Annotation of Chimp Chunk 2-10

Jerome M Molleston

5/4/2009

Abstract

A stretch of chimpanzee DNA was annotated using tools including BLAST, BLAT, and Genscan. Analysis of Genscan predicted genes revealed two predicted genes. One of these was a feature similar to human solute carrier family 9 member 7. Further analysis with BLAST and BLAT revealed that this is likely a pseudogene. The other feature predicted was revealed by BLAST analysis to be an ortholog of thymopoietin. EST evidence followed by downstream BLAST analysis revealed multiple isoforms of thymopoietin, specifically alpha, beta, and gamma. Finally, Repeat Masker analysis of the stretch of DNA revealed 6 repetitious elements larger than 500 bp in size. Overall, the region was successfully annotated with a reasonable degree of confidence.

Introduction:

In order to prepare for annotation of *Drosophila grimshawi* sequences as part of Biology 4342, Andrew Stein and I have annotated "chimp chunk" 2-10, a ~114 kb segment of chimpanzee DNA. Using Genscan, an *ab initio* gene finder, as well as Basic Local Alignment Search Tool (BLAST), we were able to find several regions of interest and compare them to known human genes in order to figure out where the genes and where the pseudogenes are in this chunk. By this method of analysis, we discovered the presence of one multiple-isoform gene as well as one pseudogene in chunk 2-10.

Initial Genscan predictions:

```
GENSCAN 1.0     Date run: 13-Mar-108    Time: 11:34:24

Sequence panTro2_dna : 113620 bp : 37.94% C+G : Isochore 1 ( 0 - 43 C+G%)

Parameter matrix: HumanIso.smat

Predicted genes/exons:


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: Initial Genscan predictions for chimp chunk 2-10.

Genscan (see Figure 1) predicted two genes, one having two exons and one having 7 exons. I was assigned feature 1, and Andrew was assigned feature 2, and any other regions of interest which came up were divided between us.

Results:

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: Features found and their properties.

Table 1 reveals the annotated sizes and natures of the two Genscan predictions. As can be seen, feature 1 was found to be a pseudogene, while feature 2 was revealed to be a gene with three isoforms, two of which (beta and gamma) were not predicted by Genscan.

Predicted Gene 1:

I began my analysis of predicted gene 1 by taking the first 12000 bp of my chunk and running BLASTx against the non-redundant protein database (nr). BLASTx translates the nucleotide query in all six possible reading frames and tries to align it against proteins in the database. This program was the ideal one to use, as it reveals sequences conserved at the protein level which may or may not be conserved at the nucleotide level (since some nucleotide divergence can occur which does not affect the protein sequence). I used 12000 base pars as my query rather than merely the Genscan prediction in order to make sure that any gene whose length was underestimated by Genscan would be revealed in full.

Figure 2: The highest NP (experimentally based) protein match for feature 1.

Most of the matches to this feature are to a gene called *solute carrier family 9 member 7* from various species. Figure 2 shows the best experimentally verified result in this region, which is the human solute carrier. This encompasses the feature predicted by Genscan, though the match is across the entire sequence rather than having a gap representing the intron which Genscan predicts. According to the protein database in Ensembl, this solute carrier has 17 exons. However, this sequence appears to have only one exon and has several stop codons within otherwise matching areas (boxed in red in Figure 2). In addition, the "intron" as reported by Genscan spans some areas which

match the protein according to BLAST, and spans regions which have similar numbers of amino acids. If this were truly an intron, it would produce large additional areas of sequence in the BLASTx translation which align with nothing in the protein sequence. Thus it seems more likely that this area represents a single-exon pseudogene rather than an ortholog of the solute carrier.

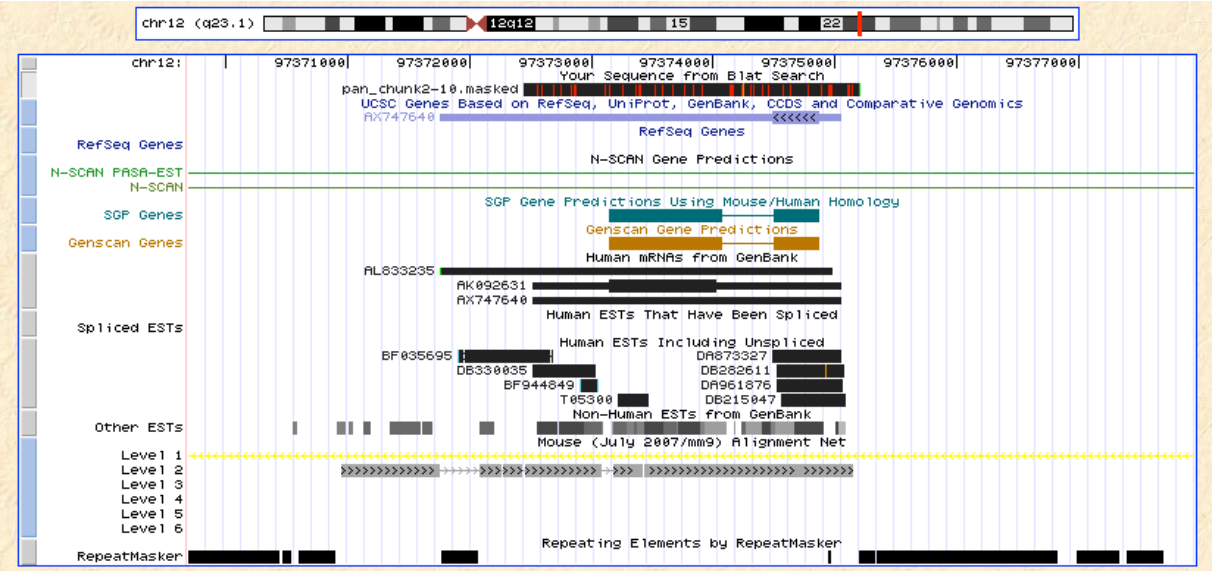

Figure 3: BLAT alignment of feature 1 to the human genome using UCSC genome browser.

To gain further evidence for feature 1's identity as a pseudogene, I used the University of California Santa Cruz (UCSC) genome browser and a tool called BLAST-Like Alignment Tool (BLAT), which looks for sequence similarity to the human genome and aligns the query sequence to its closest-matching region in human. As seen in Figure 3, the closest human match to feature 1 is in chromosome 12. However, the human solute carrier gene is located on chromosome X according to Ensembl. In this region in human (chromosome 12), there is a Genscan prediction similar to feature 1 in chimpanzee. In addition, there are no Refseq genes located there, but there are a few human mRNA sequences and ESTs visible in this region.

The ESTs and mRNA sequences in the region of figure 3 could weaken the pseudogene hypothesis, and thus must be addressed. None of the few mRNAs that appear to align here code for any known protein, and thus don't necessarily suggest a true gene, especially given the stop codons in this sequence. The EST alignments in this region are to be expected since the true solute carrier gene has a very similar sequence to this feature. This could easily produce a few individual EST sequencing reads that align here, especially if sequencing errors produced a few discrepant bases which more closely match this feature than the true gene. An actual gene in this area would produce many more EST reads and would likely have a Refseq annotation in human. In addition, the existence of a functional single-exon copy of a gene existing elsewhere in the genome is extremely unlikely. Therefore it still seems more likely that this feature is a pseudogene rather than a true gene.

The most logical conclusion is that this feature represents a retrotranscribed copy of the solute carrier mRNA which made its way into chromosome 12 sometime before the divergence of humans and chimpanzees (given that the pseudogene seems to appear in the human genome as well). Investigation of the mouse net alignment (which attempts to align the mouse chromosomes against the human ones) reveals that this area of the human chromosome does not align very well against mouse, since there is no level 1 alignment (see Figure 3 again), meaning that this area of the human chromosome does not have a high level of synteny with the mouse chromosme. Investigation of the level 2 alignment shows that the closest chromosomal match to this region is in mouse chromosome X, at the location of the functional solute carrier gene. Thus it seems likely that this pseudogene emerged after the divergence of mice and apes.

Predicted Gene 2:

Andrew Stein analyzed predicted gene 2. He began by taking the predicted protein sequence and running a BLASTp search against nr. He found a good alignment to thymopoietin alpha, with a 96% match. However, amino acids 1-157 of the Genscan prediction (the first three predicted exons) are not included in the thymopoietin alpha alignment. Those first 157 amino acids were aligned against nr using BLASTp and yielded no results, suggesting that Genscan had miscalled the first three exons. Subsequent BLAT alignment to the human genome reveals a 99.5% match to a portion of chromosome 12 of human at the location of the thymopoietin gene. The genome browser also reveals that human thymopoietin has 4 exons, as the matching portion of the Genscan prediction does. He also confirmed this with two-sequence BLASTx alignment of chimp chunk 2-10 against the human protein sequence for thymopoietin alpha, revealing four regions of matching sequence corresponding with the four exons of thymopoietin alpha.

Additional Isoforms of Gene 2:

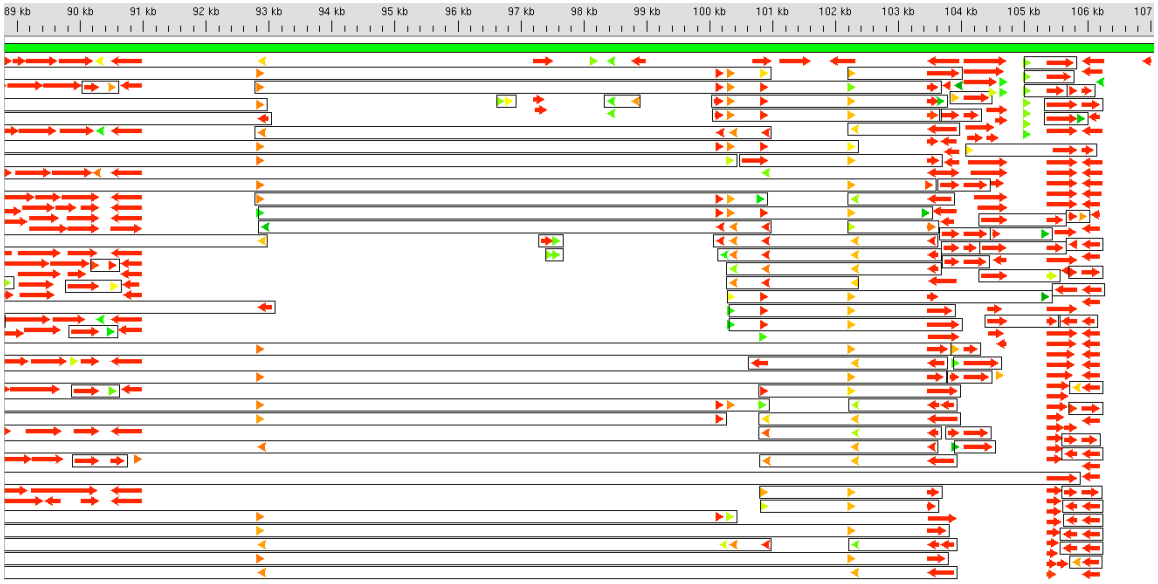

Figure 4: Alignment of human ESTs (red arrows) to chunk 2-10 in the region of 89-107 kb.

Scanning the nucleotide BLAST alignment of human ESTs to our chimp chunk, we found one suspicious region (Figure 4). Many ESTs align to the region immediately downstream of thymopoietin alpha, from 89000-107000 bp, so I decided to investigate this area by aligning it against the nr database using BLASTx.

```
>ref|NP_001027455.1| UG thymopoietin isoform gamma [Homo sapiens]
 gb|AAB60331.1| G thymopoietin gamma
 gb|AAC25390.1| G thymopoietin gamma [Homo sapiens]
 gb|EAW97589.1| G thymopoietin, isoform CRA_a [Homo sapiens]
 gb|EAW97594.1| G thymopoietin, isoform CRA_a [Homo sapiens]
Length=345

 GENE ID: 7112 TMPO | thymopoietin [Homo sapiens] (Over 10 PubMed links)

 Score =  164 bits (414),  Expect = 4e-37
 Identities = 97/106 (91%), Positives = 98/106 (92%), Gaps = 2/106 (1%)
 Frame = +2

Query  13427  FSFPPFTPN--SASCRRPIKGAAGRPLELSDFRMEESFSSKYVPKYVPFADVKSEKTKKG  13600
              F +   TP   SASCRRPIKGAAGRPLELSDFRMEESFSSKYVPKYVP ADVKSEKTKKG
Sbjct  240    FPYEASTPTGISASCRRPIKGAAGRPLELSDFRMEESFSSKYVPKYVPLADVKSEKTKKG  299

Query  13601  RSIPVWIKIllfvvvavflflvYQAMETNQVNPFSNFLHVDPRKSN  13738
              RSIPVWIKILLFVVVAVFLFLVYQAMETNQVNPFSNFLHVDPRKSN
Sbjct  300    RSIPVWIKILLFVVVAVFLFLVYQAMETNQVNPFSNFLHVDPRKSN  345


 Score = 70.1 bits (170),  Expect = 7e-09
 Identities = 34/34 (100%), Positives = 34/34 (100%), Gaps = 0/34 (0%)
 Frame = +1

Query  2824  DSKIELKLEKREPLKGRAKTPVTLKQRRVEHNQV  2925
             DSKIELKLEKREPLKGRAKTPVTLKQRRVEHNQV
Sbjct  189   DSKIELKLEKREPLKGRAKTPVTLKQRRVEHNQV  222


 Score = 62.0 bits (149),  Expect = 2e-06
 Identities = 30/32 (93%), Positives = 30/32 (93%), Gaps = 0/32 (0%)
 Frame = +3

Query  12234  VGEKTEERRVERDILKEMFPYEASTPTGIRYS  12329
              VGEKTEERRVERDILKEMFPYEASTPTGI  S
Sbjct  222    VGEKTEERRVERDILKEMFPYEASTPTGISAS  253
```

Figure 5: The top NP BLASTX result for region 90000-107000.

As seen in Figure 5, the area searched aligns relatively well to the downstream sequence of thymopoietin gamma, another isoform of the protein annotated by Andy. The match to the downstream sequence of the protein, backed up by EST data and without any stop codons to suggest a pseudogene, suggests the presence of alternative splicing in the thymopoietin gene.

Figure 6: Entrez diagram of the three isoforms of human thymopoietin (note that two of the exons of isoform beta are compressed and appear as one).

Investigating the Entrez report on thymopoietin (Figure 6), I found that it has three isoforms, alpha, beta, and gamma, produced by alternative splicing of the transcript. The first three exons are held in common by the three isoforms, with alpha having a fourth and beta and gamma having six and three additional exons, respectively. The three extra exons unique to thymopoietin gamma matched the three areas of BLAST alignment found in my search (Figure 5), and the fact that beta is transcribed from the same gene suggested that thymopoietin beta should also appear in the BLAST search.

```
GENE ID: 7112 TMPO | thymopoietin [Homo sapiens] (Over 10 PubMed links)

 Score =  164 bits (414),  Expect = 4e-37
 Identities = 97/106 (91%), Positives = 98/106 (92%), Gaps = 2/106 (1%)
 Frame = +2

Query  13427  FSFPPFTPN--SASCRRPIKGAAGRPLELSDFRMEESFSSKYVPKYVPFADVKSEKTKKG  1360
               F +   TP   SASCRRPIKGAAGRPLELSDFRMEESFSSKYVPKYVP ADVKSEKTKKG
Sbjct  349    FPYEASTPTGISASCRRPIKGAAGRPLELSDFRMEESFSSKYVPKYVPLADVKSEKTKKG  408

Query  13601  RSIPVWIKIllfvvvavflflvYQAMETNQVNPFSNFLHVDPRKSN  13738
               RSIPVWIKILLFVVVAVFLFLVYQAMETNQVNPFSNFLHVDPRKSN
Sbjct  409    RSIPVWIKILLFVVVAVFLFLVYQAMETNQVNPFSNFLHVDPRKSN  454


 Score = 85.1 bits (209),  Expect(2) = 1e-26
 Identities = 41/41 (100%), Positives = 41/41 (100%), Gaps = 0/41 (0%)
 Frame = +1

Query  10099  SYSQAGITETEWTSGSSKGGPLQALTRESTRGSRRTPRKRV  10221
               SYSQAGITETEWTSGSSKGGPLQALTRESTRGSRRTPRKRV
Sbjct  222    SYSQAGITETEWTSGSSKGGPLQALTRESTRGSRRTPRKRV  262


 Score = 65.1 bits (157),  Expect(2) = 1e-26
 Identities = 32/34 (94%), Positives = 34/34 (100%), Gaps = 0/34 (0%)
 Frame = +3

Query  10308  QVETSEHFRIEGPVISESTPIAETIMASSNESLV  10409
               +VETSEHFRI+GPVISESTPIAETIMASSNESLV
Sbjct  261    RVETSEHFRIDGPVISESTPIAETIMASSNESLV  294


 Score = 83.6 bits (205),  Expect = 6e-13
 Identities = 42/56 (75%), Positives = 45/56 (80%), Gaps = 0/56 (0%)
 Frame = +3

Query  10821  VVNRVTGNFKHASPILPITEFSDIPRRAPKKPLTRAEVNEYNLDRCYH*SFKEEIF  10988
               VVNRVTGNFKHASPILPITEFSDIPRRAPKKPLTRAEV E   +R       +E+F
Sbjct  294    VVNRVTGNFKHASPILPITEFSDIPRRAPKKPLTRAEVGEKTEERRVERDILKEMF  349


 Score = 68.6 bits (166),  Expect = 2e-08
 Identities = 33/33 (100%), Positives = 33/33 (100%), Gaps = 0/33 (0%)
 Frame = +1

Query  2824  DSKIELKLEKREPLKGRAKTPVTLKQRRVEHNQ  2922
              DSKIELKLEKREPLKGRAKTPVTLKQRRVEHNQ
Sbjct  189   DSKIELKLEKREPLKGRAKTPVTLKQRRVEHNQ  221


 Score = 62.0 bits (149),  Expect = 2e-06
 Identities = 30/32 (93%), Positives = 30/32 (93%), Gaps = 0/32 (0%)
 Frame = +3

Query  12234  VGEKTEERRVERDILKEMFPYEASTPTGIRYS  12329
               VGEKTEERRVERDILKEMFPYEASTPTGI  S
Sbjct  331    VGEKTEERRVERDILKEMFPYEASTPTGISAS  362
```

Figure 7: BLAST alignment of thymopoietin beta against chunk 2-10 region 90000-107000 bp.

As expected, thymopoietin beta is also present in the same BLAST search results (Figure 7), containing the expected six additional regions of alignment representing the six additional exons of thymopoietin beta. The alignment was overextended into areas bordering two of the exons (red boxes, Figure 7), including one region with a stop codon at amino acid 342. However, another region of alignment (likely another exon) covers these areas at the bottom of the depicted BLAST alignment, making it likely that they are in actuality intronic sequences. The Ensembl protein database confirms that the exon sizes correspond with the end of the good alignments rather than the BLAST overextension (data not pictured).

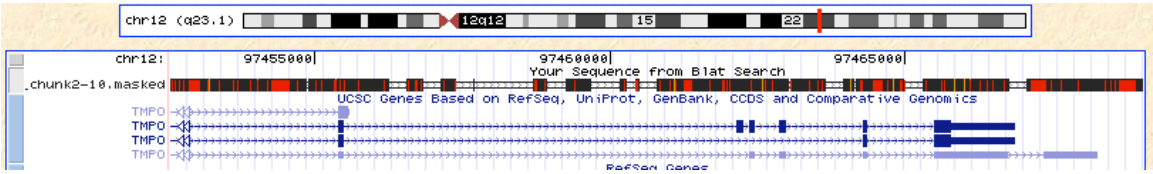

Figure 8: BLAT alignment of region 90000-107000 against human in the UCSC genome browser.

BLAT alignment of this region against the human genome (Figure 8) shows that the region maps best to the area in human where the thymopoietin isoforms (labeled TMPO in the alignment) are located, suggesting that the match to thymopoietin was the best one that could have been made. Thus I conclude that the ESTs in region 90000-107000 bp of my chunk represent the ortholog of the beta and gamma isoforms of the thymopoietin gene.

Repeat Masker Analysis:

```
==================================================
file name: pan_chunk2_10.fasta
sequences:             1
total length:     113620 bp  (104398 bp excl N/X-ru
GC level:         42.59 %
bases masked:      55163 bp ( 48.55 %)
==================================================
               number of      length   percentage
               elements*    occupied  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        7668 bp    6.75 %
      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:       15        2007 bp    1.77 %
     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 9: Repeat Masker 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/ERVL-MaLR	17826	18342	516
LINE/L1	112645	113155	510

Table 2: Repetitious elements above 500 bp in length.

Repeat Masker was run on chimp chunk 2-10 to identify and mask common repetitious elements to prevent their interference in annotation of the chunk. As seen in Figure 9, repeats were quite prevalent, with 48.55% of the sequence belonging to repetitious elements. Repetitious elements larger than 500 bp were annotated in the final map of the chunk, and are listed in Table 2. In brief, four LINEs and two LTRs were found larger than 500 bp.

Conclusion

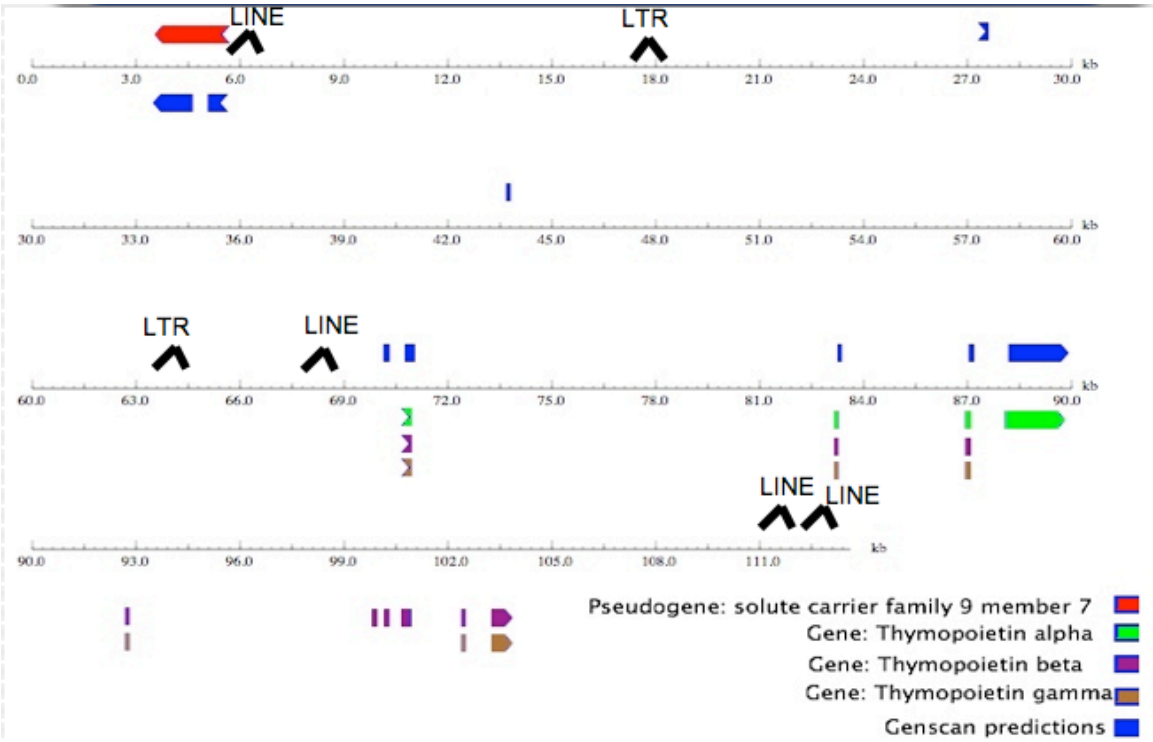

Figure 10: Final map of chunk 2-10 with all genes, pseudogenes, and repetitious elements noted.

In all, one pseudogene likely retrotranscribed from a *solute carrier family 9 member 7* message was found. In addition, the thymopoietin ortholog in chimpanzee was found, with isoforms alpha, beta, and gamma produced by alternative splicing. Finally, six repetitious elements over 500 bp were found and annotated. With the annotation project completed, we now have sufficient familiarity with BLAST, BLAT, Genscan, and the other tools of annotation to tackle the *D. grimshawi* annotation project.