Chimp BAC Analysis

Adapted by Wilson Leung and Sarah C.R. Elgin from "Chimp BAC analysis: *TWINSCAN* and *UCSC Genome Browser*" by Dr. Michael R. Brent

Prerequisites

Detecting and Interpreting Genetic Homology

Resources

The *NCBI BLAST* web server is available at https://blast.ncbi.nlm.nih.gov/Blast.cgi
The *UCSC Genome Browser* is available at https://genome.ucsc.edu
The *GENSCAN* web server is available at http://hollywood.mit.edu/GENSCAN.html
The package containing the files for this walkthrough is available through the "Chimp BAC Analysis: Genes and Pseudogene" page on the GEP website.

Introduction

Recent technological advances have dramatically increased the rate at which we can generate high-quality DNA sequence. However, characterization of features (genes, repetitive elements, etc.) found within these raw sequences remains a slow and labor-intensive process. Multiple computational methods have been devised to improve the efficiency of sequence annotation. In this walkthrough, we will focus on the identification of genes and pseudogenes in a chimpanzee BAC clone. The *GENSCAN* and *BLAST* programs, along with the *UCSC Genome Browser*, will be used to facilitate the annotation process.

GENSCAN is a program that predicts the locations of genes in DNA sequences. While potential genes identified by *GENSCAN* must still be experimentally confirmed, its predictions can nonetheless help narrow the scope of the investigation by identifying regions where genes are most likely to be found. *GENSCAN* uses a probabilistic model to predict locations of promoters, exons, and polyadenylation signals.

Discussion of the actual implementation of *GENSCAN* is beyond the scope of this tutorial. For additional information on how *GENSCAN* works and on how to use *GENSCAN*, please see the paper by Burge *et al.*[1] and the *GENSCAN* website (http://hollywood.mit.edu/GENSCAN.html).

The *Chimp_BAC_Analysis* folder contains all the files required for this tutorial. The file *ChimpBAC_sequence.fasta* contains a 170kb sequence from chimpanzee. The *ChimpBAC_sequence_RM.fasta* file contains the repeat-masked version of the same sequence using the primate library with *RepeatMasker*. *GENSCAN_ChimpBAC.html* contains the *GENSCAN* predictions for our sequence.

[1] Burge, C. and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. J. Mol. Biol. 268(1):78-94.

Overview of *GENSCAN* predictions

The *GENSCAN* analysis of our BAC is available in the file *GENSCAN_ChimpBAC.html*. To view the *GENSCAN* results, launch a web browser and open the html file. In order to reduce the number of spurious gene predictions, we have masked repetitive sequences (including low complexity sequences) in the BAC sequence prior to running *GENSCAN*. You can run the *GENSCAN* analysis yourself by submitting the sequence (*ChimpBAC_sequence_RM.fasta*) to the *GENSCAN* web server at http://hollywood.mit.edu/GENSCAN.html. (If you run the analysis on the *GENSCAN* web server, please note that the links to the PDF and PostScript files are currently broken.)

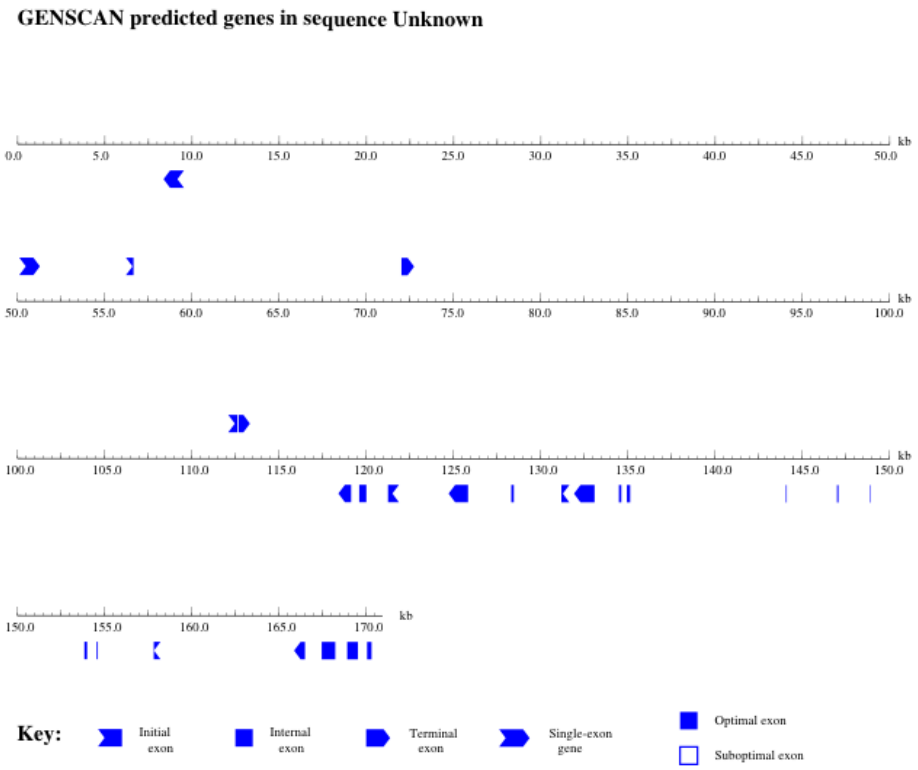


Figure 1 Schematic representation of the *GENSCAN* gene predictions for our unknown sequence. A higher resolution version of this image is available in the file GENSCAN_ChimpBAC_genes.pdf in the exercise package.

GENSCAN was run on the chimp BAC using the vertebrate parameter set. It predicted 8 genes in this sequence of about 170kb (Figure 1). However, most of these genes consist of only one or two exons, which is unusual (the median for chimp is about 8 exons).

The first part of the *GENSCAN* output (*GENSCAN_ChimpBAC.html*) is a table that summarizes the 8 predicted genes within this BAC. Explanations of the columns in the table are available at the end of the *GENSCAN* output. Illustrations of the locations of the predicted genes are also available in PDF format (*GENSCAN_ChimpBAC_genes.pdf*). The rest of the output consists of the peptide (amino acid) and coding (nucleotide) sequences of the predicted genes. We will use these sequences to evaluate the validity of the *GENSCAN* predictions.

Analysis of predicted gene 1 (single exon, 423 coding bases)

Examine conserved domain matches from the *blastp* search of the predicted peptide against the nr protein database

To get an idea of what this predicted gene could be, we will run a protein-protein *BLAST* (*blastp*) search against the NCBI's non-redundant (nr) database using default parameters. (The *blastp* results are available in the file *blastpGene1.txt* in the tutorial package.) The nr database consists of most of the known and hypothetical protein sequences that have been submitted to GenBank, with groups of essentially identical sequences reduced to a single sequence (hence the name "non-redundant"). While the *BLAST* hits produced with the nr database often contain relatively little information to help us interpret them, the names of these hits and their corresponding references nonetheless provide us with a reasonable starting point.

> *Copy the peptide sequence of this gene prediction (from the GENSCAN output) onto the clipboard. Open a new web browser window, navigate to the BLAST server at NCBI (*https://blast.ncbi.nlm.nih.gov/Blast.cgi*), and click on the "Protein BLAST" image. Paste the contents of the clipboard into the query search box and click "BLAST" (*Figure 2*). For teaching purposes, you can find the results of the blastp search in the file blastpGene1.txt.*

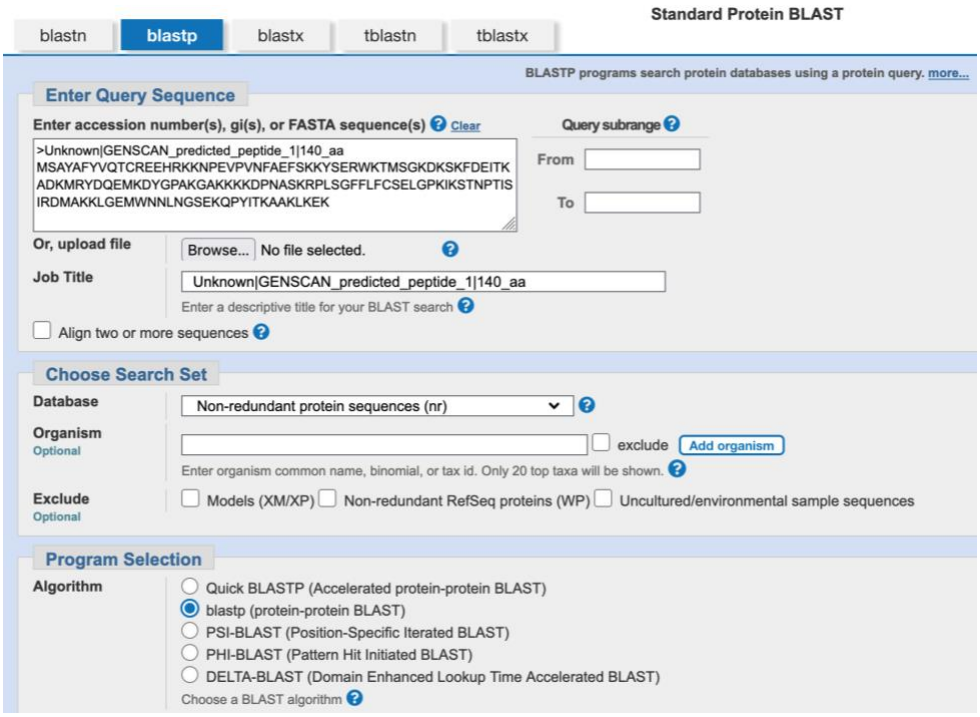


Figure 2 *blastp* search of the first *GENSCAN* prediction against the nr protein database.

Click on the "Graphic Summary" tab to obtain a graphical overview of the *BLAST* results. This tab shows that the predicted protein contained two copies of a conserved domain (the HMG-box superfamily, Figure 3). We can learn more about this conserved domain through the Conserved Domain database (CDD).

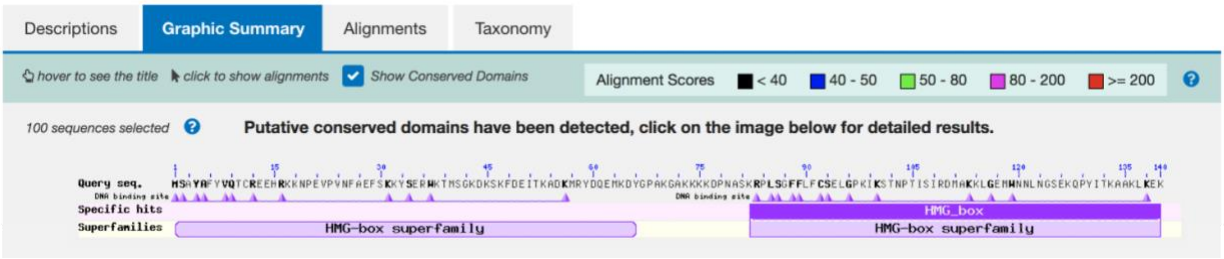


Figure 3 Conserved domain search shows the predicted protein contains two HMG-box domains.

Open a new web browser window and navigate back to the NCBI home page at https://www.ncbi.nlm.nih.gov. On the top search bar, change the database to "Conserved Domains" and search for "HMG-box" (Figure 4). Click on the fourth link with the accession number cd00084 (Figure 5). The "cd" prefix indicates that the staff at NCBI curated this record.

> Each record in GenBank has an accession number (*e.g.*, cd00084) that uniquely identifies the record in all the major sequence databases (*e.g.*, NCBI, DDBJ, EMBL).

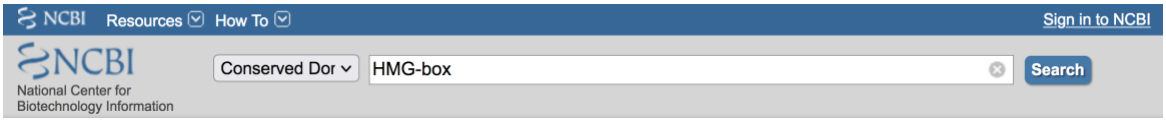


Figure 4 Search for HMG-box in the NCBI Conserved Domain database.

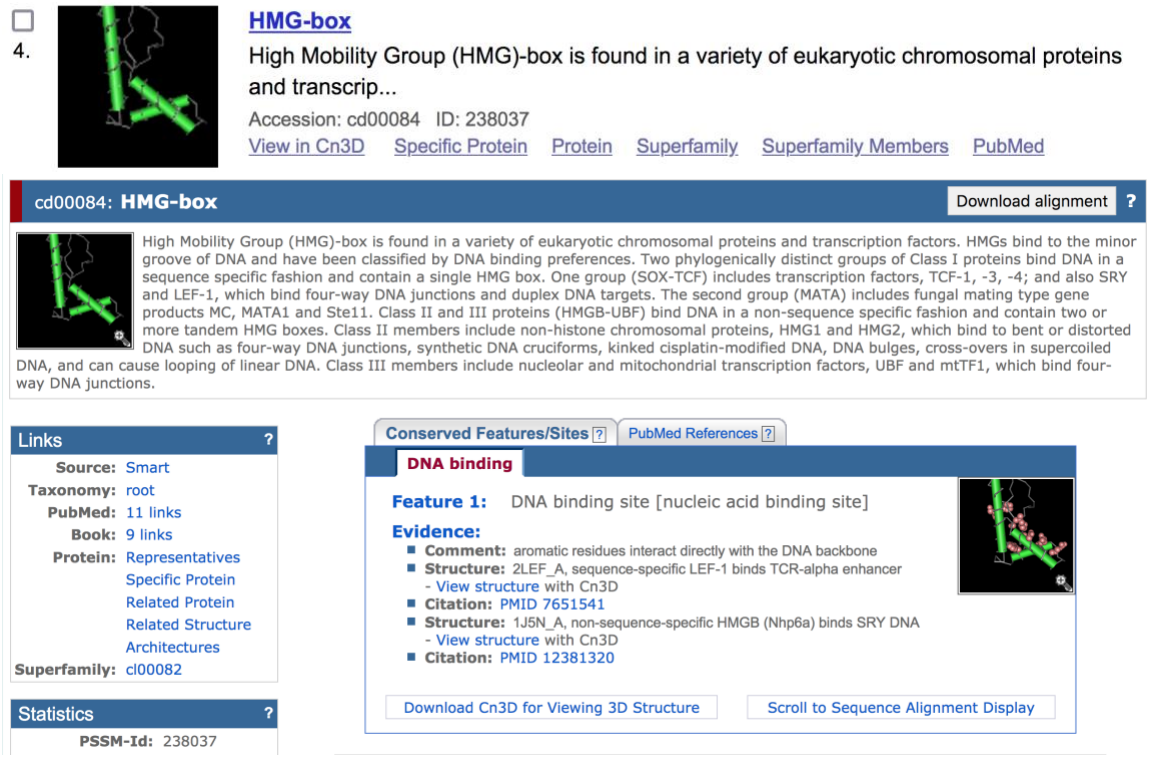


Figure 5 Detail record of the HMG-box domain in the NCBI Conserved Domain database.

The results page provides a summary and links to literature references for this conserved domain. You can also see the 3D structure of the HMG-box (*HMGbox.cn3* in the exercise package) if you have the structure viewer *Cn3D* installed (Figure 6). The bottom of the page shows the multiple sequence alignment used to construct the Position Specific Score Matrix (PSSM) for this conserved domain. The *CD-Search* program uses this PSSM to identify conserved domains in a query sequence. See the "Conserved Domains and Protein Classification Help" page (https://www.ncbi.nlm.nih.gov/Structure/cdd/cdd_help.shtml) for more information.

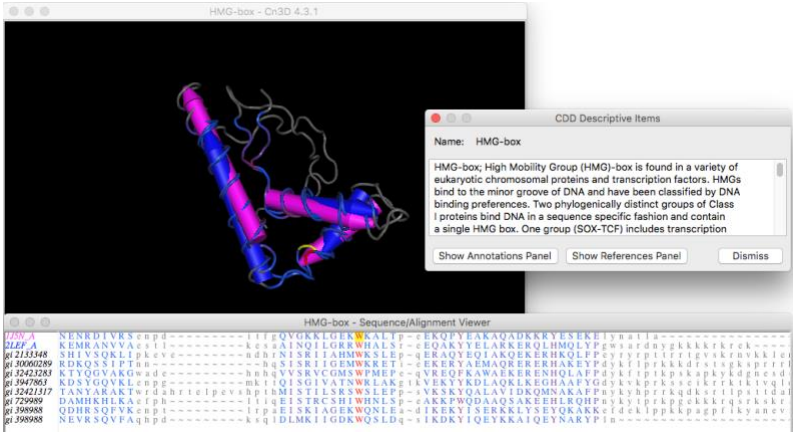


Figure 6 Cn3D display for the HMG-box domain.

We can learn the function of the HMG-box from the description of the CDD record:

> High Mobility Group (HMG)-box is found in a variety of eukaryotic chromosomal proteins and transcription factors. HMGs bind to the minor groove of DNA and have been classified by DNA binding preferences. Two phylogenically distinct groups of Class I proteins bind DNA in a sequence specific fashion and contain a single HMG box. One group (SOX-TCF) includes transcription factors, TCF-1, -3, -4; and also SRY and LEF-1, which bind four-way DNA junctions and duplex DNA targets. The second group (MATA) includes fungal mating type gene products MC, MATA1 and Ste11. Class II and III proteins (HMGB-UBF) bind DNA in a non-sequence specific fashion and contain two or more tandem HMG boxes. Class II members include non-histone chromosomal proteins, HMG1 and HMG2, which bind to bent or distorted DNA such as four-way DNA junctions, synthetic DNA cruciforms, kinked cisplatin-modified DNA, DNA bulges, cross-overs in supercoiled DNA, and can cause looping of linear DNA. Class III members include nucleolar and mitochondrial transcription factors, UBF and mtTF1, which bind four-way DNA junctions.

If you would like to learn more about the HMG-box domain, the description page also contains links to 11 original research articles. You can access the PubMed entries for these articles by clicking on the link next to the "PubMed" label in the "Links" section (column on the left). In addition to the abstracts, most of the PubMed entries contain links to the entire research article.

Now that we have a basic idea on the potential function of the predicted gene, we are ready to examine the alignments from our *blastp* search against the nr database. To summarize our conclusions from the motif portion of the *BLAST* search: the first *GENSCAN* predicted gene is probably a real gene with a DNA binding function, or a pseudogene derived from a real gene. Because this prediction is a single-exon gene (only about 7% of known human genes have a single exon), we should be vigilant looking for signs that this prediction is either a pseudogene or a mispredicted fragment of a multi-exon gene. It would not be terribly surprising if *GENSCAN* missed an exon or split two exons of the same transcript into multiple gene predictions.

Examine *blastp* alignments of predicted peptide against the nr protein database

The graphical overview tab shows that there are many significant hits to the nr database that cover the entire length of the predicted protein, supporting the gene model of the predicted gene. You can confirm this by clicking on the "Descriptions" tab and then click on the description of each hit and examining the alignments (Figure 7). However, many of the top hits have accession numbers that begin with the "XP_" prefix, which means that these predictions are unconfirmed and have no experimental evidence to support them. Generally, we do not want to rely on (possibly incorrect) computational predictions alone for any part of our annotation; instead, we will seek matches that are based on more direct evidence.

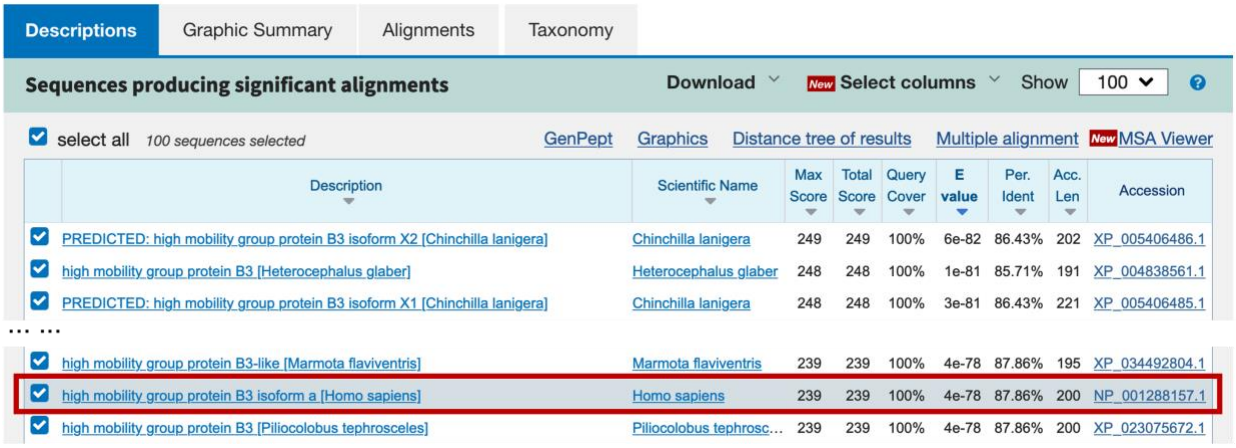


Figure 7 *blastp* hits to our predicted protein in the nr database.

The best experimentally confirmed RefSeq match (with the accession number NP_001288157.1) corresponds to the human HMGB3 protein (Figure 7). Click on the link in the Description column to jump to the alignment (Figure 8).

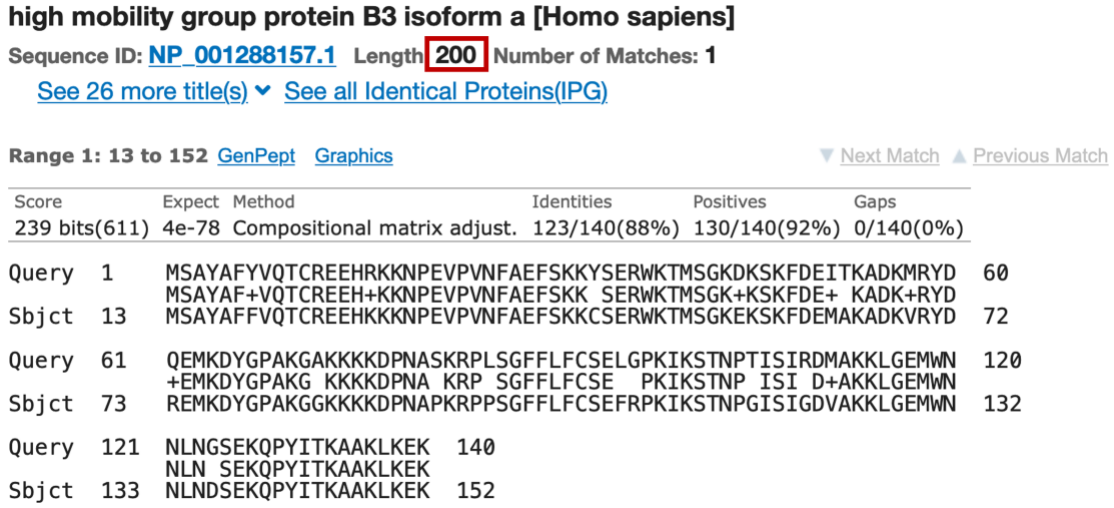


Figure 8 The *blastp* alignment of the human HMGB3 protein (subject) against our predicted protein (query). The "See 26 more title(s)" link beneath the "Sequence ID" field indicates that the subject sequence in the nr database actually corresponds to multiple sequence records at GenBank.

Looking at the "Length" field beneath the sequence name, we see that the full HMGB3 protein has a length of 200 amino acids. However, recall that our predicted peptide only has 140 amino acids. There are three potential explanations for this discrepancy. First, *GENSCAN* might have missed part of the real chimp protein. This scenario could happen if *GENSCAN* failed to identify one or more exons, or called them as part of another gene. Alternatively, the predicted gene might be a pseudogene that has acquired an in-frame stop codon. The stop codon would cause the *GENSCAN* prediction to end prematurely. A third hypothesis is that *GENSCAN* has accurately predicted a functional protein that is closely related to the human HMGB3 protein but lacks some part of it, such as one or more functional domains.

We can retrieve the GenBank entry for this *BLAST* hit by clicking on the accession number NP_001288157.1. In this page, you can display and export the GenBank record in a variety of formats, including FASTA (Figure 9). The most useful thing here is the PubMed links to the primary research papers that describes the human HMGB3 protein.

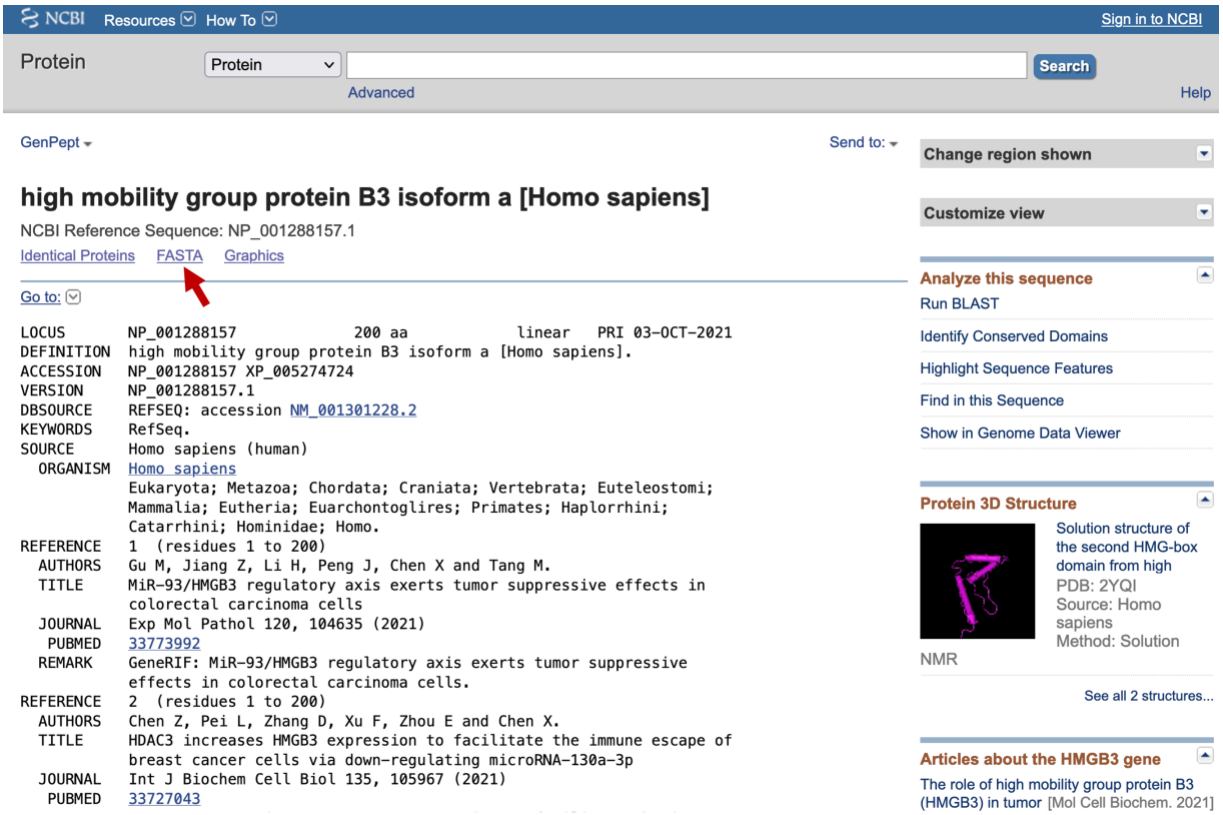
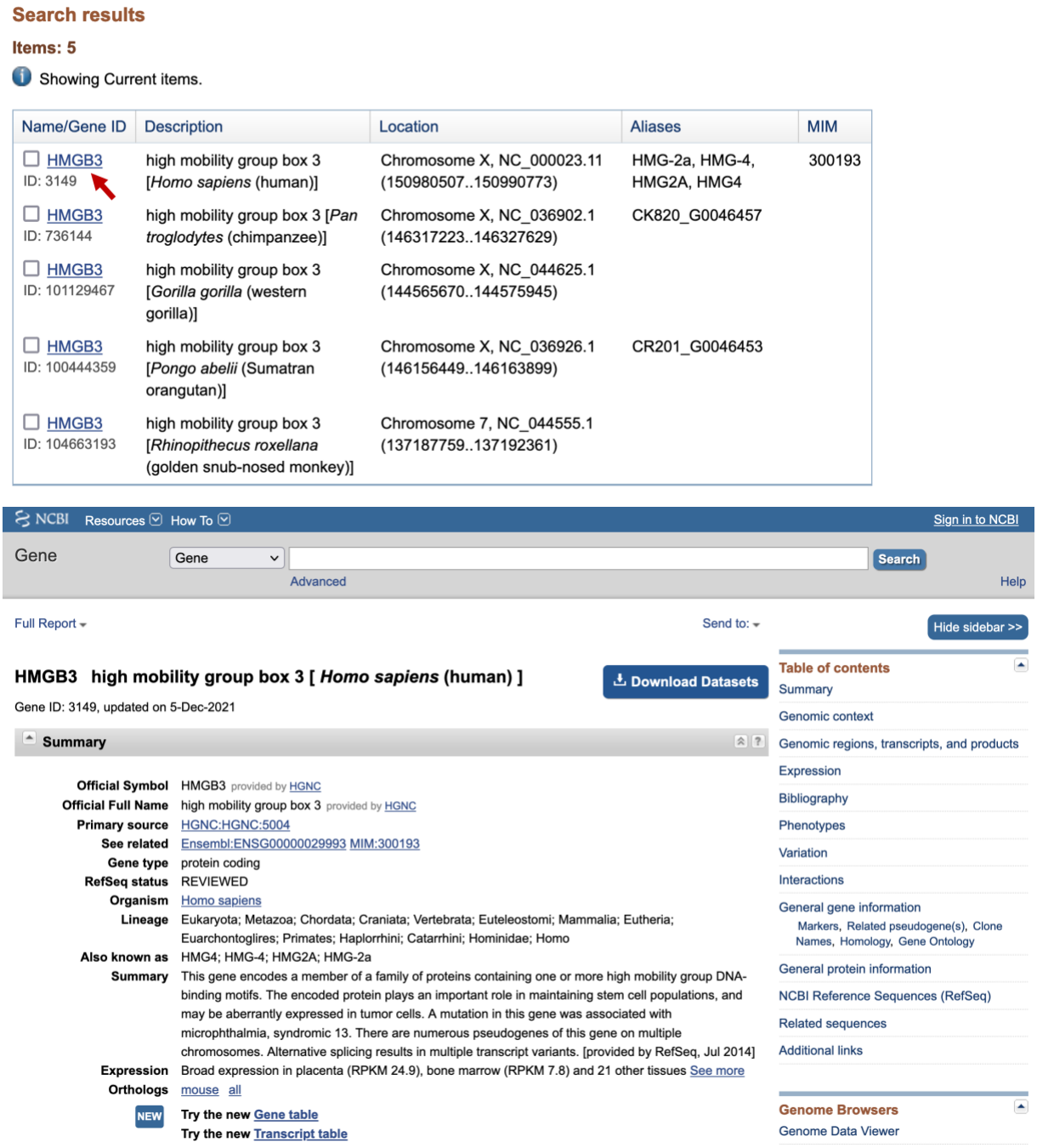


Figure 9 The GenBank RefSeq record for the human HMGB3 protein.

Go back to the *BLAST* output. Click on the "Gene" link under the "Related Information" section, and then click on the "HMGB3" link next to the entry with the description "high mobility group box 3 [*Homo sapiens* (human)]" to access this Entrez gene record (Figure 10).



Search results

Items: 5

Showing Current items.

Name/Gene ID	Description	Location	Aliases	MIM
HMGB3 ID: 3149	high mobility group box 3 [*Homo sapiens* (human)]	Chromosome X, NC_000023.11 (150980507..150990773)	HMG-2a, HMG-4, HMG2A, HMG4	300193
HMGB3 ID: 736144	high mobility group box 3 [*Pan troglodytes* (chimpanzee)]	Chromosome X, NC_036902.1 (146317223..146327629)	CK820_G0046457	
HMGB3 ID: 101129467	high mobility group box 3 [*Gorilla gorilla* (western gorilla)]	Chromosome X, NC_044625.1 (144565670..144575945)		
HMGB3 ID: 100444359	high mobility group box 3 [*Pongo abelii* (Sumatran orangutan)]	Chromosome X, NC_036926.1 (146156449..146163899)	CR201_G0046453	
HMGB3 ID: 104663193	high mobility group box 3 [*Rhinopithecus roxellana* (golden snub-nosed monkey)]	Chromosome 7, NC_044555.1 (137187759..137192361)		

NCBI Resources How To Sign in to NCBI

Gene Gene Search Advanced Help

Full Report Send to: Hide sidebar >>

HMGB3 high mobility group box 3 [*Homo sapiens* (human)] Download Datasets

Gene ID: 3149, updated on 5-Dec-2021

Summary

Official Symbol HMGB3 provided by HGNC
Official Full Name high mobility group box 3 provided by HGNC
Primary source HGNC:HGNC:5004
See related Ensembl:ENSG00000029993 MIM:300193
Gene type protein coding
RefSeq status REVIEWED
Organism Homo sapiens
Lineage Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo
Also known as HMG4; HMG-4; HMG2A; HMG-2a
Summary This gene encodes a member of a family of proteins containing one or more high mobility group DNA-binding motifs. The encoded protein plays an important role in maintaining stem cell populations, and may be aberrantly expressed in tumor cells. A mutation in this gene was associated with microphthalmia, syndromic 13. There are numerous pseudogenes of this gene on multiple chromosomes. Alternative splicing results in multiple transcript variants. [provided by RefSeq, Jul 2014]
Expression Broad expression in placenta (RPKM 24.9), bone marrow (RPKM 7.8) and 21 other tissues See more
Orthologs mouse all

NEW Try the new Gene table
Try the new Transcript table

Table of contents
Summary
Genomic context
Genomic regions, transcripts, and products
Expression
Bibliography
Phenotypes
Variation
Interactions
General gene information
Markers, Related pseudogene(s), Clone Names, Homology, Gene Ontology
General protein information
NCBI Reference Sequences (RefSeq)
Related sequences
Additional links

Genome Browsers
Genome Data Viewer

Figure 10 The Entrez Gene record for the human *HMGB3* gene.

The "Also known as" field of this Entrez Gene record shows that the other names that have been used to describe the human HMGB3 protein include "HMG4" and "HMG2A". In the "Genomic context" section, we learn that the protein is located on the X chromosome at locus Xq28 (Figure 11). We can obtain a more detailed graphical overview of the region surrounding the HMGB3 protein by clicking on the "Genome Data Viewer" link at the top right corner of this section.

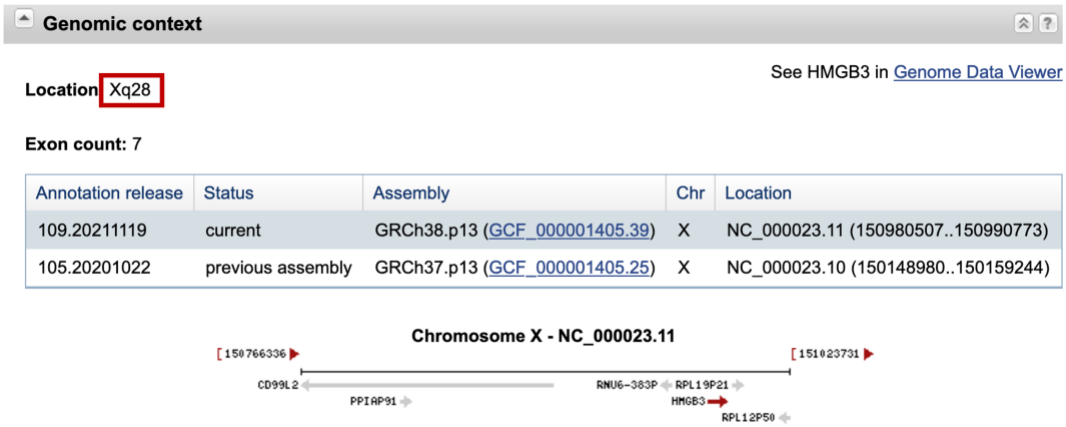


Figure 11 The "Genomic context" section shows the human HMGB3 protein is found on the X chromosome.

We can find information on all known and predicted transcripts for this gene in the "Genomic regions, transcripts, and products" section. The "NCBI *Homo sapiens* Annotation Release 109.20211119" evidence track shows that the *HMGB3* gene has five known transcripts (NM_001301231.2, NM_001301228.2, NM_005342.4, NM_001301229.2, and XM_024452369.1). We also discover that *HMGB3* is a multi-exon gene in human (Figure 12).

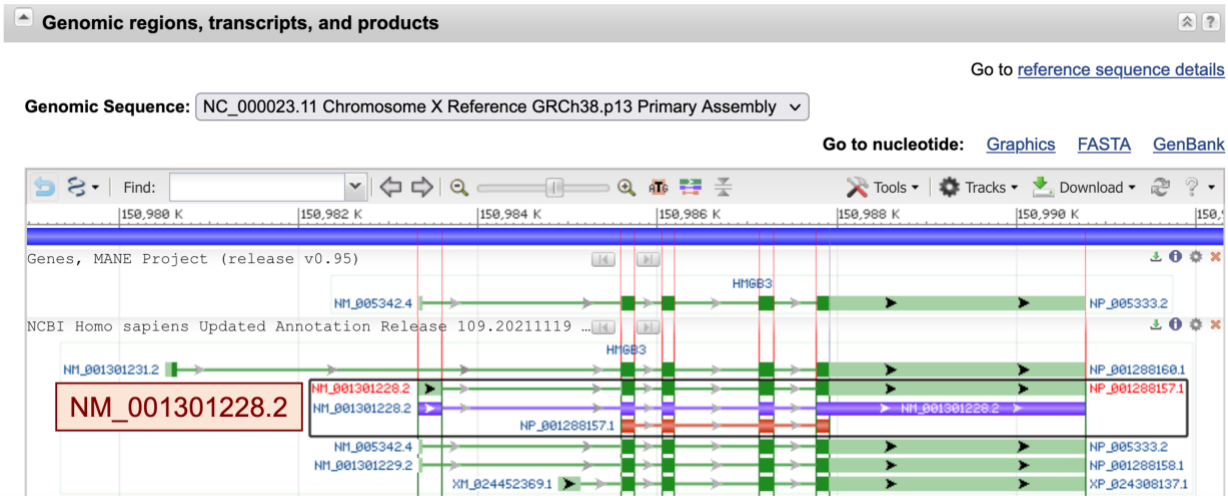


Figure 12 The "NCBI *Homo sapiens* Annotation Release 109.20211119" track in the "Genomic regions, transcripts, and products" section shows that the human *HMGB3* is a multi-exon gene on the X chromosome. The dark green boxes in this track correspond to the coding regions and the light green boxes correspond to the untranslated regions. When you click on a feature (*e.g.*, NM_001301228.2), the feature will expand into two separate records: the purple feature corresponds to the mRNA record and the red feature corresponds to the protein record.

To summarize our findings so far, we have learned that our predicted protein shows strong sequence similarity to the human HMGB3 protein but it is truncated relative to that protein. We need to decide among three possible explanations for the truncation: (1) *GENSCAN* has failed to accurately predict the gene, (2) the predicted gene is a pseudogene, or (3) the predicted gene may produce a functional product that lacks part of the human protein. To determine which of these hypotheses is correct, we need to gather additional evidence using the *UCSC Genome Browser*.

UCSC Genome Browser analysis of the *GENSCAN* prediction

*Go back to the GENSCAN output page. Copy the first predicted coding (CDS) sequence (*Figure 13*) and go to the UCSC Genome Browser (https://genome.ucsc.edu). Click on the "BLAT" link under "Our tools" and then select the "**Human**" genome and the "**Mar. 2006 (NCBI36/hg18)**" assembly. Paste the predicted sequence into the textbox and click "Submit" (*Figure 14*).*

Figure 13 Retrieve the coding DNA sequence (CDS) for the first *GENSCAN* gene prediction.

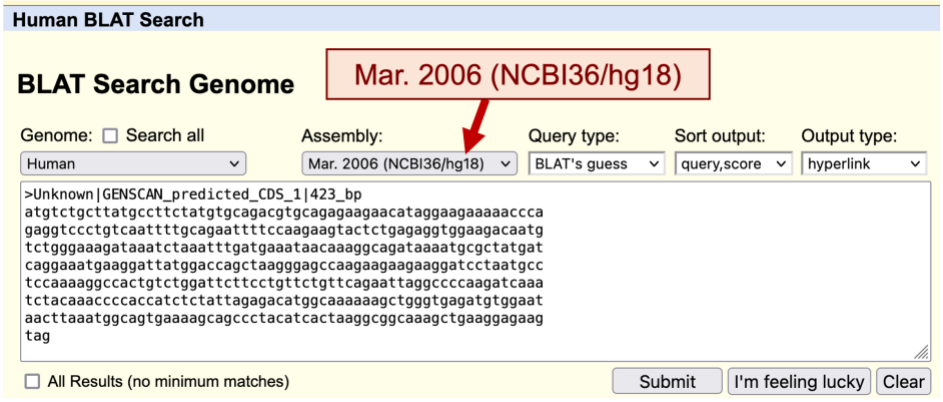


Figure 14 *BLAT* search of the CDS of our predicted gene against the human genome assembly NCBI36/hg18.

The predicted sequence matches many loci in the human genome (Figure 15), which is not surprising because it contains a conserved domain. The top hit shows an identity of 99.1% to the human genomic sequence over the length of the entire query (423bp). This is within the range of percent identity we would expect for a pair of human and chimp orthologs. The next best match has an identity of only 93.6%, which is substantially below what we expect for human and chimp orthologs. (Coding regions of chimp and human are approximately 98% identical on average.)

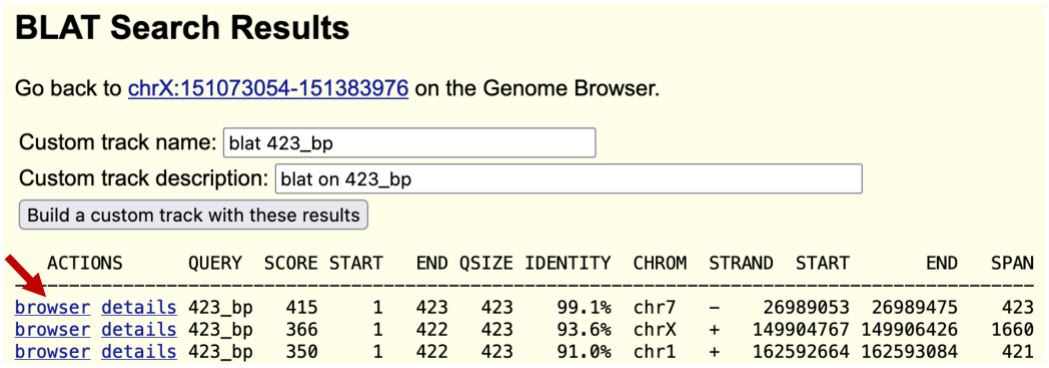


Figure 15 *BLAT* result that shows the regions of the human genome with similarity to the first predicted gene.

Click on the "browser" link next to the alignment with the best score (on chromosome 7), and the next screen will show your sequence as a black bar under the title "Your Sequence from *Blat* Search". You can change the display options by clicking on the "configure" button. (Please see the *UCSC Genome Browser* User Guide for additional information.)

Reset the display settings for the genome browser by clicking on the "hide all" button. Expand the "Mapping and Sequencing Tracks" section by clicking on the "+" sign and then configure the tracks to use the following options:

Adjust the following track to "pack" mode:

- *Under "Mapping and Sequencing": Blat Sequence*

Adjust the following tracks to "dense" mode:

- *Under "Gene and Gene Predictions": UCSC Genes, RefSeq Genes, Ensembl Genes, Genscan Genes, N-SCAN, SGP Genes* (Figure 16).
- *Under "mRNA and EST": Human mRNAs, Spliced ESTs, Human ESTs, Other ESTs.*
- *Under "Comparative Genomics": Mouse Chain/Net.*
- *Under "Variations and Repeats": RepeatMasker.*

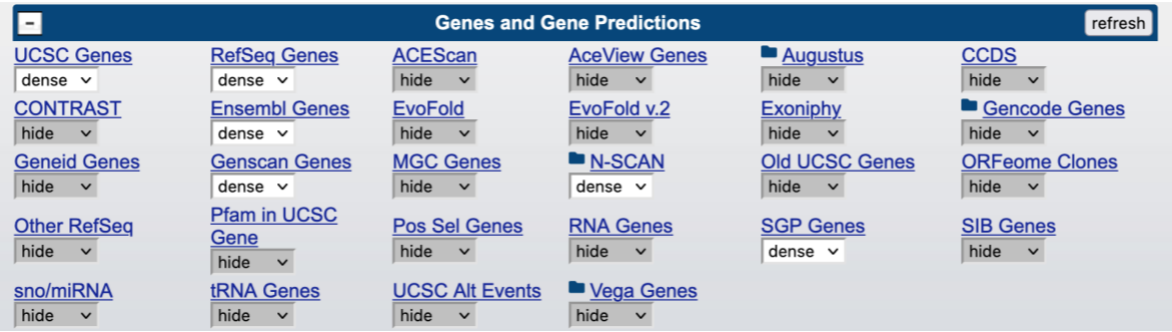


Figure 16 Configure the "Genes and Gene Predictions" section of the *UCSC Genome Browser*.

Hit "refresh" and then zoom out 3X to get a broader view of this region (Figure 17).

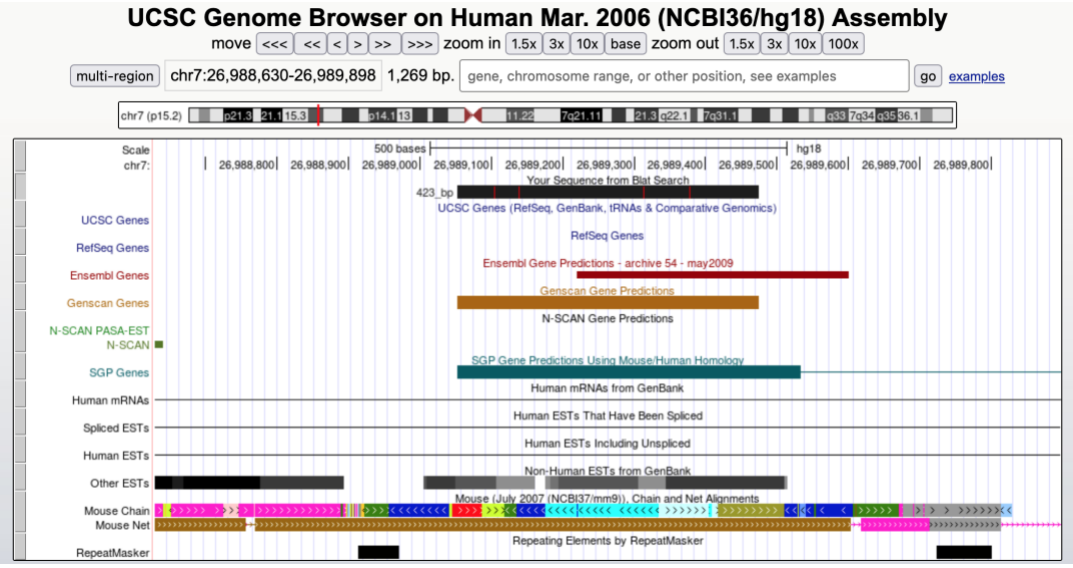


Figure 17 *UCSC Genome Browser* view of our region of interests on chromosome 7.

According to RefSeq and other curated gene lists, there are no known genes in this region. The only evidence we have for our prediction comes from hypothetical genes predicted by programs like *SGP* and *GENSCAN*. Interestingly, *N-SCAN* (another gene prediction program) did not predict a gene in this region. *SGP*, which tends to fuse multiple genes into a single feature, predicts that our putative gene is in fact part of a larger gene with two exons. (This hypothesis is plausible because the *GENSCAN* prediction is shorter than the matches to known proteins.)

There are no human ESTs or spliced ESTs that aligned to this region. The lack of spliced human ESTs is consistent with our prediction of a single-exon gene. There are also ESTs from other organisms, which are generally a more reliable guide to the presence of a real gene than unspliced human ESTs.

> *To see the EST evidence in more detail, change the "Other ESTs" track to "pack" mode and then click on the "refresh" button* (Figure 18).

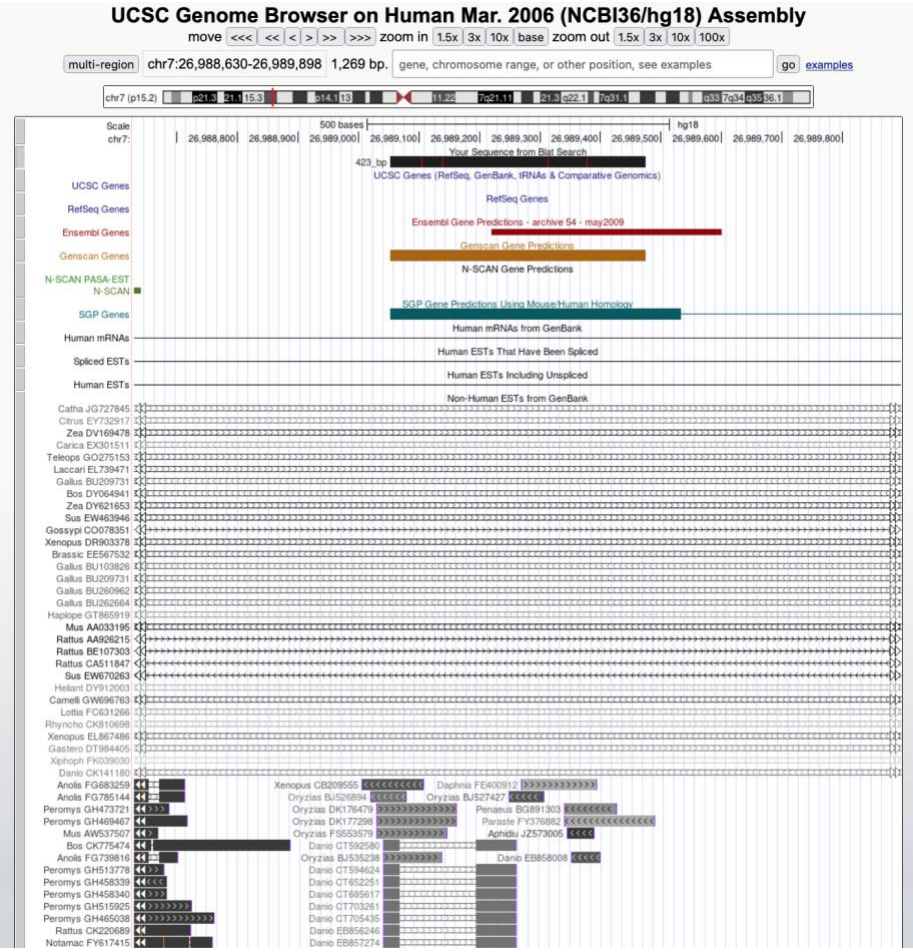


Figure 18 The "Other ESTs" track shows that ESTs from many different species mapped to the same region as the *GENSCAN* predicted feature.

The expanded "Other ESTs" track shows that there are many ESTs supporting this predicted gene from species such as Xenopus, mouse, and chicken. The shades of grey for each feature in the "Other ESTs" track correspond to the percent identity of the EST match: the higher the percent identity, the darker the feature. The presence of these ESTs from the other species

provides evidence that our prediction might be a real gene. Note that although *SGP* predicts a two-exon gene (the extra exon is off-screen to the right of Figure 18), the EST evidence from other species does not support this gene structure because none of the ESTs matching our prediction also matches the other predicted exon.

Analysis of the missing parts of the HMGB3 protein

To determine why the *GENSCAN* prediction matches only part of HMGB3, let's look at the alignment from the BLAT search.

> *Go to the GenBank record for the human HMGB3 protein. Click on the "FASTA" link under the title of the GenBank record* (Figure 19). *Copy the sequence to the clipboard. Go back to the BLAT search page (under "Tools") at the UCSC Genome Browser. Paste the sequence into the textbox, and click "Submit"* (Figure 20).

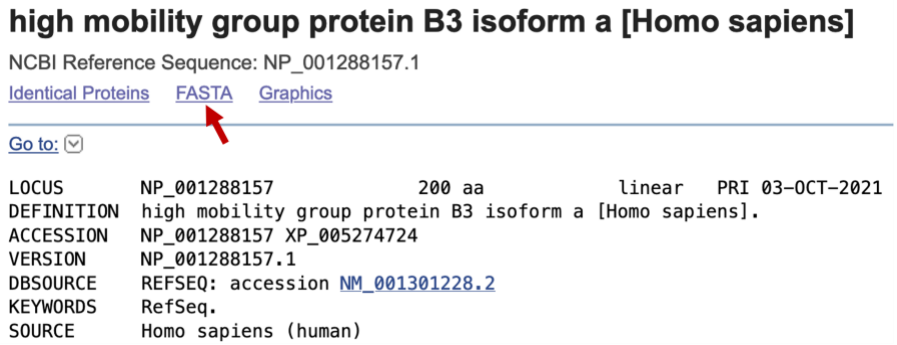


Figure 19 Click on the "FASTA" link in the GenBank record to retrieve the sequence for the human HMGB3 protein in FASTA format.

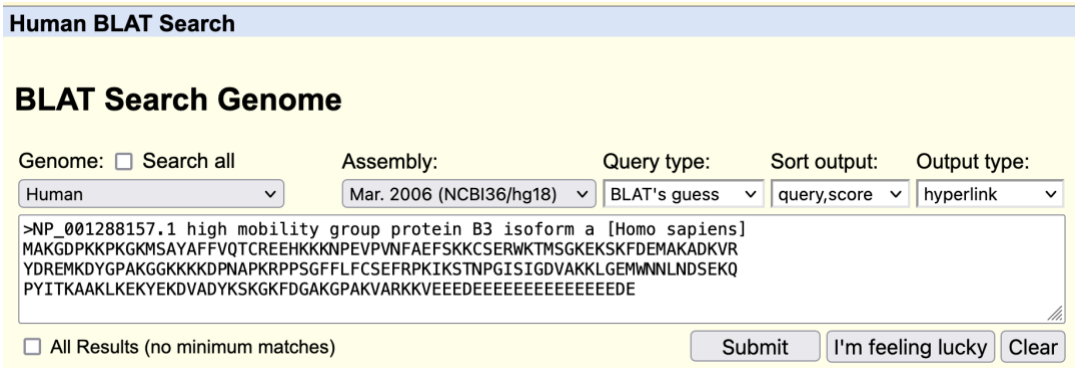
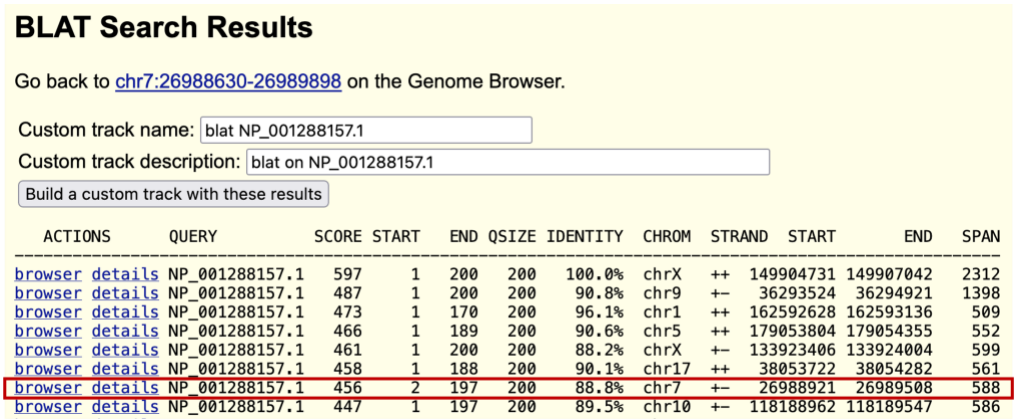


Figure 20 *BLAT* search of the human HMGB3 protein against the human genome assembly hg18.

First, you should notice that the match to the part of human chromosome 7 we have previously identified is only the 7th best match, with a sequence identity of only 88.8% (Figure 21). We have previously suspected that our predicted protein is a paralog (arise through a duplication event) of *HMGB3*, so it is not surprising that the best match from human (the likely ortholog) is not in this region. An identity of 88.8% is actually quite good for a paralog.



BLAT Search Results

Go back to chr7:26988630-26989898 on the Genome Browser.

Custom track name: blat NP_001288157.1

Custom track description: blat on NP_001288157.1

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	NP_001288157.1	597	1	200	200	100.0%	chrX	++	149904731	149907042	2312
browser details	NP_001288157.1	487	1	200	200	90.8%	chr9	+-	36293524	36294921	1398
browser details	NP_001288157.1	473	1	170	200	96.1%	chr1	++	162592628	162593136	509
browser details	NP_001288157.1	466	1	189	200	90.6%	chr5	++	179053804	179054355	552
browser details	NP_001288157.1	461	1	200	200	88.2%	chrX	+-	133923406	133924004	599
browser details	NP_001288157.1	458	1	188	200	90.1%	chr17	++	38053722	38054282	561
browser details	NP_001288157.1	456	2	197	200	88.8%	chr7	+-	26988921	26989508	588
browser details	NP_001288157.1	447	1	197	200	89.5%	chr10	+-	118188962	118189547	586

Figure 21 *BLAT* search of the human HMGB3 protein against the human genome shows that the match to chromosome 7 is the 7th best match with a sequence identity of 88.8%.

Click on the "browser" link to see the corresponding region on human chromosome 7 and then zoom out 3x. The match appears as a single block that completely overlaps with the *GENSCAN* prediction but also extends beyond it in both directions (Figure 22). This evidence suggests that there might be a longer single-exon gene at this locus than the *GENSCAN* prediction. To understand why *GENSCAN* would predict a shorter gene, we need to examine the alignment.

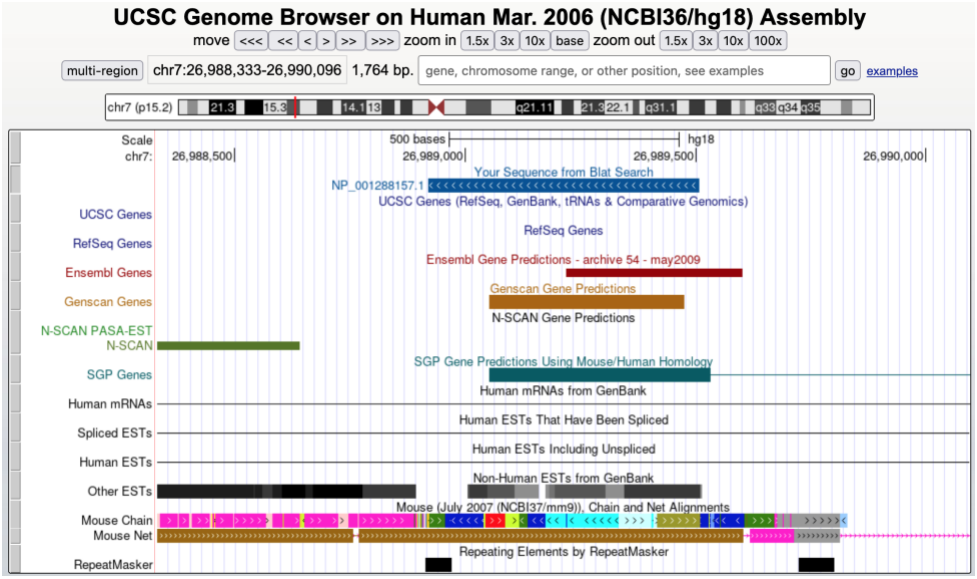


Figure 22 The *BLAT* alignment of the human HMGB3 protein (blue box) overlaps but extends beyond the *GENSCAN* gene prediction in both directions.

> *Go back to the BLAT results page, or resubmit the BLAT search using the human HMGB3 protein sequence from the GenBank record. Click on the "details" link to examine the alignment of HMGB3 to chromosome 7.*

For each alignment block, matches are in blue capital letters while mismatched or unmatched residues are in black lowercase letters. In general, the alignment looks good except for a few amino acid changes. However, when we examine some of the unmatched (black) regions, we noticed that one of these mismatches has the sequence "tag" — which would encode for a stop codon if it were in the correct reading frame (Figure 23).

Alignment of NP_001288157.1 and chr7:26988921-26989508

Click on links in the frame to the left to navigate through the alignment. Matching bases are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either sequence.

NP_001288157.1

```
mAkGDPKKPK GKMSAYAFfV QTCREEHkKK NPEVPVNFAE FSKKCSERWK TMSGKeKSKF  60
DEmaKADKvR YDrEMKDYGP AKGgKKKKDP NApKRPpSGF FLFCSEfrPK IKSTNPgISI  120
gDvAKKLGEM WNNLNDSEKQ PYITKAAKLK EKyEKDVaDY KSKGKFDGAk GPAkVArKKV  180
EEEDEEeEEE EEEEEEEede
```

Human.chr7 (reverse strand):

```
GCTaacGGTG ACCCCAAGAA ACCAAAGGGC AAGATGTCTG CTTATGCCTT CtatGTGCAG  26989449
ACGTGCAGAG AAGAACATag gAAGAAAAAC CCAGAGGTCC CTGTCAATTT TGCAGAATTT  26989389
TCCAAGAAGT GCTCTGAGAG GTGGAAGACA ATGTCTGGGA AAgatAAATC TAAATTTGAT  26989329
GAAataacaA AGGCGGATAA AatgCGCTAT GATcagGAAA TGAAGGATTA TGGACCAGCT  26989269
AAGGGAgccA AGAAGAAGAA GGATCCTAAT GCCtccAAAA GGCCActgTC TGGATTCTTC  26989209
CTGTTCTGTT CAGAAttagg cCCCAAGATC AAATCTACAA ACCCCaccAT CTCTATTaga  26989149
GACatgGCGA AAAAGCTGGG TGAGATGTGG AATAACTTAA ATGACAGTGA AAAGCAGCCC  26989089
TACATCACTA AGGCGGCAAA GCTGAAGGAG AAGtagGAGA AGGATGTTcc tGACTATAAG  26989029
TCGAAAGGAA AGTTTGATGG CGCAgagGGT CCTGCTaatG TTGCCtggAA AAAGGTGGAA  26988969
GAGGAAGATG AAGAAggcGA AGAAGAAGAA GAGGAGGAGG AGGAGGAG
```

Figure 23 A potential stop codon (tag) in the *BLAT* alignment to the human chromosome 7.

We can verify that this "tag" sequence is an in-frame stop codon by examining the "Side-by-Side Alignment" between the human HMGB3 protein and chromosome 7 (Figure 24). The color scheme for the Side-by-Side Alignment is as follows: a line for match, green for similar amino acids, and red for dissimilar amino acids. We see a red "X" (a stop codon) in the translated region of chromosome 7 that is aligned with a "Y" (tyrosine) in the human HMGB3 protein.

Side by Side Alignment*

```
00000004  A  K  G  D  P  K  K  P  K  G  K  M  S  A  Y  A  F  F  V  Q  00000063
<<<<<<<<  |  N  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  Y  |  |  <<<<<<<<
26989508 gctaacggtgaccccaagaaaccaaagggcaagatgtctgcttatgccttctatgtgcag 26989449

00000064  T  C  R  E  E  H  K  K  K  N  P  E  V  P  V  N  F  A  E  F  00000123
<<<<<<<<  |  |  |  |  |  |  R  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<<
26989448 acgtgcagagaagaacataggaagaaaaacccagaggtccctgtcaattttgcagaattt 26989389

00000124  S  K  K  C  S  E  R  W  K  T  M  S  G  K  E  K  S  K  F  D  00000183
<<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  D  |  |  |  |  |  <<<<<<<<
26989388 tccaagaagtgctctgagaggtggaagacaatgtctgggaaagataaatctaaatttgat 26989329

00000184  E  M  A  K  A  D  K  V  R  Y  D  R  E  M  K  D  Y  G  P  A  00000243
<<<<<<<<  |  I  T  |  |  |  |  M  |  |  |  Q  |  |  |  |  |  |  |  |  <<<<<<<<
26989328 gaaataacaaaggcggataaaatgcgctatgatcaggaaatgaaggattatggaccagct 26989269

00000244  K  G  G  K  K  K  K  D  P  N  A  P  K  R  P  P  S  G  F  F  00000303
<<<<<<<<  |  |  A  |  |  |  |  |  |  |  |  S  |  |  |  L  |  |  |  |  <<<<<<<<
26989268 aagggagccaagaagaagaaggatcctaatgcctccaaaaggccactgtctggattcttc 26989209

00000304  L  F  C  S  E  F  R  P  K  I  K  S  T  N  P  G  I  S  I  G  00000363
<<<<<<<<  |  |  |  |  |  L  G  |  |  |  |  |  |  |  |  T  |  |  |  R  <<<<<<<<
26989208 ctgttctgttcagaattaggccccaagatcaaatctacaaaccccaccatctctattaga 26989149

00000364  D  V  A  K  K  L  G  E  M  W  N  N  L  N  D  S  E  K  Q  P  00000423
<<<<<<<<  |  M  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<<
26989148 gacatggcgaaaaagctgggtgagatgtggaataacttaaatgacagtgaaaagcagccc 26989089

00000424  Y  I  T  K  A  A  K  L  K  E  K  Y  E  K  D  V  A  D  Y  K  00000483
<<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  X  |  |  |  |  P  |  |  |  <<<<<<<<
26989088 tacatcactaaggcggcaaagctgaaggagaagtaggagaaggatgttcctgactataag 26989029

00000484  S  K  G  K  F  D  G  A  K  G  P  A  K  V  A  R  K  K  V  E  00000543
<<<<<<<<  |  |  |  |  |  |  |  |  E  |  |  |  N  |  |  W  |  |  |  |  <<<<<<<<
26989028 tcgaaaggaaagtttgatggcgcagagggtcctgctaatgttgcctggaaaaaggtggaa 26988969

00000544  E  E  D  E  E  E  E  E  E  E  E  E  E  E  E  E  00000591
<<<<<<<<  |  |  |  |  |  G  |  |  |  |  |  |  |  |  |  |  <<<<<<<<
26988968 gaggaagatgaagaaggcgaagaagaagaggaggaggaggaggag 26988921
```

Figure 24 Side-by-Side Alignment of the human HMGB3 protein against a region of the human chromosome 7 contains an in-frame stop codon (X).

The fact that the protein alignment runs right over this stop codon, with no deterioration in similarity, strongly suggests that our prediction is a recently retrotransposed pseudogene, which has been inactivated by the mutation but has not yet diverged significantly from its source gene.

To confirm this hypothesis, go back to the *BLAT* results and click on the "browser" link to find the place where *HMGB3* best matches the genome (i.e. 100% identity at the X chromosome) and then zoom out 10x. We observe that *HMGB3* actually has four coding exons in the human genome (Figure 25)! Case closed.

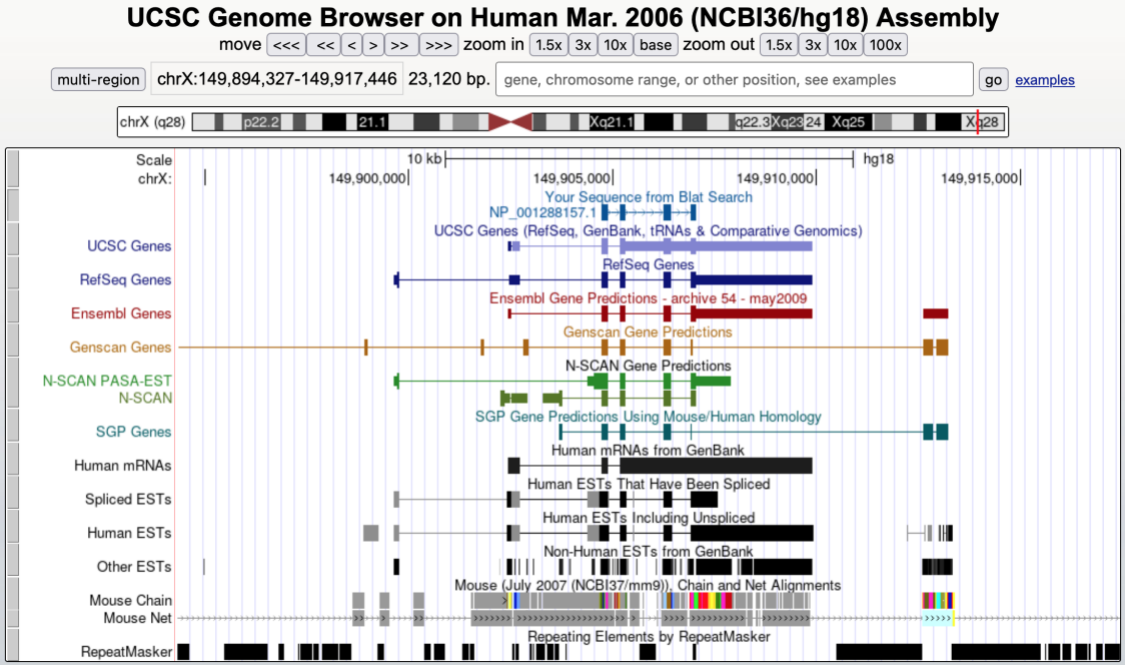


Figure 25 The best match to *HMGB3* on the X chromosome shows that it is actually a gene with four coding exons.

Based on all the evidence we have collected, we can conclude that, as a processed cDNA, the four-exon *HMGB3* gene from the X chromosome was retrotransposed into chromosome 7 and acquired a stop codon (i.e. a nonsense mutation) prior to the split of the chimpanzee and human lineages. However, this transposition event is relatively recent, because the pseudogene still retains 88.8% identity to its source protein.

But how do we explain all the supporting evidence (i.e. ESTs) for our prediction, given that it turned out to be a pseudogene? They are almost certainly due to the fact that our pseudogene has a high degree of similarity to the functional HMGB3 protein. To confirm this hypothesis, we can perform additional pairwise *BLAST* alignments of the ESTs against the pseudogene and the functional protein. If our hypothesis is correct, the ESTs should be more similar to the functional protein than to the pseudogene.

In order to estimate the age of this pseudogene, we can do a *BLAT* search using the pseudogene or the HMGB3 protein against the mouse or rat genome browsers. If we find the corresponding pseudogene in the syntenic regions of the rodent genomes, then the transposition event likely occurred prior to the divergence of rodents and primates.

Estimating the age of the pseudogene

To estimate the age of the pseudogene, go back to the *BLAT* results showing the human *HMGB3* sequence compared to homologous region of human chromosome 7. We will examine the "Mouse Chain/Net" track more closely. The "Mouse Chain/Net" track actually consists of two different tracks. The "Chain" track attempts to map every part of the human genome to the mouse genome. It was built by "chaining" individual alignments together into long stretches that indicate regions of orthology between mouse and human. The "Net" track attempts to summarize the chains into multiple levels, with the longest chain classified as level 1. Chains at subsequent levels fill in the gaps in the Net levels above. Boxes in this track represent ungapped alignments, while lines signify the presence of gaps.

Because there are a large number of alignments in the "Chain" track, we will configure the *UCSC Genome Browser* so that it only shows the "Net" track. Scroll down to the track configuration section and click on the "Mouse Chain/Net" link under the "Comparative Genomics" section. Change the "Maximum display mode" to "full", the "Chain" drop-down box to "hide" and the "Net" drop-down box to "full." Click "Submit" and zoom out 10x (Figure 26).

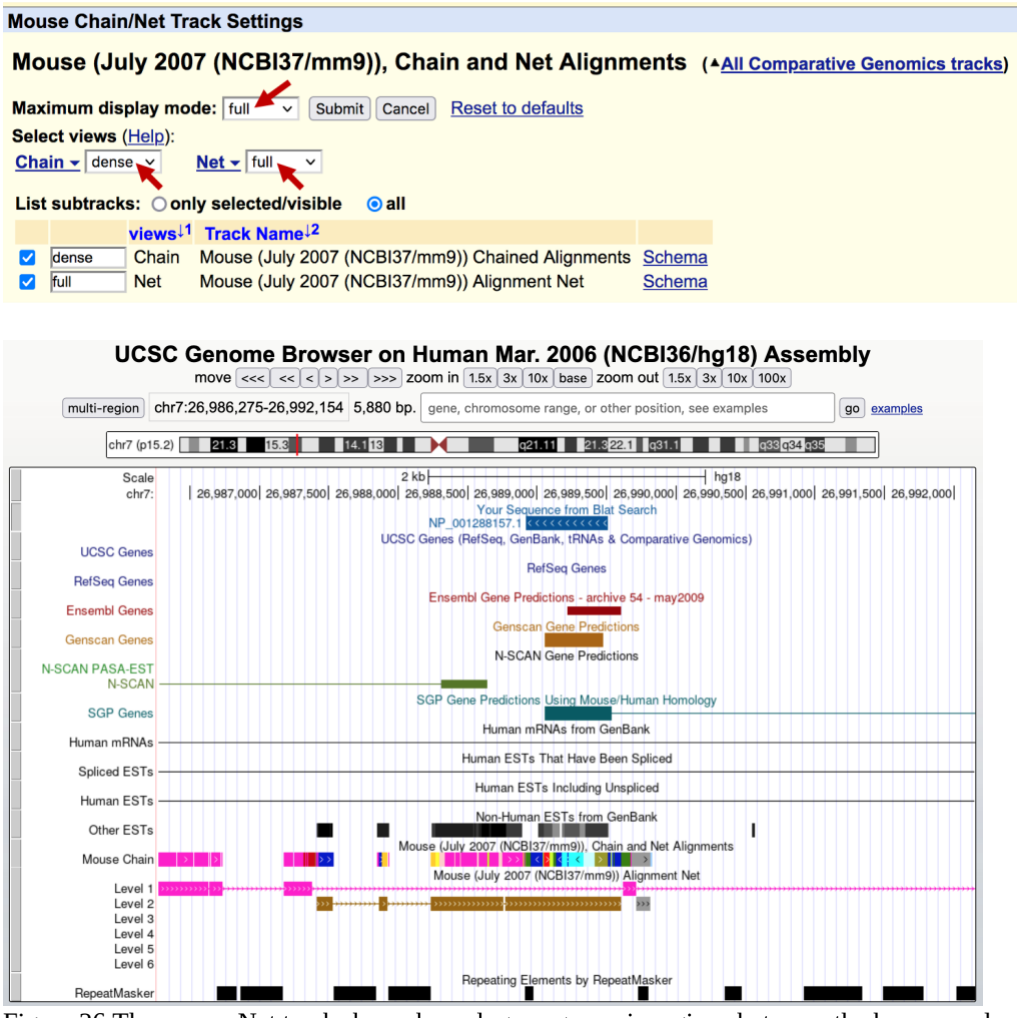


Figure 26 The mouse Net track shows homologous genomic regions between the human and mouse chromosomes.

When we click on the level 1 feature in the Mouse Alignment Net track (i.e. the boxes in pink), we find that the region surrounding the pseudogene is most similar to chromosome 6 of the mouse genome. By contrast, when we click on the feature in level 2 of the Mouse Alignment Net track (i.e. the boxes in brown), we find that the region that overlaps with the pseudogene is most similar to chromosome 1 in the mouse genome (Figure 27).

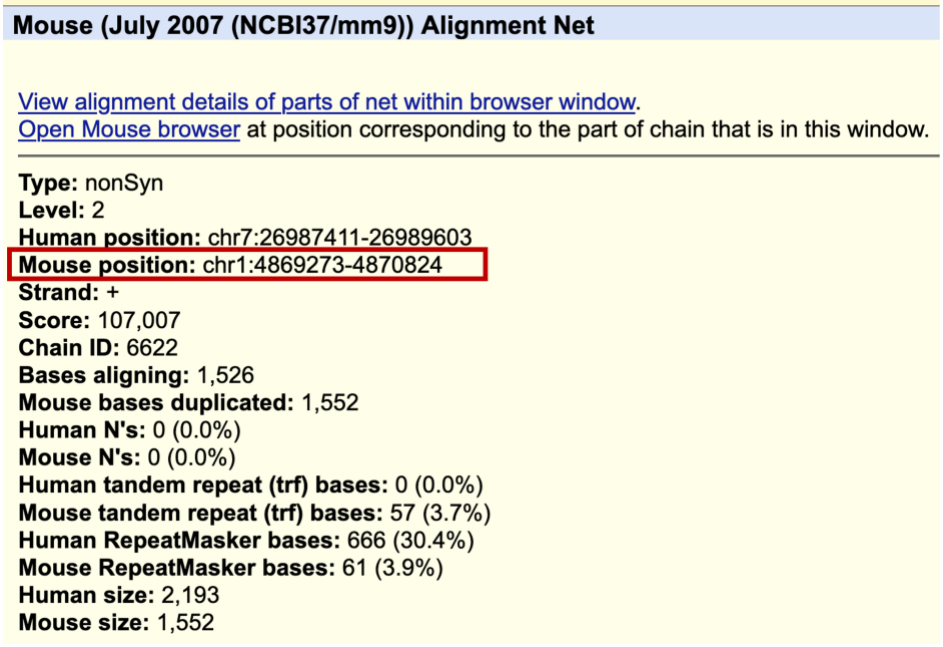


Figure 27 The *HMGB3* pseudogene region in human is homologous to chromosome 1 in mouse.

To investigate this difference further, we will perform a *BLAT* search against the "**Mouse**" genome ["**July 2007 (NCBI37/mm9)**" assembly] using the human HMGB3 protein sequence (Figure 28).

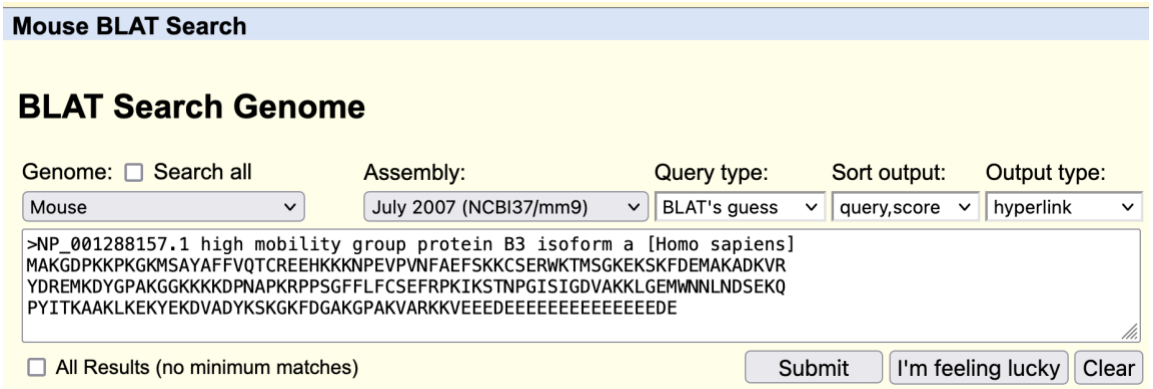
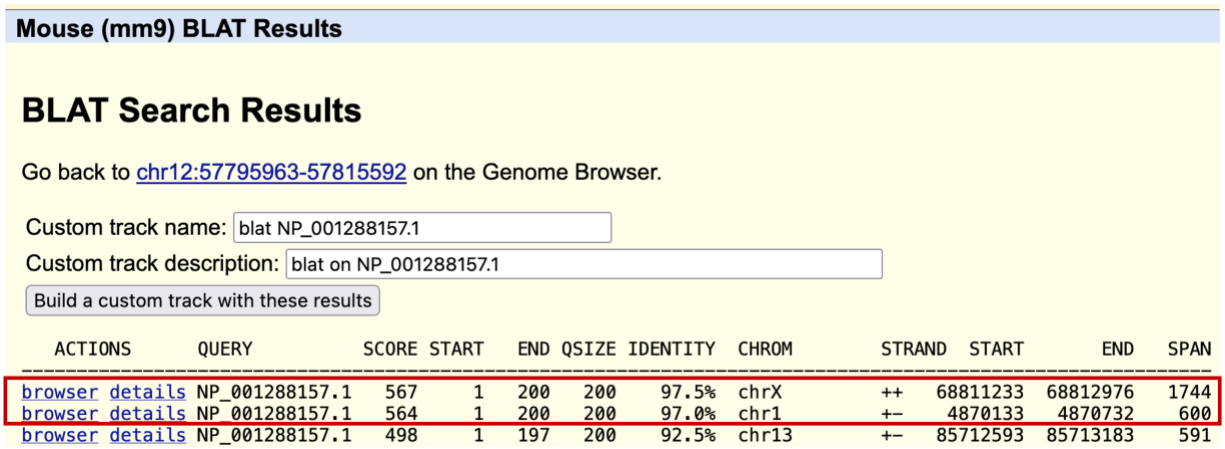


Figure 28 *BLAT* search of the human HMGB3 protein against the mouse assembly.

The *BLAT* results show two regions in the mouse genome with high (>= 97%) sequence identity to the human HMGB3 protein (Figure 29). Both of these hits match the entire length of the HMGB3 protein.



Mouse (mm9) BLAT Results

BLAT Search Results

Go back to chr12:57795963-57815592 on the Genome Browser.

Custom track name: blat NP_001288157.1

Custom track description: blat on NP_001288157.1

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	NP_001288157.1	567	1	200	200	97.5%	chrX	++	68811233	68812976	1744
browser details	NP_001288157.1	564	1	200	200	97.0%	chr1	+-	4870133	4870732	600
browser details	NP_001288157.1	498	1	197	200	92.5%	chr13	+-	85712593	85713183	591

Figure 29 *BLAT* alignment of the human HMGB3 protein against the mouse genome revealed two full-length matches with high percent identity.

Click on the "browser" link for the match with the highest percent identity (97.5%) on the X chromosome and then zoom out 3x. We observed that this feature is supported by RefSeq mRNAs and spliced ESTs. The feature has four coding exons, similar to what we saw for the real ortholog of the *HMGB3* gene in the human genome (Figure 30).

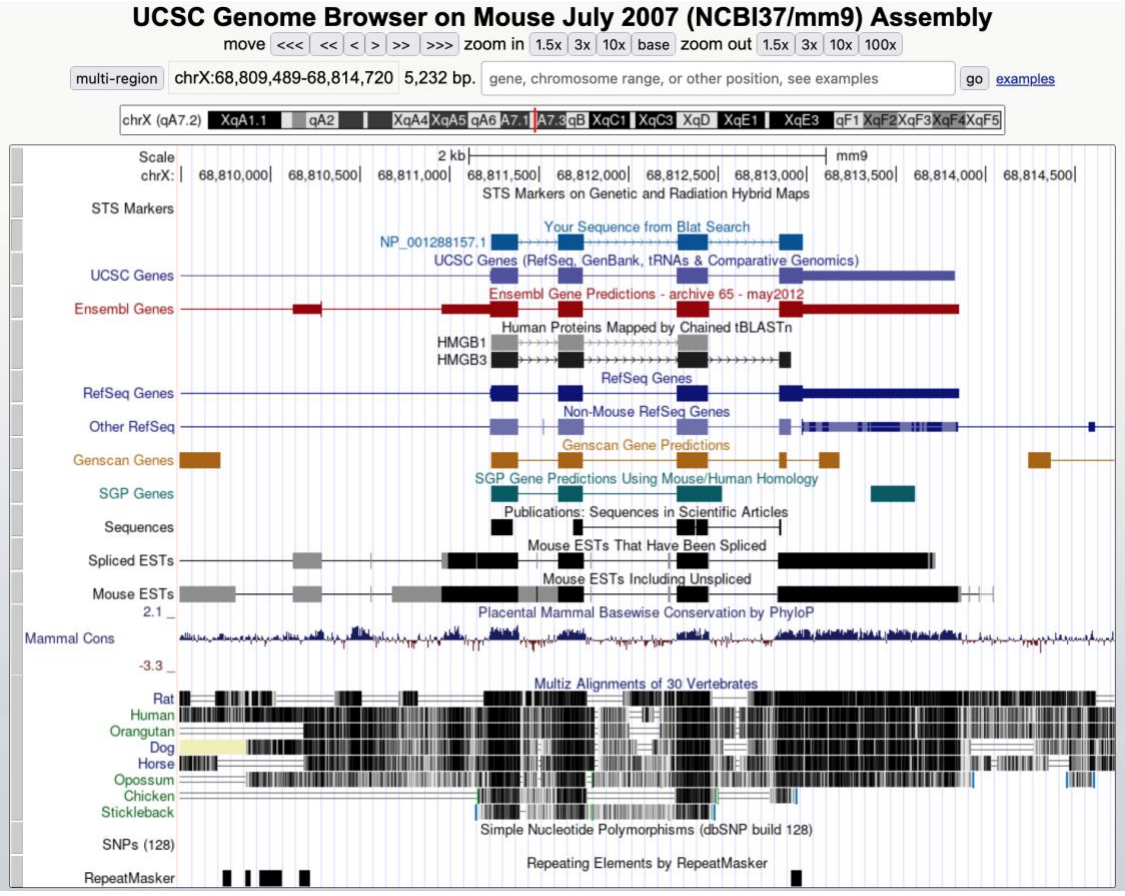


Figure 30 Putative ortholog of the human HMGB3 protein is found on the X chromosome of the mouse genome.

Now go back and click on the “details” link to examine the alignment to the match in chromosome 1 (with 97.0% identity). We see in the Side-by-Side Alignment that even though the feature consists of only a single exon, the in-frame stop codon has not yet been introduced into this sequence (Figure 31). Our analysis suggests the functional mouse ortholog of the HMGB3 protein is located on the X chromosome while the ortholog to the pseudogene is found on chromosome 1. In addition, our results suggest that the stop codon mutation was introduced into the pseudogene after the split of the primate and rodent lineages.

Side by Side Alignment*

```
0000001  M  A  K  G  D  P  K  K  P  K  G  K  M  S  A  Y  A  F  F  V  0000060
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870732 atggctaaaggtgaccccaagaaaccaaagggcaagatgtctgcttatgccttctttgtg 4870673

0000061  Q  T  C  R  E  E  H  K  K  K  N  P  E  V  P  V  N  F  A  E  0000120
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870672 cagacatgcagggaagaacataagaagaaaaacccagaggttcccgtcaattttgctgag 4870613

0000121  F  S  K  K  C  S  E  R  W  K  T  M  S  G  K  E  K  S  K  F  0000180
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  S  |  |  |  |  |  |  |  <<<<<<<
4870612 ttctccaagaagtgctcggagaggtggaagaccatgtctagcaaagagaaatcaaagttt 4870553

0000181  D  E  M  A  K  A  D  K  V  R  Y  D  R  E  M  K  D  Y  G  P  0000240
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870552 gatgaaatggcaaaggcagataaagtccgatatgatcgggagatgaaagattatggacca 4870493

0000241  A  K  G  G  K  K  K  K  D  P  N  A  P  K  R  P  P  S  G  F  0000300
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870492 gctaaaggaggcaagaagaagaaggacccaaatgcccccaaaagacctccgtctggattt 4870433

0000301  F  L  F  C  S  E  F  R  P  K  I  K  S  T  N  P  G  I  S  I  0000360
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870432 ttcttattctgttctgaattccgccccaagatcaaatccacaaaccctggcatctccatt 4870373

0000361  G  D  V  A  K  K  L  G  E  M  W  N  N  L  N  D  S  E  K  Q  0000420
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  |  |  |  S  |  N  |  |  |  <<<<<<<
4870372 ggagatgtggcaaaaagctgggtgagatgtggaataacttaagtgacaatgaaaagcag 4870313

0000421  P  Y  I  T  K  A  A  K  L  K  E  K  Y  E  K  D  V  A  D  Y  0000480
<<<<<<<  |  |  V  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870312 ccttatgtcaccaaggcagcaaagctgaaagagaagtatgagaaggatgttgctgactat 4870253

0000481  K  S  K  G  K  F  D  G  A  K  G  P  A  K  V  A  R  K  K  V  0000540
<<<<<<<  |  |  |  |  |  |  |  |  |  |  |  V  |  |  |  |  |  |  |  |  <<<<<<<
4870252 aagtctaaagggaagtttgatggtgccaagggtcctgttaaagttgcccggaaaaaggtg 4870193

0000541  E  E  E  D  E  E  E  E  E  E  E  E  E  E  E  E  E  D  E  0000600
<<<<<<<  |  |  |  E  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  <<<<<<<
4870192 gaagaagaggaagaggaggaggaagaggaagaagaggaggaggaagaggaggaagatgaa 4870133
```

Figure 31 The Side-by-Side alignment of the human HMGB3 protein against the mouse chromosome 1 shows that the stop codon has not yet been introduced into this potential retrotransposed pseudogene.

We can iteratively conduct the same experiment (i.e. perform *BLAT* searches against other genomes using the human HMGB3 protein) to obtain a better estimate as to when the pseudogene first acquired the stop codon. This analysis is left as an exercise for the reader.

Analysis of predicted gene 5 (three exons, 1017 coding bases)

Examine conserved domain matches from the *blastp* search of the predicted peptide against the nr protein database

We begin our analysis of the fifth *GENSCAN* prediction with a *blastp* search against the nr database. (The *blastp* results are available in the file *blastpGene5.txt* in the tutorial package.) The CDD search result under the "Graphic Summary" tab shows the predicted protein contains a conserved Homeobox domain (Figure 32).

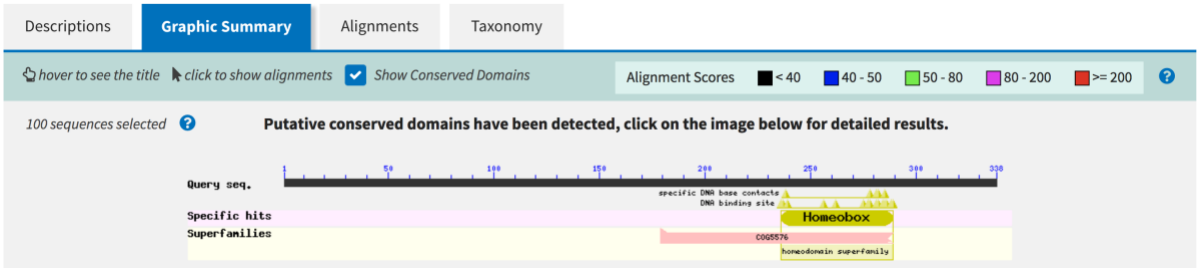


Figure 32 The fifth *GENSCAN* prediction contains a conserved Homeobox domain.

To obtain additional information regarding the function of the Homeobox domain, open a new web browser window and navigate back to the NCBI home page. On the top search bar, change the database to "Conserved Domains" and then search for "Homeobox" (Figure 33). Click on the first hit (with the accession number "pfam00046") to access the CDD record (Figure 34).

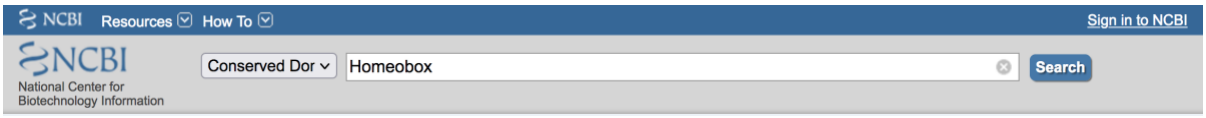


Figure 33 Search for "Homeodomain" in the Conserved Domain database.

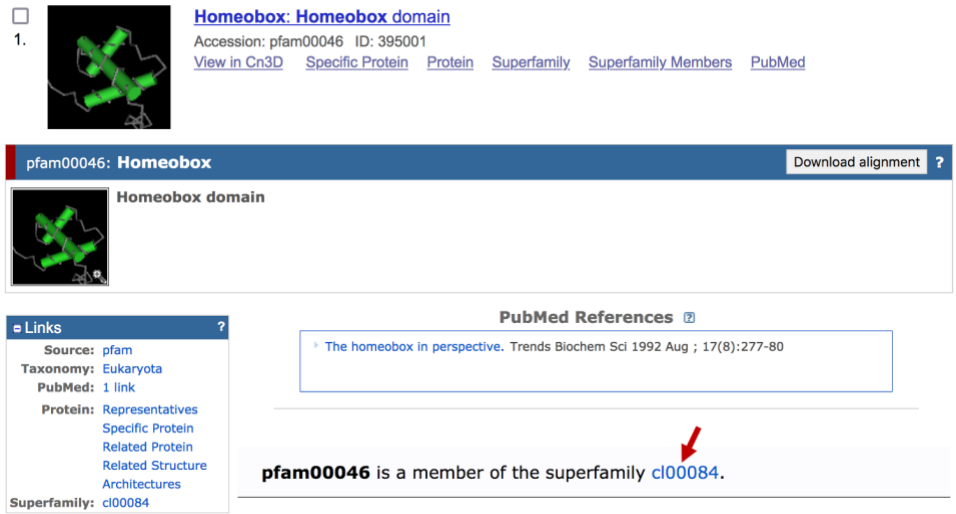


Figure 34 The Homeobox domain record at the NCBI Conserved Domain database.

Click on the “cl00084” link to navigate to the superfamily record for the Homeobox domain. We can find out more about the function of the homeodomain superfamily from the description of the CDD record:

> Homeodomain; DNA binding domains involved in the transcriptional regulation of key eukaryotic developmental processes; may bind to DNA as monomers or as homo- and/or heterodimers, in a sequence-specific manner.

Examine *blastp* alignments of predicted peptide against the nr database

Go back to the *blastp* results page and then click on the “Descriptions” tab. The first RefSeq match with the "NP_" prefix is NP_005513.2 and it corresponds to the human homeobox protein Hox-A1 isoform a (Figure 35).

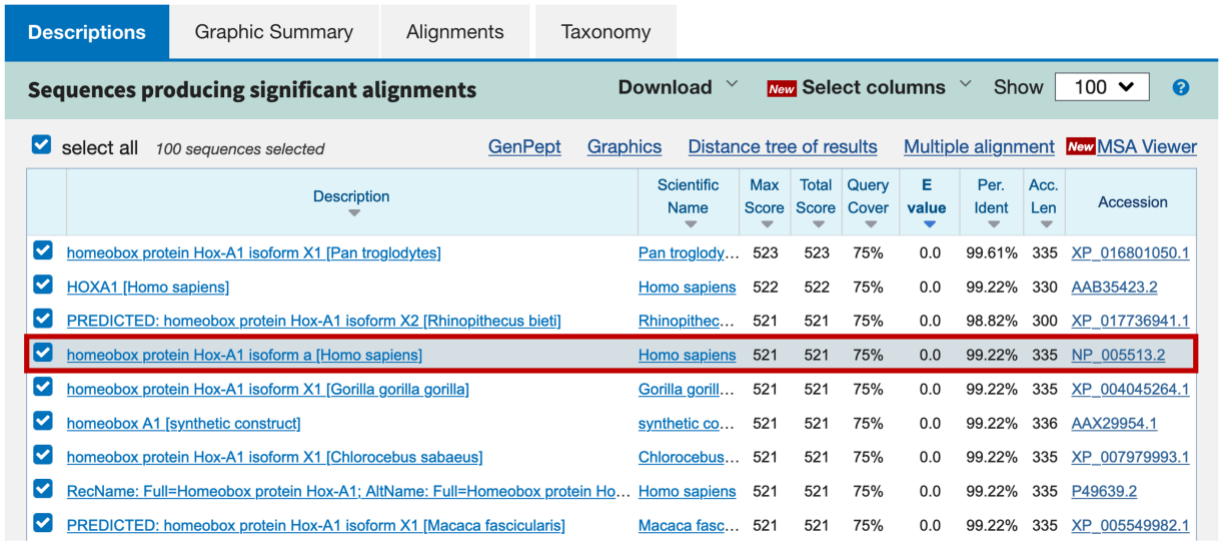
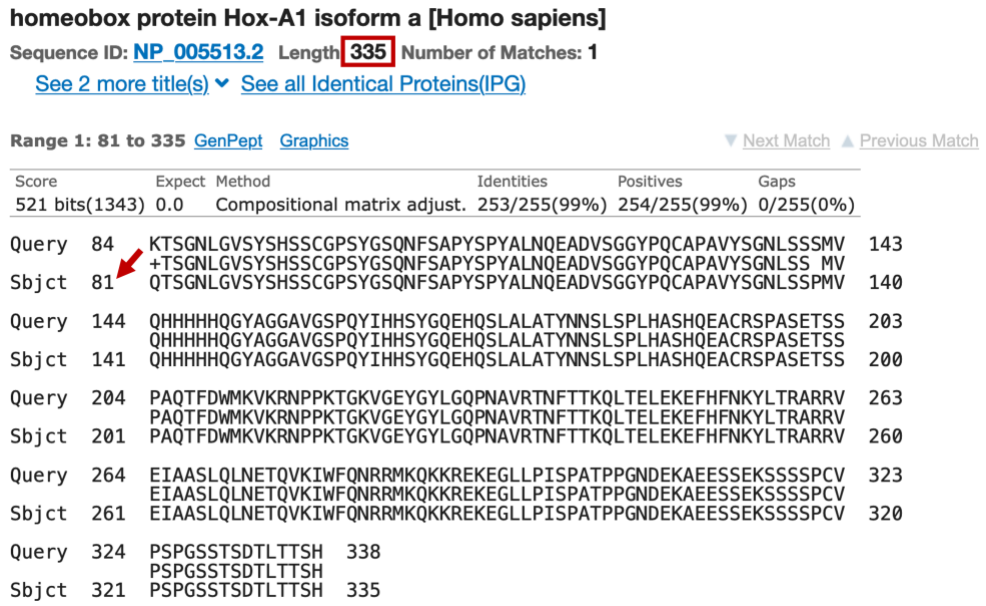


Figure 35 The fifth *GENSCAN* prediction is similar to the homeobox protein Hox-A1 in human.

Click on the description link to jump to the alignments for this match (Figure 36). We find that the human Hox-A1 protein has a total length of 335 amino acids but the *blastp* alignment to the *GENSCAN* prediction only contains the last 255 amino acids (i.e. the first 80 amino acids of the protein are missing). However, the high degree of sequence similarity between the *GENSCAN* prediction and the human protein suggests the match is not spurious. To figure out what happened, we will have to perform *BLAT* searches with both the human Hox-A1 protein and our *GENSCAN* predicted peptide against the human genome.



homeobox protein Hox-A1 isoform a [Homo sapiens]

Sequence ID: NP_005513.2 Length 335 Number of Matches: 1

See 2 more title(s) See all Identical Proteins(IPG)

Range 1: 81 to 335 GenPept Graphics ▼ Next Match ▲ Previous Match

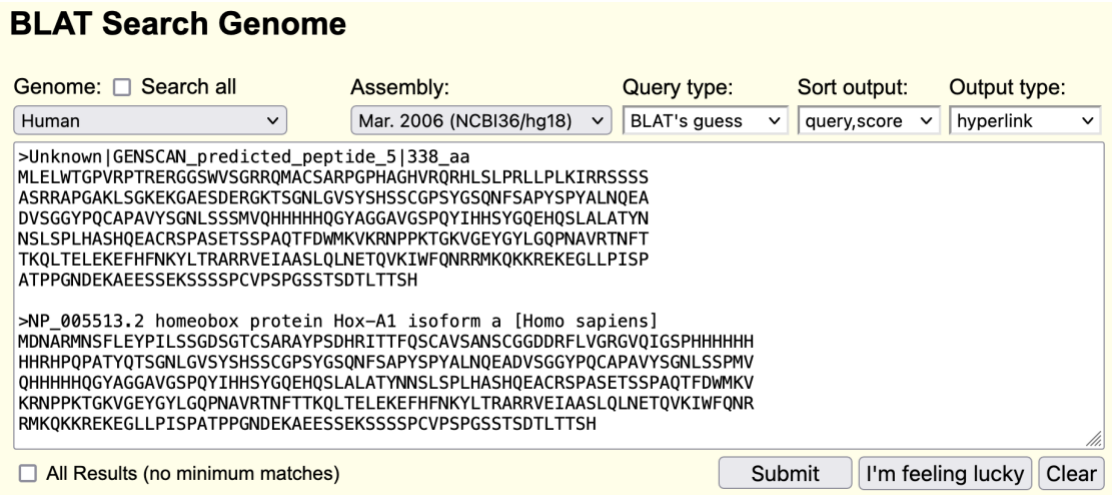
Score	Expect	Method	Identities	Positives	Gaps
521 bits(1343)	0.0	Compositional matrix adjust.	253/255(99%)	254/255(99%)	0/255(0%)

```
Query  84   KTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQCAPAVYSGNLSSSMV  143
            +TSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQCAPAVYSGNLSS MV
Sbjct  81   QTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQCAPAVYSGNLSSPMV  140

Query  144  QHHHHHQGYAGGAVGSPQYIHHSYGQEHQSLALATYNNSLSPLHASHQEACRSPASETSS  203
            QHHHHHQGYAGGAVGSPQYIHHSYGQEHQSLALATYNNSLSPLHASHQEACRSPASETSS
Sbjct  141  QHHHHHQGYAGGAVGSPQYIHHSYGQEHQSLALATYNNSLSPLHASHQEACRSPASETSS  200

Query  204  PAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFTTKQLTELEKEFHFNKYLTRARRV  263
            PAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFTTKQLTELEKEFHFNKYLTRARRV
Sbjct  201  PAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFTTKQLTELEKEFHFNKYLTRARRV  260

Query  264  EIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISPATPPGNDEKAEESSEKSSSSPCV  323
            EIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISPATPPGNDEKAEESSEKSSSSPCV
Sbjct  261  EIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISPATPPGNDEKAEESSEKSSSSPCV  320

Query  324  PSPGSSTSDTLTTSH  338
            PSPGSSTSDTLTTSH
Sbjct  321  PSPGSSTSDTLTTSH  335
```

Figure 36 The *blastp* alignment between the human homeobox Hox-A1 protein (subject) and the *GENSCAN* prediction (query) shows the first 80 residues of the human Hox-A1 protein is missing from the alignment.

Analysis of predicted and known proteins using the *UCSC Genome Browser*

For this analysis against the human genome, we will perform a *BLAT* search using both the human Hox-A1 protein and our predicted protein. To demarcate the two sequence records, each of these two sequences must be preceded by a line starting with the ">" symbol followed by a description of the sequence. Obtain the predicted protein from the *GENSCAN* page and the Hox-A1 peptide sequence from its GenBank entry as we have done previously. Paste the peptide sequences from the *GENSCAN* prediction and the human Hox-A1 protein into the *BLAT* search box (Figure 37). Change the "Genome" to "**Human**", the "Assembly" to "**Mar. 2006 (NCBI36/hg18)**", and then click "Submit".



BLAT Search Genome

Genome: ☐ Search all | Assembly: | Query type: | Sort output: | Output type:

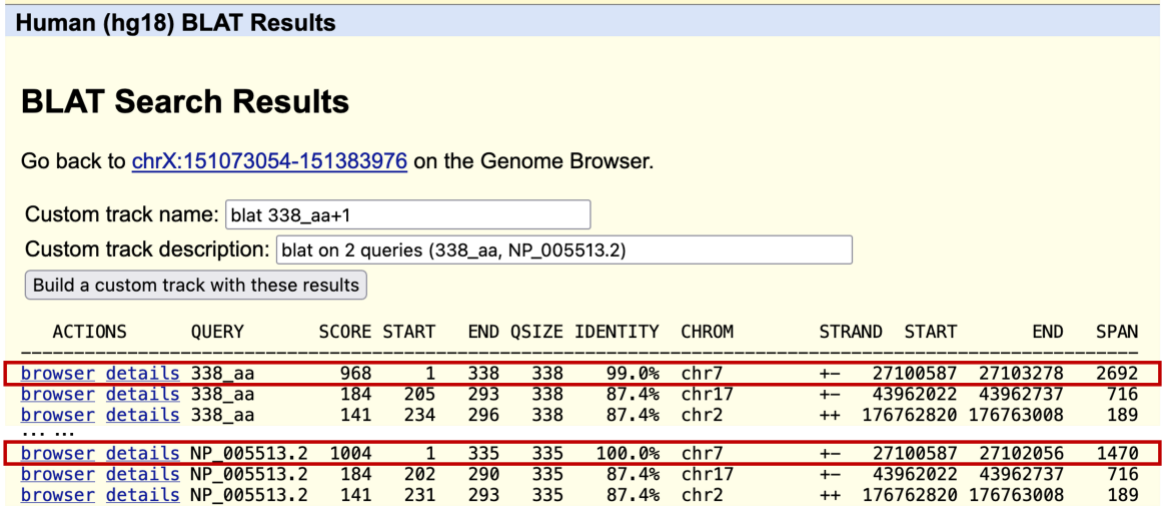
Human | Mar. 2006 (NCBI36/hg18) | BLAT's guess | query,score | hyperlink

```
>Unknown|GENSCAN_predicted_peptide_5|338_aa
MLELWTGPVRPTRERGGSWVSGRRQMACSARPGPHAGHVRQRHLSLPRLLPLKIRRSSSS
ASRRAPGAKLSGKEKGAESDERGKTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEA
DVSGGYPQCAPAVYSGNLSSSMVQHHHHHQGYAGGAVGSPQYIHHSYGQEHQSLALATYN
NSLSPLHASHQEACRSPASETSSPAQTFDWMKVKRNPPKTGKVGEYGYLGQPNAVRTNFT
TKQLTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWFQNRRMKQKKREKEGLLPISP
ATPPGNDEKAEESSEKSSSSPCVPSPGSSTSDTLTTSH

>NP_005513.2 homeobox protein Hox-A1 isoform a [Homo sapiens]
MDNARMNSFLEYPILSSGDSGTCSARAYPSDHRITTFQSCAVSANSCGGDDRFLVGRGVQIGSPHHHHHH
HHRHPQPATYQTSGNLGVSYSHSSCGPSYGSQNFSAPYSPYALNQEADVSGGYPQCAPAVYSGNLSSPMV
QHHHHHQGYAGGAVGSPQYIHHSYGQEHQSLALATYNNSLSPLHASHQEACRSPASETSSPAQTFDWMKV
KRNPPKTGKVGEYGYLGQPNAVRTNFTTKQLTELEKEFHFNKYLTRARRVEIAASLQLNETQVKIWFQNR
RMKQKKREKEGLLPISPATPPGNDEKAEESSEKSSSSPCVPSPGSSTSDTLTTSH
```

☐ All Results (no minimum matches) Submit | I'm feeling lucky | Clear

Figure 37 Perform a *BLAT* search with multiple query sequences. Use the '>' symbol followed by a description of the sequence to demarcate the two sequence records.

The top hits for our predicted protein and the human Hox-A1 protein are to the same region on human chromosome 7 (Figure 38). This result is consistent with the hypothesis that our predicted protein is the chimpanzee ortholog of the human Hox-A1 protein.



Human (hg18) BLAT Results

BLAT Search Results

Go back to chrX:151073054-151383976 on the Genome Browser.

Custom track name: blat 338_aa+1

Custom track description: blat on 2 queries (338_aa, NP_005513.2)

Build a custom track with these results

ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHROM	STRAND	START	END	SPAN
browser details	338_aa	968	1	338	338	99.0%	chr7	+-	27100587	27103278	2692
browser details	338_aa	184	205	293	338	87.4%	chr17	+-	43962022	43962737	716
browser details	338_aa	141	234	296	338	87.4%	chr2	++	176762820	176763008	189
... ...											
browser details	NP_005513.2	1004	1	335	335	100.0%	chr7	+-	27100587	27102056	1470
browser details	NP_005513.2	184	202	290	335	87.4%	chr17	+-	43962022	43962737	716
browser details	NP_005513.2	141	231	293	335	87.4%	chr2	++	176762820	176763008	189

Figure 38 The best match to the predicted protein (338_aa) and the human Hox-A1 protein (NP_005513.2) are both located in the same region on human chromosome 7.

Click on the "browser" link for the best match on human chromosome 7 to examine the genomic region that contains the best hit to the *GENSCAN* prediction. Zoom out 3x and compare the two features under "Your sequences from Blat search." The names of the two features are identified by the names you entered after the ">" in the query textbox. The prediction (338_aa) and the Hox-A1 gene (NP_005513.2) cover the same region of the genome, except that the prediction has an additional intron stuck in the first exon of Hox-A1 (Figure 39).

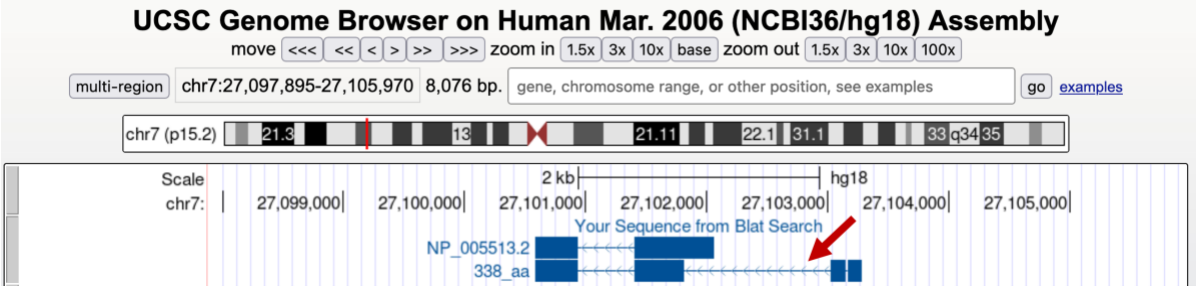


Figure 39 Extra intron in our predicted protein (338_aa) compared to the human Hox-A1 protein (NP_005513.2).

If you change the "*RepeatMasker*" track to "full" and look at that track below the spurious intron, you will see why *GENSCAN* predicted this intron (Figure 40). There is a simple repeat [(TGG)n] that covers the end of the predicted intron. It is not *GENSCAN*'s fault that it got the prediction wrong; rather, it was because *GENSCAN* was run after low complexity sequences were removed. The presence of this masked sequence caused *GENSCAN* to make a faulty prediction. You can try to run *GENSCAN* without masking the sequence first. However, this could substantially increase the amount of time needed to complete the search.

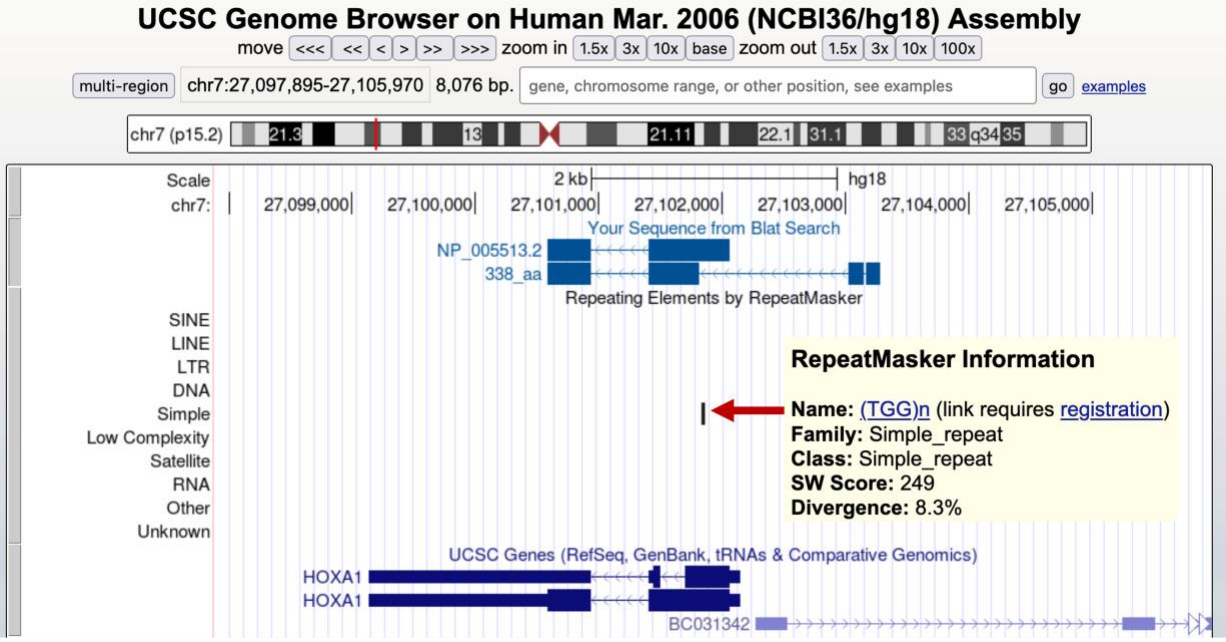


Figure 40 *RepeatMasker* identified a simple repeat at the end of a predicted intron.

Annotation of the remaining *GENSCAN* gene predictions

Now that you have seen the process of how to annotate potential genes using *GENSCAN*, *BLAST*, and the *UCSC Genome Browser*, you can try to annotate the remaining six genes predicted by *GENSCAN* in our chimp BAC, or in other regions of interests. Are they genes or pseudogenes? What kind of evidence can you obtain to support your hypothesis?