# Chimp Chunk 2-8 Annotation Report

Alan Tseng

3/30/07

*Introduction*

In order to better understand the process of annotating a genome, my partner and I worked to annotate chimp chunk 2-8 in an attempt to characterize the key features contained in the given sequence. The first step in our annotation analysis was to run RepeatMasker on our sequence with the –nolow switch which allows us to mask repeat regions without masking low complexity regions. With this masked sequence, we then carried out a GENSCAN prediction of the possible genes in our chimp chunk, resulting in two preliminary features, outlined in Figures 1 and 2. Further analysis with BLAST and the UCSC database of these two features would help us confirm or reject of the actual presence of orthologs of genes in humans.

Predicted genes/exons:

| Gn.Ex | Type | S | .Begin | ...End | .Len | Fr | Ph | I/Ac | Do/T | CodRg | P.... | Tscr.. |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1.01 | Init | + | 2319 | 2400 | 82 | 2 | 1 | 93 | 72 | 55 | 0.026 | 5.53 |
| 1.02 | Intr | + | 4970 | 4998 | 29 | 0 | 2 | 89 | 115 | 26 | 0.020 | 3.23 |
| 1.03 | Intr | + | 24197 | 24230 | 34 | 1 | 1 | 87 | 43 | 44 | 0.120 | -2.40 |
| 1.04 | Term | + | 24408 | 24532 | 125 | 1 | 2 | 11 | 42 | 224 | 0.704 | 8.65 |
| 1.05 | PlyA | + | 36821 | 36826 | 6 | | | | | | | 1.05 |
| 2.00 | Prom | + | 58234 | 58273 | 40 | | | | | | | -5.86 |
| 2.01 | Init | + | 62948 | 63260 | 313 | 1 | 1 | 63 | 105 | 368 | 0.960 | 31.49 |
| 2.02 | Intr | + | 65930 | 66268 | 339 | 0 | 0 | 52 | -14 | 436 | 0.649 | 25.25 |
| 2.03 | Intr | + | 67760 | 67934 | 175 | 0 | 1 | 78 | 100 | 183 | 0.797 | 17.50 |
| 2.04 | Intr | + | 71727 | 71808 | 82 | 0 | 1 | 76 | 41 | 67 | 0.707 | 0.44 |
| 2.05 | Term | + | 72260 | 72433 | 174 | 1 | 0 | 75 | 54 | 92 | 0.380 | 2.26 |
| 2.06 | PlyA | + | 74115 | 74120 | 6 | | | | | | | 1.05 |

Click here to view a PDF image of the predicted gene(s)
Click here for a PostScript image of the predicted gene(s)

Predicted peptide sequence(s):

>chimp2-8.fasta|GENSCAN_predicted_peptide_1|89_aa
MPLELHPVGGLTRRDDVVGTDTTNRIAEHNWSQRVHKLWTSSGAGVETGSADMRFERKAH
ALKAMKQYYGTPLAGRPVNIQLVTSQIDT

>chimp2-8.fasta|GENSCAN_predicted_peptide_2|360_aa
MGRGLWEAWPPAGSSAVAKGNCREEAEGAEDRQPASRRSAGTTAAMAASGPGCRSWCLCP
EVPSATFFTALLSLLVSGPRLFLLQQPLAPSGLTLKSEALRNWQVYRLVTYIFVYENPIS
LLCGAIIIWRFAGNFERTVGTVRHCFFTVIFAIFSAIIFLSFEAVSSLSKLGEVEDARGF
TPVAFAMLGVTTVRSRMRRALVFGMVVPSVLVPWLLLVSLNTPTSDGLTYCYSIDLSERV
ALKLDQTFPFSLMRRISVFKYVSGSSAERRAAQSRKIVEPATQALAAVTWSRRSIWFRSC
CEQNHFGPNPTSSSVYPASAGTSLGIQPPTPVNSPGTVYSGALGTPGAAGSKESSRVPMP

Figure 1: GenScan textual output

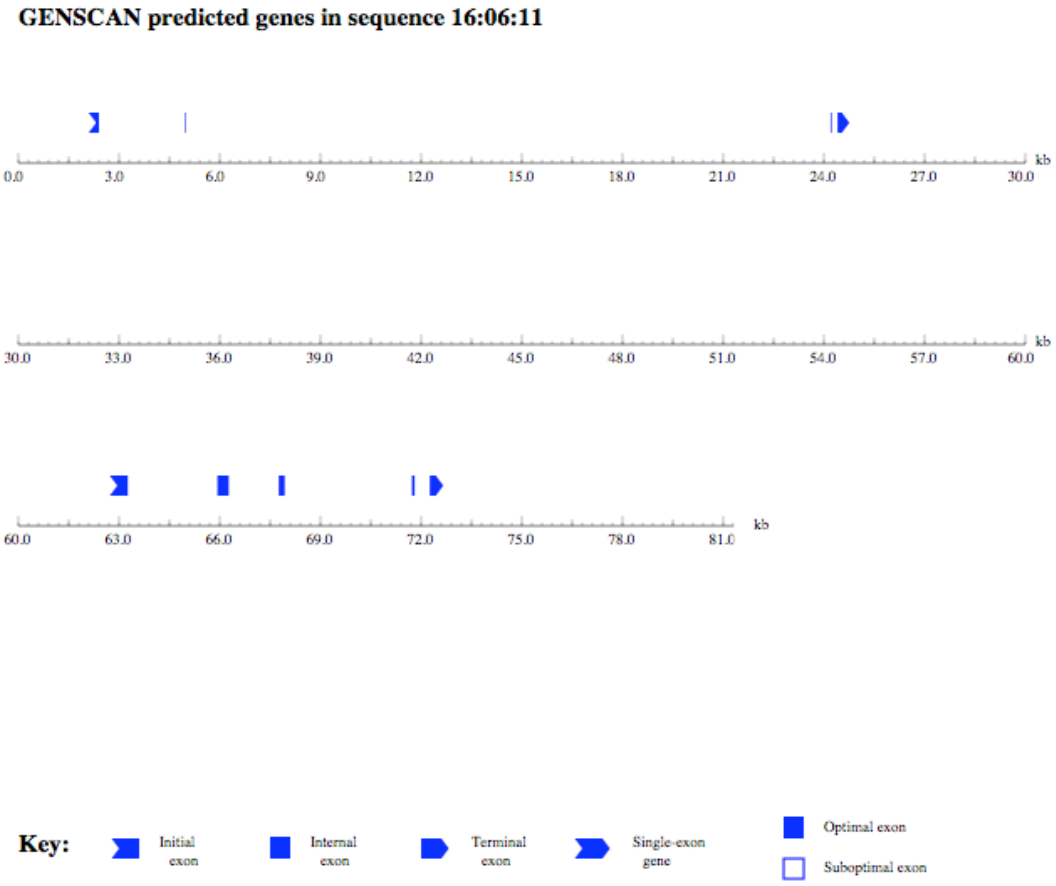
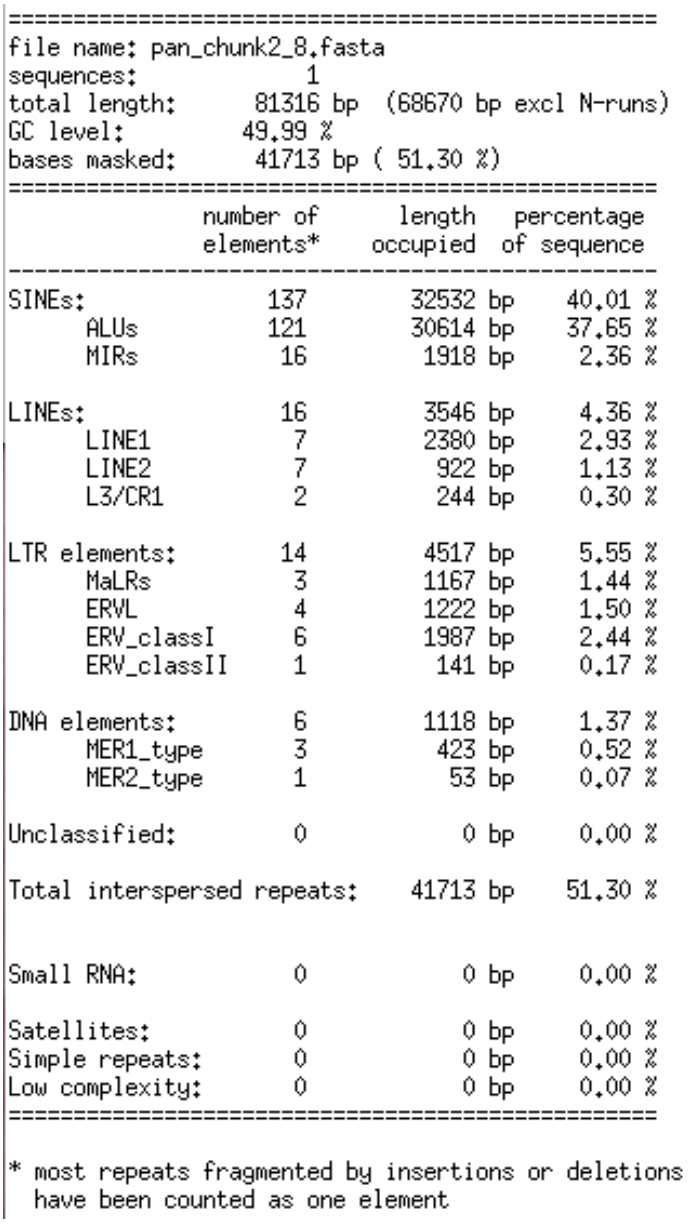



Figure 2: GenScan map of predictions



```
==================================================
file name: pan_chunk2_8.fasta
sequences:             1
total length:      81316 bp  (68670 bp excl N-runs)
GC level:         49.99 %
bases masked:      41713 bp ( 51.30 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              137        32532 bp    40.01 %
      ALUs          121        30614 bp    37.65 %
      MIRs           16         1918 bp     2.36 %

LINEs:              16         3546 bp     4.36 %
      LINE1          7         2380 bp     2.93 %
      LINE2          7          922 bp     1.13 %
      L3/CR1         2          244 bp     0.30 %

LTR elements:       14         4517 bp     5.55 %
      MaLRs          3         1167 bp     1.44 %
      ERVL           4         1222 bp     1.50 %
      ERV_classI     6         1987 bp     2.44 %
      ERV_classII    1          141 bp     0.17 %

DNA elements:       6         1118 bp     1.37 %
      MER1_type      3          423 bp     0.52 %
      MER2_type      1           53 bp     0.07 %

Unclassified:       0            0 bp     0.00 %

Total interspersed repeats:    41713 bp    51.30 %


Small RNA:          0            0 bp     0.00 %

Satellites:         0            0 bp     0.00 %
Simple repeats:     0            0 bp     0.00 %
Low complexity:     0            0 bp     0.00 %
==================================================

* most repeats fragmented by insertions or deletions
  have been counted as one element
```

Figure 3: Results of RepeatMasker

*Repeats*

From the results of RepeatMasker, there is one non-SINE repetitious element that is over 500 bp in this chimp chunk. This element was identified by RepeatMasker as follows:

Position: 62200 – 62806, Repeat Type: LTR26, Class/Family: LTR/ERV1

*Feature 1: CCL24*

The first GenScan prediction consisted of two parts, one from 2319 to 4998 bp, and the second from 24197 to 36826 bp. Running a BLAT search in the UCSC database resulted in the two parts matching to different parts of chromosome 7 of the human genome; the first part (2319 to 4998 bp) to a small inducible cytokine A24 precursor gene in humans and the second to a THO complex subunit 4 protein (Figure 4). To confirm that the first part of this feature was in fact a gene, a BLASTP search also resulted in the small inducible cytokine gene (CCL24). However, this match only matched the last 55 residues of a 119 residue long protein. To account for the first 64 residues, we returned to the original unmasked chimp chunk sequence to run a BLAST2 search using the region of interest. The resulting match extended the previous match by another 31 amino acid residues, this time with high identity (99.0%). These results from BLAST2, when run on BLAT in the UCSC database, aligned to two out of three exons as seen from the Refseq evidence. To try and match the last 30 residues of the missing exon, we looked further into the unmasked chimp sequence. What was discovered was that the chimp chunk sequence was truncated by a series of N's (denoting regions have yet to be sequenced) at a point where the first exon of CCL24 should be located. Due to this truncation, it was very likely that BLAST2 did not display the alignment for the last exon due to low alignment score (due to short length of the exon, truncation, lower identity compared to the other results). However, performing BLAT on the sequence near the truncation revealed high identity to the human genome (98.0%). The results of this BLAT search is shown in Figure 5. The first residue, a methionine, was truncated due to incomplete sequencing.

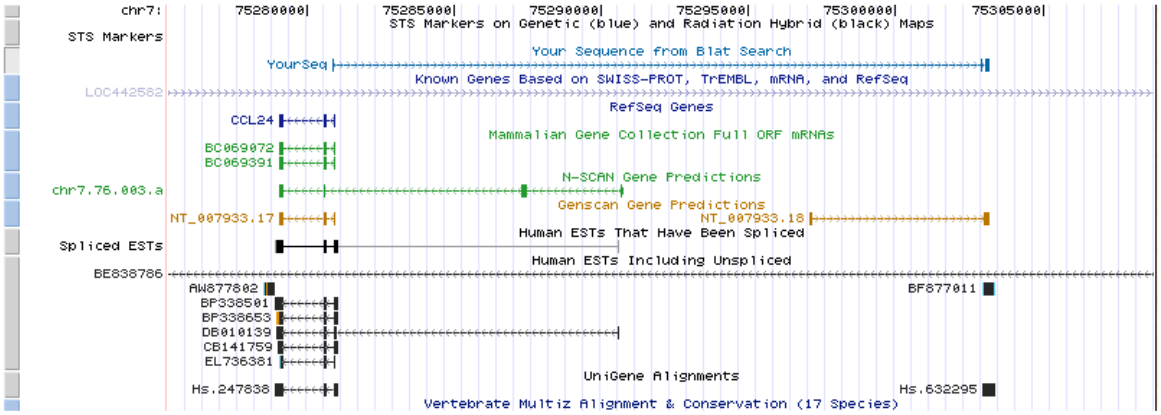



Figure 4: BLAT result of prediction 1

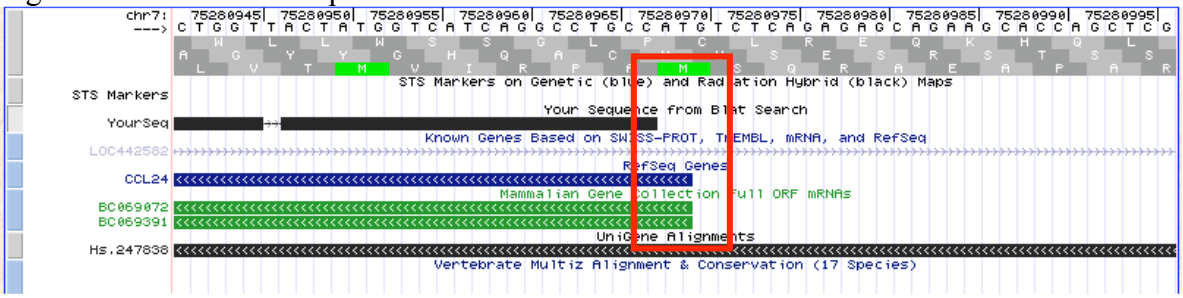



Figure 5: Truncation of the first residue

To provide further evidence that this feature is in fact the ortholog of CCL24, a BLAT search was carried out with the human CCL24 Refseq sequence to search for the actual location of the CCL24 gene. What resulted from this BLAT was only one match to the same location as feature one on chromosome 7, confirming the location of our feature. Finally, EST evidence from Herne also supported our hypothesis, with alignments to the last two exons of CCL24, but none in the truncated exon 1. From these results, we concluded that our first feature was the ortholog of gene CCL24, a small cytokine CC gene from the cytokine family of secreted proteins, important in immunoregulatory and inflammatory processes.

Now that we have ascertained the presence of an ortholog for the first part of feature one, we must now look into the secdon part of feature one of the GenScan predictions.

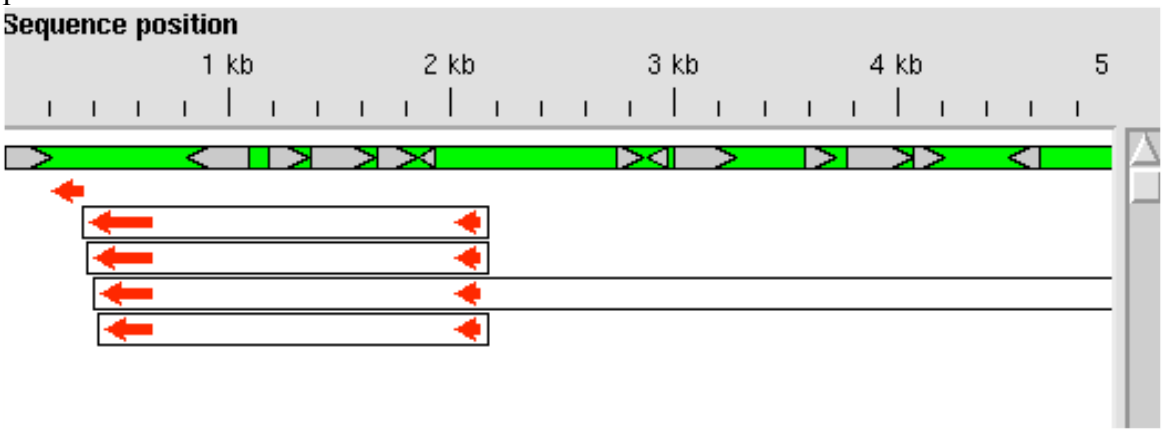



Figure 6: Herne alignment of EST matches in CCL24

*Feature 2: THOC 4*

The next region on our first GenScan prediction (from positions 24197 to 36826 bp) matched to a THO complex subunit 4 protein. Running a BLASTP search matched the GeneScan prediction to the human THOC4 sequence with high identity, but only 46 out of 230 residues were matched. Running a BLAST2 alignment between the Refseq THOC4 sequence and the unmasked chimp chunk sequence produced a longer match, from residue 81 to 174 and 212 to 230, but the identity deceased down to 50% due to significant gaps in the alignment. These misalignments suggested that our second feature might be a pseudogene of the THOC4 gene in humans.

To ascertain our hypothesis, we ran a BLAT alignment with the Refseq THOC4 sequence in the human genome. This resulted in a 100% identity match to the THOC4 gene on chromosome 17 (shown in Figure 7), which showed that the actual THOC4 gene did not lie within our chimp chunk on chromosome 7.

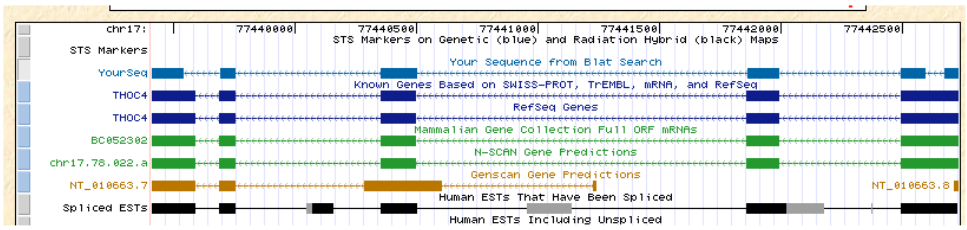



Figure 7: BLAT of Refseq THOC4 with human genome on chromosome 17

Further investigation by a BLAT alignment with the human THOC protein and the chimp genome revealed that the Refseq THOC4 sequence also matched to a region on chromosome 7 chimp, at same location predicted by GenScan. This match, however, is low in identity (80.3%), and contains only one out of five exons in the THOC4 transcript (Figure 8). Therefore, we conclude that this GenScan prediction is in fact a pseudogene of the THO complex subunit 4 protein. In addition, this pseudogene region is flanked by SINE elements on each side, which strongly suggests that this pseudogene occurred via a transposition event. However, further analysis must be done to confirm the events that resulted in this pseudogene.

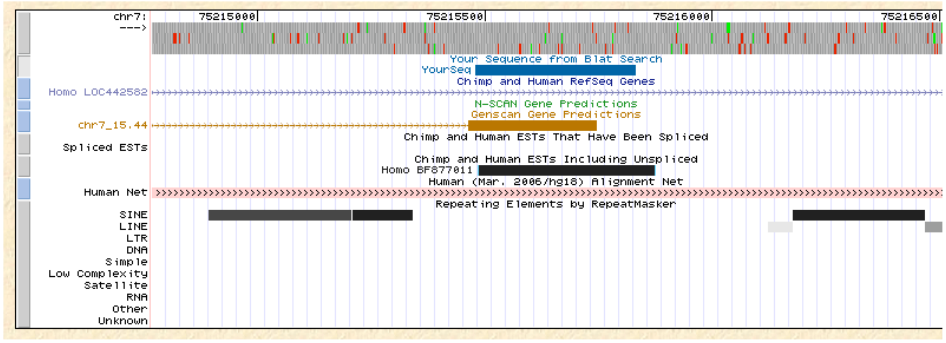



Figure 8: BLAT of Refseq THOC4 with chimp sequence on chromosome 7

*Feature 3: Rhomboid*

The last feature produced by GenScan was a gene identified from the Rhomboid family of proteins. This feature extended from 62948 to 72433 bp and contained five exons. An initial BLAT alignment of the prediction aligned the sequence at a location on chromosome 7 that had Refseq evidence for a 4 exon Rhomboid gene. This alignment produced an identity score of 99.5% and also covered 360 amino acids out of a total of 364 amino acid protein. Thus, we were fairly confident that this GenScan prediction was in fact a gene ortholog. To provide further evidence for this hypothesis, we used the unmasked chimp chunk sequence in a BLASTX search to compare our entire sequence to the protein database on NCBI. The results of the BLASTX showed matches to two major isoforms of a rhomboid containing protein 2 (RHBDD2) gene, isoform CRA a and isoform CRA b. The results of this BLASTX is show in Figure 9 below. Since isoform CRA a matched completely to our unmasked chimp chunk sequence (100% identity), it was determined that this isoform was the correct isoform to be annotated.

In addition, there was a considerable amount of EST evidence for this gene. An alignment on Herne displayed four specific columns of EST clusters that suggested 4 exons. These clusters were mirrored in the Refseq alignment on Herne.

```
>gi|119592179|gb|EAW71773.1|  rhomboid domain containing 2, isoform CRA_a [Homo sapiens]
Length=409

 Score =  257 bits (657),  Expect = 8e-66
 Identities = 118/119 (99%), Positives = 119/119 (100%), Gaps = 0/119 (0%)
 Frame = +1

Query  9127  RLNPVPGSYPTQSCHPHLSPSHPVSQTQHASGQKLASWPSCTPGHMPTLPPYQPASGLCY  9306
             +LNPVPGSYPTQSCHPHLSPSHPVSQTQHASGQKLASWPSCTPGHMPTLPPYQPASGLCY
Sbjct  291   KLNPVPGSYPTQSCHPHLSPSHPVSQTQHASGQKLASWPSCTPGHMPTLPPYQPASGLCY  350

Query  9307  VQNHFGPNPTSSSVYPASAGTSLGIQPPTPVNSPGTVYSGALGTPGAAGSKESSRVPMP  9483
             VQNHFGPNPTSSSVYPASAGTSLGIQPPTPVNSPGTVYSGALGTPGAAGSKESSRVPMP
Sbjct  351   VQNHFGPNPTSSSVYPASAGTSLGIQPPTPVNSPGTVYSGALGTPGAAGSKESSRVPMP  409


 Score =  204 bits (518),  Expect = 1e-49
 Identities = 136/136 (100%), Positives = 136/136 (100%), Gaps = 0/136 (0%)
 Frame = +3

Query  2982  VYRLVTYIFVYENPISLLCGAIIIWRFAGNFERTVGTVRHCfftvifaifsaiiflsfEA  3161
             VYRLVTYIFVYENPISLLCGAIIIWRFAGNFERTVGTVRHCFFTVIFAIFSAIIFLSFEA
Sbjct  105   VYRLVTYIFVYENPISLLCGAIIIWRFAGNFERTVGTVRHCFFTVIFAIFSAIIFLSFEA  164

Query  3162  VSSLSKLGEVEDARGFTPVAFAMLGVTTVRSRMRRALVFGMvvpsvlvpwlllGASWLIP  3341
             VSSLSKLGEVEDARGFTPVAFAMLGVTTVRSRMRRALVFGMVVPSVLVPWLLLGASWLIP
Sbjct  165   VSSLSKLGEVEDARGFTPVAFAMLGVTTVRSRMRRALVFGMVVPSVLVPWLLLGASWLIP  224

Query  3342  QTSFLSNVCGLSIGLA  3389
             QTSFLSNVCGLSIGLA
Sbjct  225   QTSFLSNVCGLSIGLA  240


 Score =  173 bits (439),  Expect = 2e-40
 Identities = 103/104 (99%), Positives = 103/104 (99%), Gaps = 0/104 (0%)
 Frame = +1

Query  1    MGRGLWEAWPPAGSSAVAKGNCREEAEGAEDRQpasrrsagttaamaasgpgCRSWCLCP  180
            MGRGLWEAWPPAGSSAVAKGNCREEAEGAEDRQPASRR AGTTAAMAASGPGCRSWCLCP
Sbjct  1    MGRGLWEAWPPAGSSAVAKGNCREEAEGAEDRQPASRRGAGTTAAMAASGPGCRSWCLCP  60

Query  181  EVPSATFFTALLSLLVSGPRLFLLQQPLAPSGLTLKSEALRNWQ  312
            EVPSATFFTALLSLLVSGPRLFLLQQPLAPSGLTLKSEALRNWQ
Sbjct  61   EVPSATFFTALLSLLVSGPRLFLLQQPLAPSGLTLKSEALRNWQ  104


 Score = 99.8 bits (247),  Expect = 3e-18
 Identities = 50/50 (100%), Positives = 50/50 (100%), Gaps = 0/50 (0%)
 Frame = +3

Query  4839  GLTYCYSIDLSERVALKLDQTFPFSLMRRISVFKYVSGSSAERRAAQSRK  4988
             GLTYCYSIDLSERVALKLDQTFPFSLMRRISVFKYVSGSSAERRAAQSRK
Sbjct  242   GLTYCYSIDLSERVALKLDQTFPFSLMRRISVFKYVSGSSAERRAAQSRK  291
```

Figure 9: BLASTX of unmasked chimp chunk

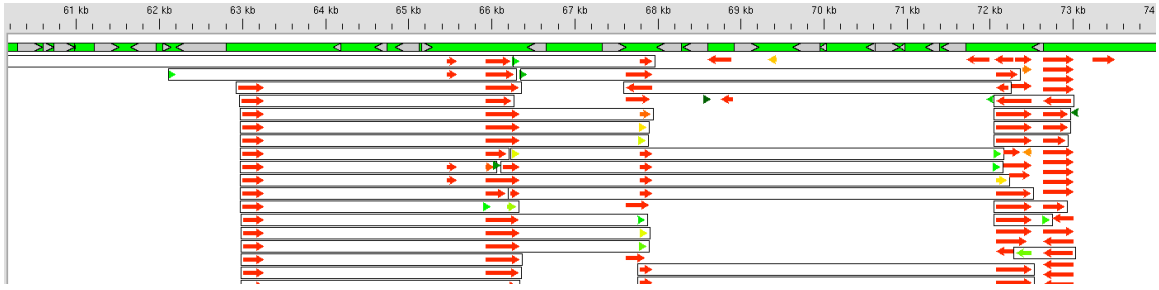



Figure 10: Herne alignment of ESTs in RHBDD2

With all of the collected evidence, we annotated this region as an ortholog to the gene RHBDD2 CRA a. This gene codes for a protein from the Rhomboid family of proteins that contain integral membrane proteins found in bacteria and eukaryotes. Rhomboid promotes the cleavage of the membrane-anchored TGF-alpha-like growth factor Spitz, allowing it to activate the Drosophila EGF receptor.

*Conclusions*

The following Figure is a map of the annotated chimp chunk 2-8. This map shows three predicted genes: a CCL24 ortholog, a THOC4 pseudogene, and a RHBDD2 CRA a ortholog. In addition, the 607 bp repeat found by RepeatMasker was also included.

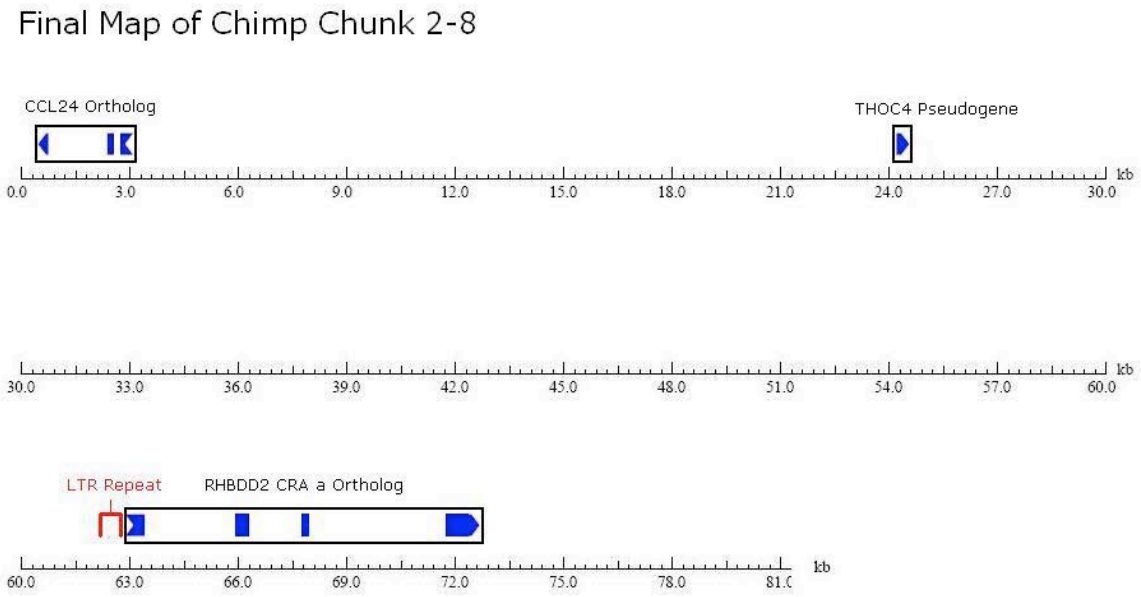



Figure 11: Final map of Chimp Chunk 2-8