Annotating Chimpanzee Chunk 03-11

Jimmy Ma
Bio 434W
Professor Elgin
April 15, 2010

Introduction

Chimpanzees and humans are very closely related, sharing about 98% identity between the two genomes. Because the chimp genome is poorly mapped and is so similar to its human counterpart, we use the human genome as a guide to annotate that of the chimp. The goal of annotation is to decode an organism's genome by identifying its constituent features and functions such as genes, repetitious areas, and pseudogenes. Here, we attempt to annotate chunk 03-11, an approximately 100 kB fragment of chimp chromosome 21, using a variety of computational tools including GENSCAN, BLAST, BLAT, and the UCSC Genome Browser.

Initial GENSCAN Results

After using RepeatMasker to hide all known repetitious regions in the DNA sequence, GENSCAN was used to predict putative coding regions in chunk 03-11. GENSCAN initially predicted three putative coding regions—A, B, and C—that were divided among the three people in the group (Figure 1). From our initial examination, feature A may be a gene with six exons. Feature B runs in reverse orientation compared to fragments A and C and may be a gene with three exons. The GENSCAN results also suggest that feature C, a possible gene with 20 predicted exons, lacks an initial exon. This may be due to the starting exon being outside of the 100 kB region or an incorrect prediction by GENSCAN. None of these genes had only a single exon, which makes it difficult to call a region a possible pseudogene upon first inspection.

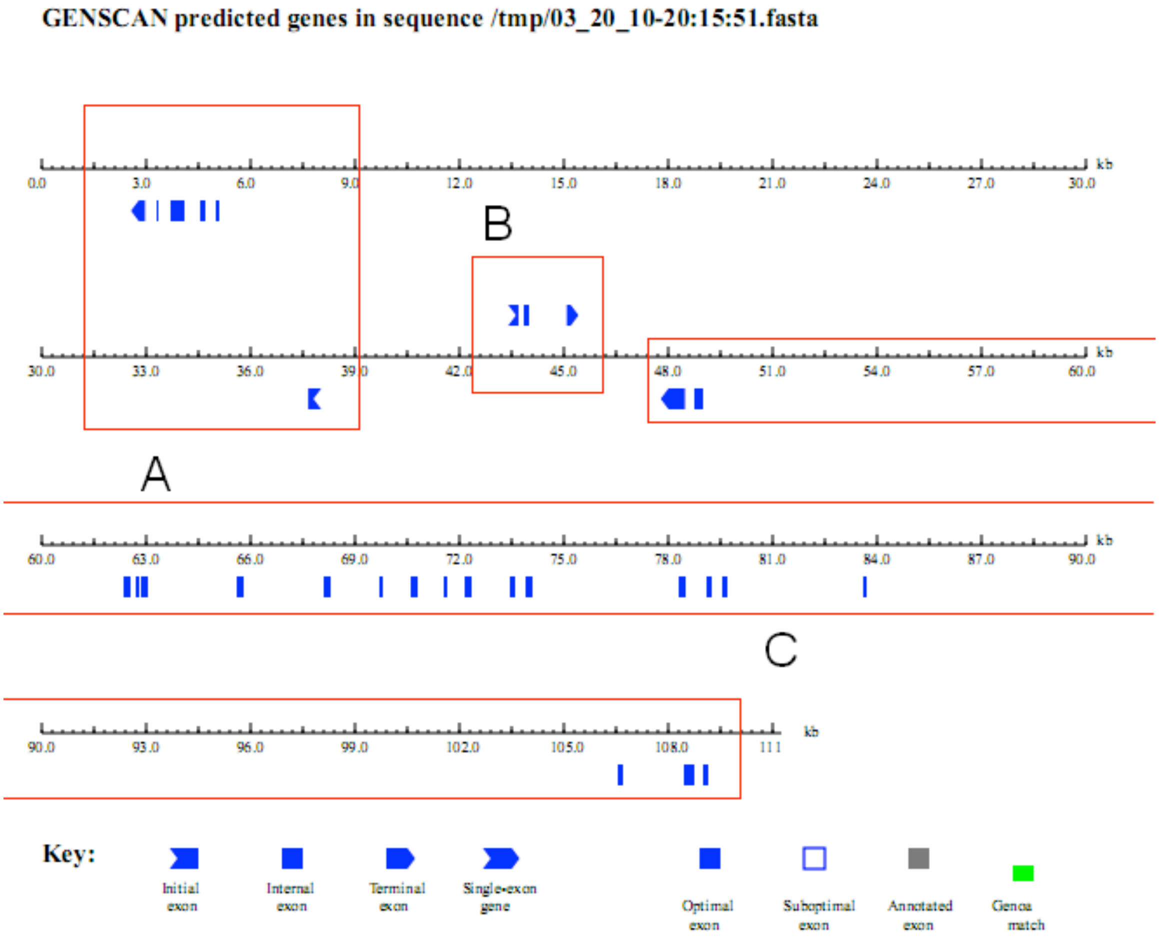


Figure 1. Initial GENSCAN gene map. Feature A and C run opposite to the coordinate system here while feature B runs in the same orientation. Feature C also lacks an initial exon.

Feature	Gene	Summary	Location (bp on chimp)	Strand	Accession #
A	FBXW11	Pseudogene	171-39169	-	NM_033645.2
B	SOD1	Ortholog	42227-45466	+	NP_000445.1
C	SFRS15	Ortholog	45597-108978	-	NM_001145445.1, NM_001145444.1, NM_020706.2

Table 1. Initial predictions based on GENSCAN map.

Feature A

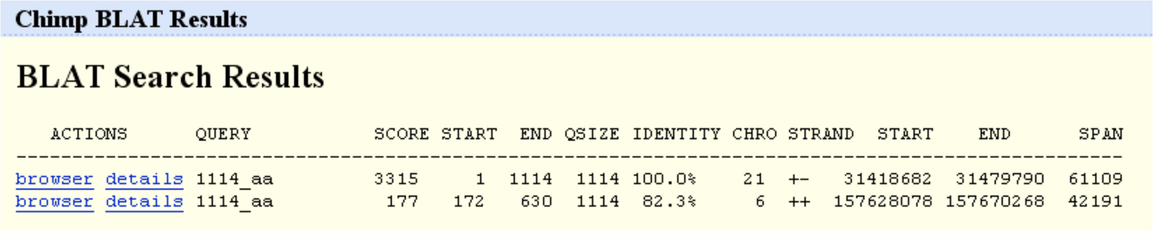
Feature A is a pseudogene of human ortholog FBXW11 which functions as a phosporylation-dependent ubiquitin. Young-In Kim used BLAT to align the GENSCAN predictions to chimp chromosome 21 and 5. *Blastp* confirmed these results in human and found the full-length gene on chromosome 5. *Blastx* of the whole masked DNA sequence of chimp chunk 03-11 against the GenBank FBXW11 peptide sequence confirmed feature A on chromosome 21 as a pseudogene. This analysis also showed hits with low percent identity in different coding frames. Because these hits were too close to be intronic and occur in both human and chimp, Young-In concluded that feature A was most likely a pseudogene that arose from a retrotransposition event before chimps and humans diverged.

Feature B

Feature B is the chimp ortholog of human gene superoxide dismutase (SOD1). Amanda Hay used BLAT to align feature B against the human genome and found a 99% identity match on human chromosome 21. Although *blastp* showed a match to human SOD1, BLAT later revealed that feature B did not completely cover all of the exons of SOD1. Using the unmasked chimp chunk DNA sequence, GENSCAN found two more exons that added an extra internal exon and the initial exon for the gene. After incorporating the extra exons, BLAST confirmed that all exons of human SOD1 protein were accounted for and feature B was concluded to be a gene.

Feature C

To verify that feature C, exists in chimp, I used BLAST-Like Alignment Tool (BLAT) at the start to align the GENSCAN output protein sequence against the published chimp genome from March 2006 (Figure 2). Figure 2 shows two general alignments matching on chimp chromosomes 21 and 6. As expected, one of the hits had complete coverage and identity. I decided to investigate the hit on chromosome 21 in more detail.



Chimp BLAT Results

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START    END      SPAN
---------------------------------------------------------------------------------------------------
browser details 1114_aa          3315     1 1114  1114 100.0%    21  +-    31418682  31479790  61109
browser details 1114_aa           177   172  630  1114  82.3%     6  ++   157628078 157670268  42191
```

Figure 2. BLAT with GENSCAN feature C protein sequence against chimp.

Figure 3 shows the browser view of the BLAT search. Though the previous BLAT results show complete coverage and identity, the UCSC Genome Browser does not display any corresponding chimp RefSeq RNA data. This can arise because feature C may not be expressed or, more likely, there is no corresponding entry in the chimp RefSeq RNA database. Because of the high identity and coverage among feature C and the three other human features, feature C is most likely expressed and lacks a RefSeq entry. The alignments suggest that there are three isoforms of human SFRS15 gene that correspond to feature C. These all have very similar alignments but differ at only a few sites. Because all three sequences have high similarity, the basic genetic sequence is likely the same but alternative splicing may slightly change the RNA among the three sequences to create the isoforms. Most of the GENSCAN predicted exons match the human counterparts and even in the two areas of inconsistency, the gap in the terminal exon (Figure 5) and the missing exon in feature C (Figure 7), are covered by matching DNA sequence. As such, most of the differences between feature C and the human sequence seem to come from inaccuracies of GENSCAN prediction over the entire DNA sequence. These results highly suggest that feature C reflects a real gene.

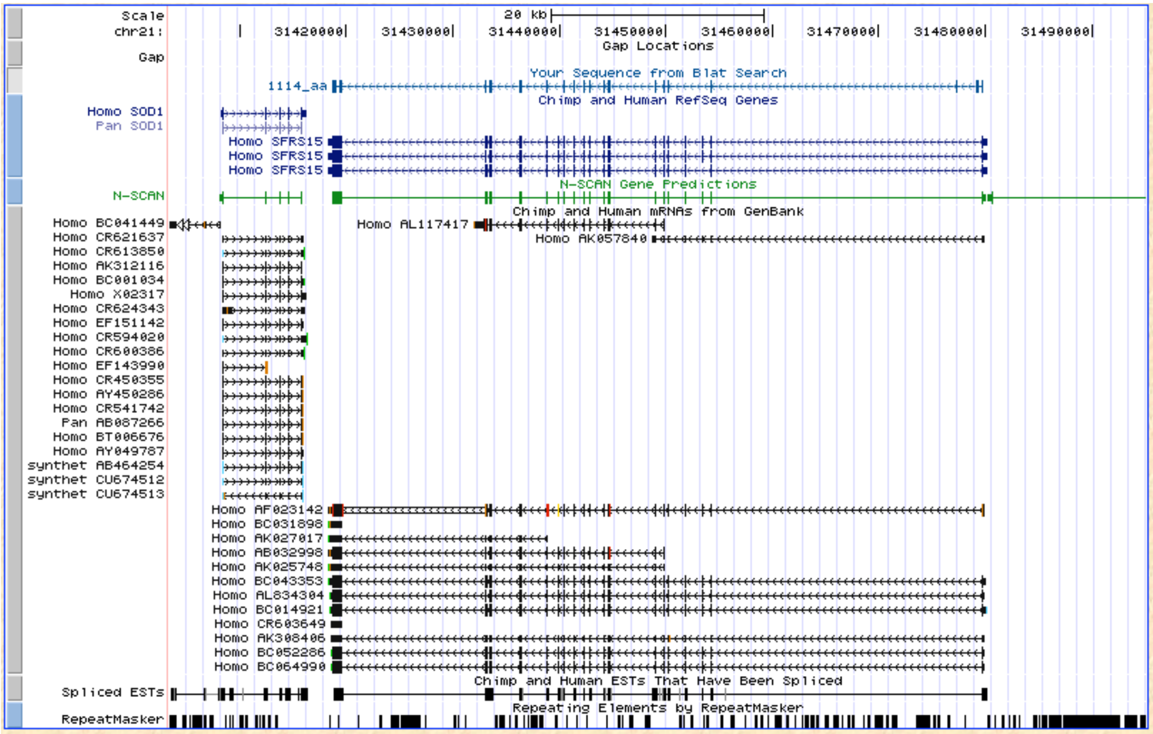
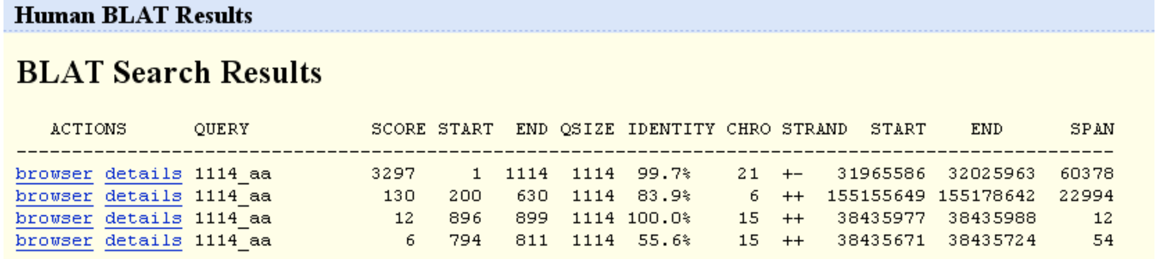


Figure 3. Browser view of 100% similarity match of chimp to chimp BLAT.

The alignment of feature C against human DNA (March 2006) on BLAT also has chromosomes 21 and 6 as hits and almost the same coverage and identity as the chimp BLAT results (Figure 4). The 99.7% identity and complete coverage between feature C and human chromosome 21 suggest that feature C really may be a real gene. The browser view of this hit shows the same three isoforms of human SFRS15 aligned against feature C as Figure 3.



Human BLAT Results

BLAT Search Results

```
   ACTIONS      QUERY           SCORE START  END QSIZE IDENTITY CHRO STRAND  START     END       SPAN
---------------------------------------------------------------------------------------------------
browser details 1114_aa          3297     1  1114  1114  99.7%    21  +-    31965586  32025963  60378
browser details 1114_aa           130   200   630  1114  83.9%     6  ++   155155649 155178642  22994
browser details 1114_aa            12   896   899  1114 100.0%    15  ++    38435977  38435988     12
browser details 1114_aa             6   794   811  1114  55.6%    15  ++    38435671  38435724     54
```

Figure 4. BLAT with GENSCAN feature C protein sequence against human.

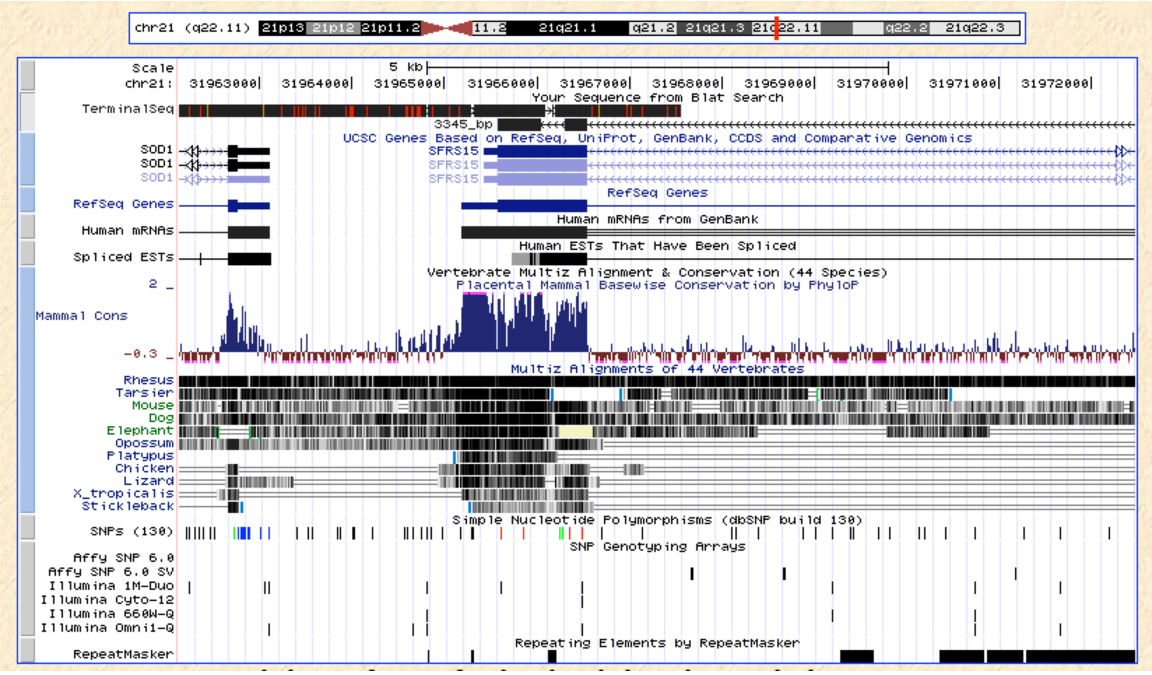


Figure 5. BLAT alignment of masked chunk 03-11 DNA sequence (TerminalSeq), GENSCAN results for feature C (3345_bp), and human sequence. In TerminalSeq row, black denotes match between query and human subject sequence and red denotes mismatch. The gap in TerminalSeq is masked sequence.

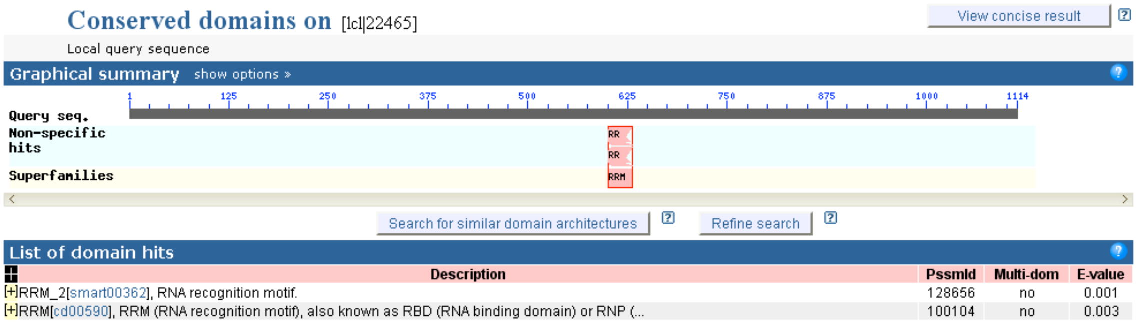
When run in *blastp* against the non-redundant (nr) protein database, the sequence matches with three different isoforms of human SFRS15 gene as part of its results (Table 2). Furthermore, aligning the GENSCAN CDS data for feature C against the entire human genome using *blastn* produces three mRNA transcript matches with high coverage and identity for SFRS15 (Table 3). These results all suggest that Feature C most likely was the SFRS15 gene. According to GenBank, the SFRS15 gene is a conserved member of the arginine/serine-rich splicing factor family. Furthermore, BLAST (Basic Local Alignment Search Tool) shows a putative conserved region between residues 601 and 631 that matches an RNA recognition motif (Figure 6). The Conserved Domain Database adds that this putative conserved region is found in the RNA binding domain in certain proteins such as RNA polymerase II in rats.

Accession	Description	Max Score	Total Score	Query Coverage	E value
NP_065757.1	splicing factor, arginine / serine-rich 15 isoform 1 [Homo sapiens]	976	1313	81%	0
NP_001138916.1	splicing factor, Arginine / serine-rich 15 isoform 2 [Homo sapiens]	993	1270	81%	0
NP_001138917.1	splicing factor, Arginine / serine-rich 15 isoform 3 [Homo sapiens]	919	1256	81%	0

Table 2. *blastp* hits from Feature C protein alignment to nr protein database. Note all three isoforms are represented.

Accession	Description	Max Score	Total Score	Query Coverage	E value
NM_001145445.1	Homo sapiens splicing factor, arginine/serine-rich 15 (SFRS15), transcript variant 3, mRNA	2383	5321	86.00%	0
NM_001145444.1	Homo sapiens splicing factor, arginine/serine-rich 15 (SFRS15), transcript variant 2, mRNA	2383	5447	88.00%	0
NM_020706.2	Homo sapiens splicing factor, arginine/serine-rich 15 (SFRS15), transcript variant 1, mRNA	2383	5442	88.00%	0

Table 3. *blastn* transcript hits from Feature C nucleotide alignment against human genome. Note all three isoforms are represented.



Conserved domains on [lcl|22465]

Local query sequence

Description	PssmId	Multi-dom	E-value
[+]RRM_2[smart00362], RNA recognition motif.	128656	no	0.001
[+]RRM[cd00590], RRM (RNA recognition motif), also known as RBD (RNA binding domain) or RNP (...	100104	no	0.003

Figure 6. Conserved domains on Feature C. The peptide sequence matches hits in the

CDD that correspond to a possible conserved RNA binding or recognition motif.

To investigate the missing initial end of the feature C from Figure 1, the last 36284 bp of the masked sequence were run against human in BLAT (Figure 7). The results show a 99.1% identity region on human chromosome 21 and also coverage of the supposed initial exon of feature C (based on the first exon in the SFRS15 RefSeq data). As such, this suggests that GENSCAN incorrectly omitted the initial exon for feature C.

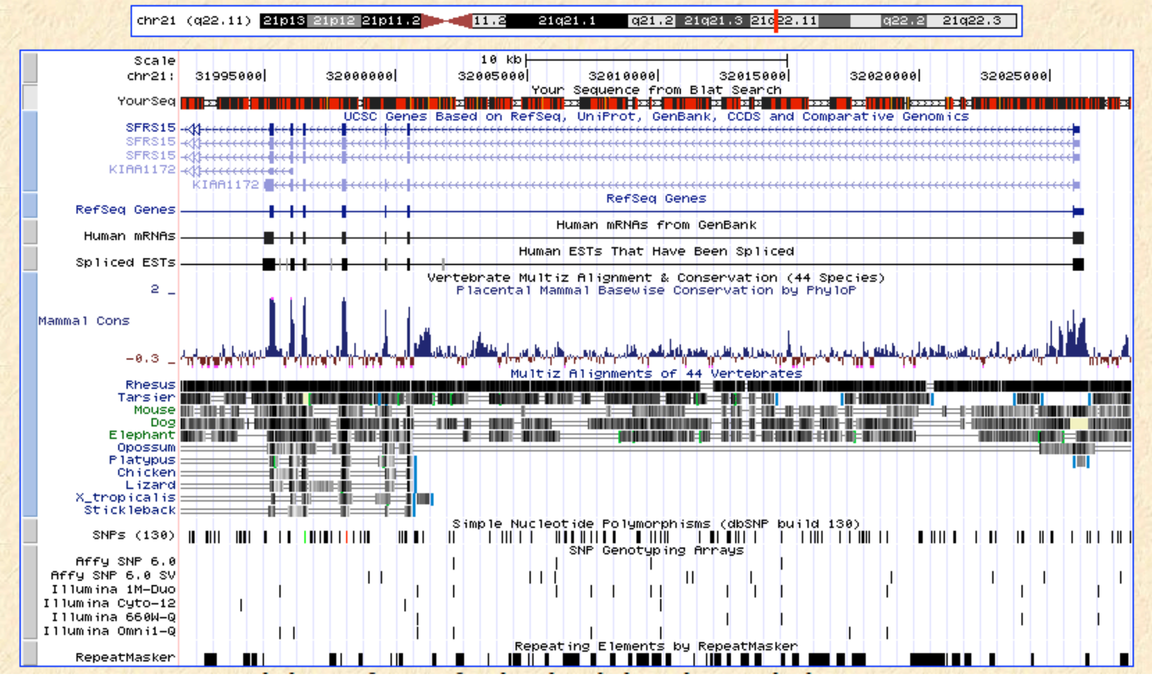


Figure 7. Alignment of last 36284 bp in masked chunk 03-11. The rightmost exon (initial exon) in the SFRS15 RefSeq data is covered by chunk 03-11 DNA sequence.

Interestingly, BLAT alignments of the final GenBank protein sequence of all three isoforms of SFRS15 also omit the initial exon of SFRS15 (Figure 8). Although Figure 7 shows alignment of mRNA at the initial exon with a small amount of coding sequence, the complete lack of the initial exon in the final GenBank protein sequence indicates that the coding interpretation may have been a miscall by BLAT. All of this and the presence of a methionine at the start of the second exon highly suggest that the initial exon may be part of a UTR region which is transcribed but not translated. Additionally, the initial peptide product from the full mRNA may undergo post-translational modification and the portion of the peptide corresponding to the initial exon may have been cleaved out prior to determination of the peptide sequence. However, this seems less likely due to the much greater evidence pointing to a putative UTR.

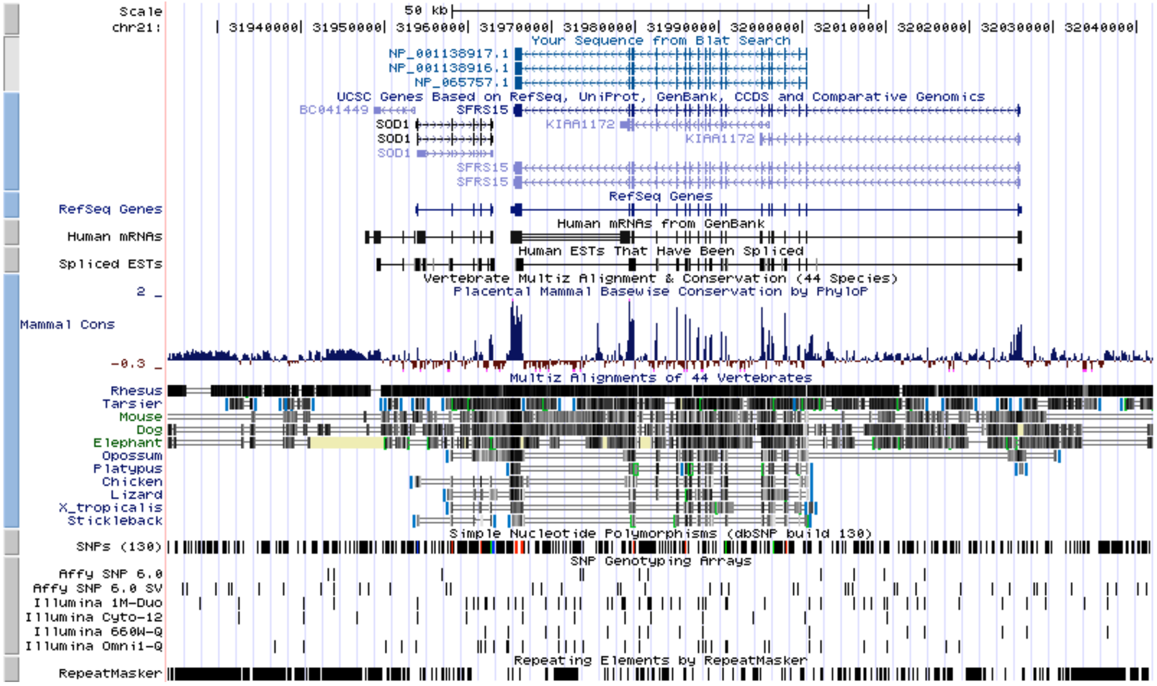
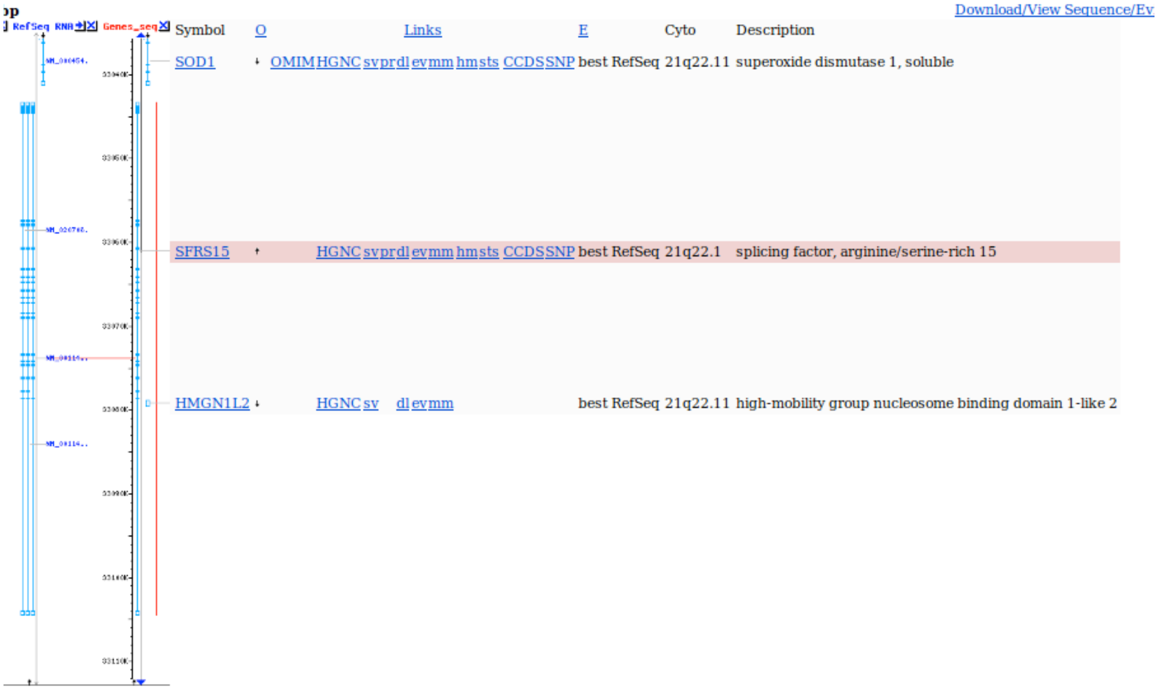


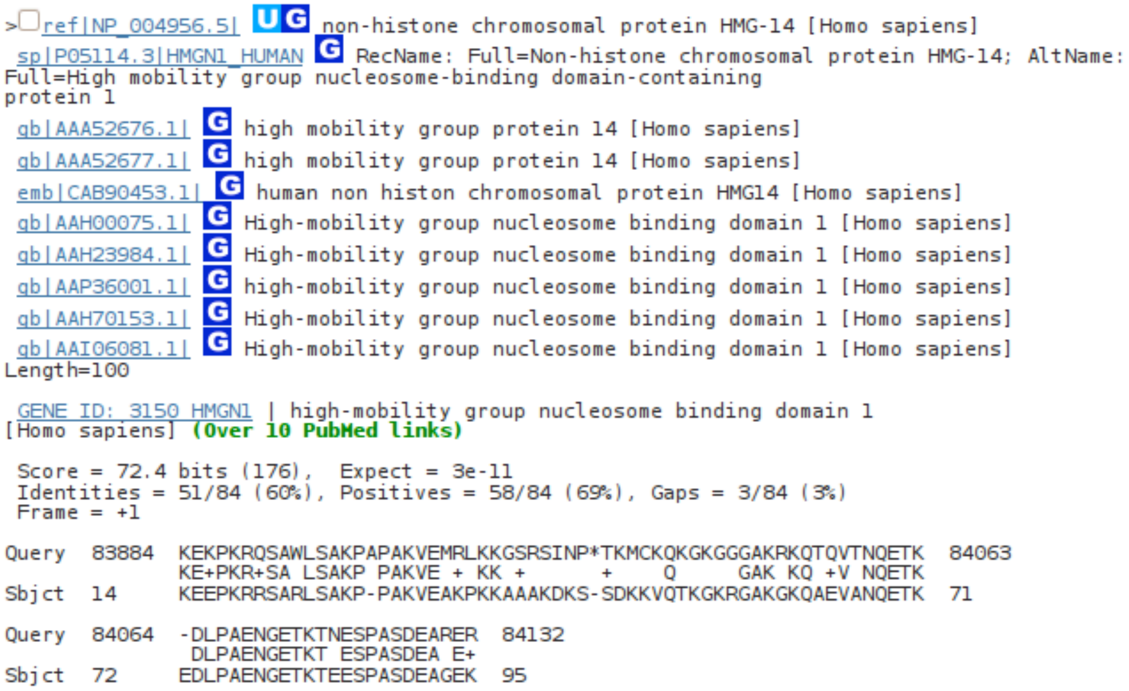
Figure 8. Alignments of the three isoforms of SFRS15 GenBank protein sequence in BLAT. Note the rightmost exon is unaligned compared to the full protein sequences.

To check for any other features in feature C that GENSCAN missed, NCBI Map Viewer and *blastx* were used to visualize the surrounding features on human chromosome 21. The NCBI Map Viewer shows the high-mobility group nucleosome binding domain 1-like 2 pseudogene (HMGN1L2) within the SFRS15 gene (Figure 9). According to GenBank, the HMGN1L2 pseudogene runs in the opposite orientation to the SFRS15 gene and has not been confirmed with experimental evidence. I then used *blastx* to sequentially align 10,000 bp regions from 45,000 to 111,000 bp of the masked DNA sequence against the human genome. *Blastx* results show two significant hits including a 81 residue alignment between feature C and the non-histone chromosomal protein HMG-14 (Figure 10) as well as a match to putative RNA-binding protein 16 (Figure 11).



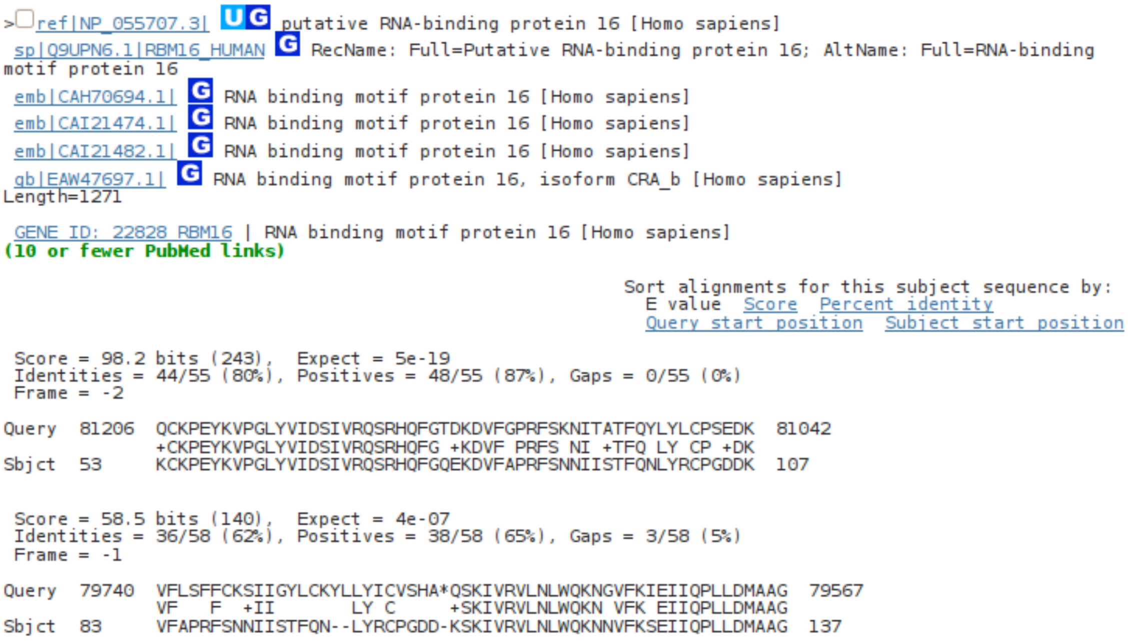
Symbol		O	Links	E	Cyto	Description
SOD1	↓	OMIM	HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	21q22.11	superoxide dismutase 1, soluble
SFRS15	↑		HGNC sv pr dl ev mm hm sts CCDS SNP	best RefSeq	21q22.1	splicing factor, arginine/serine-rich 15
HMGN1L2	↓		HGNC sv dl ev mm	best RefSeq	21q22.11	high-mobility group nucleosome binding domain 1-like 2

Figure 9. NCBI Map Viewer of region around SFRS15 gene on chromosome 21. Note the presence of HMGN1L2 in the length of SFRS15 (vertical red line).



```
>□ref|NP_004956.5| UG non-histone chromosomal protein HMG-14 [Homo sapiens]
 sp|P05114.3|HMGN1_HUMAN G RecName: Full=Non-histone chromosomal protein HMG-14; AltName:
Full=High mobility group nucleosome-binding domain-containing
protein 1
 gb|AAA52676.1| G high mobility group protein 14 [Homo sapiens]
 gb|AAA52677.1| G high mobility group protein 14 [Homo sapiens]
 emb|CAB90453.1| G human non histon chromosomal protein HMG14 [Homo sapiens]
 gb|AAH00075.1| G High-mobility group nucleosome binding domain 1 [Homo sapiens]
 gb|AAH23984.1| G High-mobility group nucleosome binding domain 1 [Homo sapiens]
 gb|AAP36001.1| G high-mobility group nucleosome binding domain 1 [Homo sapiens]
 gb|AAH70153.1| G High-mobility group nucleosome binding domain 1 [Homo sapiens]
 gb|AAI06081.1| G High-mobility group nucleosome binding domain 1 [Homo sapiens]
Length=100

 GENE ID: 3150 HMGN1 | high-mobility group nucleosome binding domain 1
[Homo sapiens] (Over 10 PubMed links)

 Score = 72.4 bits (176),  Expect = 3e-11
 Identities = 51/84 (60%), Positives = 58/84 (69%), Gaps = 3/84 (3%)
 Frame = +1

Query  83884  KEKPKRQSAWLSAKPAPAKVEMRLKKGSRSINP*TKMCKQKGKGGGAKRKQTQVTNQETK  84063
              KE+PKR+SA LSAKP PAKVE + KK +       +     Q     GAK KQ +V NQETK
Sbjct  14     KEEPKRRSARLSAKP-PAKVEAKPKKAAAKDKS-SDKKVQTKGKRGAKGKQAEVANQETK  71

Query  84064  -DLPAENGETKTNESPASDEARER  84132
               DLPAENGETKT ESPASDEA E+
Sbjct  72     EDLPAENGETKTEESPASDEAGEK  95
```

Figure 10. *blastx* match to non-histone chromosomal protein HMG-14.

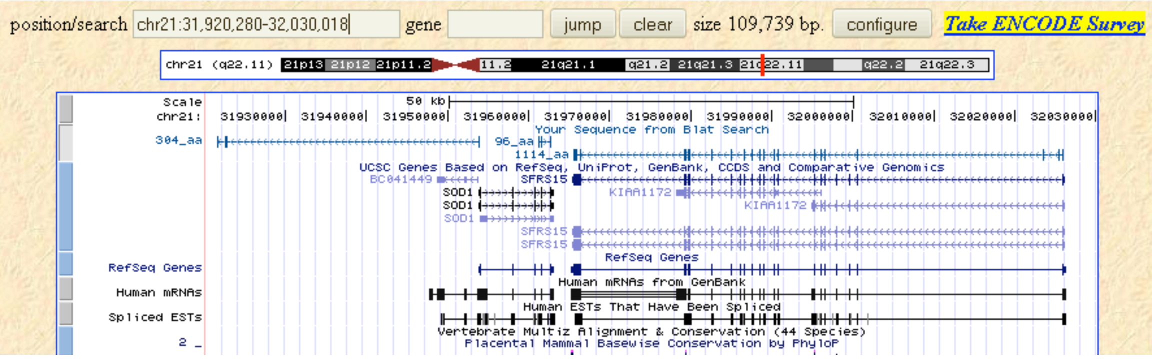


```
>□ref|NP_055707.3| UG putative RNA-binding protein 16 [Homo sapiens]
 sp|Q9UPN6.1|RBM16_HUMAN G RecName: Full=Putative RNA-binding protein 16; AltName: Full=RNA-binding
motif protein 16
 emb|CAH70694.1| G RNA binding motif protein 16 [Homo sapiens]
 emb|CAI21474.1| G RNA binding motif protein 16 [Homo sapiens]
 emb|CAI21482.1| G RNA binding motif protein 16 [Homo sapiens]
 gb|EAW47697.1| G RNA binding motif protein 16, isoform CRA_b [Homo sapiens]
Length=1271

 GENE ID: 22828 RBM16 | RNA binding motif protein 16 [Homo sapiens]
(10 or fewer PubMed links)

                                                Sort alignments for this subject sequence by:
                                                  E value  Score  Percent identity
                                                  Query start position  Subject start position

 Score = 98.2 bits (243),  Expect = 5e-19
 Identities = 44/55 (80%), Positives = 48/55 (87%), Gaps = 0/55 (0%)
 Frame = -2

Query  81206  QCKPEYKVPGLYVIDSIVRQSRHQFGTDKDVFGPRFSKNITATFQYLYLCPSEDK  81042
              +CKPEYKVPGLYVIDSIVRQSRHQFG +KDVF PRFS NI +TFQ LY CP +DK
Sbjct  53     KCKPEYKVPGLYVIDSIVRQSRHQFGQEKDVFAPRFSNNIISTFQNLYRCPGDDK  107


 Score = 58.5 bits (140),  Expect = 4e-07
 Identities = 36/58 (62%), Positives = 38/58 (65%), Gaps = 3/58 (5%)
 Frame = -1

Query  79740  VFLSFFCKSIIGYLCKYLLYICVSHA*QSKIVRVLNLWQKNGVFKIEIIQPLLDMAAG  79567
              VF   F  +II         LY C     +SKIVRVLNLWQKN VFK EIIQPLLDMAAG
Sbjct  83     VFAPRFSNNIISTFQN--LYRCPGDD-KSKIVRVLNLWQKNNVFKSEIIQPLLDMAAG  137
```

Figure 11. *blastx* match to putative RNA-binding protein 16.

Synteny and Repeats

Synteny is the relative order of genes on a specific region of a genome. When two sequences share high synteny, it offers some evidence of relationship between two sequences. Using the GENSCAN output of the masked DNA sequence, BLAT shows that all three features map to both chimp and human chromosome 21 in the same order and orientation (Figure 12). When aligned against orangutan (Figure 13) and mouse (Figure 14) genomes, the three features also share synteny although the features map to chromosome 16 in mouse. This provides some evidence that the gene order has been conserved since mouse diverged from primates.



position/search chr21:31,920,280-32,030,018 gene jump clear size 109,739 bp. configure *Take ENCODE Survey*

Figure 12. Three features aligned against human show the same relative orientation and order as in chimp (Figure 1). All map to human chromosome 21.

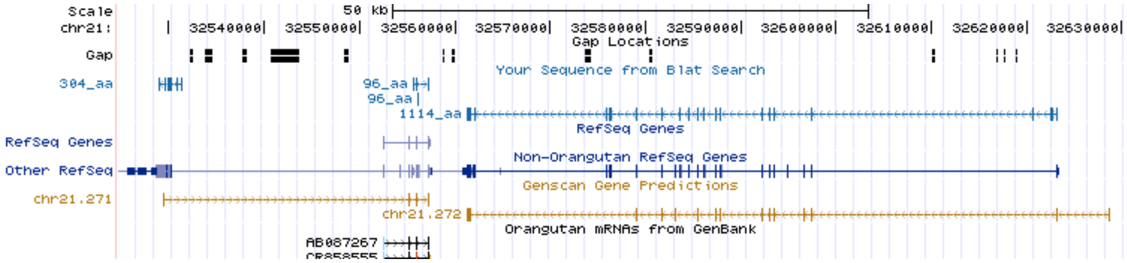


Figure 13. Three features aligned against orangutan show the same relative orientation and order as chimp and human. All map to orangutan chromosome 21.

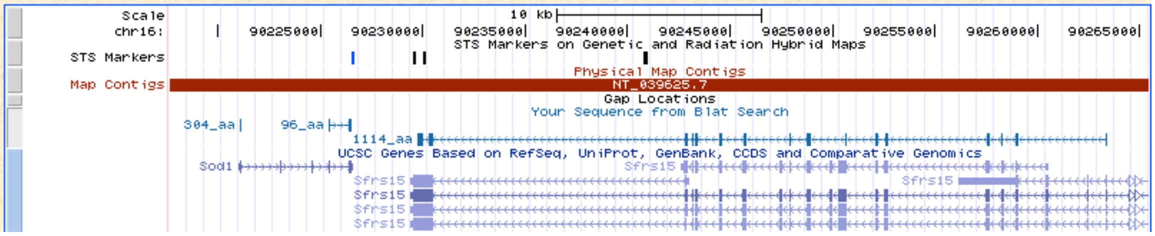
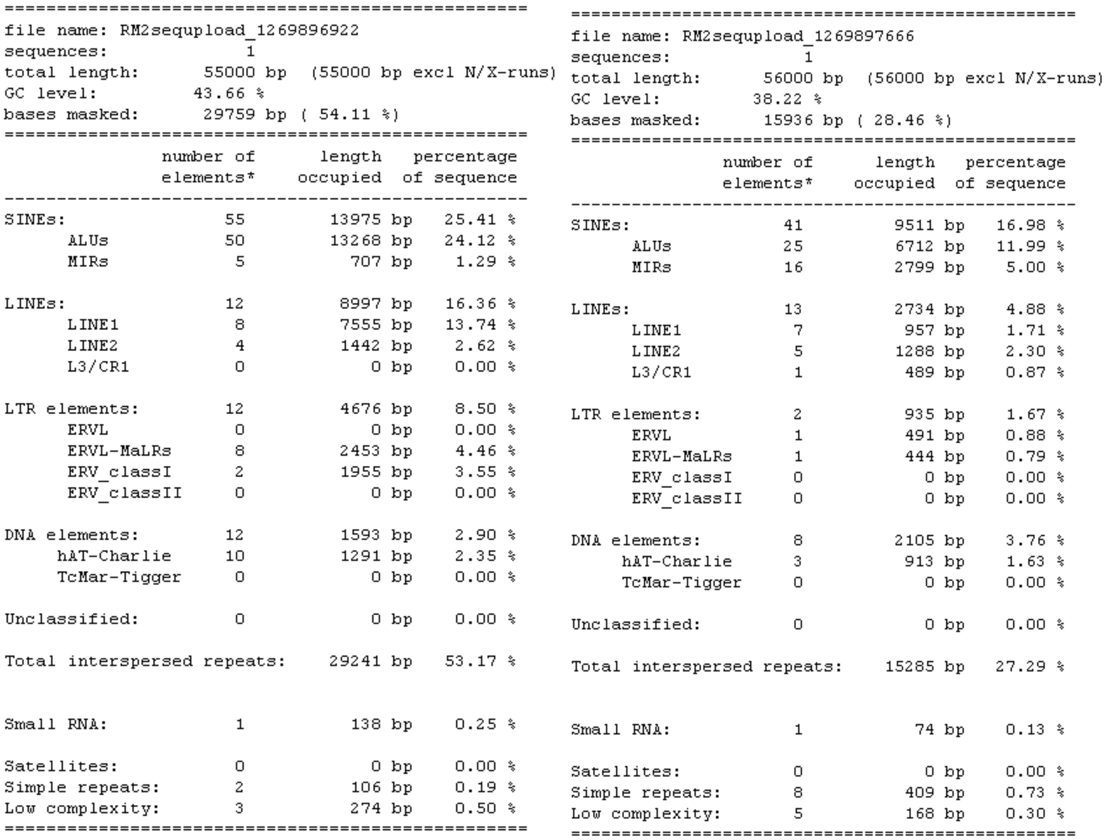


Figure 14. Three features aligned against mouse show same relative orientation and order as chimp, human, and orangutan. All map to mouse chromosome 16.

Running the unmasked sequence into RepeatMasker showed that there were several areas of long repeat in chunk 03-11. In general, the repeats were significant only in the annotation of feature B where two possible exons were masked. Two other repeats of greater than 1 kB were found at bp 11211-12198 (LINE/LI) and at bp 31928-33254 (LTR/ERV1). Figure 15 summarizes the results.



```
==================================================
file name: RM2sequpload_1269896922
sequences:            1
total length:       55000 bp  (55000 bp excl N/X-runs)
GC level:         43.66 %
bases masked:       29759 bp ( 54.11 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              55        13975 bp   25.41 %
      ALUs          50        13268 bp   24.12 %
      MIRs           5          707 bp    1.29 %

LINEs:              12         8997 bp   16.36 %
      LINE1          8         7555 bp   13.74 %
      LINE2          4         1442 bp    2.62 %
      L3/CR1         0            0 bp    0.00 %

LTR elements:       12         4676 bp    8.50 %
      ERVL           0            0 bp    0.00 %
      ERVL-MaLRs     8         2453 bp    4.46 %
      ERV_classI     2         1955 bp    3.55 %
      ERV_classII    0            0 bp    0.00 %

DNA elements:       12         1593 bp    2.90 %
     hAT-Charlie    10         1291 bp    2.35 %
     TcMar-Tigger    0            0 bp    0.00 %

Unclassified:        0            0 bp    0.00 %

Total interspersed repeats:    29241 bp   53.17 %


Small RNA:           1          138 bp    0.25 %

Satellites:          0            0 bp    0.00 %
Simple repeats:      2          106 bp    0.19 %
Low complexity:      3          274 bp    0.50 %
==================================================
```

```
==================================================
file name: RM2sequpload_1269897666
sequences:            1
total length:       56000 bp  (56000 bp excl N/X-runs)
GC level:         38.22 %
bases masked:       15936 bp ( 28.46 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              41         9511 bp   16.98 %
      ALUs          25         6712 bp   11.99 %
      MIRs          16         2799 bp    5.00 %

LINEs:              13         2734 bp    4.88 %
      LINE1          7          957 bp    1.71 %
      LINE2          5         1288 bp    2.30 %
      L3/CR1         1          489 bp    0.87 %

LTR elements:        2          935 bp    1.67 %
      ERVL           1          491 bp    0.88 %
      ERVL-MaLRs     1          444 bp    0.79 %
      ERV_classI     0            0 bp    0.00 %
      ERV_classII    0            0 bp    0.00 %

DNA elements:        8         2105 bp    3.76 %
     hAT-Charlie     3          913 bp    1.63 %
     TcMar-Tigger    0            0 bp    0.00 %

Unclassified:        0            0 bp    0.00 %

Total interspersed repeats:    15285 bp   27.29 %


Small RNA:           1           74 bp    0.13 %

Satellites:          0            0 bp    0.00 %
Simple repeats:      8          409 bp    0.73 %
Low complexity:      5          168 bp    0.30 %
==================================================
```

Figure 15. Repetitious areas found by RepeatMasker. Complete chunk was divided in half to fit the limits of software.

Summary

Chimp chunk 03-11 had a total of three features predicted by GENSCAN. Feature A is a pseudogene of human ortholog FBXW11 which functions as a phosporylation-dependent ubiquitin. *Blastx* analysis showed matching hits that were in different frames and too close to be intronic, suggesting a retrotranposition event for the creation of the pseudogene. Feature B is the chimp ortholog of human gene superoxide dismutase (SOD1). RepeatMasker masked sequence that corresponded to two exons. Afterwards, BLAST confirmed the existence of all exons. Feature C is the chimp ortholog of human gene SFRS15, an RNA splicing factor that has three different isoforms. GENSCAN incorrectly missed the initial exon but matching coverage confirmed its existence. The HMGN1L2 pseudogene and a portion of the non-histone chromosomal HMG-14 are also found within the Feature C region. BLAT shows that these three features have the same relative order in human, chimp, orangutan, and mouse. Figure 16 shows the final map with features labeled.

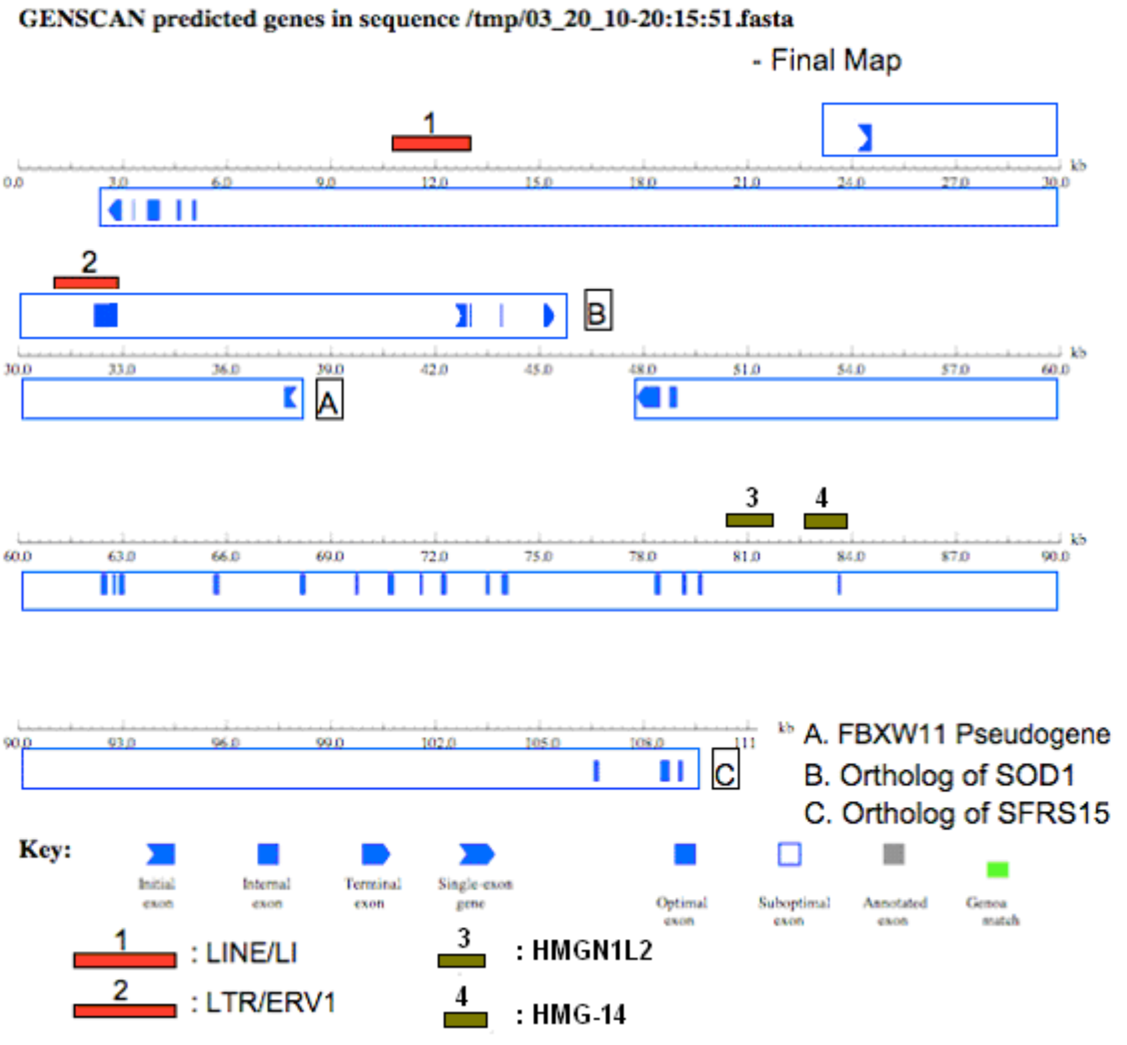


Figure 16. Final map of annotated chimp chunk 03-11.