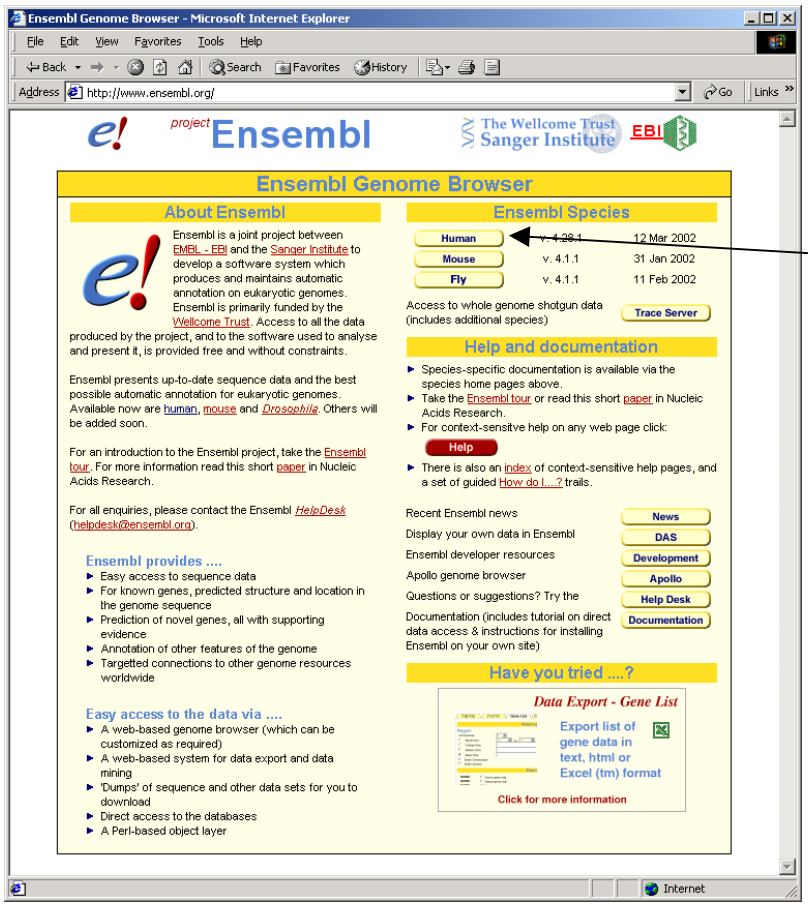
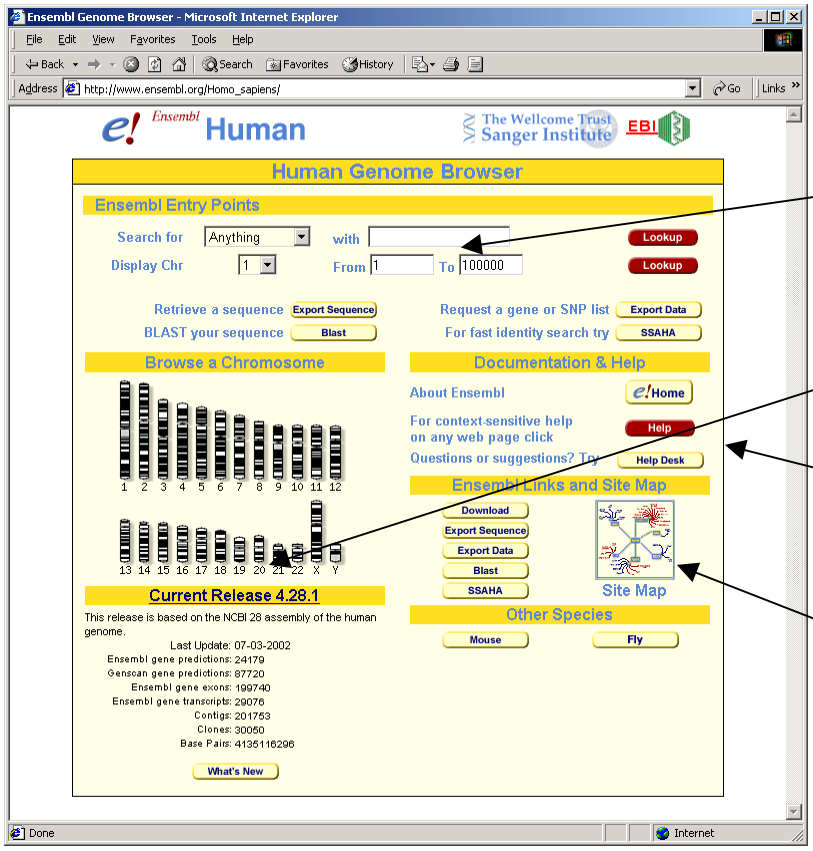
Worked example: From a chromosomal region to a gene



Pick Human Ensembl

'About Ensembl' plus general help and documentation – multi-species



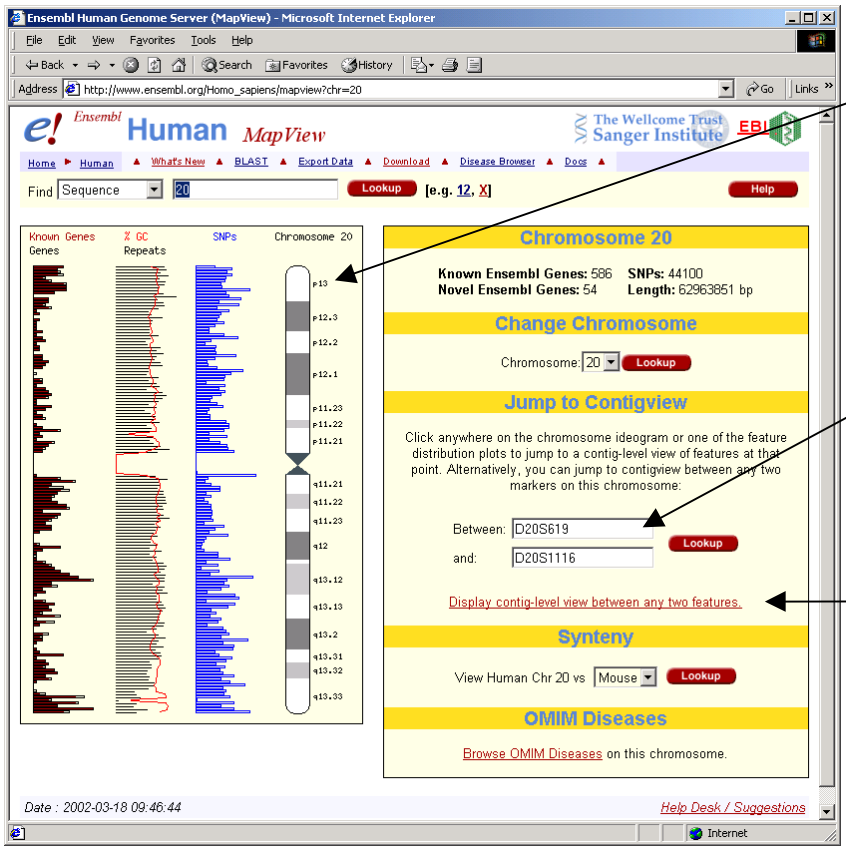
Search function

STEP 1

Pick chromosome 20

Help!

Quick access to the different "Views"

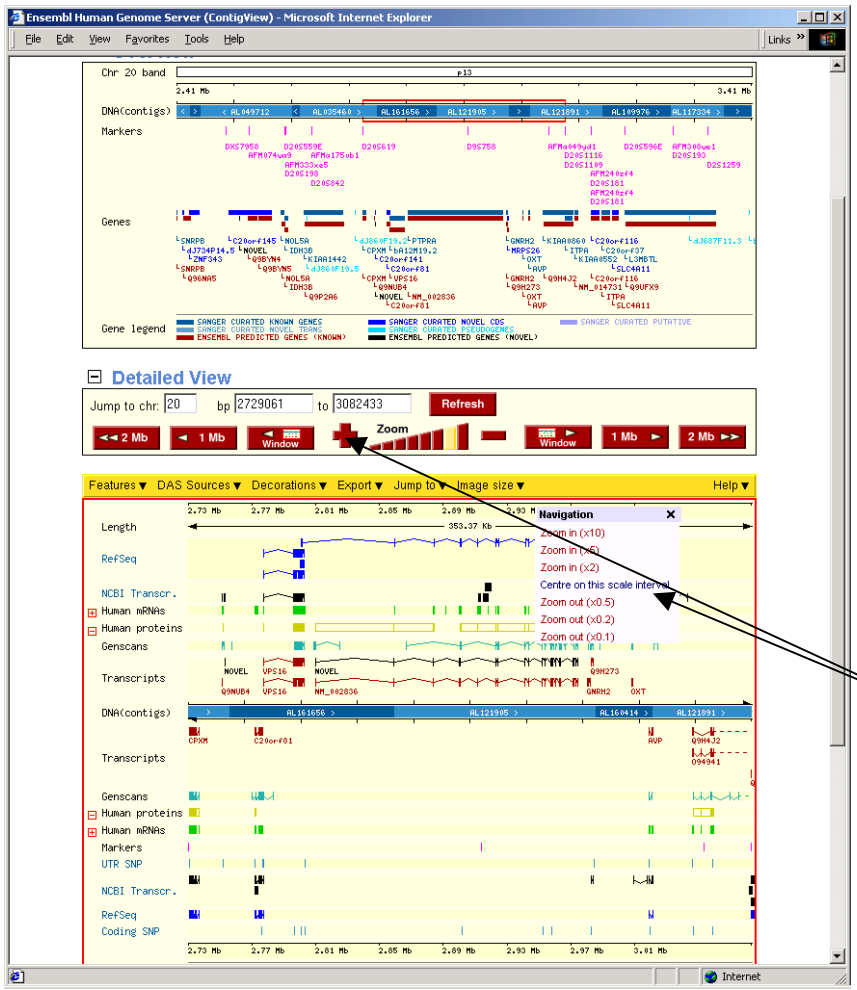


Display any region of interest

STEP 2

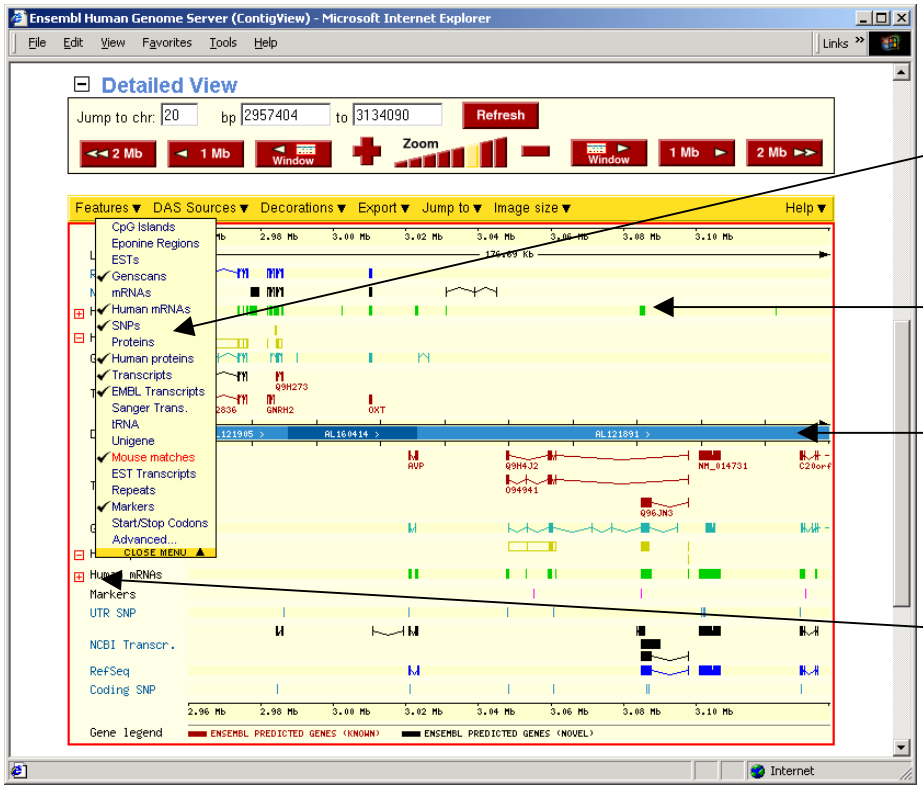
Display the region between markers D20S619 and D20S1116

Display a region between any two features



STEP 3

Centre around the AVP gene and then zoom in using the "plus" button



STEP 4

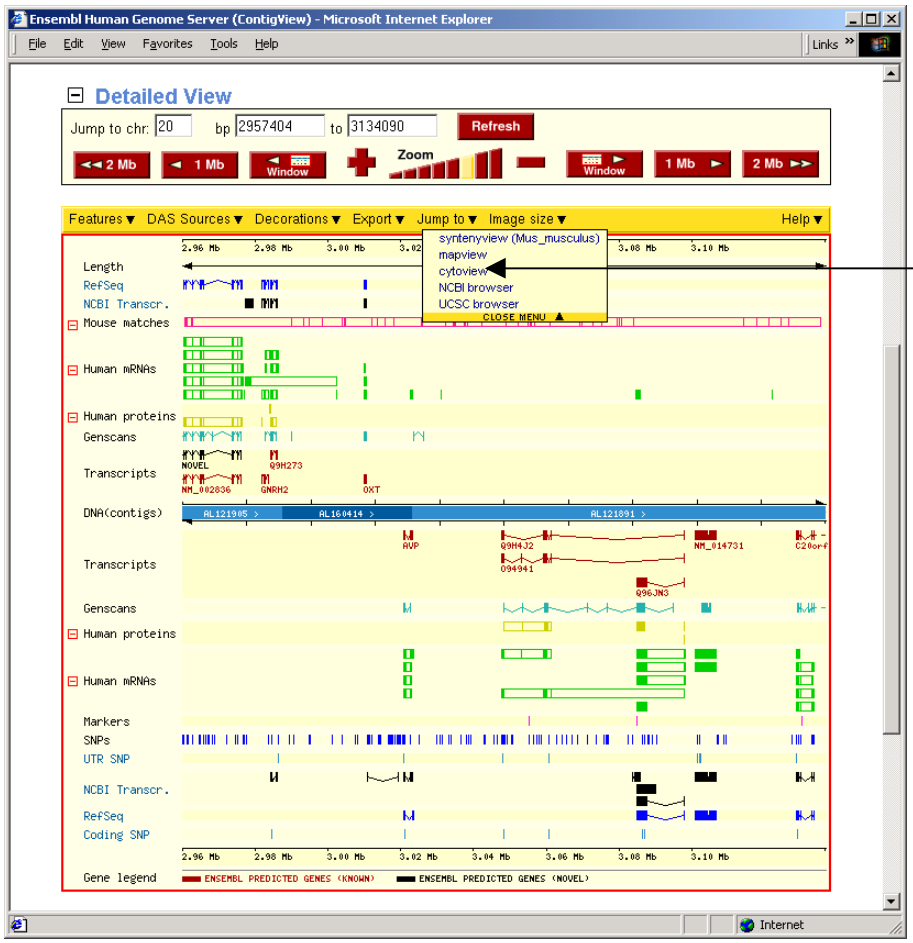
Customise the display – show SNPs and mouse matches

Evidence for gene prediction

DNA sequence - contigs

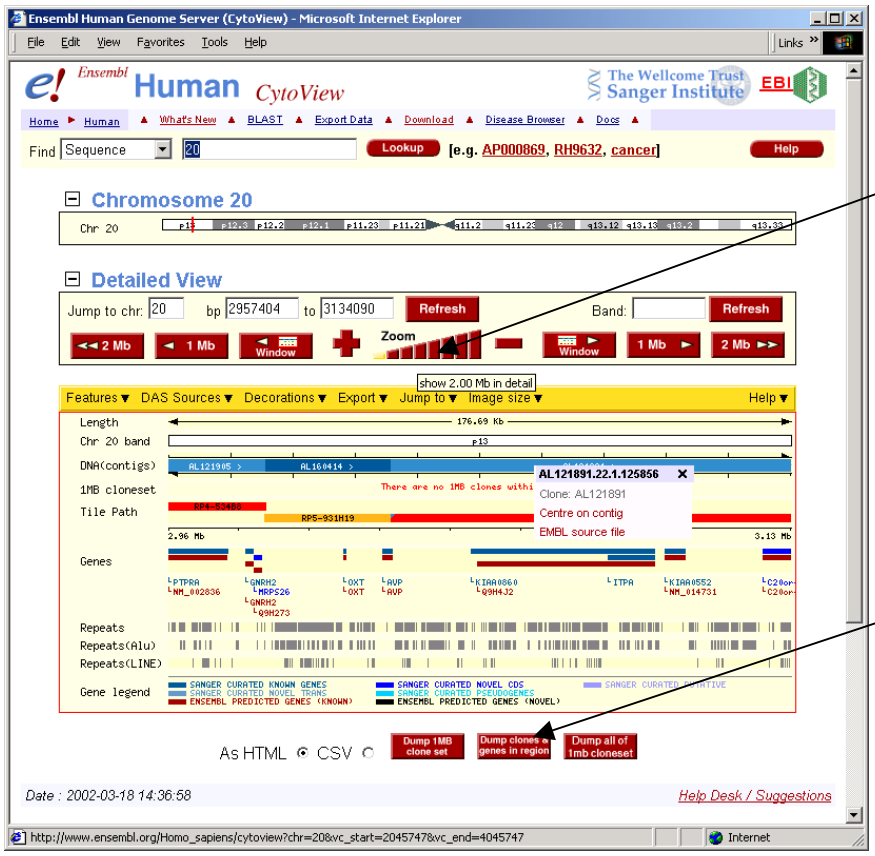
STEP 5

'Expand' the Human mRNA evidence



STEP 6

Jump to "cytoview"



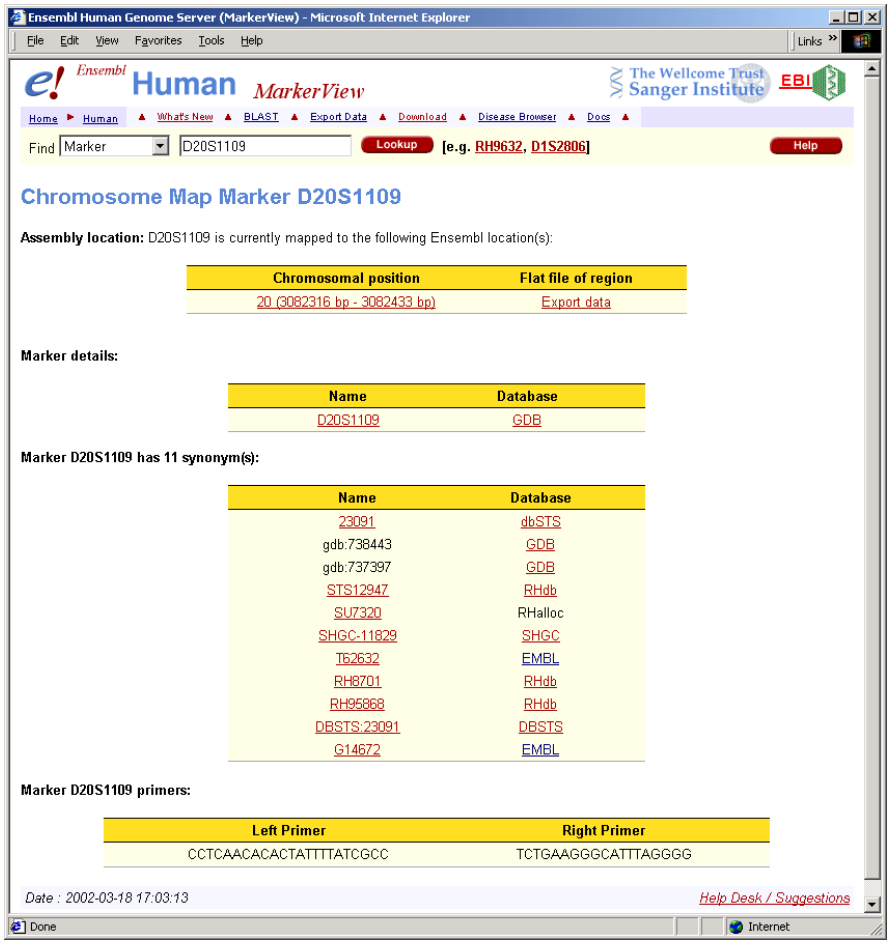
STEP 7

Zoom into larger view

Dump clones and genes

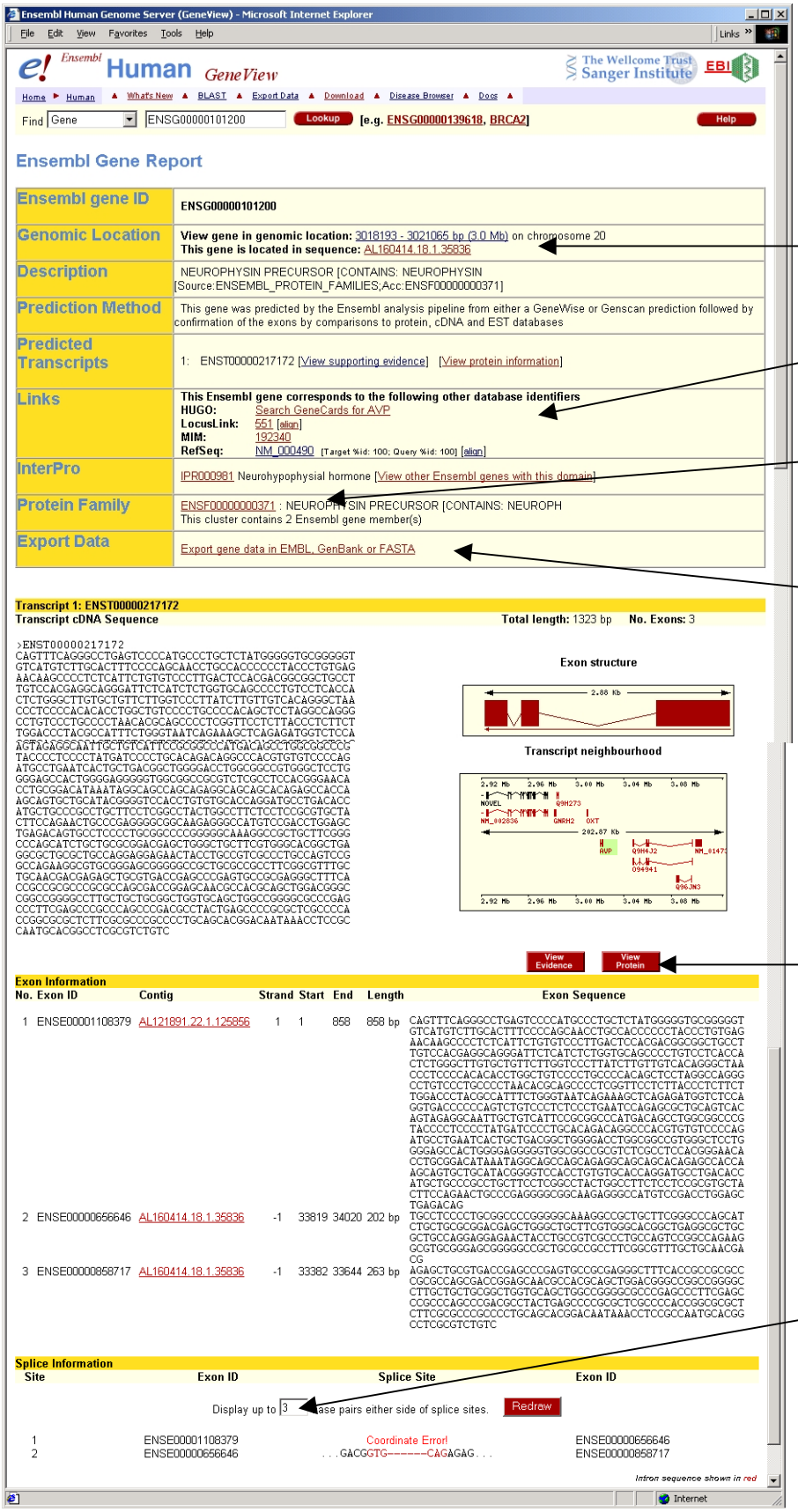
STEP 8

Go BACK to "ContigView" and then click on a marker



STEP 9

Go BACK and then click on the AVP transcript



View genomic location

Links to other databases

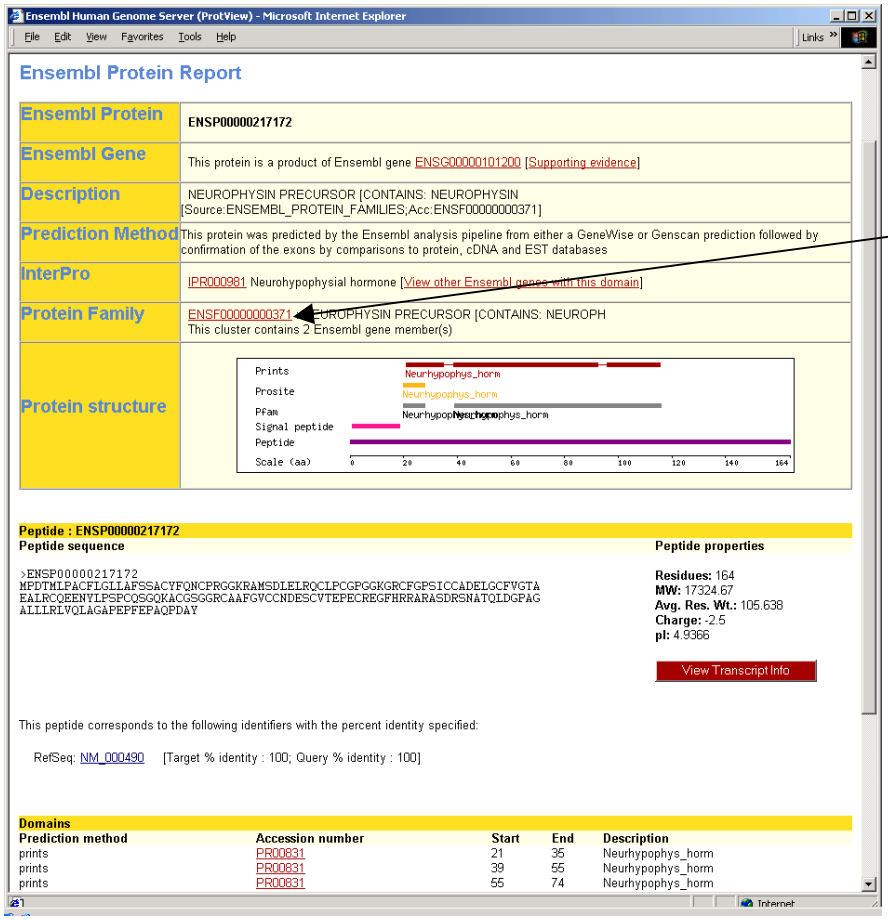
Find family members

Export data
- **retrieve sequence**
- **retrieve SNPs**

STEP 10

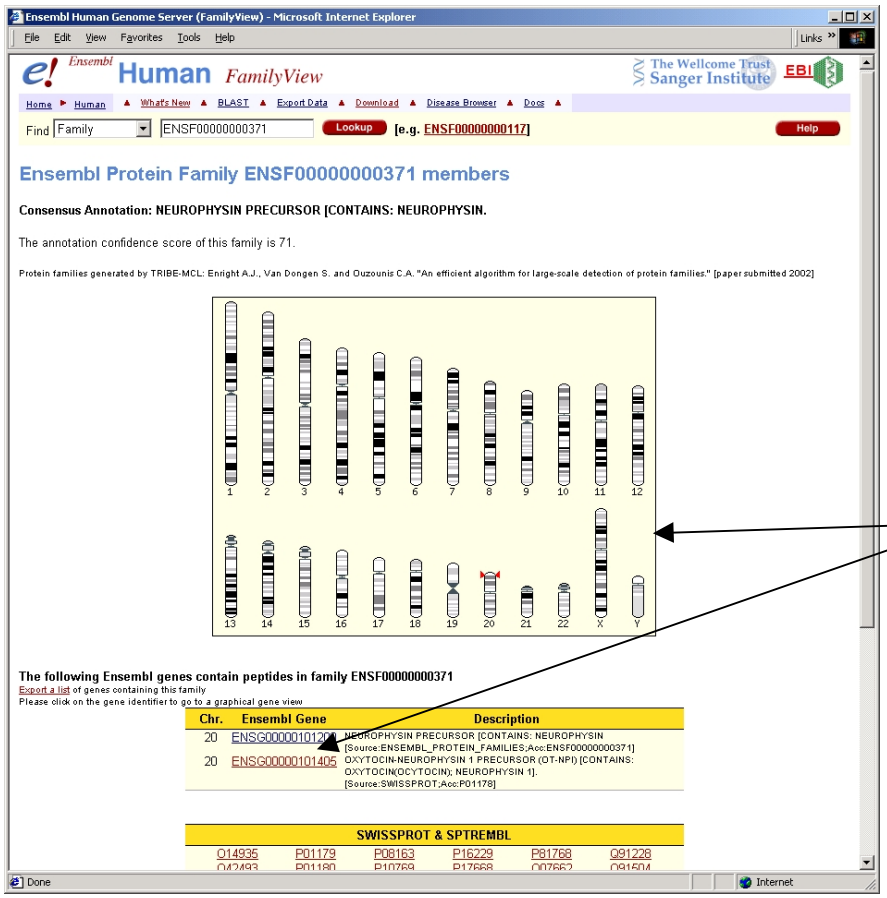
View protein

Display up to 99 base pairs on either side of the splice sites



STEP 11

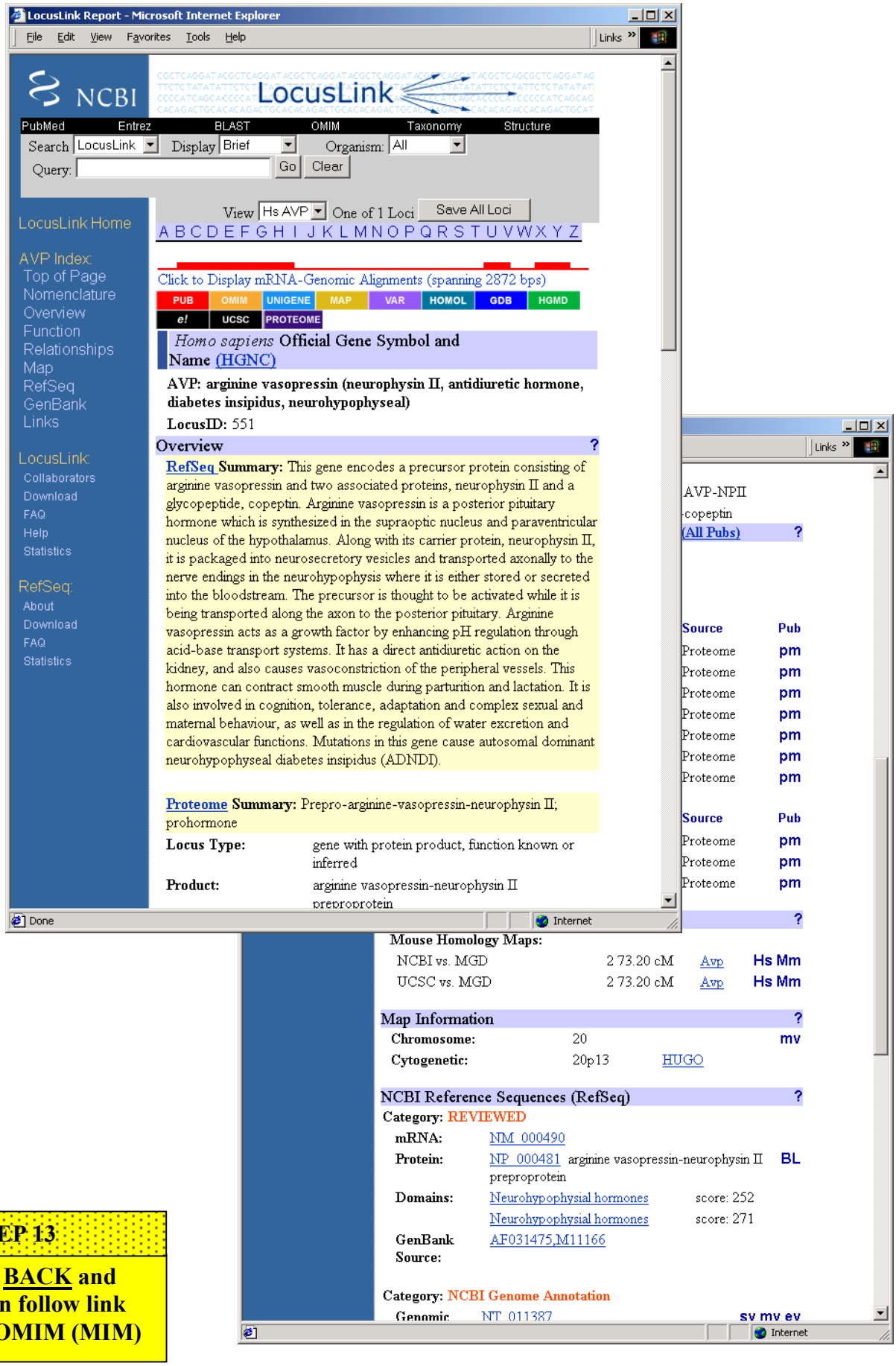
Follow link to protein family



Chromosomal location and links to other family members

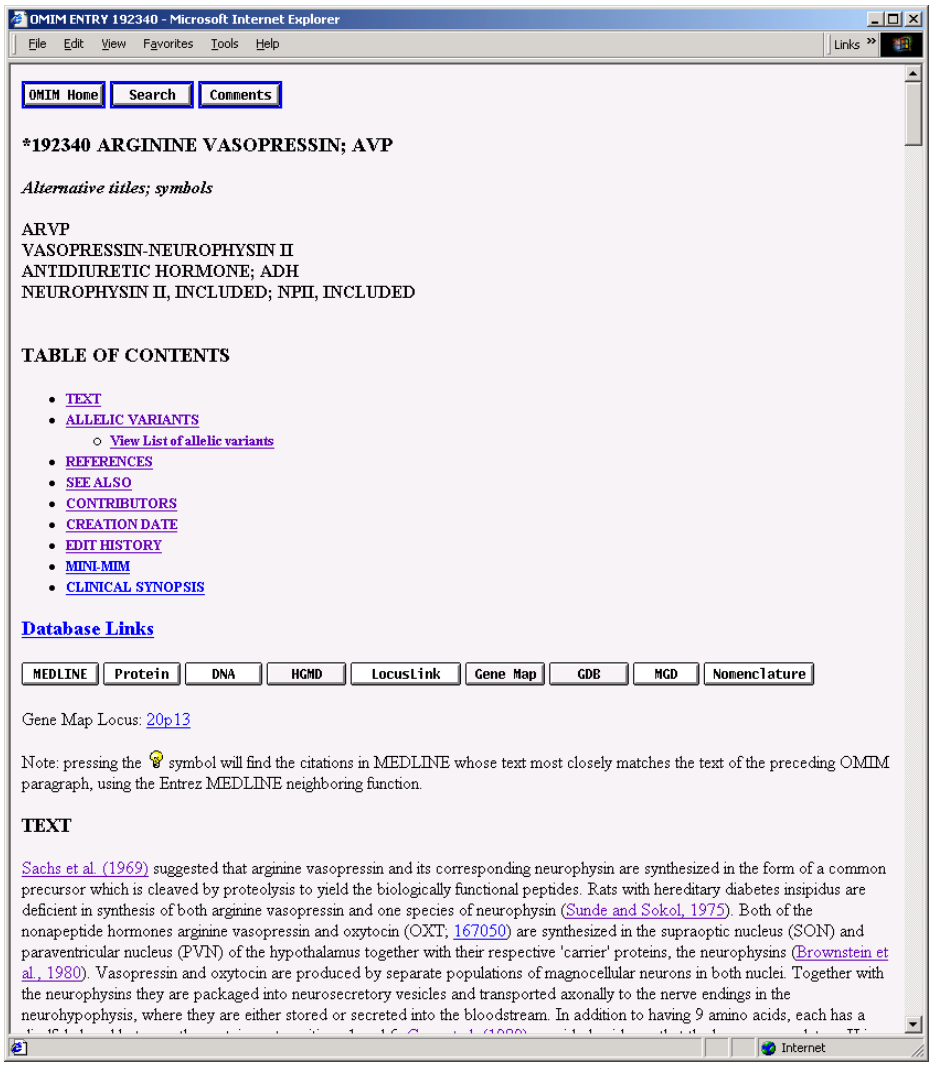
STEP 12

Go <u>BACK</u> to "GeneView" and then follow link to LocusLink



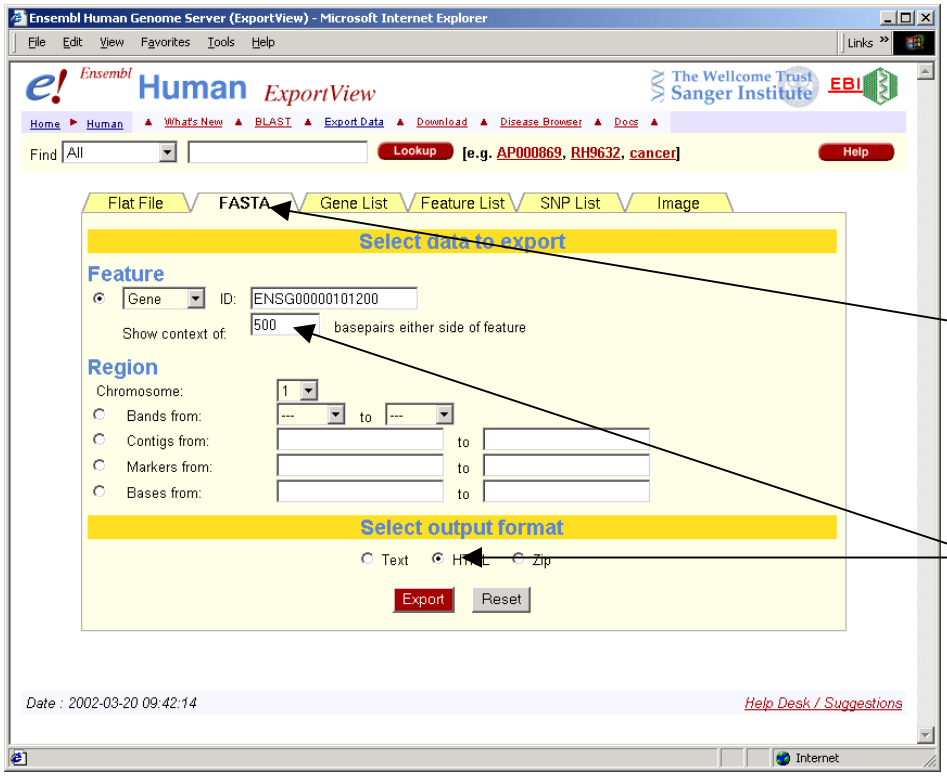
STEP 13

Go BACK and then follow link to OMIM (MIM)



STEP 14

Go BACK and then click on "Export gene data in EMBL, GenBank or FASTA"

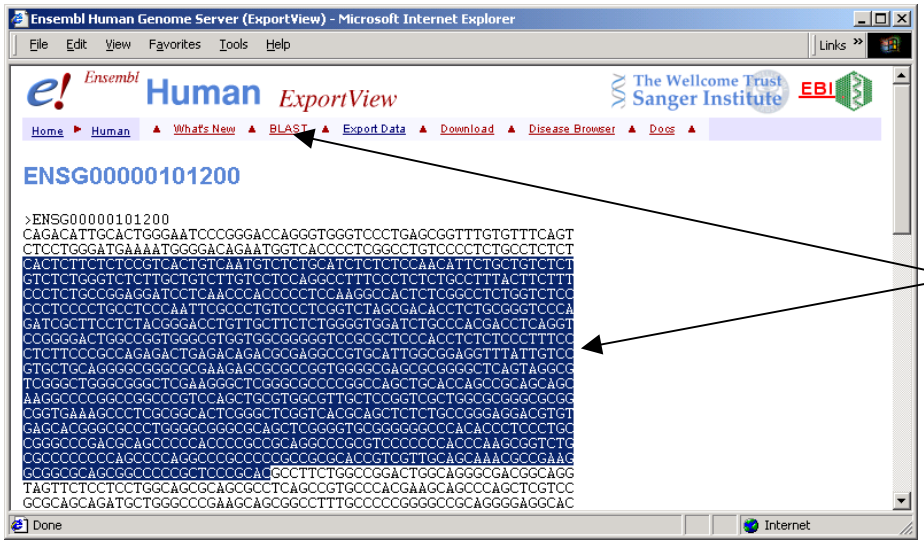


STEP 15

Choose the FASTA tab

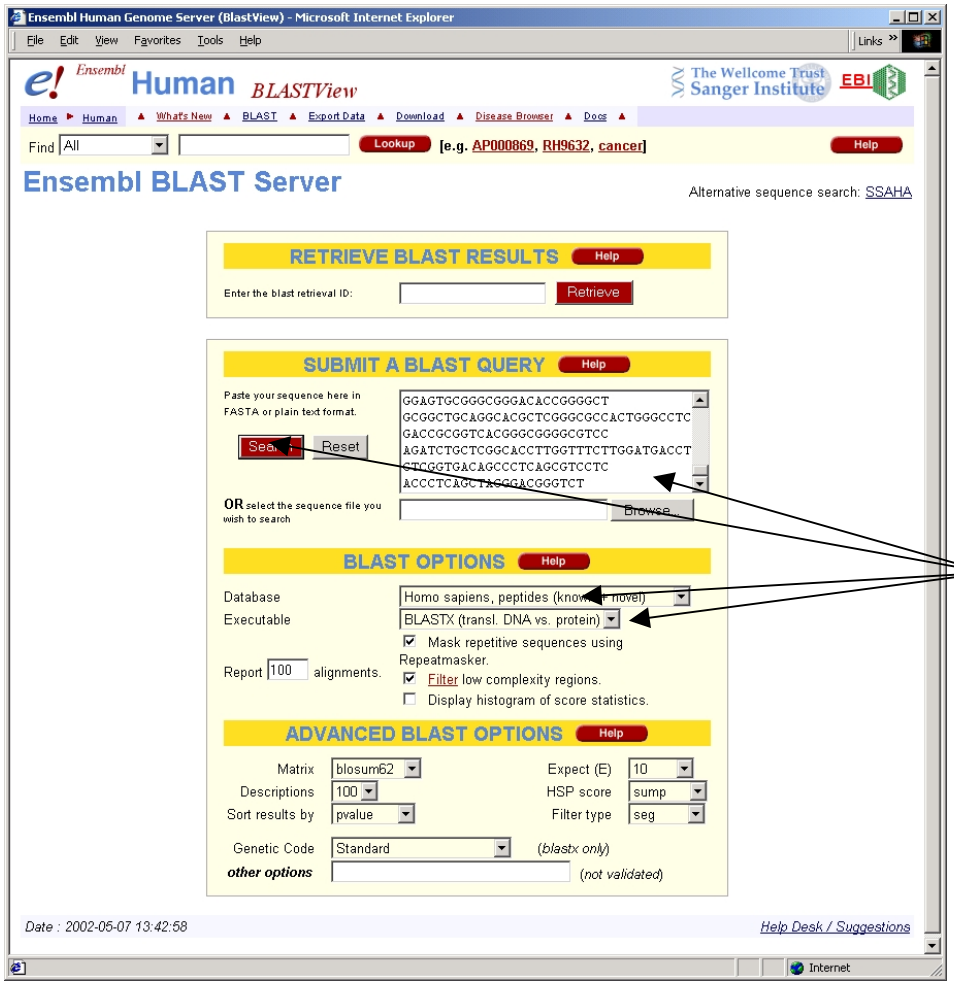
STEP 16

Include 500 bp on either side, select HTML format



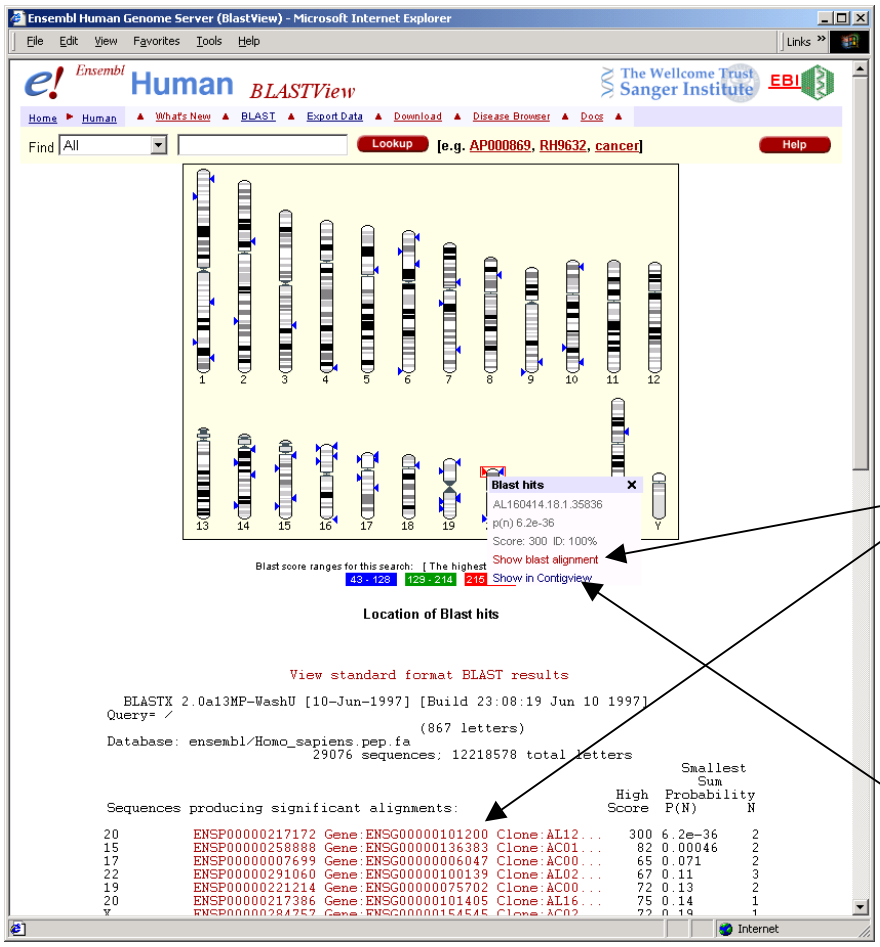
STEP 17

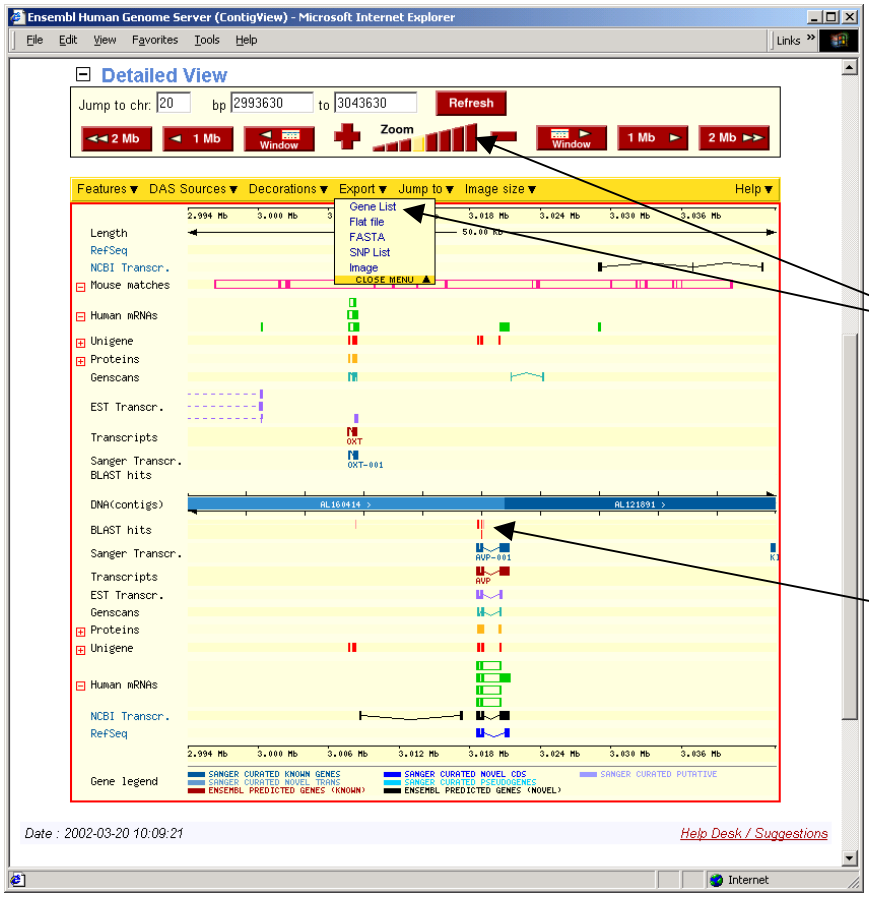
Copy part of the sequence, then click on BLAST

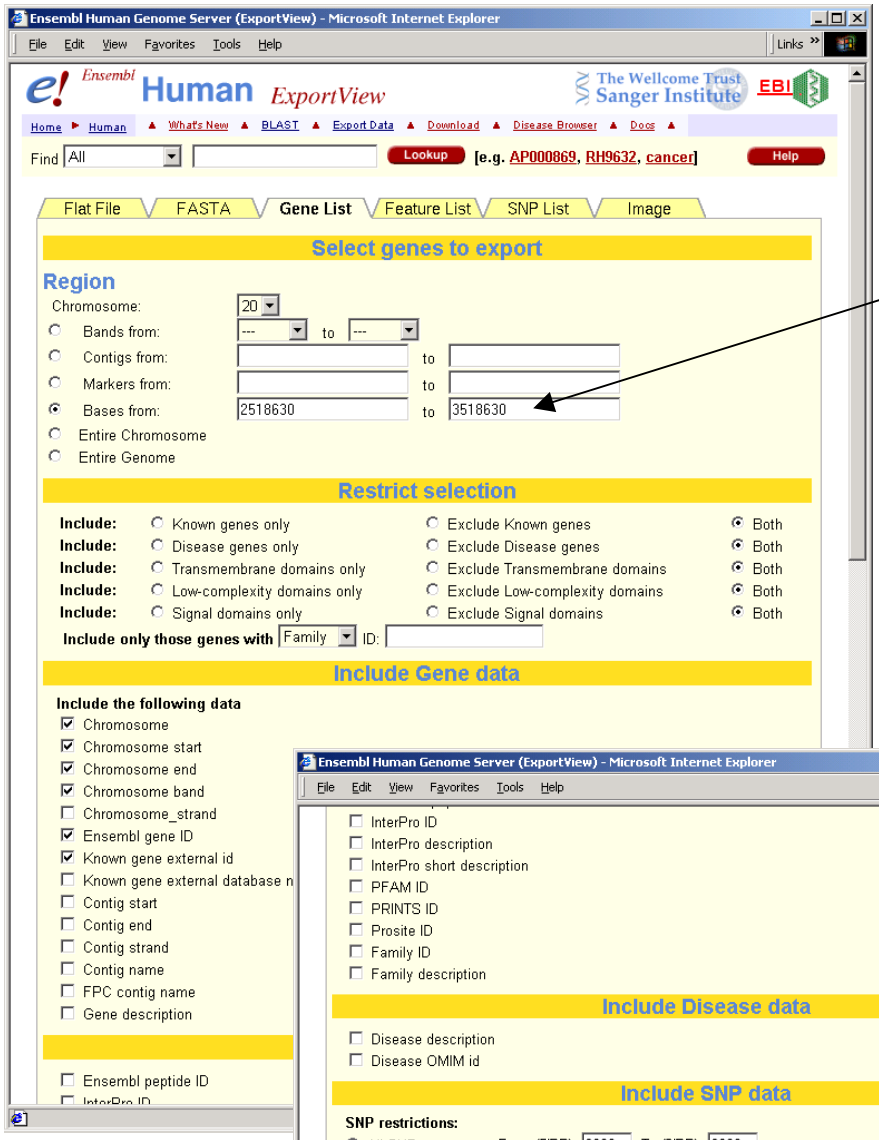


STEP 18

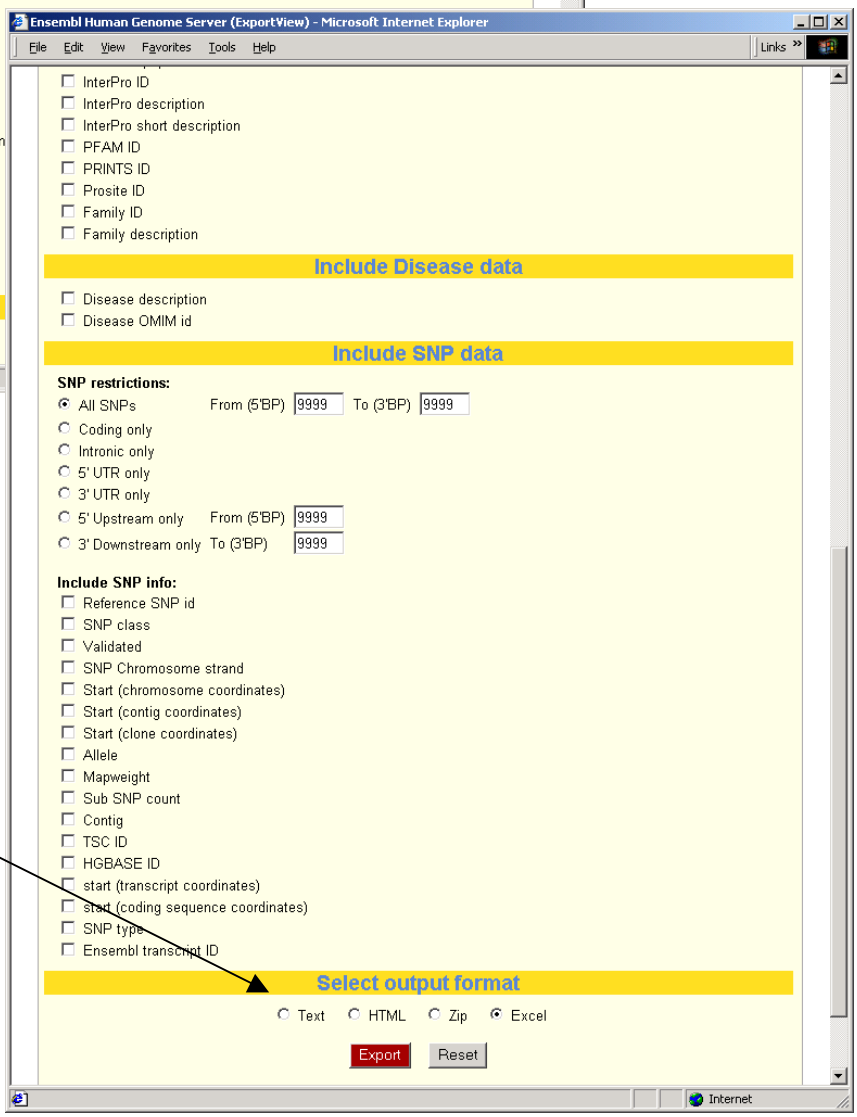
Paste the sequence, choose the 'peptides (known+novel)' database and BLASTX, then click on Search


Ensembl Human Genome Server (BlastView) - Microsoft Internet Explorer
Human BLASTView
Blast hits
Show blast alignment
Show in Contigview
Location of Blast hits
View standard format BLAST results
BLASTX 2.0a13MP-WashU [10-Jun-1997] [Build 23:08:19 Jun 10 1997]
Sequences producing significant alignments:
View the alignment
STEP 19
Display the region in ContigView


Ensembl Human Genome Server (ContigView) - Microsoft Internet Explorer
Detailed View
Jump to chr: 20 bp 2993630 to 3043630
Refresh
Zoom
Features ▼ DAS Sources ▼ Decorations ▼ Export ▼ Jump to ▼ Image size ▼ Help ▼
Gene List
Flat file
FASTA
SNP List
Image
CLOSE MENU
Length
RefSeq
NCBI Transcr.
Mouse matches
Human mRNAs
Unigene
Proteins
Genscans
EST Transcr.
Transcripts
Sanger Transcr.
BLAST hits
DNA(contigs)
Gene legend
Date : 2002-03-20 10:09:21
Help Desk / Suggestions
STEP 20
Zoom out, then export a Gene List
View the BLAST hits



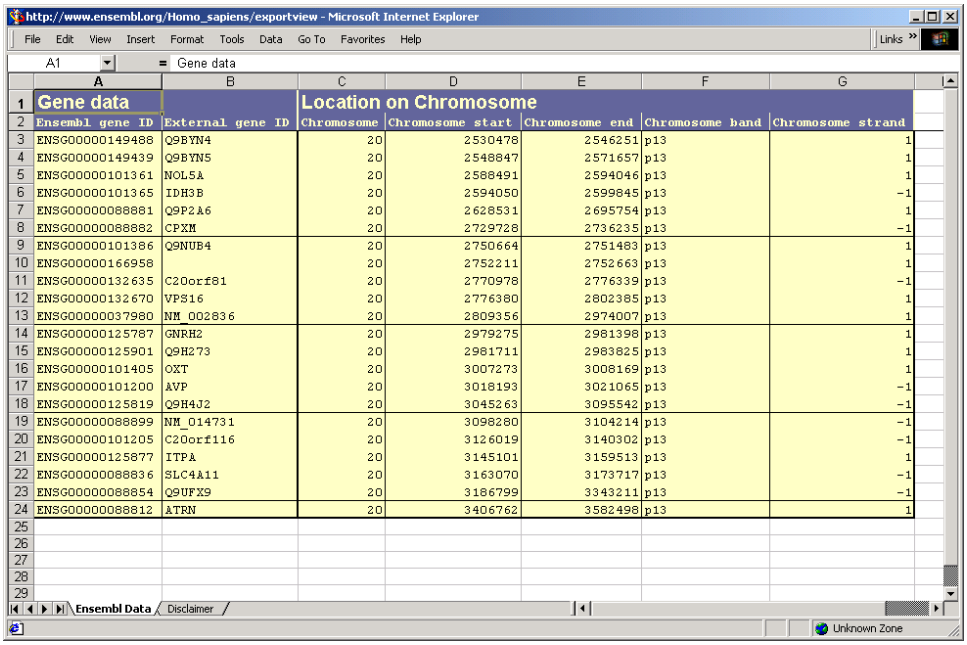
Chromosomal coordinates are filled in automatically



STEP 21

Choose an output format

Output in Excel format



http://www.ensembl.org/Homo_sapiens/exportview - Microsoft Internet Explorer

File Edit View Insert Format Tools Data Go To Favorites Help

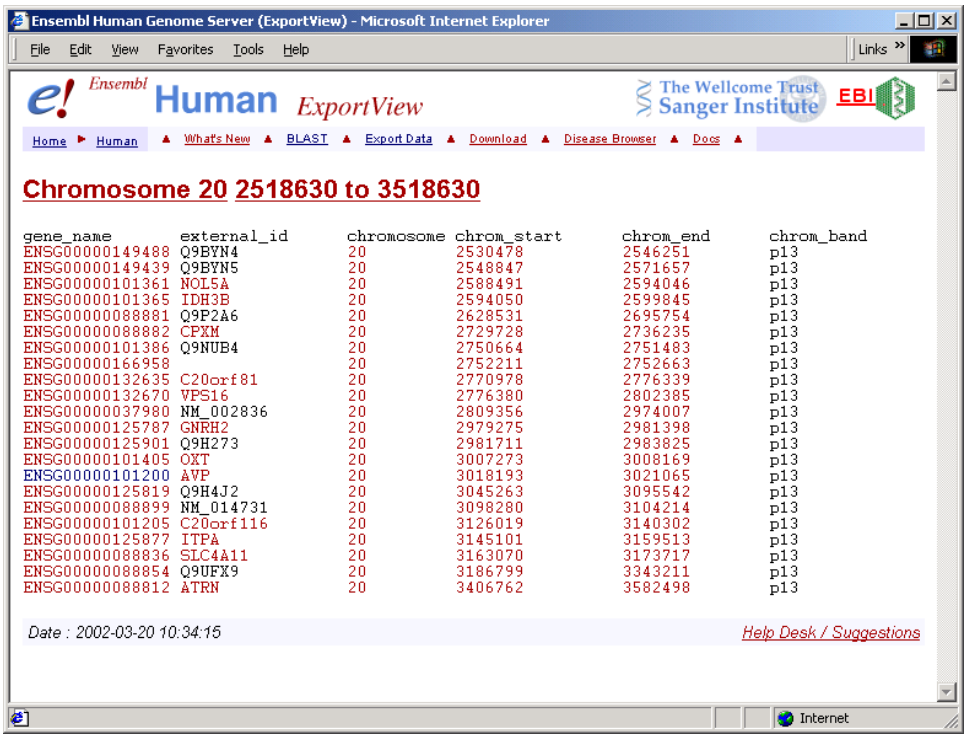
A1 = Gene data

	Gene data		Location on Chromosome				
	Ensembl gene ID	External gene ID	Chromosome	Chromosome start	Chromosome end	Chromosome band	Chromosome strand
3	ENSG00000149488	Q9BYN4	20	2530478	2546251	p13	1
4	ENSG00000149439	Q9BYN5	20	2548847	2571657	p13	1
5	ENSG00000101361	NOL5A	20	2588491	2594046	p13	1
6	ENSG00000101365	IDH3B	20	2594050	2599845	p13	-1
7	ENSG00000088881	Q9P2A6	20	2628531	2695754	p13	1
8	ENSG00000088882	CPXM	20	2729728	2736235	p13	-1
9	ENSG00000101386	Q9NUB4	20	2750664	2751483	p13	1
10	ENSG00000166958		20	2752211	2752663	p13	1
11	ENSG00000132635	C20orf81	20	2770978	2776339	p13	-1
12	ENSG00000132670	VPS16	20	2776380	2802385	p13	1
13	ENSG00000037980	NM_002836	20	2809356	2974007	p13	1
14	ENSG00000125787	GNRH2	20	2979275	2981398	p13	1
15	ENSG00000125901	Q9H273	20	2981711	2983825	p13	1
16	ENSG00000101405	OXT	20	3007273	3008169	p13	1
17	ENSG00000101200	AVP	20	3018193	3021065	p13	-1
18	ENSG00000125819	Q9H4J2	20	3045263	3095542	p13	-1
19	ENSG00000088899	NM_014731	20	3098280	3104214	p13	-1
20	ENSG00000101205	C20orf116	20	3126019	3140302	p13	-1
21	ENSG00000125877	ITPA	20	3145101	3159513	p13	1
22	ENSG00000088836	SLC4A11	20	3163070	3173717	p13	-1
23	ENSG00000088854	Q9UFX9	20	3186799	3343211	p13	-1
24	ENSG00000088812	ATRN	20	3406762	3582498	p13	1

Ensembl Data / Disclaimer

Unknown Zone

Output in HTML format



Ensembl Human Genome Server (ExportView) - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Ensembl Human ExportView — The Wellcome Trust Sanger Institute — EBI

Home ▶ Human ▲ What's New ▲ BLAST ▲ Export Data ▲ Download ▲ Disease Browser ▲ Docs ▲

Chromosome 20 2518630 to 3518630

gene_name	external_id	chromosome	chrom_start	chrom_end	chrom_band
ENSG00000149488	Q9BYN4	20	2530478	2546251	p13
ENSG00000149439	Q9BYN5	20	2548847	2571657	p13
ENSG00000101361	NOL5A	20	2588491	2594046	p13
ENSG00000101365	IDH3B	20	2594050	2599845	p13
ENSG00000088881	Q9P2A6	20	2628531	2695754	p13
ENSG00000088882	CPXM	20	2729728	2736235	p13
ENSG00000101386	Q9NUB4	20	2750664	2751483	p13
ENSG00000166958		20	2752211	2752663	p13
ENSG00000132635	C20orf81	20	2770978	2776339	p13
ENSG00000132670	VPS16	20	2776380	2802385	p13
ENSG00000037980	NM_002836	20	2809356	2974007	p13
ENSG00000125787	GNRH2	20	2979275	2981398	p13
ENSG00000125901	Q9H273	20	2981711	2983825	p13
ENSG00000101405	OXT	20	3007273	3008169	p13
ENSG00000101200	AVP	20	3018193	3021065	p13
ENSG00000125819	Q9H4J2	20	3045263	3095542	p13
ENSG00000088899	NM_014731	20	3098280	3104214	p13
ENSG00000101205	C20orf116	20	3126019	3140302	p13
ENSG00000125877	ITPA	20	3145101	3159513	p13
ENSG00000088836	SLC4A11	20	3163070	3173717	p13
ENSG00000088854	Q9UFX9	20	3186799	3343211	p13
ENSG00000088812	ATRN	20	3406762	3582498	p13

Date : 2002-03-20 10:34:15 Help Desk / Suggestions

Internet

Tasks

1. Exploring features related to a gene

Find the gene report for the human MDM2 gene (MDM2_human).

How many transcripts are there predicted for this gene? What is the size of the longest predicted mRNA? How many exons does it comprise? How many amino acids does it code for?

With which protein does the MDM2 protein interact? In which species was MDM2 first identified?

Which Interpro domains does MDM2 contain?

On which chromosomal band is MDM2 located? On which contig in the golden path?

Does MDM2 match any mouse traces?

2. Exploring a region

Display the region between markers D6S1993 and D6S2096 in ContigView (*Hint: start from chromosome 6 on the homepage*).

Centre the display around the PRDM1 gene by clicking on this gene in "Overview". In "Detailed View", why are some genes above and some below the DNA contigs?

What is the nearest marker to the PRDM1 gene? How many synonyms does this marker have?

Zoom out to the 1Mb view and export a gene list for this region. (*Hint: Use the drop-down menu at the top of "Detailed View"*). Include Ensembl gene id, known gene external id, known gene external database name, gene description, InterPro short description.

Export a SNP list for chromosomal band 6q21.

3. Examining the supporting evidence for a gene prediction

Display the ENPP3 gene (or your own favourite gene) in ContigView.

Examine the evidence for the gene prediction. (*Hint: Collapse the chromosome view and overview; compare the collapsed and expanded view of the supporting evidence, e.g. Unigene – Note that max. 7 tracks can be displayed for each section; if there are more entries in the database the display may change on multiple re-loads.*)

Tick ‘Acembly Transcripts’, ‘Ensembl RefSeq’, ‘NCBI Genomescan’ and ‘NCBI Transcripts’ under ‘DAS sources’ on the yellow bar at the top of ‘Detailed View’. Compare the Ensembl gene prediction to these gene prediction models.

Click on the ENPP3 transcript (or your own gene) and examine the links / alignments given for this gene in GeneView. Note especially the RefSeq and SpTrEMBL entries. What do the Target %id and Query %id figures indicate? (*Hint: Consult the help page by clicking on ‘Help’ in the top right corner.*)

Follow the link ‘View supporting evidence’ for the transcript ENST00000263052. From there follow the link ‘View Evidence’ to TransView. Examine the evidence from the different databases.

4. Exploring the mouse genome with Ensembl

Hint: Use what you know about navigation in human Ensembl.

Find the ‘generic’ Ensembl homepage and select mouse. Bring up a ContigView display of chromosome 11 between 111.5 Mb and 111.6 Mb.

How many ‘known’ and ‘novel’ genes are predicted in this region? For one of the known genes, find some information about its function, and look at an entry for it in the MGI resource.

Can you find out anything about the possible function of the novel gene?

What supporting evidence was used to predict these genes?

Export a text file showing the gene id numbers, descriptions and chromosomal positions of the genes in this region.

Turn on the ‘Tile Path’ track to see BAC clones that have been mapped to the assembly. What is the difference between the blue bar in the middle of detailed view (labelled DNA (contigs)) - and the blue bar at the bottom (labelled Assembly Ctgs) *(Hint: bring up the help page for ContigView, and find the link to a page of mouse-specific information).*

Which human chromosome has a region of conserved synteny with this region of mouse chromosome 11? *(Hint: Jump to Syntenyview using the menus on the gold bar).* More about inter-species comparison is coming up later in the workshop.

Task answers

1. Enter the gene name into the text search box and follow the link to GeneView. The MDM2 gene has 8 transcripts. Click on the longest transcript. The length of the transcript is 1473 bp, 10 exons. Follow the 'View protein' link. The transcript codes for 491 aa. Follow the link from GeneView to Locuslink or OMIM. MDM2 was first identified in mouse, the protein binds to p53. Check the InterPro section in GeneView. The domains are ATP/GTP-binding site motif A (P-loop), RING finger, Zn-finger in ranbp and others, p53-associated protein (MDM2). Follow the link 'Genomic location'. MDM2 is located on 12q15. It is located on contig AC025423.32.1.150579 (see GeneView, Genomic Location, or mouse-over the contig in ContigView). Make sure mouse matches are ticked under 'Features' in ContigView. Yes, MDM2 matches mouse traces.

2. Start on the homepage and click on chromosome 6. Enter the markers in MapView. The genes above the contig track are in the forward orientation, the genes below are in the reverse orientation. Get marker information by mouse-over the closest marker. It is RH16128 (D6S1217E) and has 5 synonyms (Hint: zoom in first in order to distinguish single markers). Choose gene list from the drop-down menu at the top of 'Detailed View'. Enter the required data, chromosomal coordinates are filled in automatically. Choose the tab 'SNP list' and enter the chromosomal band plus the details of your choice.

3. No standard answer.

4. The quickest way to reach the region is to enter the coordinates in the browse chromosome boxes (you can use M or Mb) on the mouse homepage (http://www.ensembl.org/Mus_musculus/). There are 3 known genes and 1 novel gene. Get functional information via the brief description on the GeneView page and via links (e.g. to SwissProt and MGI). The protein and mRNA tracks in ContigView show the supporting evidence (each has a link to a database entry). Export using the pull-down menu on the gold bar. Turn on tile path from the Decorations pull-down menu on the gold bar. Click the red help icon in the top right to see the ContigView help page, and then look at the 'extra information on mouse displays'. The mouse sequence is an assembly of whole genome shotgun reads, not sequence from BAC clones. The WGS assembly contigs are shown at the bottom. For easier processing and annotation, Ensembl is presenting the final genome assembly in 5 Mb chunks.

Glossary

Golden Path term sometimes used for a genome-wide assembly of sequence contigs to make a single base sequence, including gaps, for each chromosome

Contig A stretch of contiguous (continuous) something. Confusingly, used in many different contexts. Most often, it means a stretch of continous DNA sequence. But can also be used to mean a continuous assembly of fragments, or of clones, without implying that their sequence is known

NCBI National Centre for Biotechnology Information. Part of the U.S. National Library of Medicine (NLM), National Institutes of Health (NIH).

EBI European Bioinformatics Institute, Hinxton. An outstation of the European Molecular biology Laboratory.

URLs

Ensembl http://www.ensembl.org

Map Viewer at NCBI http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/hum_srch

Human Genome Browser at UCSC http://genome.ucsc.edu