Chimp Chunk 3-14

Annotation by Matthew Kwong, Ruth Howe, and Hao Yang

Ruth Howe
Bio 434W April 1, 2010

INTRODUCTION

De novo annotation is the process by which a finished genomic sequence is searched for features of interest—eg., genes, pseudogenes, repeats—and marked accordingly. The accuracy of the annotation depends on the extent of the extant gene databases, and it involves multiple software for gene prediction, sequence comparison, and repeat masking.

The Chimp Chunks project aims to provide an introduction to gene annotation in a team learning environment as preparation for annotating the *Drosophila grimshawi* fosmids finished during the first part of the Bio 4342/434W course. Small chunks of chimpanzee sequence (fewer than 100 Kb) were assigned to teams of three to be annotated in reference to the closely related human genome, as if the chimpanzee genome were not yet annotated. This report will focus on the annotation of chunk 3-14, which was carried out in collaboration with Matthew Kwong and Hao Yang.

INITIAL *GENSCAN* FEATURES AND AREAS OF INTEREST

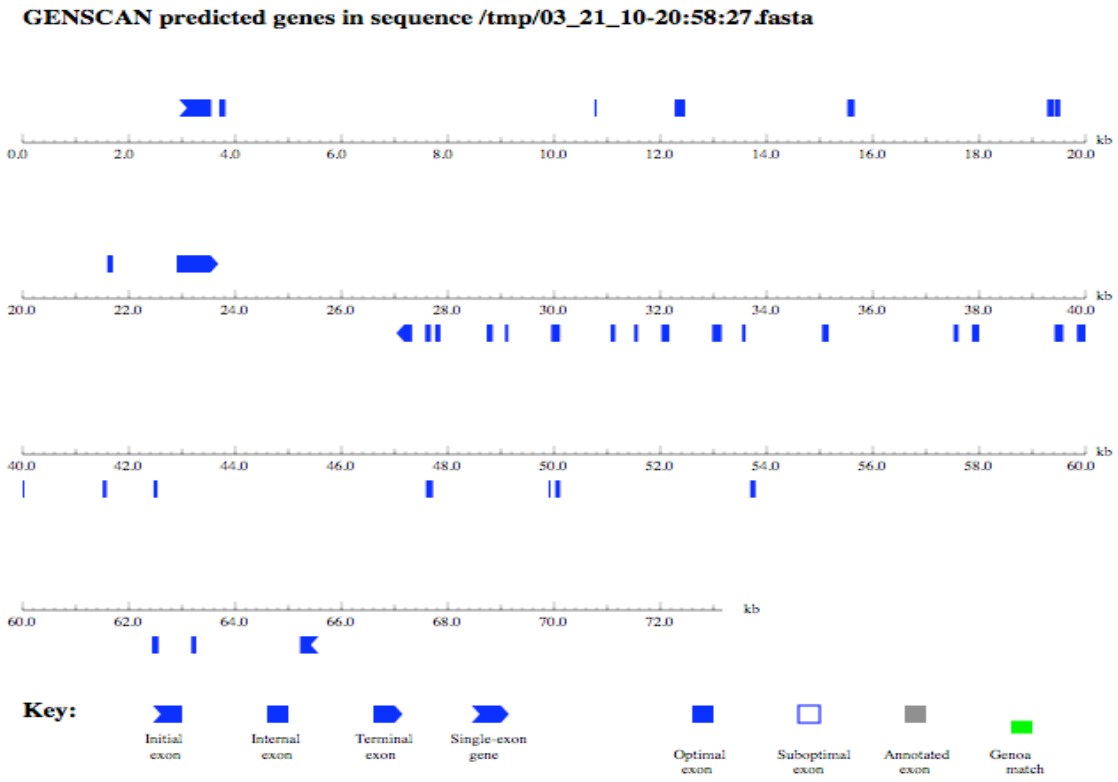


Fig. 1: Initial *Genscan* map of features predicted for chunk 3-14. The blue exons placed above the map's Kb scale indicate that the exons run in the forward direction, while those placed below the scale represent genes encoded on the reverse strand.

As shown in Fig. 1, the *Genscan* prediction for chunk 3-14 consisted of two multi-exon genes, one each on the forward and reverse strands. The predicted gene on the forward strand has 9 putative exons covering bases 3,902-23,533, while that on the reverse strand contains 25 exons running from position 65,420-27183 (Fig. 2).

Gn.Ex	Type	S	.Begin	...End	.Len	Fr	Ph	I/Ac	Do/T	CodRg	P....	Tscr..
1.01	Init	+	3092	3549	458	1	2	11	-57	521	0.488	25.22
1.02	Intr	+	3699	3816	118	0	1	48	53	191	0.852	11.15
1.03	Intr	+	10763	10802	40	1	1	109	75	49	0.892	2.58
1.04	Intr	+	12271	12466	196	2	1	22	101	170	0.862	9.25
1.05	Intr	+	15519	15654	136	0	1	64	80	62	0.913	2.65
1.06	Intr	+	19280	19414	135	1	0	52	21	134	0.809	2.94
1.07	Intr	+	19428	19526	99	2	0	44	84	157	0.858	10.29
1.08	Intr	+	21594	21698	105	2	0	108	80	139	0.966	14.59
1.09	Term	+	22901	23533	633	1	0	104	42	537	0.997	44.00
1.10	PlyA	+	24505	24510	6							1.05

2.26	PlyA	-	24766	24761	6							1.05
2.25	Term	-	27327	27183	145	2	1	85	40	223	0.999	13.60
2.24	Intr	-	27682	27577	106	2	1	81	61	119	0.964	6.85
2.23	Intr	-	27861	27762	100	0	1	113	78	87	0.956	8.96
2.22	Intr	-	28846	28735	112	0	1	35	91	62	0.514	0.66
2.21	Intr	-	29139	29078	62	0	2	88	77	39	0.506	-0.59
2.20	Intr	-	30111	29947	165	0	0	101	70	253	0.999	24.04
2.19	Intr	-	31152	31065	88	2	1	37	105	58	0.543	1.45
2.18	Intr	-	31579	31509	71	1	2	69	115	12	0.606	-0.94
2.17	Intr	-	32169	32020	150	0	0	98	80	175	0.999	17.14
2.16	Intr	-	33154	32977	178	0	1	91	79	103	0.714	8.70
2.15	Intr	-	33604	33540	65	1	2	93	93	126	0.907	10.20
2.14	Intr	-	35169	35044	126	0	0	99	103	79	0.931	10.46
2.13	Intr	-	37611	37519	93	0	0	110	119	38	0.956	8.34
2.12	Intr	-	37999	37864	136	0	1	74	-13	148	0.751	2.95
2.11	Intr	-	39583	39418	166	2	1	74	94	230	0.999	20.30
2.10	Intr	-	40030	39841	190	1	1	62	110	315	0.987	29.34
2.09	Intr	-	41583	41506	78	0	0	56	84	87	0.938	3.93
2.08	Intr	-	42537	42464	74	1	2	140	73	115	0.999	13.41
2.07	Intr	-	47721	47589	133	0	1	86	48	144	0.869	9.50
2.06	Intr	-	49932	49898	35	1	2	71	37	32	0.569	-6.58
2.05	Intr	-	50115	50023	93	1	0	84	116	150	0.861	16.42
2.04	Intr	-	53792	53691	102	0	0	74	86	41	0.481	1.73
2.03	Intr	-	62557	62435	123	2	0	47	81	83	0.412	3.14
2.02	Intr	-	63265	63175	91	1	1	83	80	11	0.632	-1.55
2.01	Init	-	65420	65220	201	2	0	73	68	98	0.494	5.22

Fig. 2: The *Genscan* predicted exons for features 1 (above) and 2 (right) of chunk 3-14. The final annotations for the features for the fosmid are tabulated below.

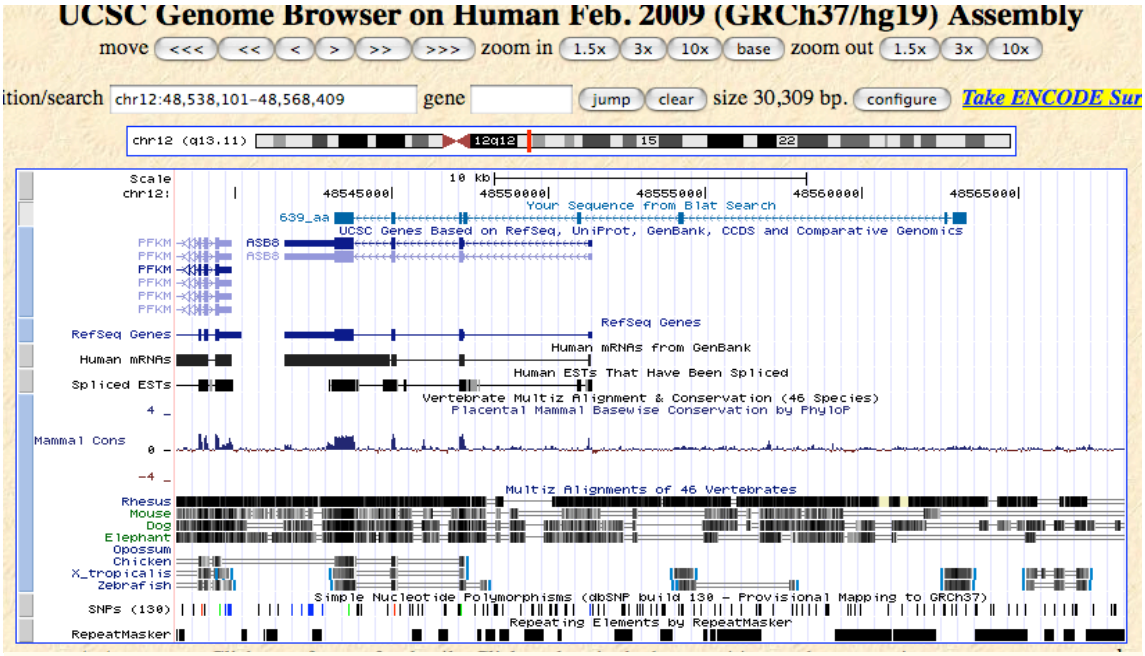
Features For Chimp Chunk 3-14				
Feature Name	Accession No.	No. exons	bp span	Orientation
prohibitin pseudogene	NP_002625	2	3092-3816	+ strand
ASB8	CAG33617	3	19520-23652	+ strand
6-PFK muscle isoform 1	NP_001160158	24	63378-27308	- strand
partial *SUMO1/SENP1*	AF149770	2	65294-70204	+ strand

INVESTIGATION OF PREDICTED FEATURE 1

The investigation of the first feature was carried out primarily by Hao Yang. Our first step was to use the BLAST-Like Alignment Tool (BLAT) to compare the predicted chimp feature to the human genome (Fig. 3). This initial scan suggested the presence of multiple features within the region, including a gene for the *ankyrin repeat and SOCS box protein 8* (*ASB8*).

Before working on the areas of *Genscan's* prediction not supported by BLAT, we first confirmed that the high degree of identity between chunk 3-14 and human *ASB8* represented a functional gene. To accomplish this, we used the Basic Local Alignment Search Tool (BLAST) to compare the predicted feature to the human genome (Fig. 4). This showed convincingly that the three exons of *ASB8* were covered by four exons in the *Genscan* prediction (two of which were shown by BLAST analysis to be merged into one true exon), and this finding was supported by a reverse alignment of *ASB8* to the

chunk 3-14 sequence (data not shown). Ensembl software subsequently confirmed that a fourth exon predicted by *Genscan* and supported by *RefSeq* was, in fact, an untranslated region (UTR) (Fig. 5).



RefSeq Gene ASB8

RefSeq: NM_024095.3 **Status:** Validated
Description: Homo sapiens ankyrin repeat and SOCS box-containing 8 (ASB8), mRNA.
Organism: Homo sapiens
UCSC browser: NM_024095 on hg19
CDS: 3' complete
Entrez Gene: 140461
PubMed on Gene: ASB8
PubMed on Product: ankyrin repeat and SOCS box protein 8

Summary of ASB8

mRNA/Genomic Alignments

SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
2602	99.6%	12	+	41577907	41587826	NM_024095	1	2610	2629

Fig. 3: The BLAT alignment of *Genscan* predicted gene 1 to human chromosome 12 showing high coverage of *ankyrin repeat and SOCS box protein 8* (*ASB8*) in Human mRNAs, Spliced ESTS, and RefSeq Genes tracks. Note, however, that there is no coverage of the extreme exons by *Genbank*, *RefSeq*, or any other gene database, suggesting a miscall by *Genscan*.

```
>lcl|57543 gi|13129098|ref|NP_077000.1| ankyrin repeat and SOCS box protein
8 [Homo sapiens]
Length=288

                                                    Sort alignments for this subject sequence by:
                                                      E value  Score  Percent identity
                                                      Query start position  Subject start position

 Score =  392 bits (1008),  Expect = 3e-112
 Identities = 194/196 (98%), Positives = 196/196 (100%), Gaps = 0/196 (0%)
 Frame = +1

Query  23020  QVNALDGYNRTALHYAAEKDEACVEVLLEYGANPNALDGNRDTPLHWAAFKNNAECVRAL  23199
              +VNALDGYNRTALHYAAEKDEACVEVLLEYGANPNALDGNRDTPLHWAAFKNNAECVRAL
Sbjct  78     EVNALDGYNRTALHYAAEKDEACVEVLLEYGANPNALDGNRDTPLHWAAFKNNAECVRAL  137

Query  23200  LESGASVNALDYNNDTPLSWAAMKGNLESVSILLDYGAEVRVINLIGQTPISRLVALLVR  23379
              LESGASVNALDYNNDTPLSWAAMKGNLESVSILLDYGAEVRVINLIGQTPISRLVALLVR
Sbjct  138    LESGASVNALDYNNDTPLSWAAMKGNLESVSILLDYGAEVRVINLIGQTPISRLVALLVR  197

Query  23380  GLGTEKEDSCFELLHRAVGHFELRKNGTMPREVSRDPQLCEKLTVLCSAPGTLKTLARYA  23559
              GLGTEKEDSCFELLHRAVGHFELRKNGTMPREV+RDPQLCEKLTVLCSAPGTLKTLARYA
Sbjct  198    GLGTEKEDSCFELLHRAVGHFELRKNGTMPREVARDPQLCEKLTVLCSAPGTLKTLARYA  257

Query  23560  VRRSLGLQYLPDAVKG  23607
              VRRSLGLQYLPDAVKG
Sbjct  258    VRRSLGLQYLPDAVKG  273


 Score = 87.4 bits (215),  Expect = 3e-20
 Identities = 43/43 (100%), Positives = 43/43 (100%), Gaps = 0/43 (0%)
 Frame = +2

Query  19520  MSSSMWYIMQSIQSKYSLSERLIRTIAAIRSFPHDNVEDLIRG  19648
              MSSSMWYIMQSIQSKYSLSERLIRTIAAIRSFPHDNVEDLIRG
Sbjct  1      MSSSMWYIMQSIQSKYSLSERLIRTIAAIRSFPHDNVEDLIRG  43


 Score = 80.5 bits (197),  Expect = 4e-18
 Identities = 36/36 (100%), Positives = 36/36 (100%), Gaps = 0/36 (0%)
 Frame = +2

Query  21716  GADVNCTHGTLKPLHCACMVSDADCVELLLEKGAEV  21823
              GADVNCTHGTLKPLHCACMVSDADCVELLLEKGAEV
Sbjct  44     GADVNCTHGTLKPLHCACMVSDADCVELLLEKGAEV  79
```

Fig. 4: Alignment of the *Genscan* predicted peptide with human *ASB8*, masked for low-complexity sequences. When not masked, the *Genscan* prediction can be seen to cover all 288 amino acids of human *ASB8* with 100% identity on the + strand (data not shown).

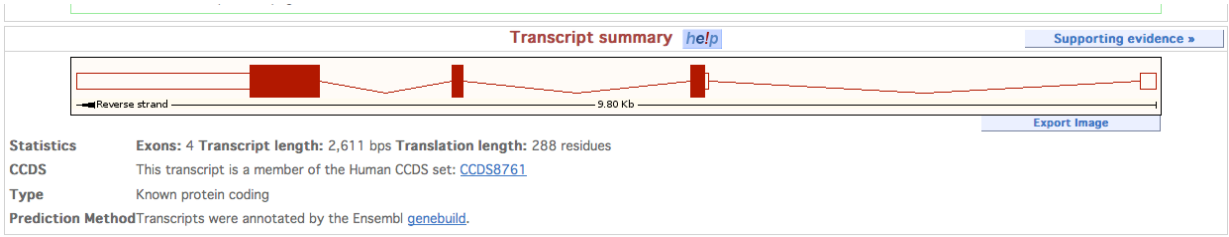


Fig. 5: Ensembl showing that the fourth exon of *ASB8* is a UTR, as are regions of exons 1 and 3. In the Ensembl display, translated exons are denoted with solid boxes, while UTRs are drawn as hollow rectangles.

On the basis of the above evidence, we decided that *Genscan's* predicted feature does contain the four exons of *ASB8* and that the chimp gene is orthologous to human *ASB8*. Hao then concentrated on the remaining five unsupported exons, using BLAST to search for homology. He found that the first two exons of *Genscan's* prediction matched to *prohibitin*, but that the match was imperfect and contained premature stop codons, suggesting it to be a pseudogene (Fig. 6). The remaining predicted exons had no significant alignments in BLAST (Fig. 7) or expressed sequence tag (EST) support (Fig. 3), and Hao concluded that *Genscan* had mispredicted them.

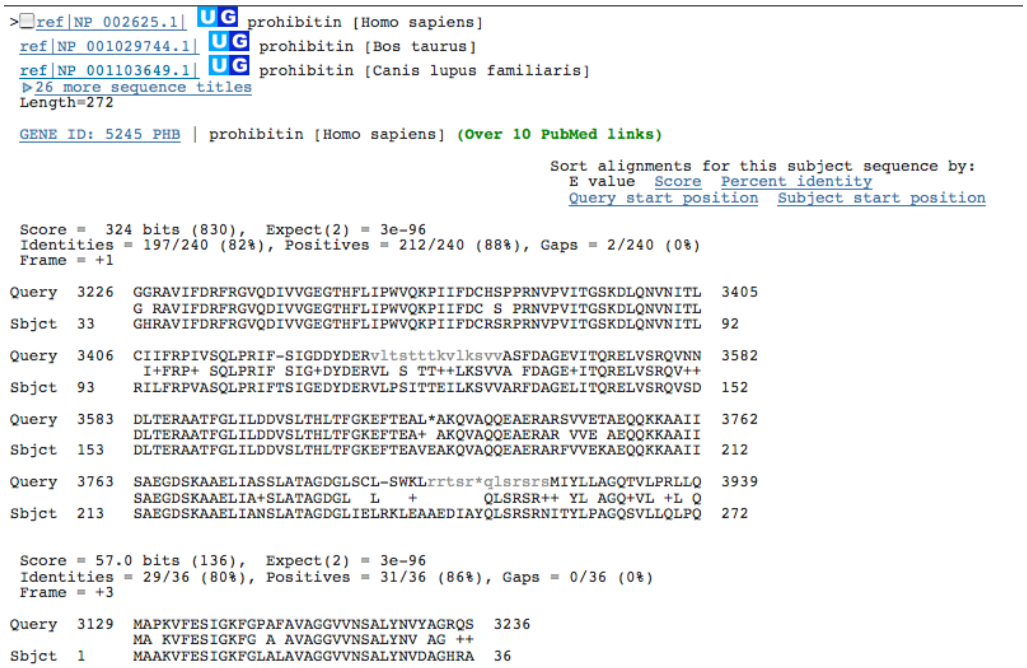


Fig. 6: A section of the alignment of *Genscan's* predicted feature 1 and human *prohibitin* showing the gaps, mutations, and premature stop codons indicative of a pseudogene.

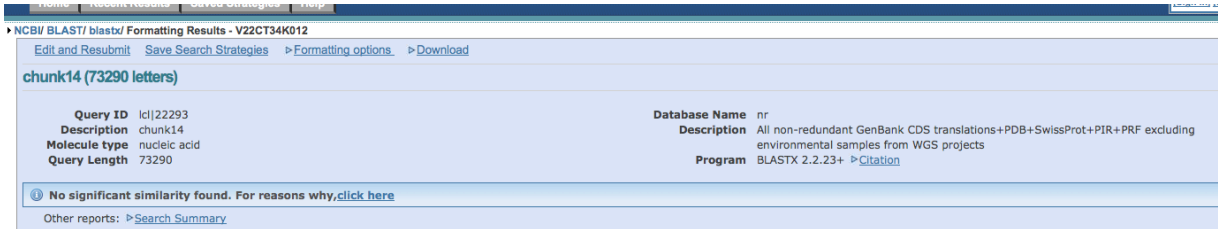
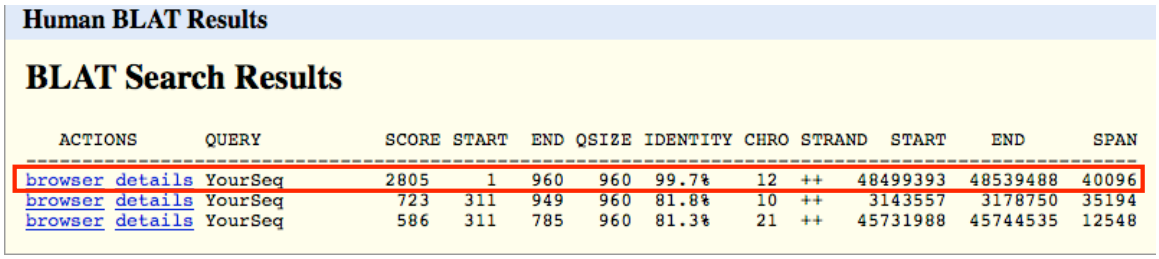


Fig. 7: Lack of BLAST alignments for the two *Genscan*-predicted exons between the *prohibitin* pseudogene and *ASB8*.

INVESTIGATION OF PREDICTED FEATURE 2

Matthew Kwong and I inspected the second feature in the *Genscan* prediction. Our first step was to BLAT the predicted chimpanzee peptide against the human genome. This revealed a close match on the minus-strand to human *6-phosphofructokinase* muscle isoform 1 (*6-PFK*) mRNA over the bulk of the prediction, but with the initial exon being covered by the plus-strand reference sequence mRNA for *SUMO-1/sentrin specific peptidase 1* (*SENP1*) (Fig. 8).



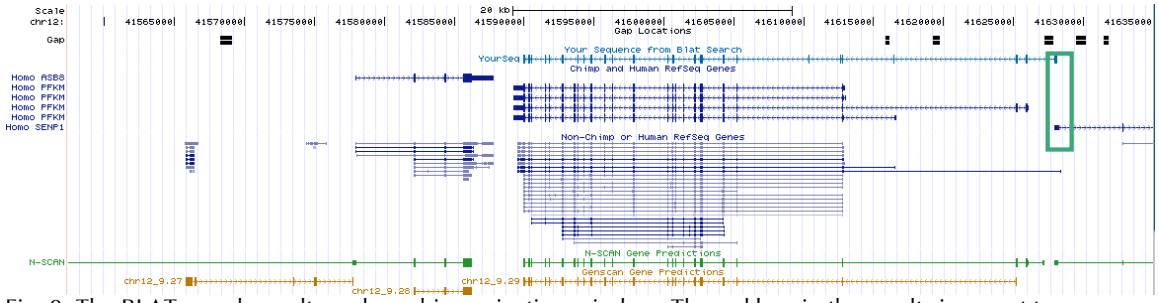


Fig. 8. The BLAT search results and graphic navigation window. The red box in the results image at top indicates the high quality of the lower alignment, well above the 97.5% identity threshold. The green box in the graphic navigation window at bottom indicates the extension of the *Genscan* prediction (light blue) beyond the termination of *6-PFK* into the mRNA coverage of *SUMO1/SENP1* (dark blue). The dark blue sequences represent *RefSeq* RNA coverage, supporting our designation of this feature as a combination of *6-PFK* and *SUMO1/SEMP1*.

To confirm these alignments, we used BLAST to align first the entire chunk 3-14 (Fig. 9), then the predicted peptide (Fig. 10) against the human genome. This showed that our primary matches in feature 2 corresponded to the BLAT alignment and supported the hypothesis that *Genscan* had lumped the initial *SUMO1/SENP1* gene in with *6-PFK*.

Sequences producing significant alignments:

Accession	Description	Max score	Total score	Query coverage	E value	Max ident	Links
Transcripts							
NM_024095.3	Homo sapiens ankyrin repeat and SOCS box-containing 8 (ASB8), mR	3986	4742	3%	0.0	100%	GM
NM_002634.2	Homo sapiens prohibitin (PHB), mRNA	1362	1362	1%	0.0	90%	GM
NM_000289.4	Homo sapiens phosphofructokinase, muscle (PFKM), mRNA	977	5464	4%	0.0	100%	GM
NM_032689.3	Homo sapiens zinc finger protein 607 (ZNF607), mRNA	965	965	1%	0.0	84%	GM
NM_014554.2	Homo sapiens SUMO1/sentrin specific peptidase 1 (SENP1), mRNA	712	805	0%	0.0	100%	GM
NM_004589.2	Homo sapiens SCO cytochrome oxidase deficient homolog 1 (yeast) (S	665	868	1%	0.0	86%	GM

Fig. 9: Major alignments of chunk 3-14 to the human genome, showing the *ASB8* and *prohibitin* matches for *Genscan* feature 1, as well as 100% identity matches for *6-PFK* and *SUMO1/SENP1*.

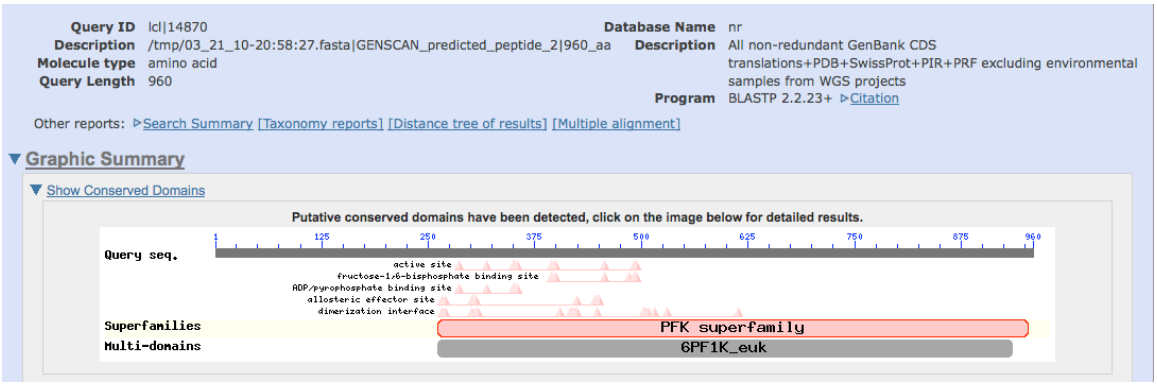


Fig. 10: BLAST result for the *Genscan* predicted peptide 2 versus the human genome, showing that the peptide matches the conserved PFK superfamily.

To test these alignments we used BLAST to align the chimp sequence to human peptide, then *vice versa*, for both putative genes (Figs. 11 and 12). In both cases, the searches resulted in the same high-quality alignments, suggesting that our hypothesis was correct. Additionally, we identified one partial *SUMO1/SENP* exon that *Genscan* had missed (likely because it spans the gap between two chunks), and included that in our final annotation assembly.

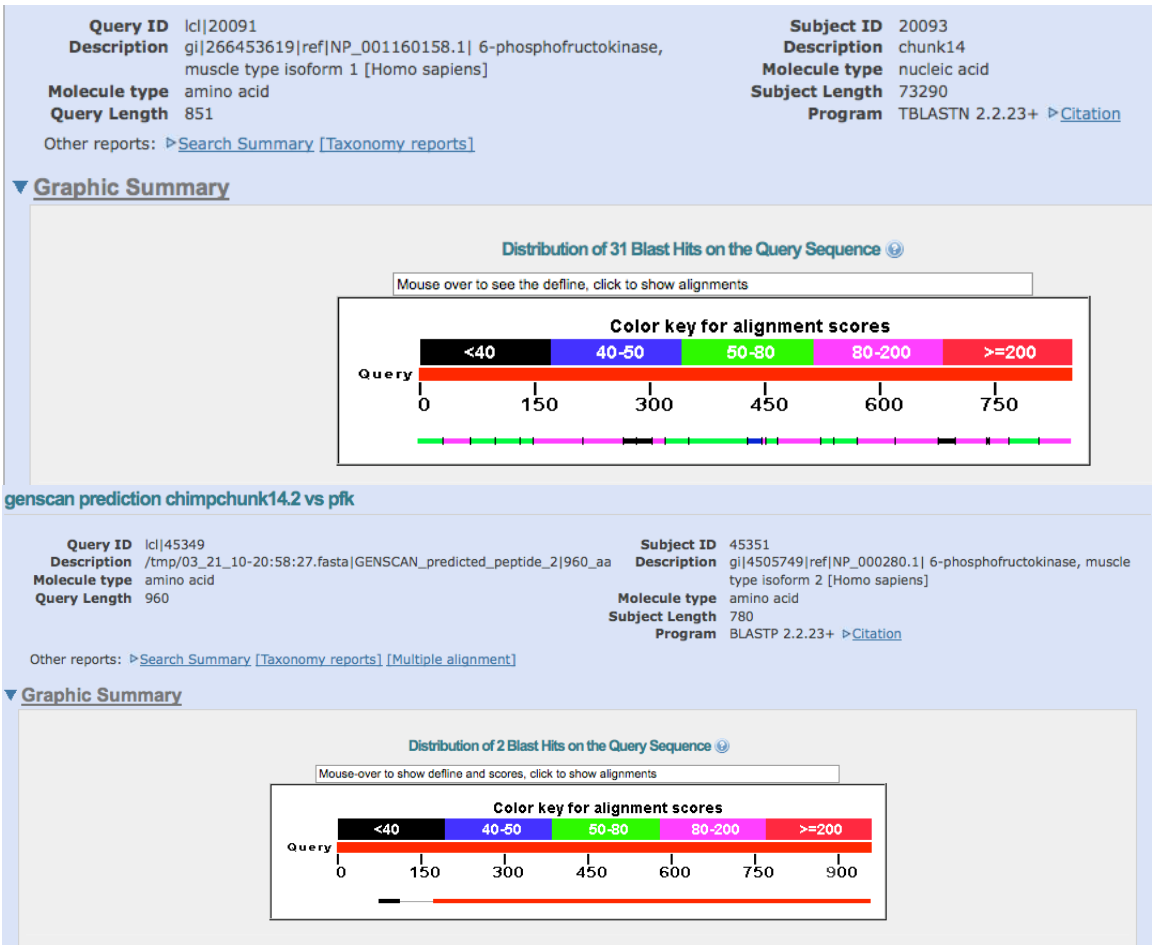
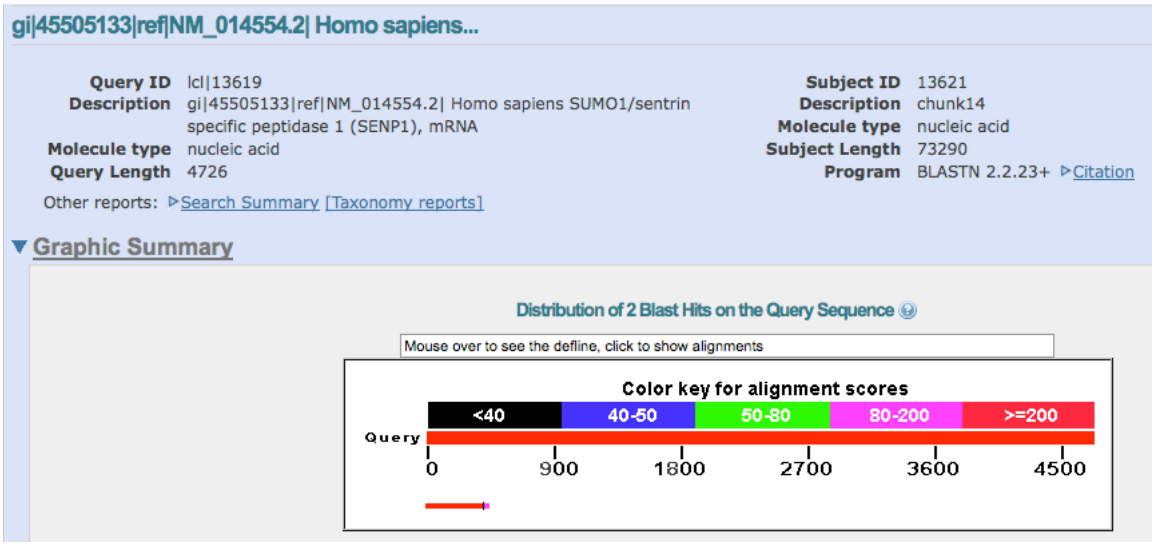


Fig. 11: BLAST results for *6-PFK* versus chunk 3-14 (top) and the predicted peptide against *6-PFK* isoform 2 (bottom). Note that the initial exon of the predicted peptide does not map to *6-PFK*, supporting the hypothesis that it belongs to *SUMO1/SENP1*. Also, this second BLAST was carried out against isoform 2, which lacks 2 small exons at the beginning of the gene and resulted in discrepancies in this area. This issue could be resolved by a BLAST search against isoform 1, supporting the decision to annotate the chunk accordingly.



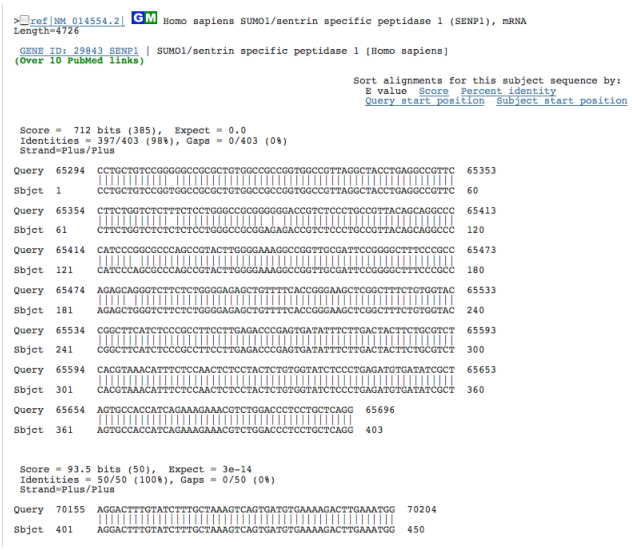


Fig. 12: The BLAST alignments of human *SUMO1/SENP1* to chunk 3-14 (top) and chunk 3-14 to the *SUMO1/SENP1* reference sequence (left). The high degree of homology and absence of acquired stop codons suggests that this is a functional gene that extends beyond the end of chunk 3-14.

The quality of the BLAST and BLAT alignments convinced us that our hypothesis was correct. We annotated the region according to the exons enumerated in Figure 13.

6-PFK amino acid #s	Chunk 3-14 bp start	Chunk 3-14 bp end	exon
01-027	63378	63298	1
28-68	62680	62558	2
69-100	50238	50143	3
100-124	42660	42586	4
125-150	41703	41628	5
151-215	40152	39958	6
215-268	39701	39542	7
269-284	38795	38748	8
285-320	38302	38195	9
321-352	38084	37986	10
353-384	37733	37826	11
384-425	35291	32142	12
426-447	33724	33661	13
447-468	33164	33099	14
464-518	32310	32142	15
518-542	31704	31630	16
542-571	31276	31190	17
572-622	31100	30948	18
623-677	30233	30069	19
677-698	29264	29199	20
698-735	28970	28859	21
735-770	27986	27879	22
769-804	27805	27697	23
805-851	27448	27308	24
SUMO1 bps			
1-403	65294	65696	1
401-450	70155	70204	2

2.26	PlyA	-	24766	24761	6						1.05
2.25	Term	-	27327	27183	145	2	1	85	40	223 0.999	13.60
2.24	Intr	-	27682	27577	106	2	1	81	61	119 0.964	6.85
2.23	Intr	-	27861	27762	100	0	1	113	78	87 0.956	8.96
2.22	Intr	-	28846	28735	112	0	1	35	91	62 0.514	0.66
2.21	Intr	-	29139	29078	62	0	2	88	77	39 0.506	-0.59
2.20	Intr	-	30111	29947	165	0	0	101	70	253 0.999	24.04
2.19	Intr	-	31152	31065	88	2	1	37	105	58 0.543	1.45
2.18	Intr	-	31579	31509	71	1	2	69	115	12 0.606	-0.94
2.17	Intr	-	32169	32020	150	0	0	98	80	175 0.999	17.14
2.16	Intr	-	33154	32977	178	0	1	91	79	103 0.714	8.70
2.15	Intr	-	33604	33540	65	1	2	93	93	126 0.907	10.20
2.14	Intr	-	35169	35044	126	0	0	99	103	79 0.931	10.46
2.13	Intr	-	37611	37519	93	0	0	110	119	38 0.956	8.34
2.12	Intr	-	37999	37864	136	0	1	74	-13	148 0.751	2.95
2.11	Intr	-	39583	39418	166	2	1	74	94	230 0.999	20.30
2.10	Intr	-	40030	39841	190	1	1	62	110	315 0.987	29.34
2.09	Intr	-	41583	41506	78	0	0	56	84	87 0.938	3.93
2.08	Intr	-	42537	42464	74	1	2	140	73	115 0.999	13.41
2.07	Intr	-	47721	47589	133	0	1	86	48	144 0.869	9.50
2.06	Intr	-	49932	49898	35	1	2	71	37	32 0.569	-6.58
2.05	Intr	-	50115	50023	93	1	0	84	116	150 0.861	16.42
2.04	Intr	-	53792	53691	102	0	0	74	86	41 0.481	1.73
2.03	Intr	-	62557	62435	123	2	0	47	81	83 0.412	3.14
2.02	Intr	-	63265	63175	91	1	1	83	80	11 0.632	-1.55
2.01	Init	-	65420	65220	201	2	0	73	68	98 0.494	5.22

Fig. 13: Table of Exons for *6-PFK* and *SUMO1/SENP1* in chunk 3-14. Note that the base pairs are close but not identical to those in the *Genscan* prediction (right) and that the *Genscan* table lists the predicted exons from terminal to initial, as they would appear along the + strand.

REPEATMASKER

Matthew Kwong tested our sequence for repetitious regions using the program *RepeatMasker*. This showed that chunk 3-14 is relatively clear of repeats. Notable repeats were defined as those near to or over 500 bp long, but neither these nor the remaining repeats overlapped with any exons. Thus, the repeats represented no threat to our hypotheses.

Summary:

```
file name: RM2sequpload_1269893996
sequences:                1
total length:       73290 bp  (70096 bp excl N/X-runs)
GC level:          43.14 %
bases masked:       23889 bp ( 32.60 %)
```

	number of elements*	length occupied	percentage of sequence
SINEs:	48	10474 bp	14.29 %
ALUs	37	9117 bp	12.44 %
MIRs	11	1357 bp	1.85 %
LINEs:	16	7062 bp	9.64 %
LINE1	4	3949 bp	5.39 %
LINE2	11	3031 bp	4.14 %
L3/CR1	1	82 bp	0.11 %
LTR elements:	4	4245 bp	5.79 %
ERVL	0	0 bp	0.00 %
ERVL-MaLRs	1	351 bp	0.48 %
ERV_classI	3	3894 bp	5.31 %
ERV_classII	0	0 bp	0.00 %
DNA elements:	7	1172 bp	1.60 %
hAT-Charlie	7	1172 bp	1.60 %
TcMar-Tigger	0	0 bp	0.00 %
Unclassified:	0	0 bp	0.00 %
Total interspersed repeats:		22953 bp	31.32 %
Small RNA:	0	0 bp	0.00 %
Satellites:	0	0 bp	0.00 %
Simple repeats:	9	480 bp	0.65 %
Low complexity:	8	456 bp	0.62 %

```
* most repeats fragmented by insertions or deletions
  have been counted as one element

The query species was assumed to be homo
RepeatMasker version open-3.2.8 , default mode
```

position begin	end	matching repeat	repeat class/family
4218	4703	L1MC4a	LINE/L1
5031	5546	L1MC4a	LINE/L1
6488	6975	L2c	LINE/L1
10712	11334	LTR78	LTR/ERV1
12741	13043	AluSx	Sine/Alu
36269	36831	L1MD1	LINE/L1
45244	45662	L2a	LINE/L2
55014	55613	MER4B	LTR/ERV1
55614	56513	L1MC1	Line/L1
57101	58303	HERVH-int	LT/ERV1
58290	59401	HERVH-int	LTR/ERV1

Fig. 14: The *RepeatMasker* readout (left) shows that the percent repeats in this region is low for a primate genome at only 32.60%, and that the makeup of the repeats is normal. The above table lists those repeats that were near or over 500 bp in length. These 11 repeats are noted as features in the final annotation image.

SYNTENY

Hao Yang tested the features we identified through assessment of *Genscan* predicted features 1 and 2 for synteny to the human genome. In humans, *ASB8*, *6-PFK*, *prohibitin* pseudogene, and *SUMO1/SENP1* are located on chromosome 12. Using BLAT, Hao was able to confirm that chunk 3-14 also belongs to chromosome 12 of the chimpanzee genome, and that the genes here are indeed syntenic with the human genes (Fig. 15).

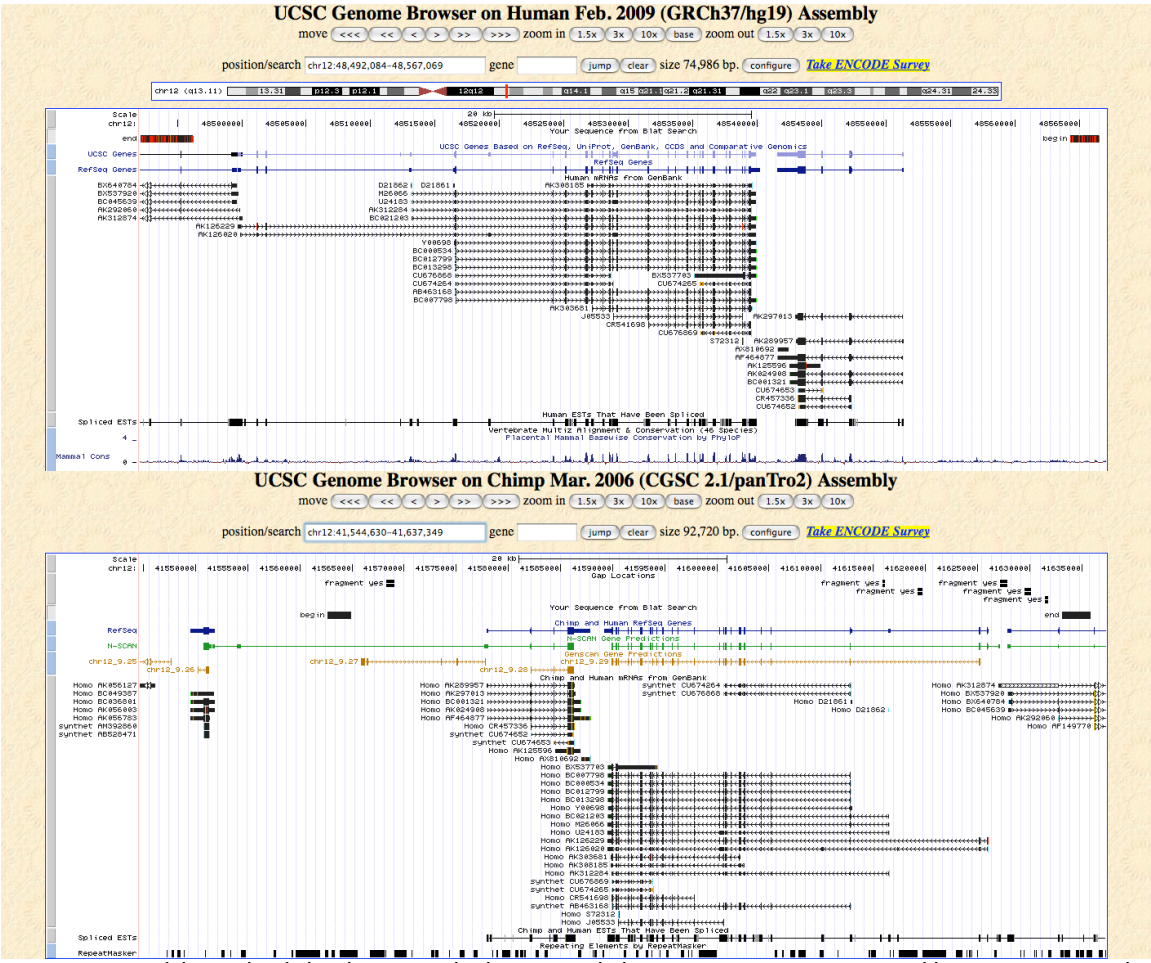
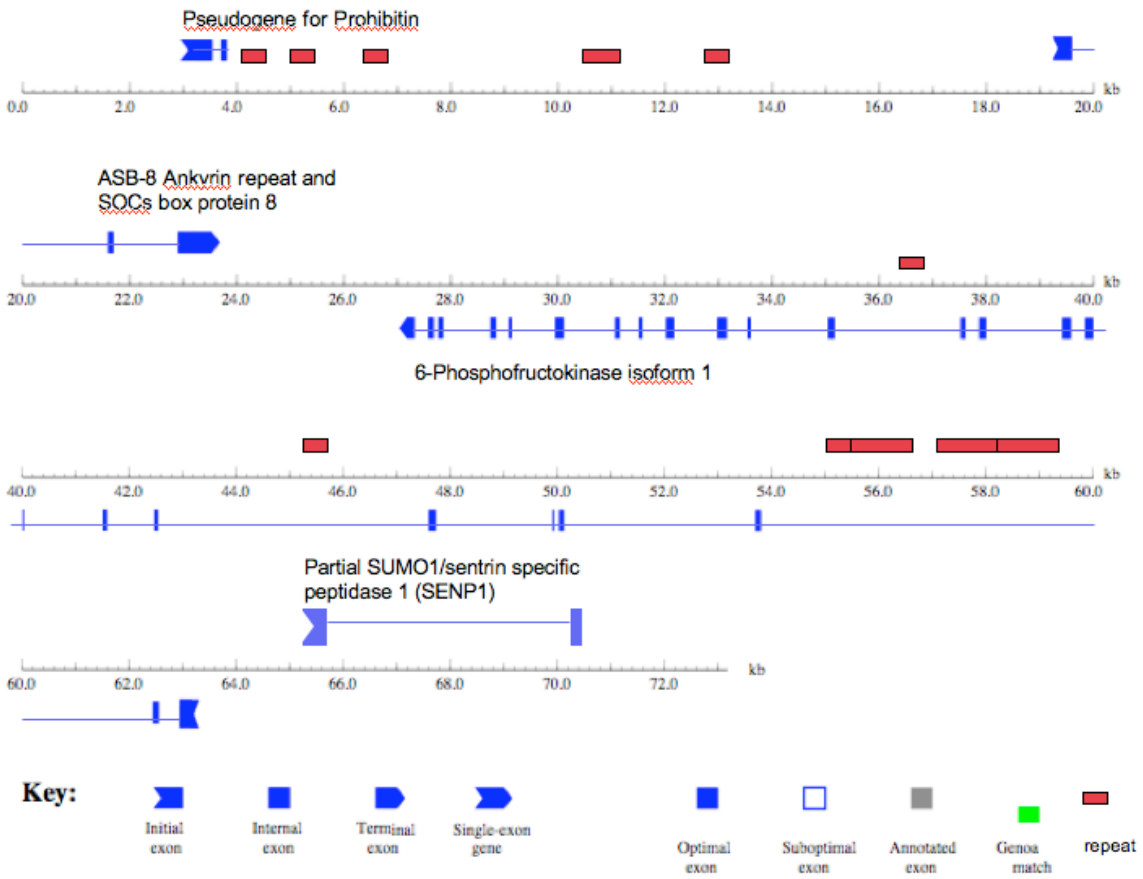


Fig. 15: BLAT of the ends of chunk 3-14 to the human and chimpanzee genomes (top and bottom, respectively) shows that the chunk is located on chimpanzee chromosome 12 and carries genes syntenic to those in humans. Note that the two images are mirror images for alignment; because chunk 3-14 is isolated, the designation of + and – strands was arbitrary and it is thus the relative positions of the genes along chimpanzee and human chromosome 12 that indicates synteny.

CONCLUSIONS

Based on the evidence illustrated above, we annotated chunk 3-14 with 11 repeats and four features: *ASB8*, *6-PFK* muscle isoform 1, *prohibitin* pseudogene, and a partial *SUMO1/SENP1* (Fig. 16). This project successfully introduced the basic principles of *de novo* annotation and the use of related tools such as *Genscan*, BLAST, BLAT, *RepeatMasker*, and the associated gene and protein databases.



Features For Chimp Chunk 3-14				
Feature Name	Accession No.	No. exons	bp span	Orientation
prohibitin pseudogene	NP_002625	2	3092-3816	+ strand
ASB8	CAG33617	3	19520-23652	+ strand
6-PFK muscle isoform 1	NP_001160158	24	63378-27308	- strand
partial *SUMO1/SENP1*	AF149770	2	65294-70204	+ strand

Fig. 16: The *Genscan* prediction readout updated to include the features verified by our inquiries. Note that repeats of fewer than 500 bp are not marked. The thin blue lines connecting exon boxes indicate that those exons are part of the same feature.