

Question 1

How does one find a gene of interest and determine that gene's structure? Once the gene has been located on the map, how does one easily examine other genes in that same region?

doi:10.1038/ng966

This question serves as a basic introduction to the three major genome viewers. One gene, *ADAM2*, will be examined using all three sites so that the reader can gain an appreciation of the subtle differences in information presented at each of these sites.

National Center for Biotechnology Information Map Viewer

The NCBI Human Map Viewer can be accessed from the NCBI's home page, at <http://www.ncbi.nlm.nih.gov>. Follow the hyperlink in the right-hand column labeled *Human map viewer* to go to the Map Viewer home page. The notation at the top of the page indicates that this is Build 29, or the NCBI's 29th assembly of the human genome. Build 29 is based on sequence data from 5 April 2002. The previous genome assembly, Build 28, was based on sequence data from 24 December 2001. To search for any mapped element, such as a gene symbol, GenBank accession number, marker name or disease name, enter that term in the *Search for* box and then press *Find*. For this example, enter 'ADAM2' and then press *Find*. The *on chromosome(s)* box may be left blank for text-based searches such as this one.

The resulting overview page shows a schematic of all of the human chromosomes, pinpointing the position of *ADAM2* to the p arm of chromosome 8 (Fig. 1.1). The search results section shows that the gene exists on two NCBI maps, *Genes_cyto* and *Genes_seq*. *Genes_cyto* refers to the cytogenetic map, whereas *Genes_seq* refers to the sequence map. Clicking on either of those two links opens a view of just that map.

Detailed descriptions of these and other NCBI maps are available at <http://www.ncbi.nlm.nih.gov/PMGifs/Genomes/humansearch.html>. To get the most general overview of the genomic context of *ADAM2*, including all available maps, click on the item in the *Map element* column (in this case, *ADAM2*). This view shows *ADAM2* and a bit of flanking sequence on chromosome 8p11.2 (Fig. 1.2). Three maps are displayed in this view, each of which will be discussed below. Additional maps, discussed in other examples in this guide, can be added to this view using the *Maps & Options* link.

The rightmost map is the master map, the map providing the most detail. The master map in this case is the *Genes_seq* map, which depicts the intron/exon organization of *ADAM2* and is created by aligning the *ADAM2* mRNA to the genome. The gene appears to have 14 exons. The vertical arrow next to the *ADAM2* gene symbol (within the pink box) shows the direction in which the gene is transcribed. The gene symbol itself is linked to LocusLink, an NCBI resource that provides comprehensive information about the gene, including aliases, nucleotide and protein sequences, and links to other resources¹⁰ (see Question 10). The links to the right of the gene symbol point to additional information about the gene.

- *sv*, or sequence view, shows the position of the gene in the context of the genomic contig, including the nucleotide and encoded protein sequences.

- *ev* brings the user to the evidence viewer, a view that displays the biological evidence supporting a particular gene model. This view shows all RefSeq models, GenBank mRNAs, transcripts (whether annotated, known or potential) and expressed sequence tags (ESTs) aligning to this genomic contig. More information on the evidence viewer can be found on the NCBI web site by clicking *Evidence Viewer Help* on any *ev* report page.
- *hm* is a link to the NCBI's Human–Mouse Homology Map, showing genome sequences with predicted orthology between mouse and human (Fig. 12.2).
- *seq* allows the user to retrieve the genomic sequence of the region in text format. The region of sequence displayed can easily be changed.
- *mm* is a link to the Model Maker, which shows the exons that result when GenBank mRNAs, ESTs and gene predictions are aligned to the genomic sequence. The user can then select individual exons to create a customized model of the gene. More information on the Model Maker can be found on the NCBI web site by clicking *help* on any *mm* report page.

The UniG_Hs map shows human UniGene clusters that have been aligned to the genome. The gray histogram depicts the number of aligning ESTs and the blue lines show the mapping of UniGene clusters to the genome. The thick blue bars are regions of alignment (that is, exons) and the thin blue lines indicate potential introns. In this example, the mapping of UniGene cluster Hs.177959 to the genome follows that of *ADAM2*, and all the exons align.

The *Genes_cyto* map shows genes that have been mapped cytogenetically; the orange bar shows the position of the gene. Although *ADAM2* has been finely mapped and is represented by a short line, other genes, such as the group below it on a longer line, have been cytogenetically mapped to broader regions of chromosome 8.

Clicking on the zoom control in the blue sidebar allows the user to zoom out to view a larger region of chromosome 8. Zooming out one level shows 1/100th of the chromosome. There are 20 genes in the region, and all 20 are labeled (displayed) in this view (Fig. 1.3). The region of *ADAM2* is highlighted in red on all maps. On the basis of the *Genes_seq* map, *ADAM2* is located between *ADAM18* and *LOC206849*.

University of California, Santa Cruz Genome Browser

The home page for the UCSC Genome Browser is <http://genome.ucsc.edu/>. At present, UCSC provides browsers not only for the most recent version of the mouse and human genome data, but also for several earlier assemblies. To use the Genome Browser, select the appropriate organism from the pull-down menu at the top of the blue sidebar (*Human*, in this case) and then click the link labeled *Browser*. On the resulting page, select the version of the human assembly to view. The genome browser from August 2001 is based on an assembly of the human genome done by UCSC using sequence data available on that date. The *Dec. 2001*

browser displays annotations based on NCBI's build 28 of the human genome, and the Apr. 2002 browser displays annotations on NCBI's build 29. As the annotations presented in this most recent human assembly are not yet as comprehensive as those from the December 2001 assembly, the examples in this text are based on the earlier assembly. Select *Dec. 2001* from the pull-down menu to access the assembly from that date (Fig. 1.4).

Supported types of queries are listed below the text input boxes. Enter 'ADAM2' in the box labeled *position* and then click *Submit*. The results of this search are presented in two categories, *Known Genes* and *mRNA Associated Search Results* (Fig. 1.5). The section marked *Known Genes* shows the mapping of the NCBI Reference mRNA sequences to the genome. The *mRNA Associated Search Results* represent the mapping of other GenBank mRNA sequences to the genome. Click on the *Known Genes* link for *ADAM2* (arrow, Fig. 1.5) to see the genomic context of the *ADAM2* mRNA Reference Sequence (NM_001464).

The resulting zoomed-in view shows a region of chromosome 8 from base pair 36234934 to 36280132, located within 8p12 (Fig. 1.6). The blue track entitled *Known Genes (from RefSeq)* shows the intron–exon structure of known genes. The vertical boxes indicate exons and the horizontal lines introns. The *ADAM2* gene seems to have 14 exons. The direction of transcription is indicated by the arrowheads on the introns. The tracks labeled *Acembly Gene Predictions*, *Ensembl Gene Predictions* and *Fgenesh++ Gene Predictions* are the results of gene predictions (see Question 7). Alignments of other database nucleotide sequences are shown in the *Human mRNAs from GenBank*, *spliced EST*, *UniGene* and *Nonhuman mRNAs from GenBank* tracks. Translated alignments of mouse and *Tetraodon* genomic sequence are in the *mouse* and *fish BLAT* tracks. Tracks displaying single-nucleotide polymorphisms (SNPs), repetitive elements and microarray data are shown at the bottom. Additional details about each track are available by selecting the track name in the *Track Controls* at the bottom.

To view the genomic context of *ADAM2*, zoom out 10× by clicking on the *zoom out 10×* box in the upper right corner. *ADAM2* is located between *TEM5* and *ADAM18* (Fig. 1.7).

Ensembl

The Ensembl⁷ project, <http://www.ensembl.org/>, provides genome browsers for four species: human, mouse, zebrafish and mosquito. Click on *Human* to view the main entry point for the human genome. The current version of human Ensembl is version 6.28.1, based on the NCBI's 28th build of the genome. To perform a text search, enter 'ADAM2' in the text box, and limit the search by selecting *Gene* from the pull-down search. Click on the upper button labeled *Lookup*. A single result is returned with a link to the *ADAM2* gene (Fig. 1.8).

Click on either of the *ADAM2* links to retrieve the GeneView window. The returned page contains four sections of data. The first section (Fig. 1.9) is an overview of *ADAM2*, including links to accession numbers and protein domains and families. Links to the Ensembl view of highly similar mouse sequences are presented in the *Homology Matches* section. Some of these fields will be described in more detail in later examples. The second section of the GeneView window provides information on the gene transcript (Fig. 1.10). The sequence of the cDNA is shown, as is a graphic of its intron–exon structure. A limited amount of the genomic context around the gene is shown schematically as well.

Exon sequences are shown in the third section of the GeneView (Fig. 1.11) and splice sites in the fourth (Fig. 1.12). If more than one transcript is predicted for the gene, each is allocated its own transcript, exon and splice-site sections.

The complete genomic context of *ADAM2* is viewed by returning to the first section of the GeneView (Fig. 1.9) and clicking on one of the two links within the *Genomic Location* box. The top portion of the resulting ContigView (Fig. 1.13) depicts the chromosome, with the region of interest outlined in red. The Overview shows the genomic context of the gene, including the chromosome bands, contigs, markers and genes that map to near 8p12. Clicking on any of these items recenters the display around that item. The section of interest is boxed in red on the DNA(contigs) map. The genes annotated by Ensembl as being around *ADAM2* are *Q96KB2* and *ADAM18*.

The bottom panel of the ContigView, the Detailed View (Fig. 1.14), shows a zoomed-in view of the boxed region, highlighting all features that have been mapped to this region of the human genome. The navigator buttons between the Overview and the Detailed View move the display to the left and right and zoom in and out. The features to be displayed can be changed by selecting the *Features* pull-down menu and then checking which features to view.

The Features shown in Fig. 1.14 are the defaults. The DNA (contigs) map separates items on the forward strand (above) from those on the reverse (below). The only feature on the reverse strand in this view is a single Genscan transcript, predicted by the GENSCAN gene prediction program¹¹ (see Question 7). The forward strand shows five types of features. Starting at the bottom, the *ADAM2* transcript is shown in red, indicating that it is a known transcript corresponding to a near-full-length cDNA sequence, protein sequence or both already available in the public sequence database. Black transcripts are predicted based on EST or protein sequence similarity. *EST Transcr.* links to individual aligning ESTs, whereas the *UniGene* track near the top displays UniGene clusters. The Genscan model on the forward strand contains many exons found in the known transcript. The *Proteins* and *Human proteins* boxes indicate protein sequences that align to this version of the genome, whereas *NCBI Transcr.* links to the NCBI Map Viewer. Positioning the computer mouse over any feature brings up the feature's name and links to more detailed information.

The NCBI, UCSC and Ensembl sometimes use different symbols for the same genes, so it can be difficult to compare the views obtained by the different browsers. Furthermore, the three sites maintain independent annotation pipelines and do not all attempt to align the same mRNA sequences to the genome. The NCBI is currently displaying build 29, Ensembl shows build 28, and UCSC offers both builds 28 (December 2001) and 29 (April 2002), although all examples from UCSC in this guide will be illustrated using the better-annotated build 28. Because of the differences between the two assemblies, there are subtle discrepancies between what is shown at the NCBI and what is available at UCSC and Ensembl. However, it is fairly easy to navigate among the three sites. The NCBI, for example, links to Ensembl and UCSC through the black boxes at the top of *LocusLink* entries for human genes, and Ensembl directs users to NCBI and UCSC through the "Jump to" link in its ContigView. Some versions of UCSC's Genome Browser have links to Ensembl and NCBI's Map Viewer in the blue bar at the top of each browser page.

Figure 1.1

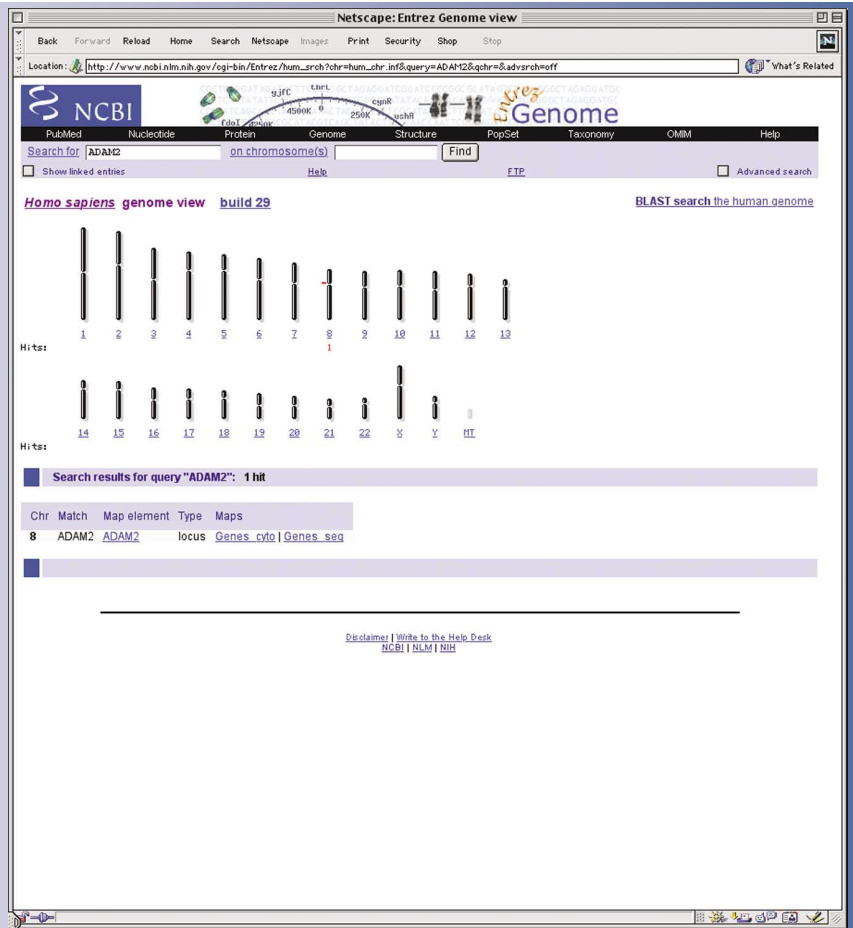
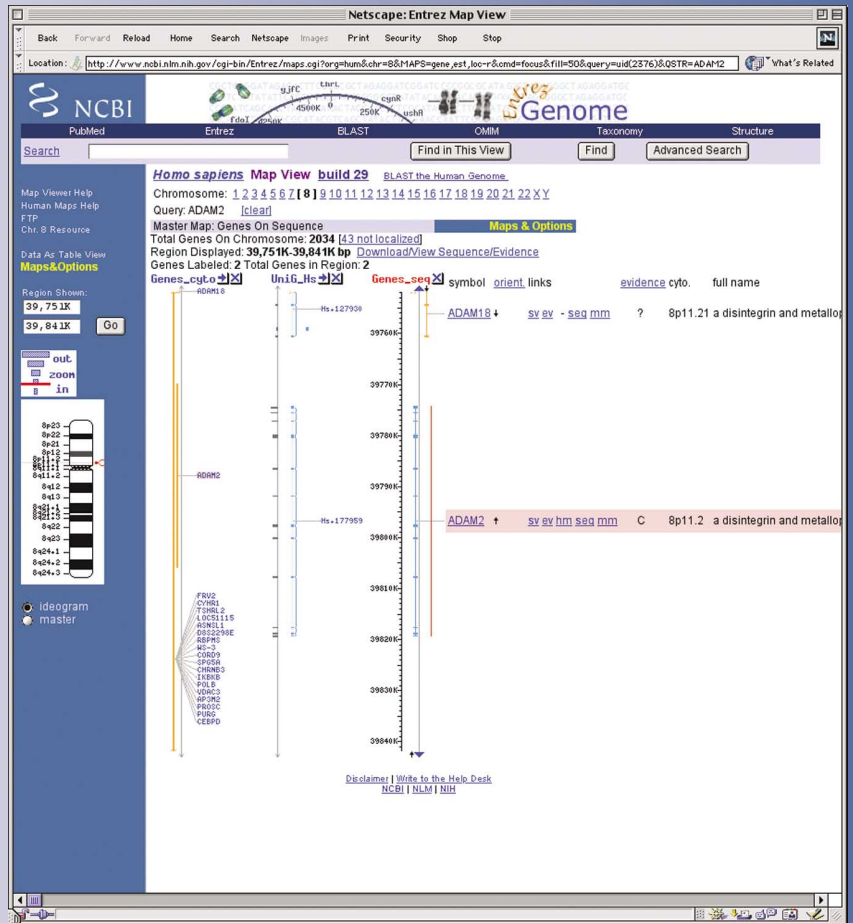


Figure 1.2



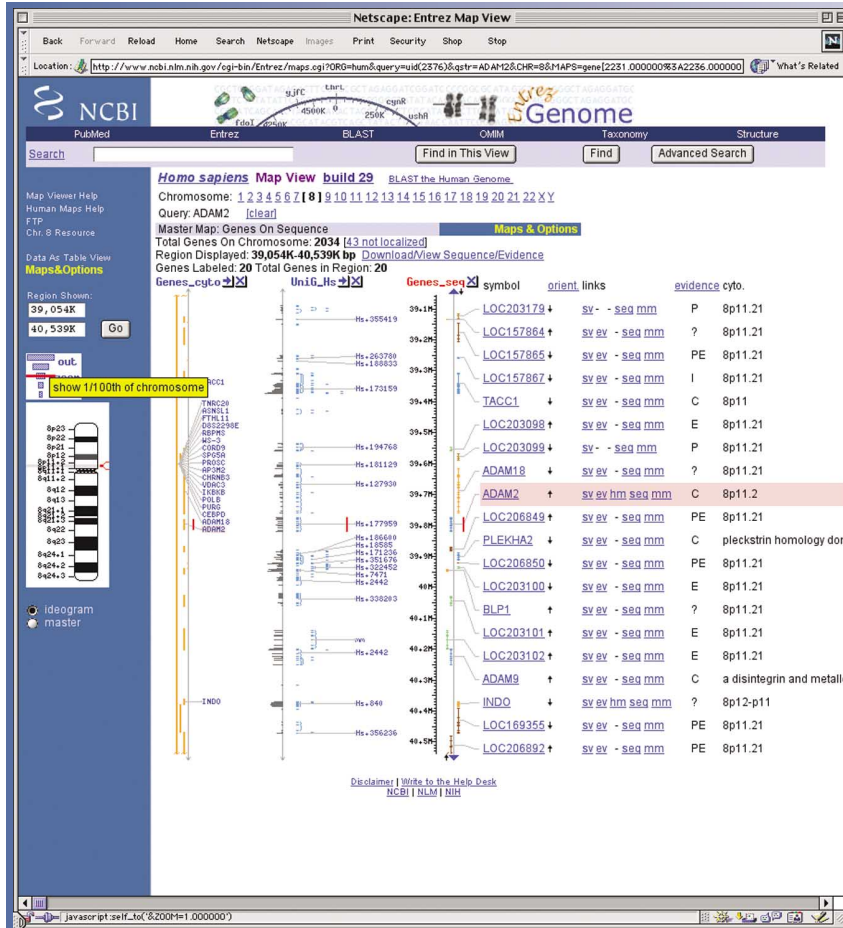


Figure 1.3

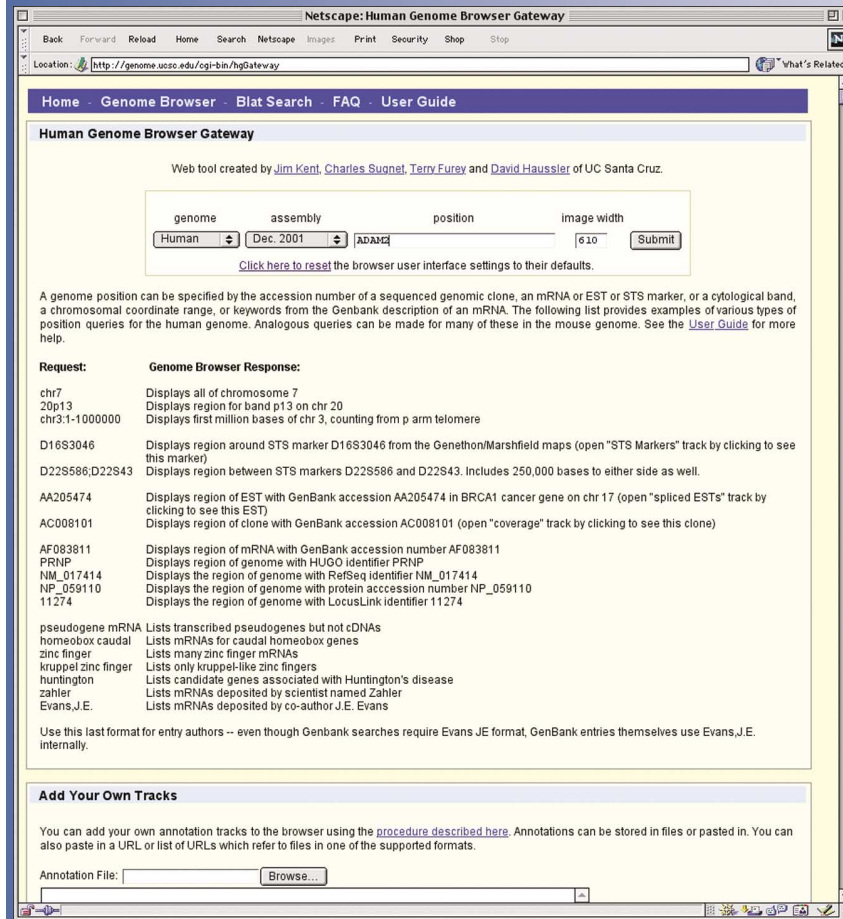


Figure 1.4

Figure 1.5

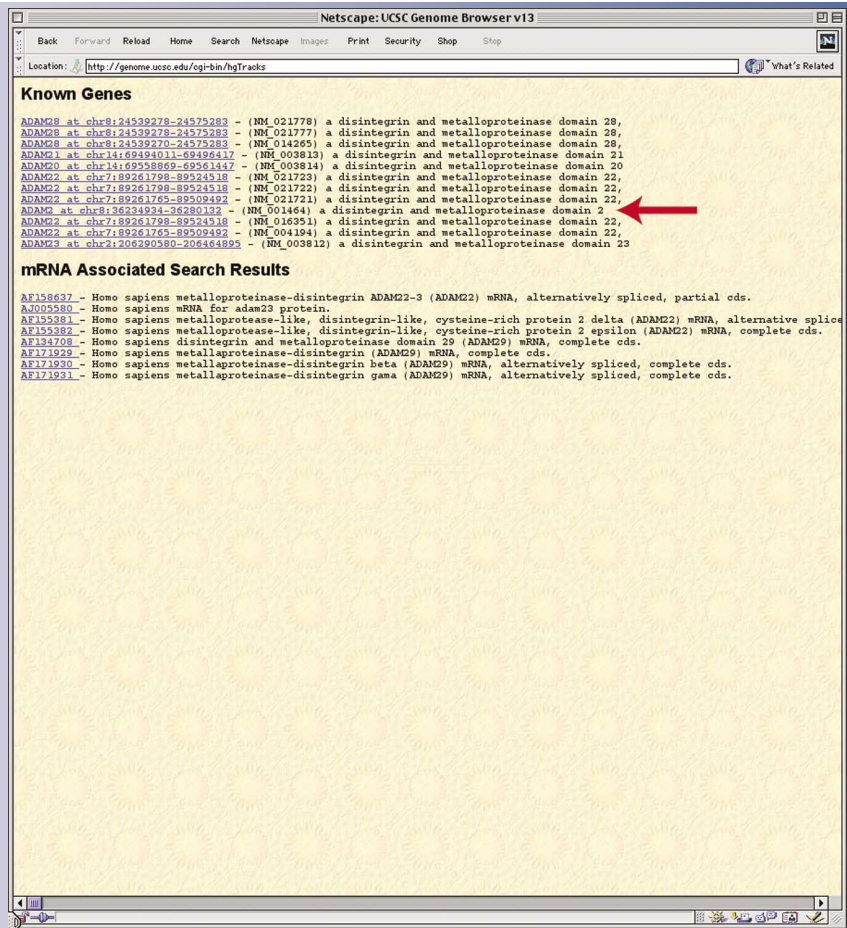
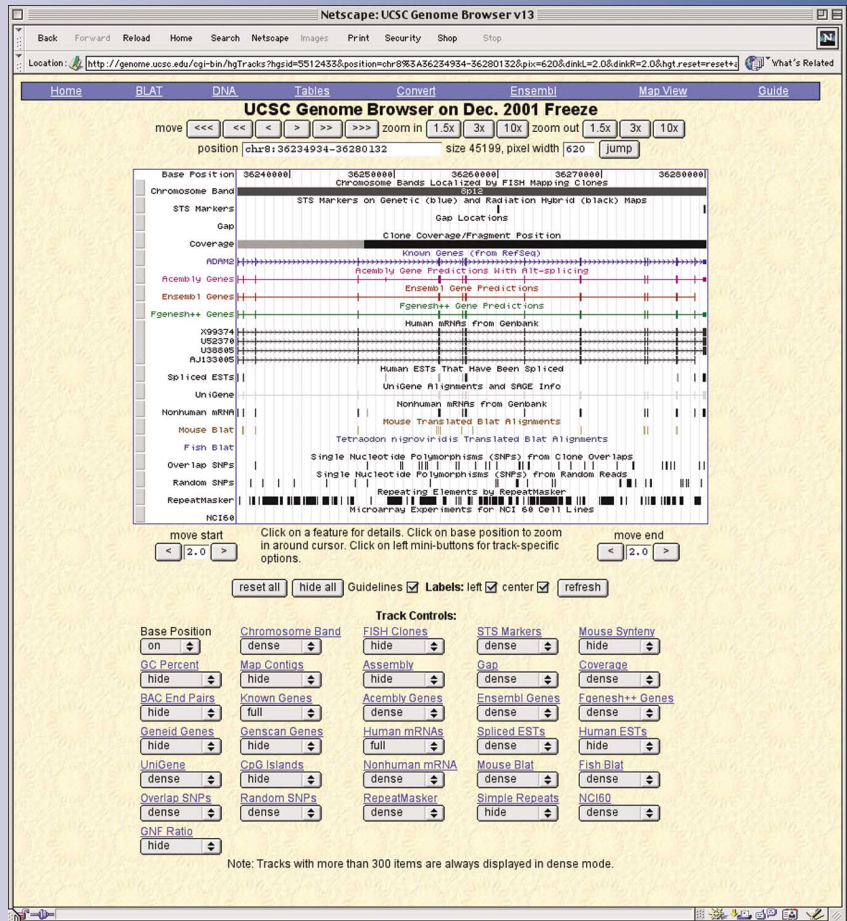


Figure 1.6



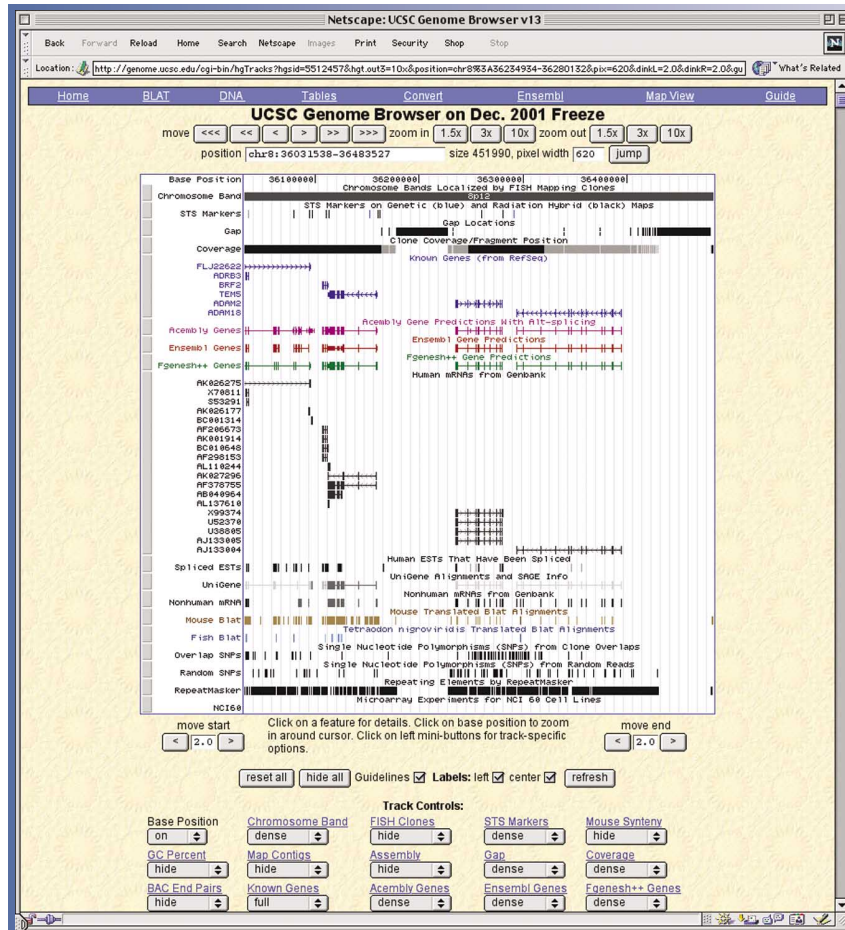


Figure 1.7

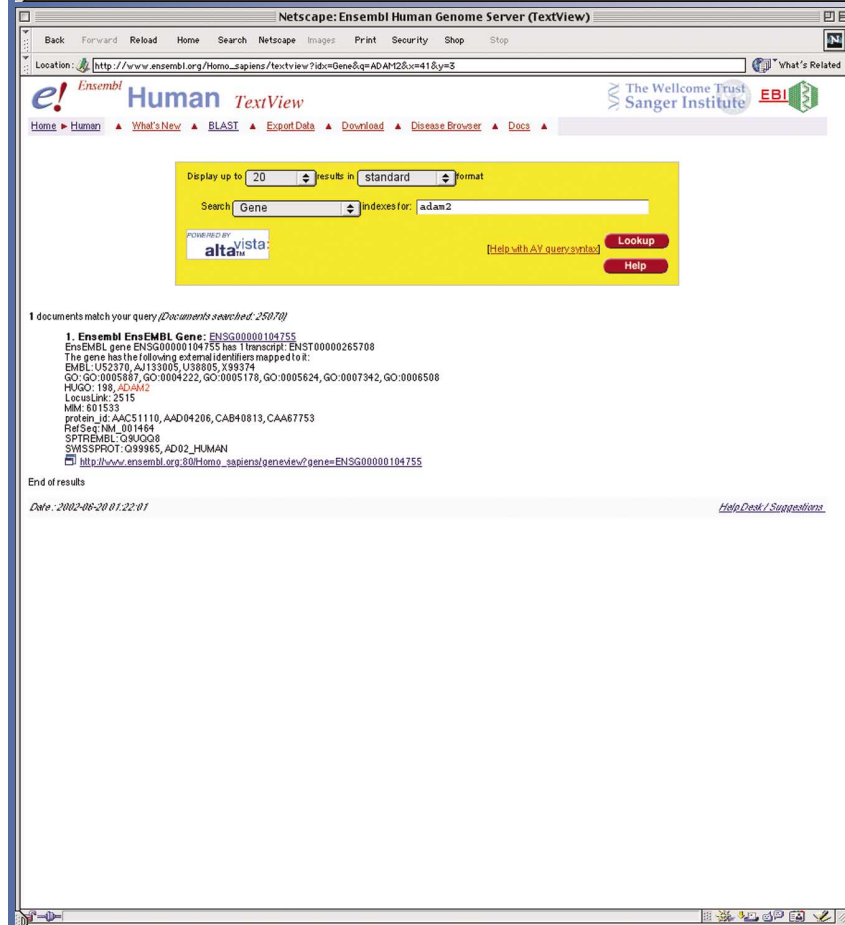


Figure 1.8

Figure 1.9

Ensembl Gene Report

Ensembl gene ID	ENSG00000104755
Genomic Location	View gene in genomic location: 36234936 - 36278004 bp (36.2 kb) on chromosome 8 This gene is located in sequence: AC_018807.5:6271236623
Description	ADAM 2 PRECURSOR [ADISINTEGRIN AND METALLOPROTEINASE DOMAIN 2] [FERTILIN BETA SUBUNIT] [PH-30] [PH30]. [Source:SWISSPROT,Acc:Q99985]
Prediction Method	This gene was predicted by the Ensembl analysis pipeline from either a GeneWise or Genscan prediction followed by confirmation of the exons by comparisons to protein, cDNA and EST databases.
Predicted Transcripts	1: ENST00000265708 [View supporting evidence] [View protein information]
Links	This Ensembl gene corresponds to the following other database identifiers EMBL: AL133005 [align] U38805 [align] U52370 [align] X98374 [align] GO: GO:0084312 [align] GO:0085118 [align] GO:0085624 [align] GO:0085897 [align] GO:0085508 [align] GO:0087342 [align] GO mapping is inherited from swissprot/spremb1 HUGO: Search GeneCards for ADAM2 LocalLink: 2515 [align] MIM: 601533 [align] RefSeq: NM_001464 [Target %id: 74; Query %id: 100] [align] SWISSPROT: A0D2_HUMAN [Target %id: 73; Query %id: 99] [align] [Search GO] SpTrEMBL: Q9LQ08 [Target %id: 74; Query %id: 100] [align] [Search GO] protein_id: AAC5110 [align] AAD04206 [align] CAA67753 [align] CAB40813 [align]
InterPro	IPR001530 Metalloprotease (ADAM type)/reprolysin (M12B) family [View other Ensembl genes with this domain] IPR001782 Disintegrin [View other Ensembl genes with this domain] IPR012436 Keratin, high sulfur B2 protein [View other Ensembl genes with this domain] IPR002839 EB module [View other Ensembl genes with this domain]
Protein Family	ENSF0000000063: ADAM PRECURSOR ADISINTEGRIN AND METALLO This cluster contains 21 Ensembl gene member(s)
Export Data	Export gene data in EMBL, GenBank or FASTA
Homology Matches	These genes have been identified as putative homologues by reciprocal BLAST analysis: Mus musculus ENSMUSG0000022038 [A0323749] ADISINTEGRIN AND METALLOPROTEINASE DOMAIN 2. [Source:SPTRMBL,Acc:Q9D4G3] Mus musculus ENSMUSG00000031555 [AJ020942] ADISINTEGRIN AND METALLOPROTEINASE DOMAIN 9 (MELTRIN GAMMA); MELTRIN GAMMA. [Source RefSeq,Acc:NM_007404]

Figure 1.10

Transcript 1: ENST00000265708
 Transcript cDNA Sequence
 Total length: 1635 bp. No. Exons: 13

Exon structure

Transcript neighbourhood

[View Evidence](#) [View Protein](#)





Exon Information							Exon Sequence
No.	Exon ID	Contig	Strand	Start	End	Length	
1	ENSE00000890080	AC018807.5.62712.86823	1	2796	2867	72 bp	TATAATCATATGGGGTCTGATACAACTGTTGTGCTCAAAAAGTTTCCA GTGGATTGGATTGGAAAGTCT
2	ENSE00000689500	AC018807.5.62712.86823	1	3285	3451	167 bp	ATTTTGTTCATTAAATATACAAATATTCTGTCTTCAATGGAGCTTGG GATAGATGAAAATAAAATTCACCACTGGAGAGCTAATGAGATTATAC ACACATTTTAAAGATGGAAACATCTTATCTTGTTTACGTCCTCATGAT GTGGCATTCTACTTGT
3	ENSE00000689502	AC018807.5.62712.86823	1	4477	4558	82 bp	TTACAGAGAAAAGTCAAAATATGTTGGTCAACCTTCAAGGGAGATGT GTGATSCAAACTACAGGAGGTGTTGCTG
4	ENSE00000689504	AC018807.5.62712.86823	1	14369	14505	137 bp	CACCCAGAACCAATAGTCTGGATCACTTCCAGTATTATTTAGCTCAAT ATTGGCTTATTATGGGATCACTTATGATGACATTAACAAATGCGAGT GCTCAGGAGCTGTCTGCATATGATCCAGAGCAAT
5	ENSE00000689506	AF178850.2.1.73384	1	7171	7354	184 bp	TCATTTCACTGGTGTGAAAGATCTTAACTGACGCTTGGAAAGCTTGG CACATTTTATTCAAAGCAGAGTCCCAAGTGTCTTCCAAATCAGCTGCG TTGATCTTTTTTCAAAGCAGCAAGCGTGTGTGTGTATGCAAGCTGG ASCAGAGAGAGATGTGACTGTGGACTGACAG
6	ENSE00000689509	AF178850.2.1.73384	1	9495	9593	99 bp	GATTGTGCGCTTATGGAGAAATGCTGTGATATTCCACATGTAGATT TAAAGCGGTTCAAACTGTGCTGAAGGACATGCTGCGAAAATGTCTA
7	ENSE00000689512	AF178850.2.1.73384	1	9703	9898	196 bp	TTTATGTCAAAGAAAGATGTGTAGGCTTCCCTTTGAGAGATGCGACCT CCTTGAATTTGCAATGGATCATCTGATCATGCTCCAGAAAACCTATG TTACAGCTGGGATCCGTGTGGACTGAAATCAATGGATCTGTATAGATGGA GTTTGTATGTTGGGATAAAAGATGTACAGACATTTGGCAAG
8	ENSE00000689515	AF178850.2.1.73384	1	15453	15558	106 bp	AAATAGATTTGGCCTTCAGAAATGTTATTTCTCACATTAATCAAGAGCT GATGTATCTGGAAACTGTGATTAAGTGTATCCAGAAATCACAGTGTGA AGCTGA
9	ENSE00000689518	AF178850.2.1.73384	1	20824	21007	184 bp	CAATCTGCACTGCGGAAATTAATATGTAATATGTAGGTAAATTTTAT TACAAATTCAGAGGCACTATATTATTTGCAAGATAAGTGGACATCTC TGCATCTGTGGAAATTTCCAGTGTACATGCAAGCAGCAAAAGATGTG GATAAAAGATGGAACTTCTGTGTTCAATAAG
10	ENSE00000689522	AF178850.2.1.73384	1	26991	27068	78 bp	GTTTCAGAAATCAAAGATGTGTGAGTCTTTCATACCTGGGTTATGATTG TACTACTGCAAAATGCAATGAGAGGTT
11	ENSE00000689527	AF178850.2.1.73384	1	27289	27427	139 bp	GTATSCAATAAAGAAAGCACTGTGCTGTAGTGTCTTATTTACCTCC AGATGCTCAGTCTCATCAGATCATGAGCTGGTGGAGATATTGACAGTG GCAATTTCCACCTGTAGCTATACAGCCAGACTCCTCG
12	ENSE00001040070	AF178850.2.1.73384	1	30108	30267	160 bp	AAAGGCTACATTTGAGACATTTACCTTCCAAACCAATGAGATGGCCA TTTTCTTATTCATCTCTCTTATTTATTTCTGTGTATGCTGATGAT AATGTTGAAATTAATTTCAAAGGAAAATAAGAGAACTGAGGACTATT CAAGGATGA
13	ENSE00001040056	AF178850.2.1.73384	1	31848	31878	31 bp	GCAACTGAAAGTGGAGTGAACCTAAAGGG

Figure 1.11

Splice Information			
Site	Exon ID	Splice Site	Exon ID
Display up to 3 base pairs either side of splice sites. <input type="button" value="Redraw"/>			
1	ENSE00000890080	...TGCCTGA-----TAPATTT...	ENSE00000689500
2	ENSE00000689500	...TTGTGTA-----AAGTTAC...	ENSE00000689502
3	ENSE00000689502	...TCTGTGA-----CAGCAC...	ENSE00000689504
4	ENSE00000689504	...CAATGTA-----TAGTCAT...	ENSE00000689506
5	ENSE00000689506	...ACAGGTT-----CAGGATT...	ENSE00000689509
6	ENSE00000689509	...TCTAGTA-----CAGTTTA...	ENSE00000689512
7	ENSE00000689512	...AAGGTA-----CAGAGAT...	ENSE00000689515
8	ENSE00000689515	...CTAGGTA-----TAGCAAT...	ENSE00000689518
9	ENSE00000689518	...TAAAGTA-----TAGGTTT...	ENSE00000689522
10	ENSE00000689522	...AGGTGTA-----TAGGAT...	ENSE00000689527
11	ENSE00000689527	...CTGTGTA-----TAGAAG...	ENSE00001040070
12	ENSE00001040070	...ATGAGTA-----CAGGCA...	ENSE00001040056

Intron sequence shown in red

Date: 2002-06-20 01:22:48 [Help Desk / Signpost Users](#)

Figure 1.12

Figure 1.13

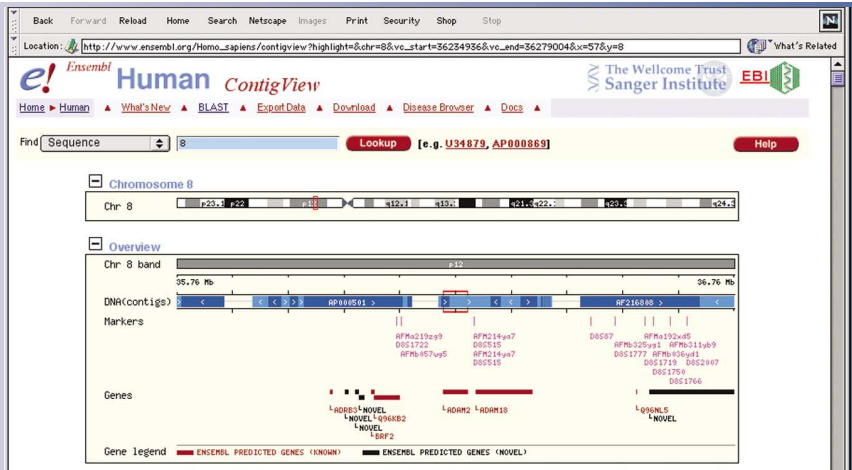


Figure 1.14

