Annotation of *Drosophila mojavensis* fosmid 8

Priya Srikanth
Bio 434W
5.1.2007

Overview

High-quality finished sequence is much more useful for research once it is annotated. Annotation is a fundamental step in obtaining an enhanced understanding of functional features present in genomic sequence data. Here I delineate the annotation of a 45 kb region, fosmid 8, of the *Drosophila mojavensis* fourth chromosome (the "dot" chromosome). I finished the sequence of fosmid 8 earlier this year. The *D. melanogaster* annotated sequence is a valuable resource for comparison to *D. mojavensis* sequence. Though it has not been annotated as exhaustively as *D. melanogaster*, the *D. virilis* sequence annotation also proves useful, as it is phylogenetically closer to *D. mojavensis* than is *D. melanogaster*. Comparison of fosmid 8 to *D. melanogaster* and *D. virilis* sequences facilitates identification of functional features in the *D. mojavensis* genomic sequence. The *ab-initio* gene finder Genscan was used to guide investigation of putative functional features. Genscan's predicted protein-coding regions were analyzed using various genome browsers and databases, as well as BLAST and BLAT searches, defined below. These tools greatly assisted the annotation process. Final annotation of fosmid 8 reveals four features: two complete genes and two partial genes (Table 1, Figure 1). The two complete genes are ankyrin and H/ACA ribonucleoprotein complex subunit 1-like protein (GCR 101 snRNP); the two partial genes are a 5' partial gene of CG32000-PD containing one exon, and a 3' partial gene of rhomboid-5 containing three exons.

Table 1. Annotated features in fosmid 8.

Feature	Exons	Location	Strand	Gene	Content[1]
8.2	1	8341-8776	(-)	CG32000-PD	Partial 5': exon 1
8.3.1	9	15176-37282	(+)	Ankyrin	Complete gene
8.3.2	3	40643-41418	(+)	CG4038-PA	Complete gene
8.4	3	41663-44312	(-)	Rhomboid-5	Partial 3': exons 4-6

[1] (Partial 5') indicates that some 3' region of the gene is absent; (Partial 3') indicates that some 5' region of the gene is absent.

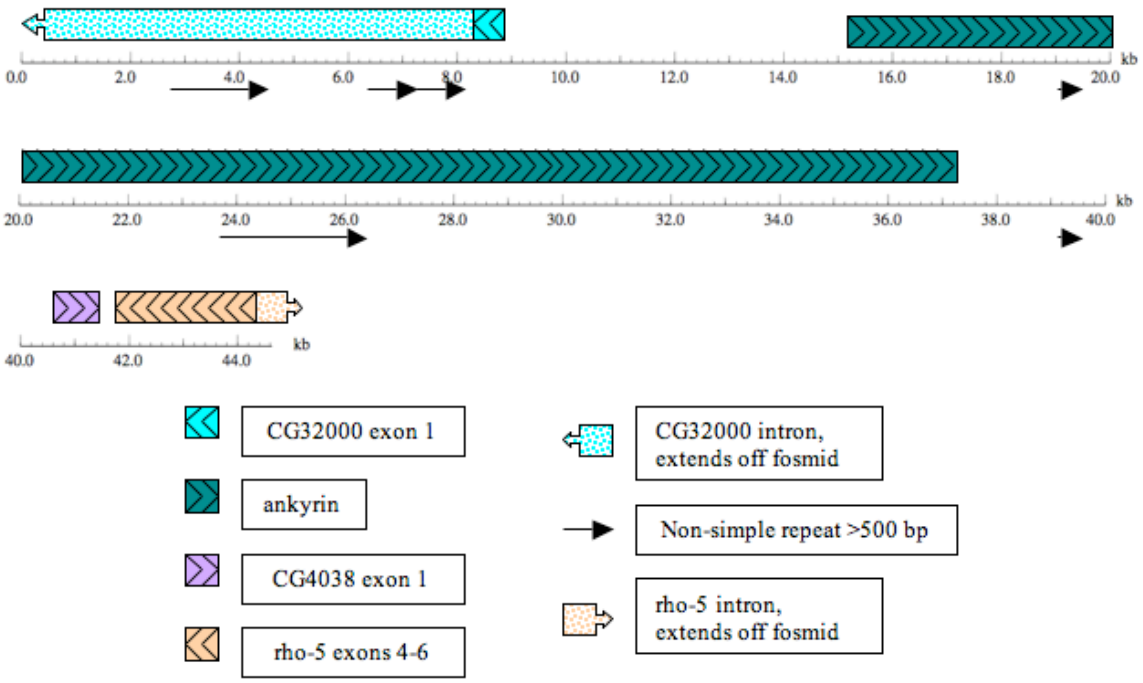

Figure 1. Annotated map of fosmid 8.

Definitions

BLAST: Basic Local Alignment Search Tool, an algorithm used to compare biological sequence data and find similarity.

blastp: Protein-protein BLAST, which compares a protein query sequence to a protein database of the user's choice. Examples of protein databases are: Swissprot, which contains curated[2] protein sequences; and nr, which is a database of non-redundant protein sequences that have been predicted and/or curated.

blastx: Translation-protein BLAST, which takes a nucleotide query, translates it in all 6 frames, and compares each translation to a protein database of the user's choice.

blastn: Nucleotide-nucleotide BLAST, which compares a nucleotide query sequence to a nucleotide database of the user's choice. Examples of nucleotide databases are: nt, a database of non-redundant nucleotide sequences that have been predicted and/or curated; EST, a database containing Expressed Sequence Tags – a sequence of transcribed mRNA that is usually a part of a transcription product, rather than the entire product; and Refseq, a database of mRNAs compiled from existing ESTs (using EST data to create models of whole transcription products) and experimentally-identified mRNAs of genes.

tblastn: Protein-translation BLAST, which compares a protein sequence query to a nucleotide database translated in all 6 frames.

tblastx: Translation-translation BLAST, which takes a nucleotide query, translates it in all 6 frames, and compares each translation to a nucleotide database translated in all 6 frames.

bl2seq: This tool compares two sequences provided by the user. The sequences are aligned using a BLAST tool specified by the user (i.e. blastn, blastp, tblastn, blastx, or tblastx).

---

[2] A curated sequence is one that has been manually reviewed by NCBI staff or collaborators. We used only curated sequences in our analysis of each feature to ensure that each sequence contained high-quality, reliable data.

BLAT: BLAST-Like Alignment Tool, which compares a nucleotide or protein query to a genome of the user's choice (e.g. human or chimp) and identifies locations in the genome that exhibit similarity to the query.

blastp search: Use of the blastp tool to find similarity of a query to proteins in the chosen database.[3]

Initial Data

RepeatMasker data show that fosmid 8 contains 37.30% GC content and 46.50% repeat density. Genscan predicted four protein-coding regions in fosmid 8 (Figure 2).

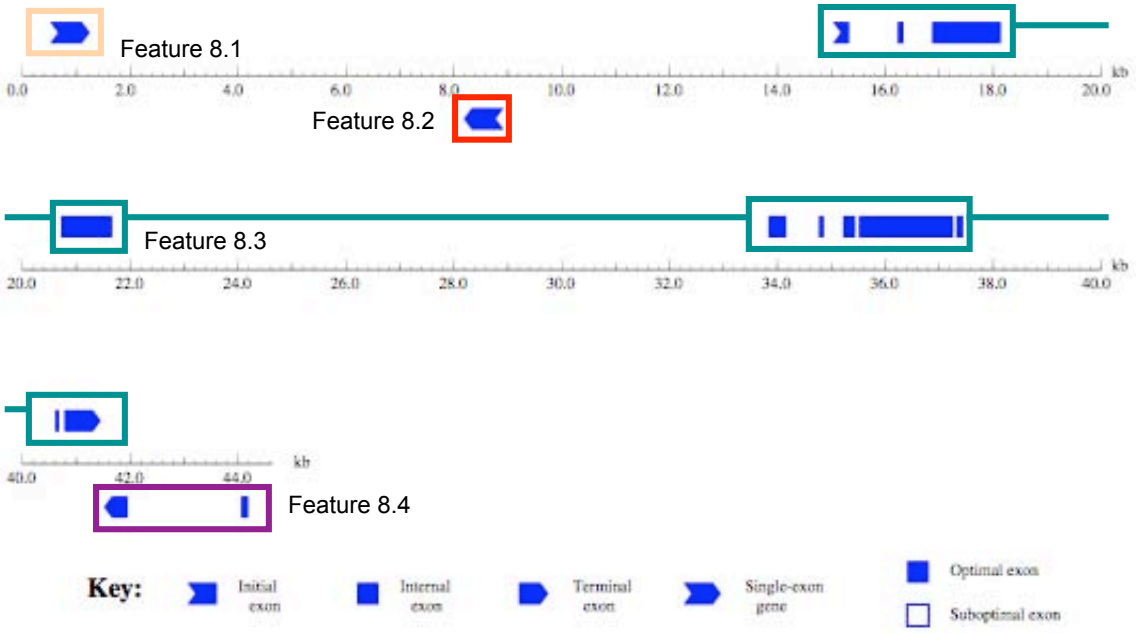

Figure 2. Genscan gene prediction map of fosmid 8.

The Goose genome browser[4] showed two known RefSeqs that aligned to fosmid 8 sequence, both of which fell within feature 8.3 (Figure 3). The peptide sequences of these genes, ankyrin and CG4038, were used in analysis of the fosmid, described below. Features 8.1 and 8.4 did not align to any known proteins or mRNAs.

[3] This "search" language will also be used with the other defined alignment tools, indicating use of the tool for its intended purpose.

[4] Available at goose.wustl.edu under "Genome Browser"

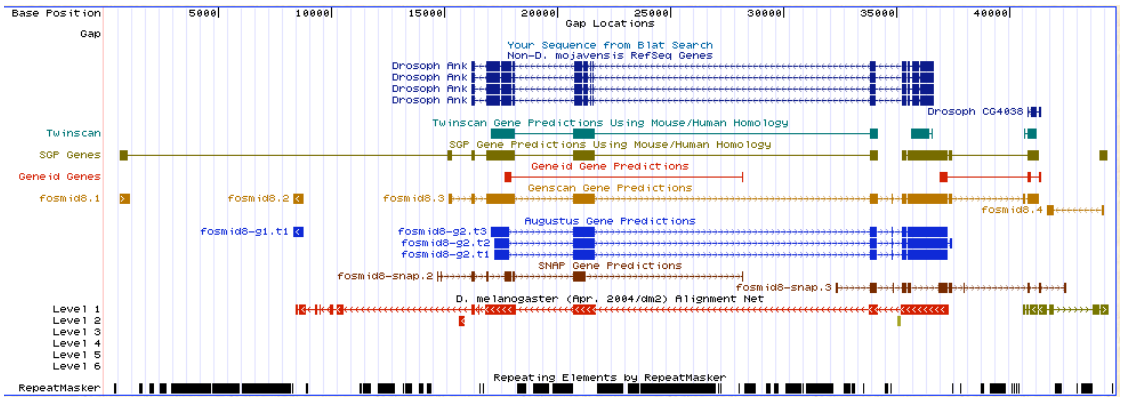

Figure 3. Genome browser view of fosmid 8.

Repeats

The library used by RepeatMasker to identify repeats in *D. mojavensis* fosmids was a combination of known *D. melanogaster* repeat sequences as well as novel *mojavensis* repeats that were incorporated into the database by Wilson Leung. RepeatMasker identified 10 repetitious features longer than 500 bp, 6 of which were not simple repeats (Table 2). Of the 44633 bp in fosmid 8, 20753 bp (46.5%) were masked. Further information on percent composition of each repeat class can be found in Figure 4.

Table 2. Long repeat regions. Non-simple repeats are in bold.

start (bp)	*end (bp)*	*length*	*repeat type*	*class/ family*
2921	**4672**	**1751**	**DMRT1C**	**LINE/R1**
6574	**7128**	**554**	**dmoj.1.77.centroid**	**DNA**
7342	**8114**	**772**	**dmoj.8.5.centroid**	**DNA**
19061	**19623**	**562**	**dmoj.1.12.centroid**	**DNA**
20056	20679	623	dmoj.7.10.centroi	Satellite
22052	22904	852	dmoj.7.10.centroi	Satellite
23125	23643	518	dmoj.41.71.centroi	Satellite
23684	**26114**	**2430**	**dvir.0.85.centroid**	**LINE**
26123	27019	896	dmoj.0.14.centroi	Satellite
39199	**39843**	**644**	**dmoj.35.53.centroi**	**d FB**

```
==================================================
file name: fosmid8.fasta
sequences:             1
total length:      44633 bp  (44633 bp excl N/X-run
GC level:         37.30 %
bases masked:      20753 bp ( 46.50 %)
==================================================
                number of      length   percentage
                elements*    occupied  of sequence
--------------------------------------------------
SINEs:                  0           0 bp     0.00 %
      ALUs              0           0 bp     0.00 %
      MIRs              0           0 bp     0.00 %

LINEs:                  2        4855 bp    10.88 %
      LINE1             0           0 bp     0.00 %
      LINE2             0           0 bp     0.00 %
      L3/CR1            0           0 bp     0.00 %

LTR elements:           2         543 bp     1.22 %
      MaLRs             0           0 bp     0.00 %
      ERVL              0           0 bp     0.00 %
      ERV_classI        0           0 bp     0.00 %
      ERV_classII       0           0 bp     0.00 %

DNA elements:          22        5648 bp    12.65 %
      MER1_type         0           0 bp     0.00 %
      MER2_type         0           0 bp     0.00 %

Unclassified:           1          79 bp     0.18 %

Total interspersed repeats:     11125 bp    24.93 %


Small RNA:              0           0 bp     0.00 %

Satellites:            17        4508 bp    10.10 %
Simple repeats:        13         621 bp     1.39 %
Low complexity:         8         342 bp     0.77 %
==================================================
```

Figure 4. RepeatMasker data summary.[5]

Methods

Annotation of each feature followed the same basic procedure. Generally, predicted peptide sequence from Genscan was used to perform a blastp search of *D. melanogaster* annotated proteins. The gene to which the predicted peptide had best alignment was then accessed in Ensembl (www.ensembl.org). The "peptide info" section in Ensembl provided the complete peptide sequence of each exon. One-by-one, each exon of the gene was aligned to the entire fosmid 8 nucleotide sequence by the bl2seq blastx tool. Alignments were evaluated for expected order and orientation in relation to the other exons of the gene and for low expect values. After identifying the exons present in fosmid 8, each exon was precisely annotated by locating the start and stop codons of the gene (if present) and the donor and acceptor sites for each exon where applicable. Start codons were identified either as the methionine that aligned to the beginning methionine of the *D. melanogaster* ortholog or the first methionine upstream of the location of the first exon alignment. Gene ends were identified as the first stop codon that occurred downstream of the last exon alignment in the proper frame.

Splice acceptor sites, present at the beginning of all exons but the first, were recognized by the presence of an "AG" on the correct strand near the beginning of the exon. Splice donor sites, present at the end of all exons but the last, were recognized by the presence of a "GT" on the correct strand near the end of the exon. Splice donor and acceptor sites of consecutive exons

[5] Most repeats fragmented by insertions or deletions were counted as one element by RepeatMasker.

must be in the same phase; i.e. they must not cause a frameshift in the coding sequence of an exon.[6] Identification of splice donor and acceptor sites was aided by the "Predicted Splice Sites" track in the Goose genome browser. This track highlighted high- and medium-confidence splice donors and acceptors by analyzing the bases around the putative site for other conserved bases often present at splice sites. By analyzing phase, distance from end/ beginning of exon, and the "Predicted Splice Sites" track of each putative donor and acceptor site, the best possible splice sites were identified. Gene models were tested using Simple Model Gene Checker (SMGC),[7] which verified that the annotated gene start, end, and splice site coordinates were correct. The Simple Model Gene Checker also provided the transcription and translation sequence for each exon of the gene model as annotated. Finally, the translated sequence of the entire gene model was used in a bl2seq blastp search against the orthologous *D. melanogaster* protein to confirm that these two protein sequences were indeed similar.

Feature 8.1

Using flybase.org, a blastp search of feature 8.1 was performed against the *D. melanogaster* annotated proteins database. This search yielded an alignment (with a high expect value of 0.034[8]) to the third exon of CG8545-PA, a three-exon gene on chromosome 2R of *D. melanogaster*. A bl2seq blastx search of exons 1 and 2 of CG8545-PA to fosmid 8 did not show any alignments in the appropriate order and orientation to agree with the exon 3 alignment. A BLAT search of the *D. mojavensis* August 2005 assembly, using the *D. melanogaster* CG8545-PA peptide sequence, showed only one hit, which did not correspond to fosmid 8's location in the 2005 assembly (identifiable by the presence of known genes ankyrin and CG4038). A BLAT search of *D. mojavensis* CG8545-PA to the March 2007 *D. mojavensis* assembly[9] showed no hits, indicating that the similarity found by the blastp search was not significant enough to be identified by BLAT search. Since the *D. mojavensis* ortholog of *D. melanogaster* CG8545-PA seems to be located at a position of the *D. mojavensis* genome not covered by fosmid 8, the blastp alignment is probably erroneous. This feature is also an unlikely gene candidate because of the location and orientation of feature 8.2. Feature 8.2 represents the first exon of a gene, CG32000-PD, which extends off the beginning of fosmid 8. CG32000-PD is oriented in the opposite direction as the alignment to CG8535-PA. It is unlikely that another gene is present in the same region as CG32000, especially as the syntenic *D. melanogaster* region containing CG32000-PD is on chromosome 4 and is not associated with CG8545-PA (on chromosome 2R) in any way. Thus, feature 8.1 is not a feature at all, but a misprediction by Genscan.

---

[6] For example, a splice donor that leaves a one-nucleotide overhang on the peptide sequence of the first exon must be paired with a splice acceptor site that leaves a two-nucleotide overhang on the peptide sequence of the second exon, making a codon and keeping the second exon sequence in frame. If the aforementioned donor site were paired with an acceptor site that leaves no overhang or a one-nucleotide overhang, the second exon peptide sequence would no longer be in frame, and the translated protein sequence would be altered.

[7] Written by Wilson Leung, available at http://goose.wustl.edu/~leung/genechecker/genechecker.html

[8] In genome-wide searches, significant e-values tend to be on the order of 1e-15 or lower.

[9] The March 2007 *mojavensis* assembly contains only the 13 fosmids finished by Bio 4342 students in Spring 2007.

Feature 8.2

A blastp search of *D. melanogaster* annotated proteins was performed using Genscan's feature 8.2 predicted peptide sequence. The best alignment this search produced was to exon 1 of CG32000-PD, a seven-exon gene on the fourth chromosome in *D. melanogaster*. CG32000 is an anonymous gene that is identified in Ensembl as a "probable cation-transporting ATPase." Using the bl2seq tool, each exon of CG32000-PD was compared to fosmid 8. Only the first exon produced a significant alignment, and any alignment to other exons was not properly oriented to agree with the orientation of the first exon's (better) alignment. A BLAT search of the *D. melanogaster* CG32000-PD sequence against the 2007 *D. mojavensis* assembly showed that the full CG32000-PD sequence was present in fosmid 7 (Figure 5). Fosmid 7 and fosmid 8 overlap by 12023 bases in opposite orientations. The first exon of CG32000-PD lies in this overlap region, while the other exons of CG32000-PD are located past base 12023 of fosmid 7[10], indicating they cannot be present in fosmid 8 (Figure 6).

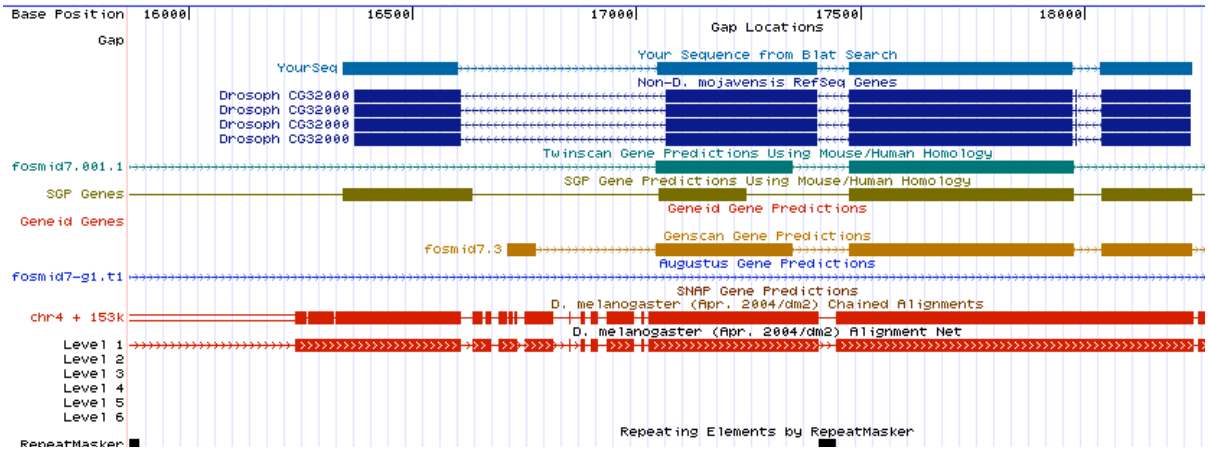

Figure 5. Alignment of *D. melanogaster* CG32000-PD protein ("Your Seq") and RefSeq sequences to *D. mojavensis* fosmid 7.

Figure 6. Location of CG32000-PD on fosmid 7 and fosmid 8.

CG32000 has seven isoforms, A – G. Of these, only isoforms D, E, and G have first exons located in the overlap between fosmids 7 and 8. The first exon of isoform E is identical to the first exon of isoform G, and will be referred to as isoform E/G exon 1. Exon 1 of isoform D was identified using a bl2seq blastx search of fosmid 8 against isoform D exon 1. Isoform E/G exon 1 has an alternate start codon, which is an internal methionine of isoform D exon 1 (Figure 7). This internal methionine is not present in the *D. mojavensis* sequence, nor are there any nearby methionines, indicating that this region of the *D. mojavensis* genome does not encode isoforms E or G, because it cannot translate the first exon of these isoforms. The above evidence indicates that feature 8.2 is a 5' partial gene of CG32000-PD, containing only the first exon at

[10] Exon locations were identified by BLAT search of each CG32000 exon sequence against the 2007 *D. mojavensis* assembly.

bases 8341-8776 (on the negative strand), and that this gene continues off the beginning of fosmid 8 into fosmid 7.

Figure 7. Comparison of *D. melanogaster* CG32000 isoform D exon 1 and isoform E/G exon 1 sequences to *D. mojavensis* fosmid 8 sequence.

The SMGC output protein sequence for the gene model in Table 1 was compared to exon 1 of *D. melanogaster* CG32000-PD using a ClustalW alignment (Figure 8). This alignment shows the modest degree of protein conservation between *D. melanogaster* and *D. mojavensis* CG32000-PD exon 1 and verifies the accuracy of the gene model.

```
isoDex1         MFASQSKACDSPSETVHLQSLPNHIEDLKFRNSDVETDDDLHFPGIAAKNKILKLSWWHS 60
feature2        MSANQSKNGGCLIAEVVPPELLLLQPLSESNVQLAEREQLLHLPRIATKNNLLKLNWWHS 60
                * *.***  ..    *   .*       : . . .* :: **:* **:**::***.****

isoDex1         PVIMHYVATESVQPKKS--DKVPSNKIKKVENNNTLVNGCSKTSARSVPLLKYNRPDQDG 118
feature2        PAQQLDIEASFLAPTRINIDANRTNTKPTEIANNTTNQSSLNTPVKSIPMPKEKHSTAK- 119
                *.*   : :. : *.:   *   :*.  .   ***  :.. :*..:*:*: * ::.  .

isoDex1         DSEENITSVLEPNVDEIYSKDSERLVDD- 146
feature2        QPNDRLSLNFDP-DEDCHLHNSQLLMKGR 147
                :.::.::  ::*  :: : ::*: *:..
```

Figure 8. ClustalW alignment of *D. melanogaster* CG32000-PD exon 1 ("isoDex1") to the peptide sequence of the feature 8.2 gene model ("feature 2"). The boxed residue is the internal M that functions as the start codon of isoforms E and G aligns.

Feature 8.3.1

A blastp search of the entire feature 3 peptide sequence against the *D. melanogaster* annotated proteins database yielded four alignments with equal scores to Ankyrin isoforms A, B, C, and D. This confirmed the alignment of a 5' region of feature 8.3 to known RefSeq ankyrin, isoforms A –D. Comparison of ankyrin isoform peptide sequences reveals that all isoforms have identical exon sequences, and only differ in untranslated sequence. Ankyrin (CG1651) contains nine exons and is on chromosome four in *D. melanogaster*. It is involved in cytoskeletal anchoring and signal transduction via actin binding, cytoskeletal protein binding, and receptor binding, and is a structural constituent of the cytoskeleton. Exon-by-exon bl2seq blastx searches of fosmid 8 sequence identified the locations of the orthologous exons 2-9 of ankyrin in *D. mojavensis* sequence. However, bl2seq blastx search did not yield any promising alignments to exon 1 of *D. melanogaster* sequence even with enormous expect values (e.g. 1e15). Instead of comparing *D. mojavensis* sequence to *D. melanogaster* annotation, *D. virilis* annotation was used as a model for comparison. Because *D. virilis* is phylogenetically closer to *D. mojavensis* than *D. melanogaster* is, one would expect the *D. virilis* ankyrin sequence to be more similar to the *D. mojavensis* ankyrin sequence than the *D. melanogaster* ankyrin sequence is. Ankyrin was annotated in *D. virilis* by Louis Lo in the spring 2006 Bio 4342 class. Using the exon 1 sequence he annotated, a bl2seq blastx search was performed against the fosmid 8 sequence. This search

yielded a good alignment, identifying the location of exon 1 in *D. mojavensis*. Gene and exon boundaries were identified as described in *Methods*.

The SMGC peptide output for the ankyrin gene model was compared to *D. melanogaster* ankyrin peptide sequence by a bl2seq blastp search. The search showed a very good alignment of the two sequences, with an expect value so low it was called 0.0 (Figure 9). Thus, feature 8.3.1 is the complete gene ankyrin, corresponding to all isoform peptide sequences (as they are identical). The region of fosmid 8 containing features 8.2 and 8.3.1 (0 kb to 37.3 kb) is syntenic with *D. melanogaster* chromosome 4 (ca. 137 kb to 156 kb). In fosmid 8 and in *D. melanogaster* chromosome 4, ankyrin and CG32000 are adjacent genes in opposite orientations.

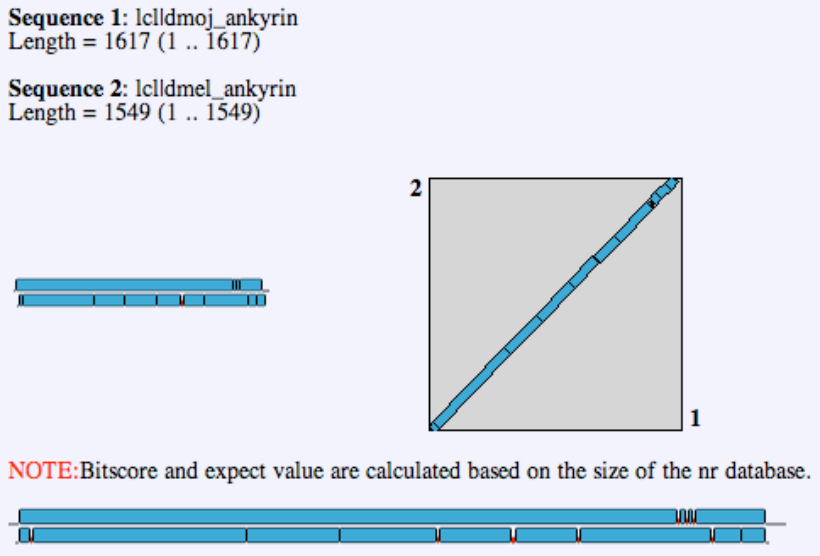

Figure 9. bl2seq blastp alignment of *D. mojavensis* (1) and *D. melanogaster* (2) ankyrin sequences.

ClustalW protein analysis

To investigate protein domain conservation, genes homologous to *D. mojavensis* ankyrin were identified by a blastp search of the *D. mojavensis* ankyrin protein sequence against the nr database. The search generated alignments to a variety of species: several mammals (rat, mouse, dog, opposum, macaque, human), insects (*Drosophila*, yellow fever mosquito, and *Anopheles gambiae* – the principal malaria vector), 2 bony fish (zebrafish and pufferfish), and 1 bird (wild chicken). Protein sequences of *D. mojavensis*, *Aedes aegypti* (yellow fever mosquito), opposum, human, dog, rat, chicken, pufferfish, and zebrafish were aligned using ClustalW. The ClustalW program generated a phylogram tree visualizing the phylogenetic distances and groupings of the query sequences (Figure 10).

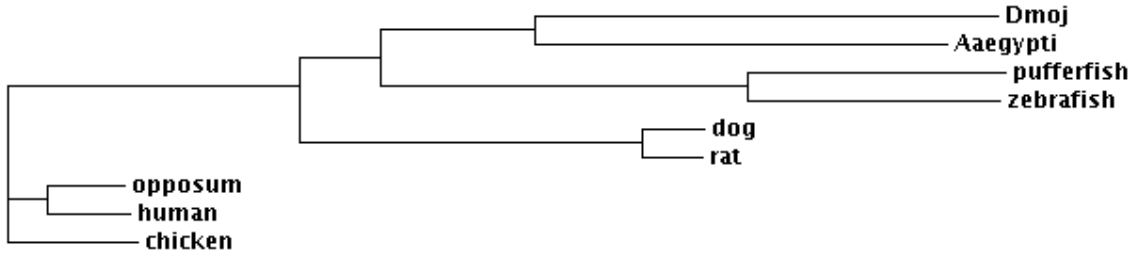

Figure 10. Phylogram tree of ClustalW sequences.

The aligned sequences showed remarkable conservation across such great species diversity. The C-terminal region of the protein showed little conservation among all the species, and the chicken sequence was unique in that it contained an insertion of ca. 1920 residues (relative to all other sequences) in the middle of the protein. Among the first 1400 residues of the *D. mojavensis* sequence, there is good conservation across all sequences. To investigate whether this conservation may be due to conserved functional domains, a blastp search was used. The blastp search of *D. mojavensis* ankyrin protein sequence to the nr database included an automatic search for conserved domains. The *D. mojavensis* ankyrin sequence contains a plethora of conserved domains of three main types: ankyrin repeats, ZU-5 domains, and death domains (Figure 11).

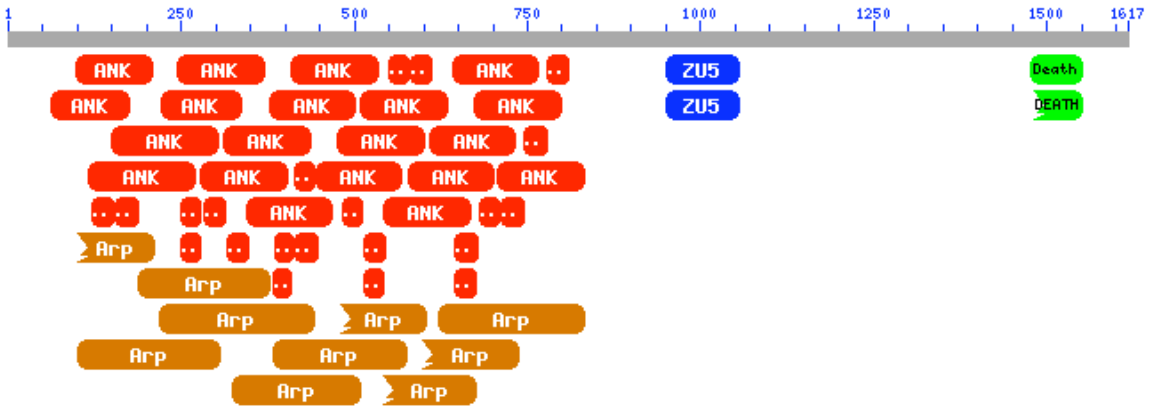

Figure 11. blastp search "conserved domain" output for *D. mojavensis* ankyrin sequence. "ANK" and "Arp" both indicate ankyrin repeat domains.

Ankyrin repeat domains mediate protein-protein interactions in a large variety of protein families. This structural repeat unit creates a specific secondary structure: two antiparallel helices and a beta-hairpin, which are often stacked in a superhelical arrangement of four consecutive repeats. The ankyrin repeat domains span residues 49-832 of ankyrin. The ZU-5 domain is present in ZO-1 and Unc5-like netrin receptors, but does not have a known function. This domain spans residues 951-1055 of ankyrin. The death domain is an alpha-helical domain found in proteins involved in apoptosis. In ankyrin, the death domains span residues 1458-1556 of ankyrin.

Many of the highly conserved regions of ankyrin identified by ClustalW analysis are within a conserved domain region. The highly conserved nature of the N-terminus of the protein may be explained by the presence of the functional ankyrin repeat domains at the protein's N-terminus. Conserved regions are found in both the ankyrin repeat and ZU-5 domain regions (Figure 12). However, no conservation is seen within the death domain of the *D. mojavensis* sequence, although many of the other ankyrin sequences also contained a death domain. It is possible that ClustalW did not align functionally homologous sequences in the death domain because the alignment score was maximized when these domains were not aligned, or because the sequence of this domain is not highly conserved.

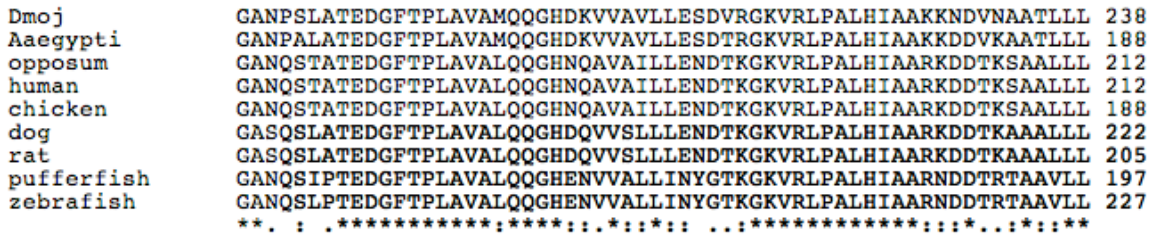

Figure 12.1. Conserved region within ankyrin repeat domain.

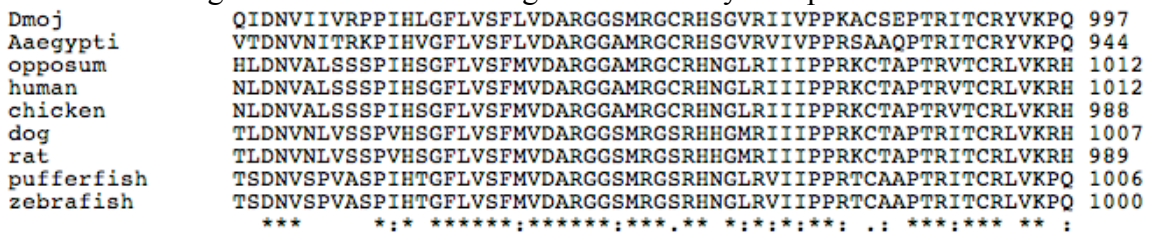

Figure 12.2. Conserved region within ZU-5 domain.

Figure 12. Conserved regions of ankyrin in ClustalW alignment.

ClustalW UTR analysis

ClustalW is also a potentially useful tool in locating untranslated regions (UTRs). By comparing the *D. mojavensis* nucleotide sequence upstream and downstream of a gene to the corresponding sequence from a closely related species, differential degrees of conservation can indicate the shift from non-functional DNA to a UTR to a coding region. This analysis was attempted with the 5' and 3' regions upstream and downstream of ankyrin, respectively.

A tblastn search of all *Drosophila* species' genome assemblies was performed at flybase.org using the *D. mojavensis* ankyrin exon 1 peptide sequence. The region from the alignment to *D. mojavensis* exon 1 to 1 kb upstream of exon 1 was extracted from genome assembly sequences of *D. virilis* and *D. grimshawi* (two species that are phylogenetically close to *D. mojavensis*). These nucleotide sequences were aligned using ClustalW. An alignment of sequences from all three species shows a drop-off in conservation from coding DNA to non-coding DNA, but no definitive change in conservation from 1 kb upstream of exon 1 to one base upstream of exon 1. A ClustalW alignment of only the *D. mojavensis* and *D. virilis* sequences does not show a change in conservation between coding and non-coding sequence, nor a significant change in conservation upstream of the gene. Regions containing the promoter sequence $tATA^{A}/_{T}A^{A}/_{T}$ (TATA box consensus sequence) were not conserved between *D. mojavensis* and *D. virilis*, even though *D. virilis* is phylogenetically closest to *D. mojavensis* of all the *Drosophila* species. Differences in conservation between non-functional and UTR sequence could not be identified in an alignment of *D. mojavensis* and *D. grimshawi* sequence.

Using a tblastn search as above, but with the *D. mojavensis* exon 9 peptide sequence as a query, homologous regions of *D. virilis* and *D. grimshawi* sequence to the 3' end of ankyrin were identified. Regions from the beginning of the exon 9 alignment to 1 kb downstream of the end of the exon 9 alignment were extracted from *D. virilis* and *D. grimshawi* genome assemblies. A ClustalW alignment of all three nucleotide sequences showed differential conservation in protein-coding versus non-coding regions, but not between UTR and non-functional sequence. Alignment of *D. virilis* and *D. mojavensis* sequences and alignment of *D. grimshawi* and *D. mojavensis* sequences yielded the same result. The alignment of *D. virilis* and *D. mojavensis* sequence shows conservation of a putative polyadenylation site, AATAAA, but it is over 800 bp away from the ankyrin stop codon (Figure 13). Polyadenylation signals are generally found

closer to the end of a coding sequence, but this does not rule out the identified sequence as a possible polyadenylation site.

```
virilis         TTTAATAATTTTAGCTTGGTGTTGAATTCAATAGGAAGCCCTCGGTTTGCACGCCTTGAA 2560
moj             TTTA-CAAAACAAACATATT--TAATATTTTTAGCTCAGCAATGAAATGAAA--CTTAAG 2599
                ****  **    * * *  *  * *  *   ***     *   *   ** *   *** *

virilis         TTCATCCGAGTAGAGTAATGCGTAATATTTTATTTCCAAAATAAAT-TGTTGAACTTTAA 2619
moj             TTTATTTGAGTA--TCAATGCGTAATGTGGTATTTCGAAAATAAAAATGTTGAACTTTAA 2657
                ** **  *****    ********** *  ****** ********  ************

virilis         TTGGATGTACTTGTATTATGATATTATAGTCAAAATTGGTATATTTCGATATAATCGTGG 2679
moj             TTA-ATGTAAGCA-ATAGCTACAT----GTTAA---TGTTGTGTTTACATTCTTCTTCAC 2708
                **  *****     **    * **    ** **   ** * * ***  **
```

Figure 13. Partial ClustalW alignment of *D. mojavensis* and *D. virilis* sequence downstream of ankyrin exon 9. The conserved putative polyadenylation site is highlighted.

Unfortunately, the phylogenetic distances between *D. mojavensis, D. virilis,* and *D. grimshawi* are not appropriate for definitively locating ankyrin UTRs in *D. mojavensis* by ClustalW analysis.

Feature 8.3.2

The sequence corresponding to ankyrin did not contain span the complete Genscan-predicted feature 8.3. A 3' section of feature 8.3, here called 8.3.2, aligned to the RefSeq gene CG4038-PA in the Goose genome browser (Figure 3). CG4038-PA is a one-isoform three-exon gene located on chromosome 2R in *D. melanogaster*. This gene has been named H/ACA ribonucleoprotein complex subunit 1-like protein (GCR 101 snRNP). The protein product is requited for ribosome biogenesis, as it is part of a complex that catalyzes pseudouridylation of rRNA. (Pseudouridine residues may stabilize the conformation of rRNAs.) Each exon of CG4038-PA was compared to fosmid 8 sequence by bl2seq blastx search. This method identified all CG40380-PA exons in fosmid 8, and exon boundaries were determined as described in *Methods*. Interestingly, a drastic intron shift occurred between *D. melanogaster* and *D. mojavensis*, shown in Figure 14. Exon 1 in *D. mojavensis* is 12 residues, while exon 1 in *D. melanogaster* is 64 residues.

```
moj             MAFGRPRGGSG---KGFRGSGGGGGR--GGGGGAFN----RSAGGGFGRGGSR--GGRGT 49
mel             MGFGKPRGGGGGGGRGFGGGGGGGGRGFGGGGGGRGGGGGRGGGGGFGRGGGGRGGGRGA 60
                *.**:****.*   :** *.******  *****. .    *..********.   ****:

moj             FDQGPPERVIALGNLSYICQNDIVCKVDIDDVPYFNAPIFLENKEQIGKIDEIFGTVRDY 109
mel             FDTGPPERVIPLGNYVYSCQNDLVCKVDIQDVPYFNAPIFLENKEQVGKIDEIFGTVRDY 120
                ** *******.***  * ****:******:****************:************

moj             SVSIKLSDNIYANSFKPNQTLFIDPGKLLPIARFLPKPPQTKGQKK-----RGGPSGG-- 162
mel             SVSIKLSDNVYANSFKPNQKLFIDPGKLLPIARFLPKPPQPKGAKKAFTNNRGGGGGGG 180
                *********:*********.********************.** **     *** .**

moj             ---VRGGRGG-AMGNRGGRGGGGFRGSSIR------GGGFNKGRGGAGGG---RGRW 206
mel             FGGRGGGRGGGGRGGGGGRGGGGFRGGAGRNGGGGGGGGFNRGRGGGGGGGGGRGRW 237
                      ***** . *. ********** .: *      *****:****.***   ****
```

Figure 14. ClustalW alignment of *D. melanogaster* and *D. mojavensis* (SMGC prediction) CG4038-PA protein sequences. Exon 1 is highlighted blue, exon 2 pink, and exon 3 green for each sequence.

The intron shifts in this protein sequence may have been facilitated by the presence of low-complexity sequence in the protein (i.e. poly-G sequences). The alignment of the complete *D. mojavensis* SMGC-predicted sequence and *D. melanogaster* CG4038 sequence confirms the accuracy of the predicted gene model, as it shows conservation between the sequences. Feature 8.3.2 is thus the full *D. mojavensis* ortholog of the *D. melanogaster* CG4038-PA gene.

Feature 8.4

A blastp search of the *D. melanogaster* annotated proteins database using peptide 8.4 as a query yielded an alignment to rhomboid-5 (CG33304-PA). Rho-5 is a one-isoform six-exon gene on *D. melanogaster* chromosome 2L. The rhomboid protein family consists of serine proteases. A bl2seq blastx search of fosmid 8 using each exon of rho-5 identified only rho-5 exons 4, 5, and 6 in fosmid 8. Exon boundaries were determined as previously described in Methods. A BLAT search of the August 2005 *D. mojavensis* assembly using the *D. melanogaster* rho-5 sequence generates only one alignment, corresponding to the location of fosmid 8 in the 2005 *D. mojavensis* assembly. Comparison of fosmid 8 to the 2005 *D. mojavensis* assembly shows that rho-5 begins off the end of fosmid 8 (Figure 15).

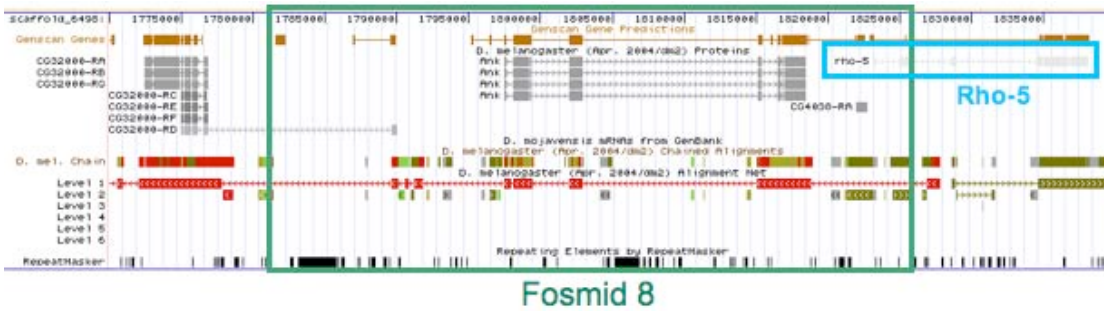

Figure 15. Genome browser view of August 2005 *D. mojavensis* assembly, comparing location of fosmid 8 to alignment of *D. melanogaster* rho-5.

This alignment confirms that rho-5 is a partial 3' gene in fosmid 8, which contains only exons 4, 5, and 6. A bl2seq blastp comparison of the SMGC-predicted partial rho-5 to exons 4-6 of *D. melanogaster* rho-5 shows a good alignment (expect = 3e-63) over most of the peptide sequence. The sequences diverge at the C-terminus (in the middle of exon 6), a phenomenon that is often evident in comparisons of orthologous protein sequences, and does not detract from the validity of the proposed gene model (Figure 16).

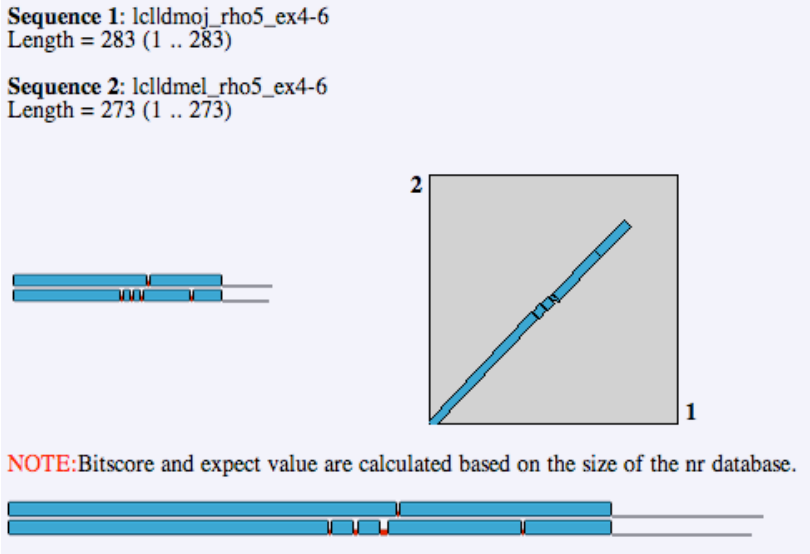

Figure 16. bl2seq blastp alignment of *D. mojavensis* (1) and *D. melanogaster* (2) sequences of exons 4-6 of rho-5.

Feature 8.4 is thus a 3' partial gene *D. mojavensis* ortholog of D. *melanogaster* rhomboid-5. The first three exons of rho-5 are present off the end of fosmid 8, as suggested by the 2005 *D. mojavensis* assembly genome browser view. Only the last three exons (4, 5, and 6) of rho-5 are present in fosmid 8.

Polymorphisms

In finishing fosmid 8, several polymorphisms were identified. None of these occur within a coding region, but several occur within intronic DNA (Table 3). These polymorphisms are unlikely to have a significant effect on genes found in fosmid 8.

Table 3. Polymorphism locations within introns in fosmid 8.[11]

Fosmid 8 position (bp)	bases	Gene location
23427-23431	deletion, T/A	Ankyrin intron (5-6)
23511	T/A	Ankyrin intron (5-6)
23522	G/A	Ankyrin intron (5-6)
23543	A/T	Ankyrin intron (5-6)
24868	A/G	Ankyrin intron (5-6)
25254	T/C	Ankyrin intron (5-6)
25284	A/G	Ankyrin intron (5-6)
25456	A/T	Ankyrin intron (5-6)
28812	extra T	Ankyrin intron (5-6)
42795	T/A	Rho-5 intron (5-6)

[11] (5-6) indicates that the intron occurs between exons 5 and 6

Synteny

The "*D. melanogaster* Chained Alignments" and "*D. melanogaster* Alignment Net" tracks in the Goose genome browser show that fosmid 8 is syntenic to regions of *D. melanogaster* chromosomes 4, 2R, and 2L. Synteny to each of these chromosomes is confirmed by the location of CG32000 and ankyrin on chromosome 4, CG4308 on chromosome 2R, and rho-5 on chromosome 2L. The respective order and orientation of the genes are shown in Figure 17. Because the true direction of fosmid 8 cannot be known until the *D. mojavensis* sequence is assembled, the orientation of fosmid 8 was flipped to minimize inversion events. This leaves just one inversion, which occurs in the 3' partial rho-5 gene.

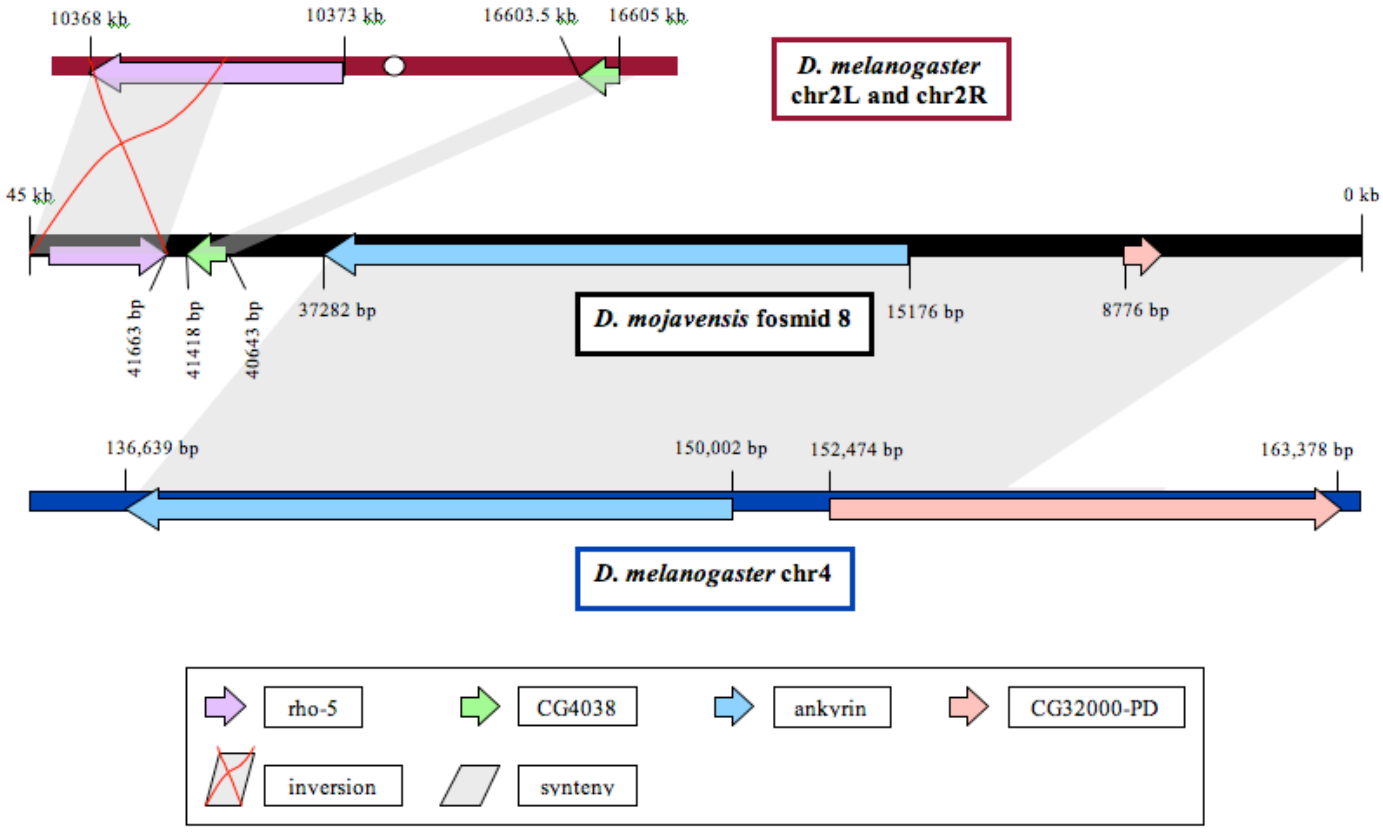

Figure 17. Synteny of *D. mojavensis* fosmid 8 to *D. melanogaster* chromosomes 4 and 2.

Breakpoints between regions syntenic to different chromosomes could not be identified by a blastn search using sections of fosmid 8 sequence (between genes from different *D. melanogaster* chromosomes) against the *D. melanogaster* genome assembly. The breakpoints were unidentifiable because the nucleotide sequences of non-coding regions of *D. melanogaster* and *D. mojavensis* are too divergent to show significant similarity of *D. mojavensis* sequence to one *D. melanogaster* chromosome over another. The gap between regions syntenic to chromosome 4 and chromosome 2R contains a repeat of unknown origin, which may have facilitated a transposition event. The gap between regions syntenic to chromosomes 2R and 2L is very small (ca. 242 bases) and does not contain any repeats, making the apparent transposition event between *D. mojavensis* and *D. melanogaster* quite remarkable. The gap between ankyrin and CG32000, which are syntenic to *D. melanogaster* chromosome 4, is about 4 kb larger in *D. mojavensis* than in *D. melanogaster* (6399 bp and 2471 bp, respectively). This region in *D.*

mojavensis contains a number of DNA elements and unknown repeats, which may explain the larger size of the intergenic region in *D. mojavensis*. Overall, there is little synteny among the genes in fosmid 8 to one region of the *D. melanogaster* genome.

Conclusion

Use of several genome browsers, databases, and alignment tools is essential to completing the annotation process. Information from Ensembl, FlyBase, NCBI, Swissprot, and the UCSC and Goose genome browsers was utilized heavily throughout annotation. With use of such tools, as well as use of existing annotation of phylogenetically close species (i.e. *D. virilis*), the process of annotating a novel sequence has become much faster than it once was. Identification of exon 1 of feature 8.3.1 may not have been possible without the use of previously annotated *D. virilis* sequence. This example emphasizes the value and importance of existing annotations of closely related species in annotating a new species' genomic sequence. As more *Drosophila* species' sequences are finished and annotated, the annotation process will continue to become easier and faster.

Appendix

Sequences

Feature 8.2, CG32000-PD exon 1 (-)

exon	8341	8776	-	0
CDS	8341	8776	-	0
start_codon	8339	8341	-	.

>Dmoj2_fosmid8_CG32000-RD_NM_166728_cds
ATGTCTGCCAACCAAAGTAAAAATGGCGGATGCCTTATTGCAGAAGTTGTGCCCCCAGAGCTTTTGTTACTGCAGCC
CTTATCGGAAAGCAACGTTCAGCTGGCAGAACGCGAACAGTTGCTGCACTTGCCGCGTATTGCTACCAAGAACAATC
TGTTGAAGCTAAATTGGTGGCATTCGCCCGCACAGCAATTGGACATCGAAGCCTCATTCCTGGCCCCCACGCGAATC
AATATTGATGCAAATCGGACCAACACAAAGCCAACTGAAATTGCCAACAACACGACCAATCAAAGCAGCCTTAATAC
ACCTGTAAAATCAATTCCTATGCCTAAGGAAAAACATTCAACTGCTAAACAGCCGAATGATAGGTTGAGTCTCAATT
TCGATCCTGATGAAGATTGTCATCTACATAATTCCCAGTTGTTAATGAAGGGTA

>Dmoj2_fosmid8_CG32000-RD_NM_166728_pep
MSANQSKNGGCLIAEVVPPELLLLQPLSESNVQLAEREQLLHLPRIATKNNLLKLNWWHSPAQQLDIEASFLAPTRI
NIDANRTNTKPTEIANNTTNQSSLNTPVKSIPMPKEKHSTAKQPNDRLSLNFDPDEDCHLHNSQLLMKG

Feature 8.3.1, CG1651 ankyrin isoforms A-D (+)

exon	15176	15331	+	0
exon	16230	16337	+	0
exon	16869	17982	+	0
exon	18040	18143	+	0
exon	20716	21663	+	0
exon	33834	34164	+	0
exon	34769	34862	+	0
exon	35219	35443	+	0
exon	35512	37282	+	0
CDS	15176	15331	+	0
CDS	16230	16337	+	0
CDS	16869	17982	+	0
CDS	18040	18143	+	2
CDS	20716	21663	+	0
CDS	33834	34164	+	0
CDS	34769	34862	+	2
CDS	35219	35443	+	1
CDS	35512	37282	+	1
start_codon	15176	15178	+	.
stop_codon	37283	37285	+	.

>Dmoj2_fosmid8_Ank_NM_175925_cds
ATGACACTAGACGATCCGCAAAACGATATACAAACGAATGCTCAGAGAAACGAAACAGCAAAAAAACGAGCACAGTT
ACAAACAAATCAGAATTGTCAGCCAACGGAAAGTGATAAATTGGATAAAAGCAATATAACGGTAAATCATCAAAAAC
AGAATGATGCTACCATATCATTTTTACGAGCTGCTCGAAGTGGAGATCTGGGCAAGGTACTGGAATTTATAGATAGT
GGACTTATAACGAACATAAACACGTGCAATGCGAATGGGTTAAATGCCCTGCACCTTGCGGCCAAGGATGGATATGT
AGACATATGCAACGAACTCCTTAAGCGCGGAATTTCGGTAGACAGTGCGACGAAAAAAGGAAACACTGCCCTACATA
TAGCTTCGTTGGCTGGCCAGCAGCAAGTGATAAAACAACTAATCCAATACAATGCGAATGTCAATGTGCAGTCGCTA

AACGGCTTTACACCTTTATATATGGCTGCACAGGAGAATCATGATAGCTGTTGCCGTCTGCTACTAAGTAAGGGTGC
CAATCCATCCCTTGCTACTGAAGATGGTTTCACACCTCTCGCTGTGGCCATGCAACAAGGTCATGATAAAGTGGTTG
CCGTGCTCCTCGAAAGCGATGTTCGTGGCAAAGTGCGTTTGCCGGCCTTACATATTGCAGCGAAAAAGAATGACGTC
AATGCGGCTACATTGCTGTTAAGGCATGATAAAAACGCCGATATTGTATCGAAGTCAGGCTTTACACCACTTCACAT
TGCTGCCCATTATGGGAATGTGGACATTGCTAACTTGTTGCTCGAATGCGGCGCCGATGTCAACTACGCGGCAAAAC
ATAACATAACACCGTTGCATGTTGCCTGTAAATGGGGAAAGGCAGCAGTATGCAGACTTTTGCTCGCAAAAGGAGCC
CGAATTGATGCCATAACTCGAGATGGACTTACACCATTGCATTGTGCGTCGCGATCCGGACATGTCGAGGTCATACA
GCTCTTACTCTCTCAACATGCTCCAATATTGTCGAAGACAAAAAATGGATTATCTGCTCTTCATATGTCAGCGCAGG
GAGAACATGACGAGGCAGCCCGTTTGCTACTCGATTACAAGGCGCCCGTTGACGAAGTGACAGTTGACTATTTGACA
GCTCTCCATGTGGCCGCCCACTGCGGCCATGTGCGAGTGGCTAAATTATTGTTGGACTATGGAGCGAATCCGAACTC
TCGAGCACTCAATGGATTTACGCCTTTGCACATTGCCTGCAAGAAAAATCGCATCAAGGTGGCGGAATTGCTGATCA
AACATGGCGCCAATATACGCGCCACCACTGAATCCGGTTTAACGCCGTTACATGTAGCCAGTTTTATGGGCTGCATG
AATATTGTTATTTACTTGTTGCAACACGATGCTAGCCCCGACATGCCTACCATACGCGGCGAAACGCCTCTTCATTT
GGCTGCACGCGCAAATCAGACTGACATTATTCGAATATTGCTTCGAAATGGAGCACAAGTGGATGCCGTTGCTCGCG
AGGGTCAAACGCCGCTGCACGTCGCCGCTCGTCTGGGCAATATTGATATTATAATGCTAATGTTGCAGCATGGTGCA
CAGGTGAATGCTTGCACTAAGGATATGTATACTGCCTTGCATATTGCTGCGAAGGAGGGCCAGGAGGAAGTATGCCA
GTTACTAATTGAACATGGGGCACAGTTAGAATCTGAAACAAAAAAAGGATTCACCCCACTGCATCTGGCTAGCAAGT
ATGGTAAGGCAAATGTGGCCGACATGTTGATTAAAAAGGGAGCCGTCATTGATTGTCAAGGTAAAAACGATGTGACA
CCTTTGCATGTTGCCACTCATTATGATCACCAGCCTGTGGTGTTACTGCTACTAGAGAAAGGCGCTTCTGCCCAAAT
AACAGCACGCAATGGCCATAGTGCCCTTCATATCGCTGCTAAGAAAAATAACCTAGAGACAGCACAGGAACTATTGC
AGCACGGTGCCGATGTTAATGCAACAAGTAAGTCGGGGTTTTCACCGGTTCATTTGGCTGCGCAGGAGGGACACGTT
GATATGGTTCAATTGCTACTTGAACAGAGTGCCAATGCCAACATATCAGCGAAAAATGGTTTGACCCCTTTGCACTT
GGCTGCTCAGGAAGGACACGTGCAGGTTTCACAGACCTTACTCAATTACGGAGCCTACATTTCAGAGCGCACTAAAG
CTGGCTATACGCCACTTCATATTGCTGCGCACTATAACCAAATAAATGAAATTAAATTTCTTCTCGAAAACGATGCG
AATATTGAAATGACAACAAATGTTGGTTATACTCCCCTTCATCAGGCAGCACAACAAGGTCATACGATGGTAATCAA
TCTACTTCTTCGACACAAAGCCAACCCCGATGCGTTAACAAATAATGGTCAAACAGCCCTTAATATCGCTCATAACT
TGGGCTACGTAACAGCTGTTGAGACACTTAAGATTGTGACGGAAAAATCCGTAATTAATACCACAACCGGAGTTCTT
GAGGAAAAATACAAAGTTGTCGTACCGGAATTTATGCACGAAACGCTACTTTCTGATTCCGACGACGAGGGCGGAAA
CGAAGCTTTGGATCACAATCAATACAAATACATGGCGACCGACGACTTAAAAGCCACCAATGATCATGATAACCATA
ACTTTAATACAACTGATCCGGAGATCGATCGATTGGATGGACTAGTCGGTCGAAATATTGAAAAAGAATTAACCAGA
ACAGATGAAACAATTTTAAATACATCTCCGGTGGAACGACAGATTGATAATGTGATTATAGTTCGGCCGCCCATTCA
TCTGGGATTTCTTGTGTCCTTCCTGGTGGACGCGCGTGGCGGTTCAATGCGCGGTTGCCGACACAGTGGTGTTCGGA
TAATTGTACCTCCCAAGGCGTGTTCGGAGCCAACCCGTATAACTTGCCGCTATGTCAAGCCGCAGCGAGTGGCTAAT
CCGCCGCCATTGATGGAGGGTGAAGCTTTGGTTAGTCGCATATTGGAAATGTCTCCAGTAGAAGGAAAGTTTCTAAG
CCCGATTGTATTGGAGGTGCCACATTTTGGATCGCTGCGCGACAAAGAGCGTGAAATAATCATTTTAAGATCGGACA
ATGGCGAAAGCTGGCGCGAACATAGTGTCTATGAAGACGAAGGTAATTTACTCGAAACCTTAAAAGAAACGATGGCA
GCTGATGTTAATCCATTGGAAGATTTACACACCAGTCGAATTATAAGAATCGTCACACAAAATGTGCCACATTTTTT
TGCGGTGGTATCTCGCGTTCGCCAGGAGGTTCACGCAATAGGACCAGATGGTGGCACTGTATCATCTATAGCTGTGC
CCCAGGTACAGTCCATATTTCCACCCCATGCATTAACAAAGAAGATTCGAGTTGGTCTTCAGGCACAGCCAGTGGAC
TTAATTGGTTGTTCCAAGCTACTAGGTCAGGGCGTTGCAGTATCTCCGGTCGTTACCGTCGAGCCACGACGGCGTAA
ATTTCACAAGGCCATCACATTAAGCATTCCGGCGCCAAAAACTTGCAATCAAGGCATGGTAAACTCACCATATAACG
CCACGAATGGAAACGTAAGTGCCCCTACTCTGCGACTGTTGTGCTCTATAACAGGTGGACAAAATCGTGCTGTATGG
GAGGACGTCACTGGTTCGACGCCTTTGGCATTTGTAAAGGATAGCGTTAGCTTTACAACAACGGTATCGGCACGTTT
TTGGCTAATGGATTGTCGCAATGTTGCGGACGCCGGTCGCATGGCAACCGAACTTTATTCGTATATGGCAAAGGTGC
CGTTTATTGTAAAATTTGTAGTATTTGCAAAACAAATATCCGCAACGGAAGCAAAATTATCTGTCTTTTGTATGACA
GATGATAAAGAGGATAAGACGCTCGAACAGCAAGAGTATTTTGCCGAAGTAGCAAAGAGTCGGGATGTTGAGGTATT
GCAAGATCAAAATATTTACCTAGAGTTTGCCGGAAATTTGGTGCCGGTTTTAAAATCGGGCGAACAACTGAACACAA
AGTTTCAAGCATTTCGTGAAAATCGCCTCTCTTTCATAGTTTACATTAAAGATCAAGAACAACCACATGCACGTATT
TCCTTTATGAGCGAGCCGAAAGTAACATCAGGAGAGGCGCCGTTGCGGCCAATTTGTACGCTGAATGTATCACTTGA
ATCGCAAAAATTAAATCGAGTTAAACTACATCCGAATGGCATCAATACTAATGGAAATAACAATACGGATCATGACC
TTAACTACAAAGATTTAATCAATGCCGAAATAAAATGTAATACAGAATCAAGAGGAAATTCAAATATAAATCCGTTA
GAAGATGGAATACAGCGTGCCAATATCCGTTTATCAGACATATCGAATCTGCTGGGTAGTGACTGGCCGCAGCTTGC
TGAAGAATTGGGTGTACCCGAGACAGACATCAACATGGTTGAAACTGAGTACGCCGATCAACCTGTTGCACAACAGG
GCTTGGTCATGCTACGTCTATGGCTAAAACAAGAGGGCAATCGCGCCACTGGTAATGCCTTGTCGCAGGCTTTAAAC
AAAATCAGACGTGAAGACGTTGTTGAAAAGTGCATCTTTAACTTAGAACTGGTCACAGATCAACTGGAGCGTGGCTT
GGCCACGACGCGAATGCAGCATGACCACACTTTGATGCAATATCTAGCGGATGGCCTGAATGAAACATTGAACATTG
AACAGCAAACTAACGAAGACGAATTGTGTCGAAGCAATGAAAAATCTAATAATGGTAAGAGACAAGAAATATTCAAT

>Dmoj2_fosmid8_Ank_NM_175925_pep
MTLDDPQNDIQTNAQRNETAKKRAQLQTNQNCQPTESDKLDKSNITVNHQKQNDATISFLRAARSGDLGKVLEFIDS
GLITNINTCNANGLNALHLAAKDGYVDICNELLKRGISVDSATKKGNTALHIASLAGQQQVIKQLIQYNANVNVQSL
NGFTPLYMAAQENHDSCCRLLLSKGANPSLATEDGFTPLAVAMQQGHDKVVAVLLESDVRGKVRLPALHIAAKKNDV
NAATLLLRHDKNADIVSKSGFTPLHIAAHYGNVDIANLLLECGADVNYAAKHNITPLHVACKWGKAAVCRLLLAKGA
RIDAITRDGLTPLHCASRSGHVEVIQLLLSQHAPILSKTKNGLSALHMSAQGEHDEAARLLLDYKAPVDEVTVDYLT
ALHVAAHCGHVRVAKLLLDYGANPNSRALNGFTPLHIACKKNRIKVAELLIKHGANIRATTESGLTPLHVASFMCMN
IVIYLLQHDASPDMPTIRGETPLHLAARANQTDIIRILLRNGAQVDAVAREGQTPLHVAARLGNIDIIMLMLQHGAQ
VNACTKDMYTALHIAAKEGQEEVCQLLIEHGAQLESETKKGFTPLHLASKYGKANVADMLIKKGAVIDCQGKNDVTP
LHVATHYDHQPVVLLLLEKGASAQITARNGHSALHIAAKKNNLETAQELLQHGADVNATSKSGFSPVHLAAQEGHVD
MVQLLLEQSANANISAKNGLTPLHLAAQEGHVQVSQTLLNYGAYISERTKAGYTPLHIAAHYNQINEIKFLLENDAN
IEMTTNVGYTPLHQAAQQGHTMVINLLLRHKANPDALTNNGQTALNIAHNLGYVTAVETLKIVTEKSVINTTTGVLE
EKYKVVVPEFMHETLLSDSDDEGGNEALDHNQYKYMATDDLKATNDHDNHNFNTTDPEIDRLDGLVGRNIEKLTRTD
ETILNTSPVERQIDNVIIVRPPIHLFLVSFLVDARGGSMRGCRHSGVRIIVPPKACSEPTRITCRYVKPQRVANPPP
LMEGEALVSRILEMSPVEGKFLPIVLEVPHFGSLRDKEREIIILRSDNGESWREHSVYEDEGNLLETLKETMAADVN
PLEDLHTSRIIRIVTQNVPHFFAVVSRVRQEVHAIGPDGGTVSSIAVPQVQSIFPPHALTKKIRVGLQAQPVDLIGC
SKLLGQGVAVSPVVTVEPRRRKFHKAITLSIPAPKTCNQGMVNSPYNATNGNVSAPTLRLLCSITGGQNRAVWEDVT
GSTPLAFVKDSVSFTTTVSARFWLMDCRNVADAGRMATELYSYMAKVPFIVKFVVFAKQISATEAKLSVFCMTDDKE
DKTLEQQEYFAEVAKSRDVEVLQDQNIYLEFAGNLVPVLKSGEQLNTKFQAFRENRLSFIVYIKDQEQPHARISFMS
EPKVTSGEAPLRPICTLNVSLESQKLNRVKLHPNGINTNGNNNTDHDLNYKDLINAEIKCNTESRGNSNINPLEDGI
QRANIRLSDISNLLGSDWPQLAEELGVPETDINMVETEYADQPVAQQGLVMLRLWLKQEGNRATGNALSQALNKIRR
EDVVEKCIFNLELVTDQLERGLATTRMQHDHTLMQYLADGLNETLNIEQQTNEDELCRSNEKSNNGKRQEIFN

Feature 8.3.2, CG4038-PA: H/ACA ribonucleoprotein complex subunit 1-like protein (GCR 101 snRNP) (+)

exon	40643	40679	+	0
exon	40778	41254	+	0
exon	41315	41418	+	0
CDS	40643	40679	+	0
CDS	40778	41254	+	2
CDS	41315	41418	+	2
start_codon	40643	40645	+	.
stop_codon	41419	41421	+	.

>Dmoj2_fosmid8_CG4038-RA_NM_057695_cds
ATGGCTTTTGGACGTCCTCGTGGAGGTAGTGGCAAGGGCTTTCGTGGATCTGGAGGTGGTGGAGGGCGAGGAGGTGG
TGGTGGTGCCTTCAACAGATCTGCAGGTGGAGGATTCGGCAGGGGCGGATCTCGTGGTGGTCGTGGGACTTTTGATC
AGGGCCCACCCGAACGCGTTATTGCATTGGGAAACCTCAGTTATATATGCCAGAATGATATAGTTTGTAAAGTAGAT
ATAGACGATGTACCATATTTCAATGCACCAATATTTCTGGAAAATAAGGAGCAGATCGGAAAAATTGACGAAATTTT
CGGAACAGTTCGTGACTACTCTGTTTCCATAAAGCTGTCTGATAATATATATGCAAATAGTTTTAAGCCGAATCAAA
CATTGTTCATTGACCCTGGGAAGCTGCTGCCGATTGCAAGATTTCTGCCAAAGCCACCACAGACAAAAGGTCAAAAA
AAGAGAGGCGGTCCTAGTGGTGGTGTAAGAGGTGGACGTGGTGGAGCCATGGGAAATCGTGGAGGACGTGGCGGCGG
TGGATTTAGAGGTAGCTCAATTCGTGGAGGCGGATTTAATAAAGGACGTGGTGGTGCAGGCGGAGGACGTGGGCGTT
GG

>Dmoj2_fosmid8_CG4038-RA_NM_057695_pep
MAFGRPRGGSGKFRGSGGGGGRGGGGGAFNRSAGGGFGRGGSRGGRGTFDQGPPERVIALGNLSYICQNDIVCKVDI
DDVPYFNAPIFLENKEQIGKIDEIFGTVRDYSVSIKLSDNIYANSFKPNQTLFIDPGKLLPIARFLPKPPQTKGQKK
RGGPSGGVRGGRGGAMNRGGRGGGGFRGSSIRGGGFNKGRGGAGGGRGRW

Feature 8.4, CG33304-PA, rhomboid-5 exons 4-6 (-)

exon	44055	44312	-	0
exon	43678	43977	-	0

exon	41663	41953	-	0
CDS	44055	44312	-	0
CDS	43678	43977	-	0
CDS	41663	41953	-	0
stop_codon	41952	41950	-	.

>Dmoj2_fosmid8_rho-5_NM_205957.1_cds
GTGGGACCCTCAGCATCTCTTTGTGGTGTAGTGTCGTCGCTGGTTGCACTCCTCTTATGGATGCATTGGAAGCATGT
GAAAAAGCCATACATGTCATTATTTAAGATGTTGCTTTTGACAACAGTTCTTTTCGGAATTGGTACACTGCCGTATC
AACTAAATTTTGCTGGACTTCTGGCTGGGTTTGGATGTGGTACATTTCTAACAATAGCGCTGGTACCATTTGCGTCT
TTCACCAAGTACAGACGCAGGAAAAAGATAAATCTTATTTGGACATGCCTGCTGTTCCATTTCTTTATGTATATGAC
TCTGGCGACAACTTTTTACATTTACCCCAGCGAATTTAACACGTTTAGCTTTGTGGATGATATATTCGGGAGCAATA
ATGGCAACAAATACATTGTTCCTACAAACAGTAATATAGGTCAGCACCATGGCGAGGTAAGCAGCACAACCCGCAGA
TATTCCGAGACACAAAAGCCTCAATATTATTATCACCATCACTCGGAAGACATAATCCGGAACACCGTAGCATATCC
TGAAGTTAATACTAAATCGCACATCTATGCAGAAACTAAGGCGCGTCGGCAAAAAGATTTTGATCTCTGGCAAGAAG
GACTTTACCCTCGATCATTTCGCTACAGTTCCAACTATAGCGATCGCATATATAACAAAATTCAATCGAATCTACTG
TCGGAAACAAATCTATTAAGCCACAAGACACAAGACTCGCTTGTCCCATCTTCGACCCGAAAGAAAAGCATTTCTCG
AAGTATAAAGGAGGCGTTGGATCTTAATACCAACACTAAACATGTTGAAAATATAAATGATAAAAATTTAAAGGGCC
TC

>Dmoj2_fosmid8_rho-5_NM_205957.1_pep
VGPSASLCGVVSSLVALLLWMHWKHVKKPYMSLFKMLLLTTVLFGIGTLPYQLNFAGLLAGFGCGTFLTIALVPFAS
FTKYRRRKKINLIWTCLLFHFFMYMTLATTFYIYPSEFNTFSFVDDIFGSNNGNKYIVPTNSNIGQHHGEVSSTTRR
YSETQKPQYYYHHHSEDIIRNTVAYPEVNTKSHIYAETKARRQKDFDLWQEGLYPRSFRYSSNYSDRIYNKIQSNLL
SETNLLSHKTQDSLVPSSTRKKSISRSIKEALDLNTNTKHVENINDKNLKGL

ClustalW UTR sequences

D. mojavensis ankyrin exon 1 + 1 kb upstream
>dmoj_ankyrin5'
TGGTACACAAAAAGAGTTATCACTCTTAATTTAGTAGAGAGTAGCTGTAAGAGTCGTAAA
CCTATTAGGGGTTCTAAAATACAATTTTTACTGTATTTGTTTTAGGTGAGGAAACTGCCA
AGAAGGTTAGTTCGATGAATTATTACGCATATTGTTTGATGATTCGTGAATCTGACAATT
ATCTATCTAAATATTTTTTTGTATTATTGTATTACGTTATATGTATATATAGAGACTGTT
AGGCGGTACGAAGCTCGCCGGAACAACTAGTATATAATATATAACCAAAATGATCTTAAA
GCTTTAATTATAAAATATACTTAGTCAGTTAGTCATAGATCCATTTTCTTTCCATTTTTG
TCAACAACAAATATATTTTATAGTAATTCATTGAAACTAAAATAATAACAAAAGCAATCG
CCTCCATGGCACTAAGGCAAAACCTATTGGTTCTTATTCAGAAAAGTTAGAAATACAATT
TTTATAAAATTTTGCAAAATTTCAAAAATGAGCTCTTACGCAGTCAAAACCAACTTCGAC
AGTAACCATTTTTTCGTTATATCGATACTAAGGAAAACCAATATCGATGTTTTTCCAAAA
TTCCAAAATATCGACTATAACGATACCGATCCGATATTTTTGCAGCACTATGTCTGATTG
AGATTCGTATATGCATTGAAATTCACTCGCACAAGATTAAGGTAGGTGGAGAAATGACGT
CATAGTAAGTGTAATAATTGCCTTGCATAAAAACATTAATATTAAGCATACACCAGTAAG
GTAATAAATTATGTGCATATGTATCATTGTATTCTAAACATAAATTTACGTTATGCATCC
AATAGGCGATACATTTTTTACTACTGATTGTATGTATATGTACATGTGTATGTGTCTGAT
AACATTAAAAATTATAATATCTATAATTCCTTTTTGTAAAAAACTTCAACGAATAGTCTT
CTTAGTCTTCTTTCTTTCAGGAAAAAAATTTGTTTTTAAAATGACACTAGACGATCCGCA
AAACGATATACAAACGAATGCTCAGAGAAACGAAACAGCAAAAAAACGAGCACAGTTACA
AACAAATCAGAATTGTCAGCCAACGGAAAGTGATAAATTGGATAAAAGCAATATAACGGT
AAATCATCAAAAACAG

D. virilis ankyrin exon 1 + 1 kb upstream
>dvir_ankyrin5'
TATTTATGAAAATTTTTTTAACAAAATGTATGTAAGTTTAAGTTCGCAACCATTGGTGTC
AAACCTCTGCCACCAGAACCACGAGGGCGTCCAAAAGCCATTTTAATGCGAAAAGGCGCA
ACAAAGTCAGAAACAAATATAAACACGTGCGACTACTCGTGAGCTGTTAGTGTTGTATTA

AATAGTTCAGTATAGGTTCAATTGAGCGGTTCATTAAAATCATGTCCCTTACTATCATCG
TTCATTTAGTTCATTTGTCGGATTCTTTTAATGCTATCAGTTCATTCTTAATTGCTGTGC
GTCAAGTTAGCGCAGGAGACTTATCATCGATTATTGGAACTTGTTTCACTCTATTTCAAA
TGCATTCAATAATGTTTAACGGCAATTTTAAACAGCTGGACCTACCCCCATTTAAGCGCA
GCTGTGTGCAGCAAAACGACGAAATCTAATCTTAAAATAAAAACTGTTGAAGTCAAGTTC
TTGCATCAAAATTTTTAATAGGTTCGGTTACAGGTTCGTTTTTTTTTGAATAAGATAGTA
TATTTGATGAACTGCCGCTGATAATAAAACTTGTAAGGTATCCAGACTCGATATATCGCC
GCAAAAATGTAACACGATATTTAACTTTTTGTGGAAAATATCGATATTCCGATATAATTT
CGATAGGCCTAATTTCGTATTAGCATTCCAATTCGTTCGCACAAAATTTAAGGTAGGTGA
AGAAATGACGTCATATGTGAAGAAATAATTAAATAAATTCGCTGCTTGAACAAAATTGTA
AACAAGTTATAGAAGTACATAGGAAATTCTTGTAAAAAAGCATGCAAGCATGTTTATACT
AAAGTGCATGTATATAGAAGATACATACATACGCTTGTATACATTTTGTTTGTATGTATT
TGATATGGTAAAAGTTTTATTTTCTATGCTCCGGATATCCATTTGTGGTCAACTTAGTAA
TTAGTCAGTGCTTTATTTTCAGGAATTCATTTAAAAAAAAATGACCCTAGGCGAGACACT
AAACGAGATACAAACGAAAAGCCAGGGCCACGAAACAGCAACAGCAATTAAACGAACGCA
GATACAAATAAATCAGCACAGTGACAGTATGGACAACGCGTACATTGATAAAGCCAATAT
TAATGCAAAGCATCAAAAACAG

D. grimshawi ankyrin exon 1 + 1 kb upstream
>dgrim_ankyrin5’
CCACCTCCAACGGACTTGTTGAATCCTCCACCACCTCCGCCACGAAAACCTAATTAAATA
AATATACATTTGTCATTTCTACGCTTAATATATTTATTTATTTACACACCTCTGCCGCCA
CCACCTCCACGACCGCGTCCGAATGACATTTTTATATTATTAATTTTCAAAACTGACTAT
AAACACGTGCGGCAATTAGTGATGTGCTACGGCGAGCCATTTTCCGACACTTATACGATG
TGTAAACATCGCAGCGACATCTGCCGATAAGTACACAATGTGAAAGCAACGCTGCTGGCT
CAAATATTATATATCAAGCAATATATATCGCGCTTGCGTGCGCCTTGTCACGCATGAACA
TACATCGATTAATGGCTATGTTGGGCTCAGAGCTGACAGTCTTTCGGTCATTTTCCATTA
ACAATTGTTAGCTGGTCTTAAAGTAATTTTAGACAGCGAAAAATACAGTAGTGTATTAAC
AAAAATAGTAATGTAATTAATAAAATAATTCGAACAATTTTCCAAGAAGATATCGTTTTA
CAAATGGTTTATCGTGTGCCCTAAATTATCGACAGATTTCGATATCGCATGTGATATAAA
CGATCGAATACATTTGGGATTTAGCATTGCAAATCGTTCGCACAAAAAATTAAGGTAGGT
GAAGGAATGACGTCATACGGCGCATAGAAAAAGTGAAGCAGTGAATGGTGTGAATTAGTG
AAAATTTAATTAAATCAATTCTGTTTTGTGTGAATTAAATGTTGAGTAAAGAAATTGTAA
GCACTCAAATACACATAAACAGCGGATTAGAACAGATAATTAAGTGTTGTATGTAAATAG
ACTATATATATATATATATACATATGTATATATCTATACGCATATGTATGCACACACCTA
AATTTTTATGTACATAGATTACAAGTGTTGCTCTAGCCGAAAACGAAAAAATAATTCGAA
TTTGTGCTTTTATTTATTATTGTAGAGTTCAATTTGAAAAATGACAATTGACGAGCCACA
AAACGATTTACAAGCGACGACAGTAATTAAACGAAATCAATTACAAATAAATCCAAGTAG
CGATTCGATGGACAACGGCAACATTGATAAATCGAATAACAGTGGCCAGCACCAAAAACA
G

D. mojavensis ankyrin exon 9 + 1 kb downstream
>dmoj_ankyrin3’
CCCGATTGTATTGGAGGTGCCACATTTTGGATCGCTGCGCGACAAAGAGCGTGAAATAAT
CATTTTAAGATCGGACAATGGCGAAAGCTGGCGCGAACATAGTGTCTATGAAGACGAAGG
TAATTTACTCGAAACCTTAAAAGAAACGATGGCAGCTGATGTTAATCCATTGGAAGATTT
ACACACCAGTCGAATTATAAGAATCGTCACACAAAATGTGCCACATTTTTTTGCGGTGGT
ATCTCGCGTTCGCCAGGAGGTTCACGCAATAGGACCAGATGGTGGCACTGTATCATCTAT
AGCTGTGCCCCAGGTACAGTCCATATTTCCACCCCATGCATTAACAAAGAAGATTCGAGT
TGGTCTTCAGGCACAGCCAGTGGACTTAATTGGTTGTTCCAAGCTACTAGGTCAGGGCGT
TGCAGTATCTCCGGTCGTTACCGTCGAGCCACGACGGCGTAAATTTCACAAGGCCATCAC
ATTAAGCATTCCGGCGCCAAAAACTTGCAATCAAGGCATGGTAAACTCACCATATAACGC
CACGAATGGAAACGTAAGTGCCCCTACTCTGCGACTGTTGTGCTCTATAACAGGTGGACA
AAATCGTGCTGTATGGGAGGACGTCACTGGTTCGACGCCTTTGGCATTTGTAAAGGATAG
CGTTAGCTTTACAACAACGGTATCGGCACGTTTTTGGCTAATGGATTGTCGCAATGTTGC
GGACGCCGGTCGCATGGCAACCGAACTTTATTCGTATATGGCAAAGGTGCCGTTTATTGT
AAAATTTGTAGTATTTGCAAAACAAATATCCGCAACGGAAGCAAAATTATCTGTCTTTTG

TATGACAGATGATAAAGAGGATAAGACGCTCGAACAGCAAGAGTATTTTGCCGAAGTAGC
AAAGAGTCGGGATGTTGAGGTATTGCAAGATCAAAATATTTACCTAGAGTTTGCCGGAAA
TTTGGTGCCGGTTTTAAAATCGGGCGAACAACTGAACACAAAGTTTCAAGCATTTCGTGA
AAATCGCCTCTCTTTCATAGTTTACATTAAAGATCAAGAACAACCACATGCACGTATTTC
CTTTATGAGCGAGCCGAAAGTAACATCAGGAGAGGCGCCGTTGCGGCCAATTTGTACGCT
GAATGTATCACTTGAATCGCAAAAATTAAATCGAGTTAAACTACATCCGAATGGCATCAA
TACTAATGGAAATAACAATACGGATCATGACCTTAACTACAAAGATTTAATCAATGCCGA
AATAAAATGTAATACAGAATCAAGAGGAAATTCAAATATAAATCCGTTAGAAGATGGAAT
ACAGCGTGCCAATATCCGTTTATCAGACATATCGAATCTGCTGGGTAGTGACTGGCCGCA
GCTTGCTGAAGAATTGGGTGTACCCGAGACAGACATCAACATGGTTGAAACTGAGTACGC
CGATCAACCTGTTGCACAACAGGGCTTGGTCATGCTACGTCTATGGCTAAAACAAGAGGG
CAATCGCGCCACTGGTAATGCCTTGTCGCAGGCTTTAAACAAAATCAGACGTGAAGACGT
TGTTGAAAAGTGCATCTTTAACTTAGAACTGGTCACAGATCAACTGGAGCGTGGCTTGGC
CACGACGCGAATGCAGCATGACCACACTTTGATGCAATATCTAGCGGATGGCCTGAATGA
AACATTGAACATTGAACAGCAAACTAACGAAGACGAATTGTGTCGAAGCAATGAAAAATC
TAATAATGGTAAGAGACAAGAAATATTCAATTAATTAAAATTATATGTGTATTTATATGA
TATTTTAGATAAGTCTTGTGCTACGCCTCCTTCAACGCCAATAGAAGTCAACGTCGACTC
GCAGGAATATCAGTATAATATTAGTACGCCGATTCCTGAAATTATGCAAAAAATTATTGA
AGAACCAAGAGGTAAGGGATACAAATCTAAAAAATTTGTAATTTCTTATATATATTCAAA
TTTATTATATATTTTTATTATATATATTTATATTTATTTAAATTTAGTAAATGAAGTAGA
AAACGATTTAATTAAACCTTTGAACGAGAAGGATATTGAAAAGATTCCCCAATCATCTGA
AGAACAACTTTTACATACCCTGGCAGAGAAATACAAGGAGCCAACAACTTATTCCACGTT
GGAGGGCGATTTAAATGTTTTCGAAACGGAAAATAATAGTAAAACAAAGATAACGAACAT
AGATCCAAAAATGAAAAGCAATTTTCTAAAGAAATTAACAGAAGATTGAAATGTTTGCAA
CTAGTTTTCAAGTTATGTGCTTATTTATTTATTTATATATGTGTATGTTTTTTTGCTTTT
ATTTGACAACAGTATCAAATTAAGATTAGTGCAGATTATAATGTATGCGGTACATAATAT
ATCACTTTCCAATCTAAACCCTTTCTTTCAATTACAATTATTAAATTTTATTTGATTTAA
AGGCAATCTCTTATTTTGTATCTACGTATGCTAAAATAACGACAATTGAAATTGGTTGTC
TAGCAATATTTAATCAAAAGTGTTTTTACAAAACAAACATATTTAATATTTTTAGCTCAG
CAATGAAATGAAACTTAAGTTTATTTGAGTATCAATGCGTAATGTGGTATTTCGAAAATA
AAAATGTTGAACTTTAATTAATGTAAGCAATAGCTACATGTTAATGTTGTGTTTACATTC
TTCTTCACTACAAGTACTTGTATGTAGAAGTACATTATTATTATTATATTGATGATTGAC
CTATTTATACT

D. virilis ankyrin exon 9 + 1 kb downstream
>dvir_ankyrin3'
CCCATTGTACTAGAGGTGCCGCATTTTGGATCGCTGCGCGAGAAGGAGCGGGAAATTATC
ATATTACGATCAGATAATGGCGAGAGCTGGCGGGAGCATAGTGTTTACGAGGACGAAGAG
CACTTATTGGGCGCCTTAAACGAAACGATCGATGCCGATTTAAATCCACTGGAAGATCTA
CACACAAACCGTATAATACGTATTGTGACACAAAATGTGCCACATTTCTTTGCCGTTGTC
TCGCGCATTCGCCAAGAGGTCCATGCAATTGGGCCCGATGGCGGTACCGTTTCATCTACA
GCCGTGCCGCAGGTGCAGGCGATATTTCCACCACATGCATTGACTAAGAAAATTCGAGTC
GGTCTTCAGGCACAGCCAGTAGACTTGATTGGCTGCTCCAAGCTGCTTGGTCAGGGCGTT
GCAGTCTCGCCCGTTGTCACCGTGGAGCCACGCCGACGCAAGTTTCACAAGGCGATAACA
TTGAGCATTCCAGCGCCAAAAACATGCAATCAAGGCATGGTTAATGCACCATATAGCGGT
ACGAATGGTAACGCCGCGCCTACTTTGAGGCTCTTATGCTCGATAACTGGTGGACAGAAC
CGGGCCATTTGGGAGGACGTAACTGGTTCGACGCCCTTGGCGTTTGTTAAGGATAGCGTA
AGCTTTACCACTACCGTATCGGCACGCTTTTGGCTAATGGATTGTCGCAATGTTGCCGAC
GCTGGACGAATGGCCACAGAACTTTATACGTATATGGCTCAGGTGCCATTTATGGTCAAG
TTTGTGGTATTTGCAAAGCAAATATCCGCTACGGAAGCAAAGCTATCCGTGTTCTGTATG
ACTGACGATAAAGAGGACAAGACTCTCGAGCAGCAAGAGTATTTCAGCGAGGTTGCAAAG
AGTCGGGACGTCGAGGTATTACAGGATCAAAATATATATCTTGAGTTTGCTGGCAATCTA
GTGCCCGTATTGAAATCGGGCGAGCAACTGAATACCAAGTTCCAGGCATTTCGTGAGAAT
CGTTTGTCTTTTATAGTGCATATCAAGGATCAAGAGCAGCCGCACGCTCGCCTCTGTTTT
ATGAGCGAGCCGAGAGTTGGACCCGGTGAGGCACCATTGCAGCCAATCTGTGCTCTAAAC
GTGTCCCTGGCAGCCCACGAGGTTAATCAGGTGTTTAACAGGTCAAATGAAAATGGTATC
GATAATGGATACAACATGGATCACGGCCTGAACTACAAGGATATGATCAATGCCGGCATA
AAGGCCAGCACAAAATCGCCCGACAATACAATAATACGGCCTATTGCCCACGTAACGGAA

GACATACAGCGCGCTGATATCAGATTATCTGATATATCGAATTTGCTGGGAAGTGACTGG
CCGCAGCTGGCAAAAGAATTGGGTGTACCCGAGACCGATATCGAGCTAGTCAAAGCTGAG
TACGCTGATCAACCGGCTGCCCAGCAGGGTCTGGTTATGCTGCGACTCTGGCTAAAACAG
GAGGGCACTCGCGCCACCGGCAATGCCATGGCTCAGGTACTAAACAAAATTGGCCGAGAC
GATATTGTGGAACAGTGCATTTTTAACTTGGAGCCAGTCACCGATAAACTGGAGCGCGGC
CTGGCCACAGCCAGATTGCAACAGAATCAAACCAATCTCGCTGACGGCCTCAATGAAACA
TTGAATATAGATCAGCTTAGCCAAGATGATGAATTGTGCCCAAAAGCATGCCAATCCCAG
AATGGTAAGTTATCTCTCTGTCAACAAATATTAAACGCAAATGTATATATTCTTTATAGA
CACAACAGATAGACTCTGCTCGACACCGCCGCCAACACCGAGTGAACCGCAAGAATATCA
GATTCCCGAATATGTGCATGAAAATATCGTGTCAAAAGATAGTAACTATTAGATAGAAAG
GAATCCTGTATTTTTATTGGTCTCTAATTGTTCTTTTAATTTATTAAGGCATTGAGACGG
AGGACAATTTAATAAAATCATTGCACCTAAAAGATTTGGACAGGATTCCAGACTCAGTTG
AGGACACGCTATTGCATTCACTGGCGGACAAGTGCCAACCAGCCGGCAATATACATGTTA
AAGCAATTACGACGAGTAATTCCGATTTGCAAAGCAATATTTCAAACAAAACAAAAGATT
GAACAGTTGATATATCATGTTTTATTTATTTATATGGATGTACTCTTTTTTTGTTTGCAA
CTACAAACGAAGCTCATGTAGTATACGGTACATGTTCAAATGTATTAAAAGCAGATTTAT
TTCGACGGACGGGACGATTTCAATTTTGTTTTTTTGCAGCTATAGTTAAAAAGAAAAAAA
ATATTTTCATTGCTTTTAATATGGATTCTAATTAAAATAGAAGTCCGACGAGATATCTTC
CAATTTTATTGCACCAGTAAAAACTATGATGTTGTCTTGCAATATTTATTCAAAGTACAT
AATTAATTTGTTTTTCGTGTGTGTTTAAATAAAAAACATATTTAATAATTTTAGCTTGGT
GTTGAATTCAATAGGAAGCCCTCGGTTTGCACGCCTTGAATTCATCCGAGTAGAGTAATG
CGTAATATTTTATTTCCAAAATAAATTGTTGAACTTTAATTGGATGTACTTGTATTATGA
TATTATAGTCAAAATTGGTATATTTCGATATAATCGTGGGGAGACTATCCTTAAGAAAGA
TATGTATTTCGTGAATTCAAATAAGACTAAAATCGCCAATAGTCGCAGCTGTCGAGCTGA
ACTAAAAATAGGAATA

D. grimshawi ankyrin exon 9 + 1 kb downstream
>dgrim_ankyrin3'
CCCATTGTGCTGGAGGTACCACATTTCGGAGCATTGCGTGAACGCGAGCGTGAGATAATC
ATACTCCGCTCGGACAATGGAGAAAGTTGGCGCGAACACAGCGTCTACGAAAACGAGGAG
CATCTCCTGGAAACGCTGAACGAAACGATCGATGGGGAACTAAATCCTCTCGAGGATTTG
CACACCAGTCGCATAATACGCATTGTGACTCAAAATGTGCCCCATTTCTTTGCTGTTGTC
TCACGGGTGCGGCAAGAGGTACACGCTATTGGCCCTGACGGGGGCACTGTGGCATCAAGT
GCTGTGCCCCATGTGCAGGCCATCTTTCCGCCCCATGCGCTGACCAAGAAGATTCGCGTC
GGCTTACAGGCGCAGCCAGTGGATCTGATCGGTTGCTCCAAGCTGCTTGGACAGGGCGTA
GCCGTCTCTCCGGTGGTAACCGTGGAGCCGCGTCGCCGTAAATTCCACAAGGCAATCACA
TTGAGCATGCCGGCACCCAAAAGTTGCAATCAGGGCATGGTCAATGCTCAGTACAATGGA
AATGCGCCAACTTTGAGATTGCTTTGCTCCATAACTGGTGGCCAAAATCGCGCCATTTGG
GAAGATGTCACCGGTTCGACGCCCCTCGCCTTTGTCAAGGATAGCGTCAGTTTTACGACG
ACAGTTTCCGCTCGGTTTTGGCTCATGGATTGCCGTAATGTGGGCGATGCCAGCCGGATG
GCTGCCGATCTTTATTCGTACATGGCTCAAGTGCCTTTTATCGTAAAATTTGTCATCTTT
GCCAAACGCATTTCTTTAACGGAGGCCAAATTGTCCATCTTTTGTATGACCGATGACAAG
GAAGAGAAAACGTTGGAGCAACAGGCGTCCTTTACGGAGATCGCCAAGAGTCGCGATGTG
GAACTGTTGCAGAATCAGCAAATTTATTTGGAATTTGCTGGGAACTTGGTGCCCGTCTTA
AAATCCGGTGATCAATTGAATATCAAATTTCAAGCTTTTCGCGAGAATCGCTTATCGTGC
ATTGTCCATCTCAAGGATGCGGAGCAGACCCAGGCACGCATCTCTTTCATGGCACAGCCA
AAGGTGGAACCGGGCCAGGCTCCGCTACAACCGATCTGCGCTTTAAATCTTTGCCTGGAT
CAGTTGGAGGAGCAGGAGCAGGGTTTCAATGGAAACCTCAGCAATGGTCATGGCCTCAAC
TATAAAGATTTAATTAATGCCGAGATTAAGGCCAGCAATCAGGAGAAGCTGTCCATACAA
CGTGCCGATATACGTCTCTCGGATATATCCAATCTGTTATCTGAAGATTGGCCTAGGCTT
GCCCATGAGCTGGGTGTTACCAGCAGCGACATTGATTTGGTCAAAGCGGAGTATGTCGAT
CAGCCACAGGCACAGCAGAGTCTCGTCATGCTGCGCCTCTGGCTGGAGCAGCAAGGAAAC
GGCGCAACTGGCAATGCTATGTCGCAGGCTCTATCCAAAATTGGACGCGATGATATCGTC
GAGCAGTGTATTTATAATCTTGAACTGGTCACCGATAAATTAGAACGTGGTCTAGCTAAT
GTTCGAATGCGATCGGCGCCACCAGCTTCGACTCCTATTCTAACTGATGGCCTCAACGAA
ACCCTCAATATTGAACAACTTTGCGAAAGTAAGCAAAATAATGGTTAGTAAAATTGAATT
AATTCCTTTCAATTCGATTATAATAAATGTTCAATTGTAGTTCTACTCTGCTCGACGCCG
CCAACAACGCCCATTGAGCTGCAGGATTCTCAATATCAAATCAATTTTCAGCGAAGCGCT

GGTAAGCATTAAATATTGTCTTTAGTTTAAGTTGCAATAAGTTTATATTTTATTTAGCGC
TCTTTTTCTTTAGTTCAATTATCGACTATCGATTAGCAGAGTTGGTTTTAATAGCGCGTT
TAACAGCGTTGCTTAGCATACACAGACCATTTCGAGTAATTAACAATATGGCGCTTAACA
TTGAGTAATTAATATGATGGTGCTTATATATGAATGAGCTAAAAGAGCTAAAATCGTATA
CATTTAAATAAGCTAACGGAATTGCATAACTTAGAAGTTAAGCAATGATTACAGGGCTTT
TATTAGTGCTTAACAGCGTTGCTTAACAAATTAGTTATACATATGTGGTCACACATACAC
ATACCATTTCGCCGCAATAACAATATGGCGCTCATATTTAAATGAGGGCATTACATATCG
ATTAGCAGTGTTGCTGAACATAATTAGAGATATATTAACACGCGCCAACATAACTAGTAC
GTATTAACATCTATCTATTTTAATTTCACAGACAATTCAATAATTAAACCGATTGAAGTG
GAGGGTGCTCTGCCCGCTTCCTCGGAAGACATGCTGCTGAAGTCCCTGGTGGACAAATGC
GCCATAATAGAACCCCCGAATTCAAAATGGACTGGCGTTAATCCAGCAAATGATTTGTTA
CAAAGTGAAATTGAAGAATCGCATATCAAAACCATAACGAAGCTTTCTCCCCCATCCGAT
ATCGAATGCAATATGTCCAAGCCGACAACAACAATAACCGATTGACCAATCGATAACTTT
TAAAATTTTGCATGTGTGTGTGTTTGTGTTTAGCTCTGTGTATCGCCTATATTTATTCTT
GTGTATTGTATGTATTTTAAATTTA