Fine Song
Bio 4342
April 2, 2006

Chimp Chunk 2.5 Annotation

Summary

A finished genome sequence has limited practicality unless its key features are elucidated. We attempted to better understand the process and principles behind characterizing key features of a genome by annotating chimp chunk 2.5. After creating a repeat-masked fasta file of our chimp sequence, we used GENSCAN, a gene prediction program, and various BLAST searches to identify features and determine possible functions of the genes in chimp chunk 2.5. The GENSCAN output in table and map format can be seen in Figs. 1 and 2, respectively.

Sequence Pan : 82926 bp : 38.07% 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	-	8	3	6							1.05
1.02	Term	-	3857	3315	543	2	0	38	37	352	0.333	18.38
1.01	Init	-	4299	4276	24	0	0	65	78	36	0.218	-1.35
1.00	Prom	-	8976	8937	40							-7.75
2.00	Prom	+	11023	11062	40							-6.45
2.01	Init	+	12453	12509	57	2	0	58	44	146	0.303	8.56
2.02	Intr	+	16506	16660	155	2	2	30	46	116	0.244	-0.45
2.03	Intr	+	22641	22755	115	0	1	34	86	67	0.289	0.63
2.04	Intr	+	50168	50329	162	1	0	77	84	244	0.998	22.15
2.05	Intr	+	52351	52444	94	0	1	31	110	103	0.999	5.42
2.06	Intr	+	52941	53006	66	1	0	82	116	8	0.656	0.96
2.07	Intr	+	53515	53666	152	2	2	74	105	91	0.993	8.36
2.08	Intr	+	54959	55165	207	1	0	120	39	185	0.993	15.15
2.09	Intr	+	58031	58135	105	1	0	66	99	67	0.966	5.09
2.10	Intr	+	60237	60418	182	2	2	82	59	157	0.999	9.94
2.11	Intr	+	61489	61582	94	1	1	25	94	92	0.986	2.55
2.12	Intr	+	62034	62090	57	2	0	107	115	-10	0.763	1.76
2.13	Intr	+	66812	66888	77	1	2	93	86	44	0.988	2.09
2.14	Term	+	67485	67662	178	0	1	98	48	138	0.772	6.98
2.15	PlyA	+	70653	70658	6							1.05

Figure 1. GENSCAN output table

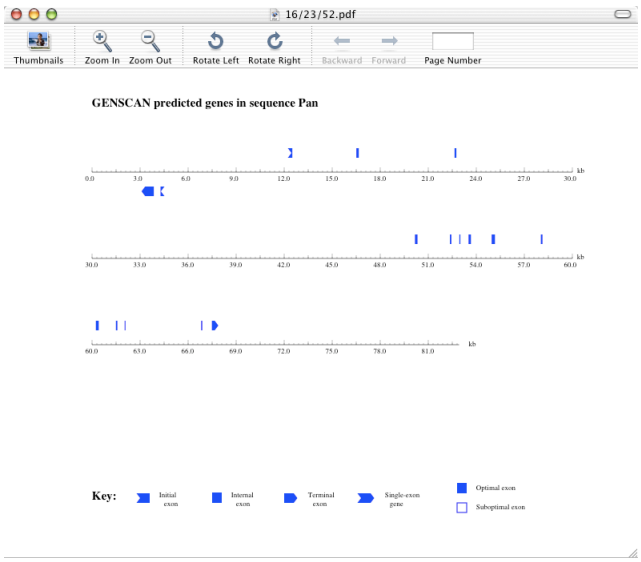


Figure 2. GENSCAN output map

On completion of our annotation, we found two gene-like features:

Feature	Location (b.p.)	Type of Feature	Related Function
1	3315-4299	Pseudogene	Heterogeneous Nuclear Ribonucleoprotein K
2	50168-76125	Gene	Apoptosis Inhibitor 5

Table 1. Features ofchimp chunk 2.5 and their location, type, and related function.

Feature 1 appears to be a pseudogene derived from the heterogeneous nuclear ribonucleoprotein K gene in humans (HNRPK, GenBank accession number NM_002140). HNRPK influences pre-mRNA processing and other aspects of mRNA metabolism and transport. It is thought to have a role during cell cycle progression. Feature 2 is a gene that encodes apoptosis inhibitor 5 (API5, GenBank accession number NM_006595), which prevents programmed cell death.

Our ~83kb (82926bp) chimp chunk has 40.65% GC and 44.86% (or 37203 b.p.) masked repeated base content (Fig. 3). There are ten significant repeat regions, defined in this case as non-Alu repeats spanning more than 500 bp:

Location (b.p.)	Length	Type of Repeat
12654-15729	1071	LTR/ERVK
26796-27598	802	LTR/ERV1
31121-31789	668	LTR/ERV1
32774-36323	2006	LINE/L1
35077-36323	1246	LINE/L1
56603-57858	1255	DNA/Mariner
63707-64656	949	LTR/ERV1
71521-72813	1292	LINE/L1
72868-73437	569	LINE/L1
77249-77838	589	LINE/L2

Table 2. Ten significant repeat regions in chimp chunk 2.5.

```
file name: pan_chunk2_5.fasta
sequences:             1
total length:      82926 bp  (79728 bp excl N-runs)
GC level:         40.65 %
bases masked:      37203 bp ( 44.86 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                68         15858 bp    19.12 %
      ALUs            42         11556 bp    13.94 %
      MIRs            26          4302 bp     5.19 %

LINEs:                13          7943 bp     9.58 %
      LINE1            6          6493 bp     7.83 %
      LINE2            7          1450 bp     1.75 %
      L3/CR1           0             0 bp     0.00 %

LTR elements:         22          9993 bp    12.05 %
      MaLRs           12          4287 bp     5.17 %
      ERVL             3           879 bp     1.06 %
      ERV_classI       6          3755 bp     4.53 %
      ERV_classII      1          1072 bp     1.29 %

DNA elements:         12          3351 bp     4.04 %
      MER1_type        5           948 bp     1.14 %
      MER2_type        6          1147 bp     1.38 %

Unclassified:          0             0 bp     0.00 %

Total interspersed repeats:    37145 bp    44.79 %


Small RNA:             1            88 bp     0.11 %
```

Figure 3. RepeatMasker summary table.

Feature 1

GENSCAN predicted a two-exon gene between 3315 bp – 4299 bp in the chimp chunk 2.5. The two-exon gene fell within a 2kb – 4kb region, where Herne output indicated alignments with human ESTs. When the GENSCAN-predicted protein was put into the NCBI BLASTp search, the results indicated that this feature is related to HNRPK. Looking at the BLAST alignments in detail, there was only ~60% amino acid homology between the GENSCAN-predicted gene and HNRPK amino acid sequences. When the predicted chimp protein sequence was used in a BLAT search of the human genome, the best match, a site on human chromosome 11, had only 95.1% identity, giving greater possibility for a pseudogene or a paralog rather than a gene. When we used the HNRPK protein in a BLAT search of the human genome, we found a site with 100% identity on human chromosome 9, rather than chromosome 11 (Fig. 4).

Human BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	YourSeq	1331	1	453	464	100.0%	9	+-	83814633	83822721	8089
browser details	YourSeq	1202	1	463	464	93.4%	3	++	97551088	97552477	1390
browser details	YourSeq	1043	1	458	464	87.8%	2	+-	136790434	136791805	1372
browser details	YourSeq	1024	2	458	464	87.1%	5	+-	126875068	126876428	1361
browser details	YourSeq	520	1	416	464	73.6%	11	+-	43240370	43241551	1182
browser details	YourSeq	98	1	51	464	81.1%	5	+-	14930875	14931026	152

Figure 4. Human BLAT results for HNRPK protein.

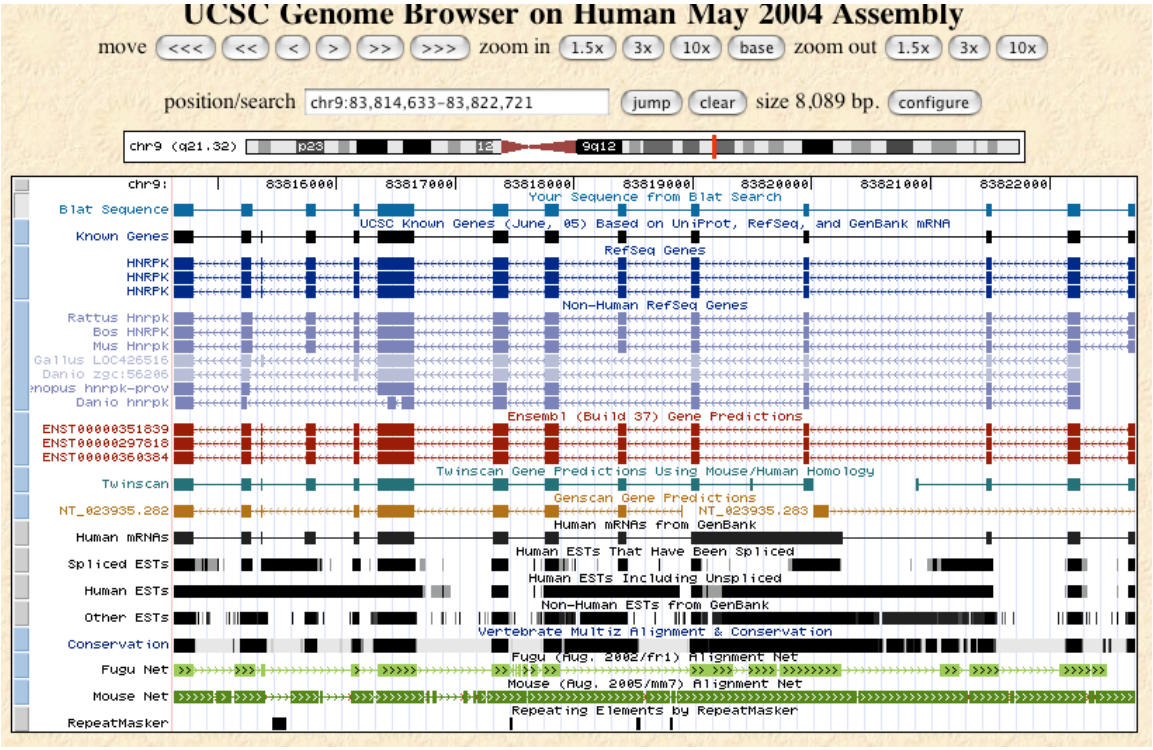


Figure 5. BLAT browser showing the gene for human HNRPK protein.

From the browser, it is clear that the HNRPK gene has 13 exons, which is far more than the two exons predicted by GENSCAN (Fig. 5). Additionally, upon examining the best match for the HNRPK protein sequence on chromosome 11 (73.6% sequence match, quite low), a stop codon was observed. Both of these observations confirmed our prediction that this is a pseudogene (Fig 6).

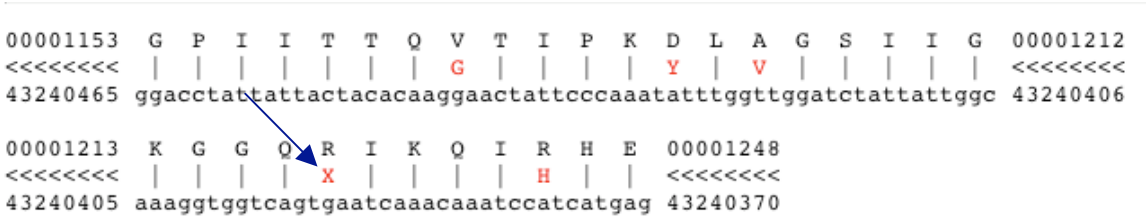


Figure 6. Presence of stop codon in HNRPK match for human chromosome 11.

To estimate the age of this pseudogene, we searched for the HNRPK gene in the mouse. The functional mouse ortholog for HNRPK was on chromosome 13 (100% sequence match) and the ortholog of the pseudogene was on chromosome 7 (98.9% sequence match). No stop codon was observed in the latter case. This data suggested that the stop codon mutation was introduced into the pseudogene after the split between the primate and rodent lineages.

Feature 2

GENSCAN predicted a 14-exon gene between 12509 bp –67662 bp. Both the BLAST and human BLAT alignments for this predicted protein were impressive, with an E-value of 0.0 and 93% amino acid homology with apoptosis inhibitor 5 (API5), and 99.9% amino acid identity with human chromosome 11, respectively. When the actual API5 protein sequence from humans was used in a BLAT search of the human genome, the gene identified showed 14 exons as well. Additionally, a BLAST2 match between a chimp chunk fragment 50168-76125 and human API5 showed 98-100% nucleotide homology.

However, the distribution of the exons was a bit troublesome to annotate for two main reasons. First, the Herne output shows no human EST matches in the region between 4 kb and 50 kb, which is where three of the exons in the series (11023-11062, 12453-12509, 16506-16660) fall (Fig. 7). Since EST matches usually indicate a high degree of conservation, the lack of EST matches indicates a lack of conservation, which might suggest that the prediction of these three exons is not very reliable. Second, there is an EST-rich region between 68kb and 76kb that was not predicted by GENSCAN to contain any exons (Fig. 8).

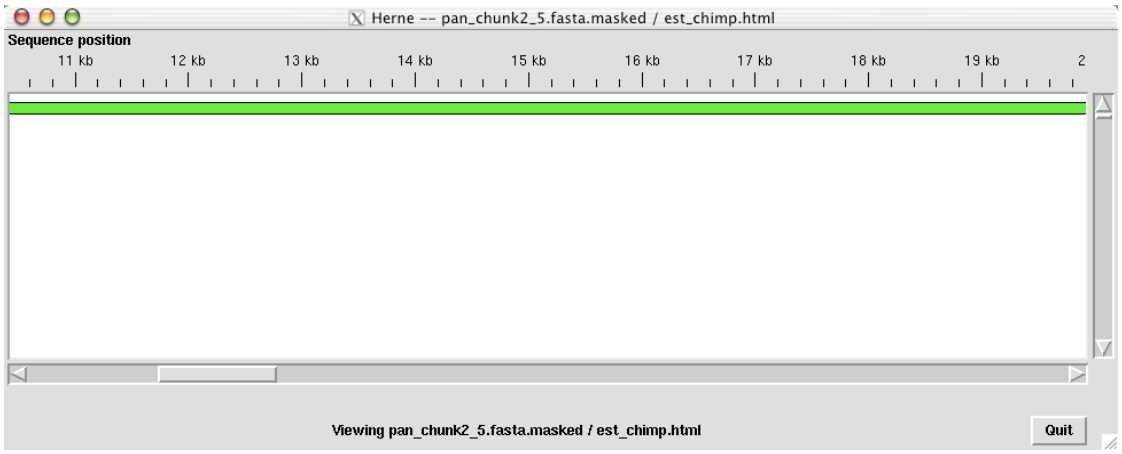


Figure 7. Herne output showing no human EST matches to parts of feature 2.

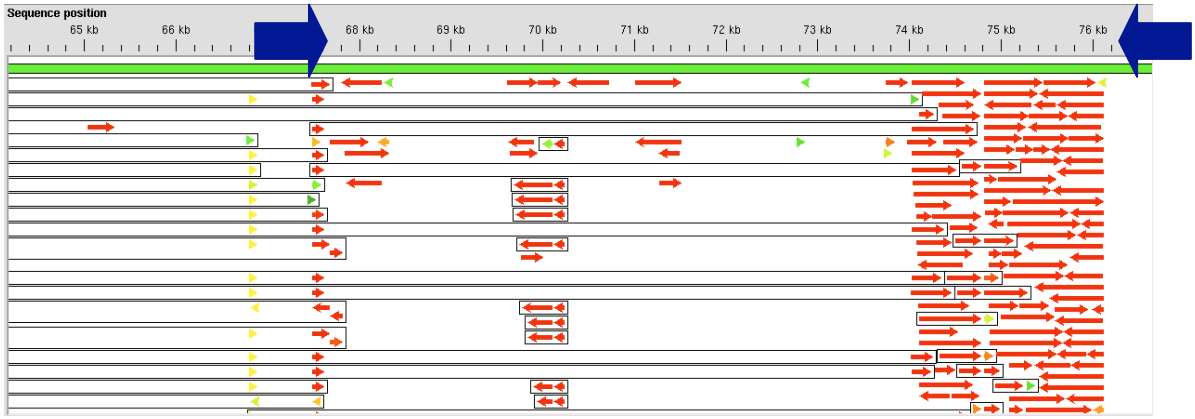


Figure 8. Herne output with significant human EST matches to parts of feature 2.

Considering these observations, perhaps the three exons in the EST-free 4kb - 50kb region can be deemed to be mistakes made by GENSCAN, since GENSCAN only predicts correctly 20-25% of mammalian genes. Looking at the alignment of the predicted protein to the human API5 gene (BLAT search), the region that seems to be missing from the predicted protein might be found among sequences from the 68-76kb region (Fig. 9).

Most likely the 68-76kb region correlates with the terminating exon and/or UTR, but was not predicted by GENSCAN due to the gene predictor's limitations. The human API5 gene has a 3' untranslated region (UTR) right next to a terminating exon (Fig. 10). Since UTRs can have regulatory function, the fact that the region has high conservation, as indicated by good quality EST evidence, seems consistent with our observations.

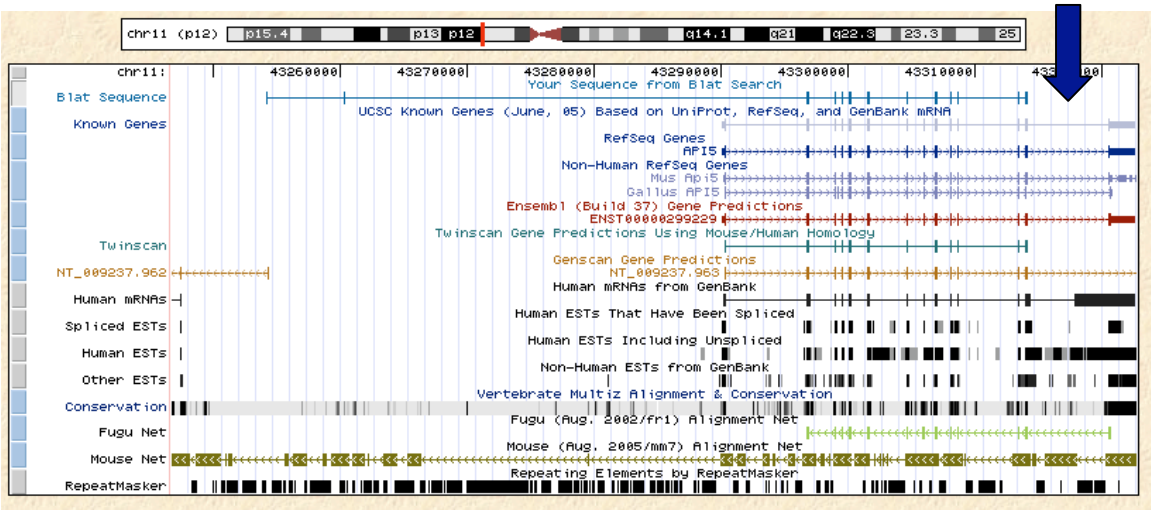
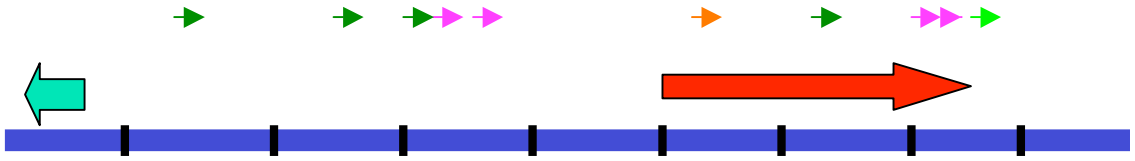


Figure 9. BLAT browser showing GENSCAN-predicted gene for Feature 2.

Figure 10. Map of human API5 gene from NCBI.

Overall, since 11 of the 14 exons predicted by GENSCAN in the chimp DNA matched perfectly with human API5 exons, and that the discrepancies that were present occurred in the less conserved 5' and 3' ends, it is still safe to characterize this second feature as encoding the API5 gene. Additionally, the BLAT search indicated a high percent identity to confirm that this portion of chimp chunk 2.5 is a human ortholog, with the best orthologous gene being the API5, as predicted by GENSCAN, Ensembl, and Twinscan and confirmed by RefSeq. API5 falls within 50168 bp –76125 bp in chimp chunk 2.5.

Final Map of Chimp Chunk 2.5



Feature 1 -Pseudogene
Feature 2 -Gene
Line/L1
Line/L2
LTR/ERV
DNA/Mariner