Priya Srikanth
Bio 434W
4.13.2007

Chimp Chunk 2_7 Annotation: Revision 1

Introduction

Annotation is critical for obtaining an enhanced understanding of functional features present in genomic sequence data. Here, we demonstrate the annotation of an 80 kb region of the *Pan troglodytes* genome, Chimp Chunk 2_7. Because of the high similarity of chimp and human sequences and low phylogenetic distance between these species, we expect to find a degree of synteny between our Chimp Chunk 2_7 and human genomic sequence. The human genome sequence and annotation will be essential to chimp annotation as we compare our Chimp Chunk to the human genome to find functional features as well as artifacts of functional sequence, such as pseudogenes. Upon receiving chimp sequence, we began analysis by running RepeatMasker using the "-nolow" switch to mask repetitive sequence but not low complexity regions. The masked sequence was used to run Genscan, an *ab initio* gene finder that predicts putative protein-coding regions. These regions, as well as anonymous EST matches from blastn, were used to carry out analysis and eventual annotation of the chimp sequence. A summary of our predicted features for the region is shown in Table 1.

Final Feature	Exons	Position (bp)	Characterization
1	1	6288-7220	Pseudogene of RBMX2
2	1	18781-17553	SPZ1 (NP_115956.2)
3	1	35992-37033	Pseudogene of ACTB
4	3	41899-40661	Pseudogene of KRT18
5	1	23390-23238	Pseudogene of RPL39
6	--	76164-79827	Anonymous EST region

Table 1. Final features predicted after analysis of Chimp Chunk.[1]

Definitions

BLAST: <u>B</u>asic <u>L</u>ocal <u>A</u>lignment <u>S</u>earch <u>T</u>ool, an algorithm used to compare biological sequence data and find similarity.

blastp: Protein-protein BLAST, which compares a protein query sequence to a protein database of the user's choice. Examples of protein databases are: Swissprot, which contains curated[2] protein sequences; and nr, which is a database of non-redundant protein sequences that have been predicted and/or curated.

blastx: Translation-protein BLAST, which takes a nucleotide query, translates it in all 6 frames, and compares each translation to a protein database of the user's choice.

blastn: Nucleotide-nucleotide BLAST, which compares a nucleotide query sequence to a nucleotide database of the user's choice. Examples of nucleotide databases are: nt, a database of non-redundant nucleotide sequences that have been predicted

[1] "Final feature" indicates the annotation of a feature after analysis. In discussion of Genscan predictions, "feature" indicates the feature as Genscan predicted it.

[2] A curated sequence is one that has been manually reviewed by NCBI staff or collaborators. We used only curated sequences in our analysis of each feature to ensure that each sequence contained high-quality, reliable data.

and/or curated; EST, a database containing Expressed Sequence Tags – a sequence of transcribed mRNA that is usually a part of a transcription product, rather than the entire product; and Refseq, a database of mRNAs compiled from existing ESTs (using EST data to create models of whole transcription products) and experimentally-identified mRNAs of genes.

BLAT: BLAST-Like Alignment Tool, which compares a nucleotide or protein query to a genome of the user's choice (e.g. human or chimp) and identifies locations in the genome that exhibit similarity to the query.

blastp search: Use of the blastp tool to find similarity of a query to proteins in the chosen database.[3]

Initial Data

Different sources provided conflicting evidence of potential features in the Chimp Chunk. Genscan predicted final features 1 through 4 in our Chimp Chunk (Figures 1 and 2). Final features 1, 3, and 4 were also identified using a blastx search against the Swissprot database. Final features 1 through 4 were present in a blastn search against the Refseq database, and all six of the final features were identified by a blastn against the human EST database. Differences in feature identification depending on the chosen search database emphasize that many sources of data should always be used to analyze sequence data for annotation. After receiving the Genscan output, we proceeded to analyze each feature using the BLAST and BLAT tools.[4]

```
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------

 1.01 Sngl +   6288   7220  933  2  0   84   54   957 0.999  88.10
 1.02 PlyA +   8927   8932    6                                1.05

 2.03 PlyA -  10182  10177    6                                1.05
 2.02 Term -  18802  17553 1250  2  2    9   32   960 0.152  73.25
 2.01 Init -  36038  35879  160  2  1   91   72   262 0.756  24.93
 2.00 Prom -  36186  36147   40                              -17.34

 3.00 Prom +  36292  36331   40                              -17.61
 3.01 Sngl +  36374  37033  660  1  0   77   48   782 0.516  69.02
 3.02 PlyA +  37598  37603    6                                1.05

 4.06 PlyA -  40034  40029    6                                1.05
 4.05 Term -  41098  40661  438  1  0  -17   53   401 0.861  20.39
 4.04 Intr -  41527  41213  315  1  0  -28   61   530 0.932  34.04
 4.03 Intr -  41899  41631  269  2  2  111   42   287 0.692  22.73
 4.02 Intr -  77852  77657  196  2  1   31   57   136 0.137   2.87
 4.01 Init -  78129  77992  138  0  0   49   33   283 0.999  17.09
```

Figure 1. Genscan gene predictions.

[3] This "search" language will also be used with the blastn, blastx, and BLAT tools, indicating use of the tool for its intended purpose.

[4] All BLAT searches were performed using protein sequence unless otherwise stated.

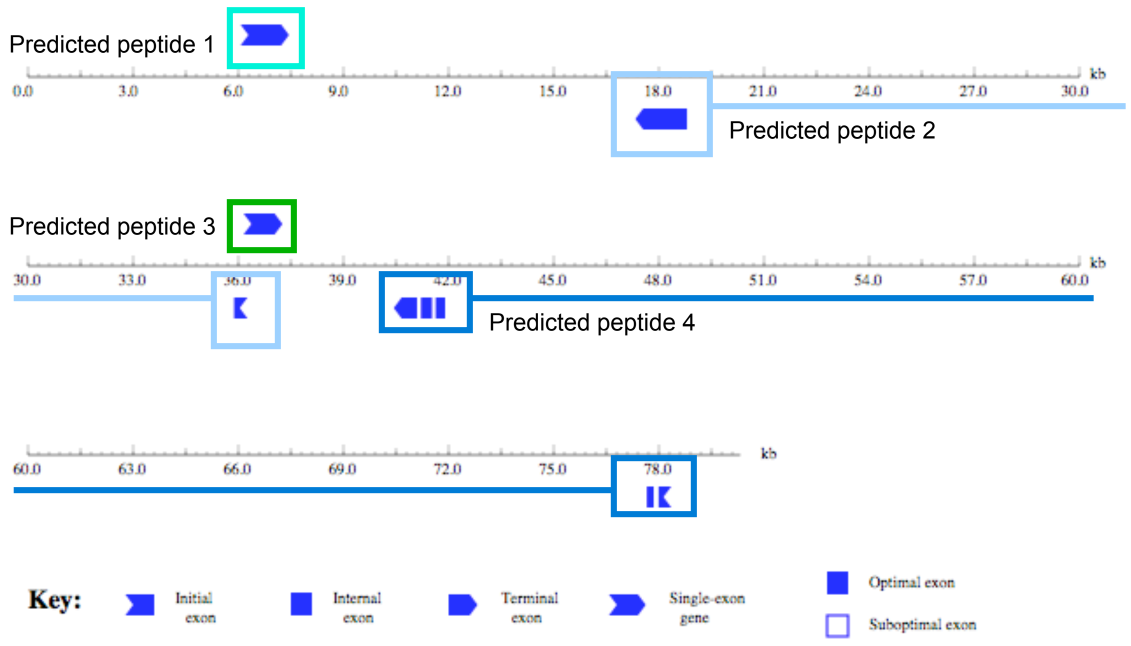


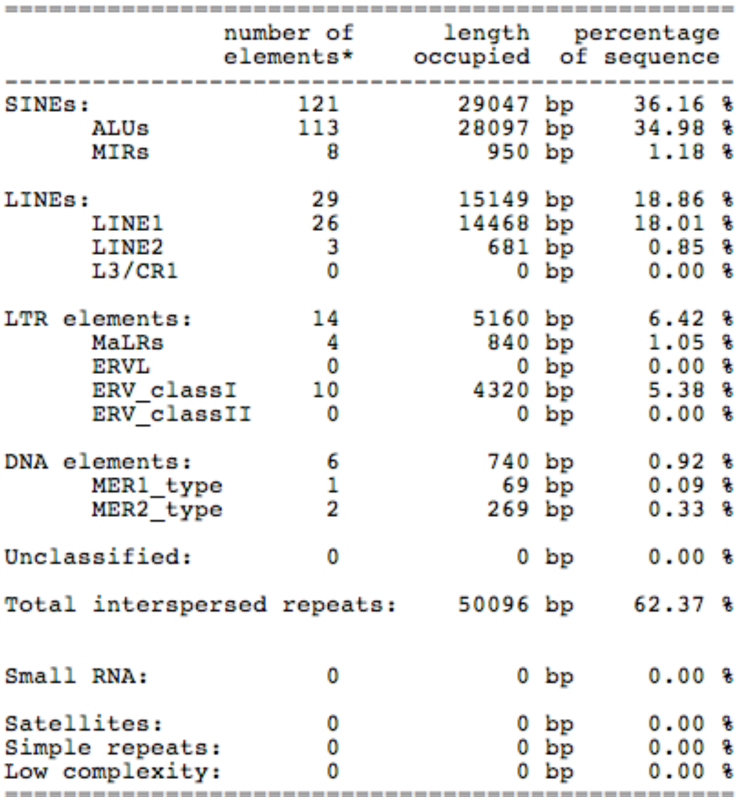
Figure 2. Genscan gene prediction map.

RepeatMasker

RepeatMasker identified nine repetitious features longer than 500 bp that were not SINEs (Table 2). The 80323 bp Chimp Chunk had 43.79% GC content and 62.37% masked bases. See Figure 3 for a summary of percentages of each type of repeat.

Location (bp)	Length	Class/Family	Repeat Type
1330-1839	509	LINE/L1	L1MD1
5125-5765	640	LINE/L1	L1ME2
10292-10911	619	LINE/L1	L1MB2
11083-11690	607	LINE/L1	L1MB2
15325-16186	861	LINE/L1	L1MDa
24478-25863	1385	LINE/L1	L1PA5
28273-29589	1316	LTR/ERV1	LTR12C
51365-52012	647	LINE/L1	L1M4c
73781-74395	594	LTR/ERV1	MER41B

Table 2. Long repeat regions.



```
==================================================
                    number of      length   percentage
                    elements*     occupied  of sequence
--------------------------------------------------
SINEs:                 121        29047 bp    36.16 %
      ALUs             113        28097 bp    34.98 %
      MIRs               8          950 bp     1.18 %

LINEs:                  29        15149 bp    18.86 %
      LINE1             26        14468 bp    18.01 %
      LINE2              3          681 bp     0.85 %
      L3/CR1             0            0 bp     0.00 %

LTR elements:           14         5160 bp     6.42 %
      MaLRs              4          840 bp     1.05 %
      ERVL               0            0 bp     0.00 %
      ERV_classI        10         4320 bp     5.38 %
      ERV_classII        0            0 bp     0.00 %

DNA elements:            6          740 bp     0.92 %
      MER1_type          1           69 bp     0.09 %
      MER2_type          2          269 bp     0.33 %

Unclassified:            0            0 bp     0.00 %

Total interspersed repeats:       50096 bp    62.37 %


Small RNA:               0            0 bp     0.00 %

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

Figure 3. RepeatMasker summary.

Feature 1

Using Genscan's feature 1, we performed a blastp search with the nr database. The best curated alignment was to human RNA binding motif protein, X-linked 2 (RBMX2). This gene contains an RRM, or RNA recognition motif, confirming its function as an RNA binding protein. The BLAST alignment of RBMX2 to feature 1 showed 84% sequence identity. The Entrez Gene entry for RBMX2 revealed that the human gene contains five exons, rather than one exon as predicted by Genscan. A BLAT search of the predicted peptide sequence against the human genome showed a 97.5% alignment in chromosome 5. The BLAT search produced an alignment of the human RBMX protein to the X chromosome and confirmed that RBMX is a five-exon gene. We then used the BLAT tool to compare the human RBMX2 protein sequence to the region of chromosome 5 where feature 1 aligned (Figure 4). A side-by-side alignment of the human protein to chromosome 5's similar region revealed a stop codon in the corresponding ORF of chromosome 5 (Figure 5).

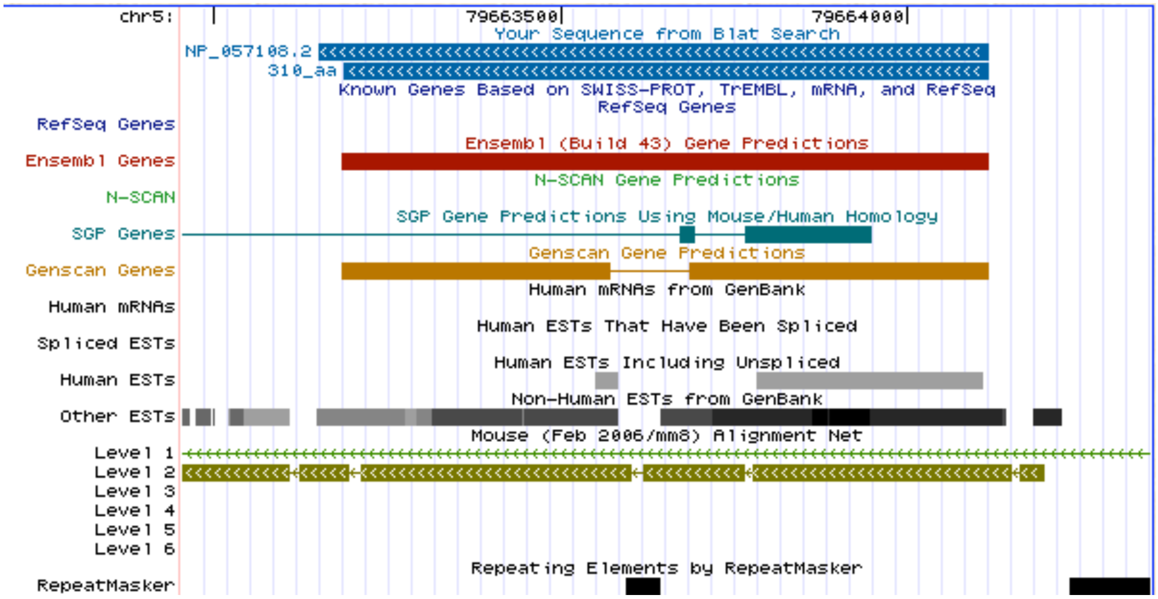
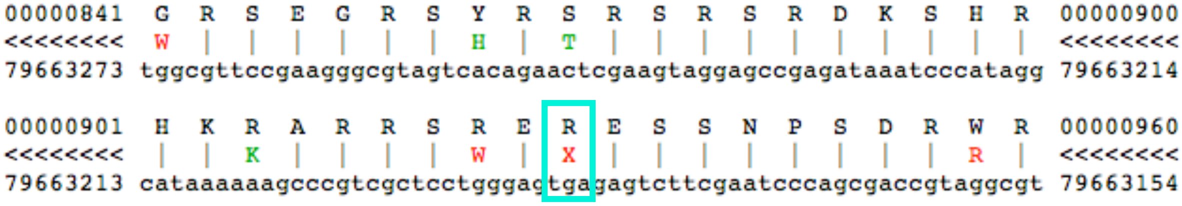

chr5:
79663500
79664000
Your Sequence from Blat Search
NP_057108.2
310_aa
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
RefSeq Genes
RefSeq Genes
Ensembl Genes
Ensembl (Build 43) Gene Predictions
N-SCAN
N-SCAN Gene Predictions
SGP Genes
SGP Gene Predictions Using Mouse/Human Homology
Genscan Genes
Genscan Gene Predictions
Human mRNAs
Human mRNAs from GenBank
Spliced ESTs
Human ESTs That Have Been Spliced
Human ESTs
Human ESTs Including Unspliced
Other ESTs
Non-Human ESTs from GenBank
Mouse (Feb 2006/mm8) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Repeating Elements by RepeatMasker
RepeatMasker

Figure 4. UCSC genome browser showing feature 1 ('310_aa') and human RBMX2 ('NP057108.2) aligned to human chromosome 5. MouseNet tracks are shown at bottom.



```
00000841  G  R  S  E  G  R  S  Y  R  S  R  S  R  S  R  D  K  S  H  R  00000900
<<<<<<<<  W  |  |  |  |  |  |  H  |  T  |  |  |  |  |  |  |  |  |  |  <<<<<<<<
79663273 tggcgttccgaagggcgtagtcacagaactcgaagtaggagccgagataaatcccatagg 79663214

00000901  H  K  R  A  R  R  S  R  E  R  E  S  S  N  P  S  D  R  W  R  00000960
<<<<<<<<  |  |  K  |  |  |  |  W  |  X  |  |  |  |  |  |  |  |  R  |  <<<<<<<<
79663213 cataaaaaagcccgtcgctcctgggagtgagagtcttcgaatcccagcgaccgtaggcgt 79663154
```

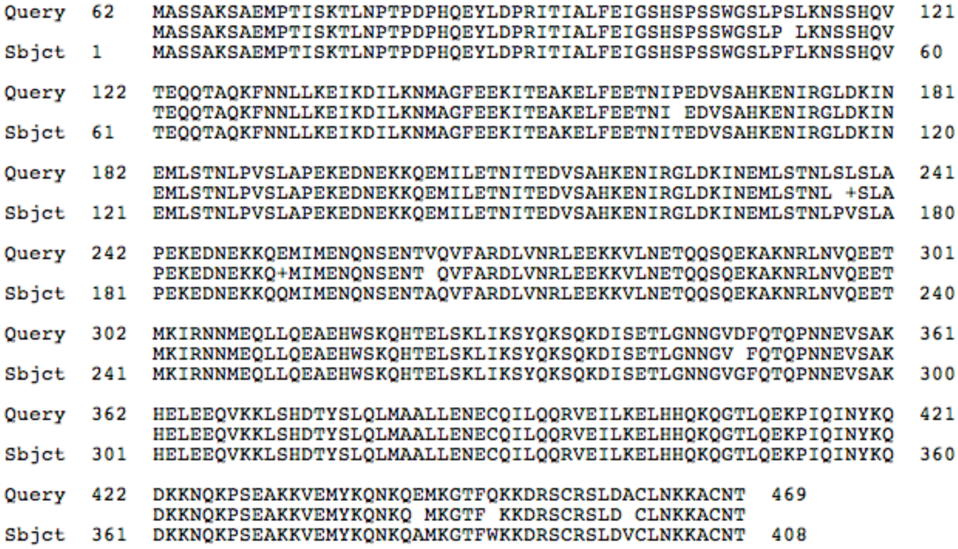
Figure 5. Side-by-Side alignment of Chimp Chunk 2_7 (top) to human chromosome 5 (bottom).

Several lines of evidence suggest that this feature is a pseudogene. The human genomic DNA most similar to feature 1 contains a stop codon, which would produce a truncated protein if translated. Feature 1 maps to chromosome 5, though it shows homology to RBMX2, a gene that is present on the X chromosome. The sequence identity within the aligned region is below the expected identity between orthologous human and chimp protein sequences (>99%). Finally, the presence of only one exon supports the theory of a pseudogene resulting from a retrotransposition event.

To investigate the age of this pseudogene, we looked at the MouseNet track in the UCSC genome browser (Figure 4, bottom). MouseNet showed that the pseudogene region of the human genome best corresponds to chromosome 2 in the mouse genome. According to MouseNet, if there is an orthologous pseudogene in mouse, it should occur on chromosome 2. We then performed a BLAT search of human RBMX2 against the mouse genome. The match of RBMX2 to mouse chromosome 2 is in the region MouseNet listed, and contains only one exon. This suggests that this is indeed the orthologous pseudogene in mouse, so the retrotransposition event that created the pseudogene must have occurred prior to the divergence of primate and rodent lineages. The side-by-side alignment of human RBMX2 to the mouse pseudogene does not show a stop codon, suggesting that the stop codon was introduced into the chimp pseudogene ortholog after the primate and rodent lineages split.

Feature 2

A blastp search using feature 2's Genscan predicted peptide sequence was performed with the nr database. BLAST showed an alignment to a curated entry, human spermatogenic leucine zipper 1 (SPZ1). SPZ1 is a transcription factor that binds to specific DNA sequences known as the E-box and G-box motifs. This protein is highly expressed in adult testis and is suspected to play a role in cell proliferation and differentiation during spermatogenesis. The Entrez Gene entry for SPZ1 showed that it contains only one exon. Genscan, however, predicted two exons for this feature (Figure 2). The BLAST alignment of chimp sequence to SPZ1 was very good, with 97% identity over 408 residues (Figure 6). However, this is less similarity than is expected to occur between chimp and human orthologs.



```
Query  62   MASSAKSAEMPTISKTLNPTPDPHQEYLDPRITIALFEIGSHSPSSWGSLPSLKNSSHQV  121
            MASSAKSAEMPTISKTLNPTPDPHQEYLDPRITIALFEIGSHSPSSWGSLP LKNSSHQV
Sbjct  1    MASSAKSAEMPTISKTLNPTPDPHQEYLDPRITIALFEIGSHSPSSWGSLPFLKNSSHQV  60

Query  122  TEQQTAQKFNNLLKEIKDILKNMAGFEEKITEAKELFEETNIPEDVSAHKENIRGLDKIN  181
            TEQQTAQKFNNLLKEIKDILKNMAGFEEKITEAKELFEETNI EDVSAHKENIRGLDKIN
Sbjct  61   TEQQTAQKFNNLLKEIKDILKNMAGFEEKITEAKELFEETNITEDVSAHKENIRGLDKIN  120

Query  182  EMLSTNLPVSLAPEKEDNEKKQEMILETNITEDVSAHKENIRGLDKINEMLSTNLSLSLA  241
            EMLSTNLPVSLAPEKEDNEKKQEMILETNITEDVSAHKENIRGLDKINEMLSTNL +SLA
Sbjct  121  EMLSTNLPVSLAPEKEDNEKKQEMILETNITEDVSAHKENIRGLDKINEMLSTNLPVSLA  180

Query  242  PEKEDNEKKQEMIMENQNSENTVQVFARDLVNRLEEKKVLNETQQSQEKAKNRLNVQEET  301
            PEKEDNEKKQ+MIMENQNSENT QVFARDLVNRLEEKKVLNETQQSQEKAKNRLNVQEET
Sbjct  181  PEKEDNEKKQQMIMENQNSENTAQVFARDLVNRLEEKKVLNETQQSQEKAKNRLNVQEET  240

Query  302  MKIRNNMEQLLQEAEHWSKQHTELSKLIKSYQKSQKDISETLGNNGVDFQTQPNNEVSAK  361
            MKIRNNMEQLLQEAEHWSKQHTELSKLIKSYQKSQKDISETLGNNGV FQTQPNNEVSAK
Sbjct  241  MKIRNNMEQLLQEAEHWSKQHTELSKLIKSYQKSQKDISETLGNNGVGFQTQPNNEVSAK  300

Query  362  HELEEQVKKLSHDTYSLQLMAALLENECQILQQRVEILKELHHQKQGTLQEKPIQINYKQ  421
            HELEEQVKKLSHDTYSLQLMAALLENECQILQQRVEILKELHHQKQGTLQEKPIQINYKQ
Sbjct  301  HELEEQVKKLSHDTYSLQLMAALLENECQILQQRVEILKELHHQKQGTLQEKPIQINYKQ  360

Query  422  DKKNQKPSEAKKVEMYKQNKQEMKGTFQKKDRSCRSLDACLNKKACNT  469
            DKKNQKPSEAKKVEMYKQNKQ MKGTF KKDRSCRSLD CLNKKACNT
Sbjct  361  DKKNQKPSEAKKVEMYKQNKQAMKGTFWKKDRSCRSLDVCLNKKACNT  408
```

Figure 6. Blastp alignment of feature 2 (Query) to SPZ1 (Sbjct). The alignment begins with the first amino acid of Genscan feature 2, exon 2 (residue 62 of feature 2) and ends prematurely at residue 408 of the human protein (SPZ1 has 430 amino acids).

A BLAT search of feature 2 against the human genome showed a 96.4% identity match to chromosome 5. This is consistent with feature 1 analysis, as it indicates synteny with feature 1 on chromosome 5, as expected. The BLAT showed that the second exon Genscan predicted aligned to a Known Gene – SPZ1 (Figure 7). Alignment of feature 2 to the location of the Known Gene in the human genome suggests that this feature may be an ortholog of the human gene. However, only the second exon of the predicted feature aligned to SPZ1, beginning at the 62nd amino acid of feature 2 (Figure 6). The first predicted exon is 17 kb away from the second exon. We isolated the second exon sequence and performed a blastx search against the nr database. The best curated alignment was identical to the alignment in Figure 6. Thus, the first predicted exon does not contribute to the similarity of feature 2 to SPZ1. A blastx search of exon 1 against the nr database yielded an alignment to human beta actin, of which feature 3 is a pseudogene. The alignment to beta actin was in the opposite orientation relative to exon 2 in feature 2 (alignment of exon 1 to beta actin was 5' to 3' in the Chimp Chunk, while exon 1 as predicted in feature 2 was 3' to 5'). The 5' to 3' orientation of exon 1's alignment to beta actin is consistent with the 5' to 3' orientation of feature 3's alignment to beta actin. Eliminating the first 14 amino acids from exon 1 of feature 2 yields a short peptide that aligns to the beginning of the beta actin sequence (Figure 10). Extracting sequence from the middle of feature 2's first exon through the end of feature 3 shows that this sequence aligns over the full length of human beta actin. This will be further analyzed in discussion of feature 3. For the analysis of feature 2, however, this provides strong evidence that exon 1 of feature 2 is in fact a part of feature 3. Exon 2 of feature 2 aligns well to the majority of human SPZ1, but appears to be missing the last 21 residues of the human protein (see Figure 6).

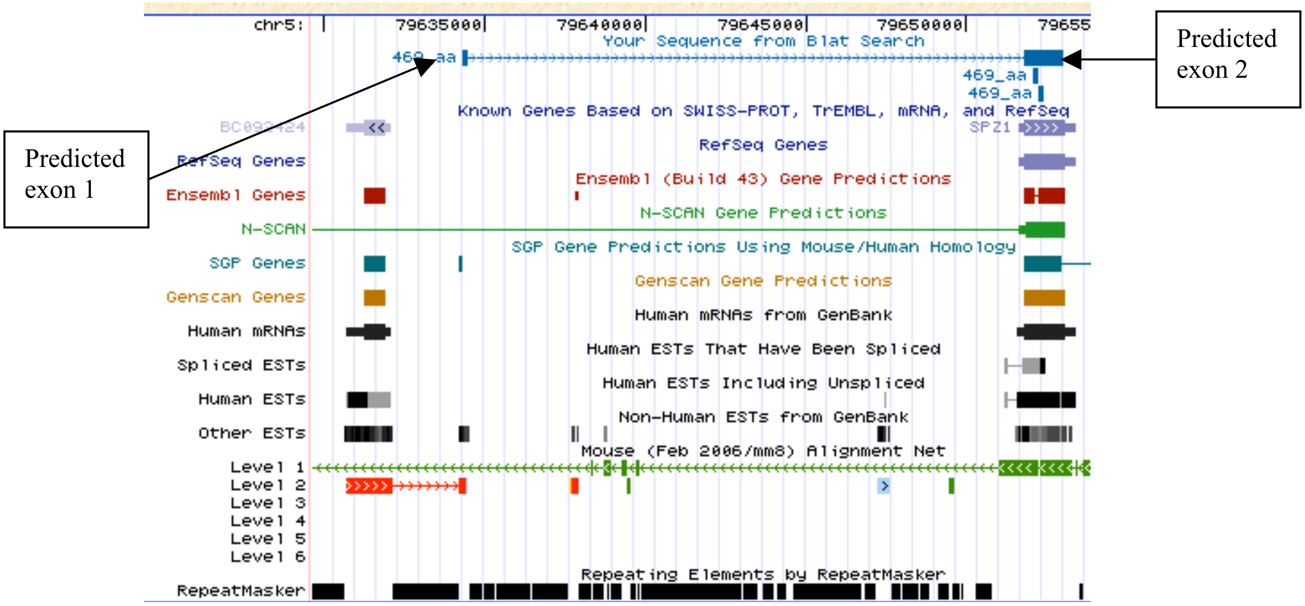
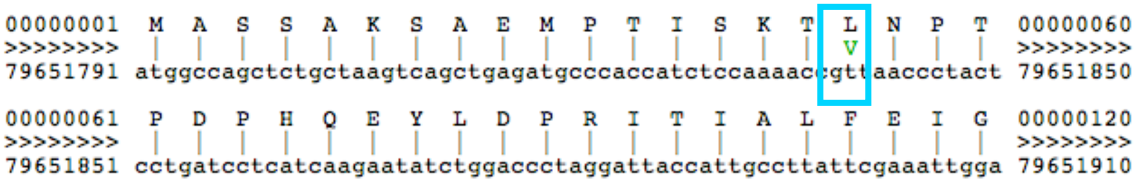


Figure 7. BLAT showing feature 2 ('469aa') aligned to human chromosome 5.

Using the human SPZ1 protein sequence, we performed a BLAT search against the human genome. The best match of the human protein mapped to the same location as the best match of exon 2 from our initial BLAT search (Figure 7). Interestingly, the human protein sequence we obtained from Refseq had a one amino acid discrepant with the human genome sequence – a leucine/ valine difference (Figure 8). This could result from conflicting protein sequences in different databases – it suggests a polymorphism present in the human sequence.



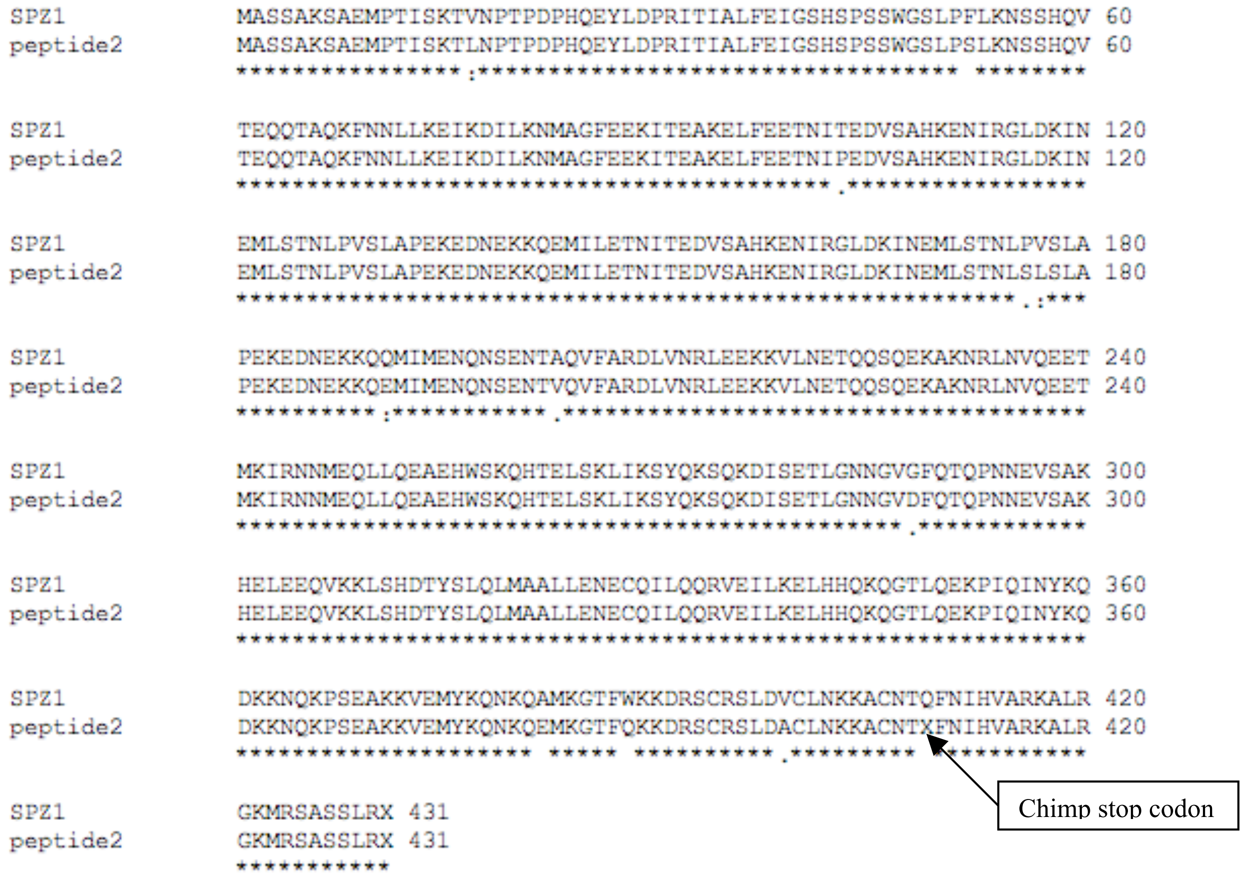
```
00000001 M  A  S  S  A  K  S  A  E  M  P  T  I  S  K  T  L  N  P  T  00000060
>>>>>>>> |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  V  |  |  |  >>>>>>>>
79651791 atggccagctctgctaagtcagctgagatgcccaccatctccaaaaccgttaaccctact 79651850

00000061 P  D  P  H  Q  E  Y  L  D  P  R  I  T  I  A  L  F  E  I  G  00000120
>>>>>>>> |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  >>>>>>>>
79651851 cctgatcctcatcaagaatatctggaccctaggattaccattgccttattcgaaattgga 79651910
```

Figure 8. Side-by-Side alignment of human SPZ1 (top) to human chromosome 5 (bottom) with 99.8% identity. Note the L/V discrepancy.

Alignment of the functional human protein and chimp peptide sequence to the same location in the genome further indicates that feature 2 may be an ortholog of human SPZ1. To explain the 21 missing residues of chimp sequence with respect to human sequence, we compared the BLAT search output from both sequences against the human genome to the BLAT search output from both sequences against the chimp genome. Comparison of alignment to each genome shows that the 21 residues in genomic chimp sequence after its (premature) stop codon are the same as the genomic human sequence before its stop codon. This can also be shown in a ClustalW alignment of the human sequence to the chimp genomic sequence, including the 21 residues after the premature stop codon (Figure 10).

Though the similarity of the chimp and human sequences is not as high as human/chimp ortholog similarity is expected to be, there is evidence that this feature is the chimp ortholog of human SPZ1 nonetheless. The strongest support for the ortholog hypothesis is that the BLAT tool shows the best match for each sequence maps to an identical region in both the human and chimp genomes; feature 2 is the best match in the chimp genome to human SPZ1. This evidence may indicate that these are orthologous genes, or it could suggest that the gene has lost functionality in chimps, releasing the sequence from negative selection and resulting in divergence from the functional human sequence. It seems unlikely that a protein that may be required for spermatogenesis could lose functionality without negatively impacting reproductive fitness, so we conclude that feature 2 is the chimp ortholog of human SPZ1. If this is true, the gene happens to have an unusually high mutation rate (to accrue more than 1% difference from human), and the protein must maintain function without the last 21 residues. It is also possible that a sequencing error caused the presence of the stop codon in the chimp sequence, since this is a one bp difference (TAG in chimp versus CAG in human), and the chimp genome sequence is only draft quality. In this case, the chimp protein would retain the final 21 amino acids.



```
SPZ1            MASSAKSAEMPTISKTVNPTPDPHQEYLDPRITIALFEIGSHSPSSWGSLPFLKNSSHQV 60
peptide2        MASSAKSAEMPTISKTLNPTPDPHQEYLDPRITIALFEIGSHSPSSWGSLPSLKNSSHQV 60
                ****************:********************************** ********

SPZ1            TEQQTAQKFNNLLKEIKDILKNMAGFEEKITEAKELFEETNITEDVSAHKENIRGLDKIN 120
peptide2        TEQQTAQKFNNLLKEIKDILKNMAGFEEKITEAKELFEETNIPEDVSAHKENIRGLDKIN 120
                ******************************************.*****************

SPZ1            EMLSTNLPVSLAPEKEDNEKKQEMILETNITEDVSAHKENIRGLDKINEMLSTNLPVSLA 180
peptide2        EMLSTNLPVSLAPEKEDNEKKQEMILETNITEDVSAHKENIRGLDKINEMLSTNLSLSLA 180
                *******************************************************.:***

SPZ1            PEKEDNEKKQQMIMENQNSENTAQVFARDLVNRLEEKKVLNETQQSQEKAKNRLNVQEET 240
peptide2        PEKEDNEKKQEMIMENQNSENTVQVFARDLVNRLEEKKVLNETQQSQEKAKNRLNVQEET 240
                **********:***********.*************************************

SPZ1            MKIRNNMEQLLQEAEHWSKQHTELSKLIKSYQKSQKDISETLGNNGVGFQTQPNNEVSAK 300
peptide2        MKIRNNMEQLLQEAEHWSKQHTELSKLIKSYQKSQKDISETLGNNGVDFQTQPNNEVSAK 300
                ***********************************************.************

SPZ1            HELEEQVKKLSHDTYSLQLMAALLENECQILQQRVEILKELHHQKQGTLQEKPIQINYKQ 360
peptide2        HELEEQVKKLSHDTYSLQLMAALLENECQILQQRVEILKELHHQKQGTLQEKPIQINYKQ 360
                ************************************************************

SPZ1            DKKNQKPSEAKKVEMYKQNKQAMKGTFWKKDRSCRSLDVCLNKKACNTQFNIHVARKALR 420
peptide2        DKKNQKPSEAKKVEMYKQNKQEMKGTFQKKDRSCRSLDACLNKKACNTXFNIHVARKALR 420
                ********************* ***** **********.********* ***********

SPZ1            GKMRSASSLRX 431
peptide2        GKMRSASSLRX 431
                ***********
```

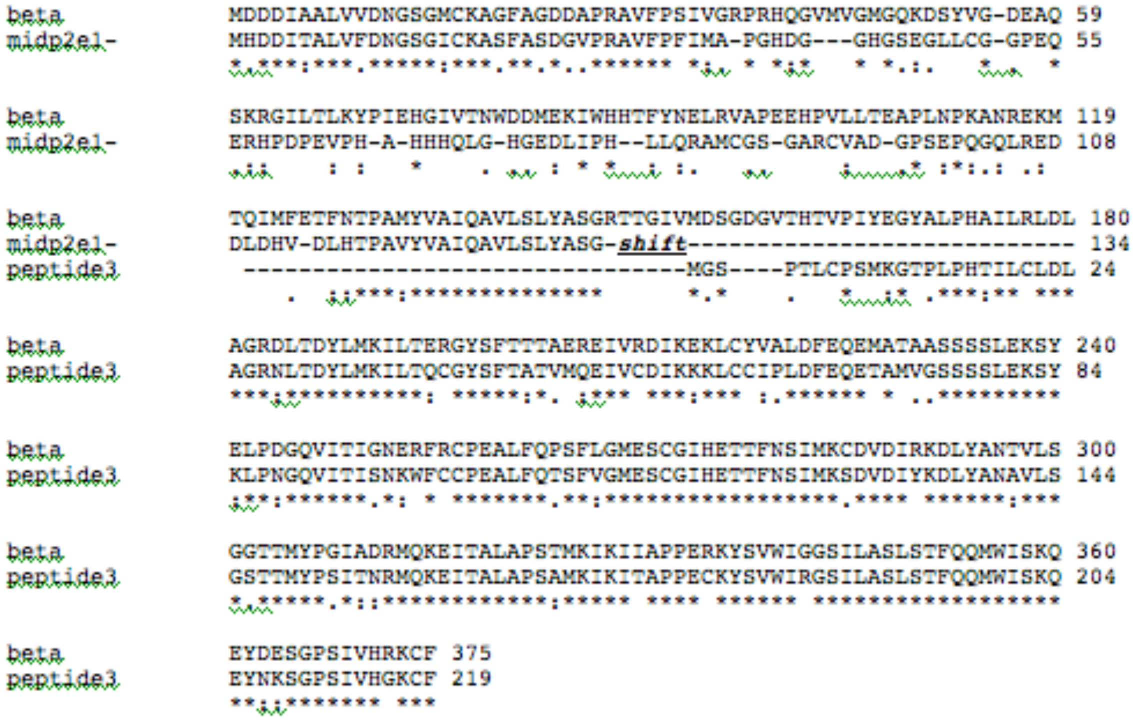
Chimp stop codon

Figure 9. ClustalW alignment of human SPZ1 and chimp feature 2.

Feature 3

Genscan predicted one exon for feature 3. A blastp search of the peptide sequence against the nr database aligned to a curated human beta actin (ACTB) sequence. There are six different genes for different actin isoforms, so we must be aware of the possibility that this feature matches a different actin isoform. However, because the best matches were all beta actin, we can be fairly confident that this is the correct isoform. Actins are highly conserved proteins involved in cell structure and cell motility. The Entrez Gene entry for ACTB showed 5 exons, while Genscan predicted only one exon.

As discovered in analysis of feature 2, the complete Chimp Chunk sequence corresponding to ACTB extends from the first exon of Genscan feature 2 ("exon 2.1") to the end of feature 3. In the region between exon 2.1 and feature 3, there has been an insertion of 10 bp relative to the human sequence, causing a shift in frame (indicated in Figure 10). This frameshift ensures that feature 3 is not an ortholog of human ACTB, since translation of the resultant mRNA would not produce a protein similar to ACTB.



```
beta            MDDDIAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQKDSYVG-DEAQ 59
midp2e1-        MHDDITALVFDNGSGICKASFASDGVPRAVFPFIMA-PGHDG---GHGSEGLLCG-GPEQ 55
                *.***:***.*****:***.**.*..****** *:. * *:*   * *.:.   * .  *

beta            SKRGILTLKYPIEHGIVTNWDDMEKIWHHTFYNELRVAPEEHPVLLTEAPLNPKANREKM 119
midp2e1-        ERHPDPEVPH-A-HHHQLG-HGEDLIPH--LLQRAMCGS-GARCVAD-GPSEPQGQLRED 108
                .:.    : :   *    .  .. : * * : :.   ..     :    * :*:.: .: 

beta            TQIMFETFNTPAMYVAIQAVLSLYASGRTTGIVMDSGDGVTHTVPIYEGYALPHAILRLDL 180
midp2e1-        DLDHV-DLHTPAVYVAIQAVLSLYASG-shift---------------------------- 134
peptide3         --------------------------------MGS----PTLCPSMKGTPLPHTILCLDL 24
                    .  ::***:**************      *.*    .   *  :* .***:** ***

beta            AGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALDFEQEMATAASSSSLEKSY 240
peptide3        AGRNLTDYLMKILTQCGYSFTATVMQEIVCDIKKKLCCIPLDFEQETAMVGSSSSLEKSY 84
                ***:**********: *****:*. :*** ***:*** :.****** * ..*********

beta            ELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSIMKCDVDIRKDLYANTVLS 300
peptide3        KLPNGQVITISNKWFCCPEALFQTSFVGMESCGIHETTFNSIMKSDVDIYKDLYANAVLS 144
                :**:******.*: * *******.**:*****************.**** ******:***

beta            GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLSTFQQMWISKQ 360
peptide3        GSTTMYPSITNRMQKEITALAPSAMKIKITAPPECKYSVWIRGSILASLSTFQQMWISKQ 204
                *.*****.*::************:***** **** ****** ******************

beta            EYDESGPSIVHRKCF 375
peptide3        EYNKSGPSIVHGKCF 219
                **::******* ***
```

Figure 10. ClustalW alignment of human beta actin ('beta'), Genscan feature 3 (36374-37033; 'peptide3'), and Genscan feature 2 exon 1 (35922-37003; 'midp2e1-'). "***shift***" in middle of figure indicates 28 bp that are not represented from chimp sequence that would cause a frameshift (insertion of 10 bp versus human sequence). Sequence resulting after the frameshift is not shown to emphasize that this region exhibits similarity to actin beta.

We used the BLAT tool to compare the nucleotide sequences of feature 3 and ACTB to the human genome. The best human ACTB match showed 100% identity to chromosome 7, where the functional protein-coding sequence is located. The best alignment of feature three showed 98.0% similarity to chromosome 5, maintaining synteny with features 1 and 2 (Figure 11).

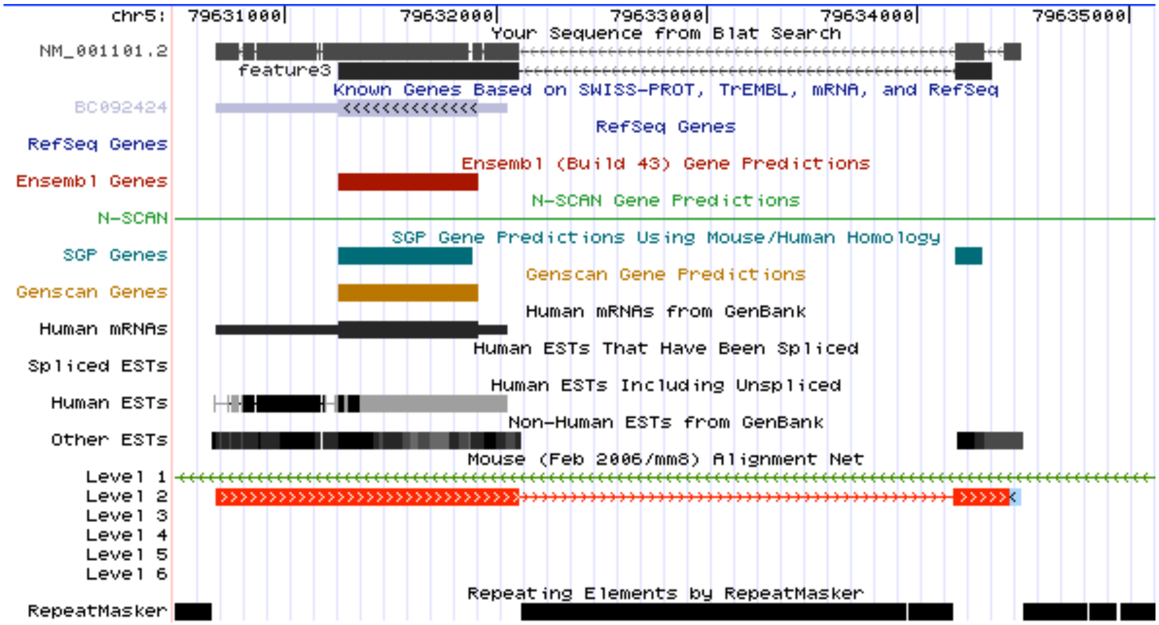
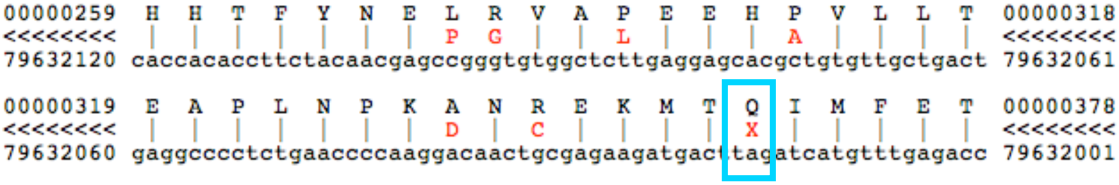

chr5:
79631000
79632000
79633000
79634000
79635000
Your Sequence from Blat Search
NM_001101.2
feature3
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
BC092424
RefSeq Genes
RefSeq Genes
Ensembl (Build 43) Gene Predictions
Ensembl Genes
N-SCAN Gene Predictions
N-SCAN
SGP Gene Predictions Using Mouse/Human Homology
SGP Genes
Genscan Gene Predictions
Genscan Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Human ESTs Including Unspliced
Human ESTs
Non-Human ESTs from GenBank
Other ESTs
Mouse (Feb 2006/mm8) Alignment Net
Level 1
Level 2
Level 3
Level 4
Level 5
Level 6
Repeating Elements by RepeatMasker
RepeatMasker

Figure 11. UCSC genome browser showing alignment of feature 3 and human ACTB ('NM_001101.2') to human chromosome 5.[5]

The best alignment of human ACTB to the human genome showed five exons, supporting the Entrez Gene data. The BLAT tool was then used to compare human ACTB peptide sequence to the region of chromosome 5 homologous to feature 3 (Figure 12). A side-by-side alignment shows the presence of a stop codon in the genomic DNA.



```
00000259  H  H  T  F  Y  N  E  L  R  V  A  P  E  E  H  P  V  L  L  T  00000318
<<<<<<<<  |  |  |  |  |  |  |  P  G  |  |  L  |  |  |  A  |  |  |  |  <<<<<<<<
79632120 caccacaccttctacaacgagccgggtgtggctcttgaggagcacgctgtgttgctgact 79632061

00000319  E  A  P  L  N  P  K  A  N  R  E  K  M  T  Q  I  M  F  E  T  00000378
<<<<<<<<  |  |  |  |  |  |  |  D  |  C  |  |  |  |  X  |  |  |  |  |  <<<<<<<<
79632060 gaggcccctctgaaccccaaggacaactgcgagaagatgacttagatcatgtttgagacc 79632001
```

Figure 12. Side-by-Side alignment of human ACTB to human chromosome 5.

There is now overwhelming evidence that feature 3 is a pseudogene of ACTB. The sequence contains an insertion that causes a frameshift, which would change the translated amino acid sequence. Feature 3 does not possess as high similarity to human ACTB as orthologs usually do, nor does it map to the same chromosome as functional ACTB. This feature does not contain 5 exons as the human gene does. Also, the genomic sequence that feature 3 maps to contains a stop codon in the ORF that aligns with ACTB. We can confidently conclude from these data that feature 3 is an ACTB pseudogene. The pseudogene likely resulted from a retrotransposition event, as indicated by lack of intronic DNA.

To investigate pseudogene age, MouseNet (seen in Figure 13) was used again. The best match of MouseNet to this region of human chromosome 5 was in mouse chromosome 5, indicating that the orthologous mouse pseudogene should be present in

[5] Paradoxically, the "Known Gene" shown in this browser is a protein listed in Swissprot as "hypothetical protein."

mouse chromosome 5. However, human ACTB protein sequence matches this region of the mouse genome with 100% identity and contains 5 exons, indicating that this is probably the mouse ortholog of ACTB. It is possible that no corresponding pseudogene exists in mice, so the best match for both the pseudogene and human ACTB is the functional mouse gene. It is also possible that MouseNet misguided our search – the match that guided us to chromosome 5 was "nonSyn" level 2, meaning that the alignment shown maps to a different chromosome than the gap in level 1. A level 1 match in MouseNet is the best, longest match of the mouse genome to the region of interest in the current genome. Thus, in the mouse genome the best alignment to the region of human chromosome 5 around feature 3 is not syntenic with the best alignment to that particular part of chromosome 5 containing the human ortholog of feature 3. Perhaps the region of the mouse genome containing the ortholog to feature 3 is not on the chromosome indicated in the level 2 alignment. There are other sequences similar to ACTB that appear to be pseudogenes simply because of their single exon character, but they are not in the region listed by MouseNet as orthologous to our pseudogene's location. It is also difficult to determine if there is an orthologous mouse pseudogene because of the huge number of BLAT hits of this peptide (total of 39 hits). Given that this phylogenetic comparison has not been informative, all we can say is that feature 3 is probably an old pseudogene on the basis of the many mutations the pseudogene possesses relative to human ACTB.

Feature 4

An initial blastp search of Genscan's feature 4 against the nr database yielded an alignment to the curated human keratin 18 (KRT18) sequence. KRT18 is a type I intermediate filament that is expressed in single layer epithelial tissues. It is one of the most prevalent members of the intermediate filament gene family. Again, because KRT18 is a member of a protein family, we must be careful not to associate our feature with the wrong group member. Since the best matches from the blastp search align specifically to KRT18, there is little chance that this feature is an artifact of a different isoform. The Entrez Gene entry shows that KRT18 has alternative splicing sites and 7 exons.

A BLAT search of feature 4 against human genomic sequence produced a match to chromosome 5. All Genscan predicted features are syntenic, aligning to human chromosome 5 in the same order and orientation as in the Chimp Chunk. The match of peptide 4 from the BLAT search showed that two sets of exons seemed to match different regions of the chromosome (Figure 13).

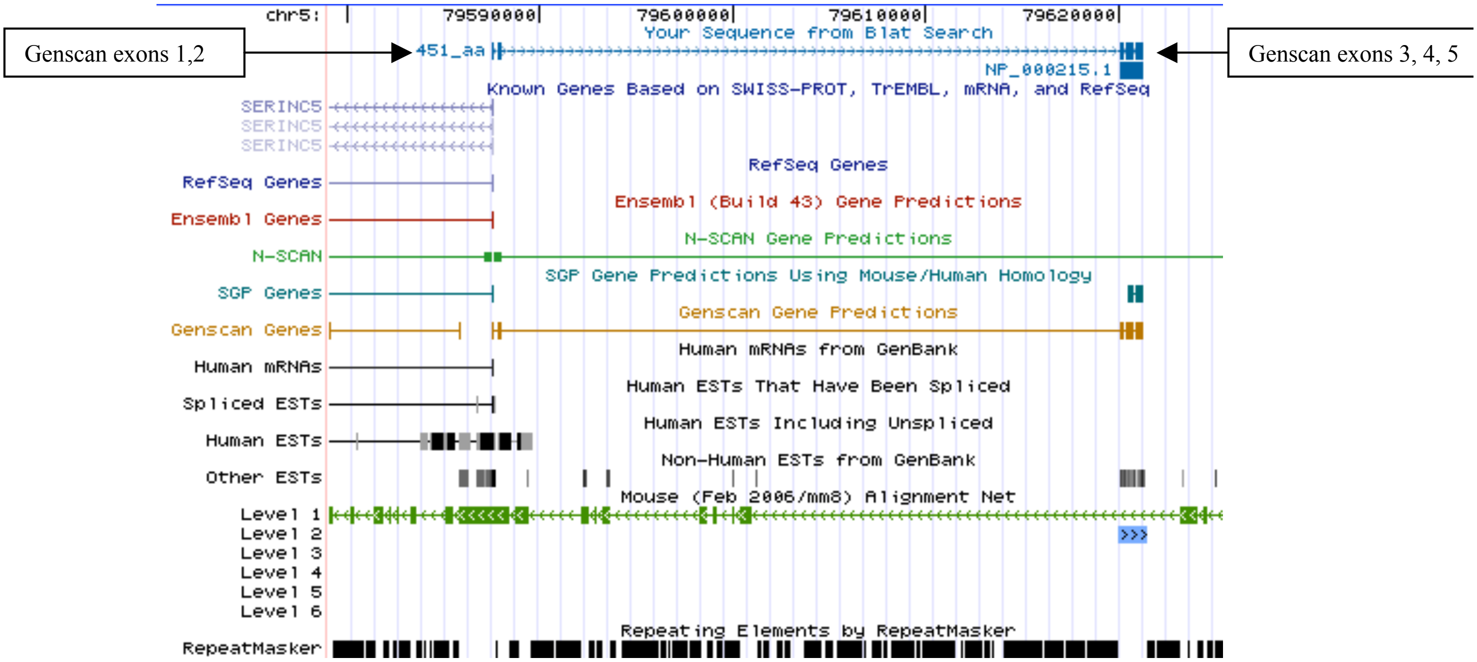


Figure 13. UCSC genome browser showing alignment of feature 4 ('451_aa') and human KRT18 ('NP_000215.1') to human chromosome 5.

Known Genes, RefSeq Genes, human mRNAs and all human ESTs that matched in the region of exons 1 and 2 did not extend or connect in any way to exons 3, 4, and 5. No other gene prediction programs connected these two sets of exons. Human KRT18 only aligns to exons 3, 4, and 5, adding doubt to the characterization of predicted exons 1 and 2 as part of the same feature as exons 3, 4, and 5. Furthermore, of the BLAST searches we carried out against the Swissprot, Refseq, and EST databases, none showed good evidence of transcription or translation in the region containing exons 1 and 2. There are a few ESTs scattered within some kb of the exons, but they are not consistent or numerous (Figure 17). Genscan data show that there are nearly 36 kb between the two sets of putative exons. The above evidence led us to conclude that exons 1 and 2 were extraneous to feature 4. The region that aligns with exons 1 and 2 is within a few kb of the end of our Chimp Chunk. It is possible that this region should align with the Known Genes, RefSeqs, etc., going in the opposite direction (off the end of our Chimp Chunk) but cannot because the sequence ends. The region containing exons 1 and 2 may be part of the 5' UTR or regulatory elements of Serinc5, as they align to the 5' UTR and to a region further upstream of the gene (Figure 14).

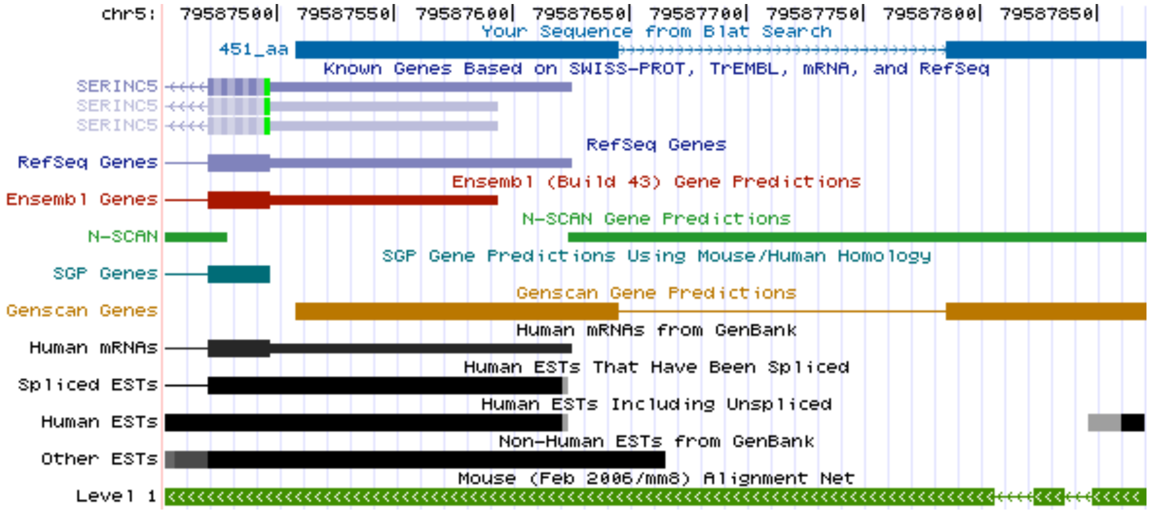
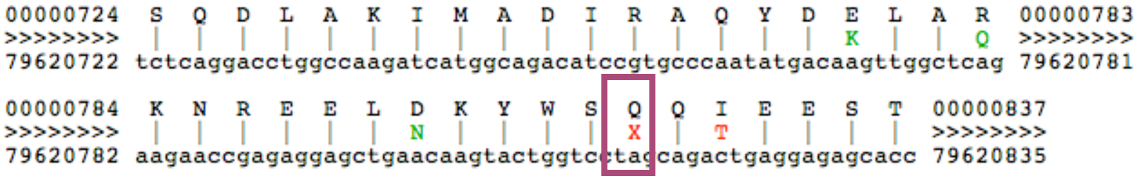

chr5:
79587500|
79587550|
79587600|
79587650|
79587700|
79587750|
79587800|
79587850|
Your Sequence from Blat Search
451_aa
Known Genes Based on SWISS-PROT, TrEMBL, mRNA, and RefSeq
SERINC5
SERINC5
SERINC5
RefSeq Genes
RefSeq Genes
Ensembl (Build 43) Gene Predictions
Ensembl Genes
N-SCAN Gene Predictions
N-SCAN
SGP Gene Predictions Using Mouse/Human Homology
SGP Genes
Genscan Gene Predictions
Genscan Genes
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
Human ESTs Including Unspliced
Human ESTs
Non-Human ESTs from GenBank
Other ESTs
Mouse (Feb 2006/mm8) Alignment Net
Level 1

Figure 14. Alignment of exons 1 and 2 of feature 4 to Serinc5 5' UTR and upstream region.

Human KRT18 aligned perfectly to chromosome 12 and contained 7 exons, confirming Entrez Gene data. The side-by-side alignment of KRT18 to chromosome 5 showed a stop codon in chromosomal DNA (Figure 15).

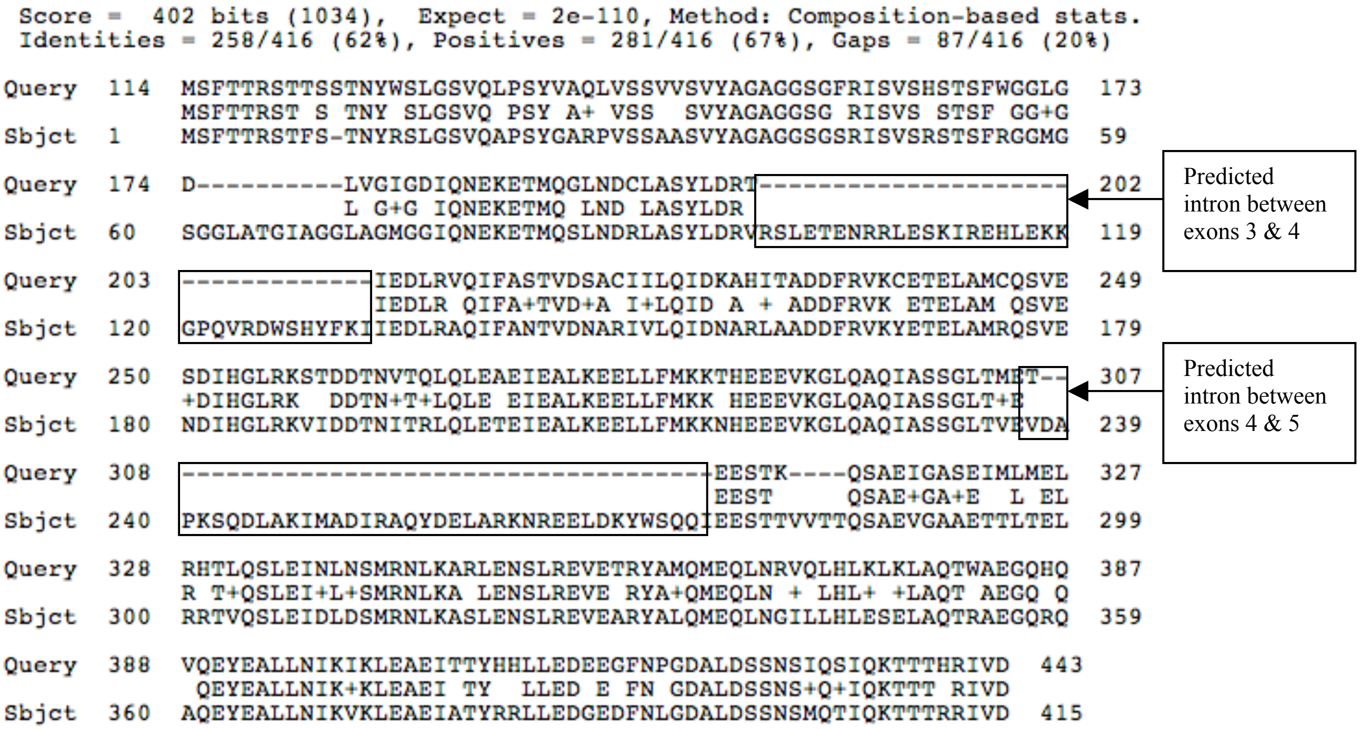


```
00000724  S  Q  D  L  A  K  I  M  A  D  I  R  A  Q  Y  D  E  L  A  R  00000783
>>>>>>>>  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  K  |  |  Q  >>>>>>>>
79620722 tctcaggacctggccaagatcatggcagacatccgtgcccaatatgacaagttggctcag 79620781

00000784  K  N  R  E  E  L  D  K  Y  W  S  Q  Q  I  E  E  S  T  00000837
>>>>>>>>  |  |  |  |  |  |  N  |  |  |  |  X  |  T  |  |  |  |  >>>>>>>>
79620782 aagaaccgagaggagctgaacaagtactggtcctagcagactgaggagagcacc 79620835
```

Figure 15. Side-by-Side alignment of human KRT18 to human chromosome 5.

Feature 4 appears to be a pseudogene of KRT18 based on reduced exon number, different map location than human KRT18, and presence of a stop codon in genomic DNA. But if this pseudogene resulted from a retrotranspositional event, why would it contain multiple exons? Extracting this region and translating it to protein sequence shows that within Genscan's intronic sequences, there are stop codons and frameshifts (resulting from indels). Predicting multiple exons actually facilitated identification of this region as a pseudogene – if Genscan's predicted sequence was shifted or truncated, recognition and classification of this feature would have been much more difficult. The boundaries of the Genscan-predicted exons do not match the exon boundaries of human KRT18 (Figure 16). Also, the introns predicted by Genscan occur at gaps in the alignment of feature 4 to KRT18. Each intron is almost exactly the same size (within one bp) as the corresponding gap in the alignment. This indicates that the predicted intron sequences correspond to sequence in human KRT18. Thus, even though this feature is predicted to have multiple exons, it is still most likely the result of a single retrotransposition event.



```
 Score =  402 bits (1034),  Expect = 2e-110, Method: Composition-based stats.
 Identities = 258/416 (62%), Positives = 281/416 (67%), Gaps = 87/416 (20%)

Query  114  MSFTTRSTTSSTNYWSLGSVQLPSYVAQLVSSVVSVYAGAGGSGFRISVSHSTSFWGGLG  173
            MSFTTRST S TNY SLGSVQ PSY A+ VSS  SVYAGAGGSG RISVS STSF GG+G
Sbjct  1    MSFTTRSTFS-TNYRSLGSVQAPSYGARPVSSAASVYAGAGGSGSRISVSRSTSFRGGMG  59

Query  174  D----------LVGIGDIQNEKETMQGLNDCLASYLDRT---------------------  202
                       L G+G IQNEKETMQ LND LASYLDR
Sbjct  60   SGGLATGIAGGLAGMGGIQNEKETMQSLNDRLASYLDRVRSLETENRRLESKIREHLEKK  119

Query  203  -------------IEDLRVQIFASTVDSACIILQIDKAHITADDFRVKCETELAMCQSVE  249
                         IEDLR QIFA+TVD+A I+LQID A + ADDFRVK ETELAM QSVE
Sbjct  120  GPQVRDWSHYFKIIEDLRAQIFANTVDNARIVLQIDNARLAADDFRVKYETELAMRQSVE  179

Query  250  SDIHGLRKSTDDTNVTQLQLEAEIEALKEELLFMKKTHEEEVKGLQAQIASSGLTMET--  307
            +DIHGLRK  DDTN+T+LQLE EIEALKEELLFMKK HEEEVKGLQAQIASSGLT+E
Sbjct  180  NDIHGLRKVIDDTNITRLQLETEIEALKEELLFMKKNHEEEVKGLQAQIASSGLTVEVDA  239

Query  308  ------------------------------------EESTK----QSAEIGASEIMLMEL  327
                                                EEST     QSAE+GA+E  L EL
Sbjct  240  PKSQDLAKIMADIRAQYDELARKNREELDKYWSQQIEESTTVVTTQSAEVGAAETTLTEL  299

Query  328  RHTLQSLEINLNSMRNLKARLENSLREVETRYAMQMEQLNRVQLHLKLKLAQTWAEGQHQ  387
            R T+QSLEI+L+SMRNLKA LENSLREVE RYA+QMEQLN + LHL+ +LAQT AEGQ Q
Sbjct  300  RRTVQSLEIDLDSMRNLKASLENSLREVEARYALQMEQLNGILLHLESELAQTRAEGQRQ  359

Query  388  VQEYEALLNIKIKLEAEITTYHHLLEDEEGFNPGDALDSSNSIQSIQKTTTHRIVD  443
             QEYEALLNIK+KLEAEI TY  LLED E FN GDALDSSNS+Q+IQKTTT RIVD
Sbjct  360  AQEYEALLNIKVKLEAEIATYRRLLEDGEDFNLGDALDSSNSMQTIQKTTTRRIVD  415
```

Predicted intron between exons 3 & 4

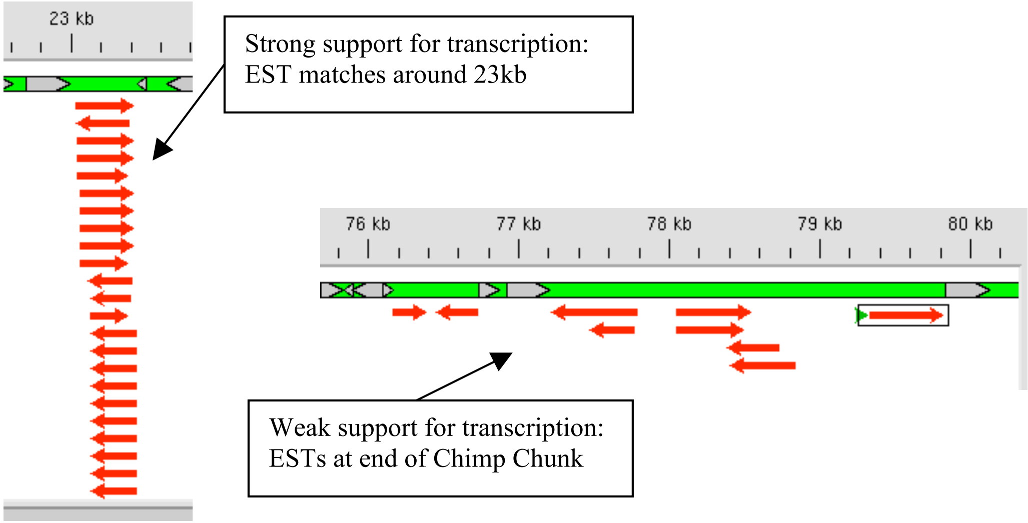
Predicted intron between exons 4 & 5

Figure 16. Blastp alignment of feature 4 to human KRT18. Residue 114 of query corresponds to the start of exon 3 of feature 4. The gaps indicated occur where Genscan predicted introns.

MouseNet maps this region of the human genome to chromosome 15. A BLAT search against the mouse genome shows that the most similar match to human KRT18 is in that same region, and contains multiple exons. This indicates that the mouse ortholog of human KRT18, not a KRT18 pseudogene, is on chromosome 15. There are other matches to the mouse genome likely to be psuedogenes because they contain only one exon and contain stop codons in side-by-side alignments to KRT18. However, as with feature 3, it is difficult to determine if one of these is the definite ortholog of feature 4. (This was a Level 2 nonSyn match, so MouseNet may have misguided us, as discussed in feature 3 analysis.) From the presence of multiple mutations, frameshifts, and stop codons, it seems likely that this pseudogene was retrotransposed into the genome fairly long ago.

Anonymous ESTs

There are two regions of anonymous ESTs matching around bp 23214-23414 and 76000-80000 bp (Figure 17). The latter region contains a total of 9 ESTs, with no more than two ESTs in one region. The small amount of EST data provides very weak evidence for transcription in this region. This region also happens to overlap with the exons 1 and 2 of Genscan's predicted feature 4. This minimal evidence of transcription may be a sign of a regulatory element or UTR for a gene off the end of the Chimp Chunk (i.e. Serinc5 from Figure 14).

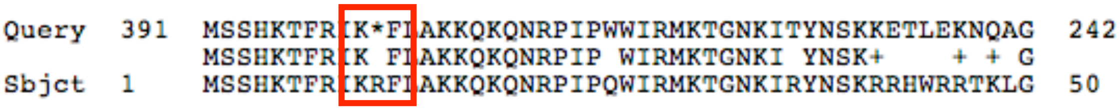


Strong support for transcription: EST matches around 23kb

Weak support for transcription: ESTs at end of Chimp Chunk

Figure 16. Herne images of anonymous EST data from blastn.

The region around 23 kb contains many ESTs, providing support for transcription in this region. We extracted chimp sequence from 23000-24000 kb. A blastx search of this 1 kb chimp sequence against the nr database showed a curated match to *Homo sapiens* 60S ribosomal protein L39 (RPL39). RPL39 is a 3-exon gene in humans. However, the chimp ESTs covered a continuous region, indicating that there are no intronic regions. In addition, the chimp sequence contained a stop codon in its alignment to RPL39 (see Figure 17). The presence of this stop codon and the single exon character of the chimp sequence strongly suggest that this region contains a pseudogene of RPL39. A BLAT search of the human RPL39 and chimp sequences reveal that while RPL39 maps to the X chromosome, the best alignment of the chimp sequence is on chromosome 5, again evoking synteny of the Chimp Chunk to the human genome. Side-by-side alignment of RPL39 to this region of chromosome 5 shows the same result as Figure 17 – there is a stop codon in the human genomic DNA. This is not surprising considering that there are multiple pseudogenes for most ribosomal proteins throughout the human genome, as well. Genscan probably did not detect this pseudogene as a putative protein-coding domain because a stop codon occurs 11 residues into the sequence, ending the ORF.



```
Query  391    MSSHKTFRIK*FLAKKQKQNRPIPWWIRMKTGNKITYNSKKETLEKNQAG  242
              MSSHKTFRIK FLAKKQKQNRPIP WIRMKTGNKI YNSK+    + + G
Sbjct  1      MSSHKTFRIKRFLAKKQKQNRPIPQWIRMKTGNKIRYNSKRRHWRRTKLG  50
```

Figure 17. Blastx alignment of human RPL39 (subjct) to ChimpChunk (query).

MouseNet aligns this region of human chromosome 5 to mouse chromosome 16. A BLAT comparison of human RPL39 to the mouse genome shows a match to mouse chromosome 16 that has a single exon, indicating that there is an orthologous mouse RPL39 pseudogene. Thus, the retrotransposition that created the pseudogene occurred before the separation of primate and rodent lineages. However, there is no stop codon present in the mouse chromosomal DNA. Thus, the stop codon in human DNA was introduced after the evolutionary divergence of primates and rodents.

Conclusion

Through using multiple bioinformatics tools and a little prior knowledge of annotation methodology, we were able to identify 4 pseudogenes and 1 functional gene in Chimp Chunk 2_7. All features exhibited synteny with human genomic sequence on chromosome 5.

Final Annotation Map

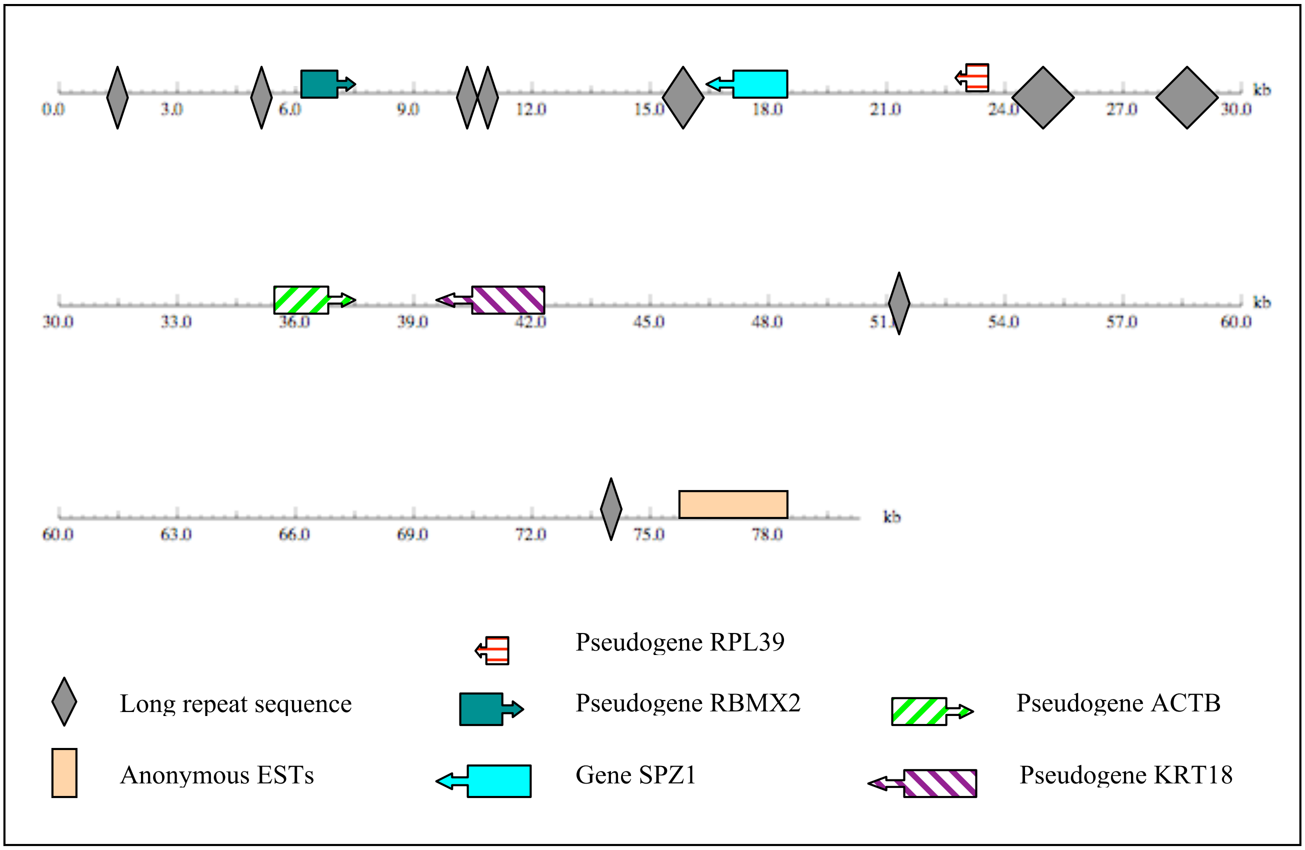


Figure 18. Map of identified features.