Chimp Sequence Annotation: Region 2_4

David Desruisseau
April 17, 2006

BIO434W Genomics

Introduction

We received region 2_4 of the ChimpChunk sequence and started our annotation analysis by running RepeatMasker with the –nolow option enabled, which tells the program not to mask low complexity regions. RepeatMasker did, however, mask regions of the sequence based on other default parameters, such as interspersed repeats. Following this, we ran an NCBI *nr* (all GenBank- and refseq-inclusive) Blastn alignment of this masked sequence file and directed that these Blast results be put into an html file. This *nr* NCBI database was used in all subsequent Blastn searches unless otherwise noted. These initial results gave an initial description of known genes that might be present, and provided the starting point for the rest of the research.

Summary of Predicted Features

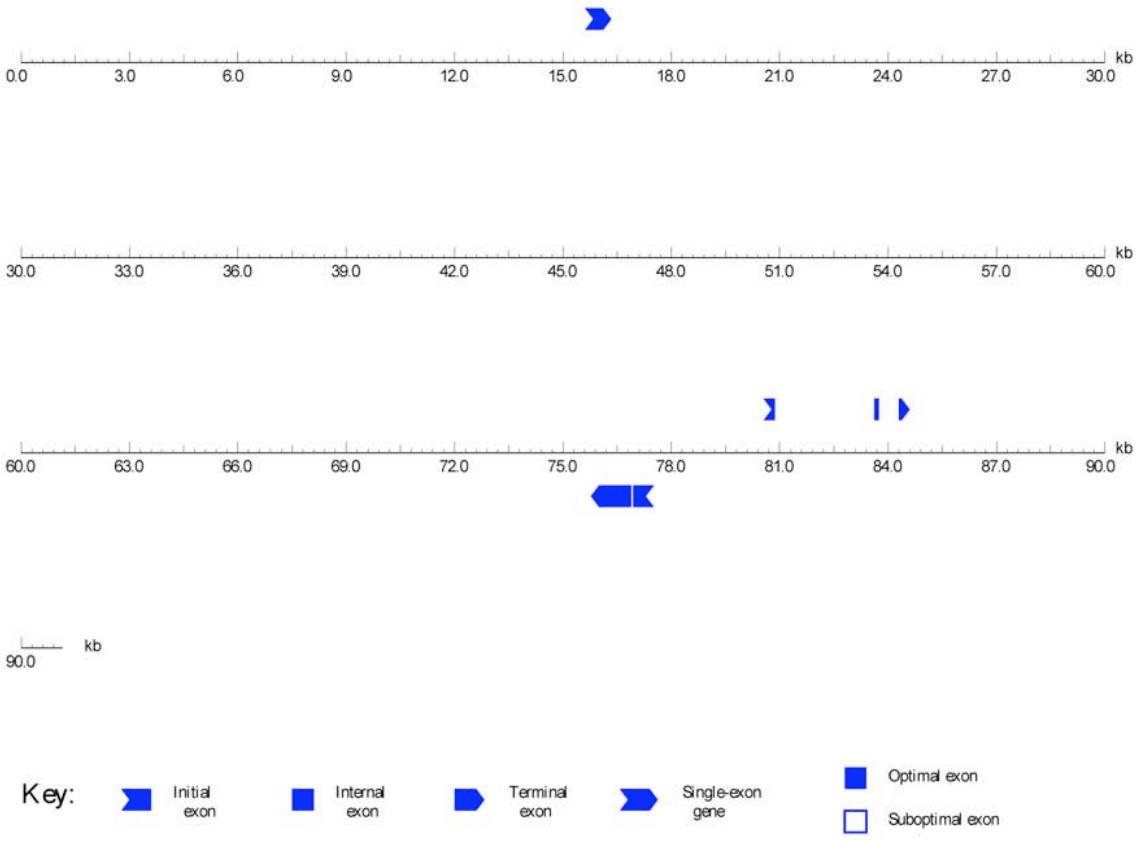
Feature	Exons	Position (bp)	Characterization
1	Single	15 839 – 16 111	Pseudogene
2	2	76 887 – 76 001	Pseudogene of Human Zinc Finger
3	3	80 780 – 80 869	Hip2 NM_005339.3

GenScan Results

We ran a GenScan search on our masked chimp DNA sequence; this identified three gene features based on criteria for predicted exon-intron structures. The first predicted feature codes for a 90 amino acid sequence contained in one exon. The second predicted feature is a two-exon region coding for 411 amino acids. The third feature is a relatively short region coding for 97 amino acids that is comprised of three exons. The GenScan results, including the corresponding predicted peptide sequences, are shown below:

```
Gn.Ex Type S .Begin ...End .Len Fr Ph I/Ac Do/T CodRg P.... Tscr..
----- ---- - ------ ------ ---- -- -- ---- ---- ----- ----- ------
 1.01 Sngl +  15839  16111  273  1  0   60   41   306 0.368  18.18
 1.02 PlyA +  18685  18690    6                                1.05

 2.03 PlyA -  18906  18901    6                                1.05
 2.02 Term -  76887  76001  887  1  2   81   42   928 0.999  79.16
 2.01 Init -  77291  76943  349  2  1   82   47   514 0.999  44.09
 2.00 Prom -  78581  78542   40                              -10.05

 3.00 Prom +  78989  79028   40                              -10.55
 3.01 Init +  80780  80869   90  1  0   72  110   105 0.991  11.64
 3.02 Intr +  83617  83745  129  0  0   51  110    82 0.980   6.67
 3.03 Term +  84298  84372   75  0  0  127   54    41 0.994   1.46
 3.04 PlyA +  84537  84542    6                                1.05

>Pan|GENSCAN_predicted_peptide_1|90_aa
MMGIIVVAINVAEQKRTAVDFLGLHQLKDGPDRDPGERFPALDTAFSLRVPGLAESGRRSSRVRRRKPVRVRPNLRSEASAG
SPDHSTRV

>Pan|GENSCAN_predicted_peptide_2|411_aa
MNKLRAEKWFCDVTIVADSLKFRGHKVILAACSPFLRDQFLLTPSSELQVSLMHSARIVADLLLYCYTGTLEFAVRDIVNYL
TVTSYLQMEHVVEKCRNALSQFTEPKIGLKEDGVPRTPKLAPKPPPPPPLSPPLLRPVKLEFPLDEDLELKAEEEDEDEDVS
DICIVKVESALDIAHRLKPPGGLGGGLGIGGSVGGHLGELAQSSVLPSTVAPPQGVVKACYSLSENAEGESLLLTPGGRASV
GATSGLVEAAAVAMVARGAGGSQGPLPGSFSGGNPLKNIKCTKCPEVFQGVAKLVFHMRQQHFIFMCPRCGKEFNHSNNLNH
HRNVHRGVKSHPCSRCGKCFTQKSTLHDHLNLHSGAQPYRCSYCDMRFAHKPAIRRHLKEQHGKTTAENVLETSVAEINVLI
R

>Pan|GENSCAN_predicted_peptide_3|97_aa
MTLRTVLLSLQALLAAAEPDDPQDAVVANQYKQNPEMFKQTARLWAHVYAGAPVSSPEYTKKIENLCAMGFDRNAVIVALSS
KSWDVETATELLLSN
```

Original GenScan Prediction Map for Chimpchunk Region 2_4



Repeats

The two non-SINE repetitious features over 500 bp in length identified by RepeatMasker are given below:

Position (bp)	Repeat Type	Class/Family
21815 - 23213	LTR12C	LTRERU1
63301 - 63921	L2	LINE/L2

Full Summary Table

```
==================================================
file name: pan_chunk2_4.fasta
sequences:             1
total length:      91150 bp  (86189 bp excl N-runs)
GC level:         41.91 %
bases masked:      45188 bp ( 49.58 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              141         34062 bp    37.37 %
      ALUs          132         32815 bp    36.00 %
      MIRs            9          1247 bp     1.37 %

LINEs:               25          6990 bp     7.67 %
      LINE1          14          4948 bp     5.43 %
      LINE2          10          1955 bp     2.14 %
      L3/CR1          1            87 bp     0.10 %

LTR elements:         7          2675 bp     2.93 %
      MaLRs           4          1059 bp     1.16 %
      ERVL            1            97 bp     0.11 %
      ERV_classI      2          1519 bp     1.67 %
      ERV_classII     0             0 bp     0.00 %

DNA elements:         9          1175 bp     1.29 %
      MER1_type       5           667 bp     0.73 %
      MER2_type       1           240 bp     0.26 %

Unclassified:         0             0 bp     0.00 %

Total interspersed repeats:    44902 bp    49.26 %


Small RNA:            2           286 bp     0.31 %

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

Initial Impressions of the Herne Browser Output

The majority of the EST matches displayed in the Herne browser were cDNA clones, as revealed by clicking on an arrow to view each result's entry. However, these cDNA clones are not as useful in identifying genetic features in our region, since they can simply indicate the presence of transcribed product from nearby promoters (read-through). Instead we focused on locating good EST matches to known genes. The rest of the chunk did not initially reveal any significant EST matches, except for a highly EST-rich region towards the right end of the chunk. At the far right of the chunk, there is an area of approximately 5 kilobases displaying a high number of adjacent EST matches in both directions (Figure 1), likely indicating an exon or other significant genetic feature, which will be discussed later in the paper.

Figure 1
Initial Herne view

Analysis of Feature 1

Using the predicted GenScan protein amino acid sequence, we performed an *nr* NCBI protein Blast search using this single exon feature, which resulted in no matches. Next, we extracted the DNA sequence from this region (15,839 to 16,111 bp) and ran the sequence through a Blat search on the UCSC database site against the human genome. This showed the predicted Feature 1 matching the right end of a significantly longer GenScan predicted human gene (brown) on chromosome 4, as shown in Figure 2.

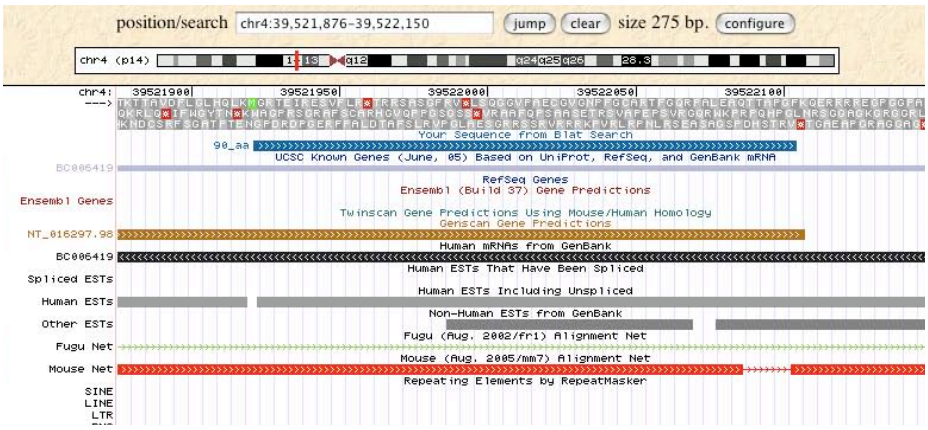


Figure 2
Feature 1 Blat

In order to verify the validity of this relationship, we ran Blast2 to compare the DNA sequence of our predicted Feature 1 to the predicted GenScan human gene referred to in Figure 2. The DNA sequence of the predicted GenScan gene was obtained by clicking on the gene hit in the UCSC Genome Browser. The resulting comparison showed a high match between these two sequences, but with some significant insertions and deletions. These misalignments suggested that our Feature 1 might in fact be a pseudogene of the anonymous gene predicted by GenScan in human DNA. Note, however, that as there is no human RefSeq gene, the human gene might also be a mispredicted region or pseudogene.

Next, we ran a ClustalW match between the two genes in order to see a global alignment of the entire sequence. This search resulted in a poorly matching region flanked by two well-aligned regions, suggesting that our chimp DNA was in fact derived from a two-exon gene, but missing the middle intron. This further supported Feature 1 as being a pseudogene of the predicted GenScan gene. Taking all of this evidence into account seems to indicate that Feature 1 is a pseudogene of the predicted GenScan gene present on chromosome 4 in humans.

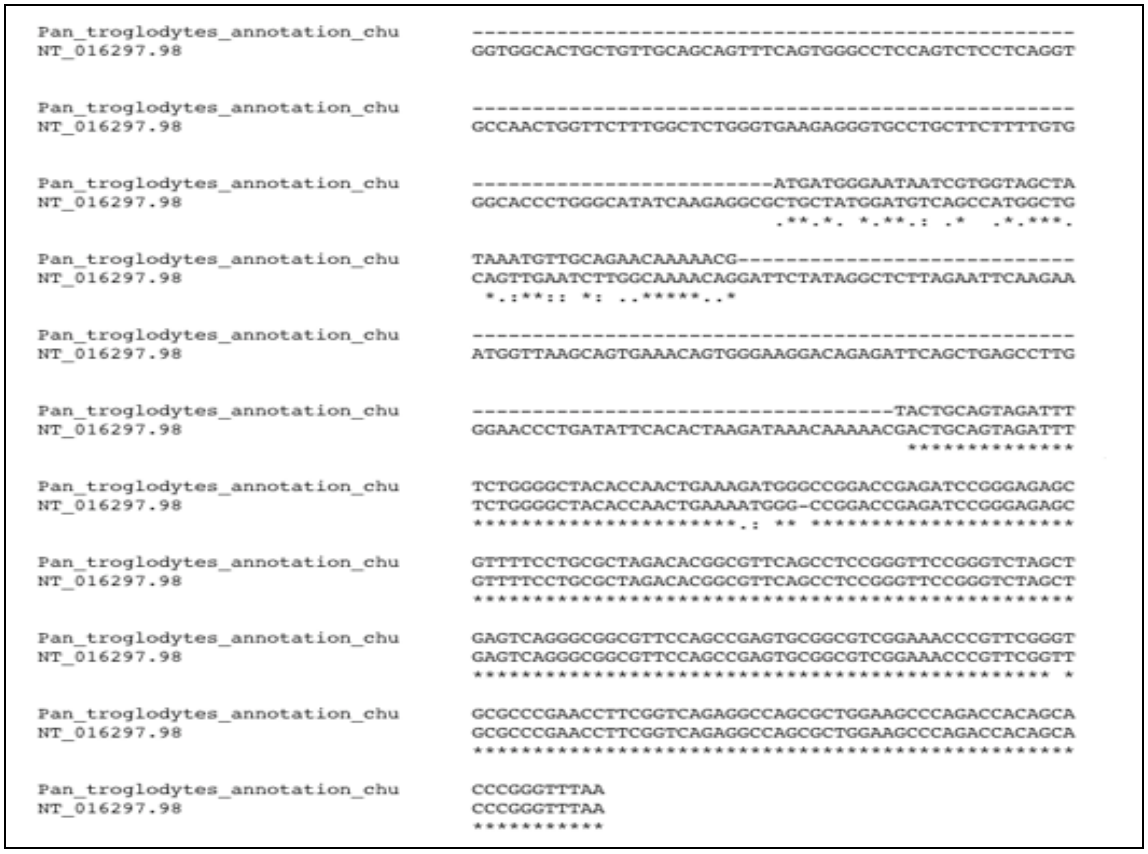


Figure 3
Feature 1 ClustalW result

Analysis of Feature 2

To characterize the second feature, we began by running an NCBI Blast search with the amino acid sequence of the two-exon gene predicted by GenScan. This resulted in several relatively low-quality matches to mouse zinc finger genes. In order to see if we could find matches to human sequence, we extracted the DNA of this putative GenScan gene region from 76,001 to 77,291 bp to run a Blastn search on the nucleotide sequence itself. This search resulted in high quality matches to human ZBT zinc finger genes.

A true orthologue of the human gene in our chimp sequence would require a match to the same gene on chromosome 4 of the human sequence. The Blat search resulted in a very good 98.7% match on chromosome 4 in the appropriate region, but only matched to an anonymous GenScan predicted gene, not an annotated zinc finger gene. However, there was a match of 95.7% to the actual refseq human zinc finger gene on chromosome 6. Although a bit low, considering the average 98% baseline similarity between humans and chimps, this match suggested that the true zinc finger gene we were looking for was likely located on chromosome 6. In order to verify this, we ran a Blat search using the Refseq entry for human ZBT and did indeed find a 99% match to the location on human chromosome 6, as we had suspected. This lack of synteny indicated the true orthologue of the ZBT gene would match to chromsome 6, while Feature 2 is on chromosome 4.

The ClustalW alignment between the Refseq ZBT entry and our extracted chimp DNA supports these findings. This alignment indicated a poor match between the two sequences. Therefore, it appears that our Feature 2 is a pseudogene of human ZBT, though it is still possible that this feature is actually a paralogue. Locating stops or indels in this region would help confirm whether or not this is the case.

Analysis of Feature 3

GenScan predicted the third feature as being comprised of three exons, despite its relatively short sequence length of 97 amino acids. As with features 2 and 3, we started out by running an NCBI Blast search of the predicted polypeptide sequence, which showed several matches to cow and mouse Hip2 genes. To confirm these results, we ran a Blat search in the UCSC database with this protein sequence as the query. The sole result matched at 100%, but was significantly shorter than that of the actual Hip2 gene in humans, indicating that GenScan had found only a portion of the true gene.

To confirm that our gene was indeed the suspected Hip2 gene, we ran a Blast2 Tblastn analysis to compare the Hip2 Refseq sequence against our entire masked chimp sequence. This resulted in very good matches to not only the already-matched region, but also to significant sequence upstream of our GenScan-predicted sequence. Further investigation revealed that there were in fact several small exons of the Hip2 gene upstream of the predicted region. These exons account for approximately 100 amino acids of sequence upstream of the previously matched region, which would explain where the rest of the Hip2 gene is located. Examining ESTs in this region using Herne corroborated this evidence, with columns of identically-matched ESTs present at locations corresponding to these exons. It seems that GenScan failed to include the first portion of the Hip2 gene in its prediction because the exons in that region were simply too small to be considered. However, taking into account the extremely good match of

the 97 amino acids that were predicted, as well as the corroborating EST evidence all but confirms that Feature 3 is in fact an orthologue of Hip2.

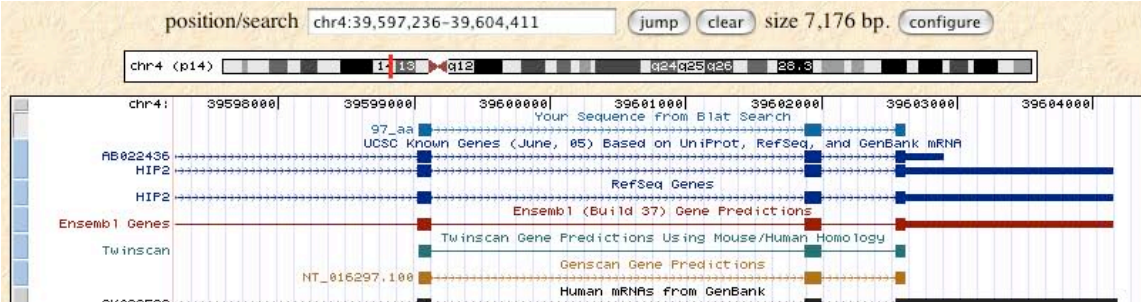


Figure 4
Feature 3 Blat result

Analysis of EST-rich Region

After characterizing all three predicted GenScan regions, we were left with an unusually EST-rich region at the right end of our chimp sequence matching numerous cDNA clones. In order to make sense of this region, we extracted the DNA from 85,400 to 88,800 bp. Running an NCBI Blastn search with this query resulted in a single match to Hip2. Since we had already established a Hip2 gene as Feature 3 of our chimp sequence, it was initially unclear as to why there was a match to Hip2 sequence in this downstream region.

To further investigate this, we did a Blat search with this extracted DNA sequence using the UCSC Browser, which showed Hip2 matches to the left end of our chimp sequence, as well as the many EST matches further downstream (Figure 5). These matches strongly suggest that these cDNA clones are possibly derived from transcriptional read-through products resulting from a weak adenylation site at the terminal end of the Hip2 gene. This would allow continuation of transcription past this site, without the actual translation of a protein product. This seems to explain the disproportionately high level of EST read depth in this otherwise seemingly featureless region. It is also quite possible that the existing annotation for Hip2 at this location is simply incorrect, and that it should in fact extend to the end of this EST-rich region.

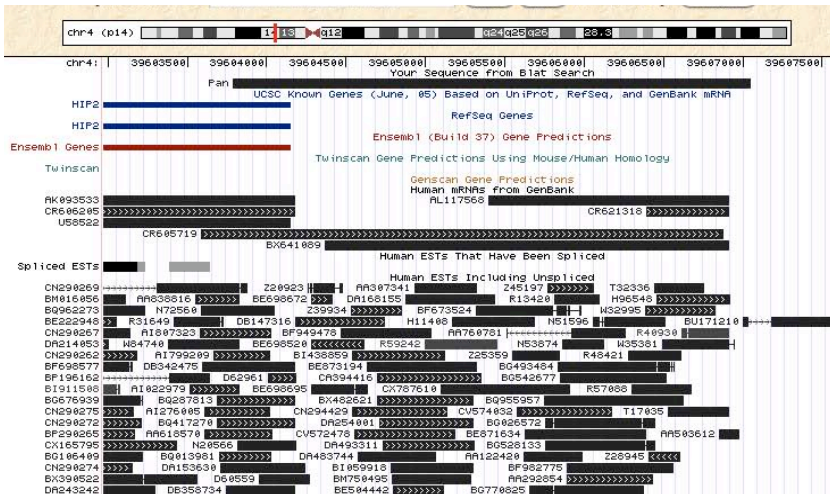


Figure 5
EST-rich Blat

Chimpchunk Map

The following is a map of the chimp region 2_4 and the three predicted genes discussed in this paper, as well as the two repeats found with RepeatMasker.