ABSTRACT

RNA GENOME ANNOTATION WITH A FOCUS ON T. BRUCEI

by
Brett Bucci

The goal of this project is to identify untranslated regions (UTRs) and UTR-indicating patterns in the genome of *T. brucei*. *T. brucei* is an interesting organism, and as the cause of African sleeping sickness—which infects 300,000-500,000 people and a significant number of cattle annually—is currently the subject of considerable research. Using existing algorithms, several patterns have been found that may lead to more complete UTR annotations in the *T. brucei* genome. The most encouraging sequence is the 11-base sequence GAGGG[CG]TGGGG, which appears in five hypothetical genes near the tail. Discovery of several such sequences could guide laboratory experimentation toward more useful results and a better allocation of time and resources.

RNA GENOME ANNOTATION WITH A FOCUS ON T. BRUCEI

by
Brett Bucci

A Thesis
Submitted to the Faculty of
New Jersey Institute of Technology
in Partial Fulfillment of the Requirements for the Degree of
Master of Science in Computational Biology

Department of Computer Science

January 2008

APPROVAL PAGE

RNA GENOME ANNOTATION WITH A FOCUS ON T. BRUCEI

Brett Bucci

Dr. Jason T. L. Wang, Thesis Advisor
Professor of Computer Science, NJIT
Date

Dr. Andrew Sohn, Committee Member
Associate Professor of Computer Science, NJIT
Date

Dr. Dimitrios Theodoratos, Committee Member
Associate Professor of Computer Science, NJIT
Date

BIOGRAPHICAL SKETCH

Author: Brett Bucci

Degree: Master of Science

Date: January 2008

Undergraduate and Graduate Education:

- Master of Science in Computational Biology,
New Jersey Institute of Technology, Newark, NJ, 2008

- Bachelor of Science in Science,
Pennsylvania State University, University Park, PA, 2000

Major: Computational Biology

ACKNOWLEDGMENT

I would like to thank Dr. Wang for his patience during my research, as well as his insight into interesting ideas and direction for my work. His willingness to share ideas has helped my studies and has hopefully helped his laboratory as well. I'd also like to thank his doctoral students for helping with questions I had along the way. And of course my committee members, Dr. Sohn and Dr. Theodoratos, deserve thanks for reading and critiquing my work.

My girlfriend, Vicky, deserves a special thank you. She kept me on track when I frequently lost focus. She kept me laughing when I otherwise wasn't in the mood. Vicky, your perseverance on your dissertation was incredible and was inspiration for me to finish a much smaller task. I couldn't ask for anyone better.

TABLE OF CONTENTS

LIST OF FIGURES

Figure **Page**

CHAPTER 1

INTRODUCTION

The goal of this project is to identify UTRs and UTR-indicating patterns in *T. brucei.* Current UTR annotations are limited, and are mostly focused on chromosome I. Several algorithms exist to predict UTRs, and many have been predicted by sequence homology and other methods, but without experimental evidence functionality cannot be verified. One of the aims of this project is to determine the best UTR candidates, and perhaps guide laboratory experimentation toward more useful results. One of the sequences found to recur in putative UTR regions also seems to be present toward the end of several hypothetical proteins, and may be a good indication of where to direct laboratory resources.

CHAPTER 2

METHODS

The two main sources of annotated UTR sequences were GeneDB [1] and NCBI [5]. To find UTRs in NCBI, the author performed the following steps. This search will produce approximately 34 results.

1. Enter *trypanosoma brucei* in the search field
2. Select Organisms on the Limits tab and click Go
3. Enter *5'UTR* in the search field
4. Click on the Limits tab and change to All Fields
5. Click History
6. Click the numbered link next to ***Search trypanosoma brucei Field: Organism***
7. Click AND in the pop-up menu, and then click Go

To find UTRs in GeneDB, the following will yield about 15 results.

1. Select *T. brucei* from the Protozoa menu on the right
2. Enter *UTR* under Full Content Search , and then click the Full Content Search button

CHAPTER 3

PUTATIVE UTRS

The sequences in Appendix A have been annotated as UTRs by either GeneDB or NCBI. Where possible, some subsequent sequence information has been provided. The key to the annotations is as follows:

The **bold** portion of the sequence is what's annotated as a UTR by GeneDB or NCBI. The underlined portion is one of the following highly conserved sequences that appears four times in this data A[AT]AG[CT]AGAGG), or twice GAGGG[CG]TGGGG (see note below).

The sequence GAGGG[CG]TGGGG appears 11 times in the *T. brucei* genome according to a BLAST search using *The T. Brucei Genome Project* (8) website:

*Tb10.389.1530 (741 bp) positions 621-631
Tb927.2.2070 (474 bp) positions 132-142
Tb11.22.0002 (486 bp) positions 264-274
Tb10.329.0010 (513 bp) positions 281-291
Tb927.8.3080 (3915 bp) positions 1143-1153
*Tb927.3.2780 (3309 bp) positions 3146-3156
*Tb927.3.3050 (3096 bp) positions 2926-2936
Tb10.05.0160 (1569 bp) positions 528-538
*Tb927.3.1910 (1776 bp) positions 1738-1748
Tb11.01.6770 (2172 bp) positions 1536-1546
*Tb11.02.0020 (1941 bp) positions 1719-1729

Figure 3.1 Locations of GAGGG[CG]TGGGG sequence in *T. brucei* genome.

The above sequences marked with asterisks (*) are good candidates for further exploration because the likely UTR indicator appears in the last 20% of the sequence. There are five such sequences. Although the sample is small, this is noticeably more than the statistically expected number of appearances, which is approximately two. It is important to note that each of these sequences is currently a hypothetical protein, and that laboratory experimentation would be required to confirm functionality. With further UTR information, the five sequences with potential UTR regions might be good targets. This could be a good indicator of 5' UTRs.

CHAPTER 4

UTR SEARCH

To search for coding regions in unannotated sequences, two main tools were used. The first tool, UTRscan [9], was developed by researchers at Istituto di Tecnologie Biomediche in Italy. UTRscan searches for approximately 30 patterns that are believed to indicate 3' or 5' UTR regions. More information about the patterns, including descriptions and sequence permutations, can be found at UTRsite [10]. These descriptions include functionality, mentions of conservation in other species, references, and historical information.

Another resource, BlastUTR [7], is maintained by the same researchers and looked promising, but has not been functioning properly.

MEME [3], the second tool used to analyze sequences for UTRs, was developed by three researchers at the University of California, San Diego. It searches input sequences for motifs and provides detailed output including locations, regular expressions, and p-values. MEME has the ability to find quite a few motifs depending upon the input parameters. These motifs nearly always have quite a bit of variability in the actual sequence, with only certain sequence locations being fixed. The sequences below show some of the MEME hits with the least variability.

The sequences in Appendix B were obtained from NCBI and run through UTRscan. Each sequence included a UTR in the annotation. The <u>underlined</u> regions are hits from UTRscan. The **bold** regions are identified as UTRs in the sequence's annotation. The *blue italicized* regions are motifs found by MEME.

The sequence CACACATACAC (which appeared twice in the UTR of AM168497) appears 24 times in the *T. brucei* genome according to a BLAST search using *The T. Brucei Genome Project* website:

Tb09.v1.0620 (117 bp) 74-84
*Tb927.1.1320 (231 bp) 11-21
*Tb927.1.3440 (246 bp) 3-13
Tb09.160.3650 (297 bp) 202-212
Tb927.1.4060 (306 bp) 245-255
Tb927.1.4510 (306 bp) 59-69
Tb09.211.2690 (312 bp) 106-116
*Tb927.1.1250 (312 bp) 30-40
Tb927.5.2330 (13254 bp) 3048-3058
Tb927.4.4800 (393 bp) 264-274
Tb09.160.4060 (456 bp) 159-169
Tb09.211.3260 (510 bp) 417-427, 409-419
Tb09.211.4260 (537 bp) 275-285
tmp.1.100 (8300 bp) 3217-3227
Tb09.160.1410 (543 bp) 258-268
Tb927.3.1190 (6984 bp) 6671-6681
*Tb927.6.3210 (678 bp) 35-45
*Tb927.2.4440 (714 bp) 53-63
Tb11.02.4490 (714 bp) 454-464
*Tb927.4.3550 (1029 bp) 5-15
*Tb927.4.4810 (1095 bp) 162-172
Tb927.4.3280 (1233 bp) 566-576
Tb11.01.3740 (2637 bp) 813-823
*Tb11.01.6760 (1788 bp) 19-29
Figure 4.1 Locations of CACACATACAC sequence in *T. brucei* genome.

The above sequences marked with asterisks (*) are good candidates for further exploration because the likely UTR indicator appears in the first 20% of the sequence. There are eight such sequences. The sample is small as above, but again this is noticeably more than the statistically expected number of appearances, which is approximately five. Since these appear in the front of the sequences, these are more likely to indicate 3' UTRs.

To get an idea how common the sequences in UTRsite are (these are the sequences that UTRscan searches for), the author submitted the first 330,000 bases of chromosome II as input into UTRscan. The results are shown in Appendix C.

The most common UTRsite sequence found in this section of Chromsome II was 15-LOX-DICE, with 153 occurrences. The following is a histogram of the 15-LOX-DICE locations as output by UTRscan above. Each bucket represents 27,500 bases. Therefore, bucket 1 counted sequence locations 1-27,500, bucket 2 counted locations 27,501-55,000, etc. The distribution is fairly even, with occurrences an average of 2,170 bases apart in this sample. Submitting each segment of the genome sequentially (in roughly 330,000 base sections, since the limit imposed by UTRscan is 350kb) could yield more interesting patterns.

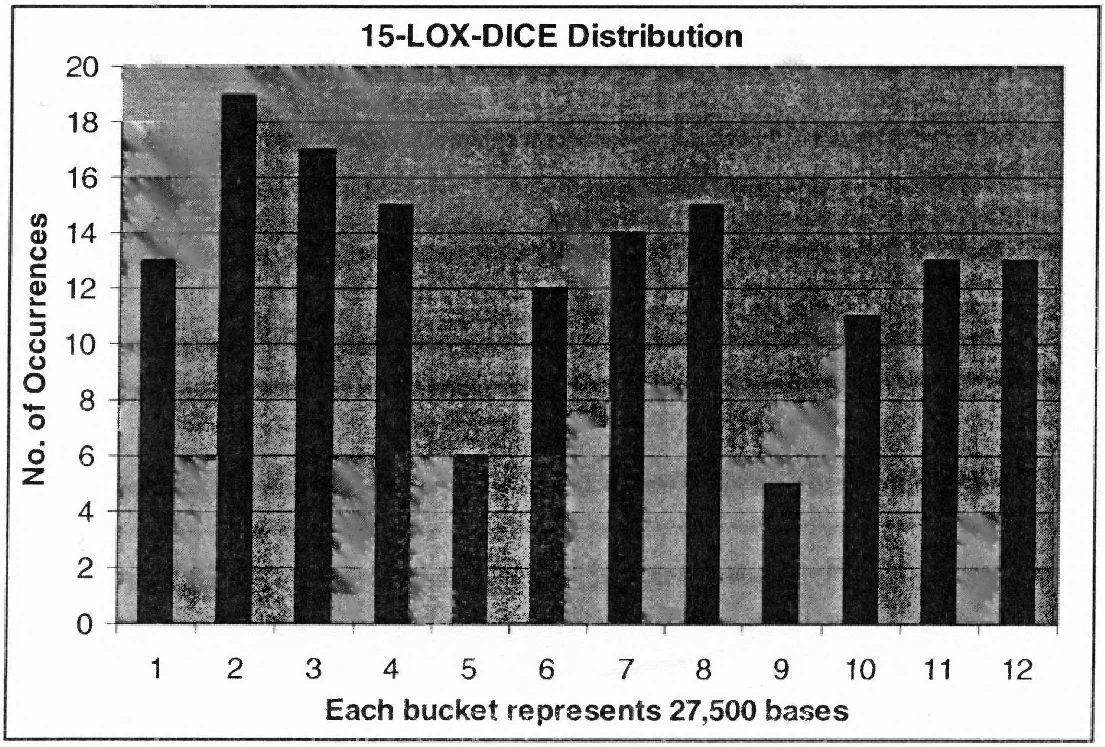

Figure 4.2 Distribution of 15-LOX-DICE locations on chromosome II.

The next most common UTRsite sequence was K-Box with 34 occurrences. Since the 15-LOX-DICE appear to distributed relatively uniformly, the author became curious about how some of the other output sequences line up with respect to these. The locations and distances of the 34 K-Box sequences were compared to the 15-LOX-DICE sequences and an interesting relationship was found.

The K-Box sequences tend to precede the 15-LOX-DICE sequences. In 23 of 34 instances (62%), the nearest 15-LOX-DICE sequence was "behind" the K-Box in question. In other words, from the K-Box's starting position, it was usually more likely to find a nearby 15-LOX-DICE sequence in the forward direction. This could mean that the combination of a K-Box followed closely by a 15-LOX-DICE provides a stronger indication of a potential UTR segment than either sequence alone. The average distance from a K-Box to the next 15-LOX-DICE ahead of it is 1457 bases, while the average distance from a K-Box to the previous 15-LOX-DICE sequence is 6080 bases. Although the sample is small, this is more than a four-fold increase.

To confuse matters, the average forward distance from a K-Box to a 15-LOX-DICE is 3,225 bases, while only 2,619 bases in the backward direction. This might suggest that in the 62% of instances in which the forward 15-LOX-DICE is closer the sequences are in some way correlated. Since the average distance between 15-LOX-DICE sequences in this sample is 2,170 bases, it would also appear that the K-Box segments are occurring in the larger gaps between 15-LOX-DICE hits. This makes sense probabilistically, since if one assumes a uniform K-Box distribution, longer spans for the K-Box sequences to fall in would yield more hits. Otherwise, the expected average distance from a K-Box to the nearest 15-LOX-DICE would be about 1,000 bases. The actual average of 1,733 bases is greater, but not alarmingly so.

CHAPTER 5

DR. GOPAL'S RESEARCH

ORBIT [6], one of the tools used here to help predict whether an RNA sequence is coding or non-coding, was developed by Dr. Shuba Gopal, currently at the Rochester Institute of Technology. Dr. Gopal's work has focused on creating an alternative to annotation by sequence homology because organisms that are long since evolutionarily diverged tend to yield many false positive coding regions [2]. *T. brucei*, for example, are thought to be more than 800 million years diverged from *S. cerevisiae*, its nearest evolutionary neighbor.

Her paper (*An organism-specific method to rank predicted coding regions in Trypanosoma brucei*) describes a method that separates coding and non-coding regions based on nucleotide composition. Using standard sequence homology, more than 500 coding regions have been noted on *T. brucei* chromosome I, yet barely one-fourth of these have assigned functions. The reason so many regions remain unassigned is because there is little evidence for function besides homology, and experimental determinations of function for so many regions is unfeasible. However, if educated guesses could be made as to which regions to look at first (i.e., that were the most likely to be true coding regions), then the effort might be worthwhile. This is what ORBIT attempts to accomplish.

ORBIT identifies differences in nucleotide composition between coding regions and the region immediately upstream. This upstream region is rich in thymine and cytosine; an abundance of these pyrimidines appears to signal a trans-splice site. These trans-splicing signals are assumed to indicate non-coding regions because it is very unlikely that they will occur in the middle of a coding sequence.

To determine whether or not a region codes for a protein, ORBIT uses linear discriminant analysis (LDA). While LDA may not be as sophisticated as other pattern recognition methods, it was the optimal classifier for this simple coding vs. non-coding decision. Transition probabilities at the dinucleotide level were calculated using maximum likelihood estimation. The codon level, which comprises groups of three nucleotides and has three potential reading frames in each direction, did not provide useful classification information for the coding vs. non-coding decision.

Dr. Gopal's other tool is Motif-er. Motif-er's primary use is for genome visualization. *T. brucei* chromosomes I and III are currently mapped with coding regions predicted by ORBIT as well as current public annotations. Sequence information can be downloaded along with the coding likelihood score as predicted by ORBIT's LDA classifier.

CHAPTER 6

DISCUSSION

The results are encouraging if not entirely concrete. With a relatively limited data set, enough clues and motifs have emerged to continue searching using similar methods. The motifs found by MEME in the UTR-only data set have yielded clues as to other possible UTRs, as shown by the five hypothetical genes in which the sequence GAGGG[CG]TGGGG appears near the tail.

The small number of UTRs that are currently annotated leaves a lot of room for improvement in this area. The current techniques are a good start, and some more advanced techniques could be a decisive step in better UTR predictions. While ORBIT's use of LDA may be optimal for a two-pattern classifier, more advanced techniques such as Support Vector Machines (SVMs) may be able to better learn the sequences and identify untranslated regions. Another advantage of using SVMs may be that there is more information than just that contained in dinucleotide transitions.

There are several sequences that show Internal Ribosome Entry Sites at their tails. There are also sequences whose annotated UTR does not agree with UTRscan's results. The 330kb section of chromosome II against which UTRscan was run gives an indication of the tool's sensitivity. In this section of bases UTRscan found 267 hits from the UTRsite list. This amounts to a potential UTR-indicating sequence every 1,236 bases. This might be slightly more than expected, however, each UTR might be composed of several different sequences, and thus this inter-UTR spacing would increase.

The sequence CACACATACAC was found by MEME to appear twice in the same UTR, and may be a promising key to other UTRs. MEME is a very valuable tool, but it will be easier to use without the 60,000 base restriction. Being able to submit an entire chromosome's sequence at a time, for example, will allow motifs that appear farther apart than 60,000 bases to be elucidated. For example, motifs that appear infrequently—perhaps only once every 100,000 bases—could be stronger indicators of UTRs than more common sequences.

If motifs could be generated by MEME and shown graphically on a map of the genome similar to the one used in Motif-er, the location of these motifs could be compared with the predicted coding regions. This could provide very valuable insight.

APPENDIX A

UTR ANNOTATIONS

This appendix contains sequences annotated as UTRs by either GeneDB or NCBI. The **bold** portion of the sequence is what's annotated as a UTR by GeneDB or NCBI. The underlined portion is one of the following highly conserved sequences that appears four times in this data A[AT]AG[CT]AGAGG), or twice GAGGG[CG]TGGGG (see note below).

Tb927.1.1000 5' UTR
Source: GeneDB
Chromosome 1
289,877 … 289,892
ORBIT:
–16 bp UTR non-coding (.840)
–1263 bp gene coding (.969)
UTRscan:
–1 IRES in 80 bp sequence centered at UTR
GAATGAAGGTAGTACTATGCGTCGCTTATTGTGTCT…

Tb927.1.700 3' UTR
Source: GeneDB
Chromosome 1
231,710 … 232,503
ORBIT:
–794 bp UTR non-coding (.999)
–1323 bp gene coding (.963)
ACTTCCAGAAAAAATATATTTCTGCAAAATACTTTTGGAAGTTTGTCTTG
TCTTTATAGATGAAGGATTTGTTTCTTTTTTGTGATGTTTTCAAGGTTAAT
TAGTTTTGGGGGTTCGTTATCTTAATTATTTTGGTGGGTGGGAGTAAATA
AAGCAGAGGTAAATTTTTTGGTGACACAAAAATTGGGAAGCTTCGTGTT
CTTACTTGTTCAACTGAAAAATGCCTTTTCAGGAATTCATATTTGGGAGT
TATTGTGGTGTAGAAGGACTGAGGAACAGAAGAAAGCAGAGGTTATTTG
CCCCTTCATGAGGAAATGTCGATGTAATTAAGTATGAGGGAGGACATGT
TGATACTGGGAAATGGACTCTAAAAATGAGAAATAAAGGGAAAGAGAAA
GGAAGAGTGATATATATTTATTTTTGGAAAAAAACACCTTTCGTTTGCTT

GCGCTGCTGAGTGGGAGATCATTCTCTGTGTTATATGTCCTTTTTCTAGT
GGTTGAGATTGTGTTGTTGTTTTTTCAATTTCTTCTGTGGATGATCTTCC
TCGTGAAGAAGACGCAGAAAGCGGGCCACACGGAGTGAATTCATACCTT
ACTTAAAATAATATAAAACGTATTAAAATATGTAATTATATATATATATAT
TTCCCTTTCTTTTTTAAAAAATCTCTCTTTTGTGCTTCTTGCTTCTCTCAT
TTTCTAAACTGGGCAATTAATATGCTCGAAAGTAAATATTGAGGTTATTG
AAGAGGGCTGGGGTGTGAATGCTTTTCTTTTT
CCTTTGCCTGTGTTACCGGTGGAGCTCTCTTTAA ...

Tb927.1.700 5’ UTR
Source: GeneDB
Chromosome 1
233,825 ... 233,904
ORBIT:
–80 bp UTR non-coding (.918)
GTTCAGCTCTTTGGTGATATCAAAGCATAATTGCTGCGGAGATACGTTTT
TCCACCTAATAAGTAATTGTGATACAAGATCAAATCGTTTGGACTGTAGG
...

Tb927.1.710 3’ UTR
Source: GeneDB
Chromosome 1
234,043 ... 234,175
ORBIT:
–133 bp UTR non-coding (1.000)
–1263 bp gene coding (.978)
TATTCATCCTGTTACGGGCCTGTTTTATGGAATTGTGTTTTTTAGTCCTTT
TTATTTGTTGGTTAGGTATTGGTTCGTACGTGACTATTATTTTTTTTTTAG
GATAACATTTATGTTTTTTCTACTCATTTTAATTGGACGAAAAGGAGTAAT
...

Tb927.1.710 5’ UTR
Source: GeneDB
Chromosome 1
235,437 ... 235,552
ORBIT:
–116 bp UTR non-coding (.996)
CAACATACTTGTATTTTTTGTTTCAAAACATTAAAAAATTGTAACAAGGG
AGTTTCTTATTTTTTTGAAAAAACTATATATATCGATATATACTTATCTGA
TCACAAATCAAATATCAACGTTTTCTCACTTAGCC ...

Tb927.1.4100 3’ UTR

Source: GeneDB
Chromosome 1
862,869 … 863,029
ORBIT:
–161 bp UTR non-coding (.999)
–1062 bp gene coding (1.000)
**TGGAATGGCTCTTTTACCCGCGTAGGTTTTGTTTATTAGTCTATTTATAT
ATTTACCTATTCGTTTGTGTATGCAATGGAGTTAGTTTGTAGCAAAGGGG
GAAGGAGGGGTGGGGAGGGGAGGTCCCAGAGAGAAAGTGAAGGAAATA
GAGGGAAGAAGAG**GCTCTCAAACAAGATTTAGT…

Tb927.1.720 3' UTR
Source: GeneDB
Chromosome 1
235,665 … 235,769
ORBIT:
–105 bp UTR non-coding (1.000)
–1530 bp gene coding (.972)
**TGTACATCAGGCGAAGGGTTTGTTTTTTTTTCTCCTGCCCTATGTTTTT
CTGATGTCGTGGGAGTTTTGAATACTTTTAGTATATCGTTTATTATTTGT
GAAC**ATTGGATGATAAGGAGTAAT…

Tb927.1.720 5' UTR
Source: GeneDB
Chromosome 1
237,298 … 237,503
ORBIT: non-coding (.785)
**GGAACGTGTGTGTGTGTGTGTCATAGAACTGCTTTCCAGCAACGCATCG
CACCAGAAAATTAATATACCTTAGTCATTCCATTTCCTATTGCGGGTACA
ACGATAACGGTGGTAAAACCGTCGGCGTTTTTTTTTCTAAGTAATCGAA
ACAACGAGAAGTAGCGGGAAGGTCAAGAACAAAAATAAGAAAAACAAGC
GGGATCAT**TCTTTTACTTACTGTTAGTG…

DQ826505 3' UTR
Source: NCBI
Chromosome 1
58 bp
ORBIT: non-coding (.996)
**ACTAGTTTCTGTACTATATTGTGAGTAGCCAGCTTTGACCAAAATATAAC
TGACTGCT**ATGTATTCGAAAAGCA…

DQ826504 3' UTR

Source: NCBI
Chromosome 1
19 bp
ORBIT: coding (1.000)
AGAAAAGACACGACCAGAAATGGCCAACACATCG...

N45755 5' UTR
Source: NCBI
Chromosome 1
376 bp
ORBIT: non-coding (.999)
**TGTNCACCCGCTGTCGNCCGCTCTAGAACTAGTNNTTCCNCTGTGNCTGC
AGGNTTTCNGNACGAGGTTGGTCGCCGCGAAGTTATNCCATACAAGGGC
GTTTTTAGGCAGCAAAANCCAAGCAAATAGCAGAGGCAAGGNGCTTCCN
CGTAAGTNTAGTTAGTGGAGCGGTTTTCTNATGCNAACAGNCGTNGCTN
TCCTGTTGNTNNTTTACAGNGGCAGTNNTTTTNTNGTNCAGTNTTTGGGG
GCCATTTNGGANAAATGCCNTTTTACAAATAACNNTGGTAAGTAGCTTGT
NTGTNGTGTTTNAGNNNACGTTGCTTCTANNGAANGTTTNNAAATTGGTN
AATGTCCCTNNTTTNTTGGTGTTGGGATT**

T26740 5' UTR
Source: NCBI
Chromosome 1
249 bp
ORBIT: coding (.942)
**TGTACATCCGCGCGCCACTCTATTCAGAGAGCCACGGATAGTAGAGGAG
GTGGGAAGGGTATATNAGGGACACGCGTACCATGATGTGGGATGTATTG
GGGTCCCTGTCTGTCCTTACGTGACTATGTATGAACCGTNACGTGTAAG
ATGAGCTAGTGAGATCAACAGTACAACTCATTAAACACGNCTTCTTCTCG
TTAAATGTACACAATCTTGNTCCTCCACCTTTAAAAAAAANAAAAAAAAA
AA**

APPENDIX B

UTRSCAN OUTPUT FROM NCBI SEQUENCES

The sequences in this appendix were obtained from NCBI and run through UTRscan. Each sequence included a UTR in the annotation. The underlined regions are hits from UTRscan. The **bold** regions are identified as UTRs in the sequence's annotation. The *blue italicized* regions are motifs found by MEME.

AJ243568

GTTCCAAGTTTAGGGGGGAACCAGCGGCCTCCAACCGAATGAACCAACCTAT
ATCATCCTATATCCTCTGTGCCGCGGCCTCGCTCCAGGCGCTTTACCGCCACA
AGAGGAATTCCCTCAATGAGGGTCTCCGCTTGTTCACTTTAGGAAGGCCACA
AACCATCCGTTCCCGCAACGG*GTGGAGACCCCAG*CGTTCCCCAAACGCCGCTT
CTCCAACTCCCGAAGAACCATCACCGCTTTTCGGGCGTCACGACTCGCCATCC
ACCTCCACATGCATATCAGTCGGTCCAAAATGCGACCCTCCCTTCCCACGCAG
AACGACAGCTTTTTTCGCCACATTGGAAGGAAGGTGGACAAAACACCCATCC
ACCACACGTGCCTTTTTCCCGTTTCGGTGAAGCCGGTGGGTAAGGAAATTGG
GCGCCCAGAAAAGGGCCGTTACGGGAATTGAACCCGTGACCTCCTGCACCCA
AAGCAG*GAATCATACCACTAGACCA*AACGGCCACACCGGCGGGGCACCAGGTC
CAACTTATGCACCTATGCTGGAGTAGATTGGAGATAGCGCCGGGTCCCCGAA
GCACCGTGGCGCAGGGGAAGCGCGATGGGCTCATAACCCATAGGACGTTGG
ATCGAAACCAACCGGTGCTAAGTTTTCACATCCACCCTTTTTTCTCCAAAGGA
AAATAAAGGTCGCCGGTTCGCAAAAAGTTGACGAGAGTGGGGTTTGAACCCA
CGCCCTCGGAAGGATTGGAACCTTAATCCAACGTCTTAGACCACTCGACCAT
CTCGCCACGGGACACCGCTACAGCACAAAACATCGACACACCGCAATGAGC
AGATCGTTATCATTTTAAGCACGTCCTGGGAAGAAAACAGCCAGCCGTGGAT
TCGAACCCTCGACAGACGAAAAACCACGTGGTATGGAAGCGGCTGAACAAG
CAGCGCCAGGCGGCGGTGGTCATGGTGTTTATTGTAACAAAATATTTATTTAA
AGTGATGGTTAGTTTTTGTAACAAGTAAGTCAGTGTTGAGCACTGGCTGGCAT
CGCCGTCTCGACTTTTACTAGGCGGCGCAGCCGATTAGCGTTTAAACTTTGGG
GTGTCGGCGGTTGTTTCCGTCCGGTGTCAATATTTTTTTCGCTTTTCCCACGGA
AGGAAAGGTAGCAATTGGGTCCGCTGGAACTCGGCTTCGCGACTGCCTTCTG
TGCCAAAGTGGCCAGAGACCCTAATAAGAGACATAAAGTTGAGTCCAGCAAC
CGACTGCCGTCGCGCTTCGTCCAACAGCAAACTACGAACAAAATCCCACGGG
CGGCGAGCACATTTCTGCTAACTAAGAGTCTGCCCGACAGAAACGAAATAAG
ATGCCATAGTCCTTGCCACCTGATACTGGTCACTGGTGAAGCCGCGTCGTCAC
CCACGTCGTCCGCTGTACCTGTGACAGCCAATTCATCACTCTCAGGTGCTTCG
ACCAGGAGGATCAACATCGATTGTGCATCAGTCCCATGTGGCCGGGAAGCGG

```
GCTTGTCTGCGATCCGGCTAGCTCTAGTCCAATCGATTTTGTCGGGGGCAAAT
TTGGAGTAAAGCATGCTACCTGTTCCAAGGGATGGCCACTCCATACTAACTG
CTCCAAAAGAAGGTCCCCATTCAGTTTAGCTTGTCCAGAGGGAGGCAAAAAA
TCTCGGCTAGACCCGCGGAGTTAACACGCAACGACCGGGTAATTCATCCCAG
CAAATCGGTCAGAGCCGCCATGGCACACTCACCACGGAGCGCTTGTGTTTTC
CAGAACTGAGGAAACAACGAGCGTCGCTATTAAGGCGCAGCACTAAACAGC
ATCAGTCACGCGCCGCTAAAAGAGCGAGTCCCAGTGGAGAGGTATTAAATAA
TTAAAGTAAGGTCTATCGGGTTATGAACAGTTCAGTTGGTTATAATATCCCGG
GTGGAATCGGAATATTGAGGTCCTTTACTTTTAACACTGAGCATATATTTGCC
GACATGAAAATTGCGGGCGCGTACGCTGGAGGAAAATGCTGCGCTGAAGGG
TGCGACTCGGAAGAAGAGAGTACTCAGCCACAGCGGCCAGTATCGAACCAA
ATGCAGTGAGGCCACGGCCGGCTGCAAACAGAAAACAGGGTTCCCCATACA
ACAAACGGGTGTTCACAGGATTCACAATCCCAAAAATCTAGTACCAAACACT
GTCAGCCTTCTCGCATCCACTAAGGATCTAAAATGTAACTTTTACAAGTAAAT
ATTTGCAACTAAGAGAATTTTAACGCAGCGGAAATATAGACACAAGACAAAT
CCAAATAGCACCTGCACGCATAACCGAAGGTATACAGGAAGATTCCTCCCAT
AAATGTACAGCTCGCGCCGATCCAGATAGGCAATTGAAGTACGTATGACCTT
CATAATTACCTTTTGCGATTGCATAGAGGCAACAAAGGGGTCTTGGGAAATG
AGGATGGAGATGGGATTTAAATGGTACGAGTAGTATAGTCGATTGCGTTGCA
CAAAAAATCCGCTAACGCTGCTGCATTTCTCATTTGACTCGACCATTGGGGAG
AGTCGTGAAACGCGCGGGTAGTAGGAAGGGGAAAGAGCGTGATACCGGCCG
TTTCACCGCCTTGCAAAGGCCCAGTATCTTTTTTAAGGAAAGAAAGTTCGAAG
CCGGAAAGCTTGCAGCTGCACCACTCGCTATTCCACCTCCACGCCGTGGTATG
CGTCTTGCCTCGCTCTCTGCGGTCGACTGCGAATTATGCACAGGGAACCCAAT
ATGTCAGTTATAACACCCACAAAGGGGACCGGGAAGAGGTGGTGAACTTATC
GAAGGGTTTATTTGGATCAAGCGAATATCGTCACCTTAGGGAGGCGATTGGG
GGCGCAGAAGCTATTGCTTCCACCAACTTATGGGTTTCTATTTTCCGATGGTG
CAGGGATTTATAGGGGTGAGTGTGTCTTCCTCGATACTTCGGTCAAACGTTAA
AATGTCCTCCAACTGTGGCATCCGTAAATTCGTTACGATCTTCCACCACGGAA
GGGGTGACATACAGCCTCCTTCGGGAGTTACCGTTAAAGAGAGAGAAGTGAG
TAACACATACGCGGGCTTGTGGCAATCAAGATGCGGATTTTGCTGAAACATG
CATTTTTGCAGCCTCAGAAAAAGATCCCAACGATGCACTTTTATGGTATATCC
ACCACTCGTCAGTTCCCAGGCCCTCTGCGTGCTTGCTGAAAGGGTGAGGGAG
TTGCAAGGGTAATGTTAGAAAATCGCTGCGCGTACTTTCGGCGCGATTTGTCT
ATTTACTTTAGCTTTCGATTGGTTAGGACCTCTATTTTCGGGTAAGCAATAAC
TCATTTCATTCTTCTCGTAGAAACGGTGGCTACACGAAATTCCAATGTTCCCT
TTCCGCGGAAAAGTCGCTTGTACGAGAAACTTGTGAAGCAGACGAGGGCGCC
CACTCGTTATGGGGGCATGGGAAAGGGTGGTGGTGAGAAACCATAGGACGC
GATCAAATGGCGTTTCTTTTGCCTTTTCCTTTCTCCGAAAAAAAGTGAATACC
GGAGAAGGACGTAATGCCGGATAAAATTACATCTCTCGTGGATACGAGGGGA
AACTTTCGTTCCCCTACGCAATGCTAAAGATAAATCAGTGGATTGCGAAAAA
GGTAACGGCAACTGGTCGATAAATTAGGAATAAAACTAAAACATAATGGCAC
TGCACCAAAAATTTTTAGCGTTTTCCTTGCTGTGGTGATGCCAAAAATGGGCG
TCGCCTCAACGAATAAAAAAAAATTGACCGATAAAAAAGTGAAGTTGGCCAC
TTTTGAGGGTGGGGGAGGGGGGAGGGACGCTAACCCTCACTTGCACGGCACT
ATTCAGAGGTTCCATTCATGTGTGTTTTTAAGTTGTGATAGATTTACAAAGGA
```

GATGGACATTGACTTGGGAAGTTTTCCGCCGCTCATTAAATTGTGCGGTAGTT
GGGAAGAGACGAATGGTTACGAACTGGTTGGATGATTCCTTTTCTCTTC*GTGG*
*AGACCCCAG*GGCACCTTGCATTTCCTCATGTTCACCACTCCCTCCCTTCTTACG
TGTACCAAAGAAAAAGAAGAGGAAAGAGGGGAGCACAAGCAACAAGTACCC
TGCCAGTTTTCGCCAACCGAATCTGACTAAGCCACCGGCCCTTTGGTGTCTGC
GTTTCCCGCCTCCCCGACGAGCAAAGGCCGTTGTGAAGTGGTGAAAAACGAG
GCCACTGGGGGGATCGAACCCCCGACCTCCGTCTTACTAGGACGGCGCTCTG
CCATTGAGCTAAGCGGCCACGCATGCCGTTGACCTTCACCGCCTTGTTTGCCT
TATTCCTGACAAGGAATCGAAAAGAGGGCCCAGCCGGGAATTGAACCCGGG
ACCTCTCCCACCCTAAGGGA*GAATCATGCCACTAGACCA*CTGGGCCGCGGGAA
CACCCCAATTTTTATTTTTCTTTTCTACTTTTAGAACGTTCATCCTGTTAGGTG
GAAAGGGTCTCTGCATCTCCCTATCCCGATTGGGCCGTACCGCACACTCAGA
GCGCTTCGGTGAGCACGTCTCCAGTCCCACTCCACAGGTTGTCTGACCTGCAT
CTTCATATATACAGAGAAATAGCAACTTAAATGAAAAAGAAAGCACATCCGC
AGGCAGGAAGAGAGGCTGCA

DQ246439
CGCTATTATTAGAACAGTTTCTGTACTATATTGAATCCACTACAAGACAG
CAGGCACAAGCTTCGATACCATCCAAATTAACAACAATTATGAGCATCAC
TTTTCATAGTTTATGGCTACTTCTGACAGTGTTGTGCACTGCAGGTATTCGTG
GTGATCGAGTCTGGTACGATTGTCCAGAAAAAGGTGTGGACACATCGCGGGA
CGGCATACAAGCTTTGTGCCGTGCGGCGGAGCAGTTCAGAGGCCTATCACAA
ACAGTAACATCTGCTGTGGAAACTTCCGCTACTGCTTCGAGTAAAGCATTTGA
AGCAAAGGTACAAGCAGAGGAAGCTGTGGAACTTGCCGAGTCAAAAGGCCT
AAACGTTACGAAAGCGAAGGAAGCTGCTGTGAGAGCAACACTCGCTGCTGA
AGCTGCGGCTACGGCTGCAAGTAATGTGGAAATTAACGCTGCAAATATTGCT
GCGGTGCCGTGGTCACAACCAAGCAGTGATGCAGGTTTACAGAAGCTGGCAC
TATGTGAAAACATCGACAAGAGTTTACGACAGTTGGCATCAGAGTGCTCGAA
GAGAGCGGAAAACGTGACAGCCCAATCGCTCAGTGAGGCGTTGGAGGGACT
AAGAAAACTACGCTACAATGATGTATACGTTAAGGAAATATTAGAGAGAGA
AGATGTTGAGTTCCACAAAGAGTTTATGTGGTTGCAGCACCACCTCCGGGAG
GCAGTTCATGCACGAAAACAGGCTGAGGATGCAGCTGCCGAGGCGAATGAA
ATTGCCGGCACAAATACGGGACCAGTTGGAAGTTCCGTTGCATCACCTGAAG
GGTCAGTGCTGCTACTGATGGCTGGACTGTTTCTCAGTTCTCTACTGTAAGA**G**
GTTACAACTCCATGAATTGTGATCCCAGCCATTCCACTTGTCTCCACACA
GTTATGTGCACATACAAGTGGGGCAGCAAACATTCTTTCATATTAACCTT
ATTTTCTATCAGTGACTATTAATATTGTTATTTTATTTTGTTTCCTTTATA
CTTTCACTTTTATTATTCTTTTTTACTTTACTTTTCCTTTTACAATGTTATC
TTTCGAACAGTTGTTGTAACTTTTATCCTTCATTTTTATTACTTCTTAACT
AGTGTGTCTATTGTCCTACTTTTTAGCTATATTTTTTCCTTTATTACCTTT
GTTATTTCATTATTTTATTCATGTAATGCACTGAGGTGATTGCGTGTAAG
CATGGCCACCTATGCGTTACAATAGTGACACTTTTTAAATCAAAATCTAT
CATGTCTACGCATCCATAATGTCCTATTGTATGCGCAAGGCTAACTGATT
CTGAGGGTTTTTATTTTATCATGAAAAAA

Z15031

TTTCTGTACTATATTG**CGTCCTTTCCCACAACGGTAAACATCTAAGAAATA**
ATGGCAGAGGAGACATCGTTGGTTGCAGATAAGGTTCCAGAGCCAGCGGTGA
TTGATGCCGTCGCAGATGCAATGCCGGACAGCCTCGAAGACGCTCTCCGGAT
TGTGTTGATGAAAGCTCGTGAGACGAATGGCCTCATTTGCGGCCTTTCAGAG
GTCACACGGGCCTTGGACCGCCGCACGGCTCACCTTTGTGTACTTGCTGATGA
TTGCGAGGATGAGGAGTACAAAAAGCTTGTTACTGCTCTCGCCAAACAAAAC
AACATTGACCTCGTAAGCATGGA*CGAGCGC*GAGAAACTCGCTCAATGGGCAG
GACTCACCAGAATGGCCGCCGACGGTTCGGTACGGAAGACGTTGAAGTGTTC
CTGCCTTGCTGTAAGGGATTTTGG*CGAGCGC*ACAAAGGCTCTTGACTACCTTC
TGTCGCAACTGCAGT**AATGTAGTGAACTGTCGCGGCACCGACATCAGCAC**
TGGAGTTTGTAGGAGTTTGTGGCACATGAGTGAAGGAGAGAGTACCTCG
GGAGTACGAGGGCGGGACTGTCATGTTTGTGTTTGCATTGTTGAGGTGG
TGTAGCATGTTGGAGGAAACGTTTTCCTATGTCCTCCATTACAGCTCGTT
CCATTTGGACGTTGCCCCTCTTCTGAGCTTATTGATATTTTCGTTGTTTA
AACAATGAAATTGATACTCCCCTTTTTCCCTTGTTGAACCCTCGATTCCT
CTGTGATTTTGCTCTTCTTACTTGTCTGCCGACCTTTCTCTTTGAGGACG
TCGATCTGTGGGAGGGAGACCACTTCTCTCTTAAAATGGTTTACTTTATT
ATTTTATTCAGTGATATAAACGAAAAAGAAAACAACAACAACAAAAAAAC
CAAAAGCTATCAGGAGTGACAAGGGTCTACGATGCATTACTTCAGTAAA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAA

AM295303

CGATCCCTAGAGTGATTGCACCTTGAGAACTGGTGGCAGCAGAGTTGTG
GATTACAATTGGTCTAGTGAAGTGAAAGTCTACCATATTGAGTGAGGGA
AGCTTTTACTAAAAGAAAAAAGGCCACGTGATCGACTTGTAATCAACCG
CTTTTGACAGGACGCGCGTGCAATAATAAGGGGGTTCGTATTCTTTTTTG
TTTTTTTTTTCACCAGGCAGTCCGTTGGATGCTAGCTCTTACATCACGCCG
GTTGCTGCTGCAGCAAACATTTATGCGATGCTGTAAGAGTGTAAACAGTGTG
ACCCTTGTTGGCGTCGTTCACGATATTCAGAGCGGCTTTGTATATGAAGATGC
CGTCACGCAGTTTACACTAACAACGACGAGCATTGACACGACACATCCAACT
CAGGAGGTTGTCGTCGAAAAGGACCACCACACGATCCGCTGCTTCGGCGAGC
TCTTCTCAGCGGAAGTGAAACAAAAGGTAAAGGAGGGAAACGTAGTGTGTGT
GAATGGGAGGCTTCGTCTTACCCCTCAACTTGAGCCTTCCTGCAACAAACACT
TTTACTTTCCATACATTCAGGTTCAGCCACCTCATGGCCAGGTGGCGGTCATC
CATGGCGACAGGCGGACGGTTCCCGCGGCAGTGAACCCTGCGGTGGAGGAT
ATAAAGTCGGAAAAGGAGGGCGCTGGTGGTGACCAGAGTGGTGTGCCGTCGT
AA**TGCCCTTGTGTTTCCGCAGTGTGAAGGAACTGTGATGGTTTTCTTCCT**
CACAGATGTGGAGGTTTCGTGAAGTGAATGGATTGGTGGGAAATCGCT
TTCCTCTGCTTTGTTTCGCGCTTCGGCGACCTTCGGTGTGTTTGGAATGC
CGGCGTTGTCAACTAAATTAATGGTGTGTGACTTTGTGGTGCTTCGGTTG
GTGAGTTTAATCGCAGGTCTCTGTGTTTC

AM168497

TTTCTGTACTATATT*GGAGGA***CTAGTATATTCGACCCTCAAGCGCGCAGCA**
TGCTTTCGTATACAGTGAA*GGAGGA*AGTAAAGGATGAAAAGCTCCCCGGAGC
TAACAACTTCGCTGACGTC*GGAGGA*TCATT*GGAGGA*TGAGGAACCCATGACG
CTGTACAC*GTGCGCG*AATTGCACGCGCGTTGTAGCATTATGCCCGTCCTCCCA
ACTCATGTGTACGCATTGCGTTCATCGTACGGGTTCCTCTACAGTGTTCTACA
AGGTTCGCCAGCAACCCACAACATATGACACTGTTTGATAACACAAAGCCAT
TTAGGAGATGGTGCCC**TTCCGGTTGATGCCTACATTGCCTCGGTTGTGGC**
GGCTCCCTGTTGTGCGAGTACGGAAGAGGTGACTGAATTGATTCTCTTT
CGCGCACTCCTCCGCGACGAACGCCGTCAGTTG*CACACATACACACACAC*
***ATACAC*AATTCTCTATGCAAAAAAAAAAAAAAAAAAAAAAAAAAAA**

AJ879575

ACAGTTTCTGTACTATATTG**TTATGAAATAGGTCCGGCGTGTTGGTACAAT**
AATGCGTAGGAACAATCGCACACGTTGCACTATTGATGATGTGAATGGAATG
CTTGCCAGAAACGCGCAGCTACGCAACGCCTTACAGGAAAGGTACAAGCAGT
TGAAAATGCGATATGAACAGCTGGCAGCTCTGAGAGCGGCCCTCTATCCTTC
GCGGGGGGTGACTCTGAGGAAACTAGGTGTAAGACAGGAAACTGCTGACGG
TGCTGAGGAGGTACGTTTCCTTGACGATTATACAACGGGAGGTGTGGGTAAT
CCCCCCTTTCGTGACGCTGGTATTTATTCAGCCAAGAATATTGTATGTTACGC
ACCGGTTCCACCTA*CACGAGACGAG*TTGCGATGGAAAGGAGTGACGTTGGCA
TTTCCACAGTTGGCATACGTACATTCTTTGGCAATGCTGCCGGAAACAGCGAA
CAGTTTTTCTAAAGCTCTCCAGTGGAGTCGAGAAGAAGATAGTGCACTCAGA
GAGCAGGTTCACGCCTACAA*GGGCGCAAGGTGTGGA*CCTTCTTTTGGAAAGC
ACTTGGTGCACCCGGTCAATCACGATTCGAAGTGGCGAGTCATTATATTAGGT
TGCAGCAGCTTGGACTCATCGAACCGGGGAAAAACGACAAAACGATGATGT
GGATCCCAGAGGAACAACGCGATATTGCTTTGCGGGAGGCAGTGTGGCGTCA
TTTAGGTGATGAGGGTGGCATTATGGCGGCTTACGTGGAAATTATTAGCGTTG
CAGCACGGAAATGTGTGTACCTTTCGGAGGTAACATGTAATGAATCACTTGT
GTTCCCTCCGTATGTGTGGGTGAAGCGTACCACCTTCAATTGGCTAACTTCAC
TTCTCATACAGAAAGCGAAAGCTCCTCTCGTGCGATGTGAAGGAGAATTTTC
AGATGATACACCTTTGTGCATGCACTTTAAGAGTGACATTTCCCTAAAACTTT
CTGCAGAAGATATGATGGCGTGTTTGTTGGCTTTCAAGGGTGAGGTGTTCGG
AGAGGTGGGTGGACTGCGGTTTATTGAGCGCGCGTTTTTGCCGAAGAATAGG
ACCTATGCACTGAAGGCCACAGACTTCATGGCTAAAAGAGAGGGCAAGAAA
CATAAACTGGATGAAGAAGTTTGAG**GTGCCTGGGGGAGATCAAAAATAAA**
GGAGCATAGTGCCAAGTAATGTGATTATCATGCACAATATATCACTGTTC
CTCACCTTGCGCGCACTCGCATATGAACTTGCTGTTGCGCCATATCTATG
TGTGCGTGGGTGTTACTGAGAACCCTCGCTTTCCTTTGCTTGTTTGCCGA
TGCTGTTGCTGCTGCGGTTCCTCCTGTCCATTTGTGTGAGTACTGCGTAT
AAATTAATGCGTAAATATATACGGCACTGAGGAGTTTGTATGTTTGTCTC
TGTAGTTTATCTCNCTGGGACTGAGTGTTTGAAAAAAAAAAAAAAATAAA
AAAAAAAAAAAAAAAAAA

L30155

TATTATTAGAACAGTTTCTGTACTATATTGGGTGTCAAACTACTGCCGCA
TAAACTACGGTTATCCCAAATTTAAGAGAAAGCAATAAAGCATCAATGTG
TGGAAAGGAAGTTGAAGGTGTTGTGAGTCCTGCGGCACAGCAGCAGCCAGCC
GTCCCGGAGGTAACAGATATCACGCTGGAGGCCGCCCGCAAGCAGAAAATTC
ACAACCTGAAGTTGAAGACCGCCTGCCTTTCGAATGAGGAATATGTCCAGGA
CCTGCACGTATCCGAGTGGAGTGAGACGCAGAAGCAGAAGCTGCAGGCTGC
ACACGAGAAAGCGCATGAATTGCTTGCCTCAGTGGAGGGTGGACGAAGTG
GAGCCTGACAGAGGCGTATGACATCAAGAAGCTGATGCGCGTCTGTGGTCTT
GAGATGTCTGTGCGTGAACTGTACAAGCCGGAGGACAAGCCACAGTTCATGG
AGATTGTTGCACTCAAGAAGACAATGAACGAACTGAAGCAACATCACAACA
AGACTCGCACGGTGTCTTTCACCGGCATGATCGACAATGCCATCGCCAAACT
GGAGAAAATCCAAGACGAACTGCGCCGGTCCCAGCTCGACGCTTCTGAGATG
GCGCAAGTTCCTGTGGCTGCACTGAAGAATATTGAGGACACGATGAACGTGG
CTGTTGTGCAGACGGCTCTTCTTGGGAACGAGGAGCAGATCAAAGCCCAACT
TGCAGCCGTTGAGAAGGCGAACGAAATCCGTAATGTTGCCATTGCCGATGGT
GAGATGGCGATTGCTGAGGAACAGTATTACATTAAGGCGCAGCTGTTGGAGC
ACCTTGTGGAGCTTGTGGCCGACAAGTTTCGCATCATTGGGCAAACTGAGGA
TGAGAATAAGAGCTTCAGTAAGATCCACGAGGTACAGAAGAAGTCATTTCAG
GAATCTGCCTCAATCAAGGACGCGAAGCGCCGCCTTAAGCAACACTGCGAGG
ACGACCTACGTAACCTTCACGATGCCATCCAGAAAGCTGACTTGGAGGACGC
CGAAGCCATGAAACGGTTCGCCACGCAGAAGGAGAAGTCGGAGCGGTTCAT
CCACGAGAACCTCGACAAACAGGACGAGGCATGGCGTCGCATTCAGGAACT
GGAGCGCGTGTTGCAGCGCCTTGGGACGGAGCGTTTTGAAGAGGTGAAGCGC
CGTATTGAGGAGAACGACCGCGAGGAGAAGCGTAAGGTGGAGTACCAACAG
TTCCTCGATGTATGTGGCCAGCATAAAAAGCTGCTGGAACTGTCTGTGTACAA
CTGCGACCTTGCGCTTCGCTGCATGGGTATGCTGGAGGAGATCGTAGCCGAG
GGCTGCAGTGCCGTCAAGTCACGCCATGACAAGACGAACGATGAGTTGTCTG
ACCTTCGGCTGCAGGTGCACCAGGAGTACCTGGAGGCATTCCGTCGCCTGTA
CAAAACTCTTGGCCAGCTTGTGTACAAGAAAGAAAAGCGCCTGGAGGAGATT
GATCGCAACATCCGCACCACACACATTCAACTGGAGTTTGCCATTGAGACCT
TTGACCCCAACGCGAAACTACACTCCGATAAGAAGAAAGACCTATACAAACT
TCGTGCGCAGGTGGAGGAAGAGTTGGAGATGCTGAAGGACAAGATGGCGCA
GGCGTTGGAGATGTTTGGACCTACTGAGGATGCGCTGAACCAGGCTGGTATC
GATTTTGTTCACCCTGCTGAGGAGGTTGAGTCCGGCAACATGGATCGCCGCA
GCAAGATGGTGGAGTACCGTGCACACCTGGCGAAGGAGGAGGAGGTGAAGA
TTGCCGCGGAGCGCGAGGAGCTGAAACGATCTAAGATGCTCCTGAGCCAGCA
GTACCGCGGCCGCACGATGCCCGAGATCACTCAGTAGC**GCTGCGCTTAAAT**
GTCTTTCATTATAATCAATGTATAACCTTTATGTAGTATTTCAATCTATGC
CGCTGTGTACGTGCACTGCGGTGCCTATCCTTCGGCATTAGAGAGTCAC
TGTTTGTGTAGATCGTAGCTGCATGTCTG

AY157307
AACGCTATTATTAGAACAGTTTCTGTACTATATTGGCAAGACATACTGGG
GGTAATATCAAGGGTGGCACACACTAAGCAGAGGAGGGGACGCCAAAC
GAAAAAAGAAAGAAACCTGCACTGACCAAAAAGACTGAACGAAACGAAA
TTGAGGCGATTGAGACGCCCCTTTCTGTAAGCGGGGTTTAGTTCATATTC
GAACGGAATGGGAGGGTGTACCTCACGTGGGCTCTCAGAAGAGAAACTCGC
ATGTTACTCCCACCGTACCGGCAACCTCGTTGATGAGCATCTTT*CGACAGGGG*
*C*AGTTGAGGCACACGAGCTTCAGCCCTTCTTTTCTTCTCTGCTGGGAGCCATC
ACGGACCTTTTGAAGTGCAGCCGGGAGGATGCCGTCGAGTTCCTTGCATGCT
CAAGTAGTGCAAACCCACGGGCGGCAGAGTTGTTTACTTCCTTTTGTGCTGCA
AATCCGCTGAATCTCATAAAATGGGATGTGAATCACGCGAAGTTCATGATGA
TATGGATAAAGTACGACGACGACAACAGCGGGGACATATGCGTTCGTGAATT
GAGGAAAATCTTGAAGGGTTTGAGCTTCCCTGAAAGGCTCTCACAAAAGATG
ATTGACGAGCTGGAAGCTACAGGGGGGAGGGCAAGCTACAAGTTGATGCAG
GGGACATTTATGTCTCTGACAAGACTTAATGAACTGACATATGCAATGCGAA
ATGTCGTGGGTCCCGATCGGGACGTGATGACAAAGGCTGAGTTTGTTACCTTC
CTGAAGGAAACTCAAGGAGAGGGTGCCGATGGTGAGGAGCTGCACGTATTTC
TAGACGCTATTGGCTGTACCGAAGAGCATCCTCTACATTTGGACGCATTCTTA
TCATTTCTCAG*CGACAGGCGC*TTTAACTCCATTGTGAACAACAGAAAGGTGTC
TAGTGTTTACCACGATATGACTCGCCCGATATGTGAGTATTTCATCAATTCCT
CGCACAATACCTACCTTACGGGTGATCAACTCTTGAGCAAATCTTCCACGGAT
ATGTACAAGAGGGTTCTACTGGATGGCTGCCTCTGCGTTGAACTTGATTGTTG
GGATGGTCGCAAGGGTCAGCCTGTAGTTTATCATGGTTACACAAGGACTTCC
AAGCTTTGGTTCCGGGACTGTATTAGCACGATCAAGAAGTATGCTTTCGTTAA
TTCAATATACCCTGTCATTTTGTCGCTTGAGGTTCACACTAGCCTCCGCCAAC
AGGATCGAATGGCGGAAATTTTGTGTGAAACGCTCGGAGATATGCTATTCTG
CAGTCCTTGGGGTGCTGGTGAACAGACTTCTTTCACGTTCTCGCCGGAAGCGC
TAAAGGGAAAAATTCTGCTAAAGAGCAAACGGGCTACTACACCTACCGATGG
GGTACAGGTTGATGATGACGACGATGAGGATGAGGAAGCCGATGGTGTGGT
GGAAAATTTCGTACCACCTGAAACTGCTCGGCGTTGTCGCGGTGGTGGAAAA
ACGAATTCAAGGGGTGCTGAAAAGAAAAAGAAGGTTTCAAAGGTTTCAGAG
AAACTATCTCGTCTTATTTCAATCGAATCCATTGGTTATAAGGGTGTTGAGGA
TCTAAGTTATCTTGAAACGCGTCAACCATATCACTGCAGCTCTTTTACTGAAG
GGAAAGCGGGGAAAATTGCCTCTTCTAACCAGGAGGAGTTTGTTGCCGTCAA
TAATCGGTGTTTGAGTCGCATATACCCCACAGGAACTCGCATCGGTAGCAGT
AACTTTCATCCTCAAACGTTTTGGAATTGTGGGTGCCAACTAGTTGCACTCAA
TTGGCAGAATTACAAGTCATACCAGCTTAGGCTAAATAGGGGGTTCTTCAGC
GACAATGGTAACTGCGGCTACCTTCTCAAACCGACTGCTGTGGACATTGCAC
GTGCAAGGGGGCCAAAACGGCAGTCACGGTTGCTCACAATAGAAATTATATC
AGCTTTCTGTCTTCCCAGGCGGAAAAATGCATCCGGTAGCAGTATTGTGGATT
CTCGCGTAGCCGCCTTGATTGAGGGCCCCGGCATGGAGAAAAGCCAACGAAA
CACATCTCCCATTCACAATAATGGCTTTCATCCTGTCTGGCGGGGTGAGCGCC
TAAACAACGAGTTCTGCTGGAAGGTGTACGAGTGGGAACTGTCCACCCTTGT
CATGCAGGTGTATGACGAGGATACCAAAAGCAATAACCTTCTGGGTGAATAT
GTTGTGCCATTACGTGCCCTAAAAGAAGGAATTCGCCAGGTCCCCCTTCGAG
ACCTCAAAGGATCTATTATACATGGCTCTTTTTAATGGTTCAAGTATCTTATC

AGTAGGA**GTTTGAGAATATCGTGTTCTTCAATTGGGTTACAAGTGTGCGT**
TGCATGTACCCGACTTTATGGCACCATGTTGTGCCATGGTCTTCACAGCG
TAGCTCATTTTTACTGATATATATATATATATTTAAACATATATTTTTCTT
GACACTGTTATTTTCATTTTTGTTGGCCGTTCTATCGTGCCTTCGGTGAG
TATTACACTCATCACGACATCATATATGAACACACCGACGGTGTTGTTTT
GTTTTGTTTTTTTTTTGCGTGTGGGTAAAATCACCAGGGGATTCCCAATG
TTTCGACCGAGGATAATGTTTCATTCAGTATCTACTTTTCCAATTCATCA
CTTCATTTGAGTTTATTTCTTTGTTGTCTCCGCTTACTTTCATGTCCTCAC
ACCGTGAGGCAGGATAGGTGAGCTGAACGATTTTTTTTTCCTTTGTGTGT
GGTGGTTAGAAGGAAGGAACGTAATATGGTAATTGGGCTTATATGGGTC
TGATGCTTCATTTTGTCCTTCCCTTCGCACACGTGCTTCCAGCTACGTGT
TCGTATTTCTGTAAATGTAAATTCGTTTCGTAGAACGGATGCCTGTTTCC
TCTGCCGATTGTTTCGCTGCTCAAGGGACTGCGGCGGTCGCAGTTAGGT
AACTGATTTTCCGTTCCACATCCCTACTTTCTAGTTGATCGAATATACGG
AAATTAAGCTACAAAAGATATCGGAAGTCAATGATGGTCCAGACGAAAA
AGTGCATCGCCACTTT*CCGGGGCC***ACAGGCAAGATATGTGGATGGCAA**
TAACAAAGAAATTTTCCTCAGCCCCTTTTCCCTTTTCAAAAAAAAAAAAAA
AAAAAAAAAAAAAAAAAAAAAAA

Z54338

CTGTTTCCACATTG*TCGTCGGG*GGTTGTGTCGCTTCTGTAGACTTTCCATTTTC
TATTTGGGGACTCTTTTGAAACCTCTCTGACCATAATTTGCTTTCATTCCTTCT
TCTGTTTCGTCTTTTCCAATACTCGATGGCAGCCTCCCTTTCCAGTAGCACAAT
GGCAAAGAAGGTCAAGTCGAAGGTGGACACCATCAA*CACCAAGATCC*AACTG
GTGATGAAATCCGGCAAATACGTTCTCGGGACGCAGCAGTCACTCAAGACAC
TTCGTCAGGGCCGCAGTAAACTCGTTGTCATTTCCGCTAACTGCCCGCCGATC
CGCAAGGCGGAGATTGAGTACTACTGCACTTTAAGCAAGACGCCAATTCACC
ACTACAGCGGCAACAACCTTGACCTTGGAA*CGGCATGCGGA*AGGCATTTCCG
TGCTTGCGTACTTTCCATTACGGATGTTGGTGACTCTGACATCACTTCTGCAT
AATCGCAACGGTGTAGGTGTGTGCCGTATGTCCTTACCGAGAGTCGTTCAAGT
GATTT

M81386

TACGCAATCAGGCAAACTCATATAAATTGTCCAGTGACCCAAAACAATGA
CTGAGCGTCGTGATAACGTTTCCCACGCACCTGATGCCATCGAGGGCCCAAA
CGATGGTGCTCACGCCGAGGACACATCACCGGGTTTCTTCTCATTCGAGAATC
TTGGTGTGGCGCAGGTACAAGTCGTCGGAGGAACTCTGAACGGATATGTTAT
TGGGTATGTTGCTGTGTATTTGCTGCTGTACCTGACGGCAACTGAGTGCAAAT
TTACTACGGAGGGTGCATGTGGCGGAGCTAAGATATATGGGTGCAAATGGAG
CGGCACGACATGCAAATTCGAGAATCCCAAATGTAGTGAAGGCTCCGATCCT
TCCGATTCTTGCAAAAACGAGGTGGCCTACACGTCTGTTTACAGCGGTATCTT
TGCCTGCGCCATGATCGTTGGTTCAATGGTTGGGTCGATTATAGCCGGGAAGT
GCATCACTACGTTTGGACTGAAGAAATCATTTATCATTGTCTCAATCACTTGT
ACTATAGCTTGTGTTGTGGTGCAGGTAGCGATTGAGTACAATAATTACTACGC

ACTGTGCACTGGACGAGTGCTCATAGGTCTCGGCGTTGGTATTTATGCTCTG
TTTGCCCCATGTATGTGAATGAGAACGCACATCCCAAACTCTGCAAGATGGA
CGGTGTGTTGTTTCAGGTGTTCACAACACTTGGCATTATGCTTGCCGCGATGC
TGGGTCTGATTTTGGACAAAACAGGAGCTAGTAAAGAAGAGGCAAACAT*GG*
*CTGGGCG*GTTACACGTTTTTTCAGCGGTACCGCTTGGATTGTCCGTCGCCATG
TTCCTAGTGGGCATGTTCCTCCGCGAAAGCACTGCAACATTTGCCCAAGACG
ATGATGGTAAGGCTGATGGCGGAATGGACCCCAACGAGTAT*GGCTGGGGG*CA
GATGTTGTGGCCACTGTTCATGGGCGCTGTAACCGCTGGTACGCTGCAGCTGA
CTGGGATCAACGCGGTAATGAACTATGCGCCGAAGATTACAGAGAACCTCGG
AATGGATCCATCACTTGGCAACTTTCTGGTTATGGCATGGAATTTTGTGACAT
CCCTTGTGGCTATTCCACTTGCGTCACGCTTTACGATGCGTCAAATGTTTATC
ACCTGTTCCTTTGTTGCGTCATGTATGTGCTTGTTCCTATGTGGAATCCCAGTG
TTCCCCGGTGTTGCAGGAAAAGAGGTGAAGAATGGTGTGGCAACTACTGGTA
TCGCCCTGTTCATAGCTGCATTTGAGTTTGGTGTTGGGTCGTGCTTCTTCGTGC
TTGCACAGGACCTTTTCCCACCATCATTCCGACCTAAGGGCGGTTCGTTTGTT
GTCATGATGCAGTTTATCTTTAACATCCTTATTAACCTATTGTATCCCATTACA
ACTGAAGCTATATCTGGTGGGCCAACTGCCAACCAGGACAAGGGACAGGCC
GTTGCATTCATACTGTTTGGTTTAATTGGCTTGATTTGTTCCGTTCTGCAGTTC
TTTTACTTGTATCCATATGATGCCAATCAGGACCATGAGAATGACCATGGTGG
TGAGCCTGTGGAACAGAAGACATATCCCGTTGAAGCATCTCCGCGGAACTAA
AGCACTGAAGCTTAATTCATCTGGTGAGGTATTGTTTGTCTCGCTCGCAT
GACTCATGTGCTGGGGAGGTGTAAAGGGGGGATGGCGACGAAGTTGTTT
CTTGCATATTTCTCGCGCATCTGATGAATTAAAAAAAAACGATTATTGCA
TAACATGATTATCTGACCACAAAACGTTTTGTAGTTTGAAGGAGGTAATT
GGGTAATGTTTTAGAGGTCGTCAATATTAGTGGCGTTAATGAAAACGGA
TTTTAAAATTTACTTCTTTTTGCTGTTTTATGTTGTCTATATATACTTTTG
TTTTTCCATCAAGTCGACTGTGCCTATTATTATCTGCTCGGTTTTGTAGC
AGCGGATGGACAGATGGATGAAGTGATATATGAGGGCAGTATGCTGTTA
GTGTGTATGTGCACTCTAAAGCTGCTGCTGTGTCGGGATAGTGATTTAC
GTAGGGCAGTTGATTTTTCTTTTTTCCTTTTTTTTTGTAATATTATACAAT
TGAAGACGTTTCTTAATAGTTTTTTGAAAAAAAAACATTGTGTTGTTTTTT
ATTGGTGTACAGGGGACAACTTGTTTTATT

APPENDIX C

UTRSCAN OUTPUT FROM CHROMOSOME II

This appendix shows the results of submitting the first 330,000 bases of chromosome II as input into UTRscan.

```
--------------------------------------------------
          Pattern = Histone 3'UTR stem-loop structure
--------------------------------------------------
     Pattern not found

--------------------------------------------------
          Pattern = IRE
--------------------------------------------------
     Pattern not found

--------------------------------------------------
          Pattern = SECIS-1
--------------------------------------------------
     Pattern not found

--------------------------------------------------
          Pattern = SECIS-2
--------------------------------------------------
     Found 3 matches in 1 sequences

seq :[152066,152127] :GATA ATGTATGGA A TGAA AGTGTGGA   AAC
AAGGTTGAGAGAAA  TCATGTG TGAG ACTG TATT
seq :[171238,171301] :ATGA CAAAT A TGAT GTTACCAT  TTA AAA
AATAAGTAGCAATA  GTAGTAG GGAG AAAGCTA TTGT
seq :[313819,313877] :CGCT GACAACT A TGAT GTTTTAGA   AAG
CAGCGGCATGGTCG
GTGAAAC GGAC TAC AGCG

--------------------------------------------------
          Pattern = APP
--------------------------------------------------
     Pattern not found
```

```
--------------------------------------------------
            Pattern = CPE
--------------------------------------------------
     Pattern not found

--------------------------------------------------
            Pattern = TGE
--------------------------------------------------
     Found 3 matches in 1 sequences

seq :[16493,16527] :GTCA ATTGAATATCT CA TTTC TT GTATGTT TTTCT
seq :[53295,53331] :CTCA CACTGAGGCCGCA CA TTTC TT TCAATTG TATCT
seq :[138215,138251] :CTCA CACTGAGGCTGCA CA TTTC TT CCAATTA TATCT

--------------------------------------------------
            Pattern = NANOS_TCE
--------------------------------------------------
     Pattern not found

--------------------------------------------------
            Pattern = 15-LOX-DICE
--------------------------------------------------
     Found 153 matches in 1 sequences

seq :[6,21] :CCCTATCGCT CAA ATG
seq :[266,281] :CTCCACCCCT TTC AGG
seq :[2743,2757] :CCCCACCTCG AT ATG
seq :[3062,3078] :CCTTACCCCT CACA ACG
seq :[5563,5581] :CCCTTCCTTC TCCATT AAG
seq :[9871,9889] :CCCTCACTCT GAGTAA ATG
seq :[11047,11060] :CCCCGCCCGT A ACG
seq :[18933,18952] :CCCTTCATCC TCTGCGC ACG
seq :[22569,22582] :CTCTGTCTCT G AGG
seq :[23118,23137] :CCTCGCCCTT CCCCGGG ACG
seq :[23191,23209] :ACCGTCCTCC TTCACT ATG
seq :[24782,24801] :CCCCTTCCCT GCCAGTA AAG
seq :[26891,26907] :CCCTAACTCT GCCA ACG
seq :[32386,32405] :CCCCGTACCC TTCCTAG AAG
seq :[32735,32750] :CCCCACCTCT TCG ACG
seq :[34158,34176] :CCCTTCCTTC TCCATT AAG
seq :[35291,35306] :CACACCCTCC GAG AGG
seq :[37159,37179] :CCCTACCCTTC ACAAAAT AAG
seq :[37460,37476] :CCAACCCTCC TGCG AGG
seq :[39542,39560] :CCCTTCCTTC TCCATT AAG
seq :[42814,42833] :TCCCGTCTCC ATTCAAA AAG
seq :[45668,45682] :CCCTATCTCT AC ACG
seq :[45933,45952] :CCTCACCTCC GGCACTC AAG
seq :[46083,46097] :CCGCTCCTCC TA ATG
seq :[46104,46118] :TCCCTCCTCT TT AAG
seq :[47343,47357] :CGCCTCCTCC GA AAG
seq :[48263,48281] :CTCCATCTCT CAGTCC ACG

seq :[48712,48728] :CCCCTTCCTT CAAT AGG
seq :[51793,51807] :CACATCCTCC CT AAG
seq :[52928,52943] :CCCATCCTGT CGT ATG
seq :[54150,54165] :CCCTATCGCT CAA ATG
seq :[54412,54427] :CCCCACCCCT TTC AGG
seq :[56896,56910] :CCCCATCTCG AT ATG
seq :[57215,57231] :CCTTACCCCT CACA ACG
seq :[59739,59757] :CCCTTCCTTC TCCATT AAG
seq :[60892,60907] :CACACCCTCC GAG AGG
seq :[65131,65149] :CCCTTCCTTC TCCATT AAG
seq :[66259,66274] :CACACCCTCC GAG AGG
seq :[68127,68147] :CCCTACCCTTC ACAAAAT AAG
seq :[68428,68444] :CCAACCCTCC TGCG AGG
seq :[70510,70528] :CCCTTCCTTC TCCATT AAG
seq :[73798,73817] :TCCCGTCTCC ATTCAAA AAG
seq :[76652,76666] :CCTTATCTCT AC ACG
seq :[76917,76936] :CCTCACCTCC GGCACTC AAG
seq :[77088,77102] :TCCCTCCTCT TT AAG
seq :[78327,78341] :CACCTCCTCC GA AAG
seq :[79244,79262] :CTCCATCTCT CAGTCC ACG
seq :[79331,79344] :CCTCATCCTC A AAG
seq :[79693,79709] :CCCCTTCCTT CAAT AGG
seq :[86382,86400] :TCCTATCTCT ACACAG ATG
seq :[86647,86666] :CCTCACCTCC GGCACTC AAG
seq :[86818,86832] :TCCCTCCTCT TT AAG
seq :[88057,88071] :CACCTCCTCC GA AAG
seq :[88980,88998] :CTCCATCTCT CTGTCC ACG
seq :[89432,89448] :CCCCTTCCTT CAAT AGG
seq :[91140,91155] :CACACCCTCC GAG AGG
seq :[92989,93009] :CCCTACCCTTC ACAAAAT AAG
seq :[94052,94068] :CCCCTTCCTT CAAT AGG
seq :[97485,97503] :CCCTATCTCT ACACTG ATG
seq :[99759,99772] :ACCTTCCTCC G AAG
seq :[101976,101994] :CCCTTCCTTC TCCATT AAG
seq :[104542,104556] :CACATCCTCC CT AAG
seq :[108116,108130] :CCCTATCTCT AC ACG
seq :[108381,108400] :CCTCACCTCC GGCACTC AAG
seq :[112664,112682] :CCCTTCCTTC TCCATT AAG
seq :[116949,116967] :CCCTCACTCT GAGTAG ATG
seq :[118166,118184] :CACCGCCCTT GCCAAC ACG
seq :[122748,122761] :CGCTGTCCCC A ACG
seq :[122769,122785] :CCTCACCCCC GCAC AAG
seq :[124499,124512] :CCCCGCCCGT A ACG

seq :[137843,137858] :CCCATCCTGT CGT ATG
seq :[139317,139332] :CCCCACCCCT TTC AGG
seq :[141788,141802] :CCCCACCTCG AT ATG
seq :[142107,142123] :CCTCGCCCTCC ACA ACG
seq :[143895,143912] :CCCCACTCCT CTCTT ATG
seq :[144749,144764] :CTCCACCTCT TCG ACG
seq :[146166,146184] :CCCTTCCTTC TCCATT AAG
seq :[150463,150481] :CCCTCACTCT GAGTAG ATG
seq :[151680,151698] :CACCACCCTT GCCAAC ATG
seq :[156217,156230] :CGCTGTCCCC A ACG
seq :[156238,156254] :CCTCACCCCC GCAC AAG
seq :[157969,157982] :CCCCGCCCGT A ACG
seq :[172656,172671] :CCCTTTCCTCC CG AGG
seq :[173468,173483] :CCCCACCCCT TTC AGG
seq :[175945,175959] :CCCCACCTCG AT ATG
seq :[176660,176673] :CTCTGTCTCT A AGG
seq :[177209,177228] :CCCCGCCCTTC CCCGGG ACG
seq :[177282,177300] :ACCGTCCTCC GTCACT ATG
seq :[178874,178893] :CCCCTTCCCT GCCAGTA AAG
seq :[180982,180998] :CCCTAACTCT GCCA ACG
seq :[186813,186829] :CCCCTTCCTT CAAT AGG
seq :[190228,190242] :TCCTATCTCT AC ACG
seq :[190493,190512] :CCTCACCTCC GGCACTC AAG
seq :[190664,190678] :TCCCTCCTCT TT AAG
seq :[191401,191415] :CCCGGCCTCT GC AGG
seq :[191901,191915] :CGCCTCCTCC GA AAG
seq :[192597,192615] :CTCCTCCTCC GTCCAT ATG
seq :[197810,197823] :ACCTTCCTCC G AAG
seq :[198065,198083] :CCCCACTTCT TTGGCA ATG
seq :[200000,200018] :CCCTTCCTTC TCCATT AAG
seq :[201083,201096] :ACCCACCTCC T AAG
seq :[201099,201114] :CACACCCTCC GAG AGG
seq :[204477,204491] :CTCTTCCTCC TC ACG
seq :[204774,204793] :CCCCTTCCCT GCCAGTA AAG
seq :[206490,206508] :CCCCGTCCACT GCAGT AAG
seq :[206886,206902] :CCCTAACTCT GCCA ACG
seq :[212206,212224] :CTCCATCTCT CAGTCC ACG
seq :[212293,212306] :CCTCATCCTC A AAG
seq :[212655,212671] :CCCCTTCCTT CAAT AGG
seq :[219344,219358] :CCCTATCTCT AC ACG
seq :[219609,219628] :CCTCACCTCC GACAACT ATG
seq :[222123,222142] :CCCCGTACCC TTCCTAG AAG
seq :[223882,223900] :CCCTTCCTCC TCCATT AAG



seq :[228156,228174] :CCCTCACTCT GAGTAG ATG
seq :[229373,229391] :CACCGCCCTT GCCAAC ATG
seq :[233949,233965] :CCTCACCCCC GCAC AAG
seq :[257254,257268] :CCCCTGCTCT CA ATG
seq :[261219,261236] :CCCCATACTT CCCTC ACG
seq :[263044,263061] :CCCCCGCTCC GAGTA ACG
seq :[263647,263664] :CCTCACCCCC ACGCC ACG
seq :[266171,266184] :CCTCATCTCC C ATG
seq :[266572,266590] :CCCCGCTCTT TTGATC AAG
seq :[268058,268074] :GCTGCCCTCC GTAG AAG
seq :[268119,268136] :CCCCCGCTCC CGTCA ACG
seq :[271503,271521] :TCCCGCCTCC CCTCTA ATG
seq :[273981,274000] :CCCCGTCTCA TCGGGG AAG
seq :[274202,274221] :CCCTGCCCCCT CCACCG AAG
seq :[275600,275614] :CCACTCCTCT GA ATG
seq :[276178,276194] :CCCCACCTCT AGAA ATG
seq :[276779,276796] :CCCATCATCC TCTTA ATG
seq :[279558,279575] :CCCCATACTT CCCTC ACG
seq :[281383,281400] :CCCCCGCTCC GAGTA ACG
seq :[281986,282003] :CCTCACCCCC ACGCC ACG
seq :[284508,284521] :CCTCATCTCC C ATG
seq :[284909,284927] :CCCCGCTCTT TTGATC AAG
seq :[286395,286411] :GCTGCCCTCC GTAG AAG
seq :[286456,286473] :CCCCCGCTCC CGTCA ACG
seq :[287840,287858] :CCCCTTCCCT TAACTG AGG
seq :[289096,289111] :CCCCACCGCC AGG ATG
seq :[299888,299905] :ACCACCCTCC AGAAC ACG
seq :[304302,304315] :CCCCCCCCCC G AAG
seq :[305119,305132] :CCACATCCTT G AAG
seq :[305596,305613] :CGCTATCCCT TGTGG ATG
seq :[309015,309034] :CCCTCCCTGT GCTATCG AAG
seq :[315061,315074] :CCCTAACCCT C ATG
seq :[316412,316426] :CCCCATCTGC GC AGG
seq :[317728,317747] :CCCCGTCCTG AATTGCC ATG
seq :[321022,321040] :CCGCACCCCC TTGGGT ATG
seq :[322115,322129] :CGCACCCTCC AG ATG
seq :[325080,325094] :CACCTCCTCT CA ATG
seq :[326850,326865] :CTCCATCCCT TAT AGG
seq :[328116,328130] :CACCTCCTCT CA ATG
seq :[329943,329959] :TCCTGCCTCC CAAC ATG
```

```
-----> Checking repeats for  15-LOX-DICE  (min: 2)
    Found 0 matches for pattern 15-LOX-DICE

-----------------------------------------------------
             Pattern = ARE2
-----------------------------------------------------
    Found 0 matches for pattern ARE2

-----------------------------------------------------
             Pattern = TOP
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = GLUT1
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = TNF
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = VIMENTIN
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = IRES
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = MSL2-5UTR
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = MSL2-3UTR
-----------------------------------------------------
    Pattern not found

-----------------------------------------------------
             Pattern = RPMS12_TCE
```

```
---------------------------------------------------
     Pattern not found


---------------------------------------------------
             Pattern = BRE
---------------------------------------------------
     Pattern not found

---------------------------------------------------
             Pattern = ADH_DRE
---------------------------------------------------
     Found 11 matches in 1 sequences

seq :[12902,12909] :AAGGCTGA
seq :[85602,85609] :AAGGCTGA
seq :[96691,96698] :AAGGCTGA
seq :[107326,107333] :AAGGCTGA
seq :[123908,123915] :AAGGCTGA
seq :[126344,126351] :AAGGCTGA
seq :[159819,159826] :AAGGCTGA
seq :[218561,218568] :AAGGCTGA
seq :[237528,237535] :AAGGCTGA
seq :[319145,319152] :AAGGCTGA
seq :[320720,320727] :AAGGCTGA

---------------------------------------------------
             Pattern = BYDV
---------------------------------------------------
     Pattern not found

---------------------------------------------------
             Pattern = Proneural-Box
---------------------------------------------------
     Pattern not found

---------------------------------------------------
             Pattern = K-Box
---------------------------------------------------
     Found 34 matches in 1 sequences

seq :[7894,7901] :ATGTGATA
seq :[15701,15708] :GTGTGATA


seq :[18483,18490] :CTGTGATA
seq :[30753,30760] :CTGTGATA
seq :[41487,41494] :GTGTGATA
seq :[51179,51186] :GTGTGATA
seq :[56646,56653] :GTGTGATA

seq :[72471,72478] :GTGTGATA
seq :[82204,82211] :GTGTGATA
seq :[100710,100717] :ATGTGATA
seq :[103920,103927] :GTGTGATA
seq :[118548,118555] :ATGTGATA
seq :[128291,128298] :GTGTGATA
seq :[141538,141545] :GTGTGATA
seq :[152062,152069] :ATGTGATA
seq :[154474,154481] :CTGTGATT
seq :[158999,159006] :CTGTGATA
seq :[162618,162625] :GTGTGATA
seq :[165402,165409] :CTGTGATA
seq :[184835,184842] :CTGTGATA
seq :[191776,191783] :TTGTGATA
seq :[210685,210692] :CTGTGATA
seq :[214120,214127] :ATGTGATA
seq :[215167,215174] :GTGTGATA
seq :[224794,224801] :ATGTGATA
seq :[229755,229762] :ATGTGATA
seq :[232168,232175] :CTGTGATT
seq :[243103,243110] :CTGTGATA
seq :[259844,259851] :TTGTGATA
seq :[265300,265307] :CTGTGATA
seq :[278192,278199] :TTGTGATA
seq :[283639,283646] :CTGTGATA
seq :[316371,316378] :CTGTGATC
seq :[321231,321238] :CTGTGATT

-----------------------------------------------------
              Pattern = Brd-Box
-----------------------------------------------------
    Found 16 matches in 1 sequences

seq :[11993,11999] :AGCTTTA
seq :[20307,20313] :AGCTTTA


seq :[121436,121442] :AGCTTTA
seq :[125435,125441] :AGCTTTA
seq :[154951,154957] :AGCTTTA
seq :[158910,158916] :AGCTTTA
seq :[168381,168387] :AGCTTTA
seq :[207973,207979] :AGCTTTA
seq :[232648,232654] :AGCTTTA
seq :[236619,236625] :AGCTTTA

seq :[246732,246738] :AGCTTTA
seq :[256155,256161] :AGCTTTA
seq :[274599,274605] :AGCTTTA
seq :[276022,276028] :AGCTTTA
seq :[276853,276859] :AGCTTTA
seq :[303252,303258] :AGCTTTA

------------------------------------------------------
            Pattern = GY-Box
------------------------------------------------------
     Found 11 matches in 1 sequences

seq :[3482,3488] :GTCTTCC
seq :[57635,57641] :GTCTTCC
seq :[64594,64600] :GTCTTCC
seq :[80769,80775] :GTCTTCC
seq :[99110,99116] :GTCTTCC
seq :[99878,99884] :GTCTTCC
seq :[135443,135449] :GTCTTCC
seq :[144077,144083] :GTCTTCC
seq :[221800,221806] :GTCTTCC
seq :[264797,264803] :GTCTTCC
seq :[283136,283142] :GTCTTCC

------------------------------------------------------
            Pattern = Androgen-Receptor
------------------------------------------------------
     Pattern not found

------------------------------------------------------
            Pattern = Elastin G3A
------------------------------------------------------
     Pattern not found

------------------------------------------------------
            Pattern = Insulin 3'UTR stability
------------------------------------------------------
     Pattern not found

------------------------------------------------------
            Pattern = Beta-actin 3'UTR zipcode
```

```
-------------------------------------------------------
     Pattern not found


-------------------------------------------------------
              Pattern = Gap-43 stabilization element
-------------------------------------------------------
     Pattern not found

-------------------------------------------------------
              Pattern = Dendritic localization element
-------------------------------------------------------
     Pattern not found
```

REFERENCES

1. GeneDB. Retrieved March 24, 2007 from http://www.genedb.org.

2. Gopal, S., Cross, G., and Gaasterland, T. (2003) An organism-specific method to rank predicted coding regions in Trypanosoma brucei. *Nucleic Acids Research, Vol. 31, No. 20, pp. 5877-5885.* http://nar.oxfordjournals.org/cgi/content/full/31/20/5877?ijkey=02Akzed6YFYVI&keytype=ref.

3. MEME. Retrieved April 2, 2007 from http://meme.sdsc.edu/meme/meme.html.

4. Motif-er. Retrieved April 8, 2007 from http://xanthos.bioinformatics.rit.edu/~shuba/bin/motif-er.cgi.

5. NCBI. Retrieved April 6, 2007 from http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?CMD=search&DB=nucleotide.

6. ORBIT. Retrieved April 9, 2007 from http://xanthos.bioinformatics.rit.edu/~shuba/bin/orbit.cgi.

7. Pesole, G. and Liuni, S. (1999) Internet resources for the functional analysis of 5' and 3' untranslated regions of eukaryotic mRNA. ***Trends in Genetics,*** *Vol. 15, No. 9, pg. 378.* http://www.ba.itb.cnr.it/BIG/UTRScan/TIGUTR.pdf.

8. The Trypanosoma brucei Genome Project. Retrieved February 16, 2007 from http://www.sanger.ac.uk/Projects/T_brucei/.

9. UTRscan. Retrieved April 11,2007 from http://www.ba.itb.cnr.it/BIG/UTRScan/.

10. UTRsite. Retrieved March 24, 2007 from http://bighost.ba.itb.cnr.it/srs6bin/wgetz?-e+[UTRSITE-ID:*].