

## 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/ng1189

This question serves as a basic introduction to the three major genome viewers. One gene, *ADAM23*, 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 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 *Map Viewer* to go to the Map Viewer home page. NCBI provides Map Viewers for 17 organisms, including mammals, other vertebrates, plants, protozoa, fungi, and invertebrates. Select an organism from the pull-down menu, enter a search term in the text box, and press *Go!* The search term can be any element mapped in that genome, which in human includes gene symbol, GenBank accession number, marker name or disease name. For this example, change the Organism to human, and enter 'ADAM23'.

The notation at the top of the resulting page (Figure 1.1) indicates that this is Build 31, or the NCBI's 31st assembly of the human genome. Build 31 is based on sequence data from 15 November 2002. The previous genome assembly, Build 30, was based on sequence data from 28 June 2002. This overview page shows a schematic of all of the human chromosomes, pinpointing the position of *ADAM23* to the q arm of chromosome 2 (Fig. 1.1). The hit on chromosome 8 is to *ADAM28*, which was at one time called *ADAM23*. The search results section shows that *ADAM23* 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 human maps are available at <http://www.ncbi.nlm.nih.gov/mapview/static/humansearch.html>. To get the most general overview of the genomic context of *ADAM23*, including all available maps, click on the item in the *Map element* column (in this case, *ADAM23*). This view shows *ADAM23* and a bit of flanking sequence on chromosome 2q33. By default, the maps are shown compressed. To display the maps in a wider format, remove the check mark next to *Compress Map* in the left blue sidebar (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 *ADAM23* and is created by aligning the *ADAM23* mRNA to the genome. The gene appears to have 27 exons. The vertical arrow next to the *ADAM23* 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 resources<sup>10</sup> (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.7164 to the genome follows that of *ADAM23*, 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. Many genes have been broadly mapped to this region of chromosome 2.

Clicking on the zoom control in the blue sidebar allows the user to zoom out to view a larger region of chromosome 2. Zooming out one level shows 1/100th of the chromosome. There are 22 genes in the region, but only 20 are labeled (displayed) in this view (Fig. 1.3). The region of *ADAM23* is highlighted in red on all maps. On the basis of the *Genes\_seq* map, *ADAM23* is located between *KIAA1571* and *LOC151405*.

### University of California, Santa Cruz Genome Browser

The home page for the UCSC Genome Browser is <http://genome.ucsc.edu/>. UCSC provides browsers not only for the most recent version of the rat, 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 *Dec. 2001* browser displays annotations based on NCBI's build 28 of the human genome, the *Apr. 2002* browser displays annotations on NCBI's build 29, the *June 2002* browser displays annotations of NCBI's build 30, and the *Nov. 2002* browser displays annotations on NCBI's build 31. Select *Human Nov. 2002* 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 'ADAM23' in the box labeled *position* and then click *Submit*. The results of this search are presented in two categories, *RefSeq Genes* and *mRNA Associated Search Results* (Fig. 1.5). The section marked *RefSeq 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 *RefSeq Genes* link for ADAM23 (arrow, Fig. 1.5) to see the genomic context of the ADAM23 mRNA Reference Sequence (NM\_003812).

The resulting zoomed-in view shows a region of chromosome 2 from base pair 206032982 to 206207297, located within 2q33.3 (Fig. 1.6). The blue track entitled *Known Genes based on SWISS-PROT, TrEMBL, mRNA, and RefSeq* shows the intron-exon structure of known genes. The vertical boxes indicate exons and the horizontal lines introns. The ADAM23 gene seems to have 26 exons. The direction of transcription is indicated by the arrowheads on the introns. The tracks labeled *Ensembl Genes*, *Acembly Genes*, *Twinscan*, *SGP Genes*, and *Genscan Genes* 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*, and *Nonhuman mRNAs* from GenBank tracks. Translated alignments of *Fugu rubripes* genomic sequence are in the *Fugu BLAT* tracks. The *Mouse Cons* and *Best Mouse* tracks shows conservation between the human and mouse genomes. Tracks displaying single-nucleotide polymorphisms (SNPs) and repetitive elements 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 ADAM23, zoom out 3x by clicking on the *zoom out 3x* box in the upper right corner. ADAM23 is located between AF338192 and BC033509 (Fig. 1.7).

### Ensembl

The Ensembl<sup>7</sup> project, <http://www.ensembl.org/>, provides genome browsers for nine species: human, mouse, rat, zebrafish, fugu, mosquito, fruitfly, *C. elegans*, and *C. briggsae*. Click on *Human* to view the main entry point for the human genome. The current version of human Ensembl is version 11.31.1, based on the NCBI's 31st build of the genome. To perform a text search, enter 'ADAM23' in the text box, and limit the search by selecting *Gene* from the pull-down search. Click on the upper button labeled *Lookup*. As at NCBI, two results are returned, the first with a link to the ADAM28 gene, and the second with a link to the ADAM23 gene (Fig. 1.8).

Click on either of the ADAM23 links (Ensembl Gene ENSG00000114948) to retrieve the GeneView window. The returned page contains two sections of data. The Ensembl Gene Report (Fig. 1.9) is an overview of ADAM23, including a link to the genomic location of the gene, a schematic of the intron/exon structure, and links to homologous genes from other organisms. Some of these fields will be described in more detail in later examples. The *Transcripts/Translation Summary* provides information on the gene transcript (Fig. 1.10). This section of the GeneView shows links to ADAM23 in other databases, as well as

protein domain information. If more than one transcript is predicted for the gene, each is allocated its own summary section.

The complete genomic context of ADAM23 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. 1.11) 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 2q33.3. 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 known genes annotated by Ensembl as being around ADAM23 are Q9BZ60 and NM\_014929.

The middle panel of the ContigView, the Detailed View (Fig. 1.12), 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. 1.12 are the defaults. The DNA (contigs) map separates items on the forward strand (above) from those on the reverse (below). The forward strand shows seven types of features. Starting at the bottom, the ADAM23 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 and was predicted by the GENSCAN gene prediction program<sup>11</sup> (see Question 7). The *Proteins* and *Human proteins* boxes indicate protein sequences that align to this version of the genome, whereas *Human cDNAs* shows mRNA sequences in the EMBL nucleotide sequence database and NCBI RefSeqs. Positioning the computer mouse over any feature brings up the feature's name and links to more detailed information. The only features on the reverse strand in this view are portions of an EST transcript and a Genscan transcript. The Basepair view, at the bottom of the ContigView (Figure 1.13) shows a very fine view of a 101 nucleotide region of ADAM23, showing the actual nucleotide and protein sequence, as well as restriction enzyme sites.

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. All three sites are currently displaying annotations based on NCBI's build 31. However, it takes significant time to update an annotation based on a new genome assembly, so shortly after the release of a new assembly, the sites may display different versions. At present, UCSC is the only site to maintain browsers based on older assemblies. Nevertheless, 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

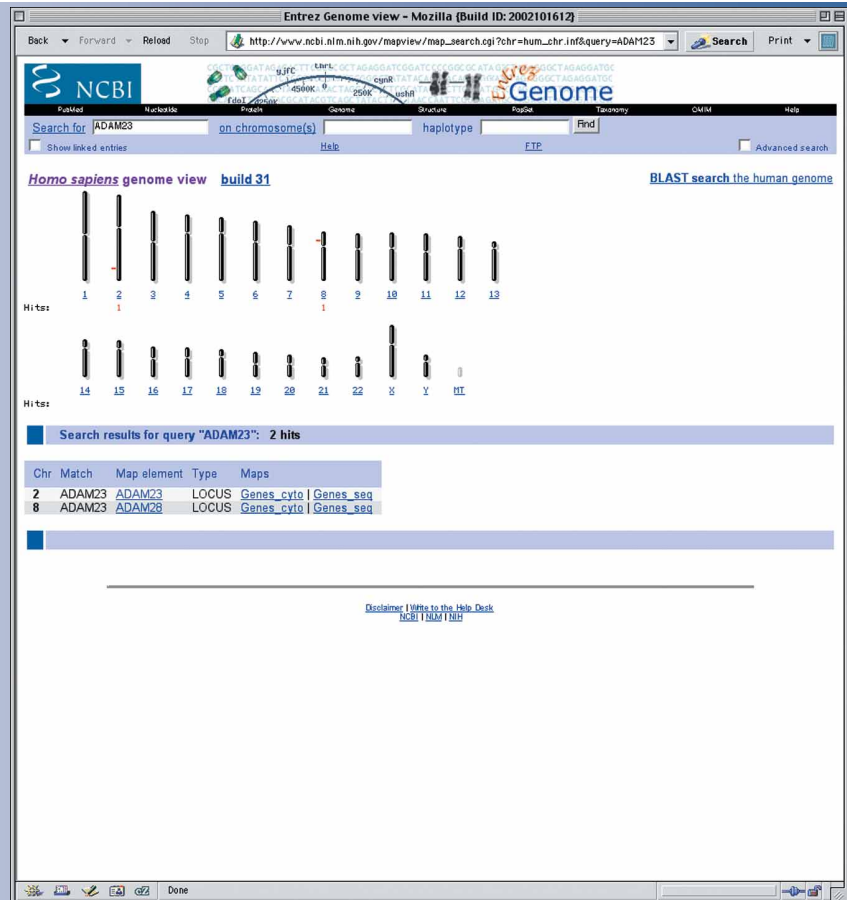
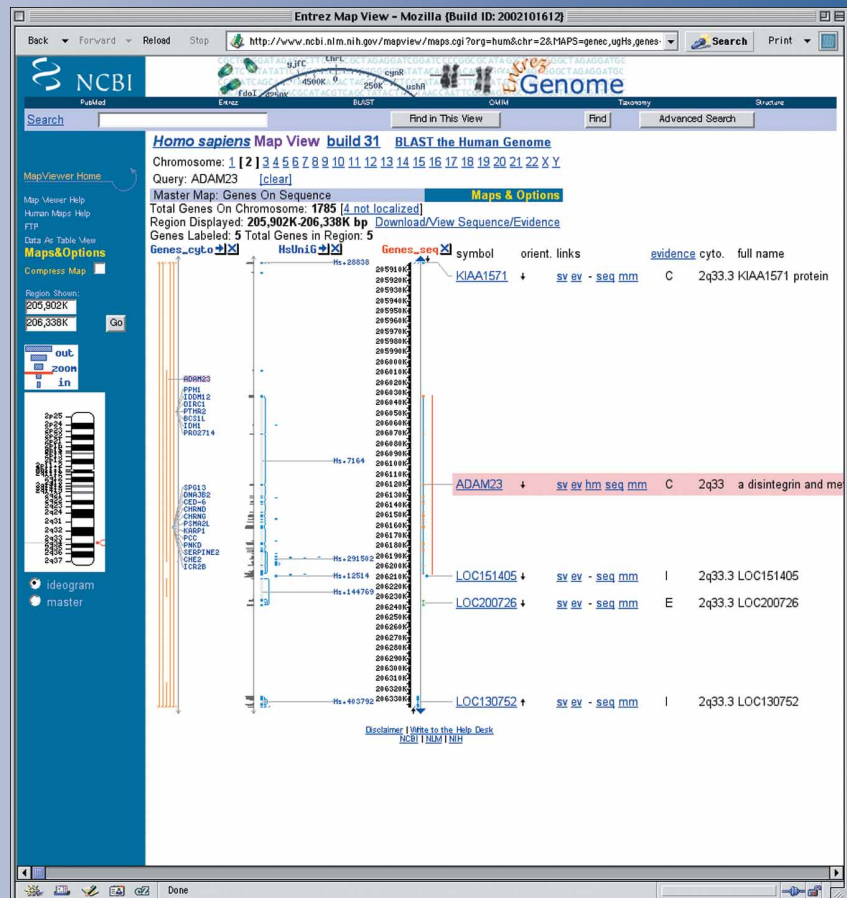


Figure 1.2



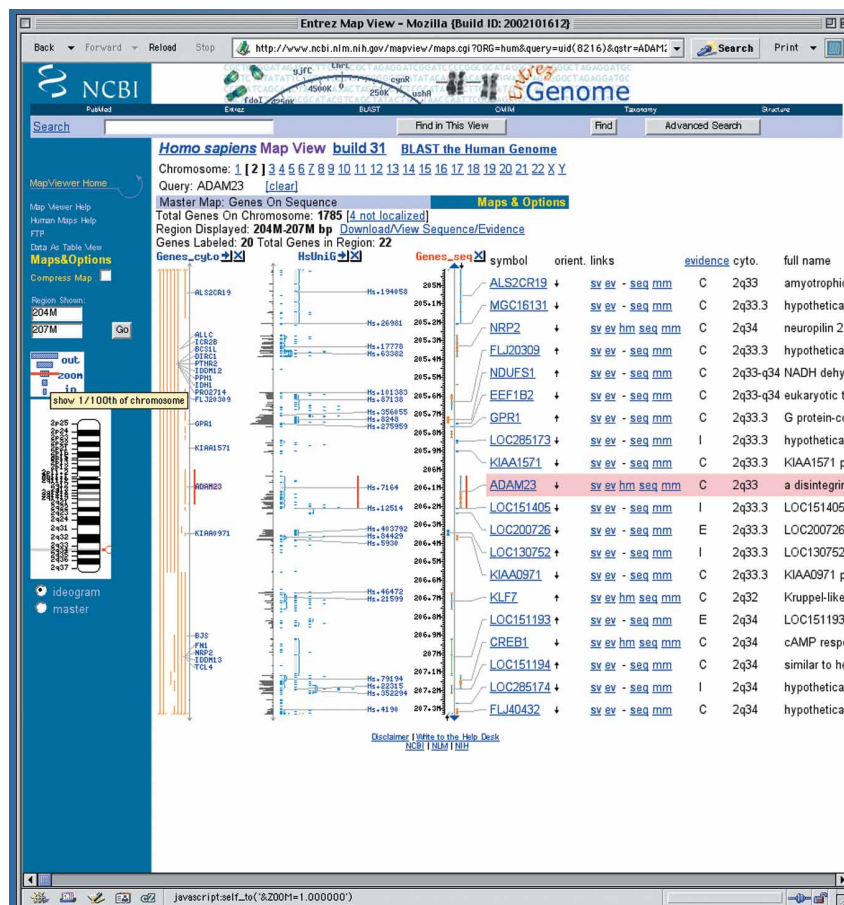


Figure 1.3

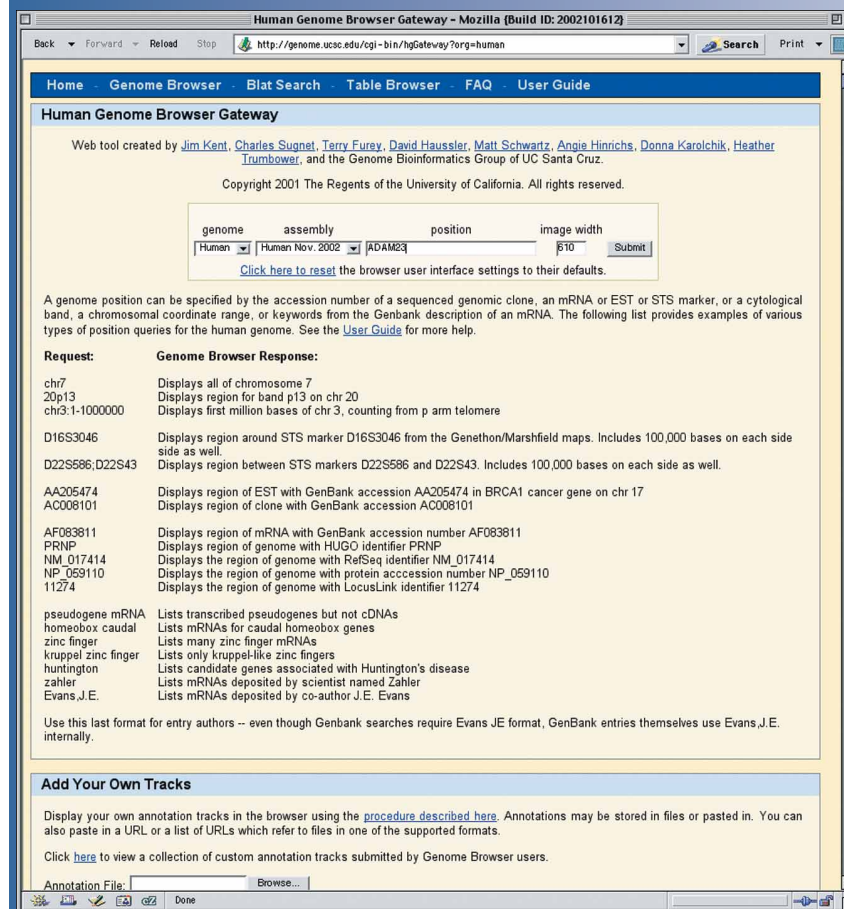


Figure 1.4



**Figure 1.6**

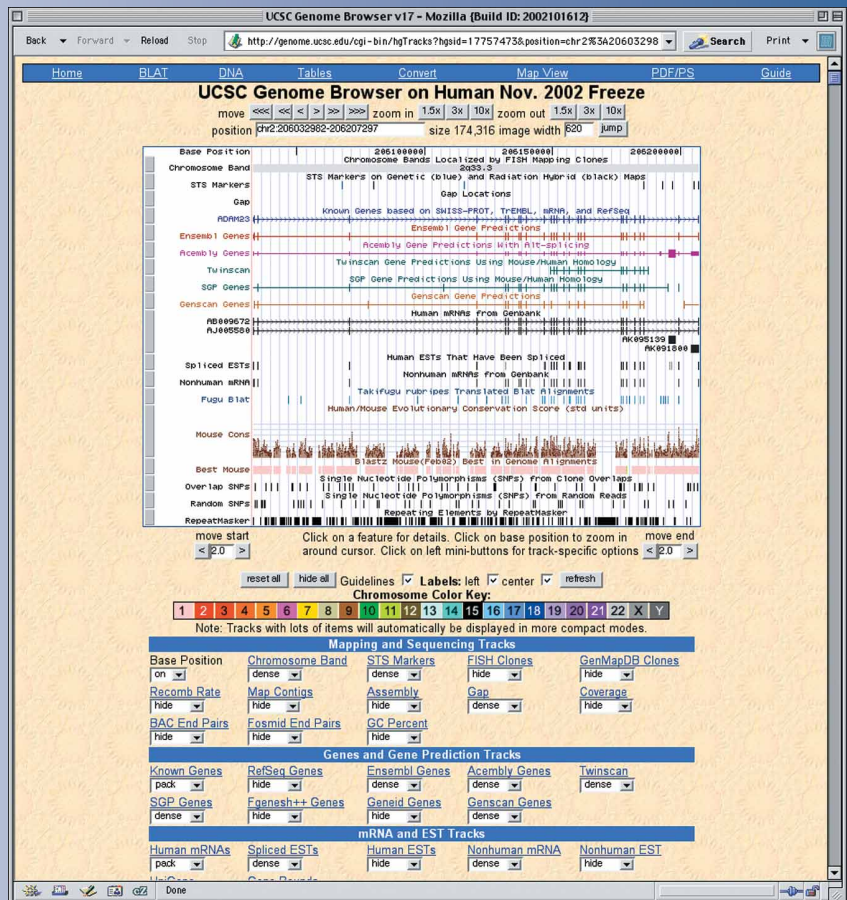






Figure 1.9

Human Genome Browser (GeneView) - Mozilla (Build ID: 2002101612)

Back Forward Reload Stop [http://www.ensembl.org/Homo\\_sapiens/geneview?gene=ENSG00000114948](http://www.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000114948) Search Print

Ensembl Human GeneView The Wellcome Trust Sanger Institute

Home Human What's New BLAST SSAHA Ensembl Export Data Download Disease Browser Docs

Find Gene  Lookup  Help

### Ensembl Gene Report

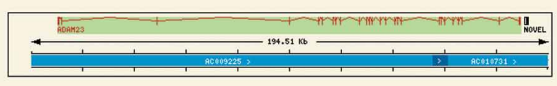
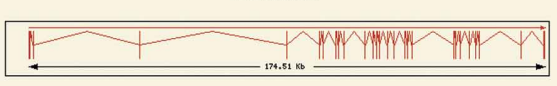
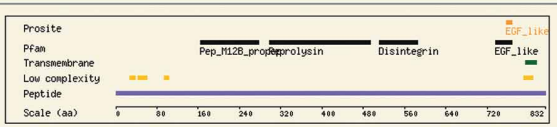
Gene	ADAM23 (HUGO ID)
Ensembl Gene ID	ENSG00000114948
Genomic Location	View gene in genomic location: 206032988 - 206207494 bp (206.0 Mb) on chromosome 2 This gene is located in sequence: <a href="#">AC079136.4.1.6026</a>
Description	A DISINTEGRIN AND METALLOPROTEINASE DOMAIN 23 PREPROPROTEIN. [Source: RefSeq (NM_003812)]
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise model from a human/vertebrate protein, a set of aligned human cDNAs followed by GenomeWise for DRF prediction or from GenScan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.
Predicted Transcripts	1: <a href="#">ADAM23</a> (ENST00000264377) <a href="#">View transcript info</a> <a href="#">View exon info</a> <a href="#">View protein info</a> 
Homology Matches	These gene(s) have been identified as putative homologues by reciprocal BLAST analysis: <i>Mus musculus</i> <a href="#">ENSMUSG00000029864</a> (Adam23) DISINTEGRIN AND METALLOPROTEINASE DOMAIN 23. [Source: RefSeq (Acc.NM_011780)] <i>Rattus norvegicus</i> <a href="#">ENSRNOG00000012424</a> (Novel Ensembl prediction) ADAM PRECURSOR DISINTEGRIN AND METALLOPROTEINASE DOMAIN <i>Rattus norvegicus</i> <a href="#">ENSRNOG00000012617</a> (Novel Ensembl prediction) ADAM PRECURSOR DISINTEGRIN AND METALLOPROTEINASE DOMAIN <i>Fugu rubripes</i> <a href="#">SINFRUG00000124740</a> (ADAM23) ADAM PRECURSOR DISINTEGRIN AND METALLOPROTEINASE DOMAIN
Export Data	<a href="#">Export gene data in EMBL, GenBank or FASTA</a>

Figure 1.10

### Transcripts/Translation Summary

ADAM23	Stable ID: ENST00000264377 Exons: 26 Transcript length: 3250 bp Translation length: 832 residues <a href="#">View transcript information</a> <a href="#">View exon information</a> <a href="#">View protein information</a>
Similarity Matches	This Ensembl entry corresponds to the following database identifiers: Affymx Microarray U133: HG-U133A.206046_at HG-U133B.244463_at Affymx Microarray U95: HG-U95A.40350_at HG-U95B.46908_at HG-U95E.73040_at EMBL: <a href="#">AB000672</a> [man] <a href="#">A1005580</a> [man] HUGO: <a href="#">ADAM23</a> [man] LocusLink: <a href="#">6745</a> [man] MIM: <a href="#">603710</a> Protein ID: <a href="#">P100190</a> [man] <a href="#">P100190</a> [man] RefSeq: <a href="#">NM_003812</a> (Target %id: 100; Query %id: 100) SpTfEMBL: <a href="#">075077</a> (Target %id: 100; Query %id: 100)
GO	The following GO terms have been mapped to this entry via Swissprot/SpTfEMBL: GO:0004222 (metalloprotease) GO:0004222 (metalloprotease) GO:0004222 (metalloprotease) GO:0004222 (metalloprotease) GO:0004222 (metalloprotease) GO:0004222 (metalloprotease) GO:0004222 (metalloprotease) GO:0004222 (metalloprotease)
InterPro	<a href="#">IPR001990</a> Metalloprotease ADAM/epsolin M12B - <a href="#">View other Ensembl genes with this domain</a> <a href="#">IPR001990</a> Disintegrin - <a href="#">View other Ensembl genes with this domain</a> <a href="#">IPR001990</a> Metalloprotease M12B - <a href="#">View other Ensembl genes with this domain</a> <a href="#">IPR006209</a> EGF-like domain - <a href="#">View other Ensembl genes with this domain</a>
Protein Family	<a href="#">ENSG00000000100</a> : ADAM PRECURSOR A DISINTEGRIN AND METALLOPROTEINASE DOMAIN This cluster contains 24 Ensembl gene member(s)
Transcript Structure	Exon structure 
Protein Structure	Prosite Pfam Transmembrane Low complexity Peptide Scale (aa) 0 80 160 240 320 400 480 560 640 720 800 

Date: 2003-03-20 22:16:44 [Help Desk / Suggestions](#)

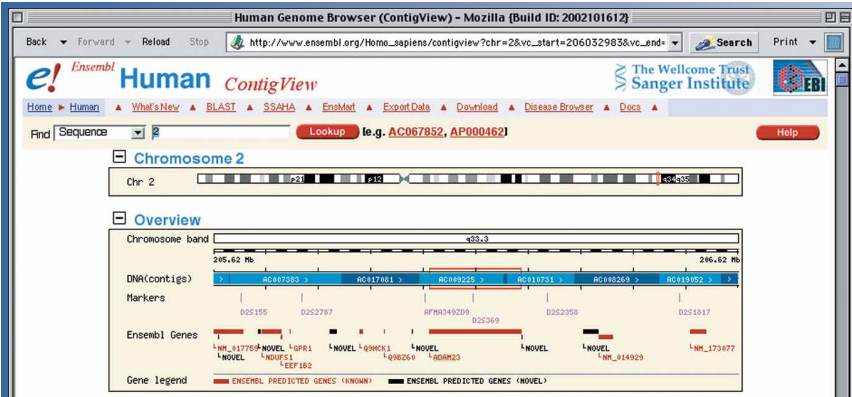


Figure 1.11

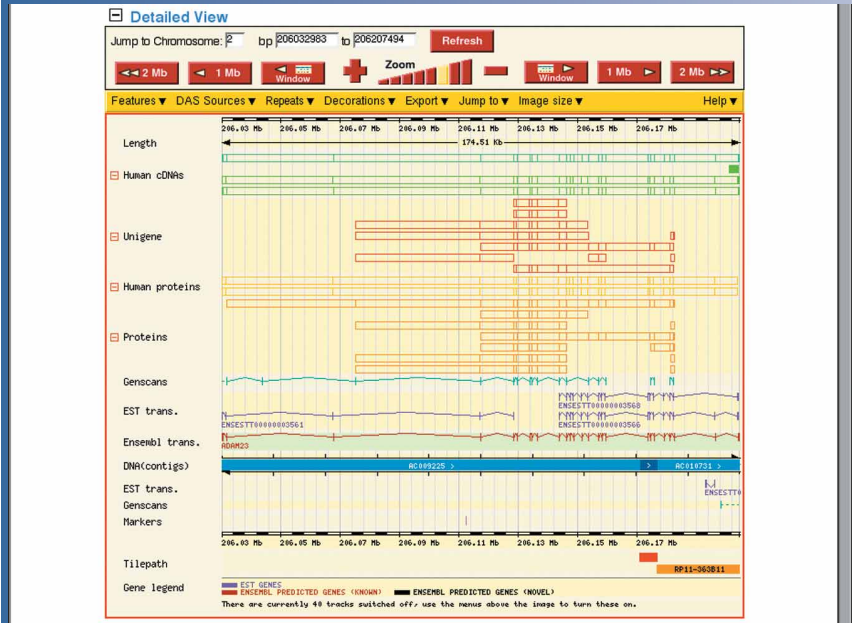


Figure 1.12



Figure 1.13

