

## Mining Genomic Sequence Data

Tyra G. Wolfsberg, Ph.D.  
NHGRI

*Current Topics in Genome Analysis*  
September 16, 2003

### Accessing the public genome sequence data

UCSC's Genome Browser ("Golden Path")  
<http://genome.ucsc.edu>

NCBI's Map Viewer  
<http://www.ncbi.nlm.nih.gov/mapview/>

Ensembl  
<http://www.ensembl.org>

## Types of data integrated in genome browsers

- Genomic sequence
- RefSeq mRNAs (non-redundant)
- GenBank mRNAs (redundant)
- ESTs
- Gene predictions
- SNPs
- Homologous sequences from other organisms
- STSs

## NCBI Reference Sequences

Accession Format

RefSeq accession numbers can be distinguished from GenBank accessions by their distinct format of [2 characters]underbar[6 digits]. For example, a RefSeq protein accession is [NP\\_015325](#).

Accession	Molecule	Method @	Note
NC_123456	Genomic	Curation	Complete genomic molecules including genomes, chromosomes, organelles, plasmids.
NG_123456	Genomic	Curation	Incomplete genomic region; primarily supplied for <i>Homo sapiens</i> and <i>Mus musculus</i> to support the NCBI Genome Annotation pipeline.
<b>NM_123456</b>	mRNA	Curation	
NR_123456	RNA	Curation	Non-coding transcripts including structural RNAs, transcribed pseudogenes, and others
<b>NP_123456</b>	Protein	Curation	
<b>NT_123456</b>	Genomic	Automated	Intermediate genomic assemblies of BAC sequence data
NW_123456	Genomic	Automated	Intermediate genomic assemblies of Whole Genome Shotgun sequence data
XM_123456	mRNA	Automated	<i>Homo sapiens</i> model mRNA provided by the Genome Annotation process; sequence corresponds to the genomic contig.
XR_123456	RNA	Automated	<i>Homo sapiens</i> model non-coding transcripts provided by the Genome Annotation process; sequence corresponds to the genomic contig.
XP_123456	Protein	Automated	<i>Homo sapiens</i> model proteins provided by the Genome Annotation process; sequence corresponds to the genomic contig.
NZ_ABCD12345678	Genomic	Automated	An ordered collection of whole genome shotgun sequence data for incomplete bacterial genomes. Accessions are not tracked between releases. The first four characters following the underscore (e.g. 'ABCD') identifies a genome project.
ZP_12345678	Genomic	Automated	Proteins, annotated on NZ_ accessions.

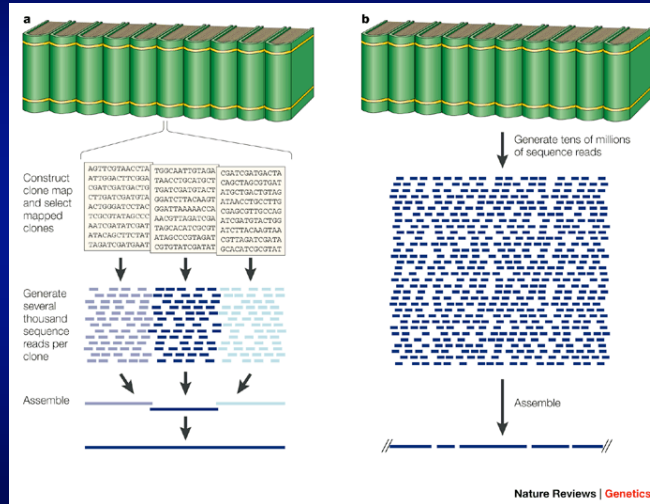
@ Method:  
*Curation*: indicates the process flow includes expert review for some of the records; analysis may be provided either by NCBI staff or collaborators.  
*Automated*: indicates records that are not individually reviewed; updates are released in bulk for a genome.

<http://www.ncbi.nlm.nih.gov/RefSeq/key.html>

## Overview of sequencing strategies

Clone-by-clone shotgun sequencing

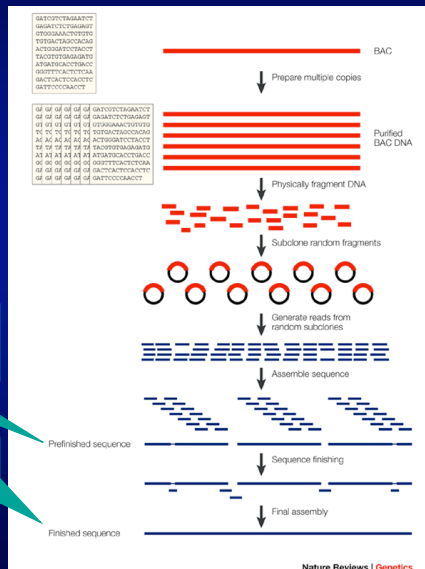
Whole-genome shotgun sequencing



Nature Reviews | Genetics

Green ED. Strategies for the systematic sequencing of complex genomes. Nat Rev Genet. 2001. 2:573-83.

## Clone-by-clone shotgun sequencing

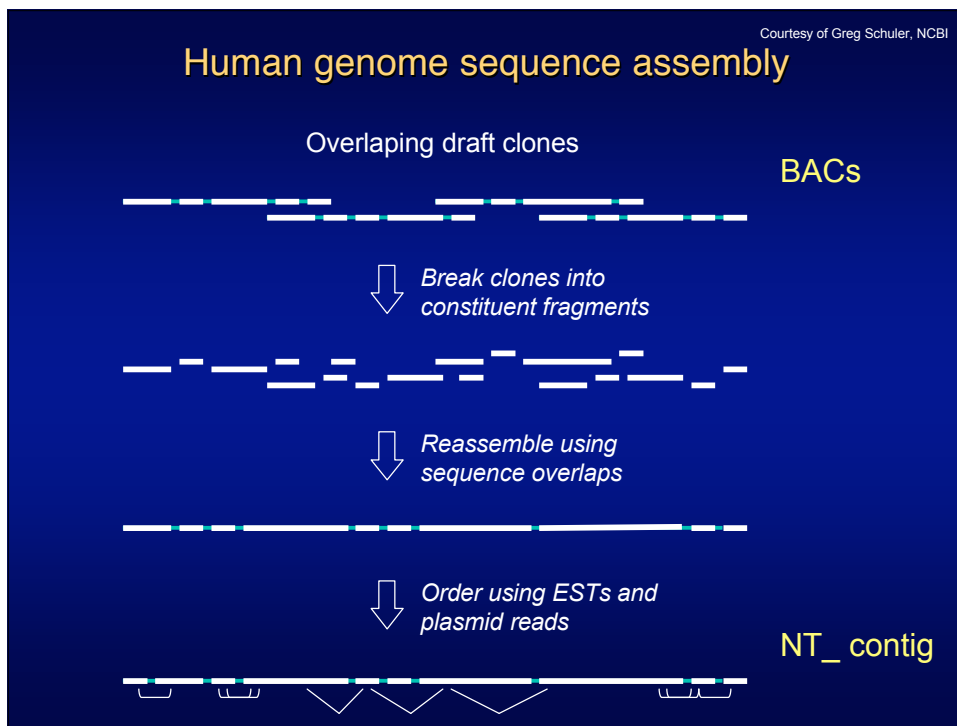


“working draft”  
Phase 0, 1, or 2  
BAC

Finished  
Phase 3 BAC

Nature Reviews | Genetics

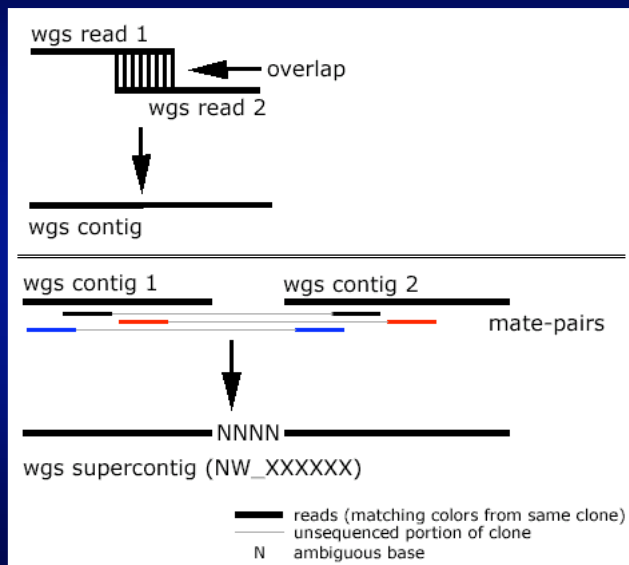
Green ED. Strategies for the systematic sequencing of complex genomes. Nat Rev Genet. 2001. 2:573-83.



### Status of the human genome sequence

- All chromosomes are now considered finished
- Build 33; April 2003
  - <400 gaps, averaging <100 Kb, representing DNA regions with unusual structures that can't be reliably sequenced
    - 138 unplaced contigs each with sequence from a single clone
    - Assembly will be updated as gaps are closed
- Build 34; July 2003
  - 11 Mb (~0.4%) more finished nucleotides than build 33
  - Covers ~99% of gene-containing regions in the genome
- NCBI and Ensembl currently display build 33; UCSC features a partially annotated build 34, as well as older assemblies
- UCSC is usually the first to display new assemblies, followed by NCBI and then Ensembl.

## Mouse and rat whole genome shotgun (WGS) assemblies



<http://www.ncbi.nlm.nih.gov/genome/seq/NCBIContigInfo.html>

## Mouse genome sequencing

- Whole genome shotgun sequence (WGS) is now completed (7x coverage)
- "MGSC Version 3" is the current assembly of the WGS
- Sequence will be finished by sequencing individual BACs and incorporating WGS
- NCBI, UCSC, and Ensembl provide browsers based on an assembly that combines MGSCv3 with finished BAC sequence (called build 30 at NCBI and Ensembl, Feb 2003 at UCSC)

## Rat genome sequencing

- Draft genome assembly produced by the Rat Genome Sequencing Consortium
- Hybrid approach combined clone by clone and whole genome shotgun methods
- Assembly covers more than 90% of the genome
- UCSC displays v. 3.1 (June 2003); not clear what assembly is shown by NCBI, or whether Ensembl shows v. 2.0 or 2.1

## Where's the genome sequence at NCBI?

<u>Status</u>	<u>BLAST database</u>	<u>Definition</u>
Phase 0 HTG	htgs	single-few pass reads of a single clone (not contigs)
Phase 1 HTG	htgs	Unfinished, may be unordered, unoriented contigs, with gaps
Phase 2 HTG	htgs	Unfinished, ordered, oriented contigs, with or without gaps
Phase 3 HTG	nr	Finished, no gaps (with or without annotations)
NT_contigs	human genome BLAST -->genome	assembled BACs from HTG phase 1-3
WGS	trace archive; mouse or rat genome BLAST -->WGS Traces	whole genome shotgun sequences
NW_contigs	rat genome BLAST -->genome	assembled WGS sequence (also called supercontigs or scaffolds)
WGS/HTG assembly	mouse genome BLAST -->genome	assembled WGS sequence, along with finished HTGs

## Accessing the public genome sequence data

UCSC's Genome Browser ("Golden Path")

<http://genome.ucsc.edu>

NCBI's Map Viewer

<http://www.ncbi.nlm.nih.gov/mapview/>

Ensembl

<http://www.ensembl.org>

## UCSC

### Genomes

Human  
Mouse  
Rat  
C. elegans  
C. brigssae  
SARS  
Zoo

### Human Assemblies

Jun 2002: NCBI  
build 30  
Nov 2002: NCBI  
build 31  
Apr 2003: NCBI  
build 33  
July 2002: NCBI  
build 34

Human Genome Browser Gateway - Mozilla

Back Forward Reload Stop <http://genome.ucsc.edu/cgi-bin/hgGateway?org=Human&db=hg15&hpid=255161> Search Print

Home Genome Browser Blat Search Table Browser FAQ User Guide

### Human Genome Browser Gateway

UCSC Genome Browser created by Jim Kent, Charles Eassey, Tom Foy, Richard Beardsall, Heather Tomblow, Anjali Mishra, Matt Schwartz, Fan He, Hiram Clawson, Eda Eisenbach, Brian Zhang, Robert Harte, Emma Partridge, Tandi Slavik, and the Genome Bioinformatics Group of UC Santa Cruz. Software Copyright (c) The Regents of the University of California. All rights reserved.

genome	assembly	position	image width
Human	Apr 2002	hg15	100
	July 2002		
	Nov 2002		
	June 2002		

[Click here](#) browser user interface settings to their defaults.  
[Add Your Own Tracks](#)

#### About the Homo sapiens assembly

The latest human reference sequence (UCSC version hg16) is based on NCBI Build 34 and was produced by the International Human Genome Sequencing Consortium. The sequence covers about 99 percent of the gene-containing regions in the genome, and has been sequenced to an accuracy of 99.99 percent. Of note in this release is the addition of the pseudoautosomal regions of the Y chromosome. This sequence was taken from the corresponding regions in the X chromosome and is an exact duplication of that sequence.

There are 2,943,433,602 finished sequenced bases in the ordered and oriented portion of the assembly, which is an increase of 0.4 percent, or approximately 11 Mb, over the Build 33 assembly. The reference sequence is considered to be "finished", a technical term indicating that the sequence is highly accurate (with fewer than one error per 10,000 bases) and highly contiguous (with the only remaining gaps corresponding to regions whose sequence cannot be reliably resolved with current technology). Future work on the reference sequence will focus on improving accuracy and reducing gaps in the sequence.

Some sequence joins between adjacent clones in this assembly could not be computationally validated because the clones originated from different haplotypes and contained polymorphisms in the overlapping sequence, or the overlap was too small to be reliable. In these instances, the sequencing center responsible for the particular chromosome has provided data to support the join in the form of an electronic certificate. These certificates may be reviewed through the link below.

#### Statistical information

- [Non-Standard Join Certificates](#)
- [Summary Statistics](#)
- [Chromosome Reports](#)
- [Genome Map Plots](#)

#### Sample position queries

A genome position can be specified by the accession number of a sequenced genomic clone, an mRNA or EST or STS marker, or a cytological band, a chromosomal coordinate range, or keywords from the Genbank description of an mRNA. The following list provides examples of various types of position queries for the human genome. See the [User Guide](#) for more help.

Request:	Genome Browser Response:
chr7	Displays all of chromosome 7
20p13	Displays region for band p13 on chr 20
chr3:1-1000000	Displays first million bases of chr 3, counting from p arm telomere
scf1:1-1000000	Displays first million bases of scaffold 1 of an unmapped genome assembly
D16S3046	Displays region around STS marker D16S3046 from the Genethon/Marshfield maps. Includes 100,000 bases on each side as well.

## UCSC: Results for ADAM2 query

**Known Genes**

ADAM20 at chr14:68979118-68981543 - (AF029899) a disintegrin and metalloproteinase domain 20  
 ADAM21 at chr14:68914800-68916644 - (AF029900) a disintegrin and metalloproteinase domain 21  
 ADAM22 at chr7:87161961-87424669 - (AF073291) a disintegrin and metalloproteinase domain 22  
 ADAM29 at chr4:176436583-176483706 - (AF134708) a disintegrin and metalloproteinase domain 29  
 ADAM23 at chr21:207277951-207446902 - (A3009580) a disintegrin and metalloproteinase domain 23  
 ADAM27 at chr8:39183088-39328434 - (AJ133004) a disintegrin and metalloproteinase domain 18  
 ADAM28 at chr8:23972184-24033174 - (AJ242015) a disintegrin and metalloproteinase domain 28  
 ADAM2 at chr8:39342200-39436675 - (BC034957) a disintegrin and metalloproteinase domain 2 (fertilin beta)

**RefSeq Genes**

ADAM29 at chr4:176481126-176706678 - (NM\_021780) a disintegrin and metalloproteinase domain 29  
 ADAM29 at chr4:176481126-176483589 - (NM\_021779) a disintegrin and metalloproteinase domain 29  
 ADAM28 at chr8:23972192-24032878 - (NM\_021778) a disintegrin and metalloproteinase domain 28  
 ADAM28 at chr8:23972192-24014179 - (NM\_021777) a disintegrin and metalloproteinase domain 28  
 ADAM22 at chr7:87161928-87424674 - (NM\_021723) a disintegrin and metalloproteinase domain 22  
 ADAM22 at chr7:87161928-87424674 - (NM\_021722) a disintegrin and metalloproteinase domain 22  
 ADAM22 at chr7:87161928-87409655 - (NM\_021721) a disintegrin and metalloproteinase domain 22  
 ADAM20 at chr14:68979115-68981693 - (NM\_003814) a disintegrin and metalloproteinase domain 20  
 ADAM2 at chr8:39342203-39436722 - (NM\_001464) a disintegrin and metalloproteinase domain 2  
 ADAM29 at chr4:176481126-176483589 - (NM\_014269) a disintegrin and metalloproteinase domain 29  
 ADAM22 at chr7:87161928-87424674 - (NM\_016351) a disintegrin and metalloproteinase domain 22  
 ADAM28 at chr8:23972184-24033171 - (NM\_014265) a disintegrin and metalloproteinase domain 28  
 ADAM22 at chr7:87161928-87409655 - (NM\_004194) a disintegrin and metalloproteinase domain 22  
 ADAM21 at chr14:68914257-68916663 - (NM\_003813) a disintegrin and metalloproteinase domain 21  
 ADAM23 at chr21:207272390-207446700 - (NM\_003812) a disintegrin and metalloproteinase domain 23

**Human Aligned mRNA Search Results**

AJ005380 - Homo sapiens mRNA for adam23 protein.  
 AF158937 - Homo sapiens metalloproteinase-disintegrin ADAM22-3 (ADAM22) mRNA, alternatively spliced, partial cds.  
 AF134708 - Homo sapiens disintegrin and metalloproteinase domain 29 (ADAM29) mRNA, complete cds.  
 AF171929 - Homo sapiens metalloproteinase-disintegrin (ADAM29) mRNA, complete cds.  
 AF171930 - Homo sapiens metalloproteinase-disintegrin beta (ADAM29) mRNA, alternatively spliced, complete cds.  
 AF171931 - Homo sapiens metalloproteinase-disintegrin gamma (ADAM29) mRNA, alternatively spliced, complete cds.  
 AF155381 - Homo sapiens metalloproteinase-like, disintegrin-like, cysteine-rich protein 2 delta (ADAM22) mRNA, complete cds.  
 AF155382 - Homo sapiens metalloproteinase-like, disintegrin-like, cysteine-rich protein 2 epsilon (ADAM22) mRNA, complete cds.

## UCSC: Default view of ADAM2

Human chr8:39342203-39436722 - UCSC Genome Browser v31 - Mozilla

UCSC Genome Browser on Human April 2003 Freeze

position chr8:39342203-39436722 size 94,520 image width 810

Base Position: 39436675

Chromosome Band: chr8:39436675-39436675

STS Markers: chr8:39436675-39436675

Gene: ADAM2

RefSeq Genes: ADAM2

Ensembl Gene Predictions: ADAM2

Accessory Genes: ADAM2

GenScan Genes: ADAM2

Human ESTs: chr8:39436675-39436675

Human/Mouse/Rat: chr8:39436675-39436675

Chromosome Color Key: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 X Y

Note: Tracks with lots of items will automatically be displayed in more compact modes.

Mapping and Sequencing Tracks

Base Position: chr8:39436675-39436675

Chromosome Band: chr8:39436675-39436675

STS Markers: chr8:39436675-39436675

FISH Clones: chr8:39436675-39436675

Recomb Rate: chr8:39436675-39436675

Map Contigs: chr8:39436675-39436675

Assembly: chr8:39436675-39436675

Coverage: chr8:39436675-39436675

RAC End Pairs: chr8:39436675-39436675

Formid End Pairs: chr8:39436675-39436675

GC Percent: chr8:39436675-39436675

Genes and Gene Prediction Tracks

Known Genes: chr8:39436675-39436675

RefSeq Genes: chr8:39436675-39436675

MGCG Genes: chr8:39436675-39436675

Ensembl Genes: chr8:39436675-39436675

Assembly Genes: chr8:39436675-39436675

Twinscan: chr8:39436675-39436675

SGP Genes: chr8:39436675-39436675

Fgenesh++ Genes: chr8:39436675-39436675

Geneid Genes: chr8:39436675-39436675

GenScan Genes: chr8:39436675-39436675

FASTA: chr8:39436675-39436675

mRNA and EST Tracks

Human mRNAs: chr8:39436675-39436675

Soliced ESTs: chr8:39436675-39436675

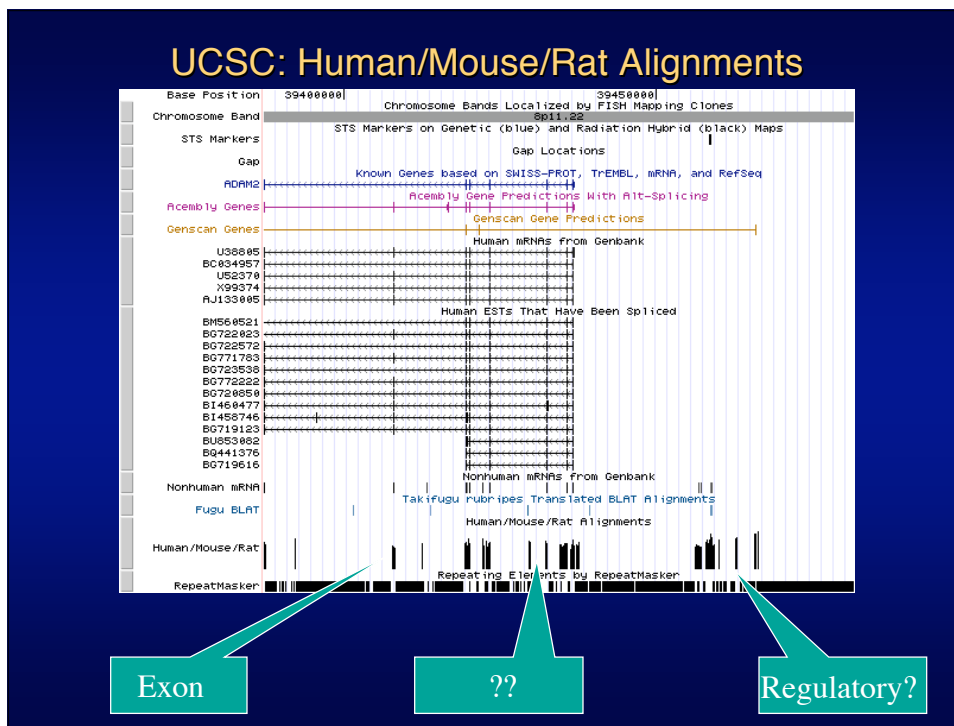
Human ESTs: chr8:39436675-39436675

NonHuman mRNAs: chr8:39436675-39436675

NonHuman ESTs: chr8:39436675-39436675







**UCSC Genome Browser on Human April 2003 Freeze**

position chr8:39387099-39481618 size 94,520 image width 620

**Get DNA in Window**

**Get DNA for**

Position

Note: if you would prefer to get DNA for features of a particular track or table, try the [Table Browser](#) and select FASTA as the output format.

**Sequence Retrieval Region Options:**

Add  extra bases upstream (5') and  extra downstream (3')

**Sequence Formatting Options:**

All upper case.  
 All lower case.  
 Mask repeats:  to lower case  to N  
 Reverse complement (get 5' strand sequence)

Note: The "Mask repeats" option applies only to "Get DNA", not to "Extended case/color options"

Track Name	Toggle Case	Underline	Bold	Italic	Red	Green	Blue
STS Markers	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Known Genes	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Assembly Genes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Genscan Genes	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Human mRNAs	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Spliced ESTs	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Nonhuman mRNA	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Fugu BLAT	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0
Human/Mouse/Rat	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	255	0	0
RepeatMasker	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	0	0	0



## UCSC BLAT results

Human BLAT Results										
BLAT Search Results										
ACTIONS	QUERY	SCORE	START	END	QSIZE	IDENTITY	CHRO	STRAND	START	END
<a href="#">browser details</a>	unknown	664	4	767	814	98.1%	3	-	120544667	120559679

### Alignment of unknown and chr3:120544667-120559679

Click on links in the frames to left to navigate through alignment. Matching bases in cDNA and genomic sequences are colored blue and capitalized. Light blue bases mark the boundaries of gaps in either side of the alignment (often splice sites).

**cDNA unknown**

```

gggAAGTAA CAGAAGTTAG AAGGGGAAT GTCCGCCCTC TGAAGATTAC 50
CCAAAGAAA AGTAAATTG CATTCCTTA TAGACTTTAA GAGGAAACA 100
CTTCAGAAAT GGAGTCTTAC CTTGAATCA AAGGATTTAA AGAAAAGTG 150
GATTTTTTTC TACGACACTT GGTAAACTTA ATCCACAAAC TTGAGGAGCC 200
CAGGACACC CTCCAATCTC TGTGTGTTT GTAACATCA CTGGAGGGTC 250
TTCACGDTA GCAATTGGAT TGTACACAG CCTCCCCGTT TTGCACCTGG 300
GAGTSCCTT GGTCTTACTT GGTTCAAAT TGTGGTCTT GACTTTGAC 350
CCTAAGCATC TAAAGCCATG GGCACACAG GAGGCAAGG AACATACCA 400
TCCAGTCTC CAACTCTAAA TTCTTTTCA CTTCTGTATC TGGTGGTCT 450
TTCACATTC TTTTCAAGTG TTTTCCACT GACCAAGGA AGTAAAGAA 500
GTGGCAAGC CTGCTCTGTA GTCCAAAGT TTTTGTGTA AGAGCTGGCA 550
CAACTTCCA TCTACTGGC AAAGAGGAA GAAGATTTG CTTACTTADA 600
TGTCTGGGG ACATGAAAtt ttggGCCGA GTCCAGAAC CcACCATcC 650
TTTGTATtc cctaattac: CTCTcGATT GTGATCTTG ggttggegc 700
cccaattgg aqaagggga cactcaagt gggatgagt cctgRAGTA 750
TggaAAAAG ACSCTTTTt aagggggaa cccactggg ctgaaatgg 800
acottatoc tttc
                    
```

**Genomic chr3 (reverse strand):**

```

acaaaagaa ctagaagaag acggagcag aggggcttc tbaaaacc 120
ctgcacact cctggcctg acaagctgt agtaactca accoctgt 1203909u
AAGTACAG AAGTTAAG GGAATATC GCCTCTGA AGATTACCA 12055630
AAGAAAAT GATTTCAT TCTTTATG ACTGTAGAA GAGACATCT 120559580
CAGATDGA GTTTACTT GAATCAAG GTTTAAGA AAAGTGGAA 120559590
TTTTTTTCA GCAgtaat acatbtaata ttatobact atggatagc 120559480
ttgtagaa tagttaccag caactcact gotttttaa caaacatcc 120559430
ctgtatoc actctatga tcaagcttg cctcaaaag saccttag 120559380
                    
```

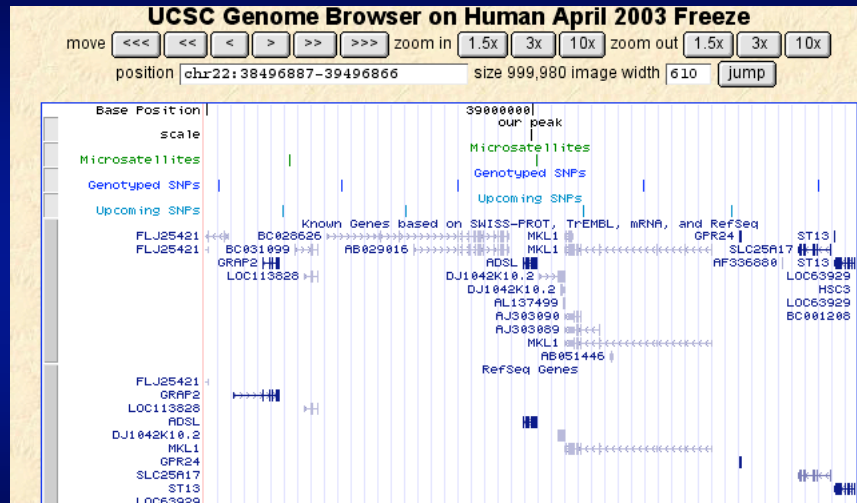
## "Add your own tracks" to UCSC

```

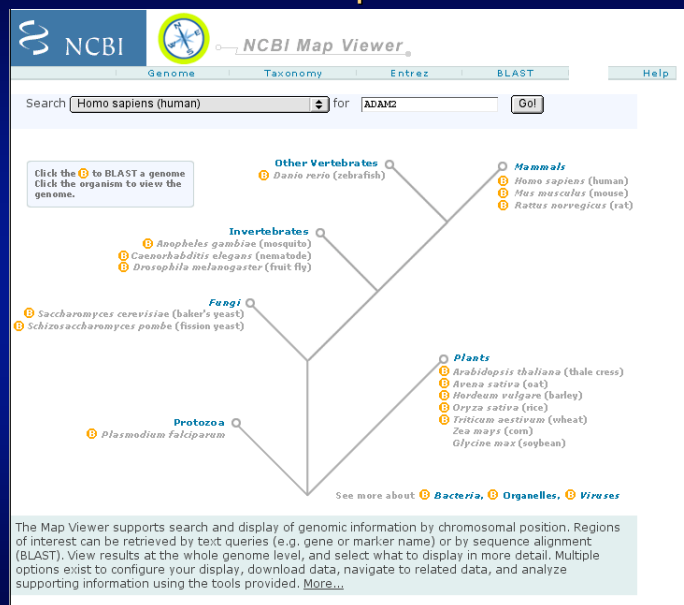
browser position chr22:38496887-39496866
browser hide cytoBand
browser hide stsMap
browser hide gap
browser hide clonePos
browser full refGene
browser dense mrna
track name="scale" description="our peak"
chr22 38996887 38996888 peak
track name="Microsatellites" description="Microsatellites" color=0,128,0
chr22 38627059 38627060 D22S276
chr22 39005417 39005418 D22S307
track name="Genotyped SNPs" description="Genotyped SNPs" color=0,0,255
chr22 38518342 38518343 ss146131
chr22 38705963 38705964 ss2941443
chr22 38884157 38884158 ss141110
chr22 39171390 39171391 ss22916
chr22 39438769 39438770 ss1479794
track name="Upcoming SNPs" description="Upcoming SNPs" color=0,128,192
chr22 38615712 38615713 ss86855
chr22 38804838 38804839 ss85533
chr22 39077895 39077896 ss141190
chr22 39305065 39305066 ss137027
                    
```

Nature Genetics User's Guide,  
Question 7

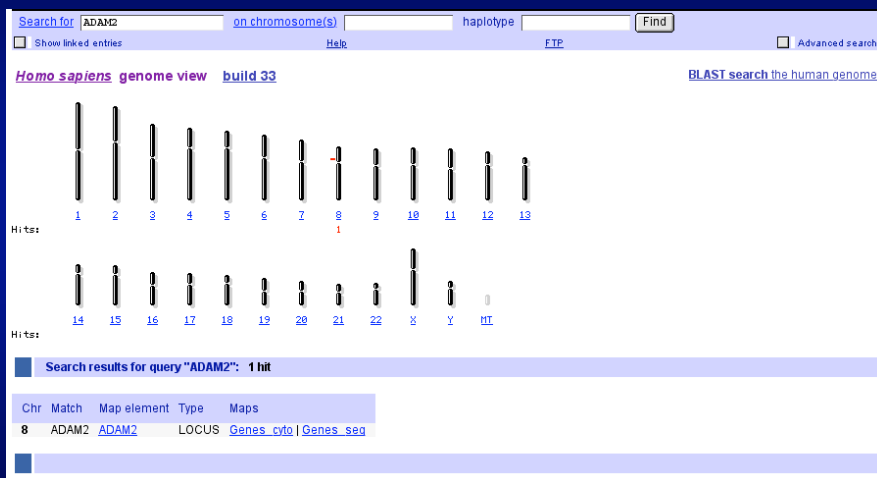
## “Add your own tracks” to UCSC



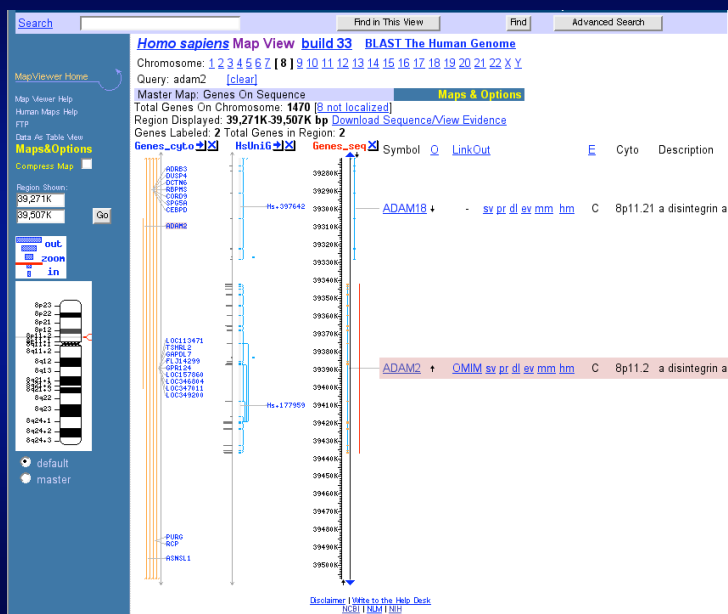
## NCBI Map Viewer



## NCBI Map Viewer search results



## NCBI default view of ADAM2



## NCBI download (dl)

Homo sapiens Genome (build 33)  
Region to retrieve (in chromosome coordinates):  
Chromosome:  Strand:   
from:  adjust by:   
to:  adjust by:

Sequence Format:

---

**This chromosome region corresponds to the contig region(s):**

Contig	start	stop	strand	
NT_008251.13	1648566	1743120	+	<a href="#">Display</a> <a href="#">Save to Disk</a> <a href="#">View Evidence</a> <a href="#">ModelMaker</a>

## NCBI model maker (mm)

The screenshot shows the NCBI Model Maker interface in a Netscape browser window. The URL is <http://www.ncbi.nlm.nih.gov/jet/ModelMaker>. The interface displays the following information:

- Evidence:** NT\_008251.13 chr 8 pos 1648566 change strand
- Putative exons (graphic view):** A graphical representation of exons and introns with a yellow highlight on the first exon.
- Your model:** A linear representation of the model with a red arrow indicating the direction.
- FASTA sequence:** A block of DNA sequence with a red circle highlighting a specific region.
- Protein model:** A diagram showing three frames (ORF#1, ORF#41, ORF#739) with amino acid sequences and start/stop codons.
- Putative exons (table view):** A table listing genomic coordinates and exon numbers.

Exon	Start	Stop	Strand
1	1743087	1743016	+
2	1742043	1741967	+
3	1738890	1738773	+
4	1729728	1729550	+
5	1726001	1725833	+
6	1714097	1714041	+
7	1693571	1693500	+
8	1693082	1692918	+
9	1691889	1691805	+
10	1681892	1681826	+
11	1674408	1674223	+
12	1674408	1674397	+
13	1674408	1674397	+
14	1673982	1673948	+
15	1671874	1671879	+
16	1661112	1660959	+
17	1660704	1660559	+
18	1654978	1654448	+
19	1654281	1654143	+
20	1651462	1651303	+
21	1649724	1649673	+
22	1648899	1648859	+

## ORF Finder and BLAST

**NCBI ORF Finder (Open Reading Frame Finder)**

Anonymous

View | 1 GenBank | Redraw | 100 | SixFr

View | 1 GenBank | Redraw | 100 | Sixframes

Frame from to Length

+3	72..2222	2151
+1	1180..1452	273
-3	1327..1512	186
+1	1855..2007	153
-3	1699..1845	147
-2	2174..2314	141
+2	2..127	126
+1	802..921	120
+1	2275..2388	114
+1	2416..2523	108

>ref|NP\_001455.2| a disintegrin and metalloproteinase domain 2 propeptide; Fertillin beta (a disintegrin and metalloproteinase domain 2); fertillin beta [Homo sapiens]  
 |c|J03480| fertillin beta chain - human  
 |b|AA04206.1| fertillin beta [Homo sapiens]  
 Length = 734

Score = 1483 bits (3840), Expect = 0.0  
 Identities = 714/735 (97%), Positives = 714/735 (97%), Gaps = 20/735 (2%)

Query: 1 MNRVLLSGLGIRMSNFDSLFPQITVPEKIRSIKEGIESQASYKIVIEGKPTVNL 60  
 M1 VLFLLSGLGIRMSNFDSLFPQITVPEKIRSIKEGIESQASYKIVIEGKPTVNL 60  
 Sbjct: 1 MNRVLLSGLGIRMSNFDSLFPQITVPEKIRSIKEGIESQASYKIVIEGKPTVNL 59

Query: 61 MQRNLFHNFRVYSVSGTGMKFLDQDFNFCVQGYIEGYPKSVVMVSTCGLRGVLF 120  
 MQRNLFHNFRVYSVSGTGMKFLDQDFNFCVQGYIEGYPKSVVMVSTCGLRGVLF 120  
 Sbjct: 60 MQRNLFHNFRVYSVSGTGMKFLDQDFNFCVQGYIEGYPKSVVMVSTCGLRGVLF 119

Query: 121 ENVVSGIEPLESSVGFHWYQVHKHKADVSLNENKLESRDLSFKLQSE----- 171  
 ENVVSGIEPLESSVGFHWYQVHKHKADVSLNENKLESRDLSFKLQSE E  
 Sbjct: 120 ENVVSGIEPLESSVGFHWYQVHKHKADVSLNENKLESRDLSFKLQSAEQDFAKYI 179

Query: 172 -----YHMGSDTPVVAQKVFQGLIGLNAIFVSNITILSSLELWIDENLIATT 221  
 YHMGSDTPVVAQKVFQGLIGLNAIFVSNITILSSLELWIDENLIATT  
 Sbjct: 180 EMHIVERQLYHMGSDTPVVAQKVFQGLIGLNAIFVSNITILSSLELWIDENLIATT 239

## NCBI HomoloGene (hm)

**NCBI HomoloGene (hm)**

PubMed | Entrez | BLAST | OMM | Taxonomy | Structure

Search: All organisms | for |

**HOMOLOGENE ENTRY**

**H.sapiens -ADAM2** a disintegrin and metalloproteinase domain 2 (fertilin beta)  
 UniGene | LocusLink | MGI | MapViewer | NM\_001464.2

**POSSIBLE HOMOLOGOUS GENES**

**M.musculus -Adam2** a disintegrin and metalloproteinase domain 2  
 UniGene | LocusLink | MGI | MapViewer | XM\_127888.2

**R.norvegicus -Adam2** a disintegrin and metalloproteinase domain 2  
 UniGene | LocusLink | ISG1 | MapViewer | NM\_020077.1

**S.scrofa -Ssc.16631** Sus scrofa mRNA for fertillin beta (FTNB gene)  
 UniGene | A559933.2

**B.taurus -ADAM2** a disintegrin and metalloproteinase domain 2 (fertilin beta)  
 UniGene | LocusLink | AF388808.1

**CALCULATED ORTHOLOGS**

Listed below are the nucleotide sequence comparisons used in determining homology. The pairs below represent reciprocal best hits; each alignment is the best one for both organisms. The percent ID below represents identity over an aligned region. When present, red arrows (↔) point out a group of sequence matches which are part of a triplet, being consistent between more than two organisms.

Organism	Gene	Organism	Gene	Percent ID
H.sapiens -ADAM2		B.taurus -ADAM2		80.9
H.sapiens -ADAM2		S.scrofa - Ssc.16631		80.8
H.sapiens -ADAM2		M.musculus -Adam2		77.6
H.sapiens -ADAM2		R.norvegicus -Adam2		77.2
M.musculus -Adam2		R.norvegicus -Adam2		89.5
S.scrofa -Ssc.16631		B.taurus -ADAM2		84.8
M.musculus -Adam2		S.scrofa - Ssc.16631		77.2
M.musculus -Adam2		B.taurus -ADAM2		76.9
R.norvegicus -Adam2		S.scrofa - Ssc.16631		75.6
R.norvegicus -Adam2		B.taurus -ADAM2		76.6

**CURATED ORTHOLOGS**

Published orthologs as reported in curated databases

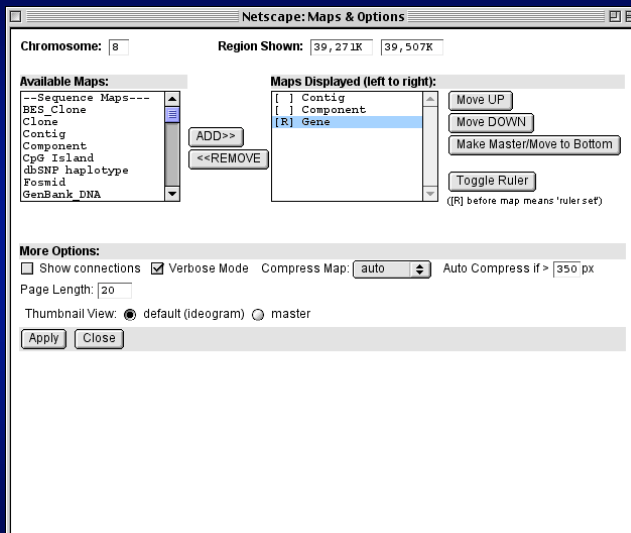
H.sapiens -ADAM2	M.musculus -Adam2	MGI
------------------	-------------------	-----

**FURTHER READING**

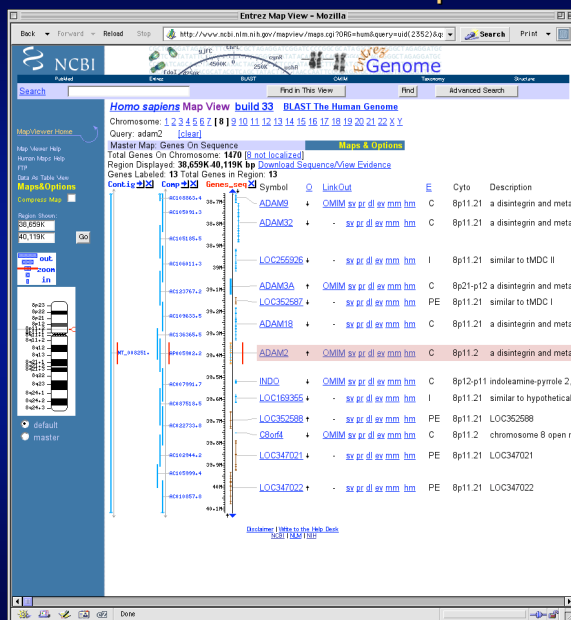
- Wolfsberg TG, et al. ADAM, a widely distributed and developmentally regulated gene family encoding membrane proteins with a disintegrin and metalloproteinase domain. Dev Biol 169: 378-383 (1995).
- Vidaaes CM, et al. Human fertillin beta: identification, characterization, and chromosomal mapping of an ADAM gene family member. Mol Reprod Dev 46: 363-369 (1997).



## NCBI Maps & Options



## Additional NCBI maps



## NCBI: Search for region between 2 STS markers

Search for  on chromosome(s)

Show linked entries Help FTP M/home

**Homo sapiens genome view build 33**

Hits: 1 2 3 4 5 6 7 8 9 10 11 12 13

Hits: 14 15 16 17 18 19 20 21 22 X Y MT

Search results for query "D21S1869 OR D21S1989": 2 hits

Chr	Match	Map element	Type	Maps
21	<a href="#">all matches</a>			
21	D21S1869	<a href="#">D21S1869</a>	STS	<a href="#">STS</a>   <a href="#">NCBI_RH</a>   <a href="#">Stanford_G3</a>   <a href="#">TNG</a>
21	D21S1989	<a href="#">D21S1989</a>	STS	<a href="#">STS</a>

## NCBI: Default view of region

**Homo sapiens Map View build 33** BLAST The Human Genome

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 [21] 22 X Y

Query: D21S1869 OR D21S1989

Master Map: STS

Total STSs On Chromosome: 1741

Region Displayed: 39M-44M bp

STSs Labeled: 20 Total STSs in Region: 327

Region Shown:

out  zoom  in

default  master

marker	Kbp	STS	NCBI	Stanford	TNG	poly
G48508	40277					
G43592	40897					
SHGC-144847	41266					
RH45177	41470					Y
SHGC-52452	41621					
G62824	41824					
RH93963	42302					Y
NIB1490	42615					Y
SHGC-30133	42787					Y
RH78209	43090					
D21S1869	43230					Y
WI-16888	43345					Y
GBR-192314	43368					
D21S1989	43604					Y
D21S1890	43705					Y
SHGC-10580	43973					
RH103858	44049					
D21S1259	44181					Y
G34627	44279					
SHGC-87686	44422					Y



## NCBI: Search for ADAM2 in the mouse genome

Search for  on chromosome(s)  strain  Find

Entrez Genomes  
MapViewer Home  
Prominent Organisms  
Maps  
Related Resources  
Sequence Data

**Mus musculus genome view build 30** [BLAST search the mouse](#)

Hits: 1 2 3 4 5 6 7 8 9 10 11

Hits: 12 13 14 15 16 17 18 19 X Y MT

Search results for query "adam2": 2 hits

Chr	Match	Map element	Type	Maps
14	all matches			
14	Adam2	Adam2	LOCUS	<a href="#">Genes_seq</a>   <a href="#">MGI</a>
14	Adam2	Adam2	STS	<a href="#">STS</a>

## NCBI: Default view of ADAM2 in mouse

**Mus musculus Map View build 30** [BLAST The Mouse Genome](#)

Chromosome: 1 2 3 4 5 6 7 8 9 10 11 12 13 | 14 | 15 16 17 18 19 X Y

Query: adam2 [\[clear\]](#)

Master Map: Genes On Sequence [Maps & Options](#)

Total Genes On Chromosome: 1326 [28 not localized]

Region Displayed: 45M-57M bp [Download Sequence/View Evidence](#)

Genes Labeled: 20 Total Genes in Region: 233

STS	MGI	Strain	Genes_seq	Symbol	LinkOut	Description
			1429_s4b...	D14Ucla2	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C DNA segment, Chr
				Psmc1	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C proteasome (prosor
				Trn9sf1	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C transmembrane 9 s
				D14Ertd484e	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C DNA segment, Chr
				Cbln3	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C cerebellin 3 precurs
				Mcpt2	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C mast cell protease
				Gzmc	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C granzyme C
				Atp12a	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C ATPase, H <sup>+</sup> /K <sup>+</sup> tra
				Cry11	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C crystallin, lamda 1
				TgN737Rpw	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C transgene insert sit
				Il17d	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C interleukin 17D
				Sacs	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C saccin
				Fdft1	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C fattyacyl diphosphat
				Gata4	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C GATA binding prote
				4930578J06Rik	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C RIKEN cDNA 49305
				F830020C16Rik	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C RIKEN cDNA F8300
				Extl3	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C exostosin (multiple
				1110020C17Rik	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C RIKEN cDNA 11100
				Adam2	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C a disintegrin and m
				2410004D02Rik	<a href="#">MGI</a> <a href="#">sv</a> <a href="#">pr</a> <a href="#">dl</a> <a href="#">ev</a> <a href="#">mm</a> <a href="#">hm</a>	C RIKEN cDNA 24100

## NCBI: View features annotated on sequence from other mouse strains

**Mus musculus Map View build 30** BLAST The Mouse Genome

Chromosome: 14 Strain: 129\_substrain  
 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 X Y

Query: adam2 [clear]

Master Map: Genes On Sequence [Maps & Options]

Total Genes On Chromosome: 91 [30 not localized]

Region Displayed: 45M 57M bp [Download Sequence/View Evidence]

Genes Labeled: 6 Total Genes in Region: 6

Symbol	LinkOut	Description
1110028A07Rik	MGJ sv pr dl ev mm hm	C RIKEN cDNA 1110028A07
Psmc1	MGJ sv pr dl ev mm hm	C proteasome (prosome)
1500005A01Rik	MGJ sv pr dl ev mm hm	C RIKEN cDNA 1500005A01
Psmc2	MGJ sv pr dl ev mm hm	C proteasome (prosome)
LOC268749	sv pr dl ev mm hm	C hypothetical protein
Isgfg	MGJ sv pr dl ev mm hm	C interferon dependent

**Available strains:**

- 129 substrain
- B6/CBAF1J
- BA/2J
- C3H
- C57BL/6J
- NOD
- Unknown
- MGSCv3 (reference)

## Ensembl Home

Ensembl Genome Browser - Mozilla

http://www.ensembl.org/

Search all species for: Anything with: [ ]

Species Selection:

Species	Version	Date
Human	v. NC21	2 Jul 2003
Mouse	v. NC21	6 May 2003
Rat	v. NC21	1 Apr 2003
Zebrafish	v. NC21	2 Jul 2003
Fugu	v. NC21	3 Mar 2003
Mosquito	v. NC21	6 May 2003
Fruitfly	v. NC21	2 Jul 2003
C. elegans	v. NC21	2 Jul 2003
C. briggsae	v. NC21	3 Mar 2003

Ensembl provides:

- Gene structure and sequence data
- Full length protein prediction to compare and highlight in the genome sequence
- Functional annotation of genes, all in support of Ensembl
- Annotation of some classes of the genome
- Detailed annotation of gene structure and exons

Ensembl provides:

- Gene structure and sequence data
- Full length protein prediction to compare and highlight in the genome sequence
- Functional annotation of genes, all in support of Ensembl
- Annotation of some classes of the genome
- Detailed annotation of gene structure and exons

Ensembl provides:

- Gene structure and sequence data
- Full length protein prediction to compare and highlight in the genome sequence
- Functional annotation of genes, all in support of Ensembl
- Annotation of some classes of the genome
- Detailed annotation of gene structure and exons

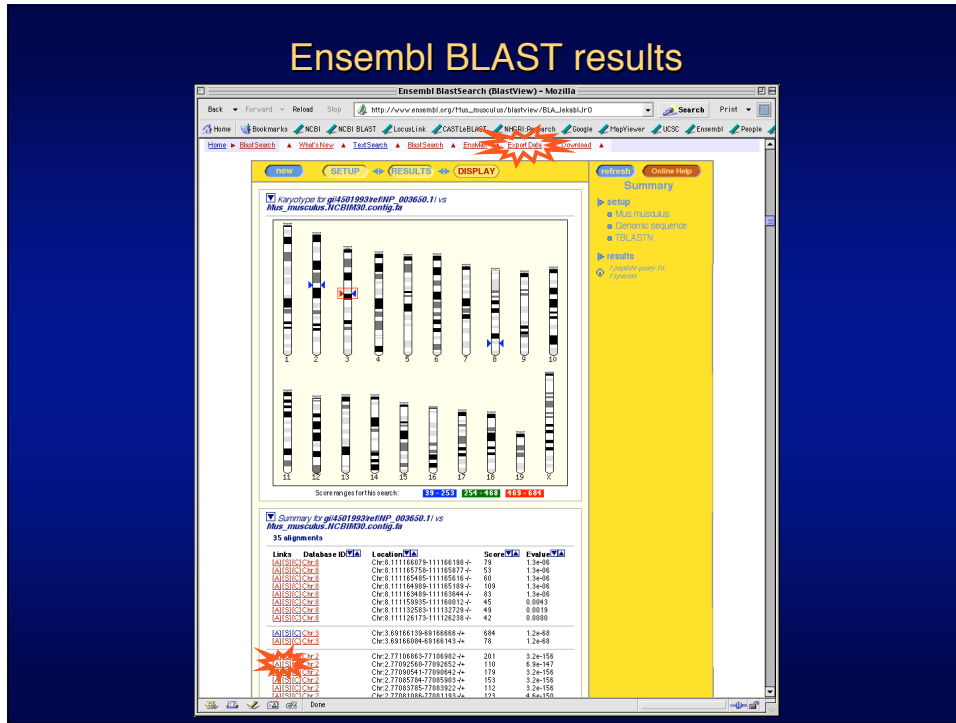
## Ensembl Mouse Home

The screenshot shows the Ensembl Mouse Genome Server homepage. The browser window title is "Ensembl Mouse Genome Server - Mozilla" and the address bar shows "http://www.ensembl.org/Mus\_musculus/". The page features a search bar with "Anything" selected and "with" and "From 1 To 10000" options. Below the search bar, there are links for "Retrieve a sequence" and "Advanced data retrieval tool: Ensembl". The main content area includes a section titled "About MGSC (16.20.1 details)" with a mouse icon and text describing the Mouse Genome Sequencing Consortium project. To the right, there is a "Browse a Chromosome" section with a chromosome ideogram. At the bottom, there are "Ensembl Links and Site Map" and "Other Species" sections with buttons for various species like Mosquito, C. briggsae, C. elegans, Zebrafish, Fruitfly, Fugu, Human, and Rat.

## Ensembl: BLAST human protein against mouse genome

The screenshot shows the Ensembl BLASTView interface. The browser window title is "Ensembl BLASTView". The page has a navigation bar with "Home", "BLASTSearch", "What's New", "Test Search", "BLAST Search", "Ensembl", "Export Data", and "Download". The main content area is divided into sections: "Enter the Query Sequence" with a text input field containing a protein sequence; "Or Upload a file containing one or more FASTA sequences" with a "Browse..." button; "Or Enter an existing Idset ID" with a "Retrieve" button; "Select the databases to search against" with checkboxes for various species and a "Genomic sequence" dropdown; "Select the Search Tool" with a "TBLASTN" dropdown and a "configure RUN" button; and "About BlastView" with a brief description of the tool's functionality.

## Ensembl BLAST results



## Ensembl BLAST details

### A (alignment) link

```

Query location      : gi|4501993|ref|NP_003650.1|      566 to 599
Database location   : 2                          77090541 to 77090642
Genomic location    : 2                          77090541 to 77090642

Alignment score     : 179
E-value             : 3.2e-156
Alignment length    : 34
Percentage identity : 97.06

Query:      566 RVTQTVDAGACIYFAPFYRNGISDPLTVFEOTE 599
           RVTQTVDAGACIYFAPFYRNGISDPLTVFE TE
Sbjct: 77090541 RVTQTVDAGACIYFAPFYRNGISDPLTVFEOTE 77090642
    
```

### S (query sequence) link

```

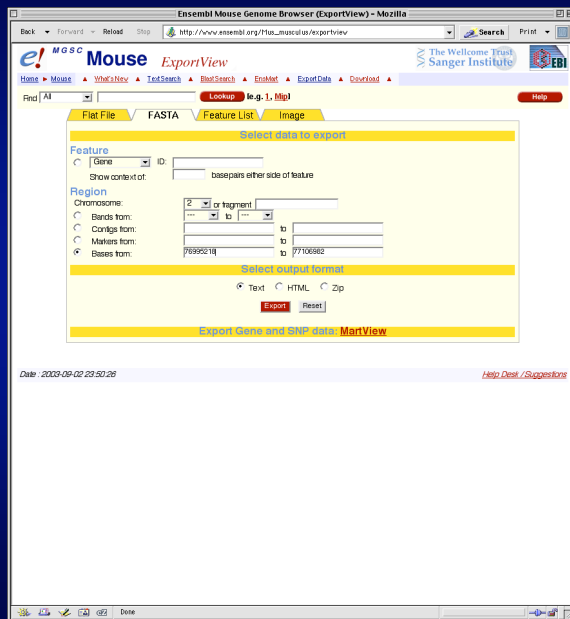
Query location      : gi|4501993|ref|NP_003650.1|      566 to 599
Database location   : 2                          77090541 to 77090642
Genomic location    : 2                          77090541 to 77090642

Alignment score     : 179
E-value             : 3.2e-156
Alignment length    : 34
Percentage identity : 97.06

THIS COLOUR: Matching bases for selected HSP
THIS COLOUR: Matching bases for other HSPs in selected hit

>gi|4501993|ref|NP_003650.1|
MAAFAAAAGGFTLQAGASTGSAADRDPPDRAGERLRLVLSGLLGRPREALSTHECA
EFAASAAATAATPAAGESGTFIKGQVTKWQWQYNDSEKIFKKGQELTKRYPFL
SGHGLPTFEWQLQNTLVNVEHKTSAKSNLSDPTFPVYVNEFLHDKETNLSYQKAD
DRVFAHGHCLKEIFLREGHFERIPDLVLPDCHDDVVKVILACKYMLCIIPIGGGTS
VSYGLCPADETPTTISLDTSQMRLIWDENRIFAHVAGITGQELRNLKESSVCTGH
EFLSLEYSTVGGWFTASGSDKMTYQWLELFTLKLMTFFGGLIESGQGFHSTGFDI
HGFHDSGTLGVIETATIKIPFPEYQKQVSTAFPNEFGVACLRIAKRCAPASIRL
MDNQVFGFHAKLPQVSSIFTFLDGLKVIITKFGDFPQLSVATLLFEGDREKVLQH
EKQVYDLAAKFGGAAEEDWQQRGVLITVYIAVIRDLALEYVVGESFETSAFWRVVDL
CRNVKERITRECKEKGQFAPFSTCRVVTQTDAGACIYFAPFYRNGISDPLTVFEOTE
AAKEEILANGGSLSHRGVGLRRLKLESI SDVGFGLKSVKEIVDPNLFGRNLL
    
```

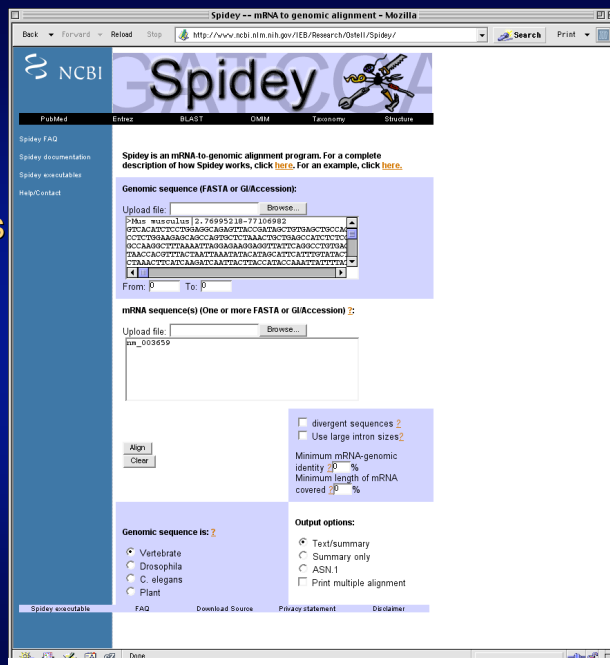
## Ensembl: Export genomic sequence of hit



## NCBI's Spidey: mRNA/Genomic alignments to determine positions of exons

Wheeler S.J., Church D.M., Ostell J.M.  
Spidey: a tool for mRNA-to-genomic  
alignments.  
Genome Res. 2001  
Nov;11(11):1952-7.

<http://www.ncbi.nlm.nih.gov/spidey>



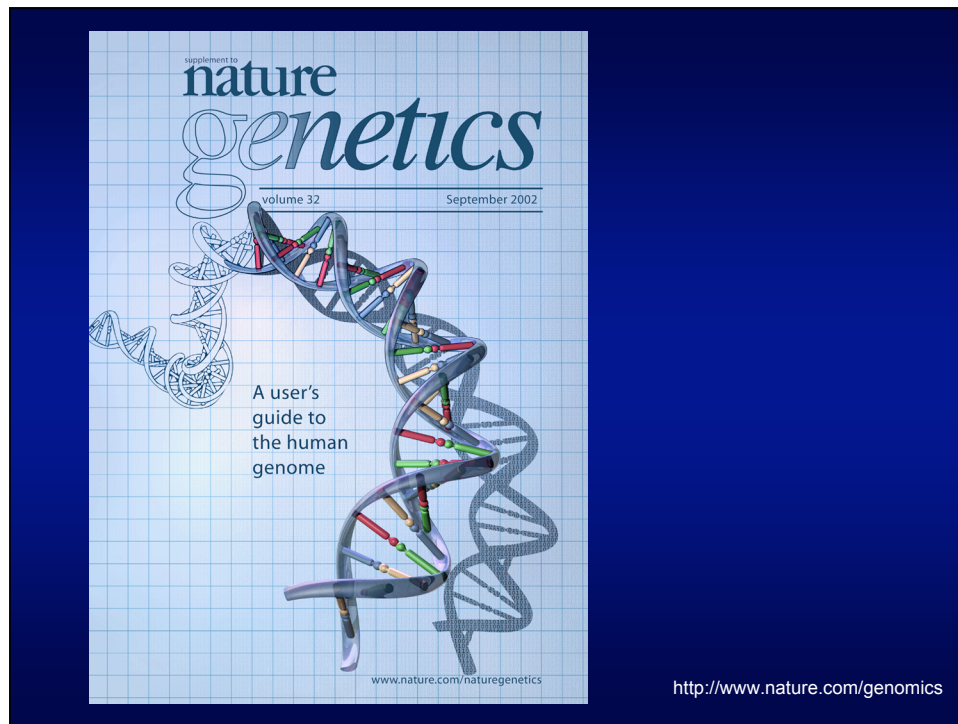




## Ensembl: Search for human gene name

## Ensembl: Human gene detail





### Additional resources

- UCSC Human Genome Browser User Guide  
<http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html>
- NCBI Genomic Biology  
<http://www.ncbi.nih.gov/Genomes/>
- NCBI MapViewer Help  
<http://www.ncbi.nlm.nih.gov/mapview/static/MapViewerHelp.html>
- Ensembl Tour  
<http://www.ensembl.org/Docs/enstour/>