Annotation of Chimp Chunk 2-7

Introduction:

My partner, Michelle Miller, and I have annotated Chimp Chunk 2-7 in preparation for annotating a segment of the *Drosophila mohavensis* genome. We were given GENSCAN (tool that predicts the location of genes in a given sequence) output for our Chimp Chunk. In order to discover the true identity of the genes predicted by GENSCAN, we analyzed a series of sequence alignments using BLAST (Basic Local Alignment Search Tool), BLAT (BLAST – Like Alignment Tool) and ClustalW (a multiple sequence alignment program).

GENSCAN 1.0 Date run: 2-Apr-107 Time: 22:00:30

Sequence chunk2-7 : 80323 bp : 40.96% 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..
1.01	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	0.757	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	0.697	22.73
4.02	Intr	-	77852	77657	196	2	1	31	57	136	0.140	2.87
4.01	Init	-	78129	77992	138	0	0	49	33	283	0.999	17.09

Fig 1: GENSCANW Output for Chimp Chunk 2-7

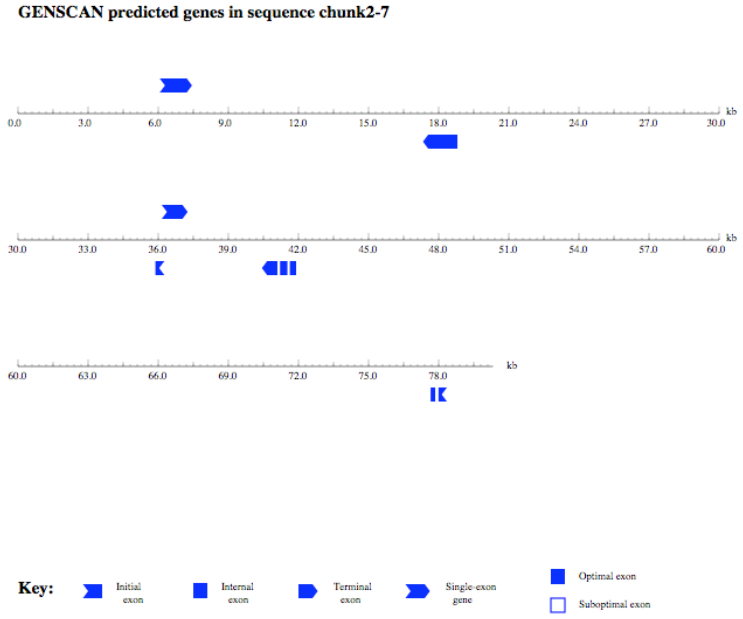

Fig 2: Initial GENSCAN Gene Map

Results:

Predicted gene number	Location in Chimp (bp)	Number of exons predicted	Type of feature and related function
1	6288-8932	1	Pseudogene: RNA binding motif---X-linked
2	10182-36147	3 estimated by GENSCAN, but in reality 2 exons. Explained later	Gene ortholog: spermatogenic leucine zipper 1 gene
3	36292-37603	1 estimated by GENSCAN, but in reality 2 exons. Explained later	Pseudogene: beta-actin. Structural support for the cell and key player in muscular contraction
4	40034-77992	5	Pseudogene: keratin18. Keratin is a high sulfur matrix protein used in feathers scales, hair, nails, etc.

Table 1: Genes Predicted by GENSCAN

Summary of Predicted Gene 1 and 2:

My partner, Michelle Miller, analyzed the first two predicted genes. She found that predicted gene 1 is a pseudogene of the RNA binding motif, X-linked 2 gene. Evidence for this was gathered from BLASTp, BLAT to both humans and mice, and ESTs in the region. Another strong indication supporting the hypothesis that this is a pseudogene in humans came from the annotation on the Entrez Gene web data base. This information increases our confidence in the liklihood that this is also a pseudogene in chimpanzees. Finally, this pseudogene arose after rodents and primates diverged.

The second feature is the orthologue to the human spermatogenic leucine zipper 1 gene. Evidence for this came from BLASTp, BLAT, and ESTs in the region. The strongest supporting evidence came from the annotation on the Entrez Gene web data base. This gene is one exon shorter than the predicted Feature 2 peptide sequence, as shown by Clustal W analysis, and further evidence was gathered to show that the erroneous exon actually belongs to the beta-actin pseudogene.

Methods for Predicted Gene 3:

After the initial analysis of the GENSCAN output, we already believe that predicted gene 1 (pg-1) and predicted gene 3 (pg-3) are probably not real functional genes, because they are predicted to only have one exon. “Only about 7% of known human genes have a single exon” (Chimp BAC analysis). Considering that chimps and humans have 98-99% gene similarity, it is expected that this statistic holds true for chimps as well.

I began by identifying possible orthologs of pg-3 by using the BLASTp tool to align my amino acid sequence from pg-3 (Fig 3) against the conserved domain data-base. The BLASTp predicted that pg-3 is a gene belonging to the actin family. After reviewing the full BLASTp report I found that the human reference sequence that had the highest e-value (Fig 4) for the alignment was a beta actin reference (NCBI accession number: NP_001092.1). Human beta actin is a 375 amino acid protein, while pg-3 only contains 219 amino acids. In addition, there is only an 84% positives[1] identity; for a functional gene I would expect around a 97%-99% positives match since humans and chimps have 98% similar DNA in coding regions. There is enough similarity between pg-3 and human beta actin to believe that pg-3 is more than a random open reading frame that happened to have some similarity. Yet, the similarity is not strong enough to be a functional beta actin gene. This leads me to believe that pg-3 is a pseudogene of beta actin, but at this point, more evidence is needed.

```
>chunk2-7|GENSCAN_predicted_peptide_3|219_aa
MGSPTLCPSMKGTPLPHTILCLDLAGRNLTDYLMKILTQCGYSFTATVMQEIVCDIKKKL
CCIPLDFEQETAMVGSSSSLEKSYKLPNGQVITISNKWFCCPEALFQTSFVGMESCGIHE
TTFNSIMKSDVDIYKDLYANAVLSGSTTMYPSITNRMQKEITALAPSAMKIKITAPPECK
YSVWIRGSILASLSTFQQMWISKQEYNKSGPSIVHGKCF
```

Fig 3: Amino Acid Sequence of Predicted Gene 3

GENE ID: 60 ACTB | actin, beta [Homo sapiens] **(Over 100 PubMed links)**

Score = 343 bits (881), Expect = 3e-93, Method: Compositional matrix adjust.
Identities = 167/213 (78%), Positives = 180/213 (84%), Gaps = 0/213 (0%)

```
Query  7    CPSMKGTPLPHTILCLDLAGRNLTDYLMKILTQCGYSFTATVMQEIVCDIKKKLCCIPLD  66
            P  +G  LPH IL LDLAGR+LTDYLMKILT+ GYSFT T  +EIV DIK+KLC + LD
Sbjct  163  VPIYEGYALPHAILRLDLAGRDLTDYLMKILTERGYSFTTTAEREIVRDIKEKLCYVALD  222

Query  67   FEQETAMVGSSSSLEKSYKLPNGQVITISNKWFCCPEALFQTSFVGMESCGIHETTFNSI  126
            FEQE A   SSSSLEKSY+LP+GQVITI N+ F CPEALFQ SF+GMESCGIHETTFNSI
Sbjct  223  FEQEMATAASSSSLEKSYELPDGQVITIGNERFRCPEALFQPSFLGMESCGIHETTFNSI  282

Query  127  MKSDVDIYKDLYANAVLSGSTTMYPSITNRMQKEITALAPSAMKIKITAPPECKYSVWIR  186
            MK DVDI KDLYAN VLSG TTMYP I +RMQKEITALAPS MKIKI APPE KYSVWI
Sbjct  283  MKCDVDIRKDLYANTVLSGGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIG  342

Query  187  GSILASLSTFQQMWISKQEYNKSGPSIVHGKCF  219
            GSILASLSTFQQMWISKQEY++SGPSIVH KCF
Sbjct  343  GSILASLSTFQQMWISKQEYDESGPSIVHRKCF  375
```

Fig 4: BLASTp Alignment: Chimp pg-3 vs. Human Beta Actin

Looking at the PubMed article on beta actin, I found that the gene for human beta actin is located on chromosome 7 (specifically 5536747-553312) and has six exons, while pg-3 only has one exon.

Next, I used the BLAT tool to align pg-3 with the human genome, using the most current version available (March 2006). The results (Fig 5) showed that the highest identity match (95%) was to a segment on chromosome 5 (79631256 - 79631912). There was also a match to chromosome 7 listed, but the identity was only 80.5%. This was expected since the identity given by the BLASTp results was only 84%. Opening the browser for the chromosome 5 match (Fig 6), I found that there were no reference sequence genes nor any spliced human ESTs. Finding a reference sequence would have meant that there was a gene in that location that had been found previously. Finding some ESTs would have meant that there was mRNA transcription previously found in the area. Finding neither gave me more confidence that pg-3 is a pseudogene. I decided to also look at the browser for the chromosome 7 match (Fig 7). It shows 5 exons with reference sequences and many EST matches, as would be expected from a functional beta actin gene.

Human BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	219_aa	591	1	219	219	95.0%	5	+-	79631256	79631912	657
browser details	219_aa	373	15	219	219	80.5%	7	+-	5533908	5534729	822
browser details	219_aa	373	15	219	219	80.5%	17	+-	77092314	77093100	787
browser details	219_aa	345	15	219	219	78.1%	2	++	131131303	131131917	615
browser details	219_aa	339	15	219	219	77.6%	14	+-	19057714	19058328	615
browser details	219_aa	333	15	219	219	77.1%	2	++	132100692	132101306	615
browser details	219_aa	332	15	205	219	79.2%	1	+-	222118099	222118667	569
browser details	219_aa	327	15	219	219	76.6%	2	+-	130936862	130937476	615
browser details	219_aa	327	15	219	219	76.6%	14	++	18655308	18655922	615
browser details	219_aa	323	15	219	219	77.3%	1	+-	227633872	227634739	868
browser details	219_aa	319	15	217	219	77.0%	6	++	46281182	46281788	607
browser details	219_aa	318	15	218	219	76.0%	2	+-	130548293	130548904	612
browser details	219_aa	315	15	219	219	75.7%	2	++	131738109	131738723	615
browser details	219_aa	306	15	206	219	76.6%	22	+-	14635391	14635966	576

Fig 5: Top Portion of BLAT Results: query = pg-3; database = human genome

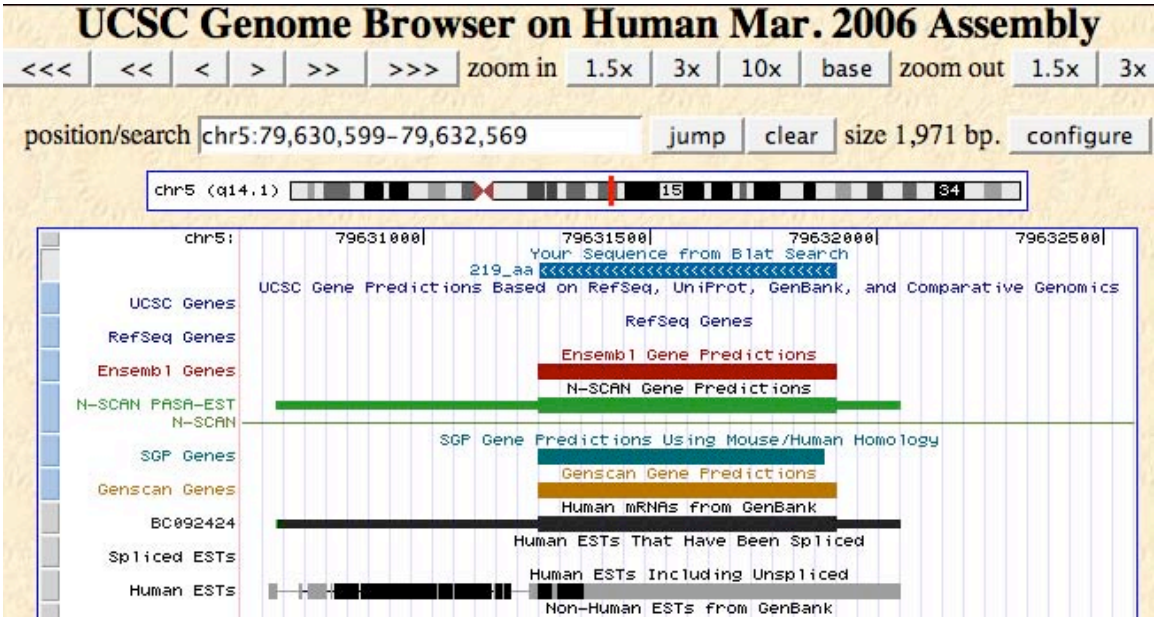

Fig 6: BLAT Browser Chromosome 5

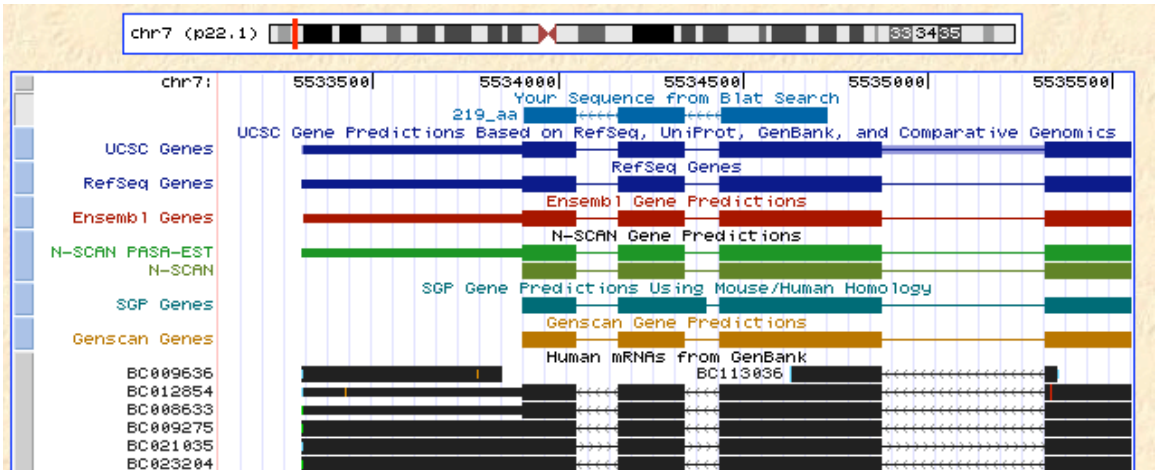

Fig 7: BLAT Browser Chromosome 7

Knowing that a pseudogene for human beta actin may have been previously annotated, I searched the NCBI data-base for an entry for humans on chomosome 5 between base 79631256 and base 79631912. I found an entry for a cytoplasmic beta-actin pseudogene [Homo sapien]

(Fig 8 and Fig 9) with a pseudogene located at 79634422-79631253 and having 2 exons. This was almost exactly the location found by the BLAT search results, which increased my confidence that pg-3 is a pseudogene. Yet, why was there a discrepancy in the number of exons that made up the pseudogene? I was informed by my partner, Michelle that the last exon in pg-2 actually belonged to pg-3. GENSCAN predicted that the third exon in pg-2 is in the negative orientation. Yet, when I aligned the third exon to the human genome using BLAT, I found that it aligns to beta actin in the positive orientation. This matches the rest of pg-3, which is also in the positive orientation. That information solved the discrepancy and allowed me to make the final conclusion that pg-3 is in-fact a pseudogene in chimps homologous to the pseudogene in humans located on chromosome 7 at 79634422-79631253.

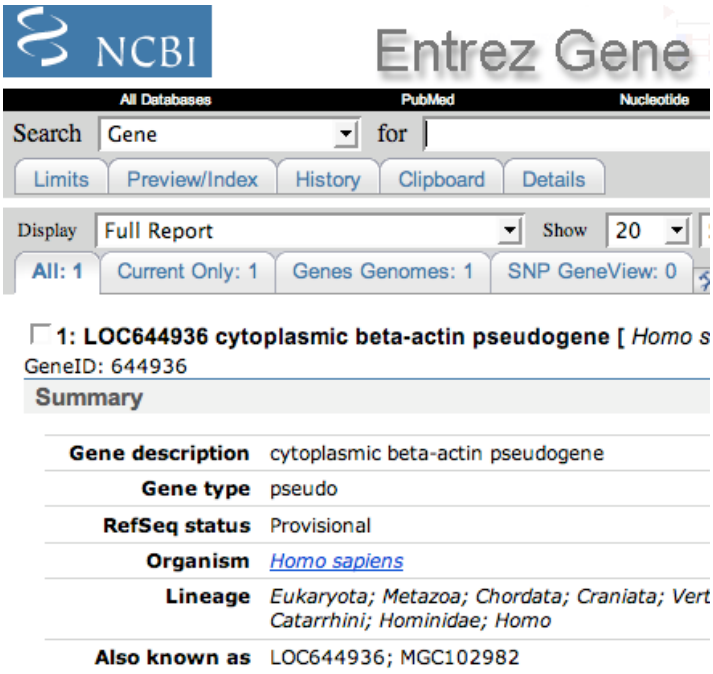

Fig 8: NCBI entry for beta-actin pseudogene

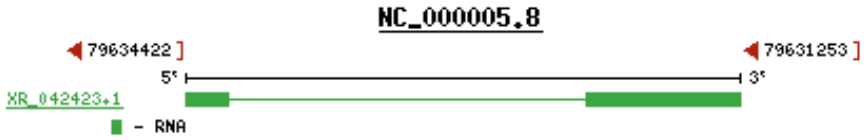

Fig 9: Human Beta-Actin Pseudogene

Having identified pg-3, I wanted to find out whether the pseudogene had evolved before or after the evolutionary divergence of rodents and primates. I did not initially know which chromosome in mice a beta-actin pseudogene would be located on. To find this information, I retrieved the FASTA formatted protein sequence of the human beta-actin pseudogene and aligned it to the mouse genome using BLAT with the mouse net feature on. Looking at the detailed Mouse Net output I found that chromosome 5 in humans is syntenic to chromosome 5 in mice. For this reason, I was expecting to find that if the pseudogene existed in mice, it would be on chromosome 5.

Mouse (July 2007/mm9) Alignment Net

View alignment details of parts of net within browser
Open Mouse browser at position corresponding to th

Type: nonSyn
Level: 2
Human position: chr5:79630674-79634434
Mouse position: chr5:143664807-143667377
Strand: +
Score: 102,743
Chain ID: 7366
Bases aligning: 1,686
Mouse bases duplicated: 2,571
Human N's: 0 (0.0%)
Mouse N's: 0 (0.0%)

Fig 10: Mouse Net Data

My next step was to use BLAT to align both pg-3 and a modified version of pg-3 to the most recent version of the mouse genome (July 2007) available. The modified pg-3 is the original amino acid sequence of pg-3 given by GENSCAN, with the last exon of pg-2 attached to the front in order to complete the beta-actin pseudogene. The BLAT results for both had the best match to chromosome 13 in mouse. The original pg-3 has an 81% identity and the modified pg-3 (mpg-3) has an 82% identity to a segment of mouse chromosome 13. This confirmed again that the end of pg-2 really did belong in pg-3, but it was strange that the best identity between pg-3 and mice was on chromosome 13 instead of chromosome 5. Interestingly, the alignment of the human beta-actin and mouse genome showed a 100% identity on chromosome 5 as well as many supporting EST matches. While I'm not sure why or how this happened, it seem as though there is potential that this could be the beta-actin pseudogene or another related peudogene is present in mice as well as primates. Therefore, the evolution of the beta-actin pseudogene likely occurred before the branching of rodents and primates.

Mouse BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	219_aa	381	15	219	219	81.0%	13	++	81204185	81204799	615
browser details	219_aa	375	15	219	219	80.5%	4	++	103236517	103237131	615
browser details	219_aa	373	15	219	219	80.5%	5	+-	143665480	143666314	835
browser details	219_aa	373	15	219	219	80.5%	11	+-	120207707	120208516	810
browser details	219_aa	369	15	219	219	80.0%	8	++	47314857	47315471	615

Fig 11: BLAT Results of pg-3 vs. Mouse Genome

Mouse BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	chunk2-7 gene	427	117	353	353	82.0%	13	++	81204059	81204799	741
browser details	chunk2-7 gene	421	117	353	353	81.5%	4	++	103236391	103237131	741
browser details	chunk2-7 gene	419	117	353	353	81.5%	5	+-	143665480	143666440	961
browser details	chunk2-7 gene	419	117	353	353	81.5%	11	+-	120207707	120208642	936
browser details	chunk2-7 gene	395	117	353	353	79.3%	6	+-	133711466	133712200	735

Fig 12: BLAT Results of Modified pg-3 vs. Mouse Genome

Methods for predicted gene 4:

The methods used to annotate pg-4 and the issues that arose were very similar to those of pg-3. I started with a BLASTp alignment of pg-4 (Fig 13) to the conserved domain data-base. BLASTp

predicted that pg-4 was a gene in the Filament family. Looking at the full results we see that the best human reference sequence match (Fig 14 and Fig 15) corresponds to keratin 18 (NCBI accession number NP_000215.1). From the BLASTp data, I noted that the keratin 18 amino acid sequence has 430 residues, while pg-4 has 451 residues. This may be the result of a mistake made by GENSCAN, over-predicting the size of the gene. I also observed that there is only a 64% positives identity between pg-4 and keratin 18. From this information, it is unlikely that pg-4 is a functional keratin gene in chimps.

```
>chunk2-7|GENSCAN_predicted_peptide_4|451_aa
MPKARSLAPRAARALLAASRLERWAQPRLTLERRSASAERLGAEAWAGGLLRSLPGWGAE
LRSRPPAFGPHAPGLVRGLAALGPSPNESAAPLRLRIRRKSLLLPLKGWSLNSMSFTTRS
TTSSTNYWSLGSVQLPSYVAQLVSSVVSVYAGAGGSGFRISVSHSTSFWGGLGDLVGIGD
IQNEKETMQGLNDCLASYLDRTIEDLRVQIFASTVDSACIILQIDKAHITADDFRVKCET
ELAMCQSVESDIHGLRKSTDDTNVTQLQLEAEIEALKEELLFMKKTHEEEVKGLQAQIAS
SGLTMETEESTKQSAEIGASEIMLMELRHTLQSLEINLNSMRNLKARLENSLREVETRYA
MQMEQLNRVQLHLKLKLAQTWAEGQHQVQEYEALLNIKIKLEAEITTYHHLLEDEEGFNP
GDALDSSNSIQSIQKTTTHRIVDSIGGPPGV
```

Fig 13: Amino Acid Sequence of Predicted Gene 4

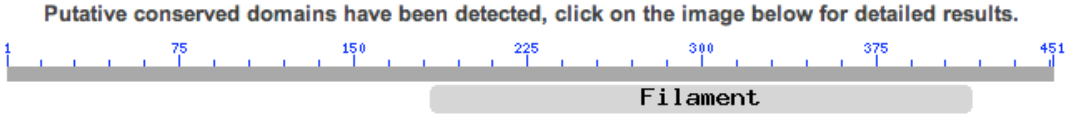

Fig 14: Blast Match of Predicted Gene 4 to the Non-redundant Database

```
Length=430

 GENE ID: 3875 KRT18 | keratin 18 [Homo sapiens] (Over 10 PubMed links)

 Score =  454 bits (1168),  Expect = 6e-126, Method: Compositional matrix adjust.
 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
```

Fig 15: predicted gene 4 vs. Keratin 18; query = human keratin 18; subject = pg-4. Keratin 18 only aligned to the latter half of pg-4 in all alignments. Further BLAST analysis of the first 113 amino acids using a lower expect value are needed to determine its identity.

Checking the PubMed entry for human Keratin 18 (KRT18), I found that there are two isoforms of the gene, one containing 7 exons and one containing 8 exons (Fig 16); pg-4 only has 5 exons. KRT18 is located on chromosome 12 (51628922 – 51632952).

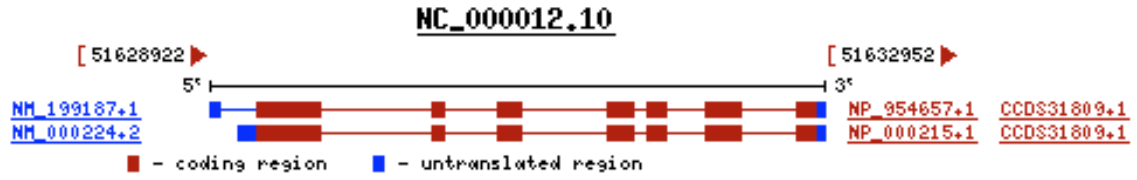

Fig 16: Map of Human Keratin 18

BLAT was the next alignment tool that I used to match pg-4 to the human genome. The results (Fig 17) showed a 96.3% identity matching 448 of the 451 residues on chromosome 5 (the same chromosome on which the human beta-actin pseudogene is located). BLAT also showed a match to chromosome 12 where KRT18 is located, with 79% identity. Looking at the browser (Fig 18) for chromosome 5, there were no reference sequence matches and no significant spliced EST matches. This is significant evidence that pg-4 is a pseudogene.

Human BLAT Results

BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	451_aa	1236	1	448	451	96.3%	5	++	79587509	79621246	33738
browser details	451_aa	491	129	443	451	78.6%	20	++	48006729	48007928	1200
browser details	451_aa	467	125	443	451	76.9%	2	+-	189884279	189885479	1201
browser details	451_aa	437	173	443	451	77.5%	11	+-	35838418	35839457	1040
browser details	451_aa	431	173	443	451	77.1%	17	+-	23627245	23628286	1042
browser details	451_aa	426	122	443	451	74.9%	20	+-	22661485	22662706	1222
browser details	451_aa	420	126	443	451	73.7%	17	++	56242342	56243557	1216
browser details	451_aa	417	110	443	451	76.7%	14	+-	35050827	35052058	1232
browser details	451_aa	407	173	442	451	75.6%	4	++	117061498	117062532	1035
browser details	451_aa	401	148	443	451	76.7%	18	+-	9668288	9669432	1145
browser details	451_aa	399	122	414	451	78.8%	2	++	65747128	65748243	1116
browser details	451_aa	377	143	443	451	73.4%	6	+-	112789380	112790537	1158
browser details	451_aa	372	173	443	451	77.4%	1	++	236721481	236722522	1042
browser details	451_aa	371	122	443	451	74.7%	5	++	36921086	36922308	1223
browser details	451_aa	368	122	443	451	73.1%	22	+-	19167150	19168371	1222
browser details	451_aa	359	173	421	451	74.7%	11	+-	5971796	5972762	967
browser details	451_aa	343	122	418	451	79.6%	12	++	51629246	51632391	3146
browser details	451_aa	341	173	438	451	76.7%	12	+-	103894719	103895732	1014
browser details	451_aa	341	125	443	451	71.2%	12	+-	64098817	64100036	1220

Fig 17: BLAT Results for pg-4 vs. Human Genome

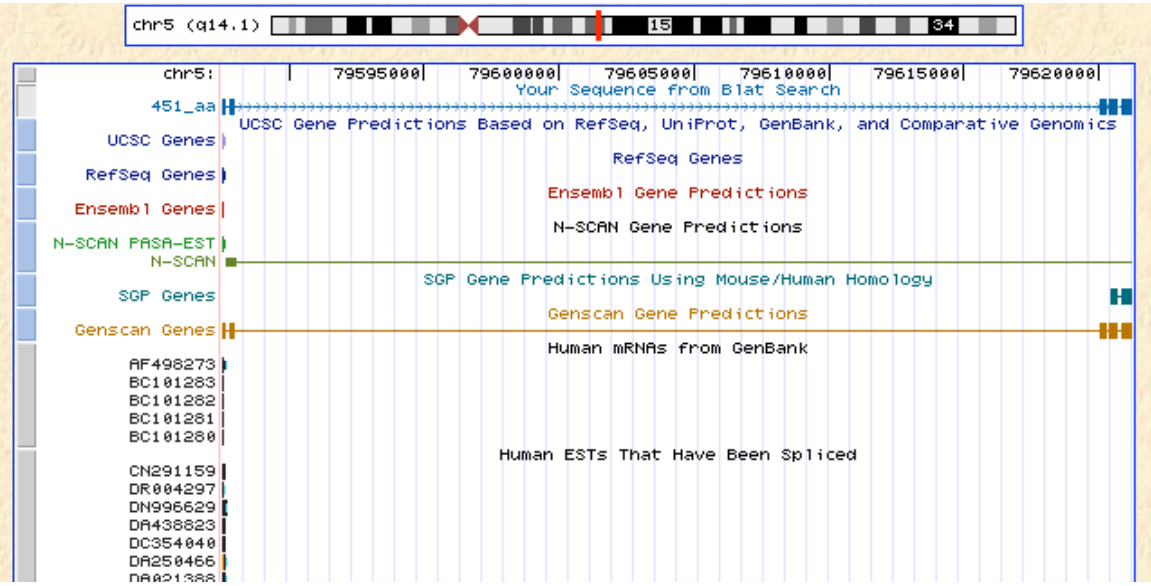

Fig 18: Browser for Chromosome 5 Match

Next I used ClustalW to align pg-4 protein sequence to KRT18 protein sequence in order to better visualize the differences between the two sequences. The ClustalW output made it easy to see that there were two methionine start codons in the predicted sequence. This explains why pg-4 has a longer sequence than KRT18. This is probably the result of GENSCAN mispredicting an early start site for this gene.

```
CLUSTAL 2.0.3 multiple sequence alignment


predicted       MPKARSLAPRAARALLAASRLERWAQPRLTLERRSASAERLGAEAWAGGLLRSLPGWGAE 60
human           ------------------------------------------------------------


predicted       LRSRPPAFGPHAPGLVRGLAALGPSPNESAAPLRLRIRRKSLLLPLKGWSLNSMSFTTRS 120
human           -------------------------------------------------------MSFTTRS 7
                                                                       *******


predicted       TTSSTNYWSLGSVQLPSYVAQLVSSVVSVYAGAGGSGFRISVSHSTSFWGGLG------- 173
human           TFS-TNYRSLGSVQAPSYGARPVSSAASVYAGAGGSGSRISVSRSTSFRGGMGSGGLATG 66
                * * *** ****** *** *: ***..********** *****:**** **:*


predicted       ---DLVGIGDIQNEKETMQGLNDCLASYLDRT--------------------------- 202
human           IAGGLAGMGGIQNEKETMQSLNDRLASYLDRVRSLETENRRLESKIREHLEKKGPQVRDW 126
                   .*.*:*.*********.*** *******.


predicted       ------IEDLRVQIFASTVDSACIILQIDKAHITADDFRVKCETELAMCQSVESDIHGLR 256
human           SHYFKIIEDLRAQIFANTVDNARIVLQIDNARLAADDFRVKYETELAMRQSVENDIHGLR 186
                      *****.****.***.* *:****:*:::******* ****** ****.******


predicted       KSTDDTNVTQLQLEAEIEALKEELLFMKKTHEEEVKGLQAQIASSGLTMETEES------ 310
human           KVIDDTNITRLQLETEIEALKEELLFMKKNHEEEVKGLQAQIASSGLTVEVDAPKSQDLA 246
                *  ****:*:****:**************.******************:*.: .


predicted       -----------------------------------TKQSAEIGASEIMLMELRHTLQSL 334
human           KIMADIRAQYDELARKNREELDKYWSQQIEESTTVVTTQSAEVGAAETTLTELRRTVQSL 306
                                                  *.****:**:*  * ***:*:***


predicted       EINLNSMRNLKARLENSLREVETRYAMQMEQLNRVQLHLKLKLAQTWAEGQHQVQEYEAL 394
human           EIDLDSMRNLKASLENSLREVEARYALQMEQLNGILLHLESELAQTRAEGQRQAQEYEAL 366
                **:*:******* *********:***:****** : ***: :**** ****:*.******


predicted       LNIKIKLEAEITTYHHLLEDEEGFNPGDALDSSNSIQSIQKTTTHRIVD----SIGGPPG 450
human           LNIKVKLEAEIATYRRLLEDGEDFNLGDALDSSNSMQTIQKTTTRRIVDGKVVSETNDTK 426
                ****:******:**::**** *.** *********:*:******:****    *  . .


predicted       V--- 451
human           VLRH 430
                *
```

Fig 19: ClustalW Output from pg-4 vs. KRT18

Other important information about Chimp Chunk 2-7:

When annotating a segment of a genome, sometimes there are interesting features to be noted outside of the genes predicted by one's program of choice. One important aspect of this Chimp Chunk is whether or not the sequence contains significant runs of repeats. Given a masked fasta file, we were able to run the DNA sequence for Chimp Chunk 2-7 through a program called repeat masker, which looks for and returns a list of repetitive segments in the sequence (Fig 20). Repeat masker found 10 significant (over 500bp in length) repeats. The summarized data can be found in table 2.

In addition to looking for repeats we also checked for matches between our DNA sequence and the human EST data base in order to see if there was any information we could gather in that way which was not available through Genescan. To do this we ran our given unmasked fasta file containing the DNA sequence against a human EST data base and created a file containing the alignment. Next a program called Herne was used to visualize the alignment results. By analyzing the results we were able to find that there is a feature (ortholog to human ribosomal

protein L39 pseudogene) between 23kb and 24kb that was missed by GENSCAN. We could also see the ESTs for two exons that GENSCAN predicted to be part of pg-4, but did not actually belong to that feature (location: 78kb).

```
==================================================
file name: pan_chunk2_7.fasta
sequences:             1
total length:      80323 bp  (78446 bp excl N/X-runs)
GC level:         43.79 %
bases masked:      50949 bp ( 63.43 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:              121        29038 bp   36.15 %
      ALUs          113        28097 bp   34.98 %
      MIRs            8          941 bp    1.17 %

LINEs:               28        15954 bp   19.86 %
      LINE1          25        15136 bp   18.84 %
      LINE2           3          818 bp    1.02 %
      L3/CR1          0            0 bp    0.00 %

LTR elements:        12         5199 bp    6.47 %
      MaLRs           4          823 bp    1.02 %
      ERVL            0            0 bp    0.00 %
      ERV_classI      8         4376 bp    5.45 %
      ERV_classII     0            0 bp    0.00 %

DNA elements:        6          785 bp    0.98 %
      MER1_type      1           69 bp    0.09 %
      MER2_type      2          269 bp    0.33 %

Unclassified:        0            0 bp    0.00 %

Total interspersed repeats:    50976 bp   63.46 %

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 %
==================================================
```

Fig 20: Repeat Masker Output

Position (bp) Start	End	Repeat Type/Class	Length
24478	25863	LINE/L1	1385
28273	29589	LTR/ERV1	1316
50903	52012	LINE/L1	1109
15325	16186	LINE/L1	861
5125	5765	LINE/L1	640
10292	10911	LINE/L1	619
73781	74395	LTR/ERV1	614
11083	11690	LINE/L1	607
29860	30399	LTR/ERV1	539
1330	1839	LINE/L1	509

Table 2: Summary of significant repeats

Conclusion:

Predicted gene 3 and predicted gene 4 were both found to be psudogenes that are homologous to human pseudogenes for beta actin and keratin 18 respectively, located on chromosome 5. I was able to deduce this by using the protein sequence given by GENSCAN and aligning it with the conserved domains data base using BLASTp to find the identity of a similar gene. Second, by following links to PubMed articles I found that beta actin is located on chromosome 7 and keratin 18 is located on chromosome 12. Third, BLAT was used to find the best alignment of both predicted genes to the entire human genome (not only the conserved domains as in the BLASTp alignment). This showed that both pg-3 and pg-4 had over 94% identity to pseudogenes for beta-actin and keratin 18. Both pseudogenes are located on chromosome 5 in humans. The lack of reference sequences and spliced ESTs in this area supports the hypothesis that pg-3 and pg-4 are pseudogenes. It is likely that the beta-actin pseudogene evolved before the divergence of rodents and primates based on BLAT alignment evidence.

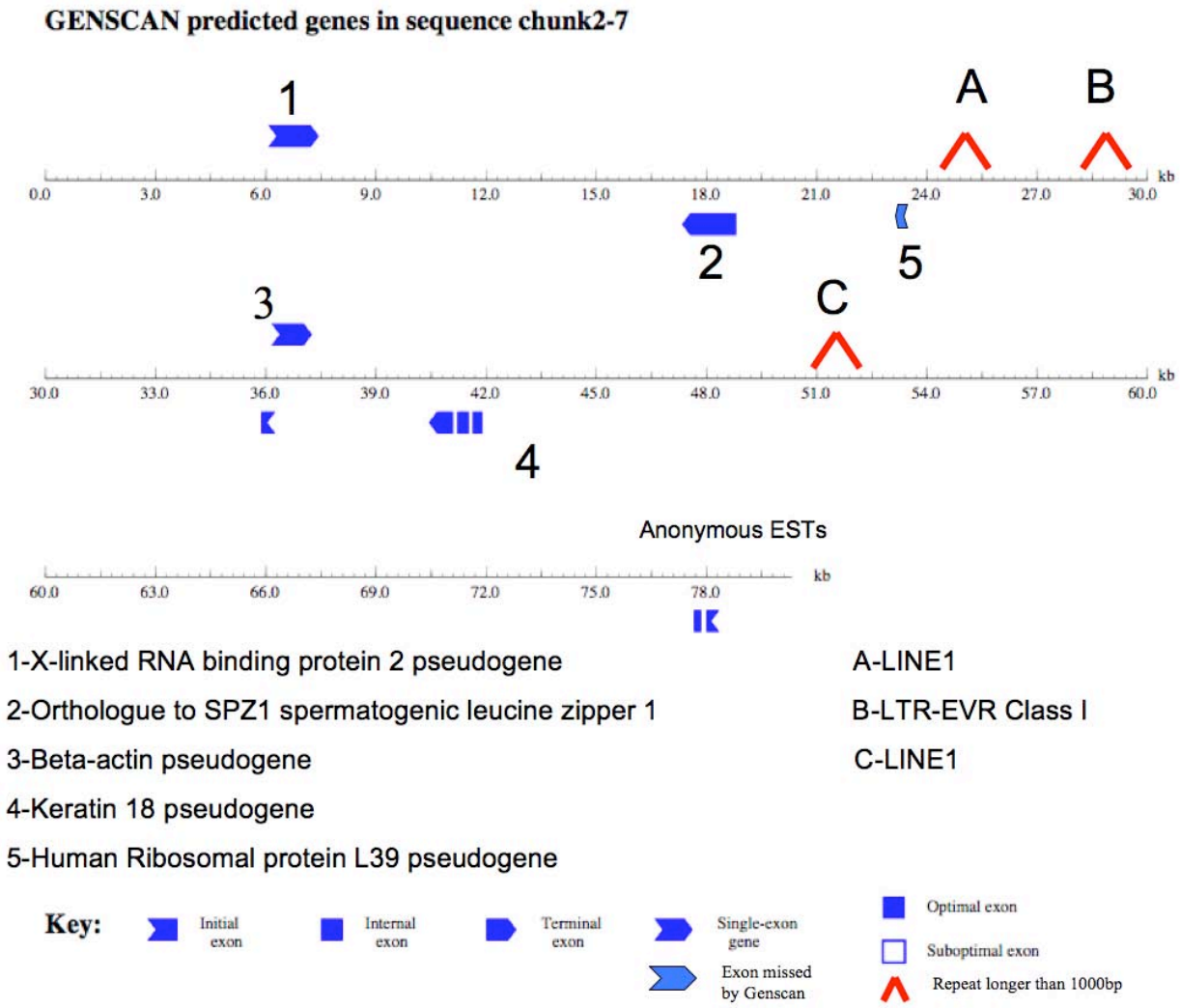

Fig 21: Final Map of Chimp Chunk 2-7: The ESTs near 78kb are the only two anonymous ESTs that were found.

Glossary

1. Positives identity: the percent of amino acids that are either the same or have the same general chemistry in the query sequence (in this case, the chimp sequence) as in the subject sequence (in this case human).