Sonal Singhal
17 April 2006
Bio 4342

Annotating the Chimp Genome: Chunk 2.6

In this paper, I discuss the annotation of a chunk of the chimpanzee genome (*Pan troglodytes*) as done to find genes, pseudogenes, and repetitive elements. In order to prepare this region for analysis, I ran RepeatMasker to mask out all repetitive elements except for low-complexity repeats. The masked sequence was used for all further analyses, unless otherwise noted. I then ran Genscan, an *ab initio* gene finder, to identify putative protein-coding domains (results shown in Appendix A and Figure 1). As an initial reference for determining whether these regions were coding and thus expressed, I also used the human expressed sequence tag (EST) database to find high-quality nucleotide matches. Further, by using blastx, I searched the Swissprot database to see if any of the open reading frames (ORFs) in our chunk coded for known proteins. The EST and blast search results were viewed by using the software Herne.

The region, chunk 2.6, consists of 92.9 Kb and has 38.4% G/C content. In all future discussion of the chunk, I will discuss locations of features with respect to the chunk rather than the chromosome from which it is derived. Twenty five percent of the region is repetitive; individually, SINE elements represent 8% of the chunk, LINEs 9%, LTRs 5%, and other DNA elements 3% (Appendix B). Initial BLAT searches of the chunk suggest that it is located on chromosome number 5 of the chimp genome and that the likely syntenic region in humans is on chromosome 6. In this region of chromosome 6 (6q21), the following features have been identified: 3' end of a novel gene, the gene for squamous cell carcinoma antigen recognized by T cells (SART2), a chromobox homolog 3 (CBX3) pseudogene, a keratin 18 (KRT18) pseudogene, a novel gene, part of a novel gene and two CpG islands. As chimp and human genomes are largely syntenic, I expect to find some of the same features in our chunk. The initial Genscan search found five features, which I discuss below (Figure 1, Appendix A).

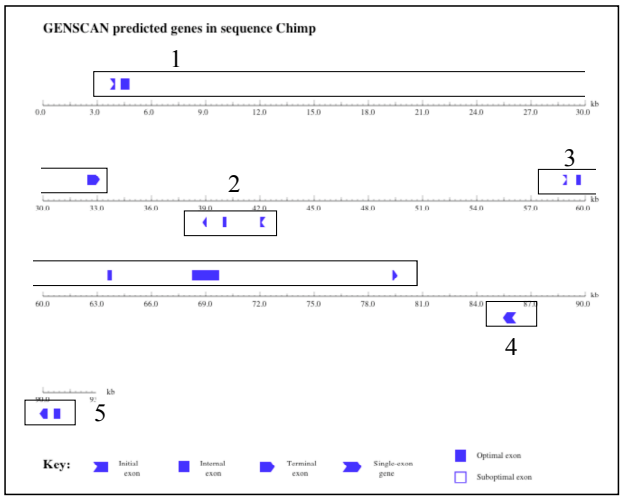

Figure 1: Genscan predicted genes

Features 1 and 3

My initial guess was that these two features were part of the same protein, as the ESTs near these two features both matched reliably to the SART2 gene. To confirm this suspicion, I entered both features into Blat and found where they were located in the chimp genome. Simultaneously, I entered the SART2 mRNA sequence for humans (Z84488) into Blat. Doing so, I found strong evidence that these two features together comprise the SART2 gene (Figure 2). However, not all regions from the human mRNA match well with the regions in chimp, and further, the first three predicted exons from human and chimp match to slightly different regions.

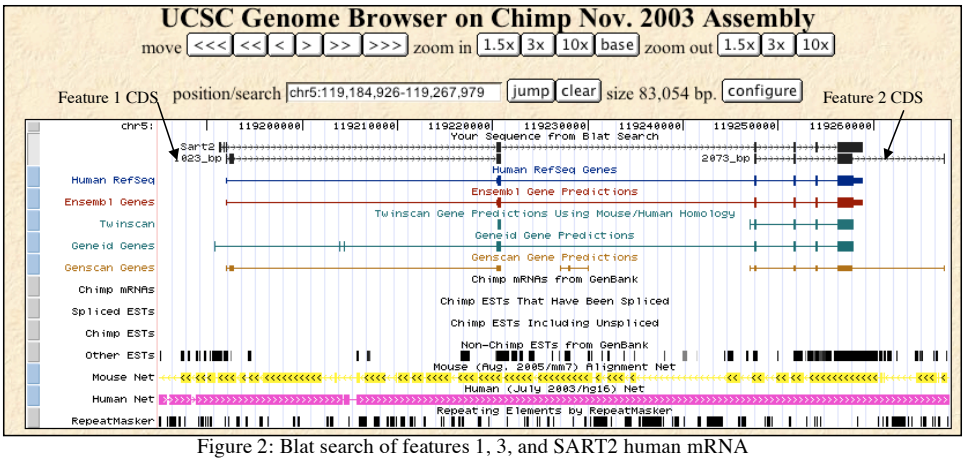

Figure 2: Blat search of features 1, 3, and SART2 human mRNA

To investigate this relationship further, I used blast2seq to consider the percent similarity of each exon of the human SART2 gene with the chimp putative SART2 coding region (3 Kb to 80 Kb). Problematically, the GenBank record for SART2 listed multiple possible splicings of the gene, and it was unclear which version had the greatest EST support. By looking at the Blat search, I saw that Genscan and Geneid each predict seven exons, and that Ensembl and Human Refseq predict 5 exons. I then used the human EST database to see which exons had sufficient and reliable EST support. This analysis indicated that the EST data support the Human Refseq Gene and Ensembl Gene predictions most strongly. I thus used the annotation of SART2 that matched these predictions (AI23408). When each exon from SART2 was compared to the chimp putative SART2 coding region, I found that the coding regions are more than 99% similar at the nucleotide level and show no evidence of insertions or deletions (indels) (Table 1). Indels result in frameshift mutations and drastic changes in peptide sequence; thus their presence suggests that the feature is no longer coding. These exons from the chimp genome also show a high percent similarity to human ESTs (Figure 3). Based on this evidence, I conclude that features 1 and 3 define the SART2 gene, which is located from 32.5 Kb to 69.9 Kb. Of course, without chimp EST evidence as support, it is possible that SART2 is non-functional in chimps.

SART2 expresses a tumor-rejection antigen that is found on squamous cell carcinomas (squamous cells are a specific class of epithelial cells). Through SART2 expression, the cancerous cell presents a small peptide at its cellular surface. This peptide targets cytotoxic T cells, which can then trigger apoptosis of the cancerous cell. SART2 has a homolog

(SART1).Searches on Genbank suggest that many other mammals (i.e., mice and rats) also express these proteins.

exon	length in bps	percent similarity	alignment in chimp
1	415	99.7%	32504-32919
2	253	99.6%	59503-59756
3	238	99.1%	63569-63808
4	206	99.0%	65962-66169
5	1747	99.3%	68208-69966

Table 1: Exons of the SART 2 human gene and their matches to chimp

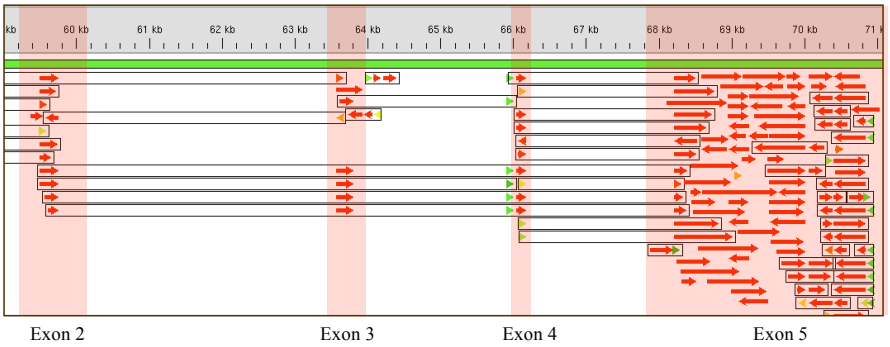

Figure 3: EST support for SART2 in chimp. Exon 1 shows a similar pattern.

Feature 2

As predicted by GenScan, this putative ORF on the negative strand begins at 52.5 Kb and ends at 33.9 Kb. Including the initial and terminal exons, the gene has three exons and encodes a predicted peptide of 92 amino acids. By using blastp and searching against the Swissprot database, I found that the Genscan predicted peptide only matches with two other proteins, both with E-values that are too high to be trusted. However, there were three high-quality EST matches in this 20 Kb region (N35795, AW899762 and AV757520). I later discovered that these match to the SART2 gene which extends through this region. Blastx searches of the predicted coding sequence found no protein matches, so I conclude that this feature is neither a pseudogene nor a coding gene. Later, I realized that this feature exists within features 1 and 3, which are really only one feature. Even though a gene can be localized within another gene, this still increases the likelihood that this region is neither a pseudogene nor a real gene. This would predict that the EST matches would have the same orientation as SART2, but I did not have time to check this out.

Feature 4

As predicted by GenScan, this putative ORF on the negative strand begins at 89.3 Kb and ends at 80.0 Kb. GenScan only predicts one exon, which encodes a peptide of 95 amino acids. A blastp search using this peptide against the Swissprot database shows that it has 100% match

to a *Mus musculus* modifier. However, this modifier is not well characterized and thus this match constitutes insufficient evidence. The same search shows that this region matches well to a human heterchromatin-1 modifier, but it only matches 70 of the 183 predicted amino acids. Incomplete matches suggest that although a protein domain is conserved between two species, the protein as a whole has not been conserved. Such a pattern can occur when one compares two non-orthologous genes. To determine why this match was incomplete, I used the predicted peptide in a Blat search. This shows us that some of the coding region might have not been predicted due to a repetitive element in the region of interest which perhaps caused Genscan to terminate the feature (as indicated in Figure 4).

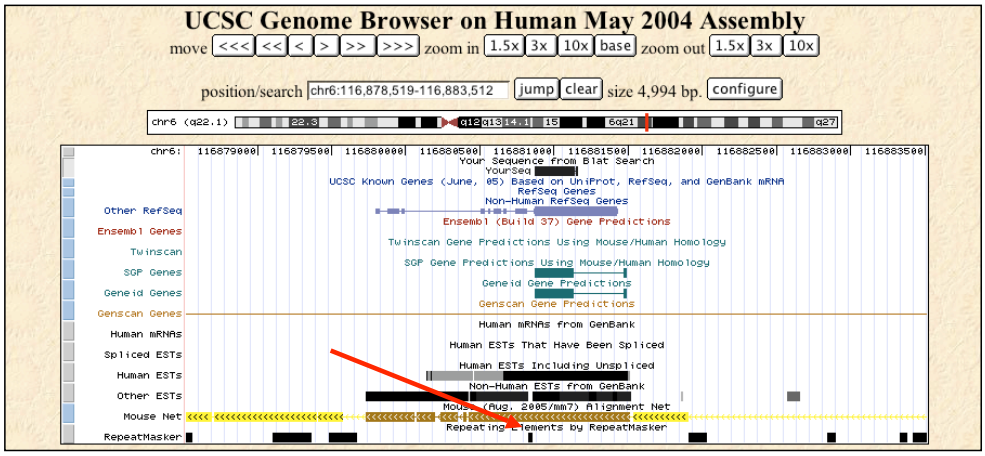

Figure 4: Repetitive element near the feature of interest

To investigate this feature further, I extracted a 10 Kb region (from 80 to 90 Kb) of unmasked nucleotide sequence and used blastx and to find all possible protein matches in this region to the Swissprot database. Using this search resulted in two matches that explained nearly all of the amino acids. However, the second part of this match includes premature stop codons; as shown by the change in reading frame between the two matches, it also includes indels within a single exon (Figure 5). Further, under the assumption of synteny, I would expect that this feature would match to a region on human chromosome 6. The feature matched to a protein encoded by a gene on human chromosome 7, which was suspicious. However, a blastn search against the human genome found that this feature is 100% similar to a region on chromosome 6 outside of a 4 base pair indel at the beginning of the putative coding region. Both the human and chimp nucleotide sequences, if translated, result in multiple premature stop codons. Thus, I conclude that Feature 4 is a pseudogene, with similarity to chromobox homolog 3 (as found in humans, rats, and orangutans). Confirming this conclusion, human ESTs match poorly to the region in question (i.e., 93-96% similarity).

Feature 5

As predicted by GenScan, this putative ORF on the negative strand begins at 90.9 kB and ends at 89.9 kB. The predicted peptide is 199 amino acids long and is encoded by two exons. By using blastp, I compared this peptide to other known proteins in the Swissprot database.

Although blastp suggests that this peptide has two conserved regions indicative of filament proteins, the blastp output does not result in a believable match. Indeed, the predicted

Figure 5: Blastx search of region containing feature 4. A premature stop codon is circled in red. The presence of an indel within the exon is indicated by the change in reading frame between the two matches, as circled in blue.

peptide has some similarity to keratin proteins—the percent identity is 58% to a *Mus musculus* keratin protein and 55% to a *Homo sapiens* keratin protein (Figure 6). For the highly similar chimp and human genomes, this difference is large, particularly as the predicted peptide is 200 amino acids shorter than the average keratin peptide. Because I also do not find any convincing human EST matches (most are 85-90% similar), I conclude that this feature is a pseudogene likely derived from a keratin-encoding gene. I suspect it is a pseudogene somehow related to Keratin-18. This feature and the mRNA for Keratin-18 are fairly similar, although the feature carries several indels in comparison to the Keratin-18 sequence. However, there is the caveat that this protein might be a unique functional gene in chimp. While this is unlikely, there is no evidence of a premature stop codon, so this putative coding domain could produce a functional protein. Without EST evidence from chimps, we cannot confirm our conclusion.

Figure 6: Blastp match of Feature 5

Fi dates not found by matched

Sally Elgin 4/26/06 3:16 PM
Deleted:

to several ESTs with high fidelity (Figure 7). However, searches of these regions (translated and untranslated) failed to recover any possible gene candidates. Because these two anonymous EST regions are close to one another, I suspect that they are derived from one unique and yet to be described gene or pseudogene. This prediction could be checked by looking for sequence conservation among species.

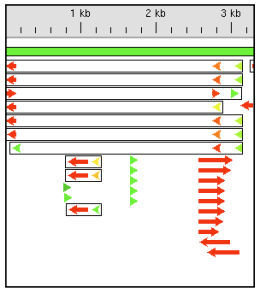

Figure 7: Anonymous ESTs

In conclusion, by using numerous homology searches and genome browsers, I was able to identify a gene (SART2) and two pseudogenes in our region of interest. A summary map and a table of our findings are shown respectively in Figure 8 and Table 2. As I predicted, the chimp chunk shows synteny with comparable sequence from humans; both regions carry a presumably functional SART2 gene and pseudogenes for chromobox homolog 3 and keratin-18. This exercise shows the power of genomic information to help characterize functional elements and also illustrates the assumptions, and thus limitations, of these tools.

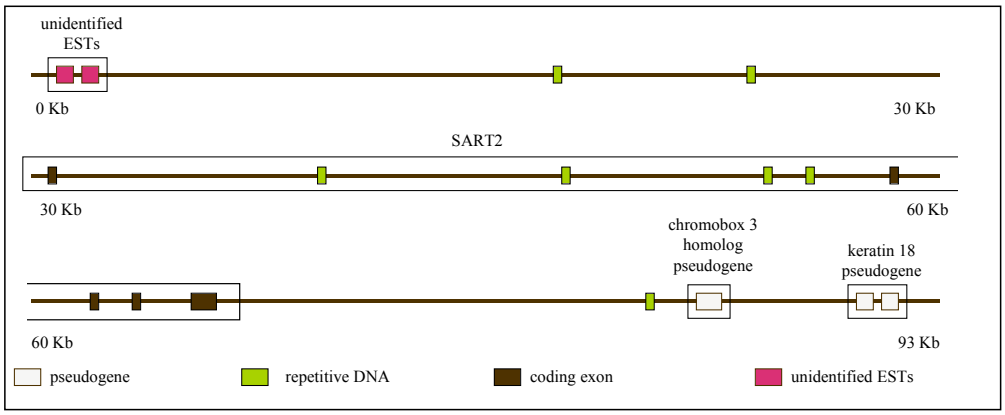

Figure 8: Summary of all identified features

feature	location	identity
1 & 3	32.5-69.9 Kb	SART2 functional gene
2	33.9-52.5	misprediction by Genscan
4	80.0-89.3 Kb	chromobox 3 homolog pseudogene
5	89.9-90.9 Kb	keratin 18 pseudogene
anonymous ESTs	2-3 Kb	no matches found to characterize

Appendix A: Genscan Predicted Genes, CDS, and Proteins

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	3927	3988	62	2	2	92	36	134	0.944	7.37
1.02	Intr	+	4295	4782	488	2	2	-4	58	251	0.121	4.63
1.03	Term	+	32450	32922	473	0	2	98	33	360	0.762	25.61
1.04	PlyA	+	33658	33663	6							1.05
2.04	PlyA	-	33984	33979	6							1.05
2.03	Term	-	39050	39042	9	2	0	131	43	0	0.555	-2.98
2.02	Intr	-	40145	39946	200	0	2	56	93	174	0.800	12.85
2.01	Init	-	42084	42015	70	0	1	62	89	27	0.802	1.46
2.00	Prom	-	52533	52494	40							-3.75
3.00	Prom	+	52584	52623	40							-6.75
3.01	Init	+	58969	58982	14	0	2	104	39	13	0.757	-2.67
3.02	Intr	+	59502	59755	254	0	2	122	98	132	0.938	13.75
3.03	Intr	+	63568	63807	240	2	0	93	80	193	0.991	15.60
3.04	Intr	+	68242	69743	1502	2	2	79	16	1496	0.562	129.97
3.05	Term	+	79333	79395	63	0	0	94	38	40	0.024	-3.49
3.06	PlyA	+	79573	79578	6							1.05
4.02	PlyA	-	80061	80056	6							1.05
4.01	Sngl	-	85953	85666	288	0	0	55	32	280	0.915	14.44
4.00	Prom	-	89332	89293	40							-10.05
5.03	PlyA	-	89902	89897	6							1.05
5.02	Term	-	90253	90017	237	1	0	44	40	279	0.702	13.88
5.01	Intr	-	90956	90594	363	2	0	60	27	530	0.636	38.56

>15:16:40|GENSCAN_predicted_peptide_1|340_aa
MAWRGRGGLGSEGAGSPGALDLEGQLFAPPHPLERRAPSREHDENLEIAFRASDLAGPAA
ISSTVLGASASLCGVGWHTLASSFPSVPQPFGGRGGGGGLCADPGAGNPTLRLRKPRKSS
PSLSQTLDSSAPGLRRHSEKLLVNLIPEAVPGRGGGGGLTHRLGSSESDATWAQDLSGLE
SELGSFEDGLAALEIWRSDATMRTHTRGAPSVFFIYLLCFVSAYITDENPEVMIPFTNAN
YDSHPMLYFSRAEVAELQRRAASSHEHIAARLTEAVHTMLSSPLEYLPPWDPKDYSARWN
EIFGNNLGALAMFCVLYPENIEARDMAKDYMERMAAQPSW

>15:16:40|GENSCAN_predicted_CDS_1|1023_bp
atggcgtggcgggggcgcggagggctcggttcggaggggccgggagcccgggcgccctg
gacctcgaggggcagcttttcgccccctccccacccccttgaacgccgagcaccgagccgg
gaacatgatgagaacctggagatcgcattccgtgcatcggatctggccggcccagccgcg
atttcctcgacggtccttggggcctcggcttcactctgcggggttgggtggcacaccctc
gcgtcctcctttccctccgtaccccagccatttgggggcaggggaggtggcggcggcttg
tgtgcagatcctggcgctggcaaccctactttgcggctgcggaagccgaggaagagctcc
ccaagcctctcccagacgctggattccagcgctcctgggctccgccggcactcggagaaa
ttgttagtaaacctcatccccgaggcagtcccggggagggggaggggcgggggcttgact
cacagactgggaagctctgaaagcgatgccacatgggctcaagatctgtcggggctggaa
tctgagctaggatctttcgaagatggtttggctgccttggagatttggagatctgatgcc
acgatgaggactcacacacgggggctcccagtgtgtttttcatatatttgctttgcttt
gtgtcagcctacatcaccgacgagaacccagaagttatgattcccttcaccaatgccaac
tacgacagccatcccatgctgtacttctccagggcagaagtggcggagctgcagcgcagg
gctgccagctcgcacgagcacattgcagcccgcctcacggaggctgtgcacacgatgctg
tccagccccttggaatacctccctccctgggatcccaaggactacagtgcccgctggaat
gaaatttttggaaacaacttgggtgccttggcaatgttctgtgtgctgtatcctgagaac

attgaagcccgagacatggccaaagactacatggagaggatggcagcgcagcctagttgg
tag

>15:16:40|GENSCAN_predicted_peptide_2|92_aa
MNTKIQGIREERRPGAGGILQRPGLCGISNDLAPDYGISSCAVQNSHVDTGKEKERWMRE
TVVSVAEMCSLNGHGRRRRRSKKAILKIPKGT

>15:16:40|GENSCAN_predicted_CDS_2|279_bp
atgaacaccaaaattcaagggatcagggaggagagaaggccaggtgcaggaggaatactt
caaagaccaggcttatgtgggataagcaatgatctggccccgactatggtatctcctcc
tgtgctgtccagaactcgcatgttgacactggaaaggaaaaggaaagatggatgagagag
actgtggtaagcgtggctgaaatgtgttcccttaatggacatggcagaaggagacgaaga
agtaaaaaagccattcttaagatacctaagggcacctga

>15:16:40|GENSCAN_predicted_peptide_3|690_aa
MVMTKLVKDAPWDEVPLAHSLVGFATAYDFLYNYLSKTQQEKFLEVIANASGYMYETSYR
RGWGFQYLHNHQPTNCMALLTGSLVLMNQGYLQEAYLWTKQVLTIMEKSLVLLREVTDGS
LYEGVAYGSYTTRSLFQYMFLVQRHFNINHFGHPWLKQHFAFMYRTILPDFGTPTLHYFE
DWGVVTYGSALPAEINRSFLSFKSGKLGGRAIYDIVHRNKYKDWIKGWRNFNAGHEHPDQ
NSFTFAPNGVPFITEALYGPKYTFFNNVLMFSPAVSKSCFSPWVGQVTEDCSSKWSKYKH
DLAASCQGRVVAAEEKNGVVFIRGEGVGAYNPQLNLKNVQRNLILLHPQLLLLVDQIHLG
EESPLETAASFFHNVDVPFEETVVDGVHGAFIRQRDGLYKMYWMDDTGYSKKATFASVTY
PRGYPYNGTNYVNVTMHLRSPITRAAYLFIGPSIDVQSFTVHGDSQQLDVFIATSKHAYA
TYLWTGEATGQSAFAQVIADHHKILFDRNSVIKSSIVPEVKDYAAIVEQNLQHFKPVFQL
LEKQILSRVRNTASFRKTAERLLRFSDKRQTEEAIDRIFAISQQQQQQSKSKKNRRAGKR
YKFVDAVPDIFAQIEVNEKKIRQKAQILAQKELPIDEDEEMKDLLDFADVTYEKHKNGGL
IKGRFGQARMVVEVNGVIVKRFQSADQCFI

>15:16:40|GENSCAN_predicted_CDS_3|2073_bp
atggtgatgaccaagttggtgaaagatgctccttgggatgaggtcccgcttgctcactcc
ctggttggttttgccactgcttatgacttcttgtacaactacctgagcaagacacaacag
gagaagtttcttgaagtgattgccaatgcctcagggtatatgtatgaaacttcatacagg
agaggatggggatttcaatacctgcacaatcatcagcccaccaactgtatggctttgctc
acaggaagcctagtcctgatgaatcaaggatatcttcaagaagcgtacttatggaccaaa
caagttctgaccatcatggagaaatctctggtcttgctcagggaggtgacggatggctcc
ctctatgaaggagttgcgtatggcagctacaccactagatcactcttccagtacatgttt
ctcgtccagaggcacttcaacatcaaccactttggccatccgtggcttaaacaacacttt
gcatttatgtatagaaccatcctgccagactttggcacccctacactgcattattttgaa
gactggggtgtcgtgacttatggaagtgcactacctgcagaaatcaatagatctttcctt
tccttcaagtctggaaaactgggggacgtgcaatatatgacattgtccacagaaacaag
tacaaagattggatcaaaggatggagaaattttaatgcagggcatgaacatcctgatcaa
aactcatttacttttgctcccaatggtgtgcctttcattactgaggctctgtatgggcca
aagtacaccttcttcaacaatgttttgatgttttccccagctgtttcaaagagctgcttt
tctccctgggtgggtcaggtcacagaagactgctcatcaaaatggtctaaatacaagcat
gacctggcagctagttgtcaggggagggtggttgcagcagaggagaaaaatggggtggtt
ttcatccgaggagaaggtgtgggagcttataacccccagctcaacctgaagaatgttcag
aggaatctcatcctcctacatccacagctgcttctccttgtagaccaaatacacctggga
gaggagagtcccttggagacagcagcgagcttcttccataatgtggatgttccttttgag
gagactgtggtagatggtgtccatggggctttcatcaggcagagagatggtctctataaa
atgtactggatggacgatactggctacagcaagaaagcaacctttgcctcagtgacatat
cctcggggctatccctacaacgggacaaactatgtgaatgtcaccatgcacctccgaagt
cccatcaccagggcagcttacctcttcatagggccatctatagatgttcagagcttcact
gtccacggagactctcagcaactggatgtgttcatagccaccagcaaacatgcctacgcc
acatacctgtggacaggtgaggccacaggacagtccgcctttgcacaggtcattgctgat
catcacaaaattctgtttgaccggaattcagtcatcaagagcagcattgtccctgaggtg
aaggactatgctgctattgtggaacagaacttgcagcattttaaaccagtgtttcagctg
ctggagaagcagatactgtcccgagtccggaacacagctagctttaggaagactgctgaa
cgcctgctgagattttcagataagagacagactgaggaggccattgacaggatttttgcc

```
atatcacagcaacagcagcagcaaagcaagtcaaagaaaaaccgaagggcaggcaaacgc
tataaatttgtggatgctgtccctgatatttttgcacagattgaagtcaatgagaaaaag
attagacagaaagctcagattttggcacagaaagaactacccatagatgaagatgaagaa
atgaaagacctttagattttgcagatgtaacatacgagaaacataaaaatgggggcttg
attaaaggccggtttggacaggcacggatggtggttgaagtgaatggagtcattgtaaaa
cgctttcaatcagctgaccaatgctttatataa

>15:16:40|GENSCAN_predicted_peptide_4|95_aa
MVKKKISDSESDDSKSKKKTDAADKPRGFARGLDPERIIGATDSSGELMFLMKWKDSDEA
DLVLAKEASMKCPQIVIAFYEERLTWHSCPEDEAQ

>15:16:40|GENSCAN_predicted_CDS_4|288_bp
atggtaaaaaagaaaatatctgacagtgaatctgatgacagcaaatcaaagaagaaaaca
gatgctgctgacaaaccaagaggatttgccagaggtcttgatcctgaaagaataattggt
gccacagacagcagtggagaattgatgtttctcatgaaatggaaagattcagatgaggca
gacttggtgctggcaaaagaagcaagtatgaagtgtcctcaaattgtaattgctttttat
gaagagagactaacttggcattcttgtccagaagatgaagctcaataa

>15:16:40|GENSCAN_predicted_peptide_5|199_aa
SEEPGDRELEAGEQNPGAPGEEGTPGQRLEPLLHDHQDLRAQIFTNTVDNARIVLQINAC
FAAVDFSVKYETELAMCQSVESDIHGVHKVIDDTNVTRLQLETEIKALKKQLLFMKNHEE
EMEQLNGILLHLESELAQTQAEGQHQAQEYEALLNIKVKLEAEIATYNNLLEDGEDFNLG
DALDNSNSMQTIQKTPPAQ

>15:16:40|GENSCAN_predicted_CDS_5|600_bp
agtgaggagcctggagaccgagaactggaagctggagagcaaaatccgggagcacctggg
gaagaagggaccccaggccagagactggagccattacttcatgaccatcaggacctgagg
gctcagatattcacaaatactgtggacaatgcccgcatcgttctgcaaatcaatgcctgt
tttgctgctgttgacttcagtgtcaagtatgagacagagctggccatgtgccagtctgtg
gagagcgacatccatggggtccacaaggtcattgatgacaccaatgtcactcggctgcag
ctggagacagagatcaaggctctcaaaaagcagctgctcttcatgaagaaccatgaagag
gaaatggaacagctcaatgggatcctgctgcacctggagtcagagctggcacagacccag
gcagagggacagcaccaggcccaggagtatgaggccctgctgaacattaaggtcaagctg
gaggctgagatagccacctacaacaacctgctggaagatggcgaggacttcaatcttggt
gatgccctggacaacagcaactccatgcaaaccatccaaaagacaccacccgcccaatag
```

Appendix B: RepeatMasker summary output of repetitive elements

	number of elements	length occupied	percentage of sequence
SINEs:	36	7445 bp	8.01%
ALUs	17	4481 bp	4.82%
MIRs	19	2964 bp	3.19%
LINEs:	28	8579 bp	9.23%
LINE1	7	3043 bp	3.27%
LINE2	18	4819 bp	5.19%
L3/CR1	3	717 bp	0.77%
LTR elements:	12	4750 bp	5.11%
MaLRs	8	2884 bp	3.10%
ERVL	2	999 bp	1.08%
ERV_classI	2	867 bp	0.93%
ERV_classII	0	0 bp	0.00%
DNA elements:	11	2649 bp	2.85%
MER1_type	9	2492 bp	2.68%
MER2_type	0	0 bp	0.00%
Unclassified:	0	0 bp	0.00%
Total interspersed repeats:		23423 bp	25.21%
Small RNA:	0	0 bp	0.00%
Satellites:	0	0 bp	0.00%
Simple repeats:	0	0 bp	0.00%