Annotation of Chimp Project 2-8

Bill Dirkes

4/10/2008

Introduction

Annotation is the process of identifying genomic features on a sequence. This includes the identification and demarcation of putative genes, pseudogenes, repeats, and non-protein coding sequences. This paper examines annotation of project 2-8 from chimpanzee by J. Kuhn and myself. While the chimpanzee is closely related to the human (98-99% for protein-coding sequences), the chimp is not close enough to the human to allow for accurate predictions relating to non-protein coding sequences. Thus, my analysis will utilize Genscan predictions, EST (Expressed Sequence Tag) evidence, BLAST searches for homology, and Repeat Masker to identify putative genes, pseudogenes, and repeats. We found that the sequence was 52.3% interspersed repeats, contains two putative functional genes, and one pseudogene (Table 1).

Rhomboid domain-containing 2 – functional gene; 4 exons; strong evidence for 3' UTR
MPIF-2 – functional gene with 3 exons identified; spans from about .5kb-2.5kb
THOC4 – pseudogene with one exon identified; spans from about 24kb-25kb

Table 1: Annotation of the features on project 2-8.

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	0.022	5.53
1.02	Intr	+	4970	4998	29	0	2	89	115	26	0.017	3.23
1.03	Intr	+	24197	24230	34	1	1	87	43	44	0.122	-2.40
1.04	Term	+	24408	24532	125	1	2	11	42	224	0.692	8.65
1.05	PlyA	+	36821	36826	6							1.05
2.00	Prom	+	58234	58273	40							-6.56
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	0.791	17.50
2.04	Intr	+	74973	75088	116	0	2	82	96	26	0.012	2.89

Predicted peptide sequence(s):

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

>chimp2-8.fasta|GENSCAN_predicted_peptide_2|315_aa
MGRGLWEAWPPAGSSAVAKGNCREEAEGAEDRQPASRRSAGTTAAMAASGPGCRSWCLCP
EVPSATFFTALLSLLVSGPRLFLLQQPLAPSGLTLKSEALRNWQVYRLVTYIFVYENPIS
LLCGAIIIWRFAGNFERTVGTVRHCFFTVIFAIFSAIIFLSFEAVSSLSKLGEVEDARGF
TPVAFAMLGVTTVRSRMRRALVFGMVVPSVLVPWLLLVSLNTPTSDGLTYCYSIDLSERV
ALKLDQTFPFSLMRRISVFKYVSGSSAERRAAQSRKQGRTHSGQRSFQTLSVSLLGWRRA
AEKQIHALLALQTSX

Table 2: Genscan's predicted proteins.

Initial Genscan Results

Genscan is a program designed to predict protein products using only the DNA sequence as input. On this project, Genscan initially predicted two genes (Table 2, Figure 1). I researched the first predicted gene, while J. Kuhn researched the second. Both predicted genes contain four exons, although the second predicted gene does not include a terminal exon. Either the intron extends beyond our project, and thus the exon may be in a neighboring project, or Genscan did not correctly predict the gene model.

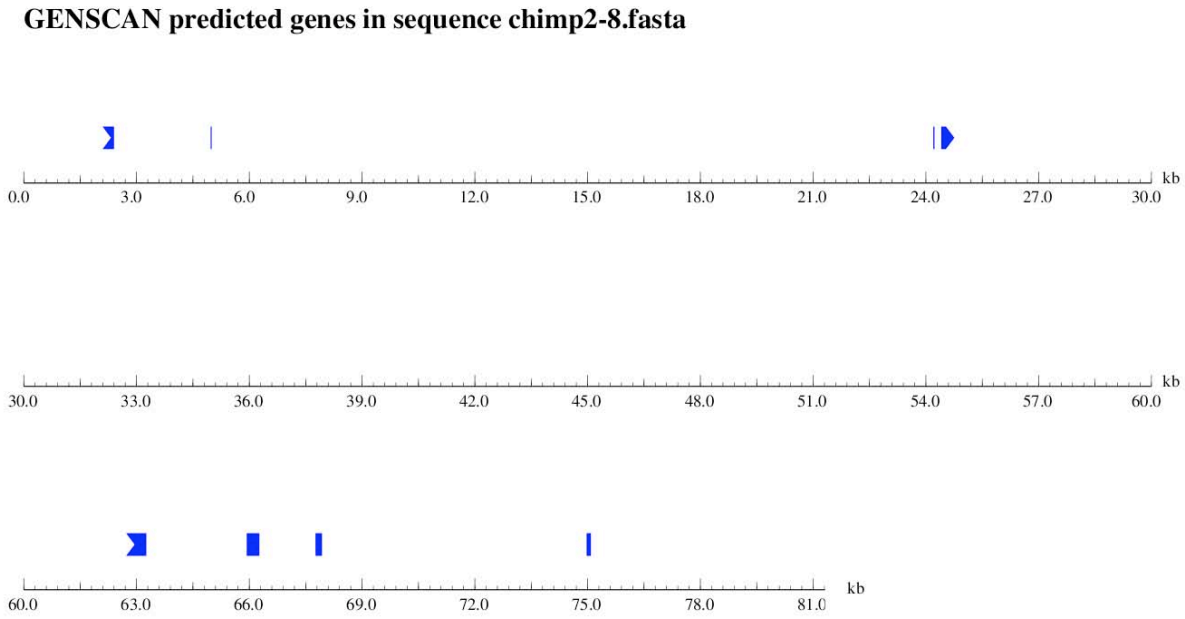

Figure 1: Genscan predicts two protein sequences.

Predicted Gene 1

Using the first predicted protein in a Blat sequence alignment search against the human genome (hg17) resulted in a match showing a gene model with two exons, including a start codon, canonical splice sites, a stop codon at the end, and no stop codons in the open reading frame (Figure 2). However, the predicted peptide does not have any matching orthologs. The first exon appears to overlap with an exon from a human gene, but the predicted peptide is transcribed from the opposite DNA strand and has different start and end coordinates. In addition, the first exon has lower conservation than would be expected (~90%). BLAST revealed further problems with Genscan's predicted protein product.

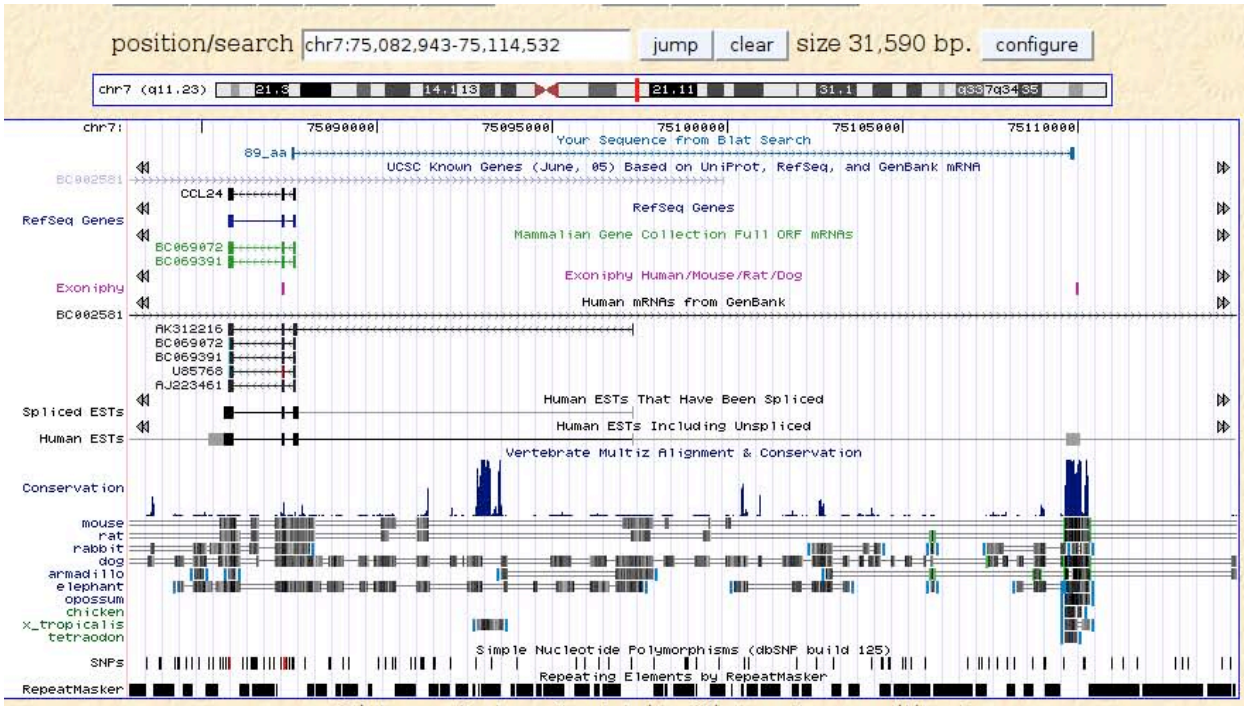

Figure 2: Blat results of the predicted peptide do not match any known human proteins.

Then I ran blastp to analyze the Genscan predicted protein sequence against the nr (non-redundant) protein database (Figure 3). BLAST found a weak match to the THO Complex 4. This match has only 71% identity, and less than one fifth of the true protein product aligns to Genscan's predicted peptide. This protein, according to the SwissProt database, "acts as chaperone and promotes the dimerization of transcription factors containing basic leucine zipper (bZIP) domains and thereby promotes transcriptional activation."

```
>ref|NP_005773.2| THO complex 4 [Homo sapiens]
 sp|Q86V81|THOC4_HUMAN THO complex subunit 4 (Tho4) (Ally of AML-1 and LEF-1) (Transcriptional
coactivator Aly/REF) (bZIP-enhancing factor BEF)
 gb|AAH52302.1| THO complex 4 [Homo sapiens]
Length=257

 GENE ID: 10189 THOC4 | THO complex 4 [Homo sapiens] (Over 10 PubMed links)

 Score = 70.5 bits (171),  Expect = 4e-11, Method: Compositional matrix adjust.
 Identities = 33/46 (71%), Positives = 37/46 (80%), Gaps = 0/46 (0%)

Query  44   AGVETGSADMRFERKAHALKAMKQYYGTPLAGRPVNIQLVTSQIDT  89
            +G   G+AD+ FERKA ALKAMKQY G PL GRP+NIQLVTSQID
Sbjct  142  SGRSLGTADVHFERKADALKAMKQYNGVPLDGRPMNIQLVTSQIDA  187
```

Figure 3: The best alignment of the predicted peptide against the nr database using blastp.

By extracting the DNA sequence of the first predicted gene exon-by-exon and running blastx against the nr database, I found that the first two exons do not match to any known proteins. The last two exons again aligned to THO Complex 4 with only 70% similarity. This evidence suggests that the first two exons predicted by Genscan are incorrect, and the last two exons actually are part of a pseudogene descended from the THO Complex 4 protein.

The last two exons lie in the region between 24kb and 25kb, which contains numerous matches to human ESTs (found using blastn of the masked project sequence against the human_est database). To ensure that these EST matches were to the THO Complex 4 protein, and not to any other functional genomic elements, I extracted this region and ran blastx of the unmasked sequence against the nr database (Figure 4). The results reveal that the region only aligns to the THO Complex 4 protein. In addition, this BLAST search reveals that the raw DNA sequence between 24kb and 25kb contains a run of N's. Running Genscan in this region following sequencing probably would have changed the peptide sequence and improved the alignment. However, even with the improved sequence data only half of the protein would have aligned, based on the length of the protein and the length of the alignment.

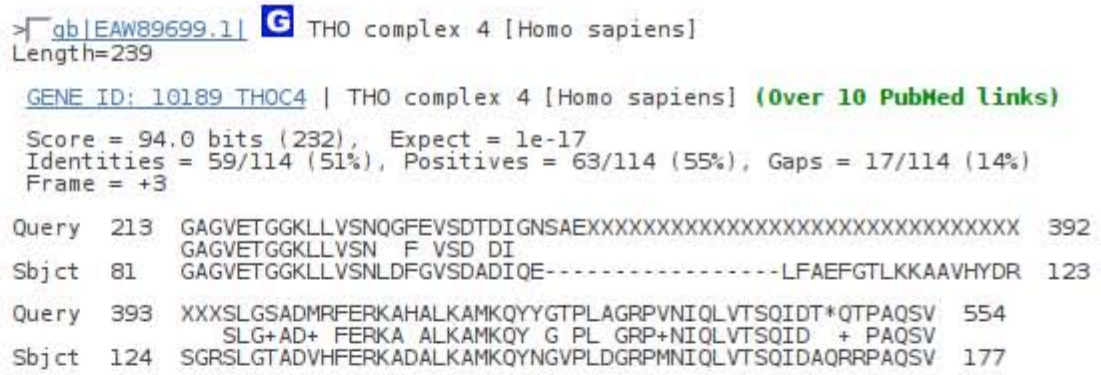

Figure 4: blastx of 24kb-25kb against the nr database.

Then, I ran tblastn, comparing the human protein sequence to the unmasked chimp-chunk sequence. This allowed me to see how well a functional ortholog maps to my project sequence (Figure 5). The BLAST results provide evidence that this feature is a pseudogene. The alignment does not span the entire protein, contains a stop codon in the Open Reading Frame (ORF), numerous mismatched residues, and a frame shift.

```
 Score =  173 bits (439),  Expect(2) = 1e-42
 Identities = 98/157 (62%), Positives = 102/157 (64%), Gaps = 17/157 (10%)
 Frame = +2

Query  99     GAGVETGGKLLVSNLDFGVSDADIQE-----------------LFAEFGTLKKAAVHYDR  141
              GAGVETGGKLLVSN  F VSD DI
Sbjct  24212  GAGVETGGKLLVSNQGFEVSDTDIGNSAEXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX  24391

Query  142    SGRSLGTADVHFERKADALKAMKQYNGVPLDGRPMNIQLVTSQIDAQRRPAQSVNRGGMT  201
                 SLG+AD+ FERKA ALKAMKQY G PL GRP+NIQLVTSQID   + PAQSVNRGGMT
Sbjct  24392  XXXSLGSADMRFERKAHALKAMKQYYGTPLAGRPVNIQLVTSQIDT*QTPAQSVNRGGMT  24571

Query  202    RNRGAGGFGGGGGTRRGTRGGARGRGRGAGRNSKQQL  238
              RNRG GGFGG GGTR GTRGG RGRGRGAGRNSKQQL
Sbjct  24572  RNRGPGGFGGNGGTRGGTRGGDRGRGRGAGRNSKQQL  24682
```

```
 Score = 26.6 bits (57),  Expect(2) = 1e-42
 Identities = 14/19 (73%), Positives = 15/19 (78%), Gaps = 4/19 (21%)
 Frame = +3

Query  239    SAEELDAQLDAYNARMDTS  257
              SAEE+DA    YNARMDTS
Sbjct  24684  SAEEMDA----YNARMDTS  24728
```

Figure 5: tblastn of THOC4 against the unmasked chimp-chunk sequence. Highlighted regions show stop codon in the ORF and a frame shift.

Returning to the initial two predicted exons, blastx of the nucleotide sequence for each exon against the nr database found no significant sequence similarity. Based on the human EST data (Figure 6), the first exon shows some transcription data while second exon does not have any matches. In addition, a blastn search of the masked project sequence to the nt (nucleotide) database (Figure 7) only found weak matches for the first exon (and no matches for the second predicted exon).

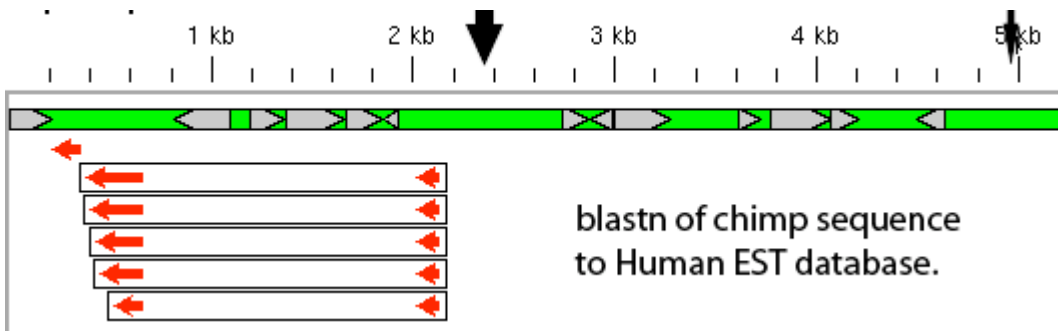

Figure 6: Herne shows the blastn matches to the human EST database do not match to the exons predicted by Genscan. Blastn alignments marked by red arrows; Genscan predicted exons marked by black arrows (arrow width represents approximate exon width).

Because the second exon did not show any matches based on the blastx search of the project sequence to the nr database, did not show any strong matches using blastn against the nt database, and there is no EST data corresponding to the region, the second exon is probably a false prediction by Genscan, and not a pseudogene.

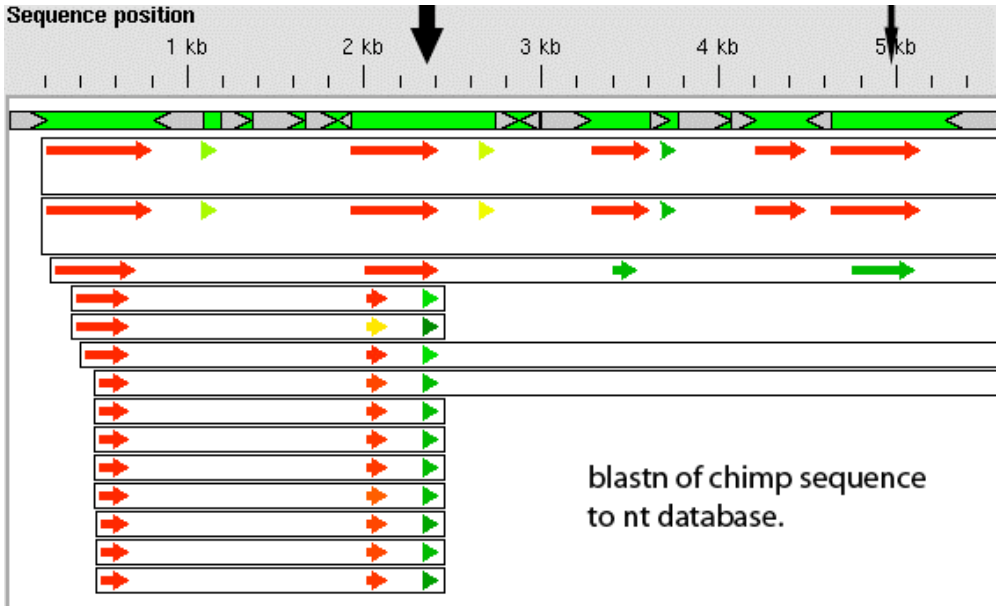

Figure 7: Blastn of the project sequence to the nt database shows the first predicted exon contains some conservation with another genomic element, but the second exon does not. Black arrows indicate locations of the exons predicted by Genscan, with arrow width representing approximate exon width.

However, the matches from the EST and nt databases reveal a feature not predicted by Genscan. Two regions contain strong matches to the MPIF-2 gene (also known as the CCL24). This protein is involved in the immune system. By running the peptide sequence of the human ortholog protein against the chimp chunk sequence using tblastn, I was able to determine how well this ortholog matches to this region (Figure 8).

```
 Score =  115 bits (287),  Expect = 5e-24
 Identities = 53/54 (98%), Positives = 53/54 (98%), Gaps = 0/54 (0%)
 Frame = -1

Query  65   FTTKKGQQFCGDPKQEWVQRYMKNLDAKQKKASPRARAVAVKGPVQRYPGNQTT  118
            FTTKKGQQFCGDPKQEWVQRYMKNLDAKQKK SPRARAVAVKGPVQRYPGNQTT
Sbjct  673  FTTKKGQQFCGDPKQEWVQRYMKNLDAKQKKTSPRARAVAVKGPVQRYPGNQTT  512

 Score = 83.2 bits (204),  Expect = 2e-14
 Identities = 48/85 (56%), Positives = 57/85 (67%), Gaps = 8/85 (9%)
 Frame = -3

Query  25    GSVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGVIFTTKKGQQFCGDPKQEWVQR  84
             GSVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGV+  +     F G  ++    R
Sbjct  2135  GSVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGVM*VS-----FMGTWQERGGGR  1971

Query  85    YMKNLDAKQKKASPRARAVAVKGPV  109
                 +  KQ + +P   A+ + G V
Sbjct  1970  CNSRVPEKQGQPAP---ALCIPGEV  1905
```

Figure 8: tblastn of the human MPIF-2 protein against the chimp chunk sequence. BLAST overextends the alignment to the right of the highlighted stop codon. The top alignment shows high identity for what may be one exon, and the bottom alignment (to the left of the highlighted region) shows what may be another exon.

The query sequence overlaps in these two alignments. The top alignment is very strong but only for the second half of the MPIF-2 protein. The 98% sequence similarity indicates that this may be a functional ortholog.

BLAST overextends the second alignment, resulting in the gaps, mismatches, and stop codons in the ORF. However, the bottom alignment also shows strong similarity for amino acids 25-64 in the reference protein sequence. Note that while both these matches are very high quality, BLAST did not align the first 24 residues in the reference protein sequence. Running the tblastn search again using only the first 25 residues of the MPIF-2 protein yielded two alignments with E=28 (Figure 9). While these regions have lower sequence similarity than expected, the other two exons show very high similarity, so it is unlikely that this feature is a pseudogene.

```
 Score = 29.3 bits (64),  Expect(2) =    28
 Identities = 10/11 (90%), Positives = 11/11 (100%), Gaps = 0/11 (0%)
 Frame = -3

Query  15    GVCAHHIIPTG  25
             GVCAHH+IPTG
Sbjct  2384  GVCAHHVIPTG  2352
```

```
 Score = 21.9 bits (45),  Expect(2) =    28
 Identities = 9/13 (69%), Positives = 12/13 (92%), Gaps = 0/13 (0%)
 Frame = -1

Query  2     AGLMTIVTSLLFL  14
             AGLMTIV ++LF+
Sbjct  2422  AGLMTIVPAILFV  2384
```

Figure 9: tblastn of 1st 24 AA of MPIF-2 against the unmasked project sequence. All but the first of the 25 residues are matched, although there is a frame shift in the alignment.

Because the first exon does not match well, the gene model was verified using the Gene Model Checker developed by W. Leung. Figure 10 shows that the gene model is correct, with the exception of the start codon. The start codon cannot be determined because of N's in the sequence (Figure 11). This low quality sequence is the reason Genscan did not make a gene prediction for this region.

Because the tblastn results shown in Figure 9 align the second residue at 2422, I predict the start codon is at 2426. This position yields the correct frame for the protein, and adds one amino acid to the protein sequence (so as to agree with the tblastn alignment above).

Checklist Item	Status	Message
Find Start Codon	FAIL	Found X in first codon instead of methionine
Find Stop Codon	PASS	
Splice Acceptor for Exon 1	SKIP	Already checked for start codon
Splice Donor for Exon 1	PASS	
Splice Acceptor for Exon 2	PASS	
Splice Donor for Exon 2	PASS	
Splice Acceptor for Exon 3	PASS	
Splice Donor for Exon 3	SKIP	Already checked for stop codon
Additional Issues	PASS	

Figure 10: Gene Model Checker verifies results. N's mask the start codon (see Fig. 11).

Exon #	Extracted Sequence	Complete Codons in Extracted Region
Exon 1 (73 bp) 2426-2354	.....NNNNN 2426 NNNGGCAGGCCTGATGACCATCGTACCAGCGATTCTGTTCGTGGTGTCTG 2377 2376 TGCCCACCACGTCATCCCTACGG 2327 GTAAG.....	XGRPDDHRTSDSVRGVCAHHVIPT
Exon 2 (118 bp) 2134-2017	.....TCTAG 2134 GCTCTGTGGTCATCCCCTCTCCCTGCTGCATGTTCTTTGTTTCCAAGAGA 2085 2084 ATTCCTGAGAACCGAGTGGTCAGCTACCAGCTGTCCAGCAGGAGCACGTG 2035 2034 CCTCAAGGCAGGAGTGAT 1985 GTAGG.....	SVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGV
Exon 3 (166 bp) 674-509	.....CACAG 674 CTTTACCACCAAGAAGGGCCAGCAGTTCTGTGGCGACCCCAAGCAGGAGT 625 624 GGGTCCAGAGGTACATGAAGAACCTGGACGCCAAGCAGAAGAAGACTTCC 575 574 CCTAGGGCCAGGGCAGTGGCTGTCAAGGGCCCTGTCCAGAGATATCCTGG 525 524 CAACCAAACCACCCTC 475 TAATC.....	FTTKKGQQFCGDPKQEWVQRYMKNLDAKQKKTSPRARAVAVKGPVQRYPG NQTTL

Figure 11: Gene contains canonical splice sites, stop codon at the end, and no stop codons in the ORF. Start codon masked by N's in sequence.

Finally, I ran the final protein from the Gene Model Checker against the nr database using blastp to determine the overall conservation of the protein (Figure 12). The results show that the protein is highly conserved after residue 36. However, the alignment only has 92% identity and is missing 11 residues. This is surprising given that in general 98-99% identity is expected when comparing chimp to human. The first 36 amino acids of the protein are probably not essential for protein function. While the amino acid sequence at the ends may not be conserved, the length of the protein is conserved, with both my model and the human ortholog having 119 amino acids.

```
>ref|NP_002982.2| UG small inducible cytokine A24 precursor [Homo sapiens]
 sp|O00175.2|CCL24_HUMAN G C-C motif chemokine 24 precursor (Small-inducible cytokine A24)
(Myeloid progenitor inhibitory factor 2) (MPIF-2) (CK-beta-6)
(Eosinophil chemotactic protein 2) (Eotaxin-2)
 gb|AAD15410.1| G unknown [Homo sapiens]
 gb|AAH69072.1| G Small inducible cytokine A24, precursor [Homo sapiens]
 gb|AAH69391.1| G Chemokine (C-C motif) ligand 24 [Homo sapiens]
 gb|EAW71772.1| G chemokine (C-C motif) ligand 24 [Homo sapiens]
Length=119

 GENE_ID: 6369 CCL24 | chemokine (C-C motif) ligand 24 [Homo sapiens]
(Over 10 PubMed links)

 Score =  204 bits (518),  Expect = 2e-51, Method: Compositional matrix adjust.
 Identities = 102/110 (92%), Positives = 103/110 (93%), Gaps = 2/110 (1%)

Query  9    TSDSVRGVCAHHVIPT-SVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGV-FTTK  66
            TS    GVCAHH+IPT SVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGV FTTK
Sbjct  9    TSLLFLGVCAHHIIPTGSVVIPSPCCMFFVSKRIPENRVVSYQLSSRSTCLKAGVIFTTK  68

Query  67   KGQQFCGDPKQEWVQRYMKNLDAKQKKTSPRARAVAVKGPVQRYPGNQTT  116
            KGQQFCGDPKQEWVQRYMKNLDAKQKK SPRARAVAVKGPVQRYPGNQTT
Sbjct  69   KGQQFCGDPKQEWVQRYMKNLDAKQKKASPRARAVAVKGPVQRYPGNQTT  118
```

Figure 12: blastp of the final gene model against the nr database. Percent identity is 92% but the alignment does not even match the entire peptide sequence.

Predicted Gene 2

J. Kuhn ran the second predicted peptide against the nr database using blastp, and found 99% identity with 315 out of 357 amino acids in the rhomboid domain-containing 2 protein. The last ~40 amino acids were not aligned. After aligning the human ortholog's peptide sequence against the chimp chunk sequence using the tblastn algorithm, the entire protein matched to the DNA with 99% identity. Genscan mispredicted an exon at 75kb, and did not identify the terminal exon in the protein, which was actually downstream at 72kb.

In addition to matching the entire rhomboid domain-containing 2 protein, blastn of the chimp chunk sequence against the nt database revealed a match with 99% similarity around 73kb. The BLAST results indicate this region belongs to the rhomboid domain-containing 2 gene. Because this is downstream of the protein-coding region (and the entire protein sequence was aligned), we suspected this was an untranslated region (UTR). After extracting the matching region and running the sequence using blastx against the nr database, we found only poor protein matches to the region (E>.1). EST data aligns with 99% similarity in this region too. This data supports the hypothesis that this region is a 3' UTR. However, further analysis is needed for verification.

Repeat Masker Results

Repeat Masker is a program designed to identify the repetitious elements contained in a DNA sequence. We ran Repeat Masker with the no-low option to prevent accidental masking of repetitious amino acid sequences. Repeat Masker found interspersed repeats in 52% of the sequence. In addition, it should be noted that nearly 12kb (out of 80kb) consists of N's or X's, which may have affected the Genscan predictions. Only 2 repeats are over 500bp in length, both being LTR elements.

```
total length:      81316 bp  (68670 bp excl N/X-runs)
GC level:         49.99 %
bases masked:      42524 bp ( 52.29 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              138        32706 bp   40.22 %
      ALUs          121        30630 bp   37.67 %
      MIRs           17         2076 bp    2.55 %

LINEs:              17         4064 bp    5.00 %
      LINE1          8         2574 bp    3.17 %
      LINE2          7         1246 bp    1.53 %
      L3/CR1         2          244 bp    0.30 %

LTR elements:       14         4486 bp    5.52 %
      MaLRs          3         1167 bp    1.44 %
      ERVL           4         1191 bp    1.46 %
      ERV_classI     6         1987 bp    2.44 %
      ERV_classII    1          141 bp    0.17 %

DNA elements:        7         1307 bp    1.61 %
      MER1_type      3          423 bp    0.52 %
      MER2_type      1           53 bp    0.07 %

Unclassified:        0            0 bp    0.00 %

Total interspersed repeats:    42563 bp   52.34 %
```

Figure 13: Repeat Masker output summary.

Synteny

Using the UCSC genome browser, I evaluated the relative orientations of the two functional genes we annotated for this project, relative to the human genome (hg17). Both genes mapped to the long arm of chromosome 7 (Figure 14). In addition, the genes have the same relative orientation in our project as in the human assembly.

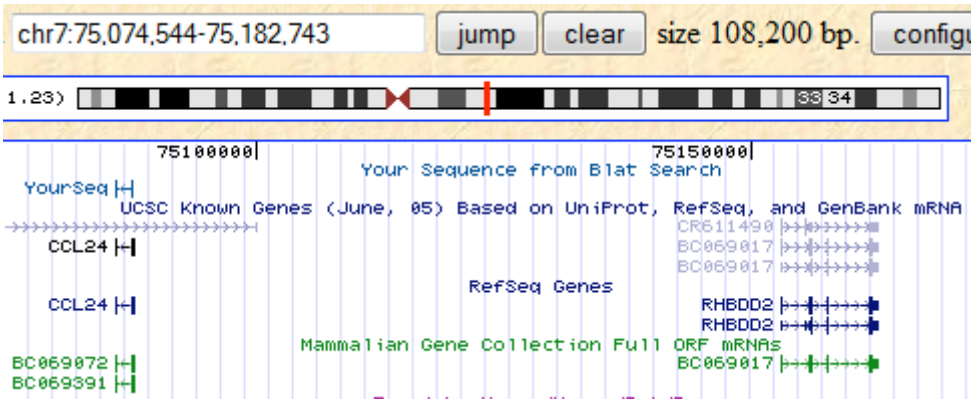

Figure 14: Human assembly showing both functional genes in our project (CCL24=MPIF-2,

CR611490=rhomboid domain-containing 2). Comparison with Figure 15 reveals synteny of these genes.

Summary

In addition to being highly repetitious, this chunk contains two genes. The first gene is an ortholog of the MPIF-2 gene with 3 exons. The protein has lower conservation (<92% identity) than expected, however the length of the protein has been conserved. The second gene is an ortholog of the rhomboid domain-containing 2 gene. This gene contains four exons and a highly-conserved 3' UTR. This project also contains one pseudogene. Genscan correctly predicted three out of four exons for the rhomboid domain-containing 2 ortholog, but incorrectly predicted the other two peptides, and failed to predict the MPIF-2 ortholog. See Figure 15 for more information. In addition, the two functional genes in the project are syntenic with the orthologous human region.

Further Work

Further work may include improvement of sequence quality. This would improve confidence in our predictions of both the start codon in the MPIF-2 ortholog and the rhomboid domain-containing 2 ortholog. In addition, experimentation could identify a 5' UTR for the rhomboid domain-containing 2 gene, and verify the 3' UTR. Furthermore, because the first 25 amino acids in the MPIF-2 ortholog are not as highly conserved as expected (<92% identity), analysis of this region will probably reveal that these amino acids are not essential for protein function.

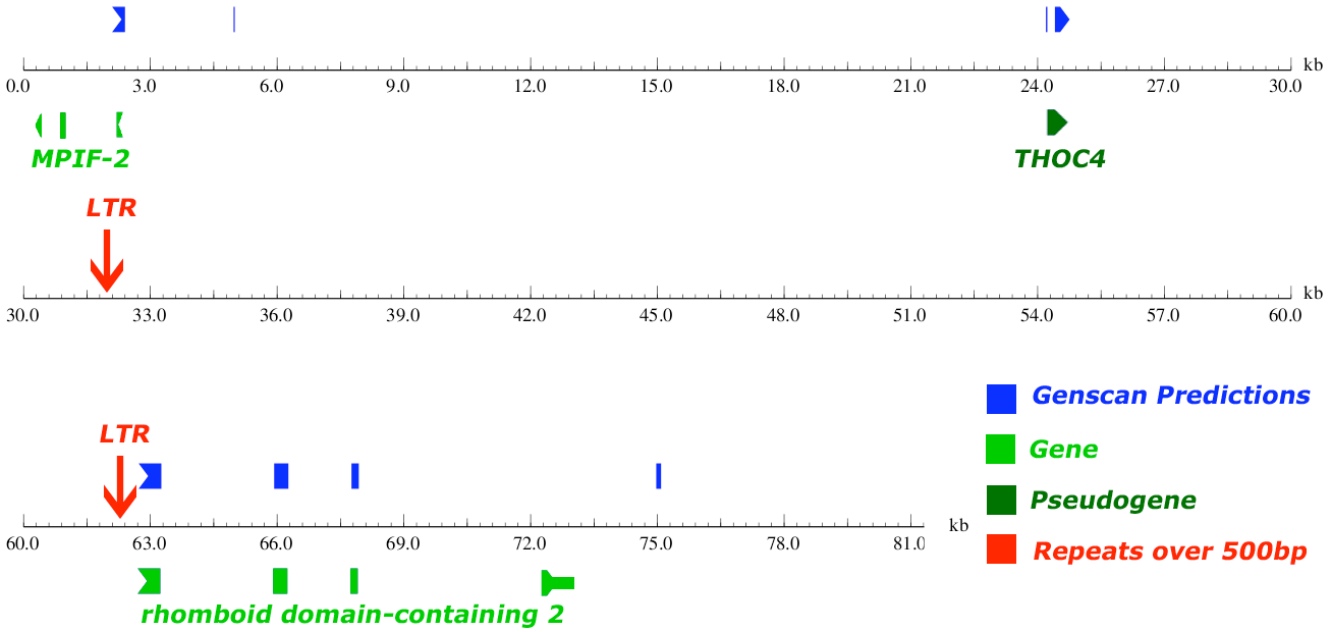

Figure 15: Annotation summary.

Rhomboid domain-containing 2 – functional gene; 4 exons; strong evidence for 3' UTR
MPIF-2 – functional gene with 3 exons identified; spans from about .5kb-2.5kb

THOC4 – pseudogene with one exon identified; spans from about 24kb-25kb

52.3% of sequence is interspersed repeats. Only two repeats spanned more than 500bp; both were LTR elements.

- 5% LINEs
- 37.7% Alus