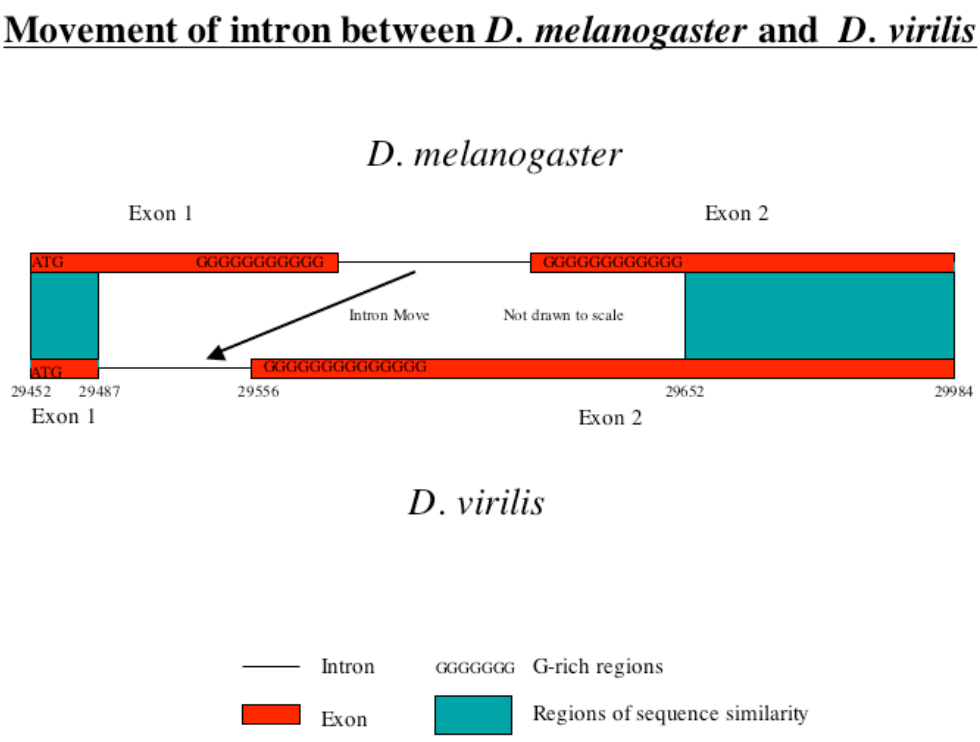

Figure 17. Model for intron movement

I was able to find strong sequence similarity between feature 4 and CG4038. I was also able to individually map all three exons of CG4038 to feature 4. Translating the exons I predicted for feature 4 shows an amino acid sequence starting with a start codon, ending with a stop codon and having no stop codons within the sequence. This strongly suggests that feature 4 is a homolog of the CG4038 gene.

Feature 3

For feature 3 Genscan predicted 10 exons spanning from 20097 to 28670. Using BlastP to search the non-redundant databse with Genscan's putative amino acid sequence I found matches to the gene Ankyrin in *D. melanogaster*. This gene had various isoforms, so I used the UCSC genome browser to visualize the isoforms (Figure 18).

Figure 18. Visualization of ankyrin isoforms. Thick boxes are coding regions

The visualization shows that all of the isoforms have the same coding exons. Because of this I decided to use the isoform that had the highest score from the initial BlastP search. Ankyrin-D has 9 coding exons and is 1549 amino acids in length. To confirm that ankyrin-D exists in my fosmid I found the boundaries of each exon using the same process as in feature 5. I was able to easily find boundaries for all of the exons except for exon 1. The similarity for exon 1 between *D. melanogaster* and *D. virilis* was weak, as was the case for exon 1 for feature 4. In fact, I used the same procedure in finding the exon boundary of exon 1 in feature 3 as I did in feature 4.

After determining the boundaries for the exons I combined the sequences from all of the exons and then translated. I found that there was a methionine to start the amino acid sequence, a stop codon at the end and no interspersed stop codons. This evidence along with the BlastP results and the presence of all exons in my fosmid suggests that this feature is a homolog of the ankyrin gene.

ClustalW analysis

I conducted a Clustal analysis of feature 4 (CG4038) to find how well the gene has been conserved through evolutionary time. CG4038 is involved in the synthesis of rRNAs by specifying sites of uridine to pseudouridine conversion. Because it is involved in the synthesis of rRNAs, which are important for protein synthesis, I hypothesized that this gene would be well conserved between distantly related species. To determine how quickly the gene was mutating, I used the nucleotide sequence of feature 4 and BlastN to search the non-redundant database. I found that the gene was mutating slowly because it had strong identity matches to a variety of organisms that are distantly related to *D. virilis*; corroborating my hypothesis. I compared the *D. virilis* gene to distantly related species: humans, zebra fish and mice. Using the BlastN results I was able to obtain the amino acid sequences of mouse, human and zebra fish nucleolar protein family A, member 1. With the amino acid sequences from the four species, I created a Clustal alignment, which aligns the amino acid sequences based on local identity (Figure 19).

Figure 19. Clustal alignment – region of strong conservation highlighted

The Clustal alignment shows strong conservation across all four species. The level of conservation, however, is not consistent throughout the entire alignment. The first row shows many gaps across the species. Human and mouse show good alignment, but when compared to zebra fish and *D. virilis*, gaps appear. This shows that this region is not under strong selection pressure. This strengthens the possibility of intron movement between *D. virilis* and *D. melanogaster* as described by my analysis of feature 4 – weaker selection pressure allows for more divergence. Although the nucleotide identity of this region may not be important for the protein to function properly, the region is still important because it contains the start codon for the protein. The highlighted region in Figure 1 shows strong conservation across all four species. The strong selection pressure in this region suggests that it likely contains the active site for the uridine to pseudouridine conversion.

Again using Clustal, I searched for promoter regions of the gene that were conserved through evolutionary time. Because promoter regions are under less selective pressure than coding regions, I chose species that were more closely related to *D. virilis*. To do the comparison I found the location of the gene in *D. melanogaster, virilis, pseudoobscura* and *yakuba*. I then extracted 1000 bases upstream from the start codon of each homolog for the alignment. Using the 1kb fragments and Clustal I created an alignment comparing upstream regions (Figure 20).

```
D.pseudoobscura      ------GCTTTTTATACAACAAAAC-GAACCTTAAGTG--CAAAATGAAAAAAATAGATG 51
D.melanogaster       CGATCCGTCATCAACGTTTCAGGGGAGAGGATCGATCA--CAAAGGGGCTACGGAAAGTA 58
D.yakuba             ------------AATATAGCTGTGTGAGCTATATATAA--TGCATCCGATGATGTATGAG 46
D.virilis            -----CTGTGCTGATTTATTTGTATCTGCGTTCGTTTAATTGCTGTTGCTGTTTCGTGGC 55
                                 *                   *

D.pseudoobscura      ACTTATTGATGCATTATCATAAACTTGACACCTT--CATATATCAGGGTAATTTA---AC 106
D.melanogaster       TCTGATGAAAATTATACAAGAAATTTTA-ATTTA--TTTGTATTTTGAAAGTGCA---GA 112
D.yakuba             ACTCTC-----TCTTACATTGTGTTTCATGTTCT--CTTA----GTGAATATTTA---GT 92
D.virilis            CCTGGCTTTTCGTTTGTATCTCGTTTAGTGTCTCGCCTAGGGTCATTTTTTTTTAAATGA 115
                      **          *         **                        *  *

D.pseudoobscura      TTTTGCTTTCTTT-CTAAAAAATTTTATTTTAACCCAACCTCAAAT---TAAGCTCGGAA 162
D.melanogaster       AATACTTTGAGCC-CACAATGACGTCAGTGGGCATAAAGATAATATA--TGAGAAGGCAA 169
D.yakuba             GCTTATTAAATAAGCACCAGGAATTTGTGCTAGTTAAACGATGTACCCATTTGGTTGGGT 152
D.virilis            ATTCCTGAAAATAAAGCACTGACTAATTACTAAGTTGACCACAAATGGATATCCGGAGCA 175
                       *                   *                 *     *    *
```

Figure 20. Portion of Clustal alignment between *Drosophila* species

Although only a small region of the 1kb alignment is shown it is representative of the entire alignment. There is little conservation in the 1kb upstream region between the four *Drosophila* species. This shows that this region has not been under strong selection pressure. The results of this alignment show that there are no conserved promoter regions between the four species. If there were conserved promoters in this region we would see areas of strong nucleotide identity between the four species. This does not mean that there are no promoters in the upstream regions, merely that there exist no homologous promoters with identical positions in the alignment.

Repeat analysis

Table 2 lists all of the repeats found within fosmid 5. Repeat Masker recognized 30 separate repeats in three different families within fosmid 5. There were no repeats that were 500 bps or longer. Collectively they account for 4230 bps (10.90%) (Table 3).

Table 2. Repeats found by Repeat Masker in fosmid 5

Begin	End	Repeat	Repeat Family	Repeat ID
989	1038	AT_rich	Low complexity	1
1377	1428	(CAGT)n	Simple repeat	2
4520	4553	(CA)n	Simple repeat	3
4797	4823	(CA)n	Simple repeat	4
10199	1227	AT_rich	Low complexity	5
11803	11854	(TA)n	Simple repeat	6
12092	12155	(CATA)n	Simple repeat	7
12313	12350	(AACTG)n	Simple repeat	8
12513	12534	AT_rich	Low complexity	9
12722	12743	AT_rich	Low complexity	10
13437	14739	Dvir.14.34.centroid	DNA	11
15050	15071	AT_rich	Low complexity	12
15119	15190	Dvir.16.2.centroid	DNA	13
15157	15220	PENELOPE	LINE	14

15218	15266	Dvir.16.2.centroid	DNA	13
15790	15817	AT_rich	Low complexity	15
15822	16138	Dvir.11.33.centroid	TRF	16
16172	16241	Dvir.11.33.centroid	TRF	16
19580	19609	AT_rich	Low complexity	17
22609	22632	(TA)n	Simple repeat	18
22995	23018	AT_rich	Low complexity	19
23022	23110	Dvir.16.2.centroid	DNA	20
23063	23117	PENELOPE	LINE	21
23118	23500	Dvir.16.2.centroid	DNA	20
23697	23717	AT_rich	Low complexity	22
24076	24147	Dvir.16.2.centroid	DNA	23
24152	24551	Dvir.16.2.centroid	DNA	23
24685	24726	Dvir.11.23.centroid	TRF	24
27166	27187	AT_rich	Low complexity	25
30089	30246	(CGG)n	Simple repeat	26
34233	34281	(CTG)n	Simple repeat	27
36229	36637	Dvir.16.2.centroid	DNA	28
36642	36706	Dvir.16.2.centroid	DNA	28
37090	37121	AT_rich	Low complexity	29
38680	38804	Dvir.16.2.centroid	DNA	30

Table 2. Summary of repeats in fosmid 5

Repeat	Number of Elements	Percentage of fosmid 5
LINEs	2	0.10%
DNA elements	6	7.64%
Simple repeats	9	1.28%
Low Complexity	11	0.78%

Interestingly a region that I had difficulty working on during the finishing of fosmid 5 is found by Repeat Masker to be within one of the repeats. The region was found between 13650 and 14005 of fosmid 5, which corresponds to repeat 11. This region caused difficulty because Consed$_1$ thought that the region had been assembled into the wrong position. Upon analyzing the region's sequence I determined that the region did belong at that position, given the possibility of repeats. Repeat Masker confirms that this region does contain repeated elements – verifying my analysis of this region during finishing.

I also analyzed fosmid 5 for novel repeats, ones that were not found through the Repeat Masker analysis. To look for candidates I used BlastN to compare the nucleotide sequence of fosmid 5 to a database of all *D. virilis* fosmids. Then using Herne, I was able to visualize the identity matches between fosmid 5 and the rest of the *D. virilis* fosmids. Novel repeat candidates would be regions of my fosmid that match to other fosmids and are 100 bps or longer. There were no regions greater than 100 bps, so I analyzed a region that was 96 bps in length (Figure 21).

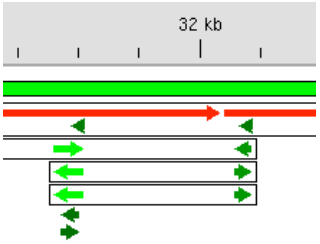

Figure 21. Novel repeat candidate region. Boxed arrows are hits.

It is found between 31521 and 31617 of my fosmid. To determine if this region was part of a repeat that had already been characterized by Repeat Masker I checked to see if the candidate region was at either end of an already characterized repeat. I found that the candidate region was not part of a Repeat Masker repeat. With BlastN, I then searched the non-redundant database with the novel repeat candidate's nucleotide sequence to see if the candidate was part of a protein. This search revealed that the candidate was not part of a protein. Not being part of an already characterized repeat or a protein suggests that the candidate is in fact a novel repeat. To verify this I extracted the nucleotide sequence of the candidate region and used Blat to determine where it existed in *D. virilis*. It was found in only one other location – fosmid 15e16, which lies adjacent to fosmid 14p24 (Figure 22).

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	97	1	97	97	100.0%	fosmid5	+	31521	31617	97
browser details	YourSeq	47	1	57	97	91.3%	fosmid10	+	30768	30824	57

Figure 22. Blat results showing areas in *D. virilis* that match to novel repeat candidate

Because the two fosmids are adjacent I concluded that the candidate was *not* a novel repeat. Rather the identity match was due to sequence overlap between fosmid 15e16 and 14p24.

Synteny analysis

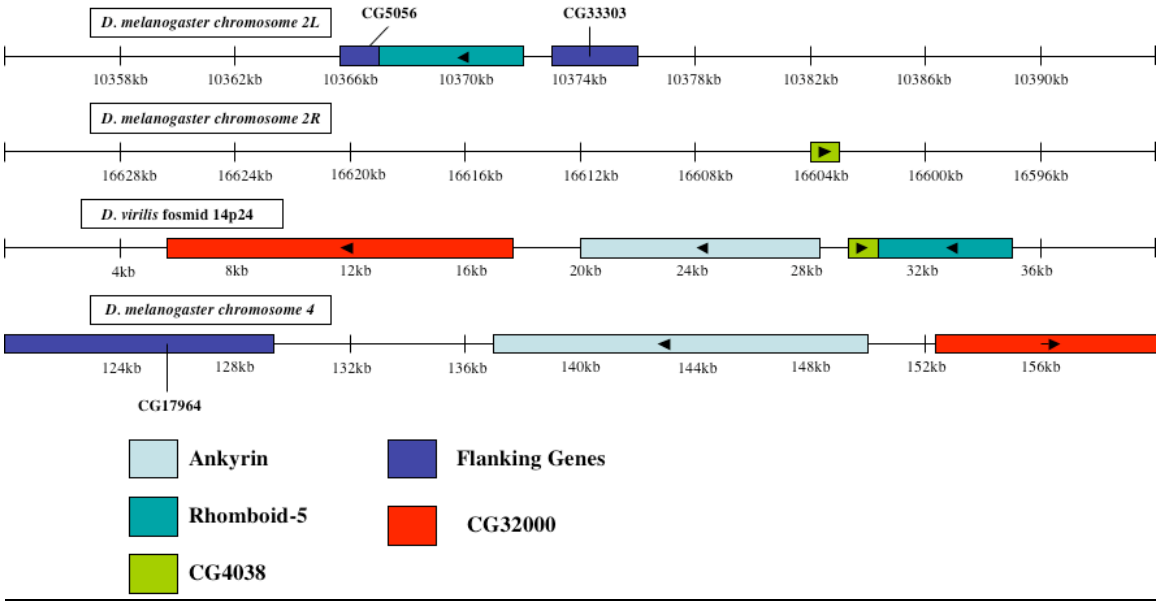

Figure 23. Diagram of synteny between *D. virilis* and *D. melanogaster*

Figure 23 shows that my fosmid does not have good synteny to *D. melanogaster* chromosome four. In fact the four genes found in my fosmid are found in three different chromosomes of *D. melanogaster*: chromosomes 2L, 2R and 4.

The rhomboid-5 gene is in chromosome 2L of *D. melanogaster* and in chromosome 4 of *D. virilis*. Examination of these regions shows that in *D. melanogaster* CG5056 and CG33303 flank the rhomboid-5 gene. These two genes do not flank the rhomboid-5 gene in *D. virilis*. This shows that rhomboid-5 is not syntenic between *D. melanogaster* and *D. virilis*. The gene frequency for the region in 2L is about 1 gene in 13kb while it is 1 gene in 10kb for my fosmid. The fact that none of the flanking genes in 2L are flanking rhomboid-5 in 14p24, and that the gene is found only once in both *Drosophila* genomes, suggests that the gene may be on different chromosomes because of DNA transposons. A transposition event may have cut off the flanking genes, leaving the flanking regions of rhomboid-5 free to differ between *D. melanogaster* and *D. virilis*.

CG4038 is also a gene that did not remain syntenic between *D. melanogaster* and *D. virilis*. In *D. melanogaster* it exists on chromosome 2R while it is in chromosome 4 of *D. virilis*. Analysis of the region surrounding the gene shows that CG4038 in *D. melanogaster* has no closely flanking genes, while the gene in *D. virilis* is flanked by ankyrin and rhomboid-5. The gene frequency for the region in 2R is 1 gene in 40kb while it is 1 gene in 13kb 14p24. Because the gene does not share similar flanking genes in the two species and is found only once in both *Drosophila* genomes, it is possible that the lack of synteny is due to DNA transposition as explained with the rhomboid-5 gene.

CG32000 and ankyrin do show synteny between *D. melanogaster* and *D. virilis*. In both species the two genes are adjacent to each other on chromosome 4, however, the relative orientation between the two genes is different in *D. melanogaster* and *D. virilis*. In *D. melanogaster* both genes are orientated in the same direction while in *D. virilis* they are orientated in opposite directions. There are no similar genes that flank the two genes between *D. melanogaster* and *D. virilis*. The difference in relative orientation of the two

genes could have been caused through chromosomal inversion. A short segment containing one of the genes and having an end in the region between the two genes could have been inverted.

Appendix

Feature 2

Exon	Begin	End
1	17871	17440
2	10534	10256
3	10155	10099
4	9820	9491
5	9428	8925
6	8859	8659
7	8591	6768
8	6693	6025

>CG32000-PG protein sequence
MNAGEYTLEKLRGRELIRFSDFHLTHWQAIGLSKKQVSPSSDNVMNSVQSATEQQEHIGSFKRRLR
KRRDGDVIDTDTDTDLDVDGAQLGSVAYIEQAAGFGLILIDLVLLQKEYVKQFYFEKRVGSGGQQ
QQKQLAGAWCLRLPKPCVALLNQDQSDQMKISGYRRSHIRSILCWICIVLTGGLLRLVLHWWRH
WYLLATCRPCSLQEAQQVLIEEDYQGNHKLYHVKSVQILHVEDFDDNKLQLPVHFASAHFKQLV
YAWDNSIKNFKKINGLDVNVPCSYYHQQRGLSVQEQLARRIVFGQNEITVPLRDVKTLLFLEVLNP
FYVFQIFSVVLWFTYDYYYYACVILLMSIFGISMSIVQTKKNQDVLQKTVLNTGNAWVVNAKGVS
VELSTKMLVPGDIIEIPSSGCTMQCDAVLLSGNCILDESMLTGESVPVTKTPLPMKRDVIFDKKEHA
RHTLFCGTKVIQTRYIGSKKVLAFVINTGNITAKGGLIRSILYPPPVDYKFEQDSYKFIQFLALIACIG
FIYTLVTKILRGTDAVKIVVESLDLITIVVPPALPAAMTVGRFYAQKRLKANNIFCISPRSINVAGSID
CCCFDKTGTLTEDGLDMWGVVPKSSTNQFQIPLKDVKRLPYDHFLFGMVTCHSITVMNGMMMG
DPLDLKMFESTGWILEDSNNIPENEKYGLIYPTILRQPNQEEINSSLAFTHHEAVITEAPPPVFRQSSV
DDLLANMGLLKSKTNFDHGIVREFPFTSNLQRMSVITRCLSAQGFNVYCKGSPEMLQQLCHPQSIP
NDYTHQLSIYAKKGFRIIAVAFKTLDPKVNYTKVQRLSREEVEHNLEFLGFVILENRLKPDTTAVIN
SLNLANIRTVMITGDNILTAMSVARDCGIVSSTQAVITVHAVPVLKGFDKLNTDQNGQYELQYTLE
LGSQTCTPALNGNRTFSAAEIALDMDLSKSTNSLVNNGGSSCASSVLPNSNSLASVKTIDTWTHND
GDDVELGATVVARNDSWRRQYIFAMDGKTWQIVKDQFPQQMEIILTRGAIYARMSPEQKQSLVM
ELQNLDYYVAMCGDGANDCGALKVAHTGISLSETESSIASPFTSRNPTIAAVPNVIKEGRAALVTS
FGIFKYMAAYSLVQFVSVMILYSIDSNLTDKQYLYIDLGLISIFAFFFGKTEAYDGQLVKQVPLSSLI
SPTPLSSLVLHLAVVVIFQWFQLHQQPWFKPFGPSDEHHLGCYENYTMFAISSFQYIILAFVFSKGA
PYRKPIWSNWPLCLTLVINACIVVYLVAYPSDWIANFFQLIVPPDMSFRYIMLQYGAAAFITH
AFLEAFVVEYLVFKRFQVRREQNLKTSNRKYMRLEYNIKSFDNWPPITEVYDLHDPADQGAELQP
TYVNLSAEQNLDGQPSGFPGFFETSEHSVENRHRIHSEVSSASPRHS-

>CG32000-PG nucleotide sequence
TCAGCTGTGCCGTGGACTCGCACTGGATACCTCTGAGTGTATGCGATGTC
GGTTTTCTACGCTATGCTCAGACGTCTCGAAAAAGCCAGGAAAACCGCTG
GGCTGACCGTCCAAATTTTGTTCAGCGCTTAGGTTAACATATGTCGGCTG
TAGCTCCGCTCCTTGGTCGGCTGGATCGTGCAGATCATAGACCTCCGTTA
TGGGCGGCCAGTTGTCAAAGGACTTAATATTATACTCCAATCGCATATAT
TTACGATTTGATGTTTTTAAGTTTTGCTCCCGTCGCACCTGGAATCGCTT
AAATACCAAATACTCCACCACAAAAGCCTCGAGAAATGCATGAGTGATAA
ATGCAGCTGCTCCATATTGGAGCATAATGTAACGAAAAGACATATCCGGT
GGCACGATGAGTTGAAAGAAGTTGGCAATCCAATCACTTGGATAGGCAAC

TAGGTATACGACAATGCAGGCATTTATCACCAGGGTCAGGCATAGAGGCC
AATTCGACCAGATTGGCTTTCGGTATGGTGCTCCCTTTGAAAATACAAAT
GCAAGTATAATATACTGGAAACTCGATATGGCAAACATGGTGTAGTTCTC
ATAGCAGCCCAAATGATGCTCGTCAGATGGCCCGAAAGGCTTAAACCAAG
GCTGTTGATGCAGTTGAAACCA
CTGGAATATGACCACAACGGCTAAATGGAGCACAAGCGAGCTTAAAGGCG
TTGGCGAAATGAGGGAACTAAGAGGCACTTGTTTAACCAGCTGACCATCG
TAGGCTTCAGTTTTGCCAAAGAAGAATGCGAAAATCGATATCAAGCCAAG
GTCAATGTAGAGATATTGTTTGTCGGTAAGATTCGAGTCGATTGAGTATA
GGATCATCACGGACACAAACTGTACCAGGGAATAGGCAGCCATATACTTA
AAAATACCAAATGATGTAACCAAGGCTGCGCGACCCTCCTTGATAACATT
TGGTACGGCCGCAATAGTCGGATTACGGGAGGTGAACGGTGATGCTATTG
AGGATTCAGTCTCGCTCAACGATATTCCAGTGTGCGCTACCTTTAGGGCA
CCACAATCGTTAGCCCCATCTCCGCACATTGCTACATAATAGTCAAGATT
CTGTAGCTCCATCACCAGCGATTGCTTTTGCTCTGGCGACATGCGTGCAT
AGATAGCACCGCGCGTTAATATGATTTCCATTTGTTGAGGAAATTGGTCC
TTAACAATTTGCCATGTCTTGCCATCCATGGCGAATATATATTGACGACG
CCAGCTATCATTTCGTGCCACAACTGTCGCTCCAAGCTCCACATCATCGC
CGTCATTGTGCGTCCATGTGTCGATGGTCTTAACACTAGCCAAGCTGTTG
CTGTTGGGCAACACACTTGAGGCACAGCTCGAGCCACCATTATTTACCAA
TGAATTGGTTGACTTTGATAGATCCATATCCAGAGCGATTTCAGCAGCCG
AAAATGTTCTATTACCATTGAGAGCTGGCGTGCACGTCTGACTGCCCAGC
TCGAGAGTATACTGTAGTTCATATTGGCCATTTTGATCGGTATTTAGCTT
ATCAAAGCCTTTCAACACGGGCACCGCATGCACTGTGATTACCGCCTGCG
TGGAACTAACAATGCCGCAATCTCGAGCAACGCTCATTGCTGTCAATATG
TTATCACCCGTAATCATAACTGTGCGAATATTTGCCAAATTTAGTGAATT
TATAACAGCCGTTGTATCGGGCTTTAGGCGATTCTCAAGTATTACAAATC
CGAGGAACTCCAAATTGTGCTCGACCTCTTCGCGCGAAAGGCGTTGCACC
TTTGTGTAGTTCACCTTTGGATCAAGTGTCTTAAATGCCACGGCAATAAT
TCGGAAACCTTTTTTGGCATAAATGGAGAGCTGATGGGTATAGTCGTTCG
GTATGCTCTGTGGATGGCACAACTGCTGCAACATTTCGGGCGAGCCCTTG
CAGTAAACGTTGAAGCCTTGCGCGCTGAGGCAGCGGGTGATGACTGACAT
ACGCTGCAGGTTGGATGTAAATGGGAACTCCCTAACAATTCCATGATCGA
AATTTGTTTTTGATTTCAACAATCCCATATTGGCCAAGAGATCATCCACC
GAACTCTGCCGAAATACGGGTGGTGGCGCCTCTGTTATTACCGCTTCATG
ATGTGTGAAAGCTAAGCTTGAGTTGATTTCCTCTTGATTCGGTTGCCGTA
ATATTGTCGGATAGATCAGGCCGTACTTCTCGTTTTCGGGTATGTTGTTG
GAATCTTCCAGTATCCAGCCTGTCGATTCAAACATCTTAAGATCCAAGGG
ATCACCCATCATCATGCCGTTCATGACGGTAATTGAGTGGCATGTGACCA
TGCCAAACAGAAAGTGATCGTATGGCAGACGCTTCACATCTTTCAGTGGT
ATTTGAAACTGATTTGTGGAAGACTTTGGAACGACACCCCACATGTCTAG
GCCATCCTCCGTTAAGGTTCCAGT
CTTATCAAAGCAACAGCAATCAATACTGCCTGCCACGTTGATGGAGCGCG
GAGATATGCAAAAGATATTGTTGGCTTTTAATCGTTTCTGGGCATAAAAC
CGACCAACAGTCATAGCTGCCGGCAAGGCTGGTGGCACAACAATTGTTAT
AAGATCCAAAGATTCGACCACTATCTTAACAGCATCTGTACCACGCAATA
T
CTTGGTGACAAGCGTATATATAAAGCCAATGCACGCAATTAGCGCCAAAA
ACTGTATGAACTTGTACGAATCCTGCTCAAACTTATAGTCCACAGGCGGT
GGATAAAGAATGGAGCGTATTAGGCCACCTTTGGCAGTTATATTGCCAGT
ATTTATCACGAAGGCCAAAACTTTCTTGGATCCAATGTATCGTGTTTGTA
TTACTTTCGTGCCGCAGAATAGTGTGTGCCTAGCATGTTCCTTTTTATCA
AATATCACATCGCGCTTCATCGGAAGCGGCGTTTTAGTAACTGGCACACT
TTCGCCAGTCAGCATTGATTCGTCTAGGATGCAGTTGCCGGAAAGTAAAA
CGGCATCGCATTGCATTGTGCAGCCGGACGATGGTATTTCAATAATATCG
CCAGGCACTAACATTTTCGTAGACAATTCCACAGATACTCCCTTGGCGTT

AACGACCCATGCATTTCCCGTGTTTAAAACTGTTTTCTGCAGAACATCTT
GATT
CTTCTTTGTTTGCACAATGGACATAGATATGCCGAATATAGACATTAACA
AGATGACACAGGCATAATAGTAATAGTCATAGGTAAACCACAGGACGACA
GAAAATATTTGGAATACATAAAAGGGATTAAGAACTTCCAGGAAGAGCAG
GGTTTTCACATCCCTCAGCGGCACGGTTATCTCATTTTGCCCAAATACAA
TGCGGCGTGCCAGCTGTTCTTGCACAGATAGTCCACGTTGCTGATGGTAA
TAAGAACAAGGCACATTTACGTCTAATCCATTGATTTTCTTAAAGTTTTT
AATGCTGTTGTCCCAAGCATAAACCAGCTG
TTTGAAATGCGCGGAAGCAAAGTGAACTGGGAGCTGTAGCTTATTATCAT
CGAAATC
CTCCACATGCAGTATCTGCACGCTCTTGACATGATACAGTTTGTGATTAC
CCTGATAGTCTTCTTCTATGAGCACCTGTTGTGCCTCCTGCAAGGAGCAC
GGCCTGCATGTGGCAAGCAGATACCAATGCCGCCACCAATGTAGGACAAG
GCGCAACAGACCTCCAGTTAACACTATGCAGATCCAGCACAATATCGAAC
GGATATGGGAGCGCCGGTAACCGCTTATTTTCATTTGATCTGACTGATCT
TGGTTGAGAAGTGCGACGCATGGTTTCGG
CAGGCGCAGACACCACGCTCCGGCCAGTTGTTTTTGTTGTTGTTGTCCTC
CGCTGCCTACGCGCTTCTCGAAATAGAATTGTTTGACGTACTCTTTCTGC
AGTAGCACCAAATCGATCAGGATCAGGCCAAAGCCGGCAGCTTGTTCAAT
ATACGCCACTGATCCCAGCTGTGCGCCGTCCACGTCCAAGTCGGTGTCAG
TGTCTGTGTCGATAACGTCACCGTCTCGACGCTTACGCAACCTTCTCTTG
AATGACCCGATATGCTCCTGCTGCTCGGTTGCACTTTGAACTGAATTCAT
TACGTTATCACTAGAAGGAGATACTTGTTTCTTTGACAACCCGATGGCCT
GCCAATGTGTTAGGTGGAAATCTGAAAATCTGATTAATTCACGTCCTCTC
AGTTTCTCAAGTGTGTATTCTCCTGCGTTCAT

>CG32000-PG coding region 6022-17871
TTTATTGAATACTTATTCAGTAGTAATTTGATAGGCTACAGTATAGATAC
ATTACAAAATAAATATTAATTACAAATGGCAACTTATGTACATAAGTGCG
TACGCAGCAGCTGTTTCTCTGTAGTCTAGTGTAATAATCAATATATCCGT
TGCATAAACTTCACCCAATCAGATGAACTAATGCAAAGGTGTTTAATTAT
GTATTTAACACACCCCTAACACCTCCGATTTTCTAAATGACAGAACATTA
TCTATGAAGGCATTTAACCAGATAAGTACATAGTAAGTAAAATTTCATTG
GAAATACCTATTAACATGCGCGATAAAATGCGATTAACGGTACAAATTAG
CAAAAATAACAGTGTCTATATGTTTATCAATAAACATATACTTAATATAT
GTAGCCTATGTAAACTTATAGTATAATGCACAAAATGTATGTACATGCAT
AAATGTATATACAAAATGTAAGTACATCTCCTCATCATTATTTTTGGACT
TCAGCTGTGCCGTGGACTCGCACTGGATACCTCTGAGTGTATGCGATGTC
GGTTTTCTACGCTATGCTCAGACGTCTCGAAAAAGCCAGGAAAACCGCTG
GGCTGACCGTCCAAATTTTGTTCAGCGCTTAGGTTAACATATGTCGGCTG
TAGCTCCGCTCCTTGGTCGGCTGGATCGTGCAGATCATAGACCTCCGTTA
TGGGCGGCCAGTTGTCAAAGGACTTAATATTATACTCCAATCGCATATAT
TTACGATTTGATGTTTTTAAGTTTTGCTCCCGTCGCACCTGGAATCGCTT
AAATACCAAATACTCCACCACAAAAGCCTCGAGAAATGCATGAGTGATAA
ATGCAGCTGCTCCATATTGGAGCATAATGTAACGAAAAGACATATCCGGT
GGCACGATGAGTTGAAAGAAGTTGGCAATCCAATCACTTGGATAGGCAAC
TAGGTATACGACAATGCAGGCATTTATCACCAGGGTCAGGCATAGAGGCC
AATTCGACCAGATTGGCTTTCGGTATGGTGCTCCCTTTGAAAATACAAAT
GCAAGTATAATATACTGGAAACTCGATATGGCAAACATGGTGTAGTTCTC
ATAGCAGCCCAAATGATGCTCGTCAGATGGCCCGAAAGGCTTAAACCAAG
GCTGTTGATGCAGTTGAAACCAACCTGAAACAATTGAAAAATCAGTAAAA
GGCACAAATATATGTTAAGAGTATTGCGATTGAGCTTACCGGCGACCTGG
AATATGACCACAACGGCTAAATGGAGCACAAGCGAGCTTAAAGGCGTTGG
CGAAATGAGGGAACTAAGAGGCACTTGTTTAACCAGCTGACCATCGTAGG

CTTCAGTTTTGCCAAAGAAGAATGCGAAAATCGATATCAAGCCAAGGTCA
ATGTAGAGATATTGTTTGTCGGTAAGATTCGAGTCGATTGAGTATAGGAT
CATCACGGACACAAACTGTACCAGGGAATAGGCAGCCATATACTTAAAAA
TACCAAATGATGTAACCAAGGCTGCGCGACCCTCCTTGATAACATTTGGT
ACGGCCGCAATAGTCGGATTACGGGAGGTGAACGGTGATGCTATTGAGGA
TTCAGTCTCGCTCAACGATATTCCAGTGTGCGCTACCTTTAGGGCACCAC
AATCGTTAGCCCCATCTCCGCACATTGCTACATAATAGTCAAGATTCTGT
AGCTCCATCACCAGCGATTGCTTTTGCTCTGGCGACATGCGTGCATAGAT
AGCACCGCGCGTTAATATGATTTCCATTTGTTGAGGAAATTGGTCCTTAA
CAATTTGCCATGTCTTGCCATCCATGGCGAATATATATTGACGACGCCAG
CTATCATTTCGTGCCACAACTGTCGCTCCAAGCTCCACATCATCGCCGTC
ATTGTGCGTCCATGTGTCGATGGTCTTAACACTAGCCAAGCTGTTGCTGT
TGGGCAACACACTTGAGGCACAGCTCGAGCCACCATTATTTACCAATGAA
TTGGTTGACTTTGATAGATCCATATCCAGAGCGATTTCAGCAGCCGAAAA
TGTTCTATTACCATTGAGAGCTGGCGTGCACGTCTGACTGCCCAGCTCGA
GAGTATACTGTAGTTCATATTGGCCATTTTGATCGGTATTTAGCTTATCA
AAGCCTTTCAACACGGGCACCGCATGCACTGTGATTACCGCCTGCGTGGA
ACTAACAATGCCGCAATCTCGAGCAACGCTCATTGCTGTCAATATGTTAT
CACCCGTAATCATAACTGTGCGAATATTTGCCAAATTTAGTGAATTTATA
ACAGCCGTTGTATCGGGCTTTAGGCGATTCTCAAGTATTACAAATCCGAG
GAACTCCAAATTGTGCTCGACCTCTTCGCGCGAAAGGCGTTGCACCTTTG
TGTAGTTCACCTTTGGATCAAGTGTCTTAAATGCCACGGCAATAATTCGG
AAACCTTTTTTGGCATAAATGGAGAGCTGATGGGTATAGTCGTTCGGTAT
GCTCTGTGGATGGCACAACTGCTGCAACATTTCGGGCGAGCCCTTGCAGT
AAACGTTGAAGCCTTGCGCGCTGAGGCAGCGGGTGATGACTGACATACGC
TGCAGGTTGGATGTAAATGGGAACTCCCTAACAATTCCATGATCGAAATT
TGTTTTTGATTTCAACAATCCCATATTGGCCAAGAGATCATCCACCGAAC
TCTGCCGAAATACGGGTGGTGGCGCCTCTGTTATTACCGCTTCATGATGT
GTGAAAGCTAAGCTTGAGTTGATTTCCTCTTGATTCGGTTGCCGTAATAT
TGTCGGATAGATCAGGCCGTACTTCTCGTTTTCGGGTATGTTGTTGGAAT
CTTCCAGTATCCAGCCTGTCGATTCAAACATCTTAAGATCCAAGGGATCA
CCCATCATCATGCCGTTCATGACGGTAATTGAGTGGCATGTGACCATGCC
AAACAGAAAGTGATCGTATGGCAGACGCTTCACATCTTTCAGTGGTATTT
GAAACTGATTTGTGGAAGACTTTGGAACGACACCCCACATGTCTAGGCCA
TCCTCCGTTAAGGTTCCAGTCTGAATAAAATGGAAATAATTGAATAAAGC
CTTAACATTTATTTTTTCGTTATCTTGGCAGACGTACCTTATCAAAGCAA
CAGCAATCAATACTGCCTGCCACGTTGATGGAGCGCGGAGATATGCAAAA
GATATTGTTGGCTTTTAATCGTTTCTGGGCATAAAACCGACCAACAGTCA
TAGCTGCCGGCAAGGCTGGTGGCACAACAATTGTTATAAGATCCAAAGAT
TCGACCACTATCTTAACAGCATCTGTACCACGCAATATCTAAAAATAGCA
TTAAATAAACATGAGCCGTAAATAATATATATTCGCATATATTTCGTACT
CACCTTGGTGACAAGCGTATATATAAAGCCAATGCACGCAATTAGCGCCA
AAAACTGTATGAACTTGTACGAATCCTGCTCAAACTTATAGTCCACAGGC
GGTGGATAAAGAATGGAGCGTATTAGGCCACCTTTGGCAGTTATATTGCC
AGTATTTATCACGAAGGCCAAAACTTTCTTGGATCCAATGTATCGTGTTT
GTATTACTTTCGTGCCGCAGAATAGTGTGTGCCTAGCATGTTCCTTTTTA
TCAAATATCACATCGCGCTTCATCGGAAGCGGCGTTTTAGTAACTGGCAC
ACTTTCGCCAGTCAGCATTGATTCGTCTAGGATGCAGTTGCCGGAAAGTA
AAACGGCATCGCATTGCATTGTGCAGCCGGACGATGGTATTTCAATAATA
TCGCCAGGCACTAACATTTTCGTAGACAATTCCACAGATACTCCCTTGGC
GTTAACGACCCATGCATTTCCCGTGTTTAAAACTGTTTTCTGCAGAACAT
CTTGATTCTGAAATCAGATATGGGTACAGTTGCTAACACTTTAATAATAA
CTTTGTCATAATATCTCACCTTCTTTGTTTGCACAATGGACATAGATATG
CCGAATATAGACATTAACAAGATGACACAGGCATAATAGTAATAGTCATA
GGTAAACCACAGGACGACAGAAAATATTTGGAATACATAAAAGGGATTAA
GAACTTCCAGGAAGAGCAGGGTTTTCACATCCCTCAGCGGCACGGTTATC

TCATTTTGCCCAAATACAATGCGGCGTGCCAGCTGTTCTTGCACAGATAG
TCCACGTTGCTGATGGTAATAAGAACAAGGCACATTTACGTCTAATCCAT
TGATTTTCTTAAAGTTTTTAATGCTGTTGTCCCAAGCATAAACCAGCTGC
TTGCAACGAAATGTTCTCAAGGATCGAAAACCTAAAAAAGCAAAAGTTTA
TTAGATGTGCCATACAATATTCAAATAGATTAAGAATAAGTCTATTATAA
AAACCAAAAATAAATAATTAATGAAAAATTACAAAGAAGATCCTCTAGCC
TAGTCAATATATTGATTACAAACTACATTTTTAATACAAATACTAATTAA
AAATCCTAGCTTTTGAGTATAATCTAATACTAATCTTATATACTCATATT
TGTTTTCTGTAAATATTTTGAATTACCTTTGAAATGCGCGGAAGCAAAGT
GAACTGGGAGCTGTAGCTTATTATCATCGAAATCCTCTCCTGGCAATATT
TGCTGCAATAATTTTCTAAAAGAGTACTTAATATTAATACATATATATTA
AATAAACAATTTTCGAAAAGACGTACTTGAACTGCTCCACATGCAGTATC
TGCACGCTCTTGACATGATACAGTTTGTGATTACCCTGATAGTCTTCTTC
TATGAGCACCTGTTGTGCCTCCTGCAAGGAGCACGGCCTGCATGTGGCAA
GCAGATACCAATGCCGCCACCAATGTAGGACAAGGCGCAACAGACCTCCA
GTTAACACTATGCAGATCCAGCACAATATCGAACGGATATGGGAGCGCCG
GTAACCGCTTATTTTCATTTGATCTGACTGATCTTGGTTGAGAAGTGCGA
CGCATGGTTTCGGATCTGTAAGAAAACAAAGGATAATGCAAAGAGAATTT
AATGATGCTGCTGTGGCAATAACAATTTGGACAAATGTAAATATATTTGC
CCAACCATACACATATGCACACCATTCACCAGTTATCAATCTTATTATTG
GGTTTTGTCTCATACCTCGCAATACTTCGCAGTTCATCATCAGAGTTTAG
CTTAAAATTTGAGCGTTGTCTTTTCAGGAATGTCTGCAATATGGCACACA
ACATTTAATTTAAACCCGCTTTTCGGATTGTCATCTTTATTCATTACATA
AAGTATACTCTCGCTAAAAATTCTATTCTTTCCTTTGTTCATACATATAA
GTAGCTCACACGACCCCACCAACCCCCCCACCCACTCCACCGACGAGGCA
CACACAGTACATTTAAAAGAATGGGAGTACTTCTACGACCCGGCTGAGCT
ATTTAATTGGCAATTGGACTCGGACGAGGCAATTGGAGTGTCGGCACTTA
AACATCACAAAAAGCAACATTGCGACAAAAGGCATTGGTACTAAGATAAA
ATTCGAGTGATAAGAAAAAAACGTGCATTTGTACAACGTCGACCCTTTAT
ACGACGGGTTATACGACGCCGAATGTACATAATTGTTTGCCTCTAGAATT
TTCGAAATAACGTCTTGTGTGTTTGCTTTGCTAAATTATTTTATTTCTTT
ATAGGGATATAATAATACTCACAATTGCACAACCGCATGACAGCTGCGAA
TGCCAATGTTCTTATGAATATGTGCGTGTGTATGAATAAGCGTGGCTTTG
CTGCGCCTTTTGTGCTTCTGCATGTGTGTGTGTATGTTTTTTTTTTGTTT
TAGTTTTTAGCCAAGGTGCCGTTAACAGTAAATAAGCAATCGAATACCCT
TGCAAGTTTAAAAATTCATGTTCATATTCAATAAATAATACACGAAAGTA
CAATGAAAAGAACTAATCAAATTGAGAACTACCCACACAATGGCAGCAGC
GACCACTTAAGGAACCATTGTTTCACACACGCATGGACAATTAATCAACA
ATATGCGTCAATTAATACATTTATTGTATGCAAGAGCGTAATGTATGGGA
GTTGATAACATTGAAAAGAACCACAAAGAGTGTATAAAACAATAGTGGAG
AGATCACAACATATATTTTTTACACATTAATCAAATAGTCTTCATAAATA
ATAAAACACTTGGGAAAGAGGAATGGCTCACGCGTGGCATGACGTTACCC
ACTGACACACATATACACATTATTCACACGCATATATAAATACATGTATT
TATATACATGCATATGTATGTATATACATATATTGTGTTTCATAAATTCA
CTCGCCCTGGTCCGAAATGGATGCTCAGTGGCCGAATGGCCGACCTCATG
CCTGACAGACACCCAAATATTGTACGTTTAAAATTGTACTGGCATGTGTG
GACAAGCACCTGTTGCTATATTGACAGCCTAGCAAGAAATTGCGATGAAA
TGTTTTGTAGTCTTAGCCACAGTTTATAAGAAATTTTGATAATATTTTAT
AGCCATGTTCATGTGTCGGGCATACATATACACATACGTACATATGTACT
TACGGACGCGTGTAAGCACACATATATGTACATAATTCACTTAATCTTGT
AGCGTCCATGTCAACGCACGAATATCAAATTGCGGTCGCTATTCAGACTC
GAAACCGAAACCGTATCACGAACCGAAATGTAAAGCCAAGAGCCGATTAC
GATAGCAAGCGCTTCGTTAACTTAAATCAATATGTACTCAAAACTGAACT
GAACTAAACAAAACTCAACTGAACTGAACCACACAGACATCGAAGAACGA
CTGCGAGCCGGCAGCAGAGACGCGCACACAATGCTCGCAGTATGATTACA
AATCGTTATTAATTTGTATATTTTTTGAATGGGCCATTTGAATTCCATTT

AATTGCTTCTAAGTTGAATTTTTTCTTGCTGCTAATCGTTGTATATATTT
ATTTTAAATTTATGTGTTTATGAATTTAAGTGCAATTGGTCTTTTTATTT
TTTCACTTAGCAGTTACTAGGGTAAATTATATCTTGTCATTGAAAAAAAA
ATCGAAAAATCACGAAAATATCTCGTCAGAACTGTTTAAAATAGGGCACT
GAAATGTAGTTGAGCTTAACGCCCGGGAATTCCTTAACAATTGGTCTCCC
AATATAATTAATAAATATATTTGCTACACTTTACGAGGATATCTCCTAGT
CAGGCGCTTTTAATTACAATTCGTTTAGTTAACAATGTAAATCAACTTTC
AGGGTTACTGCGTTTGTAGTTAAAAAGTAACTAAGCTAAGACATACAGTA
TATATATGTTTATATATAAACGCGCAAAATACTAGTTCTTTGCTGACTAC
GAATCCAGCTCAGCCAAACAAATTACCAACACTAGCTGAAACATTGAGGA
ATCAGCTGACTGCTCAGCTGCGCGCTGCGTGCTTTAACTTAATTACATGC
AAATGCAAATGCAAGTGTTCAATTCCACTTACAGAGTAGTTCGGGAATTA
AATAGAAAATATAAATAACTTTTTCAATTGTATTTGAATTAAAATTTGTG
ATAAAACGGAGTTTGTAGGAAGAAGAGTTTGCCTTGACTGACTCTGACTG
TACAGCGAAAGGGATCCGGTCGAAGCTGACAGACTGCGTACATTCCCAGC
TGTAATAAAACAGAGTTTGTAGGAAAGAAGAGTTTGCCCTGACAAACCGA
CTGTATTTTTTTTTACCCGAAACCAACAAAAGGGATCTAGTCGAAGCTGC
CAGATTGGGAGATACCCAGTACTATCCGAACTGTGATAAAACAAAGTTTT
TACGTCAAATTTAAAAATCGACAAAACCATAAAGTTGAACTTTGCAAACG
GGGAACATCATCAAACCAAATACAATATAAACACTAAGATAAGTAACGTC
CATACATTTGAATGCGACGCAAAATTTCTGGCCTGGCCTTTCATCTGGTC
TTTGAGGCTTCGTACGGCATGCCTGCCGACTGGTTGTTCGTTCTCCAAAG
ACCAGACGAATGAATGCCGAAGTCGTGGCTTGCTGATGCTGACACGAGCT
ACGAAGTAAGTCGCCAGCCTTCCTTCAGCAAAGCTTGTCGATGCTGCTTT
GCGTCGCTTCGTTTTCGAAGTACATTGAGCGAAAAAGTTTAGAACTTACG
ATCGATGCACTGTGCTTGTGCTTTATCCATGAAATTCATGCTTTATTTAA
ATGATATATTACCCTAGATATTAACGATTAATTTTACATGAAACTTTAAG
AACTTATATAATTTTTACACAAGGCGCTTCTTGCGCAATAACTATGTTAA
GAAAATAGCACACATAGATTGATTTCACTTTTCTTTTTGGACAGTGTATG
CTCGAGAAAAGTTCTTGTCTCAGCCAATAGCATGCGTGCACCTTTCGTTG
TATGCGTGAATTATACATGCATAAGATCTTTAAAGTATGTATGTATGTGT
TGTTATTCACTGCTCTAGCTTTAGATGGCAAGAAATTAATTTTGTTAGTT
TTCTAGCGCAGATCATGACCTATCCCAAATTTGTGTTGATATATGAATGC
ATTTTGTGTACTAAGGTTGGCTCCGCAGAAAATACCCGAAGGTCAGATTT
AGTTATGACCCCATTATAATTATAATGGTCTCCTCGCGGTGTTCCCTGCG
TACTTTGTTCTTTATAATTCGAACAACAAAAATAATAAAATAAAAAAAAA
AAATAATTAAATGTATCTTTGAAAAAATAATAATTTAATTAATAATTTCA
AATATAAAAAATTACATAAATATTTGCATCGACTTAAAAATACAATTTTT
TTTGTACGGAGTACACAAAATGCATTCATATATCAACAGAAATTTTGGTT
AGGTGATCTGTGCTAAAAAACTAAGAAAAATGTTTCTTTGCCAGCCAGAG
ACTTTGCCAGAGCCGTGGATAACAACACATTCAAATATGCTGTTATGTAT
GTACATATGTATATTCCCACGCATACAAACATAATTGCATGCACGGCCCT
CATAGAGCCGAAGCTGAGAGCAAATTCTATTTTTAGCTTTTTCGTTCTCT
CGGCTGTGAGAGCGCAATGCTACAGATTTCGTTTTCGTTCTCAATTTTCA
AAAAATCAAGCTGCATAGTAGCATTTTGTTGTATGGCGTTACTCATCTTA
GTGTTTATATTGTCTTTGATCAAACTGTGAGGCTTAGCTGTCAGAGCGTC
GGACTCCCAACCCAGAGGTCGTAGGTTCAACTGCAGCAGAATTTGTATAT
TTATACGTGCTATGTATTTAATTTGATAAATGAATAAAAACAAAAAGCAA
ATACAAATGGAATTCAAGTTAACACAATAATAATAACCTTTTCTTTTTGC
TCTGAATCTTTTGGGATATCAAAATTAAATTACACTCTTTTAAATTAAAT
TAAATACGAATATTAACACGAGCAACGAGTGTAATGAGTTTCTTAGCATG
CATTTTCTTAGTTACTTGATTGAGCGGCAATTATTATTTATTTATTAAAT
GGTCGACAAAATCAAGACTTCCAAAAATATACAATATATCAATTTATAAC
AAGTAAGAGTTTCTAGTCGGGAGCTCCCGACTAGGGGATACCCTGAACCC
TCTTCCTTCAATATCAAATGCATATATATATTTTATTTTAGAAGCTATAC
AATATACCCTTATACCCATTTTTAATGGGTTCAGAGTATAAAAATTAAGT

GCTAATTAGAATACACAATTTTCTAATAGAATAACTGATCAAAGTGGGGA
CAACTATCTCGTTAAATGACCTCACGATGTGAATGTCTGGGTAAAGGATC
TCGGTAGCCATGACTAACAGTACTGCTATGTCAATGTTACTGTTAATGTT
AATGGATTGCGTGCGGGAAGAGAGTACGGTACATGGGTAGACGCAGCTTA
GTATACTAAACTAAAAATATATTTTTAAGTACACATTACACAAACCGGTT
CGCTGCCGACGAAGGGTAAATACTTCTTCATATGTAATATACAAAAACAG
CGAACGAAGAGATTATACGTTTTTGTGATTTCATCTTTGCATTGTACACG
GAATGGAAACGCTTTCACGACCTGAAGACCCCTCTCAACTGTTGAACTTA
AGACCTTGTGTCGAATCGAGGTGTTAATCGTGACATCTGATTTCTGCCGC
TCTGCAGTTACCAAGATTTAAATATTATTCGAAACAAATAAAACGCTTTA
TCAACTGCTCTTTTTGTGTTTTTAATATAAATATAAATAAATATTTCCCG
TTTTTATACTCTTTCAGAGGGTATTATGATTTTGTCGTGAAATGTGAGAC
ATCTCCGCCCATATATTCTTGATCAGCATTAACAGCCGAGTCGATATAGC
CATGTCCGTCTGTCTATTTCTATGCGACCTAGTCTCTCAGTTTTAAAGCT
TAATGAAACTCTGTAGAAGTCCCTTTTTCTGTTGCACGCAGCACACATGT
CAAAACCAGCTAGATCGAACCACTATATCATAAAGCTGCCATAGGAACGA
TCGGTCGAAAATTAAATGGTTGTATGGAAAAGTTTTTTGTCAAGAAATTT
TGATTTAACACGGCATTAAAAGATTTCACTGCTCCCTTAATTTCTGCCTT
TATATCTGCCATAGGAACGATGGGTCGAAAACTAGGTTTTTGATTAAAAA
AAAATTTTTTTTTTAATATAAGGTATAAAATGCGCTATATGTATGCCTAA
CCTGGCCTTGGTCAATTGTTTTGACATATGGATATATTTGCCCGGTCTTG
GGGCTTTATACATTTTCTGACTCGTCAAACCTTTATGTAGAAATGTTTTT
AGCATTCATTCAAATAAATAAAACAAATTAGACAGACGATTGGTAGAGCC
GTGTGTATGAAGTGCAAATATATACAATCAGAAAGACACAAATAAATGAT
TAATAAATATAACATTTACGCACCCTTTGCGAGTAGCTGTGAGTCCTGCA
GATCGAAATCTTCTGCCGGATCAACGATAGAGCTGCTGTCCTTGTGCTTC
GGATGCCTTTCATTTGAACGATGATGGTGAGGCGGGGAGGAGGTTGACTT
CGCCGGCGTATTGAGGCTGCTTCGGGTAGCCGTGTTATTGTTGTTCAAAT
TGGTTGGTAGACTTTGATCTGTTGTGCTGAGGATTGAGGCTTCGATGTCG
ATGAGGTGCTGTGTCGGCGGCGTATGCCACCAGTTGAGCTTCAAAAGATT
ATTTTTGGTTGCGATTCGAGGTAGATGAAGCAGTCCCTCCTGTTCCTCCG
CATGCACGTTGTTGCCATTGAATAGCGGCTGCAGCACCAGCAGCTCCGGG
CTGACACTGTCCGCAATCAGGCACGCGGCACCAGGCTCACATTGGCTGGC
AGACATTTTCTGGCGCTTATTTTCGCGATTCTAATTCACGTCAATATATG
TATGTATTTTAATTTGTGAGGCTTTACTCTGATTTATATCGTGTTTGGGT
TTGCCCTAATTTTTAATCGCACAGTTGGTTAGTTTTCTTTTCTTTATATC
CTTTGTGTTCTGATTGACTATGTCAGCTGGACATCTCCCGTACATGCCGG
TTTGTCGGAGTGGATGCAACAACGACTACGCGAGACGCCGCCAAAAGGAA
CAAGCCATATTTAGGGAACGCGGAGACCGTGCGTGGTGGCATCTTTGACC
GTTTCACATGTGAGAGAGTGGGGCGACGGAACGATATCGAGAAAAAGCGA
ATCGAAACGAGCTGGAATGGAAGTGCCCCACCGCACTCTTTTACGTTGTC
GCAAAGCGTAAAGCTCGTACCGGTATTTCGCAATCCACTTCCAAAGGAAT
TCACCTCCGCGCGCCGGTCTACCGAGCCAAATGAACTTTAATCATCCAAA
ATTCGAGCTAGCGTTCACCAGGCGCAGACACCACGCTCCGGCCAGTTGTT
TTTGTTGTTGTTGTCCTCCGCTGCCTACGCGCTTCTCGAAATAGAATTGT
TTGACGTACTCTTTCTGCAGTAGCACCAAATCGATCAGGATCAGGCCAAA
GCCGGCAGCTTGTTCAATATACGCCACTGATCCCAGCTGTGCGCCGTCCA
CGTCCAAGTCGGTGTCAGTGTCTGTGTCGATAACGTCACCGTCTCGACGC
TTACGCAACCTTCTCTTGAATGACCCGATATGCTCCTGCTGCTCGGTTGC
ACTTTGAACTGAATTCATTACGTTATCACTAGAAGGAGATACTTGTTTCT
TTGACAACCCGATGGCCTGCCAATGTGTTAGGTGGAAATCTGAAAATCTG
ATTAATTCACGTCCTCTCAGTTTCTCAAGTGTGTATTCTCCTGCGTTCAT
ATGTACATATCCATTCATATATACATTACCAAAGTTTCGTTGTATTTTAT
AACGTTTCAATCGCGATTATATTGTCGTTGAATACACACAAACACACACA
GTTCACTAGAGTATTTTTTTTAAAGTATTTTCTTGTTAAACAGGGCGTCT
TTTAAATTATACGCTTAGCAATAATAATTGTTCAGCTATTATGTGATTTT

CCGCGAGATGATCAAATGACAACTGTTGATAGAACGTTGACTTATTATAT
TTACAAATCGTTGTTTTGTTTTGAACTACAACTTGTGATTTGCCTGGTTT
AAAAATGGGAAAATTACTGTTTTTTACGCTCATTCCTTGATGAGTAACAT
TGGATGGCATACTCATCGCGACAAATGCATGCGGCGGCTTGGCGTTGGCT
TTTTATACACACACACAAGTCAAGAAAACGTGGAAAGTGTGGAAAGCTGT
TCGCTGTAGGTTGTCTGAAATTAATAAGAGTCTGTTATTACTCATATATA

Feature 3

Exon	Begin	End
1	28552	28391
2	27396	27325
3	27124	26048
4	25975	25871
5	25818	24871
6	22957	22643
7	22552	22472
8	22398	22195
9	22092	20281

>Ankyrin protein sequence
MTLGETLNEIQTKSQGHETATAIKRTQIQINQHSDSMDNAYIDKANINAKHQKQDATISF
LRAARSGDLGKVLEFIDAGFVDICEELLKRGINVDNATKKGNTALHIASLAGQQQVIKQL
IQYNANVNVQSLNGFTPLYMAAQENHDGCCRLLLSKGANPSLATEDGFTPLAVAMQQGHD
KVVAVLLESDVRGKVRLPALHIAAKKNDVSAATLLLQHDPNVDIVSKSGFTPLHIAAHYG
NVDIASLLLERGADVNYTAKHNITPLHVACKWGKAAVCSLLLSQHARIDATTRDGLTPLH
CASRSGHVEVIQLLLSQNAPILSKTKNGLSALHMSAQGEHDEAARLLLDHKAPVDEVTVD
YLTALHVAAHCGHVRVAKLLLDYGANPNSRALNGFTPLHIACKKNRIKVAELLLKHGANI
RATTESGLTPLHVASFMCMNIVIYLLQHDASPDMPTVRGETPLHLAARANQVTDIIRILL
RNEAQVDAVAREGQTPLHVAARLGNIDIIMLMLQHGAQVDASTKDTYTALHIAVKEGQEE
VCQLLIENGAKLDAETKKGFTPLHLASKYGKVKVANLLLQKGAAIDCQGKNDVTPLHVAT
HYDHQPVVLLLLEKGASTQISARNGHSSLHIAAKKNNLEIAQELLQHGADVGATSKSGFS
PLHLAALEGHVEMVQLLLEHGANANSSAKNGLTPLHLAAQEGHVQVSHILLEHGANISGR
TKAGYTPLHIAAHYNQINEIKFLLENDANIEITTNVGYTPLHQAAQQGHTMVINLLLRHK
ANPDAITNLNIAHNLGYITAVETLKVVTQTSVINTSTGVLEEKYKVVAPEFMHETLLSDS
EDEGGDELLDHNQYKYMATDDLKAANDHDNQNFDTTNPEHDQLDGLGGRTIERAIASNLN
ASSMERQSDNVVIVRPPVHLFLVSFLVDARGGSMRGCRHSGVRIIVPPKACAEPTRITCR
YVKPQRVANPPPLMEGEALVSRILEMSPPIVLEVPHFGSLREKEREIIILRSDNGESWRE
HSVYEDEEHLLGALNETIDADLNPLEDLHTNRIIRIVTQNVPHFFAVVSRIRQEVHAIGP
DGGTVSSTAVPQVQAIFPPHALTKKIRVGLQAQPVDLIGCSKLLGQGVAVSPVVTVEPRR
RKFHKAITLSIPAPKTCNQGMVNAPYSGTNGNAAPTLRLLCSITGGQNRAIWEDVTGSTP
LAFVKDSVSFTTTVSARFWLMDCRNVADAGRMATELYTYMAQVPFMVKFVVFAKQISATE
AKLSVFCMTDDKEDKTLEQQEYFSEVAKSRDVEVLQDQNIYLEFAGNLVPVLKSGEQLNT
KFQAFRENRLSFIVHIKDQEQPPARLCFMSEPRVGPGEAPLQPICALNVSLAAHEVNQVF
NRSNENGIDNGYNMDHGLNYKDMINAGIKASTKSPDNTIIRPIAHVTEDIQRADIRLSDI
SNLLGSDWPQLAKELGVPETDIELVKAEYADQPAAQQGLVMLRLWLKQEGTRATGNAMAQ
VLNKIGRDDIVEQCIFNLEPVTDKLERGLATARLQQNQTNLADGLNETLNIDQLSQDDEL
CPKACQSQNGKLSLCQQILNANVYILYRHNR-

>Ankyrin nucleotide sequence
CTATCTGTTGTGTCTATAAAGAATATATACATTTGCATTTAATATTTGTT
GACAGAGAGATAACTTACCATTCTGGGATTGGCATGCTTTTGGGCACAAT
TCATCATCTTGGCTAAGCTGATCTATATTCAATGTTTCATTGAGGCCGTC
AGCGAGATTGGTTTGATTCTGTTGCAATCTGGCTGTGGCCAGGCCGCGCT
CCAGTTTATCGGTGACTGGCTCCAAGTTAAAAATGCACTGTTCCACAATA
TCGTCTCGGCCAATTTTGTTTAGTACCTGAGCCATGGCATTGCCGGTGGC
GCGAGTGCCCTCCTGTTTTAGCCAGAGTCGCAGCATAACCAGACCCTGCT
GGGCAGCCGGTTGATCAGCGTACTCAGCTTTGACTAGCTCGATATCGGTC
TCGGGTACACCCAATTCTTTTGCCAGCTGCGGCCAGTCACTTCCCAGCAA
ATTCGATATATCAGATAATCTGATATCAGCGCGCTGTATGTCTTCCGTTA
CGTGGGCAATAGGCCGTATTATTGTATTGTCGGGCGATTTTGTGCTGGCC
TTTATGCCGGCATTGATCATATCCTTGTAGTTCAGGCCGTGATCCATGTT
GTATCCATTATCGATACCATTTTCATTTGACCTGTTAAACACCTGATTAA
CCTCGTGGGCTGCCAGGGACACGTTTAGAGCACAGATTGGCTGCAATGGT
GCCTCACCGGGTCCAACTCTCGGCTCGCTCATAAAACAGAGGCGAGCGGG
CGGCTGCTCTTGATCCTTGATATGCACTATAAAAGACAAACGATTCTCAC
GAAATGCCTGGAACTTGGTATTCAGTTGCTCGCCCGATTTCAATACGGGC
ACTAGATTGCCAGCAAACTCAAGATATATATTTTGATCCTGTAATACCTC
GACGTCCCGACTCTTTGCAACCTCGCTGAAATACTCTTGCTGCTCGAGAG
TCTTGTCCTCTTTATCGTCAGTCATACAGAACACGGATAGCTTTGCTTCC
GTAGCGGATATTTGCTTTGCAAATACCACAAACTTGACCATAAATGGCAC
CTGAGCCATATACGTATAAAGTTCTGTGGCCATTCGTCCAGCGTCGGCAA
CATTGCGACAATCCATTAGCCAAAAGCGTGCCGATACGGTAGTGGTAAAG
CTTACGCTATCCTTAACAAACGCCAAGGGCGTCGAACCAGTTACGTCCTC
CCAAATGGCCCGGTTCTGTCCACCAGTTATCGAGCATAAGAGCCTCAAAG
TAGGCGCGGCGTTACCATTCGTACCGCTATATGGTGCATTAACCATGCCT
TGATTGCATGTTTTTGGCGCTGGAATGCTCAATGTTATCGCCTTGTGAAA
CTTGCGTCGGCGTGGCTCCACGGTGACAACGGGCGAGACTGCAACGCCCT
GACCAAGCAGCTTGGAGCAGCCAATCAAGTCTACTGGCTGTGCCTGAAGA
CCGACTCGAATTTTCTTAGTCAATGCATGTGGTGGAAATATCGCCTGCAC
CTGCGGCACGGCTGTAGATGAAACGGTACCGCCATCGGGCCCAATTGCAT
GGACCTCTTGGCGAATGCGCGAGACAACGGCAAAGAAATGTGGCACATTT
TGTGTCACAATACGTATTATACGGTTTGTGTGTAGATCTTCCAGTGGATT
TAAATCGGCATCGATCGTTTCGTTTAAGGCGCCCAATAAGTGCTCTTCGT
CCTCGTAAACACTATGCTCCCGCCAGCTCTCGCCATTATCTGATCGTAAT
ATGATAATTTCCCGCTCCTTCTCGCGCAGCGATCCAAAATGCGGCACCTC
TAGTACAATGGG
GGGGGACATTTCCAATATGCGACTAACCAGCGCCTCACCCTCCATCAATG
GCGGTGGATTAGCCACTCTCTGAGGCTTAACATAGCGACAGGTTATACGC
GTTGGTTCTGCACACGCCTTGGGCGGCACAATGATGCGAACTCCACTGTG
TCGACAGCCACGCATTGATCCGCCACGCGCGTCCACGAGGAACGAGACGA
GGAA
CAGATGAACGGGCGGTCGTACGATGACCACATTATCGCTTTGTCGTTCCA
TAGAAGATGCATTCAAGTTTGATGCAATTGC
TCTTTCGATTGTTCGGCCGCCAAGTCCATCCAATTGATCGTGTTCGGGAT
TGGTAGTGTCAAAATTTTGGTTATCATGATCGTTGGCGGCTTTTAAATCG
TCGGTCGCCATATATTTGTATTGATTATGATCTAAAAGTTCGTCGCCACC
TTCGTCTTCGGAATCGGATAGCAGCGTTTCGTGCATAAATTCCGGAGCTA
CGACTTTGTACTTTTCCTCTAGCACACCAGTCGATGTATTGATTACGGAC
GTTTGCGTTACAACTTTAAGTGTCTCTACGGCTGTTATATAGCCTAGATT
GTGAGCTATATTCAA
ATTTGTAATTGCATCGGGATTTGCCTTGTGTCTAAGCAGTAGATTGATTA
CCATTGTATGTCCTTGCTGTGCCGCCTGATGGAGCGGTGTGTAGCCAACA
TTTGTTGTAATTTCAATGTTGGCATCATTTTCGAGCAGAAATTTAATTTC

ATTTATTTGATTGTAGTGGGCAGCAATGTGAAGGGGCGTATAGCCAGCCT
TCGTACGCCCCGAAATGTTTGCGCCGTGTTCGAGTAAAATATGCGAGACT
TGAACATGCCCCTCCTGGGCAGCCAAGTGCAACGGAGTCAAGCCGTTCTT
TGCTGAACTATTGGCATTGGCACCATGTTCCAAGAGTAGTTGAACCATTT
CGACGTGTCCCTCCAGCGCAGCCAGATGCAAGGGGGAAAATCCAGATTTA
CTTGTTGCTCCCACATCGGCGCCGTGCTGCAGTAGCTCCTGCGCAATTTC
CAAGTTATTTTTTTTGGCAGCGATATGCAGAGAACTATGCCCGTTGCGTG
CGCTTATTTGCGTCGAGGCGCCCTTCTCCAGCAGCAACAGGACCACAGGC
TGATGATCGTAGTGCGTGGCCACGTGCAAAGGTGTCACATCGTTTTTACC
CTGACAATCGATGGCGGCACCTTTTTGTAGCAATAGATTTGCCACCTTTA
CCTTCCCATACTTGCTGGCCAGGTGCAGCGGTGTAAAGCCTTTTTTGGTC
TCAGCGTCCAGCTTGGCGCCATTCTCGATGAGCAGCTGGCAGACTTCCTC
TTGGCCCTCCTTGACAGCAATATGCAAAGCCGTGTACGTATCCTTCGTGC
TGGCATCCACCTGGGCCCCATGCTGAAGCATGAGCATTATAATGTCAATG
TTGCCCAGACGGGCCGCCACATGCAGCGGAGTCTGACCCTCGCGAGCAAC
TGCATCAACTTGAGCTTCGTTGCGCAGTAATATTCGAATAATATCAGT
AACCTGATTTGCACGCGCAGCCAAATGCAGCGGCGTCTCGCCGCGAACTG
TGGGCATATCGGGACTGGCGTCGTGTTGCAGCAAGTATATAACAATGTTC
ATGCA
CATAAAGCTGGCCACATGCAGTGGGGTCAAGCCGGATTCCGTGGTTGCAC
GAATGTTGGCGCCGTGTTTAAGCAACAATTCGGCAACCTTGATGCGATTC
TTCTTGCAGGCAATGTGCAAGGGCGTAAATCCGTTGAGGGCACGCGAGTT
CGGATTGGCGCCGTAGTCAAGCAACAGCTTGGCCACACGCACATGGCCAC
AGTGGGCGGCCACATGCAGGGCCGTCAAATAGTCAACGGTAACCTCATCG
ACAGGTGCTTTGTGATCAAGCAACAGGCGTGCCGCTTCGTCGTGCTCCCC
CTGCGCCGACATATGCAGGGCTGACAAGCCGTTCTTTGTCTTTGACAGTA
TTGGCGCATTCTGAGAGAGCAGCAGCTGAATGACTTCAACGTGACCGGAG
CGCGATGCACAGTGCAGCGGTGTTAAGCCATCTCGAGTTGTTGCATCTAT
GCGGGCATGTTGCGACAGCAAGAGACTGCACACTGCCGCCTTGCCCCATT
TGCAGGCCACGTGCAGCGGCGTGATGTTATGCTTGGCCGTATAGTTGACA
TCTGCGCCGCGTTCCAGCAACAAGCTGGCAATGTCTACATTGCCATAATG
GGCGGCAATGTGCAGCGGCGTGAAGCCAGACTTCGATACAATATCAACAT
TCGGATCGTGCTGCAGCAGCAACGTTGCCGCACTAACATCATTCTTTTTG
GCAGCAATATGCAACGCCGGCAAACGCACCTTGCCGCGCACATCGCTCTC
AAGCAGCACCGCGACCACCTTGTCGTGGCCCTGCTGCATGGCCACTGCCA
GAGGCGTGAAGCCGTCTTCGGTCGCGAGCGACGGATTAGCGCCCTTGCTG
AGCAGCAGTCGGCAGCATCCGTCATGGTTCTCCTGTGCAGCCATATACAA
TGGCGTAAAGCCGTTTAGCGACTGCACATTCACGTTCGCATTGTACTGGA
TTAGCTGTTTGATCACCTGCTGCTGGCCCGCCAATGAGGCTATATGCAGC
GCGGTGTTGCCCTTTTTGGTCGCGTTGTCGACGTTAATGCCACGTTTCAG
TAGCTCCTCACAAATGTCCACGAAACC
AGCATCTATGAATTCCAGAACCTTGCCCAGATCACCGCTCCGTGCAGCTC
GTAGAAATGATATTGTTGCATC
CTGTTTTTGATGCTTTGCATTAATATTGGCTTTATCAATGTACGCGTTGT
CCATACTGTCACTGTGCTGATTTATTTGTATCTGCGTTCGTTTAATTGCT
GTTGCTGTTTCGTGGCCCTGGCTTTTCGTTTGTATCTCGTTTAGTGTCTC
GCCTAGGGTCAT

>Ankyrin coding region 20281-28552
AAAACAAAATTGAAATCGTCCCGTCCGTCGAAATAAATCTGCTTTTAATA
CATTTGAACATGTACCGTATACTACATGAGCTTCGTTTGTAGTTGCAAAC
AAAAAAAGAGTACATCCATATAAATAAATAAAACATGATATATCAACTGT
TCAATCTTTTGTTTTGTTTGAAATATTGCTTTGCAAATCGGAATTACTCG
TCGTAATTGCTTTAACATGTATATTGCCGGCTGGTTGGCACTTGTCCGCC
AGTGAATGCAATAGCGTGTCCTCAACTGAGTCTGGAATCCTGTCCAAATC
TTTTAGGTGCAATGATTTTATTAAATTGTCCTCCGTCTCAATGCCTTAAT

AAATTAAAAGAACAATTAGAGACCAATAAAAATACAGGATTCCTTTCTAT
CTAATAGTTACTATCTTTTGACACGATATTTTCATGCACATATTCGGGAA
TCTGATATTCTTGCGGTTCACTCGGTGTTGGCGGCGGTGTCGAGCAGAGT
CTATCTGTTGTGTCTATAAAGAATATATACATTTGCATTTAATATTTGTT
GACAGAGAGATAACTTACCATTCTGGGATTGGCATGCTTTTGGGCACAAT
TCATCATCTTGGCTAAGCTGATCTATATTCAATGTTTCATTGAGGCCGTC
AGCGAGATTGGTTTGATTCTGTTGCAATCTGGCTGTGGCCAGGCCGCGCT
CCAGTTTATCGGTGACTGGCTCCAAGTTAAAAATGCACTGTTCCACAATA
TCGTCTCGGCCAATTTTGTTTAGTACCTGAGCCATGGCATTGCCGGTGGC
GCGAGTGCCCTCCTGTTTTAGCCAGAGTCGCAGCATAACCAGACCCTGCT
GGGCAGCCGGTTGATCAGCGTACTCAGCTTTGACTAGCTCGATATCGGTC
TCGGGTACACCCAATTCTTTTGCCAGCTGCGGCCAGTCACTTCCCAGCAA
ATTCGATATATCAGATAATCTGATATCAGCGCGCTGTATGTCTTCCGTTA
CGTGGGCAATAGGCCGTATTATTGTATTGTCGGGCGATTTTGTGCTGGCC
TTTATGCCGGCATTGATCATATCCTTGTAGTTCAGGCCGTGATCCATGTT
GTATCCATTATCGATACCATTTTCATTTGACCTGTTAAACACCTGATTAA
CCTCGTGGGCTGCCAGGGACACGTTTAGAGCACAGATTGGCTGCAATGGT
GCCTCACCGGGTCCAACTCTCGGCTCGCTCATAAAACAGAGGCGAGCGGG
CGGCTGCTCTTGATCCTTGATATGCACTATAAAAGACAAACGATTCTCAC
GAAATGCCTGGAACTTGGTATTCAGTTGCTCGCCCGATTTCAATACGGGC
ACTAGATTGCCAGCAAACTCAAGATATATATTTTGATCCTGTAATACCTC
GACGTCCCGACTCTTTGCAACCTCGCTGAAATACTCTTGCTGCTCGAGAG
TCTTGTCCTCTTTATCGTCAGTCATACAGAACACGGATAGCTTTGCTTCC
GTAGCGGATATTTGCTTTGCAAATACCACAAACTTGACCATAAATGGCAC
CTGAGCCATATACGTATAAAGTTCTGTGGCCATTCGTCCAGCGTCGGCAA
CATTGCGACAATCCATTAGCCAAAAGCGTGCCGATACGGTAGTGGTAAAG
CTTACGCTATCCTTAACAAACGCCAAGGGCGTCGAACCAGTTACGTCCTC
CCAAATGGCCCGGTTCTGTCCACCAGTTATCGAGCATAAGAGCCTCAAAG
TAGGCGCGGCGTTACCATTCGTACCGCTATATGGTGCATTAACCATGCCT
TGATTGCATGTTTTTGGCGCTGGAATGCTCAATGTTATCGCCTTGTGAAA
CTTGCGTCGGCGTGGCTCCACGGTGACAACGGGCGAGACTGCAACGCCCT
GACCAAGCAGCTTGGAGCAGCCAATCAAGTCTACTGGCTGTGCCTGAAGA
CCGACTCGAATTTTCTTAGTCAATGCATGTGGTGGAAATATCGCCTGCAC
CTGCGGCACGGCTGTAGATGAAACGGTACCGCCATCGGGCCCAATTGCAT
GGACCTCTTGGCGAATGCGCGAGACAACGGCAAAGAAATGTGGCACATTT
TGTGTCACAATACGTATTATACGGTTTGTGTGTAGATCTTCCAGTGGATT
TAAATCGGCATCGATCGTTTCGTTTAAGGCGCCCAATAAGTGCTCTTCGT
CCTCGTAAACACTATGCTCCCGCCAGCTCTCGCCATTATCTGATCGTAAT
ATGATAATTTCCCGCTCCTTCTCGCGCAGCGATCCAAAATGCGGCACCTC
TAGTACAATGGGGCTTTAAATTTGAATACAAGTACATAGAATGGAGAAGT
AAACTTCAATTATGATTTAAAATAAAACAAATCTGATAACGTACCTTAAA
AACTTTCCCTCGACGGGGGACATTTCCAATATGCGACTAACCAGCGCCTC
ACCCTCCATCAATGGCGGTGGATTAGCCACTCTCTGAGGCTTAACATAGC
GACAGGTTATACGCGTTGGTTCTGCACACGCCTTGGGCGGCACAATGATG
CGAACTCCACTGTGTCGACAGCCACGCATTGATCCGCCACGCGCGTCCAC
GAGGAACGAGACGAGGAATCTGCAATTAGTAGTACCGAAAAGCTTAACTT
GTTTTGTGACTGTGATTTAAGTCAAAAATGATGACTCACCCCAGATGAAC
GGGCGGTCGTACGATGACCACATTATCGCTTTGTCGTTCCATAGAAGATG
CATTCAAGTTTGATGCAATTGCCCTGGTTACTTCTATCGGGTGAATTAAA
TATATGTGATTAGTTGAGTAGCACTAAGTATTTATATATATATATATATA
TAACTACGCACCTCTTTCGATTGTTCGGCCGCCAAGTCCATCCAATTGAT
CGTGTTCGGGATTGGTAGTGTCAAAATTTTGGTTATCATGATCGTTGGCG
GCTTTTAAATCGTCGGTCGCCATATATTTGTATTGATTATGATCTAAAAG
TTCGTCGCCACCTTCGTCTTCGGAATCGGATAGCAGCGTTTCGTGCATAA
ATTCCGGAGCTACGACTTTGTACTTTTCCTCTAGCACACCAGTCGATGTA
TTGATTACGGACGTTTGCGTTACAACTTTAAGTGTCTCTACGGCTGTTAT

ATAGCCTAGATTGTGAGCTATATTCAAAGCTGTTTGACCATTCTGGAAGC
GAGAAAACAAGATGTTATAAATATAATATAATAAAAAACAAAAAAACAAG
TCAGAGGTTCTAGTCGGGAGCTACCGACTAGGGGATACCCTGAACCCTCT
TCTTCCAACATCAAATGCAAATTCATCAAATTCTATTATATGTCAAGTTT
GGTGACTCTAGTTCTTAATATTTGCCAAAATTGCCCAAAAAACATAATAT
CGATATCGATTTTTATCGATTGCTTGAAAACGGAGTAAGTTATCGATTAT
TGAAACAAACTCGATCTGCGCAGGCACTAGGAGCACCAACATCTAAAATT
TCAAGTCTCTAGCTCTTATAGGTTCTGAGATCCTTGCGTTCATACATATG
GACGGATGGACGGACAGACAGACAGTCGGACGTGGCTAGATCGAGTCGGC
TATTGATGCTGATTAAGAATTTATATACTTTATTGGGTCGGATATGCTTC
CTTCTAACTGTTACATACATTTGGATTTTGCACAAATACAGTATACCCTT
ATTAATAGGTTTTGGGTATATATATATAAAAAAAAACAACAAACTTTTAA
TAAAACACTTTCACTAACAATTTGCTGGAGAAGCTCAACCAACTTCAAAA
TTTATATAAAACTGTTTAACATTTTCCGAACAACGCCCGAACTTTGAGAA
ACATGAGAGCATATGTGAATGACTAATTCAATATCCAATAAAATACTCAG
AGAGCAAGAGAATTTCATATATATATATATATTTATAGATACAGAGTGTC
AGCAAGTCAGTAGAGGAGACGAGAGAAAGAGAGATAGAGATCACCAATCA
ATAGAGAATGAGTGCACCACGTGCCAAGTTTTTTGTTGATCAATTTGAGA
TTTGAATTGAATATTGTATCTATATATTTTCATTTGTATGCAGTACACAC
AGTATATAGATGGAAAATAAACAAAGTCTGTTTAAAAATAGATTGTCTAT
TGCTCTATTCTTTTGGTCTTGTTGGTCGCGTGGAAAAATAGACATAGTTT
TTAATGAACACAAATCTCTCGGTATTTCAATTGCTTAGTCTCAGTTCCAA
TCGAAACTAAGTAGTATACAGTCGAGCTATTGATTATCGCTATACAACAA
GTGAAAGGTGCTAGTCGGGAGCTCCCGACTAAGGGATACCCTGAACCCTC
TTCTTCCAAGATCAAATGCATATATGTATATTATATTTTAGAAGCTATAT
GTCAAGTTTGGTGACTCTAGCTCTTATTATTTACCAAAATTGGCAAAAAA
CAGGATATTGATATATATCTTTATCGATTGCTTGAGTAAGTTATCGATTA
TCGGAAACAAACTCGATCTGCGCAGGCACTACGAGCACCTACATCTAAAA
TTGCAAGTCTCTAGCTCTTCTAGGTTCTGAGATCCTTGCATTCATACGGA
TAGACGGACAGGCGGACAGACAGACGGTCTAGATCGACTCGGCTATTGGT
GCTCATAAAGAATATATATACTTTATGGGGTCGGAGATGCTTCCTTCTTA
CATACATTTGGATTTAGCTCAAATACAATATACCCTTACACCCATTTTTT
AGGGGTTCAGAGTATAAAAATGTAGCCAGCTAGCTGCATTCTTCACGGGA
GGCGCATCTCTTATCTATTCCGCAGCAAAGTCATGAATCATGAAAAATTT
TGTTTATTTTTACAGATATTTTTATCGAACGAGTCCTGGACTACTGAATT
CTATTATAGCTGACATACGAACAATCGGTCGGATATTACATTTTTGAAAA
GACGTGAAGATTGAGTGAAGTCAACGAGGTATTTCCCAAATTCCTGAAAT
TTGCATAAGCTTTAAACTATTAGCTCTGCGTTATTTAAAGGGATCGTGGC
AGTAAAAATATAAAGTATGGCAAAAGCTTGGGAAACTTACATTTGTAATT
GCATCGGGATTTGCCTTGTGTCTAAGCAGTAGATTGATTACCATTGTATG
TCCTTGCTGTGCCGCCTGATGGAGCGGTGTGTAGCCAACATTTGTTGTAA
TTTCAATGTTGGCATCATTTTCGAGCAGAAATTTAATTTCATTTATTTGA
TTGTAGTGGGCAGCAATGTGAAGGGGCGTATAGCCAGCCTTCGTACGCCC
CGAAATGTTTGCGCCGTGTTCGAGTAAAATATGCGAGACTTGAACATGCC
CCTCCTGGGCAGCCAAGTGCAACGGAGTCAAGCCGTTCTTTGCTGAACTA
TTGGCATTGGCACCATGTTCCAAGAGTAGTTGAACCATTTCGACGTGTCC
CTCCAGCGCAGCCAGATGCAAGGGGGAAAATCCAGATTTACTTGTTGCTC
CCACATCGGCGCCGTGCTGCAGTAGCTCCTGCGCAATTTCCAAGTTATTT
TTTTTGGCAGCGATATGCAGAGAACTATGCCCGTTGCGTGCGCTTATTTG
CGTCGAGGCGCCCTTCTCCAGCAGCAACAGGACCACAGGCTGATGATCGT
AGTGCGTGGCCACGTGCAAAGGTGTCACATCGTTTTTACCCTGACAATCG
ATGGCGGCACCTTTTTGTAGCAATAGATTTGCCACCTTTACCTTCCCATA
CTTGCTGGCCAGGTGCAGCGGTGTAAAGCCTTTTTTGGTCTCAGCGTCCA
GCTTGGCGCCATTCTCGATGAGCAGCTGGCAGACTTCCTCTTGGCCCTCC
TTGACAGCAATATGCAAAGCCGTGTACGTATCCTTCGTGCTGGCATCCAC
CTGGGCCCCATGCTGAAGCATGAGCATTATAATGTCAATGTTGCCCAGAC

GGGCCGCCACATGCAGCGGAGTCTGACCCTCGCGAGCAACTGCATCAACT
TGAGCTTCGTTGCGCAGTAATATTCGAATAATATCAGTCTAGAGTTGAAG
TAAATGTAAACATCAATCAGTTTGTCAGAAGAATATTACAAACCTGATTT
GCACGCGCAGCCAAATGCAGCGGCGTCTCGCCGCGAACTGTGGGCATATC
GGGACTGGCGTCGTGTTGCAGCAAGTATATAACAATGTTCATGCAACCTA
CAAATAATCATTTCACAGATTTCTGATTAGCTGAATTGGAAGAAAAGCAG
AAAATTCGCAACGTACCCATAAAGCTGGCCACATGCAGTGGGGTCAAGCC
GGATTCCGTGGTTGCACGAATGTTGGCGCCGTGTTTAAGCAACAATTCGG
CAACCTTGATGCGATTCTTCTTGCAGGCAATGTGCAAGGGCGTAAATCCG
TTGAGGGCACGCGAGTTCGGATTGGCGCCGTAGTCAAGCAACAGCTTGGC
CACACGCACATGGCCACAGTGGGCGGCCACATGCAGGGCCGTCAAATAGT
CAACGGTAACCTCATCGACAGGTGCTTTGTGATCAAGCAACAGGCGTGCC
GCTTCGTCGTGCTCCCCCTGCGCCGACATATGCAGGGCTGACAAGCCGTT
CTTTGTCTTTGACAGTATTGGCGCATTCTGAGAGAGCAGCAGCTGAATGA
CTTCAACGTGACCGGAGCGCGATGCACAGTGCAGCGGTGTTAAGCCATCT
CGAGTTGTTGCATCTATGCGGGCATGTTGCGACAGCAAGAGACTGCACAC
TGCCGCCTTGCCCCATTTGCAGGCCACGTGCAGCGGCGTGATGTTATGCT
TGGCCGTATAGTTGACATCTGCGCCGCGTTCCAGCAACAAGCTGGCAATG
TCTACATTGCCATAATGGGCGGCAATGTGCAGCGGCGTGAAGCCAGACTT
CGATACAATATCAACATTCGGATCGTGCTGCAGCAGCAACGTTGCCGCAC
TAACATCATTCTTTTTGGCAGCAATATGCAACGCCGGCAAACGCACCTTG
CCGCGCACATCGCTCTCAAGCAGCACCGCGACCACCTTGTCGTGGCCCTG
CTGCATGGCCACTGCCAGAGGCGTGAAGCCGTCTTCGGTCGCGAGCGACG
GATTAGCGCCCTTGCTGAGCAGCAGTCGGCAGCATCCGTCATGGTTCTCC
TGTGCAGCCATATACAATGGCGTAAAGCCGTTTAGCGACTGCACATTCAC
GTTCGCATTGTACTGGATTAGCTGTTTGATCACCTGCTGCTGGCCCGCCA
ATGAGGCTATATGCAGCGCGGTGTTGCCCTTTTTGGTCGCGTTGTCGACG
TTAATGCCACGTTTCAGTAGCTCCTCACAAATGTCCACGAAACCATCTTT
GGCAGCCAAATGCAGTGCATTTAATCCGTTCTTAGAAAAATAAAAATAAA
ATATATTGTTTGAAATGATAGCGCCAATGATTTCTGAATATTCTGATAGT
TATTATATTATTATACTTTGCAAGCGCTTTACAATAAAATACATCTGCGA
ATATGTCTTACCGCATTGCAAGTGTTAATGTCCGTGATTAGACCAGCATC
TATGAATTCCAGAACCTTGCCCAGATCACCGCTCCGTGCAGCTCGTAGAA
ATGATATTGTTGCATCGTTCTGGATATATGGATAATAAAAAACACAACTT
TAATCCCTATATATAAAACTGTTTGGTGATTAACAAACACAGCCCAAACC
GTAAGAGCTTAGACGTGGTTGTGGGTAATTCCACCCGCAAGTATGGAAAA
AACGCAAAAGATGTTTGCGTGCCATTTTTTTGTTTACCATTCGATTATTA
ACAAAGATGTTTGCTTAACTTTGAGACTCCACGATTTGGTGAATGAAGTT
TAAATGTATTCAACGTTTTGTTGAATGAAGTTTAAATGTATTTCTAAAGA
TCTTTCAAATGTTGAATGTTGAATGAGCGAAAACTCGGGTGTAAGCTTTT
GCAAAGATCTTTAAGAAATTTAATTTAAATCAAAAATCAAGGGAAAATGT
CAGCTAAAGGTAGTAAGGGTGAAGTTCACTCCATTATTAGTTAATAAATA
CTTCTCCACAGCATTGCGCTCGAATATCTAGTTTGTTGTCCCGATCATGT
GTTTTCATTGGGGTACAGAGAATAAAAACCAATTGAAAAAATAAGCATGG
GTTGCTATTCATATATTAAACCATAATTCACATACATACACATAAAATAG
TTAGTTAATTTAGAACGAATAGCACCGATTCCGTCGACTGAGCACAATAC
AGTTTAGTTTCTTTTAATGTTAGCGTACATACACACACACACATATACAC
AGATTCAGATACAACATATGCATCTCATGTATTTGAATTTACAGTTATTA
TTGGTGACCATCGCACAATAGTGTGATCAGCACTGGAAACGGCGAGGGTG
GGGGTGGGTCATGCCACAATGTACGCCAGAGATAGCTGTGCGTGAGTGTG
TGTGTTTATGTGAGGTGATCTCATCAGGCGGATGGTCGAGCTCTTCATTC
TATATAGATTTTAATCCGTTGTCTGACAGTTTTAGATACAAGGACTAAGG
TGTGCATAATTTGTAAAAATTTTGGTCATATATTAAATCATTTTTAAACT
AAGTTTTTACCTGTTTTTGATGCTTTGCATTAATATTGGCTTTATCAATG
TACGCGTTGTCCATACTGTCACTGTGCTGATTTATTTGTATCTGCGTTCG
TTTAATTGCTGTTGCTGTTTCGTGGCCCTGGCTTTTCGTTTGTATCTCGT

TTAGTGTCTCGCCTAGGGTCATTTTTTTTTAAATGAATTCCTGAAAATAA
AGCACTGACTAATTACTAAGTTGACCACAAATGGATATCCGGAGCATAGA
AAATAAAACTTTTACCATATCAAATACATACAAACAAAATGTATACAAGC
GTATGTATGTATCTTCTATATACATGCACTTTAGTATAAACATGCTTGCA
TGCTTTTTTACAAGAATTTCCTATGTACTTCTATAACTTGTTTACAATTT
TGTTCAAGCAGCGAATTTATTTAATTATTTCTTCACATATGACGTCATTT
CTTCACCTACCTTAAATTTTGTGCGAACGAATTGGAATGCTAATACGAAA
TTAGGCCTATCGAAATTATATCGGAATATCGATATTTTCCACAAAAAGTT
AAATATCGTGTTACATTTTTGCGGCGATATATCGAGTCTGGATACCTTAC
AAGTTTTATTATCAGCGGCAGTTCATCAAATATACTATCTTATTCAAAAA
AAAACGAACCTGTAACCGAACC

Feature 4

Exon	Begin	End
1	29452	29478
2	29556	29984
3	30088	30255

>CG4038 protein sequence
MAFGRPRGSGGRSRGGGGGGRGGGGFSKFGGGFNKGGGRGTFDQGPPERVIALGNFSYAC
QNDLVCKVDIDDVPYFNAPIFLENKEQIGKIDEIFGTVRDYSVSIKLSDNIFANSFKPNQ
QLFIDPGKLLPISRFLPKPPQPKGAKKKGGPSGGGRGGRGGGGFRGGSSRGGGGGGGGFN
RGRGGGGGGGGGFNRSRGGAGGGGGRGRW-

>CG4038 nucleotide sequence
ATGGCTTTTGGACGCCCTCGTGGTTCTGGTGGCAGA
TCTCGCGGAGGTGGTGGAGGTGGACGCGGCGGTGGTGGATTCAGCAAGTT
CGGCGGAGGATTTAATAAAGGCGGAGGTCGGGGTACCTTTGATCAGGGCC
CGCCCGAGCGTGTGATTGCATTGGGCAACTTTAGCTATGCATGCCAGAAC
GATCTAGTTTGTAAAGTAGATATAGATGATGTACCATATTTCAATGCCCC
AATATTCCTCGAAAATAAGGAGCAGATCGGAAAAATTGATGAAATTTTTG
GCACAGTTCGGGACTACTCCGTCTCCATTAAACTTTCTGATAATATATTT
GCAAATAGTTTTAAGCCGAACCAACAGTTGTTTATTGATCCTGGAAAATT
ACTGCCAATTTCAAGATTTCTACCAAAGCCACCACAACCAAAAGGTGCTA
AAAAGAAGGGTGGCCCTAGTGGTGGAGGA
CGTGGTGGACGCGGCGGCGGTGGCTTTAGAGGAGGGTCAAGCCGAGGAGG
AGGAGGAGGCGGCGGTGGATTTAATAGAGGACGTGGTGGAGGCGGCGGCG
GCGGCGGTGGTGGATTTAATAGAAGCCGTGGTGGCGCCGGTGGCGGAGGC
GGGCGCGGCCGTTGGTAG

>CG4038 coding region 29452-30255
CGGCGATATATCGAGTCTGGATACCTTACAAGTTTTATTATCAGCGGCAG
TTCATCAAATATACTATCTTATTCAAAAAAAAACGAACCTGTAACCGAAC
CTATTAAAAATTTTGATGCAAGAACTTGACTTCAACAGTTTTTATTTTAA
GATTAGATTTCGTCGTTTTGCTGCACACAGCTGCGCTTAAATGGGGGTAG
GTCCAGCTGTTTAAAATTGCCGTTAAACATTATTGAATGCATTTGAAATA
GAGTGAAACAAGTTCCAATAATCGATGATAAGTCTCCTGCGCTAACTTGA
CGCACAGCAATTAAGAATGAACTGATAGCATTAAAAGAATCCGACAAATG
AACTAAATGAACGATGATAGTAAGGGAGATGATTTTAATGAACCGCTCAA
TTGAACCTATACTGAACTATTTAATACAACACTAACAGCTCACGAGTAGT
CGCACGTGTTTATATTTGTTTCTGACTTTGTTGCGCCTTTTCGCATTAAA
ATGGCTTTTGGACGCCCTCGTGGTTCTGGTGGCAGAGGTTTGACACCAAT
GGTTGCGAACTTAAACTTACATACATTTTGTTAAAAAAATTTTCGTAAAT

AGGTTCTCGCGGAGGTGGTGGAGGTGGACGCGGCGGTGGTGGATTCAGCA
AGTTCGGCGGAGGATTTAATAAAGGCGGAGGTCGGGGTACCTTTGATCAG
GGCCCGCCCGAGCGTGTGATTGCATTGGGCAACTTTAGCTATGCATGCCA
GAACGATCTAGTTTGTAAAGTAGATATAGATGATGTACCATATTTCAATG
CCCCAATATTCCTCGAAAATAAGGAGCAGATCGGAAAAATTGATGAAATT
TTTGGCACAGTTCGGGACTACTCCGTCTCCATTAAACTTTCTGATAATAT
ATTTGCAAATAGTTTTAAGCCGAACCAACAGTTGTTTATTGATCCTGGAA
AATTACTGCCAATTTCAAGATTTCTACCAAAGCCACCACAACCAAAAGGT
GCTAAAAAGAAGGGTGGCCCTAGTGGTGGAGGAGTAAGGGGGCGTGGAGG
CGGAATGGGTAAGTTAAAATAATTGTATTATGTCTAGAACTATTCTCCAC
AGTAATTTTAATTCTTTACAATTATACTATTTAGGACGTGGTGGACGCGG
CGGCGGTGGCTTTAGAGGAGGGTCAAGCCGAGGAGGAGGAGGAGGCGGCG
GTGGATTTAATAGAGGACGTGGTGGAGGCGGCGGCGGCGGCGGTGGTGGA
TTTAATAGAAGCCGTGGTGGCGCCGGTGGCGGAGGCGGGCGCGGCCGTTG
GTAGTAATTCAGAATGAAGCGCGAAACTGAATAAAATGTAAAACAAATTA
TAAAAATTAATTGGTTGTTTTGATAAATCGTAAATTGGTTTAATATTTAT
TTGGAAATACTGACCACGATACATGCAATATTTTACGGATATATATATAT
GTATGTATAGTTATGTACATTTGTATAGTTGTGTGCACTTATTGTGCTAT
TGTTGCTTTAAATGATGTTAATAATTAAGTGTGGAAGATTCATAGGCCCT
TAAAATTTGTATCATTCATATTTTCGATATTTTTGTGATTAAGCTGCATC
GCATTGACAGCTTTCTCACTTATCCGCGAGTTGTTTTTCTTGTTCATGTT
TAATATTCGACCGGGCCCGTAAGAGGCGAGAGTATCTTTGTTGCCTACGA
TTGCATTTGCGATATTGTTATAAATGCGATCGCTGTAGTTGGAATTGTAG
CGAAATGACCGGGTGTATTCTTGCCAGAGTTGAATATTTTTTTGTCGCTG
CGTC

Feature 5

Exon	Begin	End
1	35,948	32,784
2	32727	32386
3	31,964	31,709
4	31,447	31,185
5	31,125	30,829
6	30,768	30,646

>Rhomboid-5 protein sequence
MHPIRSDLSSGASNQRRNGNGNGNDCNAIVAGNACAGRRKSVNYQPQLSLKANGDEPIGS
PAGRFSPHPNEIFLAQSGKLTIPLAKFDDSLINGCLPPPSPAPNSDCLPSDLSQAHAQTQ
SNFQTKIGLNTNSSGNQMQQLNLNQFQQQQQGHAQGYPLSSSNSTSSLCTNNSGSTQSQP
AESSSGIHHKHGKYSLHPLHPRAMSPNSKYRLERYRDPAQKVKLIEAMNLLSPGLAPATA
TTQSATSTRYFMPPKMVECDAYNGYLGSTVHTPVKRYVPTPPPASDIYTDSGLGPTRTAT
GTATTCAATPPLSSQYVNIPYNYRAKCCHNEHIESGQAQGSYSKNAQTYNVHNVPSSSAA
TSSNSNPCPCPSPSPASSDSLSITFPAPGTSICPPIVSTGKNIGSCNKLRSNMESARINH
PDEAESIIVIQSSALGQSEGQRQVPAQVQVQVQGQDMAGTCLHCNTTRRTTGVHQTTQTT
GPISPVPLAMPMAMPVPVTSTDLAKLKHQHDQEMMAHENASATIEGSDLSSTIAFQRQQQ
QQQVYPVAHLANSPRLQQQQQHVMPQQQQQQQALRYSCKKRIIIYMRREVARFFGVETST
ETADFALWYGRHRRLAIRRFGPLNTSSELDYNMPMPIDNRDNGNNAEAIGYHATDRPDIL
PAQDAQNVDMALHATGSCRWRKCYAGSDFSAGEFVERKASVAHMLMTGVSYLISMFNVRP
TKNGHGRLGKRFHHRQWSRSFAPIHVHGRGVDSMDHGEDMDAECSMGIANSLAALIDDEV
FFDSPCDASSTSSANEDGETNKQPPTDAVGLGLGVGVGGEGGVGVYMASERHHNGWRTSA
LNGGNGGNGDMHLIADAHQIHQVNHIHMSGSSGQGHSVLRTSNSTTNSSTTNRGNRIAAQ
LLDGVLENSRRPQTQHIKYFSVNDLDDRTDHRPFFTYWINTVQIVVLFLSIICYGIAPIG

FGTEQKTGQVLVTSLSLQTVQHIEQRNLWIGPRNNDLVHMGAKFAACMRRDIKIMEVVAK
TRRQERETACCIRNDDSGCVQSSQADCSIRGLYPTKSISTWKKWSPGESGPGGRISGSVC
GLDPKFCDAPASIAPYEWPDDITKWPICRKTNSFSQRYRYKDHTAEHMVCEVIGHPCCTG
LYGECRITTREYCDFVNGYFHEEASLCSQISCLNNVCGMFPFISVEIPDQIYRLLTSLCM
HAGILHLAITLIFQYLFLADLERLIGTLRTAVVYIMSGLAGNLTSAVLVPHRPEAFSLAL
WCGILAGCPAYMDALEKCTQAICGIVQTIAIVHSSIRDRHITISAELCWSASWSCLWHVF
NNILGAVCHFHEIRTQKKGTYTINLIWTCILFHLFVYATLITTFYIYPSEFSTFSFVDDI
FGSNGNGNSYIGATNSNIGHQNGEVSSTPRRYSQTQKPQYNYHHQSEDIIRNAFPEGNII
STYIRRRSDKKIFNSGKNTPGHFATIPTTAIAFITISQMQS-

>Rhomboid-5 nucleotide sequence
CTACGATTGCATTTGCGATATTGTTATAAATGCGATCGCTGTAGTTGGAA
TTGTAGCGAAATGACCGGGTGTATTCTTGCCAGAGTTGAATATTTTTTTG
TCGCTGCGTCTCCGGATATATGTCGAGATGATGTTGCCTTCAGGGAACGCATTCCTGATTATGT
CTTCCGATT
GATGGTGATAATTATATTGAGGCTTTTGTGTCTGAGAATATCTGCGGGGG
GTGCTGCTTACCTCGCCATTCTGGTGTCCAATATTGCTATTCGTGGCACC
AATGTAGCTGTTACCGTTACCATTACTACCGAATATATCATCCACAAAAC
TAAATGTACTAAATTCGCTGGGATAGATGTAAAACGTTGTTATTAAAGTT
GCGTATACAAACAGATGAAACAGTATGCATGTCCAAATCAGATTTATTGTATATGTGCCTTTT
TTCTGCGTCCGTATTTCGTGAAAGTGGCAAACGG
CACCAAGGATATTGTTAAAAACGTGCCACAAGCAACTCCAGCTAGCAGAC
CAGCAAAGTTCAGCTGATATGGTAATGTGCCGATCCCGAATAGAACTGTG
GACAATAGCAATAGTTTGAACAATGCCACATATGGCTTGTGTACATTTTT
CCAGTGCATCCATATAAGCAGGGCAGCCAGCGAGGATACCACACCACAGA
GCGAGGCTGAAGG
CCTCAGGTCGATGGGGTACCAGAACAGCACTGGTAAGATTTCCAGCCAAG
CCCGACATAATATATACCACGGCAGTTCGCAAGGTTCCGATTAATCTCTC
CAAGTCGGCAAGAAATAAATACTGAAATATAAGTGTAATTGCCAGATGCA
AAATGCCAGCATGCATGCATAGCGATGTCAAAAGTCTATAAATCTGATCT
GGAATTTCCACTGAGATAAATGGGAACATGCCACAAACGTTATTTAAGCA
GGAAATTTGTGAACAGAGTGAAGCTTCCTCATGAAAATATCCATTAACGAAATCAC
AATATTCGCGGGTTGTTATTCTGCATTCACCATAGAGTCCTGTACAGCAA
GGATGGCCAATTACTTCACATACCATATGCTCAGCCGTGTGATCTTTGTA
GCGGTAGCGCTGCGAGAATGAGTTTGTTTTCCGGCAAATGGGCCACTTGG
TTATGTCATCGGGCCACTCGTATGGCGCTATGGATGCGGGTGCATCGCAA
AATTTTGGATCTAAGCCGCACACAGATCCTGAAATCCGCCCGCCAGGCCC
AGACTCACCGGGTGACCACTTTTTCCAAGTCGATATTGATTTCGTAGGATAGAGACCCCGTAT
GGAACAATCCGCCTGGGAGCTCTGAACGC
AGCCGGAGTCATCGTTCCGTATGCAGCAAGCCGTTTCACGTTCCTGCCTT
CTGGTTTTTGCCACCACCTCCATGATCTTAATGTCGCGCCGCATGCATGC
CGCAAATTTTGCACCCATATGAACCAGGTCATTATTGCGTGGACCAATCC
AGAGATTGCGCTGCTCGATGTGCTGTACCGTCTGCAGGCTTAGGCTGGTC
ACCAGCACCTGTCCGGTTTTCTGTTCGGTACCGAAACCAATTGGTGCGAT
GCCGTAGCAAATGATCGATAAAAACAGTACGACAATTTGTACCGTATTTA
TCCAGTATGTGAAGAATGGTCGATGATCCGTGCGATCGTCCAGATCATTG
ACCGAAAAGTATTTGATATGCTGCGTCTGGGGGCGTCTCGAGTTCTCCAG
AACGCCATCCAGCAGCTGGGCAGCTATACGATTGCCACGATTGGTGGTTG
AAGAATTAGTTGTCGAATTGGATGTGCGCAGCACGGAGTGGCCTTGTCCC
GAGCTGCCGCTCATATGAATATGGTTCACCTGGTGTATTTGATGTGCATC
GGCTATCAAATGCATATCCCCGTTACCCCCGTTGCCACCATTTAGAGCCG
AGGTGCGCCAACCATTGTGATGCCTTTCAGACGCCATATAGACGCCAACG
CCGCCCTCGCCTCCGACTCCGACGCCCAGTCCCAAGCCGACTGCATCTGT
TGGCGGCTGTTTGTTTGTTTCGCCGTCCTCATTTGCCGATGAGGTGGACG
ACGCATCGCATGGACTATCGAAGAAGACCTCGTCATCGATCAGTGCAGCC
AGACTATTGGCAATGCCCATGCTGCACTCGGCGTCCATGTCCTCGCCATG

ATCCATGGAATCCACGCCCCGGCCGTGCACATGGATCGGCGCAAAGCTGC
GTGACCATTGCCGATGATGGAATCGCTTACCAAGGCGGCCATGTCCATTC
TTTGTGGGTCGCACATTGAACATGCTTATCAGGTAGCTAACACCTGTCAT
CAGCATGTGAGCTACCGATGCCTTGCGCTCCACAAATTCGCCCGCACTGA
AATCGCTACCCGCATAGCATTTGCGCCATCTGCAGCTACCGGTGGCATGC
AGCGCCATATCCACATTTTGGGCATCCTGAGCGGGCAGGATATCGGGCCG
ATCCGTTGCATGGTAGCCAATTGCTTCCGCATTGTTTCCATTATCACGAT
TGTCAATTGGCATGGGCATATTGTAGTCCAATTCCGAGCTTGTGTTCAAT
GGCCCAAAGCGTCGAATTGCCAGACGTCGATGGCGTCCGTACCAGAGGGC
AAAGTCCGCAGTCTCCGTGCTGGTCTCCACGCCAAAGAAACGCGCAACCT
CGCGTCGCATATAAATAATGATGCGTTTCTTGCAGGAATACCGAAGGGCC
TGCTGCTGTTGTTGCTGCTGCGGCATGACATGTTGCTGCTGCTGCTGCAG
GCGAGGCGAATTGGCCAAGTGGGCAACTGGATATACCTGCTGCTGTTGCT
GCTGACGTTGGAATGCTATTGTGCTAGACAGATCGGAGCCCTCGATTGTG
GCGCTGGCATTCTCGTGTGCCATCATTTCCTGATCATGCTGATGCTTCAG
CTTGGCCAGATCCGTTGATGTTACGGGCACGGGCATGGCCATGGGCATGG
CCAGGGGCACGGGACTGATTGGACCAGTTGTTTGTGTTGTCTGATGCACA
CCGGTTGTCCGACGCGTGGTATTGCAATGCAAACAGGTGCCAGCCATGTC
CTGACCCTGCACCTGCACCTGCACCTGAGCTGGCACCTGTCGCTGGCCCT
CACTCTGACCCAATGCCGAAGACTGTATGACTATAATGGACTCGGCCTCA
TCCGGATGGTTGATGCGTGCACTTTCCATATTAGATCGCAACTTGTTGCA
CGATCCAATGTTTTTTCCCGTGCTCACAATGGGCGGGCAGATGCTGGTGC
CTGGTGCAGGAAATGTTATGGATAATGAGTCAGATGAGGCAGGCGAAGGC
GATGGGCATGGGCATGGATTCGAGTTGGAAGATGTGGCGGCTGATGATGA
TGGCACATTGTGCACATTATACGTTTGTGCGTTCTTCGAATAGCTTCCCT
GAGCCTGCCCAGACTCAATATGTTCATTGTGACAGCATTTGGCACGATAA
TTATACGGAATATTCACGTATTGCGATGACAGCGGAGGCGTGGCAGCGCA
TGTTGTTGCCGTACCTGTTGCTGTCCTTGTTGGGCCGAGCCCTGAATCCG
TGTATATATCACTCGCCGGAGGCGGCGTCGGGACATACCGTTTGACCGGC
GTGTGCACCGTGGAGCCCAAATAACCGTTGTATGCATCACATTCCACCAT
TTTGGGTGGCATAAAATAACGTGTCGAGGTTGCACTTTGAGTTGTCGCCG
TTGCGGGTGCAAGTCCCGGCGATAATAGATTCATTGCTTCGATTAGTTTC
ACCTTTTGAGCCGGATCTCGGTAGCGCTCCAGGCGATATTTCGAGTTGGG
CGACATTGCACGTGGATGCAGTGGATGCAGGCTATATTTGCCATGCTTAT
GATGGATGCCGGACGAGCTCTCCGCCGGCTGCGACTGCGTGCTTCCCGAA
TTGTTTGTGCATAGCGACGAGGTCGAGTTCGAGGATGACAACGGATAGCC
TTGAGCGTGGCCCTGTTGCTGTTGCTGAAATTGGTTCAGATTCAGCTGCT
GCATCTGATTTCCGCTGCTGTTCGTGTTAAGGCCAATTTTGGTTTGAAAA
TTCGATTGGGTTTGAGCGTGGGCCTGACTTAGATCCGAGGGCAAACAATC
GCTATTTGGCGCAGGCGACGGCGGCGGCAAGCAGCCGTTTATTAGCGAAT
CATCAAATTTGGCCAACGGTATGGTTAGTTTGCCCGACTGCGCCAGGAAT
ATCTCATTTGGATGCGGACTAAAGCGACCAGCTGGTGACCCAATCGGTTC
ATCTCCGTTCGCCTTAAGCGATAGTTGCGGCTGATAATTAACCGATTTCC
TACGACCGGCACACGCATTCCCGGCAACTATAGCATTGCAATCATTTCCA
TTTCCATTTCCATTTCGGCGCTGGTTGCTAGCACCACTCGAAAGATCGCT
ACGAATTGGATGCAT

>Rhomboid-5 coding region 30646-35948
GCGGCGGTGGATTTAATAGAGGACGTGGTGGAGGCGGCGGCGGCGGCGGT
GGTGGATTTAATAGAAGCCGTGGTGGCGCCGGTGGCGGAGGCGGGCGCGG
CCGTTGGTAGTAATTCAGAATGAAGCGCGAAACTGAATAAAATGTAAAAC
AAATTATAAAAATTAATTGGTTGTTTTGATAAATCGTAAATTGGTTTAAT
ATTTATTTGGAAATACTGACCACGATACATGCAATATTTTACGGATATAT
ATATATGTATGTATAGTTATGTACATTTGTATAGTTGTGTGCACTTATTG
TGCTATTGTTGCTTTAAATGATGTTAATAATTAAGTGTGGAAGATTCATA
GGCCCTTAAAATTTGTATCATTCATATTTTCGATATTTTTGTGATTAAGC

TGCATCGCATTGACAGCTTTCTCACTTATCCGCGAGTTGTTTTTCTTGTT
CATGTTTAATATTCGACCGGGCCCGTAAGAGGCGAGAGTATCTTTGTTGC
CTACGATTGCATTTGCGATATTGTTATAAATGCGATCGCTGTAGTTGGAA
TTGTAGCGAAATGACCGGGTGTATTCTTGCCAGAGTTGAATATTTTTTTG
TCGCTGCGTCTCCGGATATATGTGCTGAAAGCAAGCGATAAATGTATATA
TTCGCAGATATTGGAATGTGAAGAATATTTTACCGAGATGATGTTGCCTT
CAGGGAACGCATTCCTGATTATGTCTTCCGATTGATGGTGATAATTATAT
TGAGGCTTTTGTGTCTGAGAATATCTGCGGGGGGTGCTGCTTACCTCGCC
ATTCTGGTGTCCAATATTGCTATTCGTGGCACCAATGTAGCTGTTACCGT
TACCATTACTACCGAATATATCATCCACAAAACTAAATGTACTAAATTCG
CTGGGATAGATGTAAAACGTTGTTATTAAAGTTGCGTATACAAACAGATG
AAACAGTATGCATGTCCAAATCAGATTTATCTGCGAATAGAAAACAAACA
AATTTCATTCACGGAAACCTCTTTAATATTAGTTAAGCATGTATATGTGC
CTTTTTTCTGCGTCCGTATTTCGTGAAAGTGGCAAACGGCACCAAGGATA
TTGTTAAAAACGTGCCACAAGCAACTCCAGCTAGCAGACCAGCAAAGTTC
AGCTGATATGGTAATGTGCCGATCCCGAATAGAACTGTGGACAATAGCAA
TAGTTTGAACAATGCCACATATGGCTTGTGTACATTTTTCCAGTGCATCC
ATATAAGCAGGGCAGCCAGCGAGGATACCACACCACAGAGCGAGGCTGAA
GGTCCCACCTGTTGACAAATTGAAAACGAGGTAAGAATTTGAATTTGAAC
AAATATACAGTTAAACTGAGTTAGTGATGAGTATGCATATGTCTCTCTGT
TAAATGTAATATCGATACATTCCTTTGGTTTGTAACTCTATTAAATGTAA
TATCGATACAATCCGAAGGAATGTATTGTAAAATTCATACTTTACATCGA
GTAAATCAATTAAAGAAGGCTGATCTACTTTAATATATATACATATAGTT
ACATAAATACCTACCTCAGGTCGATGGGGTACCAGAACAGCACTGGTAAG
ATTTCCAGCCAAGCCCGACATAATATATACCACGGCAGTTCGCAAGGTTC
CGATTAATCTCTCCAAGTCGGCAAGAAATAAATACTGAAATATAAGTGTA
ATTGCCAGATGCAAAATGCCAGCATGCATGCATAGCGATGTCAAAAGTCT
ATAAATCTGATCTGGAATTTCCACTGAGATAAATGGGAACATGCCACAAA
CGTTATTTAAGCAGGAAATCTGAAATAAGAAAGAATCATTAGAAAATCAT
AATTAACTATAAAATATTTGGTTTTTTGTAGTGCGCATGTAACTATGTTA
ATGTTATTCATATAAATTTCGGTTTTTTATTTTTTTACACAACTGCCAAT
TTTAAAAAAAGTGATCGCATGCTTATTCAAAATATCGATACATTCATTCGG
CTTGTATCGATATTACATTTAACAGAATTACATATACATCGGTAACTTCG
ATTAACAGTGGATGTGTTTAAATTTCAATTTAAATTACTCAAATTCTTTC
ACATCGGTCTACTTTATTATATTGCAAATTGGAGCAATCGGACAATCGGA
AAATTGTGGCTAAGTATAGAGAATATCTACAAGGCTATTTACTATTCGGC
ACACCGAACAGTCTGTGATATTAGTCTGGTTAATATTTACTTGTGAACAG
AGTGAAGCTTCCTCATGAAAATATCCATTAACGAAATCACAATATTCGCG
GGTTGTTATTCTGCATTCACCATAGAGTCCTGTACAGCAAGGATGGCCAA
TTACTTCACATACCATATGCTCAGCCGTGTGATCTTTGTAGCGGTAGCGC
TGCGAGAATGAGTTTGTTTTCCGGCAAATGGGCCACTTGGTTATGTCATC
GGGCCACTCGTATGGCGCTATGGATGCGGGTGCATCGCAAAATTTTGGAT
CTAAGCCGCACACAGATCCTGAAATCCGCCCGCCAGGCCCAGACTCACCG
GGTGACCACTTTTTCCAAGTCGATATTGATTTCTGGTGGGAATTAGAAAC
CGATAAAGACGGGCATATATATAGTATATAACACTTACCGTAGGATAGAG
ACCCCGTATGGAACAATCCGCCTGGGAGCTCTGAACGCAGCCGGAGTCAT
CGTTCCGTATGCAGCAAGCCGTTTCACGTTCCTGCCTTCTGGTTTTTGCC
ACCACCTCCATGATCTTAATGTCGCGCCGCATGCATGCCGCAAATTTTGC
ACCCATATGAACCAGGTCATTATTGCGTGGACCAATCCAGAGATTGCGCT
GCTCGATGTGCTGTACCGTCTGCAGGCTTAGGCTGGTCACCAGCACCTGT
CCGGTTTTCTGTTCGGTACCGAAACCAATTGGTGCGATGCCGTAGCAAAT
GATCGATAAAAACAGTACGACAATTTGTACCGTATTTATCCAGTATGTGA
AGAATGGTCGATGATCCGTGCGATCGTCCAGATCATTGACCGAAAAGTAT
TTGATATGCTGCGTCTGGGGGCGTCTCGAGTTCTCCAGAACGCCATCCAG
CAGCTGGGCAGCTATACGATTGCCACGATTGGTGGTTGAAGAATTAGTTG
TCGAATTGGATGTGCGCAGCACGGAGTGGCCTTGTCCCGAGCTGCCGCTC

ATATGAATATGGTTCACCTGGTGTATTTGATGTGCATCGGCTATCAAATG
CATATCCCCGTTACCCCCGTTGCCACCATTTAGAGCCGAGGTGCGCCAAC
CATTGTGATGCCTTTCAGACGCCATATAGACGCCAACGCCGCCCTCGCCT
CCGACTCCGACGCCCAGTCCCAAGCCGACTGCATCTGTTGGCGGCTGTTT
GTTTGTTTCGCCGTCCTCATTTGCCGATGAGGTGGACGACGCATCGCATG
GACTATCGAAGAAGACCTCGTCATCGATCAGTGCAGCCAGACTATTGGCA
ATGCCCATGCTGCACTCGGCGTCCATGTCCTCGCCATGATCCATGGAATC
CACGCCCCGGCCGTGCACATGGATCGGCGCAAAGCTGCGTGACCATTGCC
GATGATGGAATCGCTTACCAAGGCGGCCATGTCCATTCTTTGTGGGTCGC
ACATTGAACATGCTTATCAGGTAGCTAACACCTGTCATCAGCATGTGAGC
TACCGATGCCTTGCGCTCCACAAATTCGCCCGCACTGAAATCGCTACCCG
CATAGCATTTGCGCCATCTGCAGCTACCGGTGGCATGCAGCGCCATATCC
ACATTTTGGGCATCCTGAGCGGGCAGGATATCGGGCCGATCCGTTGCATG
GTAGCCAATTGCTTCCGCATTGTTTCCATTATCACGATTGTCAATTGGCA
TGGGCATATTGTAGTCCAATTCCGAGCTTGTGTTCAATGGCCCAAAGCGT
CGAATTGCCAGACGTCGATGGCGTCCGTACCAGAGGGCAAAGTCCGCAGT
CTCCGTGCTGGTCTCCACGCCAAAGAAACGCGCAACCTCGCGTCGCATAT
AAATAATGATGCGTTTCTTGCAGGAATACCGAAGGGCCTGCTGCTGTTGT
TGCTGCTGCGGCATGACATGTTGCTGCTGCTGCTGCAGGCGAGGCGAATT
GGCCAAGTGGGCAACTGGATATACCTGCTGCTGTTGCTGCTGACGTTGGA
ATGCTATTGTGCTAGACAGATCGGAGCCCTCGATTGTGGCGCTGGCATTC
TCGTGTGCCATCATTTCCTGATCATGCTGATGCTTCAGCTTGGCCAGATC
CGTTGATGTTACGGGCACGGGCATGGCCATGGGCATGGCCAGGGGCACGG
GACTGATTGGACCAGTTGTTTGTGTTGTCTGATGCACACCGGTTGTCCGA
CGCGTGGTATTGCAATGCAAACAGGTGCCAGCCATGTCCTGACCCTGCAC
CTGCACCTGCACCTGAGCTGGCACCTGTCGCTGGCCCTCACTCTGACCCA
ATGCCGAAGACTGTATGACTATAATGGACTCGGCCTCATCCGGATGGTTG
ATGCGTGCACTTTCCATATTAGATCGCAACTTGTTGCACGATCCAATGTT
TTTTCCCGTGCTCACAATGGGCGGGCAGATGCTGGTGCCTGGTGCAGGAA
ATGTTATGGATAATGAGTCAGATGAGGCAGGCGAAGGCGATGGGCATGGG
CATGGATTCGAGTTGGAAGATGTGGCGGCTGATGATGATGGCACATTGTG
CACATTATACGTTTGTGCGTTCTTCGAATAGCTTCCCTGAGCCTGCCCAG
ACTCAATATGTTCATTGTGACAGCATTTGGCACGATAATTATACGGAATA
TTCACGTATTGCGATGACAGCGGAGGCGTGGCAGCGCATGTTGTTGCCGT
ACCTGTTGCTGTCCTTGTTGGGCCGAGCCCTGAATCCGTGTATATATCAC
TCGCCGGAGGCGGCGTCGGGACATACCGTTTGACCGGCGTGTGCACCGTG
GAGCCCAAATAACCGTTGTATGCATCACATTCCACCATTTTGGGTGGCAT
AAAATAACGTGTCGAGGTTGCACTTTGAGTTGTCGCCGTTGCGGGTGCAA
GTCCCGGCGATAATAGATTCATTGCTTCGATTAGTTTCACCTTTTGAGCC
GGATCTCGGTAGCGCTCCAGGCGATATTTCGAGTTGGGCGACATTGCACG
TGGATGCAGTGGATGCAGGCTATATTTGCCATGCTTATGATGGATGCCGG
ACGAGCTCTCCGCCGGCTGCGACTGCGTGCTTCCCGAATTGTTTGTGCAT
AGCGACGAGGTCGAGTTCGAGGATGACAACGGATAGCCTTGAGCGTGGCC
CTGTTGCTGTTGCTGAAATTGGTTCAGATTCAGCTGCTGCATCTGATTTC
CGCTGCTGTTCGTGTTAAGGCCAATTTTGGTTTGAAAATTCGATTGGGTT
TGAGCGTGGGCCTGACTTAGATCCGAGGGCAAACAATCGCTATTTGGCGC
AGGCGACGGCGGCGGCAAGCAGCCGTTTATTAGCGAATCATCAAATTTGG
CCAACGGTATGGTTAGTTTGCCCGACTGCGCCAGGAATATCTCATTTGGA
TGCGGACTAAAGCGACCAGCTGGTGACCCAATCGGTTCATCTCCGTTCGC
CTTAAGCGATAGTTGCGGCTGATAATTAACCGATTTCCTACGACCGGCAC
ACGCATTCCCGGCAACTATAGCATTGCAATCATTTCCATTTCCATTTCCA
TTTCGGCGCTGGTTGCTAGCACCACTCGAAAGATCGCTACGAATTGGATG
CATCGTAAATGACCTGGTTGTAGCGTTTTCCTAACTGCTTTATCTGGAAA
AACGATTGCACAGAGAGAGAGAGCTATTTATTAATTGTTAGAGGACATCC
CCTAGGCTCTATCGAATCCGGATATTGTTTTAAATTGGGTATGAATATAT
ATATATGCTAGCCAAGAAATAAAGAAAGAAATACGATAAACGAGCACATA

GATTCAAAACTTAGTTTCTTTTGAATACTTCGAATACTCGATCACTCTAT
CCGCCATTGGCATATTATTATTATTACCTGTATTTTTTATACCCTGAACC
CATTAAAAATGGGTATAAGGGTATATTGTATTTGTGCGAAATCCAAATGT
ATGTAACAGGCAGAAGGAAGCATCTCCGACCCCATAAAGTATATATATTC
TTAATCAGCATCAATAGCCGAGTCGATCTAGCCACTTCCGTCTGTCCGTC
CGTCCGTGCGTATGTATGAACGCAAGGATCTCAGAACCTATAAGAGCTAG
ATA