Chimp Sequence Annotation: Region 2_3

Jeff Howenstein
March 30, 2007

BIO434W Genomics

Introduction

We received region 2_3 of the ChimpChunk sequence, and the first step we performed was to run RepeatMasker with the –nolow option enabled in order to mask out all repetitive elements except for low complexity repeats. After running RepeatMasker, we used GENSCAN, a gene prediction program, and BLAST searches to identify each feature and determine their functions. The GENSCAN output table and map are shown below.

GENSCANW output for sequence BoneKrushers

GENSCAN 1.0 Date run: 23-Mar-107 Time: 14:13:17

Sequence Pan : 101300 bp : 40.66% 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	Init	+	1661	1986	326	1	2	62	36	272	0.773	14.25
1.02	Intr	+	4302	4423	122	0	2	106	55	0	0.372	-2.28
1.03	Intr	+	5669	5852	184	0	1	82	21	138	0.574	4.42
1.04	Intr	+	7263	7393	131	0	2	49	79	62	0.317	0.82
1.05	Intr	+	8546	8741	196	0	1	18	31	152	0.559	0.05
1.06	Intr	+	8774	9053	280	2	1	84	49	119	0.749	4.16
1.07	Intr	+	9112	9258	147	0	0	45	44	119	0.601	2.71
1.08	Term	+	10390	10623	234	0	0	84	45	117	0.580	2.34
1.09	PlyA	+	15177	15182	6							1.05
2.06	PlyA	-	16115	16110	6							1.05
2.05	Term	-	22698	22416	283	2	1	65	29	208	0.964	6.51
2.04	Intr	-	25508	25208	301	0	1	75	94	329	0.997	27.07
2.03	Intr	-	27023	26860	164	1	2	90	95	97	0.992	9.30
2.02	Intr	-	28311	28227	85	1	1	87	69	114	0.998	7.36
2.01	Init	-	31458	31284	175	0	1	49	51	143	0.725	6.26
2.00	Prom	-	34543	34504	40							-4.95
3.03	PlyA	-	35159	35154	6							1.05
3.02	Term	-	38498	38225	274	1	1	84	53	207	0.843	10.76
3.01	Init	-	38730	38567	164	0	2	91	47	65	0.881	2.04
3.00	Prom	-	40418	40379	40							-8.95
4.00	Prom	+	41625	41664	40							-7.55
4.01	Init	+	41704	41904	201	0	0	60	86	115	0.520	7.42
4.02	Term	+	45670	45921	252	0	0	84	32	163	0.858	4.95
4.03	PlyA	+	46503	46508	6							1.05
5.02	PlyA	-	47163	47158	6							1.05
5.01	Sngl	-	50135	49083	1053	2	0	61	39	815	0.750	70.79
5.00	Prom	-	53828	53789	40							-4.95
6.00	Prom	+	72890	72929	40							-4.35
6.01	Init	+	73337	73419	83	1	2	73	81	63	0.713	4.67
6.02	Intr	+	84140	84382	243	2	0	37	51	227	0.014	9.59
6.03	Term	+	84521	84896	376	2	1	-41	43	279	0.458	3.73
6.04	PlyA	+	86163	86168	6							1.05

Figure 1: GENSCAN output table

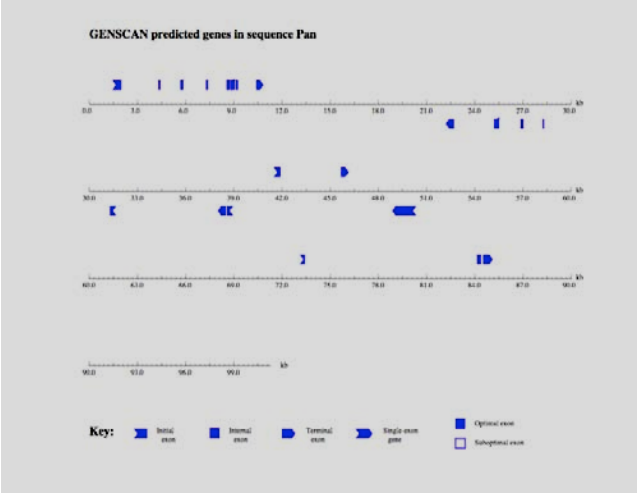


Figure 2: GENSCAN output map

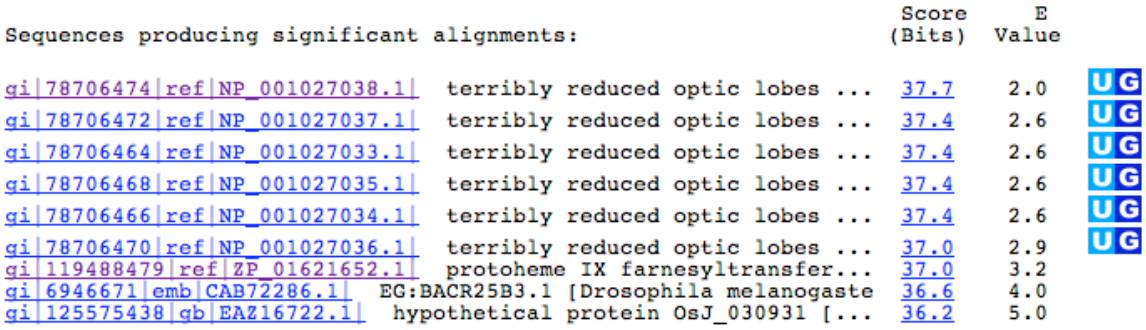
GENSCAN Results

After running GENSCAN on our ChimpChunk DNA sequence, the program predicted six features. The first feature codes for 539 amino acids contained in 8 exons. The second feature codes for 335 amino acids in 5 exons. The third predicted feature is a two exon region coding for 145 amino acids. The fourth feature is a four exon region coding for 150 amino acids. The fifth feature codes for 350 amino acids in one exon. Finally, the sixth predicted feature is a five exon region coding 233 amino acids. The corresponding predicted peptide sequences are shown below in order from 1 to 6.

```
1)>Pan|GENSCAN_predicted_peptide_1|539_aa
MLRRSRAWALTGGAGWPGPSVVGGCLLLLPSHLSGSEIGAWLRVSMEEAPSLKPFLASPS
APLVPPPPVCSQDLLALWHWGLEDTFSGQALKMAFEIEDRVVTCGWPPRGMSPPKFHCPV
PATRRPCVHSTLSTHSHRFLKSHCPELVAASSQSLGVGSVIPSYRRGRCGHADGEPAAHG
SSVGEEKGAPGPLGPPAQPPAANACSTWGFSGELTVSWRGWQTPKVPQAFPEHLTKTVHL
CTHGAGLTSVTLESGKRVSYHTFLSLWIHMTWNLPWATQMLFILLKMREPPLFWTLAILG
TAPWTTSGKHAPNLSVSEDRWPLGMPCGENQGRGPHSEVLELGRCMLCGWQLVPLETEAC
WAEPSCLVVSLLLALLIPVPHSLPTLSSGRGTGKCADAHERWGKPMTVPGRVQRPVRGWL
PRQAPRGARRFVCVEGALVVMCPTGGGTGPYCPIPSGSGRRHFPLKETNNRKGRLPKRLR
PHTWCCYSLEYSSYSLRCTEDTWTEDQVRLSYRVSVSPNVTNTHPFLSSPTPIPFWVLT

2)>Pan|GENSCAN_predicted_peptide_2|335_aa
MITKIKISVNILAAHCTQQKRIRKLKDKLEEITSVLHRERHSGGNCSSKEEKGVEDNKCS
VDEGVSEGLPTLQSTSSTNAPPDDDDRLENVQYPYQLYIAPSTSSTERPSPNGPDRPFQC
PTCGVRFTRIQNLKQHMLIHSGIKPFQCDRCGKKFTRAYSLKMHRLKHEGKRCFRCQICS
ATFTSFGEYKHHMRVSRHIIRKPRIYECKTCGAMFTNSGNLIVHLRSLNHEASELANYFQ
SSDFLVPDYLNQEQEETLVQYDLGEHGFESNSSVQMPVISQVSSTQNCESTFPLGSLGGL
AEKEEEVPEQPKSSACAEATRDDPPKSELSSITIE

3)>Pan|GENSCAN_predicted_peptide_3|145_aa
MGVRHPLEEAVCPLAEVKHCAGRTLLVSIRCSLQNPQAGTFKSTEAAPTAAPSPRDALPN
EEESKEAVWPQPLCHTVLSSAQLELPGLFSAVRGKPTTQASVMADAPPPTKLDRPRSTSD
CCAGSENFKPVVLSLLGSVGVGPAE

4)>Pan|GENSCAN_predicted_peptide_4|150_aa
MSELLFTIASKKRKYLGIQLTRDMKDLFKENYKPLLNEIKDDTNKWKNIPCSWIGRMNIV
KMAILPKGERDLKPVAKQRDREGHVKTNAYTKVQYHILELQAKEHQGLWAAATGWKRQGR
ILPYNLQREKDPSKGANTCRSDFWPLTMRE

5)>Pan|GENSCAN_predicted_peptide_5|350_aa
MGVKTFTHSSSSHSQEMLGKLNMLRNDGHFCDITIRVQDKIFRAHKVVLAACSDFFRTKL
VGQAEDENKNVLDLHHVTVTGFIPLLEYAYTATLSINTENIIDVLAAASYMQMFSVASTC
SEFMKSSILWNTPNSQPEKGLDAGQENNSNCNFTSRDGSISPVSSECSVVERTIPVCRES
RRKRKSYIVMSPESPVKCGTQTSSPQVLNSSASYSENRNQPVDSSLAFPWTFPFGIDRRI
QPEKVKQAENTRTLELPGPSETGRRMADYVTCESTKTTLPLGTEEDVRVKVERLSDEEVH
EEVSQPVSASQSSLSDQQTVPGSEQVQEDLLISPQSSSIGIMPSVFFLVL

6)>Pan|GENSCAN_predicted_peptide_6|233_aa
MQSLTVAPSILRHLGFWLKGFTSAGGNQELAMQIFGVLKELMTQHVHTYGLIMGGSNRSA
EAQKLANGINITVATPGRLLYHMQNIPGFMYKNLQCLVIDEADRILDVGVDDDKANTTVD
GLEQGYVVCPSEKRFLLLFTFLKKNRKKLVVFFSSCMSVKYHYELLNYIDLPVLAIHGKQ
KQNKRTTTFFQFCNTDLGTHCVWMWWQEDWTFLKSTGLFTMTIRMTLRNIFIV
```

Feature 1

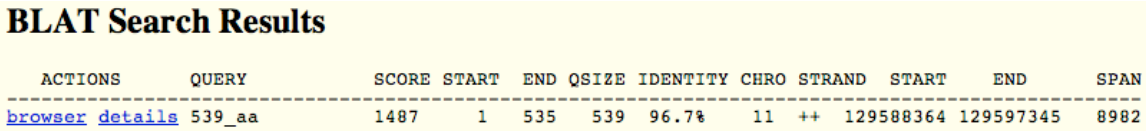
GENSCAN predicted an 8 exon gene for feature 1 with a total length of 539 amino acids. We performed an nr NCBI protein BLASTp search for this feature and the results showed no quality hits. The lowest E value obtained was 2.0, which was a relation with the reduced optic lobes CG33950-PF. Referring to the BLAST alignments in detail, this gene is actually located in the Drosophila Melanogaster genome. The sequence is ~36% homologous with the Drosophila Melanogaster reduced optic lobes CG33950-PF gene.



Sequences producing significant alignments:		Score (Bits)	E Value
gi|78706474|ref|NP_001027038.1|	terribly reduced optic lobes ...	37.7	2.0
gi|78706472|ref|NP_001027037.1|	terribly reduced optic lobes ...	37.4	2.6
gi|78706464|ref|NP_001027033.1|	terribly reduced optic lobes ...	37.4	2.6
gi|78706468|ref|NP_001027035.1|	terribly reduced optic lobes ...	37.4	2.6
gi|78706466|ref|NP_001027034.1|	terribly reduced optic lobes ...	37.4	2.6
gi|78706470|ref|NP_001027036.1|	terribly reduced optic lobes ...	37.0	2.9
gi|119488479|ref|ZP_01621652.1|	protoheme IX farnesyltransfer...	37.0	3.2
gi|6946671|emb|CAB72286.1|	EG:BACR25B3.1 [Drosophila melanogaste	36.6	4.0
gi|125575438|gb|EAZ16722.1|	hypothetical protein OsJ_030931 [...	36.2	5.0

Figure 3: BLASTp hits for feature 1

Following the BLASTP search, a BLAT search was performed with the original GENSCAN prediction in order to compare the predicted gene to the known human genes.



BLAT Search Results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	539_aa	1487	1	535	539	96.7%	11	++	129588364	129597345	8982

Figure 4: BLAT results for feature 1

To our surprise, we did find a 96.7% homologous region on the human genome that contained the predicted sequence, which is found on chromosome 11. When we looked more closely in the browser view, we discovered that the sequence was only partially aligned with the GENSCAN gene predictions. There was no evidence of a gene from RefSeq, Human mRNAs, and other databases. Because there is no other evidence than just the GENSCAN prediction and because there were no quality hits from the BLASTp search, it can be concluded that this GENSCAN prediction is a misprediction/false-positive. This means that the predicted sequence is a possible gene in the human/chimp genome, but no one has ever annotated it.

After performing this experiment, we found that we at the beginning of the project, we ran GENSCAN with the wrong library. We actually ran it against the Drosophila library, and that is why this feature shows up in our chimp chunk.

Features 2 and 5

GENSCAN predicted a 5-exon gene for feature 2 with a length of 335 amino acids. Using the predicted GENSCAN protein amino acid sequence, we performed an nr NCBI protein BLAST search, which resulted in the detection of a putative conserved domain, COG5040.

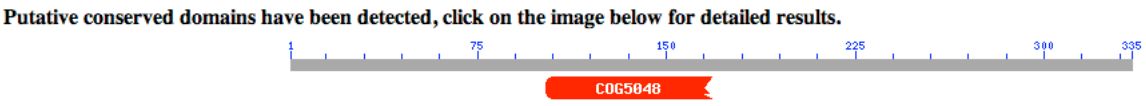


Figure 5. Cog5040 is a FOG: Zn-finger

gi|57086053|ref|XP_546395.1|	PREDICTED: similar to BTB (POZ) ...	533	4e-150	UG
gi|118101882|ref|XP_417873.2|	PREDICTED: similar to ZBTB44 prote	516	7e-145	UG
gi|126327478|ref|XP_001373788.1|	PREDICTED: similar to ZBTB44 pr	511	2e-143	G
gi|114641309|ref|XP_508864.2|	PREDICTED: similar to BTBD15 prote	451	3e-125	G
gi|119919095|ref|XP_869163.2|	PREDICTED: similar to ZBTB44 prote	447	3e-124	UG
gi|74760158|sp|Q8NCP5|ZBT44_HUMAN	Zinc finger and BTB domain-...	447	4e-124	G
gi|109109290|ref|XP_001113645.1|	PREDICTED: similar to BTB (P...	445	2e-123	G

Figure 6. BLASTP results for feature 2

When the predicted protein was put into the NCBI BLASTp search, the results indicated that this feature is related to the Zinc finger and BTB domain-containing protein 44. Next we ran a BLAT search comparing the human genome to both the predicted feature sequence as well as the ZBT44 sequence, and the best match was found on human chromosome 11 for both sequences.

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END	SPAN
browser details	335_aa	825	60	335	335	100.0%	11	+-	129609099	129614999	5901
browser details	335_aa	367	137	331	335	95.4%	13	+-	76064495	76064862	368
browser details	335_aa	66	314	335	335	100.0%	13	++	76064863	76064928	66
browser details	335_aa	60	310	335	335	88.5%	12	+-	63344812	63344889	78
browser details	ZBT44_HUMAN	1676	1	562	570	99.9%	11	+-	129609261	129636978	27718
browser details	ZBT44_HUMAN	217	418	562	570	92.3%	13	+-	76064645	76064862	218
browser details	ZBT44_HUMAN	12	180	183	570	100.0%	18	+-	62928518	62928529	12
browser details	ZBT44_HUMAN	12	196	207	570	66.7%	18	+-	62928446	62928481	36

Figure 7. BLAT search results.

The GENSCAN predicted a protein that had 100 percent homology on chromosome 11 and the ZBT44 protein showed a 99.9% homology on chromosome 11.

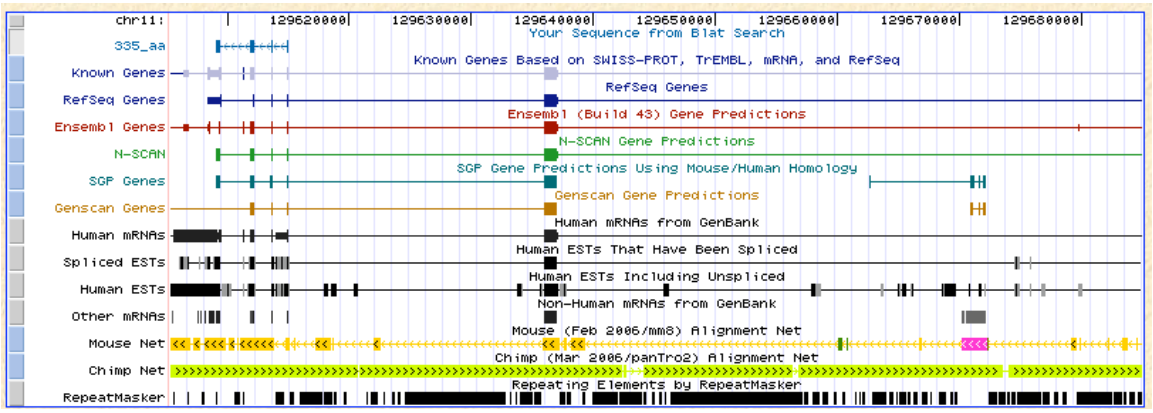


Figure 8. Browser view of the 335_aa GENSCAN predicted feature 2.

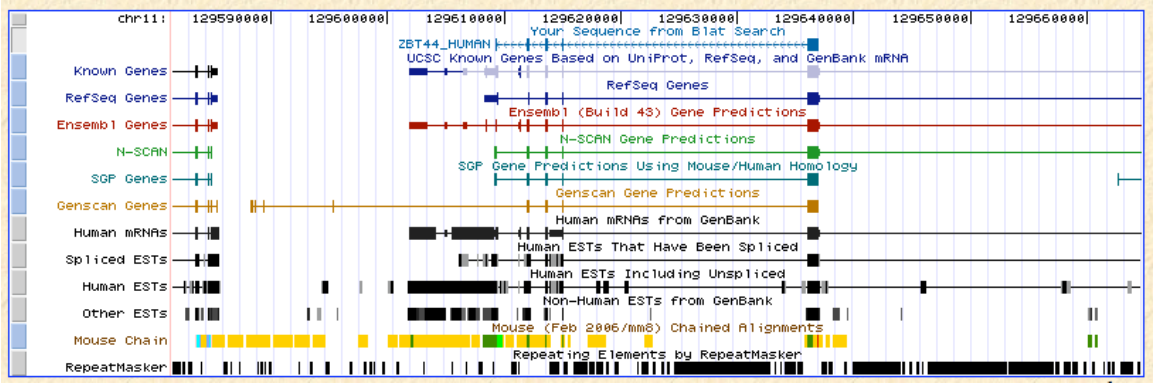


Figure 9. Browser view of the ZBT44 gene from NCBI BLASTP results.

From the browser view, feature 2 matched very well with known genes and all other databases. However, when compared to the ZBT44 gene match, feature 2 predicted from GENSCAN was missing an exon on position 129640000. The GENSCAN prediction can still be a real gene because the chimp chunks have been truncated, so the truncated chunk may have cut the last exon off. To increase our confidence that the predicted GENSCAN prediction is indeed a real gene, we used the ZBT44 sequence and ran a BLAST 2-comparison alignment with the masked fasta chunk. Using the TBLASTN program, it was discovered that all query sequence existed in my chimp chunk. Therefore, the predicted GENSCAN predicted gene is likely to be the Zinc finger and BTB domain-containing protein 44 (ZBT44) homolog on chromosome 11, location 11q24.3, but the prediction failed to predict the end exon at position 129640000.

For feature 5, GENSCAN predicted a gene encoding for 350 amino acids contained in one exon. When the GENSCAN predicted gene had a BLASTp search ran on it, the results indicated a conserved domain, the BTB protein. In fact, because we ran the GENSCAN with the wrong library, GENSCAN split features 2 and 5, which go together; feature 5 is shown below, which is the missing exon to the ZBT44 gene. Thus, features 2 and 5 go together and make a real gene.

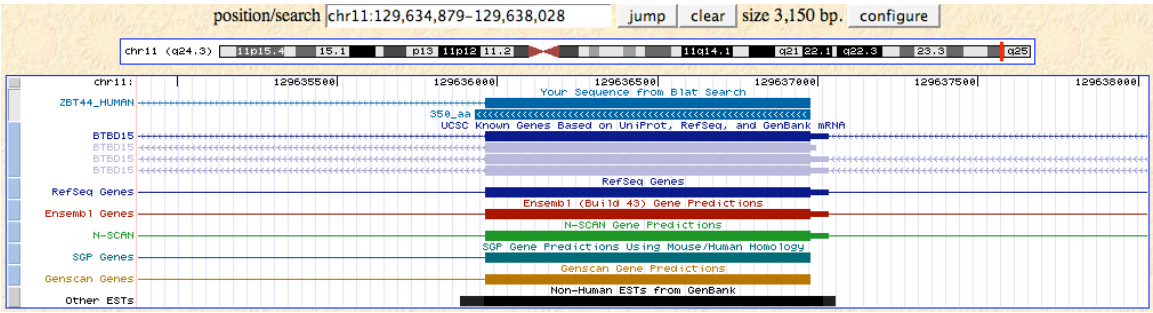
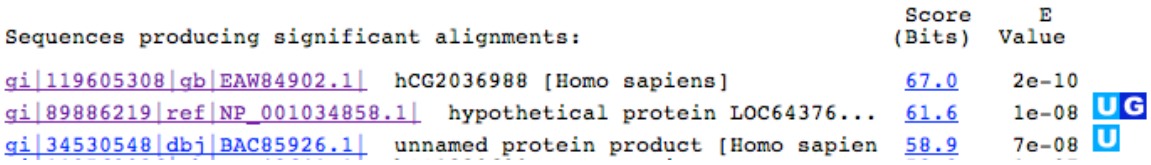


Figure 11: Blat match for feature 5

Feature 3

For feature 3, GENSCAN predicted 145 amino acids and 2 exons. After running BLASTp search, we received several low quality matches. The best match is the hCG2036988 gene.



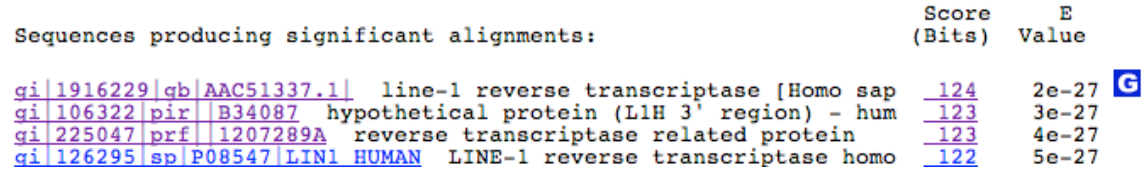
Sequences producing significant alignments:	Score (Bits)	E Value
gi|119605308|gb|EAW84902.1| hCG2036988 [Homo sapiens]	67.0	2e-10
gi|89886219|ref|NP_001034858.1| hypothetical protein LOC64376...	61.6	1e-08
gi|34530548|dbj|BAC85926.1| unnamed protein product [Homo sapien	58.9	7e-08

Figure 10. BLAST result of feature 3.

Next we compared both the GENSCAN predicted sequence and the hCG2036988 sequence to known human genes by running a BLAT search against the human genome, but there were no matches found. A possible explanation for this is that someone predicted an incorrect protein and submitted to Genbank. This incorrect prediction could have been a failure of RepeatMasker. When we looked into this, we discovered that Santacruz only searched unmasked regions and we have determined that the Santacruz RepeatMasker probably masked this feature.

Feature 4

For feature 4, GENSCAN predicted a gene coding for 150 amino acids with 2 exons. When a BLASTp search was run on the predicted protein, the results yielded no evidence for putative conserved domains, and all of the top BLAST hits relate this feature to the reverse transcriptase protein. Below are the top hits for the BLASTp search on NCBI.



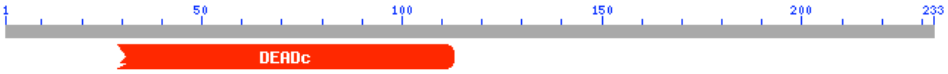
Sequences producing significant alignments:	Score (Bits)	E Value
gi|1916229|gb|AAC51337.1| line-1 reverse transcriptase [Homo sap	124	2e-27
gi|106322|pir||B34087 hypothetical protein (L1H 3' region) - hum	123	3e-27
gi|225047|prf||1207289A reverse transcriptase related protein	123	4e-27
gi|126295|sp|P08547|LIN1_HUMAN LINE-1 reverse transcriptase homo	122	5e-27

After this, the predicted chimp protein sequence as well as the line-1 reverse transcriptase were used in a BLAT search of the human genome, and there were no BLAT matches found for either sequence. Due to this, it can be assumed that the predicted protein is not a gene, but rather it simply matches some element of the reverse transcriptase protein and was reverse transcripted back into the genome. Again, because we ran GENSCAN with the Drosophila library, this is most likely some element found in drosophila that does not belong as a feature in the chimp chunk.

Feature 6

GENSCAN predicted a 233 amino acid long protein for feature 6 with 5 exons. After running a BLASTp on the predicted feature, it seems that it is related to the DEAD box protein. DEAD box proteins, characterized by the conserved motif Asp-Glu-Ala-Asp (DEAD), are putative RNA helicases. They are implicated in a number of cellular processes involving alteration of RNA secondary structure such as translation initiation, nuclear and mitochondrial splicing, and ribosome and spliceosome assembly.

Putative conserved domains have been detected, click on the image below for detailed results.



The request ID is 1174939847-25662-71126906118.BLASTQ5

From this information, I conclude that the predicted gene is probably a real gene with an ATP binding function, or a pseudogene derived from a real gene. Looking further into the BLAST hits, I note that the DEAD box protein is 660 amino acids in length, while my predicted gene is only 233. This makes me lean towards pseudogene, so I ran BLAT for both my unknown feature and the DEAD box protein, and the results are shown below.

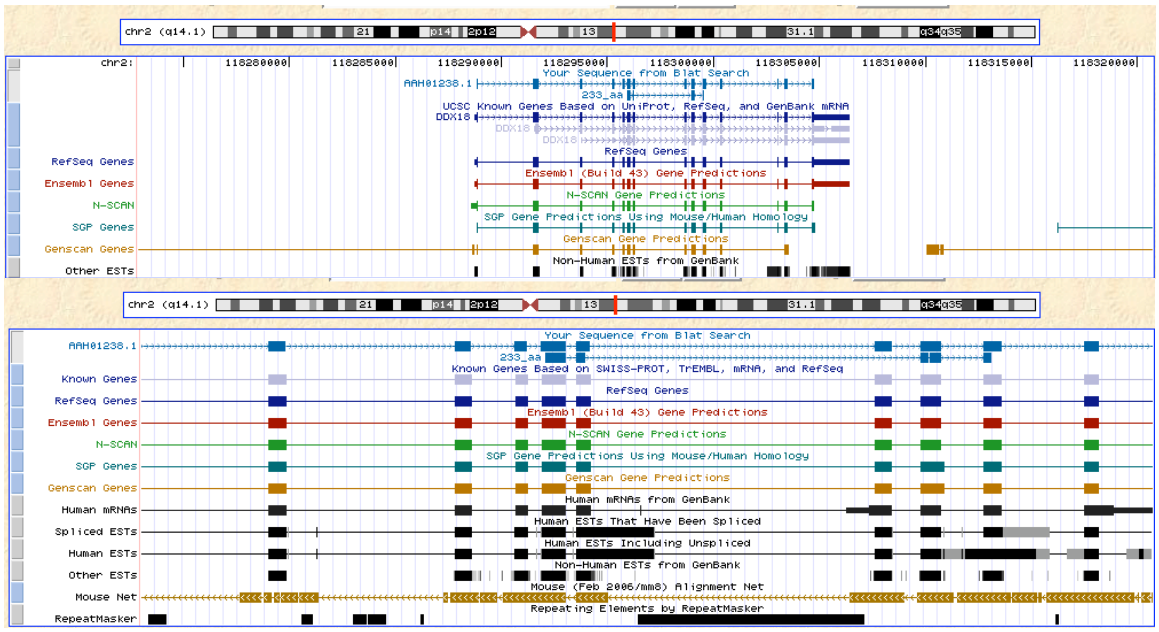


Figure 12: BLAT match for feature 6

It is interesting that there are 5 exons in my feature, but 14 in the dead box protein. It is even more interesting that the 5 exons are found in the middle of the dead box protein rather than being on one of the ends. There are exon matches in the middle, so it is a possibility that those 5 exons were alternatively spliced, and that mRNA with these five exons was then reverse transcripted back into the genome. Therefore, this could be a pseudogene that was constructed off an mRNA that got reverse transcripted back into the original mRNA, which was an alternatively spliced portion of the original gene. Looking up the DDX18 gene, the dead box gene, there is evidence of alternative splicing, and more specifically it showed in humans the exon patterns: 4,5,6,8,9, and 9,10,11,13,14. Mine follows the pattern 6,7,9(a), 9(b), 10. My pattern is not the exact same, but something may have mutated along the way in the CHIMP. I say 9a and 9b because those two exons in my feature are both on exon 9 from the dead box, a quality that there is evidence for in humans; the 10th exon in humans gets cut into isoforms, which is what it appears like in the predicted sequence. It is known that a sequence only

becomes alternatively spliced if it is a functional element, and since we found this predicted sequence on Ensembl, we know it is a functional element. In conclusion, this is probably a pseudogene that was alternatively spliced, then reverse transcripted back into the genome.