Sarah Spencer, Partner: Jien Shim 11 April 2008
Bio 4342

Annotating the Chimp Genome: Chunk 2-10
[First Revision]

Summary

In order to practice and better understand the process of gene annotation, J. Shim and I analyzed Chunk 2-10, a 113 kb portion of the chimpanzee (*Pan troglodytes*) genome. GENSCAN identified two features in our sequence, and we did not identify any additional features. J. Shim focused on RepeatMasker results and the annotation of Feature 1, both of which I summarize in this paper. My in-depth analysis of Feature 2 follows. Considered together, our work presents a complete and thorough annotation of Chunk 2-10 that improves upon the initial GENSCAN predictions. A table and a map summarizing our final annotations appear at the end of this report (Figure 12, Table 4).

All documented locations for genetic features refer to relative locations on Chunk 2-10 and not the *P. troglodytes* source chromosome. For annotation we frequently utilized the National Center for Biotechnology Information's (NCBI's) Basic Local Alignment and Search Tool (BLAST). We also used the BLAST-Like Alignment Tool (BLAT) by UC Santa Cruz (UCSC). Databases for alignment searches included NCBI's GenBank sequence database and NCBI's Reference Sequence (RefSeq) nucleotide sequence database.

GENSCAN Output

We began our analysis with the output of GENSCAN, a gene identification program that also designates probable exon and intron structures. The results of GENSCAN on Chunk 2-10 appear in Figures 1 and 2 and are summarized in Table 1. GENSCAN predicted two genes in the Chunk 2-10 chimpanzee sequence. The first predicted gene is referred to as Feature 1 in this paper and has two exons. The second predicted gene is referred to as Feature 2 and has seven putative exons. We hypothesized that Feature 1 might be a pseudogene because it has only two predicted exons (compared to an average of eight exons in functional human genes). Subsequent expressed sequence tag (EST) analysis revealed no other significant features in Chunk 2-10.

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.03	PlyA	-	1018	1013	6							1.05
1.02	Term	-	4627	3701	927	1	0	44	41	691	0.647	51.16
1.01	Init	-	5434	5069	366	1	0	104	87	335	0.981	32.07
1.00	Prom	-	15261	15222	40							-8.15
2.00	Prom	+	23920	23959	40							-3.95
2.01	Init	+	27519	27607	89	2	2	38	101	47	0.096	1.16
2.02	Intr	+	43684	43814	131	1	2	26	49	112	0.050	0.52
2.03	Intr	+	70135	70318	184	2	1	137	17	168	0.089	12.52
2.04	Intr	+	70753	71044	292	1	1	-7	-15	476	0.097	24.01
2.05	Intr	+	83244	83370	127	2	1	69	106	154	0.958	14.63
2.06	Intr	+	87034	87192	159	2	0	41	93	157	0.986	10.54
2.07	Term	+	88179	89698	1520	1	2	130	39	710	0.889	59.97
2.08	PlyA	+	90779	90784	6							1.05

Figure 1. GENSCAN output table.

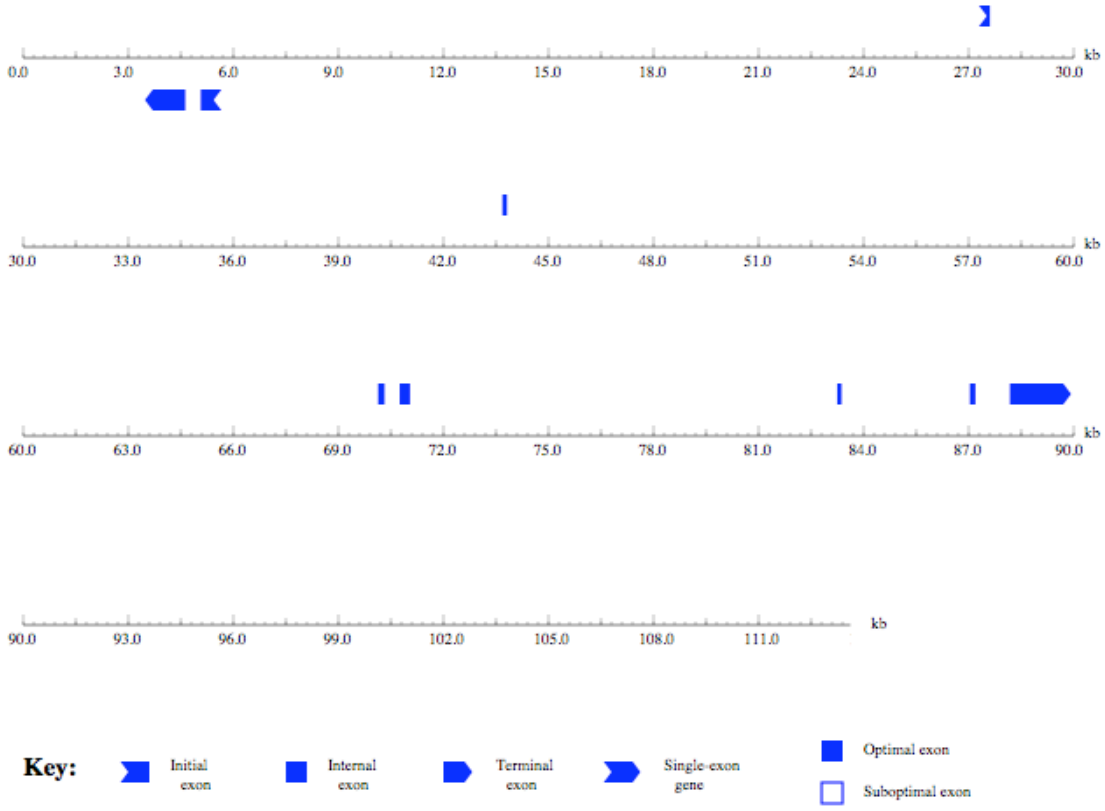


Figure 2. GENSCAN predicted genes. Feature 1 extends from 3701-5434 bp. Feature 2 extends from 27519-89698 bp.

Feature	GENSCAN Gene Prediction (bp)	Number of Exons
1	3701-5434	2
2	27519-89698	7

Table 1. Summary of GENSCAN gene predictions and designation of Chunk 2-10 features.

RepeatMasker Output

Before investigating Features 1 and 2, we characterized our chimpanzee sequence by running RepeatMasker. RepeatMasker is a program that identifies interspersed repeats and low-complexity DNA sequences. The results of running RepeatMasker on Chunk 2-10 are summarized in Figure 3. All together, interspersed repeats constitute 48.35% of our chimpanzee sequence. This is consistent with the large number of repetitive elements found in the human genome – repeats are currently estimated to make up approximately 50% of the genome. Finer classification of the interspersed repeats shows that Chunk 2-10 is composed of 31.49% SINEs, 8.29% LINEs, 6.75% LTR elements, 1.69% DNA elements, and 0.13% unclassified interspersed repeats. Other than interspersed repeats, 0.13% of Chunk 2-10 codes for small RNA molecules.

	number of elements*	length occupied	percentage of sequence
SINEs:	135	35774 bp	31.49 %
ALUs	122	33951 bp	29.88 %
MIRs	13	1823 bp	1.60 %
LINEs:	28	9420 bp	8.29 %
LINE1	17	6070 bp	5.34 %
LINE2	10	2927 bp	2.58 %
L3/CR1	1	423 bp	0.37 %
LTR elements:	23	7667 bp	6.75 %
MaLRs	13	4746 bp	4.18 %
ERVL	2	282 bp	0.25 %
ERV_classI	8	2639 bp	2.32 %
ERV_classII	0	0 bp	0.00 %

(continued)

DNA elements:	13	1925 bp	1.69 %
MER1_type	7	1189 bp	1.05 %
MER2_type	3	495 bp	0.44 %
Unclassified:	1	150 bp	0.13 %
Total interspersed repeats:		54936 bp	48.35 %
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 3. Summary of RepeatMasker results for Chunk 2-10.

Using the RepeatMasker output, we identified and analyzed five interspersed repetitive elements that are longer than 500 bp in Chunk 2-10. To determine if the repeats are still active transposons, we ran the nucleotide sequence of each repeat in BLASTx using the non-redundant protein sequence (nr) database. Three of the five repeats did not display meaningful matches to the protein database. The remaining two repeats, however, showed partial alignments with retrotransposon elements and reverse transcriptase proteins (Figure 4). Upon examination, none of the alignments extend across the entire length of a retrotransposon or a reverse transcriptase protein. We therefore concluded that the Chunk 2-10 repeats are no longer active transposons. The retrotransposon partial alignments, however, indicate that two of the five long repeats may have been active relatively recently in evolutionary history. The five repeats and their BLASTx results are summarized in Table 2.

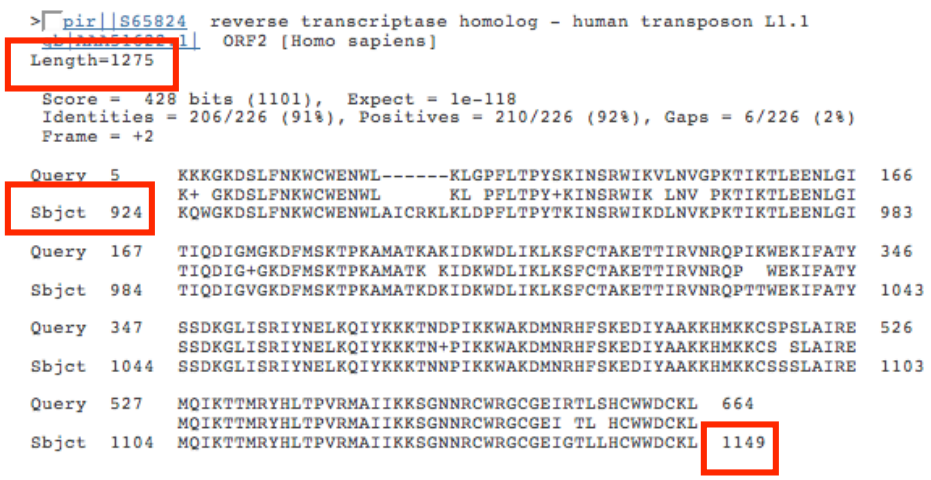


Figure 4. Alignment of a Chunk 2-10 repeat to a reverse transcriptase homolog in a human transposon. The alignment only includes 226 of the 1275 amino acids in the reverse transcriptase protein.

Repeat Class/Family	Location in Chunk 2-10 (bp)	Length (bp)	Repeat Class/Family	Retrotransposon Matches*
LINE/L1	5900-6564	665	L1P1	Yes
LTR/ERV1	63849-64452	604	MER4A	No
LINE/L2	67812-68878	1067	L2	No
LINE/L1	110955-111597	643	L1MEc	No
LINE/L1	112645-113155	511	L1M5	Yes

*from BLASTx search against non-redundant protein (nr) database; all elements appear currently inactive

Table 2. Description of Chunk 2-10 repeats longer than 500 bp.

Feature 1: Predicted Gene Investigation

We approached the GENSCAN gene prediction of Feature 1 with suspicion because of its small number of exons. A BLASTp comparison between NCBI's non-redundant protein sequence (nr) database and the predicted GENSCAN sequence revealed two putative conserved domains. These two domains, NhaP and Na^+/H^+ Exchanger, have similar functions in sodium-hydrogen ion transport.

The BLASTp alignments with the highest alignment scores matched to the human protein "solute carrier family 9 (sodium/hydrogen exchanger), member 7" or "SLC9A7" found at the human locus Xp11.23. In a subsequent BLAT alignment search between the human genome and the predicted Feature 1 gene, the highest alignment match was to human chromosome 12. The match to human chromosome 12 had a 98.8% identity compared to the 94.0% identity of the chromosome X match. A higher identity match to chromosome X would be expected for a SLC9A7 ortholog.

Starting with the lack of chromosomal synteny between the predicted gene and the true human gene, multiple lines of evidence converged to implicate Feature 1 as a pseudogene. The BLAT browser of chromosome 12 displayed a distinct lack of RefSeq matches to the region. Additionally, SLC9A7 has 15 exons in the human chromosome X while the chromosome 12 BLAT alignment showed only two exons. The loss of exons

and altered chromosomal location implies that Feature 1 is probably a pseudogene derived from a retrotransposition event.

To cross-validate these findings, a tBLASTn alignment was conducted using the functional human SLC9A7 protein sequence as a query of the Chunk 2-10 sequence. Although a large portion of the human protein matched along the chimpanzee subject, the percent identity of the alignment was only 74%. This is far below the orthology threshold of 98% identity. In addition, the alignment revealed numerous stop codons along the protein sequence, characteristic of a pseudogene (Figure 5).

Comparing human EST sequences to the sequence of Chunk 2-10 (BLASTn), displayed a region of EST matches that had an average of 95% identity. This is once again below the 98% identity threshold for homologous regions.

There is a small region of five EST matches 1 kb upstream of Feature 1, but the ESTs all have low identity matches (<90%) and most likely do not indicate a coding region. A BLASTn alignment between Chunk 2-10 and the SLC9A7 mRNA sequence revealed no evidence of an untranslated region (UTR). The fact that two of the five ESTs in question (AA284109.1, W47657.1) carry Alu elements suggests that this region of EST matches may result from confounding repeats or simply experimental error.

In view of this, according to the previously conducted tBLASTn alignment the probable pseudogene extends from 3347 bp to 5593 bp (Figure 5). Note that immediately downsteam of this region, extending from 5900-6564 bp, is a LINE element (Table 2). The presence of this long repetitive element supports the termination site of the SLC9A7 derivative and may have enabled the transposition event that led to the pseudogene's formation.

```
 Score = 1050 bits (2716),  Expect = 0.0
 Identities = 570/762 (74%), Positives = 604/762 (79%), Gaps = 52/762 (6%)
 Frame = -1

Query  2     EPGDAARPGSGRATGAPPPRLLLLPLLL-GWGLRVAAAASASSSGAAAEDSSAMEELATE  60
             EPGDAA P  GR   APP R         G GLRV A ASASSSGAA E+SSAMEEL TE
Sbjct  5593  EPGDAALPCPGRVAQAPPRRXXXXXXXXXGRGLRVTAEASASSSGAAVENSSAMEELVTE  5414

Query  61    KEAEESHRQDSVSLLTFILLLTLTILTIWLFKHRRVRFLHETGLAMIYGLIVGVILRYGT  120
             KEAEESHR DSVSLLTFILLLTLTILTIWLFK+ RV FLHETGLAMI GLIVGVILRYGT
Sbjct  5413  KEAEESHRPDSVSLLTFILLLTLTILTIWLFKYCRVHFLHETGLAMICGLIVGVILRYGT  5234

Query  121   PATSGRDKSLSCTQEDRAFSTLLVNVSGKFFEYTLKGEISPGKINSVEQNDMLRKVTFDP  180
             P T GRDK L+CTQED+AFSTL+V+VSGKFFEYTLK EISPGKINSV+QNDML KVTFDP
Sbjct  5233  PGTRGRDKLLNCTQEDQAFSTLVVDVSGKFFEYTLKREISPGKINSVKQNDMLGKVTFDP  5054

Query  181   EVFFNILLPPIIFHAGYSLKKRHFFRNLGSILAYAFLGTAVSCFII----GNLMYGVVKL  236
              VFFNILLPP+I HAGYSLK RHFFRNLGS L    LG    CF++      + YG
Sbjct  5053  *VFFNILLPPVISHAGYSLK-RHFFRNLGSRL----LGDC--CFVLPYWKSQVWYGEAH-  4898

Query  237   MKIMGQLSDKFYYTDCLFFGAIISATDPVTV---LAIFNELHADVDLYALLFGESVLNDA  293
                   L +   +T  LF    +      T        NELHAD+DLY LLFGES+LND
Sbjct  4897  EDYETALRELLHTLSLF*SNHLCH*SGTCSDCAGDINELHADMDLYVLLFGESILNDV  4718

Query  294   VAIVLSSSIV-AYQPAGLNTHAFDAAAFFKSVGIFLGIFSGSFTMGAVTGV-NANVTKFT  351
             V +  S  ++   Q    +THAFDAAAF KSVGIFLGIFSG FTMGAVTGV  A VTKFT
Sbjct  4717  VVLYFSHLLLPTSQQD*TSTHAFDAAAFLKSVGIFLGIFSGCFTMGAVTGVVTALVTKFT  4538

Query  352   KLHCFPLLETALFFLMSWSTFLLAEACGFTGVVAVLFCGITQAHYTYNNLSVESRSRTKQ  411
             KL CFPLLETALFFLMSWSTFLLAEACGFTGVVAVLFCGITQAHYT+NNL VESRSR+KQ
Sbjct  4537  KLDCFPLLETALFFLMSWSTFLLAEACGFTGVVAVLFCGITQAHYTFNNLLVESRSRSKQ  4358

Query  412   LFEVLHFLAENFIFSYMGLALFTFQKHVFSPIFIIGAFVAIFLGRAAHIYPLSFFLNLGR  471
             LFE     AENFIFS M LALFTFQKHVFSP+FIIGAFVA+FLGRAAHIYPLSFFL+LGR
Sbjct  4357  LFE-----AENFIFSCMVLALFTFQKHVFSPVFIIGAFVAVFLGRAAHIYPLSFFLSLGR  4193

Query  472   RHKIGWNFQHMMMFSGLRGAMAFALAIRDTASYARQMMFTTTLLIVFFTVWIIGGGTTPM  531
             RHKIGWNFQH MMFSGLRGA+AFALAI DTASYARQM FTTT  IVFFT+WIIGGGTTPM
Sbjct  4192  RHKIGWNFQHTMMFSGLRGAVAFALAICDTASYARQMTFTTTPFIVFFTIWIIGGGTTPM  4013

Query  532   LSWLNIRVGVEEPSEEDQNEHHWQYFRVGVDPDQDPPPNNDSFQVLQGDGPDSARGNRTK  591
             LSWLNIRV ++EPS+ED+NEHHWQYFRVGVDPDQDPPPNNDSFQVLQGD PDSARGN TK
Sbjct  4012  LSWLNIRVSIKEPSKEDRNEHHWQYFRVGVDPDQDPPPNNDSFQVLQGDSPDSARGNWTK  3833

Query  592   QESAWIFRLWYSFDHNYLKPILTHSGPPLTTTLP-------------------------  625
             QES WIFRLWYSFDHNYLKPILTHSG PLTTTLP
Sbjct  3832  QESTWIFRLWYSFDHNYLKPILTHSGSPLTTTLPPGGDTAAFH*PPLSCLV*TKQLPTNH  3653

Query  626   ---AWCGLLARCLTSPQVYDNQEPLREEDSDFILTEGDLTLTYGDSTVTANGSSSSHTAS  682
                AWC LL  CLTSPQVYDNQEPLRE +SDFILTEGDLTLTYGDSTVTANG S SHTAS
Sbjct  3652  HSPAWCSLLA*CLTSPQVYDNQEPLREGNSDFILTEGDLTLTYGDSTVTANGFSGSHTAS  3473

Query  683   TSLEGSRRTKSSSEEVLERDLGMGDQKVSSRGTRLVFPLEDN  724
             TSLEGS R KSSSEEVLERDLGMG+QKV S+GTRLVFPLEDN
Sbjct  3472  TSLEGSWRMKSSSEEVLERDLGMGNQKVLSQGTRLVFPLEDN  3347
```

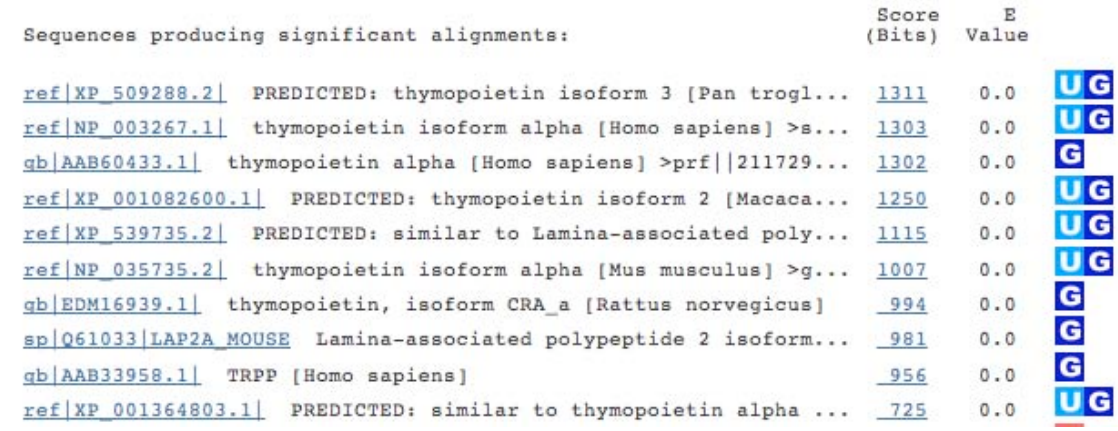
Figure 5. tBLASTn alignment of Chunk 2-10 nucleotide sequence with the human SLC9A7 protein sequence. Premature stop codons are indicated with red arrows.

Feature 2: Predicted Gene Investigation

With our chimpanzee sequence characterized for repeats and potential genes, I focused my work on the annotation of Feature 2. First, I used the predicted peptide sequence of Feature 2 from GENSCAN in a BLASTp query of the non-redundant protein sequence (nr) database. The search identified two putative conserved domains. The first was the 49 bp thymopoietin (TMPO) protein domain. TMPO proteins are ubiquitously expressed and involved in nuclear envelope organization. The second conserved domain identified was the 50 bp LEM domain. The LEM domain encodes two alpha helices and is associated with inner nuclear membrane proteins.

Moving on to the BLASTp protein alignments, I found ten hits with an E value of 0.0 and many more with extremely low E values (Figure 6). All the matched proteins fell into the thymopoietin protein family. This gave me confidence that Feature 2 contained a real gene in the thymopoietin family. The second entry bolstered my confidence, because it showed a low E value match to a manually curated human RefSeq entry. To check my hypothesis, I studied the top ten individual alignments. In each case the entire database protein sequence aligned to the Feature 2 query. This provided further evidence that the

predicted gene is a real gene: the predicted gene matched to the entire sequence of functional proteins from other organisms (including human).



Sequences producing significant alignments:	Score (Bits)	E Value
ref|XP_509288.2| PREDICTED: thymopoietin isoform 3 [Pan trogl...	1311	0.0
ref|NP_003267.1| thymopoietin isoform alpha [Homo sapiens] >s...	1303	0.0
gb|AAB60433.1| thymopoietin alpha [Homo sapiens] >prf||211729...	1302	0.0
ref|XP_001082600.1| PREDICTED: thymopoietin isoform 2 [Macaca...	1250	0.0
ref|XP_539735.2| PREDICTED: similar to Lamina-associated poly...	1115	0.0
ref|NP_035735.2| thymopoietin isoform alpha [Mus musculus] >g...	1007	0.0
gb|EDM16939.1| thymopoietin, isoform CRA_a [Rattus norvegicus]	994	0.0
sp|Q61033|LAP2A_MOUSE Lamina-associated polypeptide 2 isoform...	981	0.0
gb|AAB33958.1| TRPP [Homo sapiens]	956	0.0
ref|XP_001364803.1| PREDICTED: similar to thymopoietin alpha ...	725	0.0

Figure 6. Output of BLASTp showing top ten alignments. The GENSCAN protein sequence of Feature 2 was used as a query against the non-redundant protein sequence (nr) database.

I used the link to Entrez Gene available on the BLASTp output page to gather information on the human thymopoietin gene. The NCBI Entrez Gene page provided gene-specific information for the database subjects. I found that the thymopoietin gene in *Homo sapiens* is located on chromosome 12 at locus 12q22. Recall that the sequence of Feature 1 also maps to the human chromosome 12 (98.2% identity) according to a BLAT search. Even more, both Features 1 and 2 map to the chimpanzee chromosome 12 according to a BLAT search against the March 2006 genome assembly. This evidence indicates that Features 1 and 2 are syntenic between humans and chimpanzees. Continuing to investigate Feature 2, I found that there are three isoforms of the thymopoietin gene transcript in humans, all of which play a role in nuclear architecture. One has four exons, one has six exons, and one has nine exons.

The only complication I encountered in the BLASTp output was that none of the low E value matches aligned to the first 157 aa of the GENSCAN predicted protein (Figure 7). The most likely explanation for this observation was that GENSCAN overpredicted the size of the gene in this region. GENSCAN could have attached exons from a separate gene or identified spurious open reading frames (ORFs) in an intergenic region. Further investigation was necessary to determine the proper annotation. I have termed this region the "overpredicted region" for the purposes of this paper.

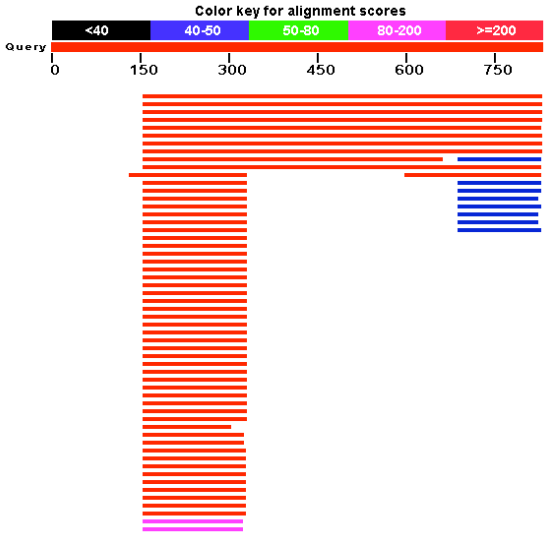


Figure 7. Distribution of low E value BLASTp hits on the second GENSCAN predicted protein query sequence (Feature 2). The non-redundant protein sequence (nr) database was used. Note the lack of matches to the first 157 amino acids of the GENSCAN predicted protein.

For the next step in my analysis, I used the UCSC Genome Browser to access the program BLAT and compare the GENSCAN predicted protein sequence with the May 2004 human genome assembly (Figure 8). BLAT returned two sequence matches, one with 88.7% identity and one with 99.5% identity.

The second match mapped to human chromosome 16 with 88.7% identity. I used a small region of well-aligned sequence to search the Conserved Domain Database for a match. The search revealed another likely LEM domain. I concluded that this region on human chromosome 16 could be paralogous to the human thymopoietin gene on chromosome 12 or it could encode a protein with a similar function in nuclear membrane organization; these possibilities require further investigation to verify.

The first BLAT match proved more informative. It showed 99.5% identity, which is within the range for human and chimpanzee orthologs (98-100%). Additionally, the match mapped to human chromosome 12, the chromosome previously found to encode the human thymopoietin protein. Since my query protein aligned with high identity to a human genomic region that encodes a functional protein, I became even more confident that Feature 2 was a human thymopoietin ortholog.

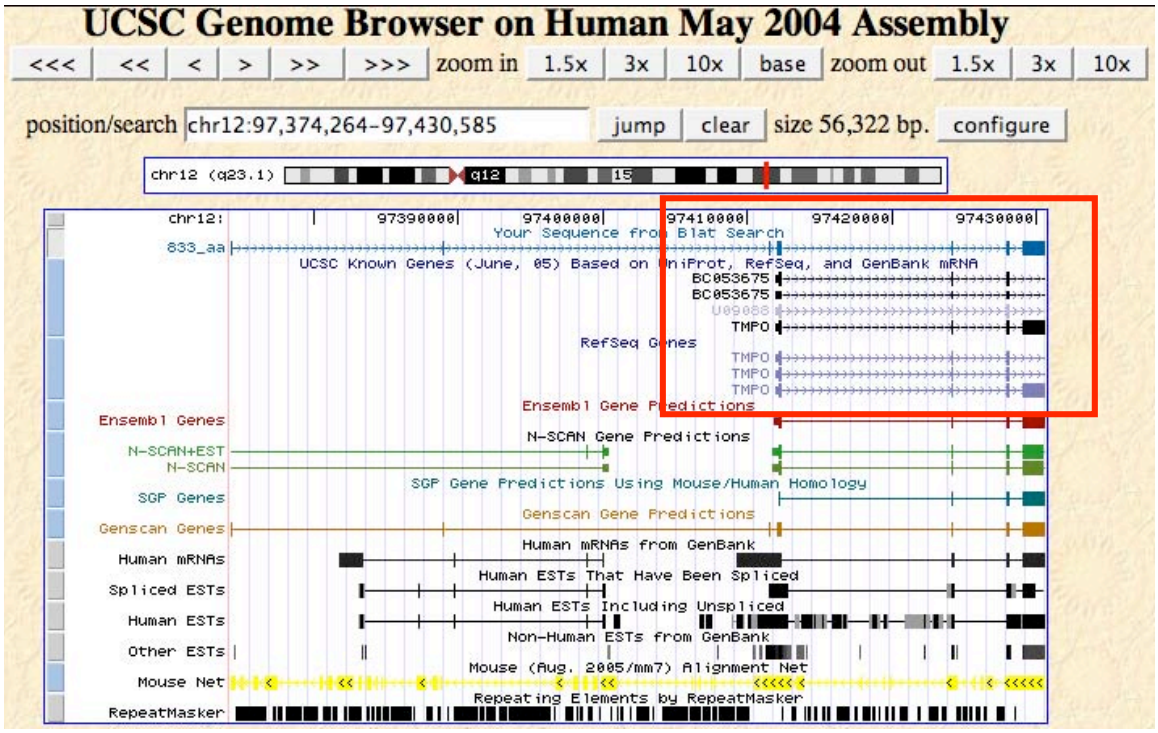
Human BLAT Results

BLAT Search Results

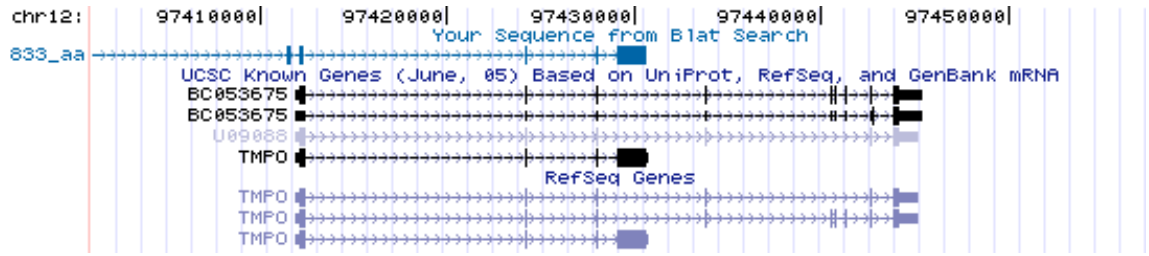
ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	833_aa	2465	1	833	833	99.5%	12	++	97374264	97430585	56322
browser details	833_aa	227	233	327	833	88.7%	16	++	73259013	73259294	282

Figure 8. BLAT results using the second GENSCAN predicted protein sequence (Feature 2) to query the May 2004 Human Genome Assembly.

I opened the UCSC Genome Browser for the top BLAT match and set the configuration of tracks seen in Figure 9.a. This BLAT browser clarified the structure and orientation of Feature 2 significantly. First, I noted that the GENSCAN prediction extends far upstream of the known TMPO gene in humans; this forms the overpredicted region I identified previously. Downstream of the GENSCAN prediction, however, it is apparent that two of three human thymopoietin isoforms extend farther than the prediction (Figure 9.b). The status of Feature 2 as a probably TMPO ortholog and the presence of well-defined TMPO RefSeq genes indicate a different gene model than GENSCAN predicted. With the new information, I felt fairly confident that Chunk 2-10 contained the ortholog of human TMPO, including all three isoforms.



(a)



(b)

Figure 9. (a) BLAT browser showing the exons of the second GENSCAN predicted protein. Three RefSeq genes match, but only to the final four exons of the predicted gene. Also, two of the RefSeq genes extend downstream of the GENSCAN prediction. (b) The BLAT browser is scrolled to show the human thymopoietin isoforms extending downstream of the GENSCAN prediction.

Feature 2: Exon Designation

As a final step to confirm my revised hypothesis, I used the protein sequences of the three human TMPO isoforms to query the entire unmasked Chunk 2-10 sequence. This required three separate iterations of tBLASTn. Each of the proteins matched with high identity along its entire length to the translated chimpanzee sequence, once again indicative of a real gene. The protein alignments were divided into discrete units (exons) by intervening regions of chimpanzee DNA (introns). I determined the location of TMPO exons in Chunk 2-10 by identifying the boundaries of 99-100% identity regions within each alignment (Figure 10). In other words, I ignored low identity regions (intronic regions) that BLAST continued off the ends of high identity (exon) matches. The results of my analysis and the probable locations of chimpanzee TMPO exons for the three isoforms appear in Table 3.

I did not identify any regions better conserved than others because all exon matches displayed almost perfect identities. I also found no evidence to suggest that the feature could not produce a working protein, such as premature stop codons or gaps that could cause frameshift mutations.

```
 Score =  113 bits (282),  Expect = 4e-22
 Identities = 58/65 (89%), Positives = 60/65 (92%), Gaps = 4/65 (6%)
 Frame = +3

Query  126    KYGVNPGPIVGTTRKLYEKKLLKLREQGTESRSSTPLPTISSSAENTRQNGSNDSDRYSD  185
              K+ +NPG    TTRKLYEKKLLKLREQGTESRSSTPLPTISSSAENTRQNGSNDSDRYSD
Sbjct  87015  KFCLNPG----TTRKLYEKKLLKLREQGTESRSSTPLPTISSSAENTRQNGSNDSDRYSD  87182

Query  186    NEEGK  190
              NEEGK
Sbjct  87183  NEEGK  87197
```

Figure 10. One region of tBLASTn alignment between human thymopoietin isoform alpha (query) and unmasked chunk 2-10. The beginning of 100% identity at chunk 2-10 base 87048 indicates the start of the third exon.

(a) Isoform Alpha - 694 aa

Exon	Alignment in Human Protein (aa)	Alignment in Chimp (bp)
1	1-93	70820-71098
2	94-136	83244-83348
3	137-190	87048-87197
4	189-694	88178-89695

(b) Isoform Beta - 454 aa

Exon	Alignment in Human Protein (aa)	Alignment in Chimp (bp)
1	1-93	70820-71098
2	94-136	83244-83348
3	137-188	87048-87191
4	189-221	92823-92921
5	222-262	100098-100220
6	262-294	100310-100408
7	294-331	100820-100933
8	331-359	102233-102319
9	360-454	103459-103737

(c) Isoform Gamma - 345 aa

Exon	Alignment in Human Protein (aa)	Alignment in Chimp (bp)
1	1-93	70820-71098
2	94-136	83244-83348
3	137-188	87048-87191
4	189-222	92823-92924
5	222-253	102233-102319
6	251-345	103459-103737

Table 3. Probable exons of three thymopoietin isoforms in chunk 2-10 as determined by tBLASTn alignment with human proteins. (a) thymopoietin isoform alpha, (b) thymopoietin isoform beta, (c) thymopoietin isoform gamma. All designated exons showed 99-100% identity in the tBLASTn alignment.

Human EST evidence provides additional support for my conclusions about Feature 2. I generated this line of evidence by aligning the repeat-masked Chunk 2-10 sequence to the human EST database with BLASTn. The results were displayed in Herne, a BLAST viewing tool (Figure 11). The human ESTs matching to Feature 2 show 99-100% identity and thus support the orthology hypothesis once again. In addition, the probable exon boundaries I determined match perfectly to regions of numerous ESTs. This supports the functionality of my predicted gene as well as the isoform exon designations.

Herne showed three 1-2 kb regions of EST matches not included in the predicted exons. One region is upstream of the initial exon for all isoforms, one is downstream of

the terminal exon for isoform alpha, and the last is downstream of the terminal exon for isoforms beta and gamma. These are likely 3' and 5' un-translated regions (UTRs) for the different TMPO isoforms. The location and size of these regions is consistent with UTR locations in human TMPO isoforms. This evidence is consistent with and supports the model of strong genome conservation between humans and chimpanzees.

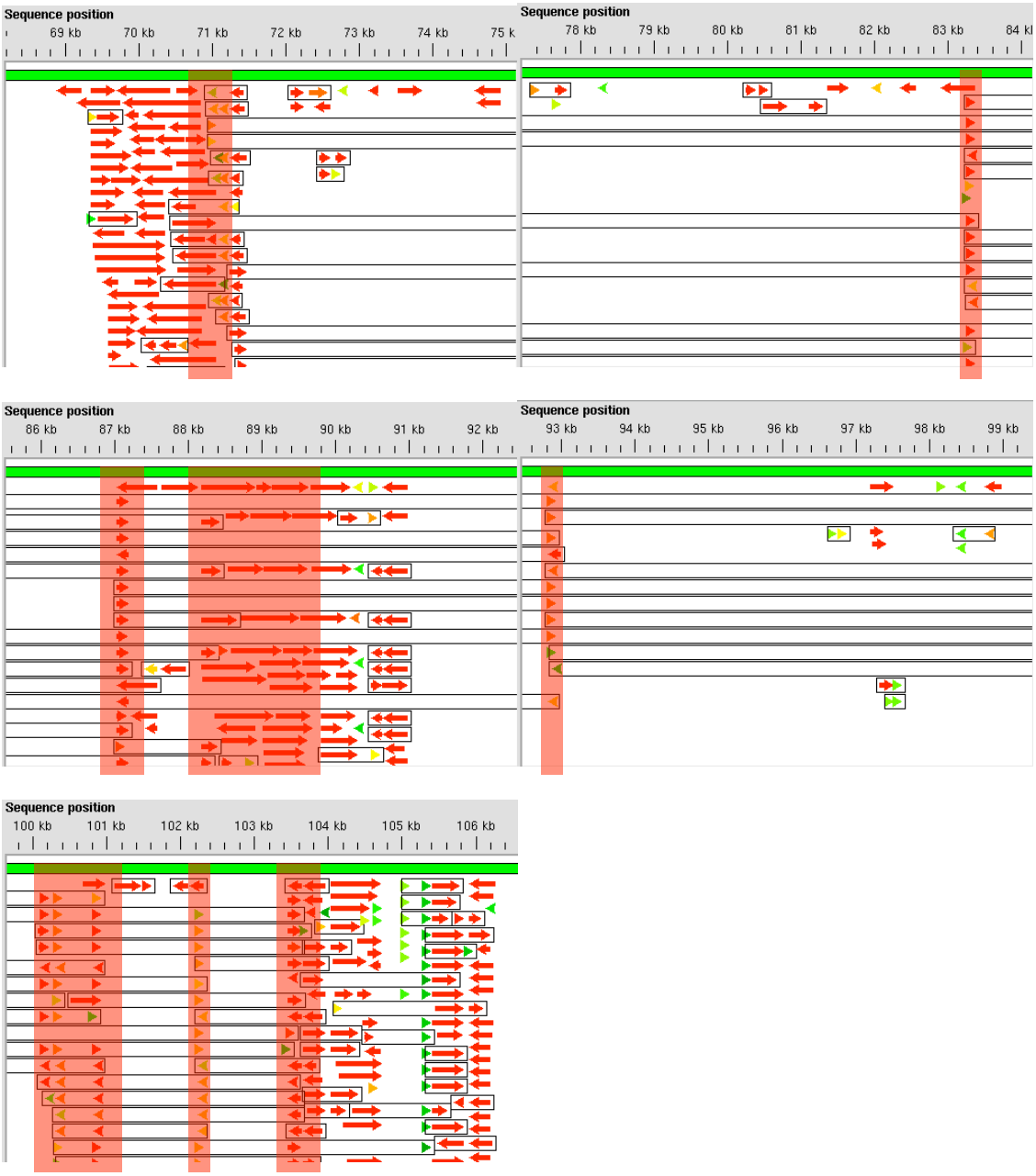


Figure 11. EST support for thymopoietin exon designations in chimpanzee. Approximate exon designations are shown in red boxed areas. Prominent EST regions not included in exon designations are most likely 3' and 5' UTRs for different isoforms The locations of EST regions not included in exon designations are consistent with the locations of human UTRs.

Feature 2: Overprediction

Confident in my TMPO annotation, I continued working on Feature 2 by investigating the GENSCAN overpredicted region upstream of the probable TMPO gene (Figure 9.a). In the high identity BLAT browser of human chromosome 12, there is one GenBank human mRNA and eight human ESTs (spliced) that map to the overpredicted region. This evidence indicated the possibility of an additional feature in the region.

Upon further examination, however, the human mRNA in the region (AK126817) only matched to hypothetical proteins. A tBLASTx comparison between the mRNA sequence and the repeat-masked Chunk 2-10 sequence revealed only poor alignments with low identities and frequent stop codons. This is a signature of intergenic sequence alignment. I also found that the human mRNA in question supported a RefSeq entry that has been retracted (XM 931792). A previous professional annotation designated a human gene in this region, but subsequent analysis led to the withdrawal of that gene designation. I conclude from this that there is currently not enough EST and mRNA evidence to support a gene model in the homologous chimpanzee overpredicted region. It is most likely an intergenic region with spurious ORFs.

As a second check of the overpredicted region, I used BLASTx to search for possible non-human protein matches to the overpredicted region. I tested the overpredicted subrange of Chunk 2-10 against the non-redundant (nr) Genbank protein database. BLASTx generated one alignment to a predicted gene in *Macaca mulatta* with a length of 71 aa and an identity of only 85%. This means it is possible that the overpredicted region contains a conserved functional sequence. More likely, however, the low identity and the high E value (2e-16) of the alignment imply a spurious match to the hypothetical protein.

Final Annotation

J. Shim and I are in complete agreement regarding our annotations for Chunk 2-10 and feel confident that we have improved the initial gene predictions from GENSCAN. Table 4 displays our final annotations for Features 1 and 2 compared to the original GENSCAN gene predictions. Figure 12 displays a schematic representation of our annotated chimpanzee Chunk 2-10, displaying our two features along with repetitive elements longer than 500 bp.

Feature	GENSCAN Gene Prediction (bp)	New Annotation (bp)	Characterization	Number of Exons
1	3701-5434	3347-5593*	Pseudogene, SLC9A7	2
2	27519-89698	70820-103737**	Gene, TMPO	4, 6, or 9

*Annotation by J. Shim, **Annotation by S. Spencer

Table 4. Final annotations for chimpanzee Chunk 2-10 after comprehensive analysis.

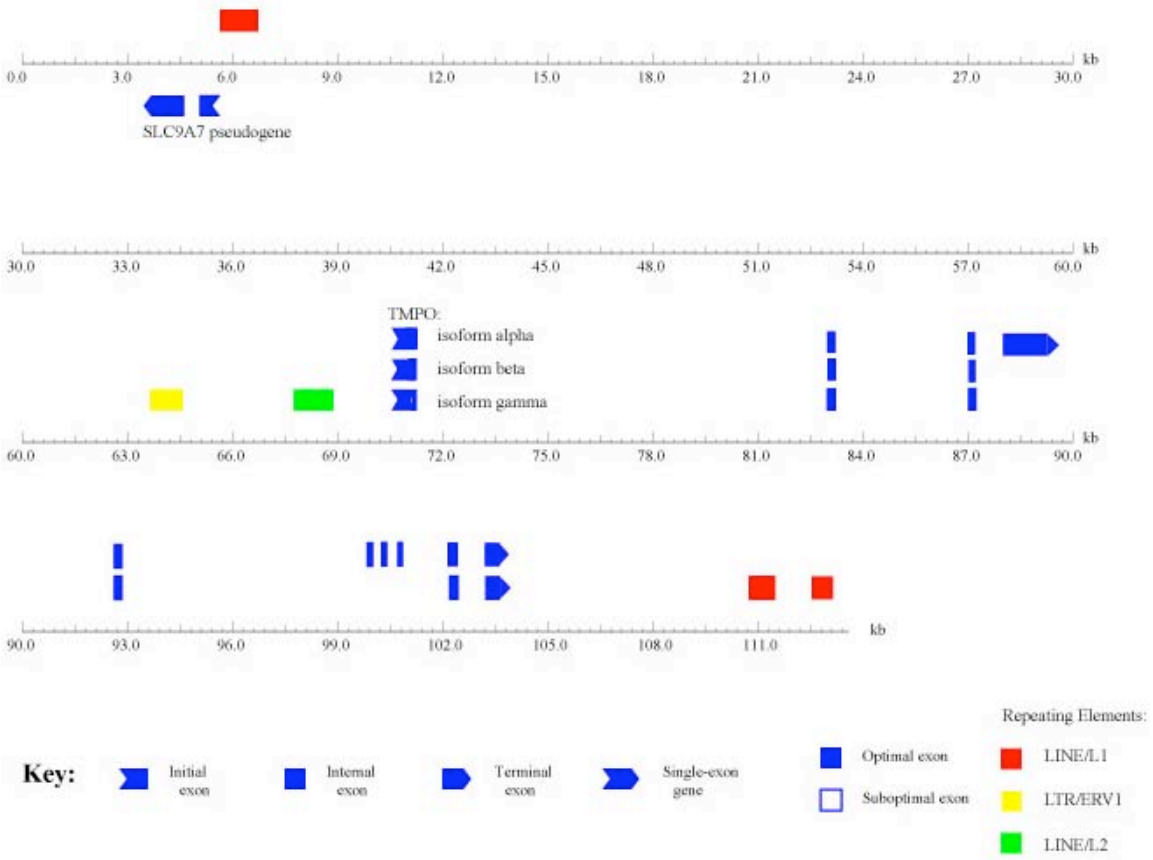


Figure 12. Final annotated map of chimpanzee Chunk 2-10.