

# Annotation of Chimp Chunk 2-10

---

**Andrew Stein**  
**April 19, 2009**

## Abstract

In this paper, I present information regarding the annotation for chimp chunk 2-10 as carried out by Jerome Molleston and me. We began our work by reviewing the GenScan output information that shows two predicted genes. The first predicted gene is a pseudogene while the second predicted gene is the chimpanzee ortholog to the human protein thymopoietin (TMP) isoform alpha. GenScan originally predicted that this gene consisted of seven exons, but my analysis proves that it is actually comprised of four exons. Furthermore, we also identified two other isoforms of the TMP protein in addition to isoform alpha. These isoforms were discovered by comparing our chimp chunk to human EST information along with BLAST searches to human protein databases. Overall, GenScan originally predicted two genes within chimp chunk 2-10, but our analysis revealed the presence of one pseudogene and isoforms alpha, beta and gamma of the human TMP protein (Table 1).

GenScan prediction number	Location of feature identified (bp)	Predicted number of exons	Actual number of exons	Feature description
1	3347-5587	2	1	Pseudogene: similar to solute carrier family 9 member 7
2	70821-103738	7	4, 6, 9	Gene: orthologous to thymopoietin isoforms alpha, beta, gamma. Thymopoietin is involved in the structural organization of the nucleus.

Table 1: Summary of features in chimp chunk 2-10

## Introduction

Annotation is the process of surveying sequence information in order to understand the biologically important aspects that are represented within the apparent nonsense of DNA. This process is used to identify genes, pseudogenes, noncoding regions, repetitive regions, and all other aspects of a genomic sequence. In this project, I utilized annotated sequences from the human genome to deduce information about our chimpanzee fragment since humans and chimpanzees are on average 98-99% identical with regard to DNA content. I used information from ESTs, BLAST, BLAT, and GenScan to produce the necessary evidence to make proper

inferences about the given chimpanzee sequence. In this paper, I discuss the evidence that Jerome and I gathered from these sources to reveal the presence of the orthologs of the three isoforms of the human TMP protein along with a pseudogene within chimp chunk 2-10.

## Project Work Flow

### *Initial GenScan Results*

To begin our work on annotating chimp chunk 2-10, Jerome and I surveyed the initial GenScan output. We saw that GenScan predicted the presence of two genes within our sequence fragment (Figure 1). The first predicted gene has two exons and is located between base pairs 3000 and 6000, while the second predicted gene spans from base pairs 27000 to 90000 and consists of seven exons (Table 2). I analyzed predicted gene number two, and Jerome examined the first predicted gene.

Predicted Gene	Number of Exons	Base Pair Location
1	2	3000-6000
2	7	27000-90000

Table 2: GenScan predictions

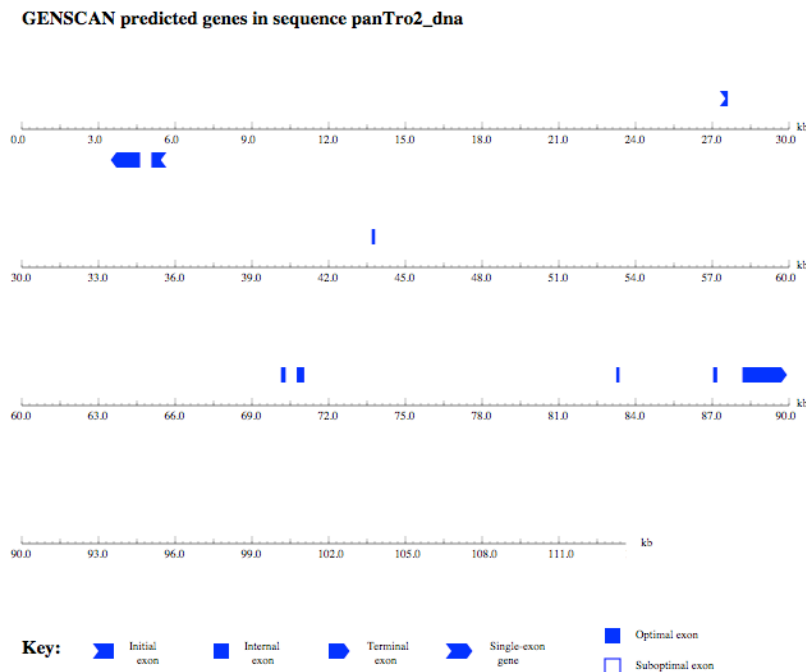


Figure 1: Initial GenScan output revealing two predicted genes

### *Annotation of Predicted Gene Number Two*

My first step was to extract the protein coding sequence of predicted gene number two from the GenScan results and run a BLASTP search with it against the nonredundant (nr) protein database. The best experimentally supported match was to the thymopoietin (TMP) isoform

alpha protein from *Homo sapiens* (Figure 2). This protein is located on chromosome 12 in humans.




<a href="#">ref XP_509288.2</a>	PREDICTED: thymopoietin isoform 3 [Pan trogl...	<a href="#">1311</a>	0.0	
<a href="#">ref NP_003267.1</a>	thymopoietin isoform alpha [Homo sapiens] >s...	<a href="#">1303</a>	0.0	
<a href="#">gb AAB60433.1</a>	thymopoietin alpha [Homo sapiens] >prf  211729...	<a href="#">1302</a>	0.0	

Figure 2: Results from BLASTP between predicted protein number two and the nr protein database

Predicted protein number two aligns to the TMP isoform alpha protein with 96% identity (Figure 3). However, amino acids 1-157 from the protein sequence of predicted gene two does not align to any portion of the TMP protein. Therefore, I extracted amino acids 1-157 from the predicted protein sequence and ran a BLASTP search with this sequence against the nr protein database. I did not find any biologically interesting results from this BLAST search, immediately suggesting that GenScan had miscalled the exons of this predicted gene. Also, amino acid 158 is methionine, which indicates that it is the first amino acid of this gene. At this point, I believed that the first three exons predicted by GenScan (correlating to amino acids 1-157) were incorrectly inferred and are not actually part of this gene.

```

GENE ID: 7112 TMPO | thymopoietin [Homo sapiens] (Over 10 PubMed links)
Score = 130.1 bits (3373), Expect = 0.0, Method: Compositional matrix adjust.
Identities = 672/694 (96%), Positives = 674/694 (97%), Gaps = 18/694 (2%)
Query 158 MPEFLEDPSVLTKDKLSELVANNVTLFAGEQQRKDVYVQLYLQHLTARNRPPLPAGTNSK 217
Sbjct 1 MPEFLEDPSVLTKDKLSELVANNVTLFAGEQQRKDVYVQLYLQHLTARNRPPLPAGTNSK 60
Query 218 GPPDFSSDEEREPTP-----KATEKTDKPRQEDKDDLDVTELTNEDL 259
Sbjct 61 GPPDFSSDEEREPTP KATEKTDKPRQEDKDDLDVTELTNEDL 120
Query 260 LDQLVKYGVNPGPIVGTTRKLYEKKLLKLRQQTESRSSTPLPTISSSAENTRQNGSND 319
Sbjct 121 LDQLVKYGVNPGPIVGTTRKLYEKKLLKLRQQTESRSSTPLPTISSSAENTRQNGSND 180
Query 320 DRYSDNEBGGKKEHKVKVSTRDVPFSELGTTPSGGGFFQGISFPEISTRPPLGSTELQA 379
Sbjct 181 DRYSDNEBGGKKEHKVKVSTRDVPFSELGTTPSGGGFFQGISFPEISTRPPLGSTELQA 240
Query 380 AKKVHTSKGDLPREPLVATNLPGRGQLQKLASERNLFIACKSSHDRCLKSSSSSSQPEH 439
Sbjct 241 AKKVHTSKGDLPREPLVATNLPGRGQLQKLASERNLFIACKSSHDRCLKSSSSSSQPEH 300
Query 440 SAMLVSTAASPSLIKETTGGYKDIVENICGREKSGIQPLCPERSHISDQSPLSKRRKAL 499
Sbjct 301 SAMLVSTAASPSLIKETTGGYKDIVENICGREKSGIQPLCPERSHISDQSPLSKRRKAL 360
Query 500 EESSESQLISPPLAQAIRDYVNSLLVQGGVGSPLGTSNSMPPLDVENIQKRIDQSKFQET 559
Sbjct 361 EESSESQLISPPLAQAIRDYVNSLLVQGGVGSPLGTSNSMPPLDVENIQKRIDQSKFQET 420
Query 560 EPLSPPKRVFRLSEKSVREERDGSFVAFQNIPEGSELMSSFAKTVVSSSLTTLGLEVAKQS 619
Sbjct 421 EPLSPPKRVFRLSEKSVREERDGSFVAFQNIPEGSELMSSFAKTVVSSSLTTLGLEVAKQS 480
Query 620 QHDKIHASELSFPFRESILKVIIEEWQVDRQLPSLACKYFPVSSREATQILSVPKVDDEI 679
Sbjct 481 QHDKIHASELSFPFRESILKVIIEEWQVDRQLPSLACKYFPVSSREATQILSVPKVDDEI 540
Query 680 LGFISEATPLGGIQAASTESCNOQLDLALCRAYEAAASALQIATHTAFVAKAMQADISQA 739
Sbjct 541 LGFISEATPLGGIQAASTESCNOQLDLALCRAYEAAASALQIATHTAFVAKAMQADISQA 600
Query 740 AQILSSDPSRTHQALGILSKTYDAASYICEAAFDEVKMAAHTMGNSTVGRRYLMLKDCKI 799
Sbjct 601 AQILSSDPSRTHQALGILSKTYDAASYICEAAFDEVKMAAHTMGNSTVGRRYLMLKDCKI 660
Query 800 NLASKNKLASTPPFGGTLFGGEVCKVIKGRGNKH 833
Sbjct 661 NLASKNKLASTPPFGGTLFGGEVCKVIKGRGNKH 694

```

Figure 3: Alignment between predicted protein number two and TMP isoform alpha

My initial results indicate that predicted gene number two consists of four exons and is the chimpanzee ortholog of the human TMP isoform alpha protein, but I decided to gather additional information from the UCSC browser. I ran a BLAT search with the predicted protein sequence

against the human database and found a region on chromosome 12 where the feature aligns with 99.5% identity (Figure 4). This provided important information about this predicted gene since the TMP isoform alpha protein is also located on chromosome 12. Furthermore, other gene predictors such as Ensembl and N-SCAN reveal the same gene prediction with only four exons (Figure 5).

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
<a href="#">browser details</a>	YourSeq	2465	1	833	833	99.5%	12	++	97395927	97452248	56322
<a href="#">browser details</a>	YourSeq	227	233	327	833	88.7%	16	++	73259013	73259294	282

Figure 4: BLAT result showing a match to the human genome at chromosome 12 with 99.5% identity

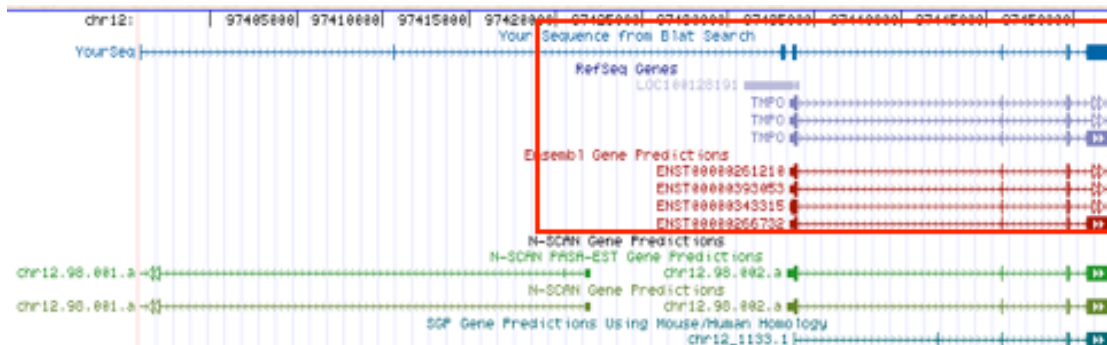


Figure 5: BLAT output showing the other gene predictor models

To confirm the presence of the chimpanzee ortholog to the TMP isoform alpha gene in my chimpanzee sequence, I decided to run a BLASTX search where I compared the entire masked chimp chunk DNA sequence to the human protein sequence of TMP isoform alpha. The BLAST output reveals four separate matches between the chimpanzee sequence and the protein sequence (Figure 6). Although some of the amino acids are found in two of the alignments in Figure 6, they represent a misalignment by BLAST and truly belong in only one of the alignments. For example, the first eleven amino acids that are present in both alignments three and four only belong in alignment four (Figure 6). However, the last sixteen amino acids that are in both alignments three and four truly belong at the start of alignment three. All in all, these alignments indicate that the ortholog to the TMP isoform alpha protein is indeed present within chimp chunk 2-10 and only consists of four exons. Specifically, these four exons are located between base pair positions 70820-89695 within the chimp chunk. Thus, this BLASTX result shows that there are indeed only four exons pertaining to predicted gene number two, instead of seven as GenScan indicates.

```

Query 89078 ELSFPFHESILKVIIEEWQVDRQLPSLACKYPVSSREATQILSVPKVDEILGFISEAT 89257
Sbjct 489 ELSFPFHESILKVIIEEWQVDRQLPSLACKYPVSSREATQILSVPKVDEILGFISEAT 548

Query 89258 PLGGIQAASTESCNOQLDLALCRAYEAAASALQIATHTAFVAKAMQADISQAAQILSSDP 89437
Sbjct 549 PLGGIQAASTESCNOQLDLALCRAYEAAASALQIATHTAFVAKAMQADISQAAQILSSDP 608

Query 89438 SRTHQALGILSKTYDAASYICEAAFDEVKMAAHTMGNSTVGRRYLWLKDKCKINLASKNKL 89617
Sbjct 609 SRTHQALGILSKTYDAASYICEAAFDEVKMAAHTMGNSTVGRRYLWLKDKCKINLASKNKL 668

Query 89618 ASTPFKGGTLPFGEVCKVIKRRGNKH 89695
Sbjct 669 ASTPFKGGTLPFGEVCKVIKRRGNKH 694

Score = 174 bits (440), Expect = 1e-45
Identities = 87/97 (89%), Positives = 88/97 (90%), Gaps = 0/97 (0%)
Frame = +2

Query 70820 MPEFLEDPSVLTDKDKLSELVANNVTLPAGEQRKDVVYQLYLQHLTARNRPFPLPAGTNSK 70999
Sbjct 1 MPEFLEDPSVLTDKDKLSELVANNVTLPAGEQRKDVVYQLYLQHLTARNRPFPLPAGTNSK 60

Query 71000 GPPDFSSDEEREPTVLGSGXXXXXXXXSRAAVGRVTR 71110
Sbjct 61 GPPDFSSDEEREPTVLGSGAAAAGRSRAAVGRKATK 97

Score = 113 bits (282), Expect = 2e-27
Identities = 58/65 (89%), Positives = 60/65 (92%), Gaps = 4/65 (6%)
Frame = +3

Query 87015 KFCLNPG----TTRKLYEKKLLKLRQVTESRSSTPLPTISSAENTRQNGSNDSDRYSD 87182
Sbjct 126 K+ +NPG TTRKLYEKKLLKLRQVTESRSSTPLPTISSAENTRQNGSNDSDRYSD 185
KYGVNPGPIVGTTRKLYEKKLLKLRQVTESRSSTPLPTISSAENTRQNGSNDSDRYSD

Query 87183 NEEGK 87197
Sbjct 186 NEEGK 190

Score = 92.8 bits (229), Expect = 3e-21
Identities = 49/69 (71%), Positives = 53/69 (76%), Gaps = 8/69 (11%)
Frame = +3

Query 83241 QKATKKTDKPRQEDKDDLVDVTELTNEDLLDQLKYGVNPGPIVG-----KLIFQIQ 83396
Sbjct 93 +KATKKTDKPRQEDKDDLVDVTELTNEDLLDQLKYGVNPGPIVG KL+K + Q 152
RKATKKTDKPRQEDKDDLVDVTELTNEDLLDQLKYGVNPGPIVGTTRKLYEKKLLKLRQV

Query 83397 YLFSEAGPP 83423
Sbjct 153 S + P GTESRSSTP 161

```

Figure 6: BLASTX output between masked chimp chunk sequence and TMP isoform alpha protein sequence

From the BLAST and BLAT outputs I examined, I am confident in asserting that predicted gene number two represents the chimpanzee's ortholog of the human protein TMP isoform alpha. Furthermore, GenScan has incorrectly predicted the gene structure of this protein, and it truly consists of four exons instead of seven. The entire protein is 694 amino acids in length and the DNA sequence aligns with greater than 99% identity to a portion of the human genome on chromosome twelve. An identity of such a high percentage is expected and necessary to make assertions regarding orthologous genes between humans and chimps given their close evolutionary relationship.

### *Expressed Sequence Tag (EST) Analysis*

After defining predicted gene number two, I analyzed the EST information to see if there were any other genes present within this chimp chunk that GenScan failed to predict. I ran a BLASTN search using the entire chimp chunk sequence against the human EST database. Then, I examined the Herne output to see if any region of the chimp chunk had multiple EST matches. I noted that the region around 69000-71000 base pairs has many EST matches (Figure 7). This indicates that there could be another feature present within this region.





Figure 7: EST matches present around base pairs 69000-71000 in Herne output

Based upon the multiple human EST matches to the region from base pairs 69000-71000, I decided to investigate this region further. I ran a BLASTX search on this region using the masked chimp chunk sequence against the nr protein database. However, the BLAST results do not reveal anything biologically interesting since the e values are quite high (greater than one). Although this indicates that there are not any proteins present within this region, the vast number of EST hits confused me. Therefore, I wanted to investigate the sequence in this area further. I ran a BLASTN search using the unmasked chimp sequence against the nucleotide (nt) database (Figure 8).

Sequences producing significant alignments:  
(Click headers to sort columns)

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
<a href="#">AC013418.32</a>	Homo sapiens 12 BAC RP11-181C3 (Roswell Park Cancer Institute Hur	9965	2.495e+04	92%	0.0	98%	
<a href="#">XM_509288.2</a>	PREDICTED: Pan troglodytes thymopoietin, transcript variant 3 (TMPO	5866	5866	52%	0.0	100%	<a href="#">G</a>
<a href="#">XM_001148930.1</a>	PREDICTED: Pan troglodytes thymopoietin, transcript variant 1 (TMPO	5866	6099	57%	0.0	100%	<a href="#">G</a>
<a href="#">XM_001148998.1</a>	PREDICTED: Pan troglodytes thymopoietin, transcript variant 2 (TMPO	5866	6099	57%	0.0	100%	<a href="#">G</a>
<a href="#">NR_027157.1</a>	Homo sapiens hypothetical protein LOC100128191 (LOC100128191),	5276	5715	52%	0.0	100%	<a href="#">G</a>
<a href="#">U18266.1</a>	Human thymopoietin (TMPO) gene, exon 1	4523	4523	41%	0.0	99%	<a href="#">E</a>
<a href="#">BC037346.2</a>	Homo sapiens cDNA clone IMAGE:5266100	4473	4473	41%	0.0	98%	<a href="#">U</a> <a href="#">G</a>
<a href="#">XM_001082868.1</a>	PREDICTED: Macaca mulatta hypothetical protein LOC693517 (LOC69	3585	3585	42%	0.0	92%	<a href="#">G</a>

Figure 8: Results from BLASTN between unmasked chimp chunk sequence and the nt database

From the output information above (Figure 8), there are many matches to hypothetically annotated refseqs from human and chimp that indicate the presence of the TMP protein in this region. The chimp hypothetical protein (variant 3) spans from base pairs 67923-71098 of the chimp chunk sequence and aligns with 100% identity. This seems to reveal that the TMP protein sequence starts earlier than I anticipated (previously thought to start at base pair 70820). I went to the Ensembl website to investigate the TMP protein in greater detail, and I discovered that there are actually four annotated versions of this protein in humans. Then, I extracted the first exons from each of these variants to see if any of them would align to the chimp chunk sequence

around base pair 68000. However, the farthest upstream that any of these variants stretch to is base pair 70579. Overall, this information does not help to explain the presence of such a high density of ESTs in this region. Based upon the evidence I gathered about this region, I developed a few possible explanations that could account for the high density of human EST hits. It is possible that the chimp version of the TMP sequence has a longer 5'-UTR (and therefore first exon) as compared to the human gene. Also, it is within the realm of possibility that the EST information is not very reliable, or the refseq annotation for this protein is not entirely complete.

Aside from EST information present within the aforementioned region, there are also EST matches between base pairs 36000 and 57000 (Figure 9).

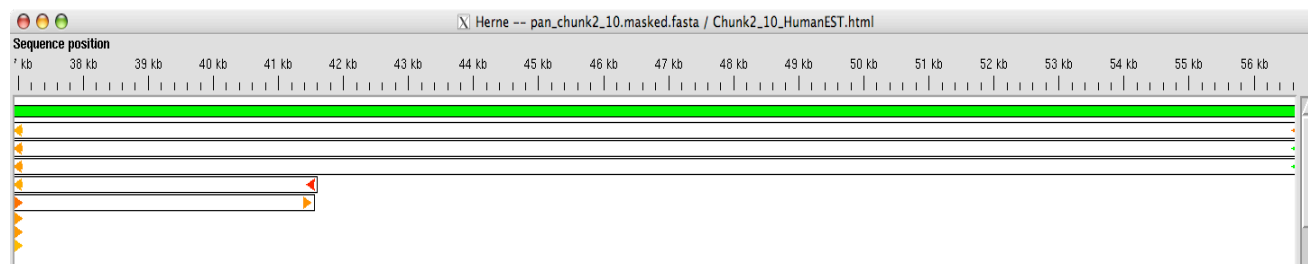


Figure 9: EST information present between base pairs 36000 and 57000

I extracted the region between base pairs 36000 and 57000 and attempted to run a BLASTX search with this region against the nr protein database. However, the NCBI website indicated that the request required too much memory and thus the task could not be completed. Similarly, I attempted to use a BLASTN search to the nt database, but this failed as well. In the end, I used a BLASTX search against the refseq protein database. This BLAST search does not reveal any significant results. Based upon this information, I concluded that there are not any genes present within the chimp chunk from base pairs 36000-57000.

### ***RepeatMasker Analysis***

Another important feature of annotation is to determine the quantity of repetitive elements present within a sequence. Repetitious elements comprise a significant percentage of many organisms' DNA (especially chimpanzees), and they provide insightful information into the evolutionary history of these organisms. For this reason, I ran RepeatMasker on the chimp chunk sequence to deduce the quantity and types of repetitive elements present (Figure 10). Chimp chunk 2-10 is comprised of 48.55% repetitious elements, the majority of which are Alu elements (29.88%). Although there are many repetitive elements present within this sequence, I only annotated the elements that are over 500 base pairs (Table 3).



```

=====
file name: pan_chunk2_10.fasta
sequences: 1
total length: 113620 bp (104398 bp excl N/X-runs)
GC level: 42.59 %
bases masked: 55163 bp ( 48.55 %)
=====
              number of      length      percentage
              elements*    occupied    of sequence
-----
SINEs:
  ALUs      122      33951 bp    29.88 %
  MIRs       13       1823 bp     1.60 %

LINEs:
  LINE1     17       6070 bp     5.34 %
  LINE2     10       2927 bp     2.58 %
  L3/CR1     1         423 bp     0.37 %

LTR elements:
  ERVL        2         293 bp     0.26 %
  ERVL-MaLRs 13       4736 bp     4.17 %
  ERV_classI  8       2639 bp     2.32 %
  ERV_classII 0          0 bp     0.00 %

DNA elements:
  hAT-Charlie 9       1231 bp     1.08 %
  TcMar-Tigger 2         377 bp     0.33 %

Unclassified: 1         150 bp     0.13 %

Total interspersed repeats: 55019 bp    48.42 %

Small RNA: 2         144 bp     0.13 %

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

```

Figure 10: Complete list of RepeatMasker results

Repetitive Element Type	Start Position	End Position	Total Length
LINE/L2	67812	68878	1066
LINE/L1	5900	6564	664
LINE/L1	110955	111597	642
LTR/ERV1	63849	64452	603
LTR/ERV1-MaLR	17826	18342	516
LINE/L1	112645	113155	510

Table 3: Repetitive elements included in the final annotation map

### *Jerome's Work and Analysis of Predicted Gene Number One*

Jerome analyzed predicted gene number one from the GenScan output. By running a BLASTX search between the DNA sequence from this area and the nr protein database, he noticed that the best match was to the human solute carrier family 9 member 7 protein. Contrary to GenScan's prediction of two exons, Jerome's BLAST output aligns the entire protein sequence as one contiguous block (most likely one exon). However, based upon information from Ensembl, the actual solute carrier family 9 member 7 protein in humans consists of seventeen exons. Since the true gene has seventeen exons and the gene within our chimp chunk has only one exon, this indicates that predicted gene number one is a pseudogene. Furthermore, from the BLAST alignment, he noticed many premature stop codons, which further demonstrates that predicted gene number one is a pseudogene. Jerome also aligned the DNA sequence against the human database using a BLAT alignment. This alignment is homologous to chromosome 12 whereas

the functional solute carrier gene is present on chromosome X. The BLAT output provides even more information to support the presence of a pseudogene in this region of chimp chunk 2-10.

Next, Jerome analyzed the human EST matches at the 3' end of our chimp chunk sequence to look for any features present in this region. He noticed many ESTs in the area between base pairs 90000 to 107000. Therefore, he ran a BLASTX search on this region versus the nr protein database, and the best match is to the human thymopoietin isoform gamma protein. Based upon this match, he went to Ensembl and found that there are three isoforms of the TMP protein. I had previously found the TMP isoform alpha within our sequence. Jerome predicted that the BLASTX results would also reveal the presence of TMP isoform beta, and he was able to find the appropriate alignments to this isoform from the output information. In short, the orthologs of isoforms beta and gamma are also present in chimp chunk 2-10. Isoform beta has a total of nine exons whereas isoform gamma has a total of six (alpha has four exons).

Overall, Jerome was successful in showing that predicted gene number one from the GenScan output is in fact a pseudogene. Also, he showed the presence of the orthologous genes to the human beta and gamma isoforms of the TMP protein.

## **Conclusion**

Jerome and I discovered that chimp chunk 2-10 contains three isoforms of the thymopoietin protein along with one pseudogene (Figure 11). The pseudogene is comprised of only one exon and is located between base pairs 3347 and 5587. This appears to be a pseudogene of the solute carrier family 9 member 7 protein that actually consists of seventeen exons in humans. The TMP isoform alpha ortholog is located between base pairs 70820 and 89695 and contains a total of four exons. The other two isoforms of the TMP protein, beta and gamma, share the first three exons with isoform alpha; however, the beta and gamma isoforms extend from base pairs 70820 to 103738 and consist of nine and six exons respectively. Aside from annotating the various pseudogenes and genes present within this chimp sequence, we also determined that it is 48.55% repetitious. Among this large percentage of repetitious elements, only six elements cover more than 500 base pairs (Figure 11). Thus, the evidence presented in this paper supports the presence of one pseudogene and the chimp orthologs of the three isoforms of the human TMP protein within chimp chunk 2-10.

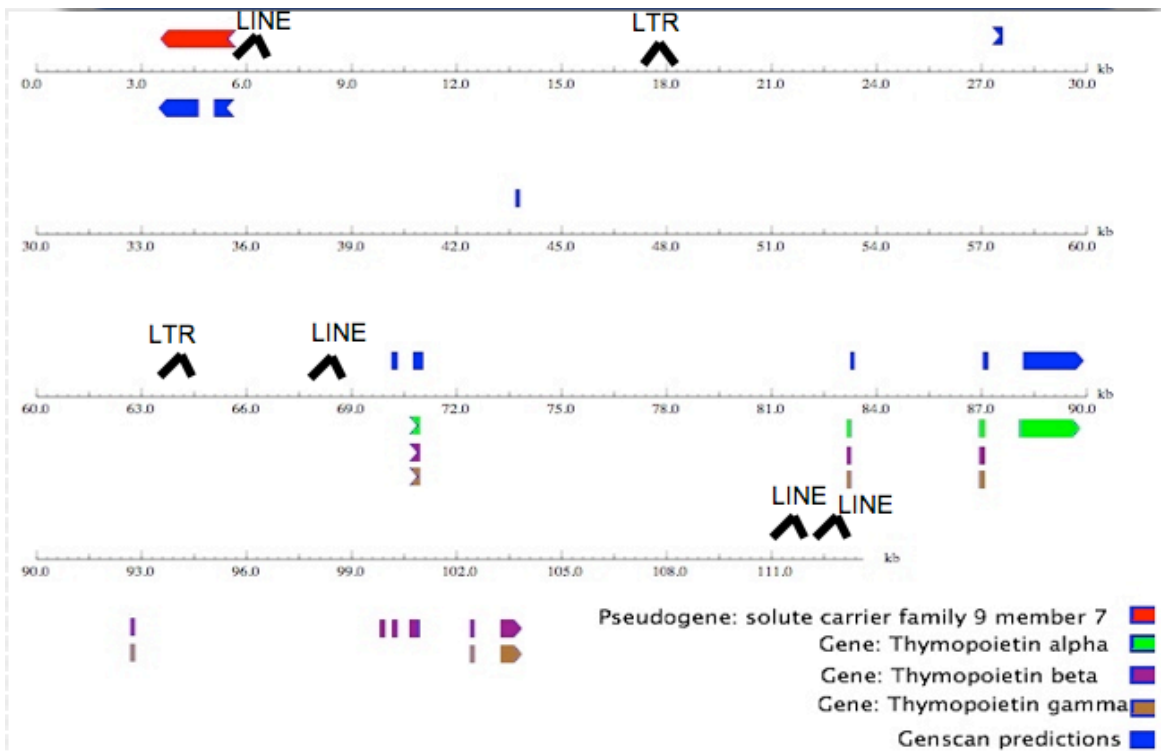


Figure 11: Final annotation map showing one pseudogene, three isoforms of one gene, and repetitive elements

## Future Work

There remains one important question that should be addressed in future work on this chimp chunk. I was unable to determine why there are so many EST matches in the region between base pairs 69000 and 71000. This region needs to be analyzed in greater detail to determine the reason why there are so many EST matches. As I discussed earlier, it is possible that the EST matches are unreliable. Therefore, it would be beneficial to analyze the specific EST matches to see where they come from and determine their importance and reliability. Furthermore, I mentioned that the annotation of the human TMP protein could be incomplete. In this manner, it would also be reasonable to investigate this annotation further and see if it is indeed missing a portion of the first exon. All in all, Jerome and I have shown many important features present within chimp chunk 2-10, but further analysis of the region from base pairs 69000 to 71000 is needed to explain the high density of human EST matches present in this region.