IMGC Bioinformatics Workshop

Genome Browsers

Strasbourg
05-11-05

Laurens Wilming

Genome Browsers

(modified by Laurens Wilming from Open Door Workshop module by Xosé M. Fernández, Bert Overduin and Michael Schuster)

Aims

- explain why it can be useful to look at the whole genome.
- discuss how genes and other features can be predicted and displayed.
- briefly present the main web-based genome browsers.
- using Ensembl, demonstrate some of the features and applications of genome browsers.
- introduce the BioMart data retrieval system.
- examples (may include location and structure of a known gene and its products; information about a defined chromosomal region; convenient export of selected information).

Introduction

Web-based genome browsers have been developed to make it easier to access comprehensive information about regions of human, mouse and other genomes and about their whole gene sets. They help you to:

- explore what is in a chromosomal region
- see features in and around a specific gene
- search & retrieve across the whole genome
- investigate genome organisation
- compare to other genomes

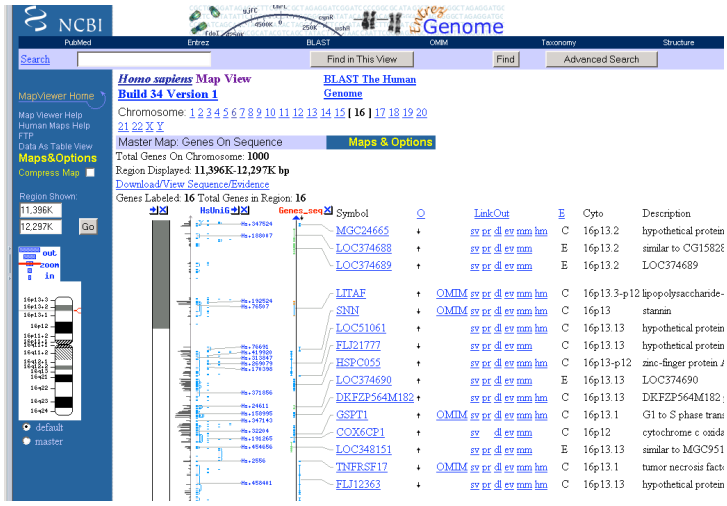
As well as the genomic sequence assembly itself, browsers attempt to show the location and structure of known genes and also to display predicted novel genes. They show information about the predicted mRNA transcripts and may also include information about the protein products. Information about genes is integrated with information about other genomic features (e.g. cytogenetic bands, markers, SNPs, repeated sequences, regions homologous to other species). Protein, mRNA and EST entries from the sequence databases may also be shown mapped onto the chromosomes.

In addition to providing annotation across the whole genome, browsers provide other resources. The browsers differ in what is provided and how it is presented. Resources that can be found include:

- **links** to other databases and resources
- **text Searching**
- **BLAST** and other sequence similarity searching
- **download** of genomic sequence, gene information and other data
- **data mining** facilities

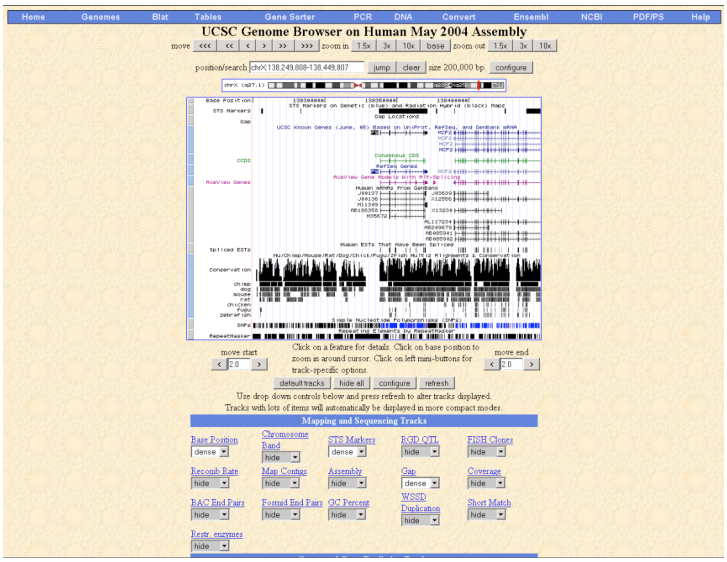
Browsers (and some of their strengths)

- **Map Viewer** – maintained by NCBI
  http://www.ncbi.nlm.nih.gov/mapview/
- **Human Genome Browser** – maintained by UCSC
  http://genome.ucsc.edu
- **G-integra Genome Browser** – maintained by DDBJ and JBIRC
  http://www.h-invitational.jp
- **Ensembl** – maintained by EBI / Sanger Institute
  http://www.ensembl.org



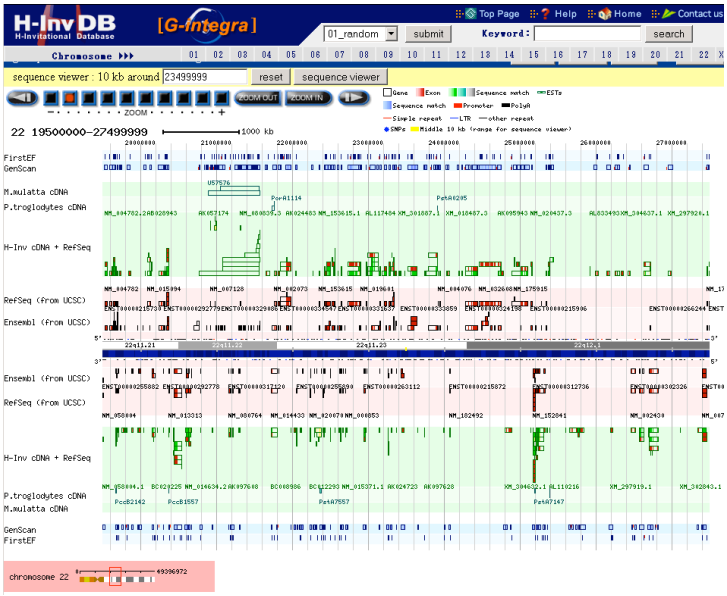
NCBI Map Viewer

- good integration with other NCBI resources



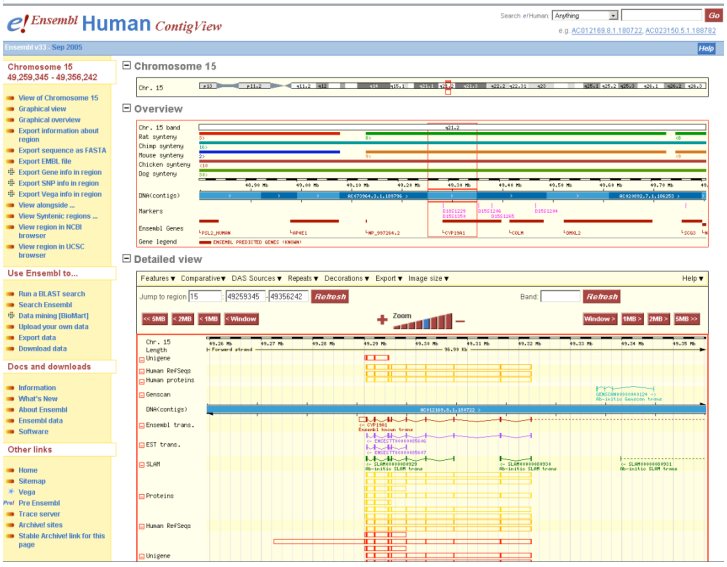
UCSC Genome Browser

- straightforward feature display
- old assemblies available
- wide range of tracks supplied by other groups



G-integra Genome Browser

- human gene database, with integrative annotation validated by full-length cDNA clones



Ensembl

- well-supported gene set with evidence
- range of different views
- easy retrieval of data sets
- archive available

While browsers can be very useful tools, they do not provide the definitive answer to every question!

Data retrieval and data mining

Genomic annotation data, due to its complexity and volume, does not lend itself to easy access. Presenting it on a web site is important, but so is providing simple but flexible ways to select and retrieve specific sets of data. NCBI has the Entrez query system and UCSC has its Table Browser.

In Ensembl, BioMart facilitates rapid retrieval of richly annotated gene lists and sequences and of SNP details, integrated with third party data and applications. Genes can be selected by chromosome region, protein domains, associated external identifiers or SNP properties, and these filters can be combined to group and refine biological data, including cross-species analyses, disease links, sequence variations and expression patterns.

BioMart is built upon a query-optimised relational database schema allowing quick and efficient access to this voluminous data through a user-friendly, interactive web interface. After selecting the biological object and the species, the results can be refined using a set of pre-defined filters. After each navigation event, the user is provided with immediate feedback on the number of matches found. Output can consist of annotated gene lists, gene structures, SNP details or various kinds of sequence sets. Output can be in HTML, text, Microsoft Excel and compressed formats.

Further reading

Birney, E *et al.* (2004)
An Overview of Ensembl.
Genome Research **14**: 925-928
http://www.genome.org/cgi/content/abstract/14/5/925

Hubbard, T *et al.* (2005)
Ensembl 2005.
Nucleic Acids Res **33**: D447-D453
http://nar.oupjournals.org/cgi/content/abstract/33/suppl_1/D447

Ashurst, JL *et al.* (2005)
The Vertebrate Genome Annotation (Vega) database.
Nucleic Acids Res **33**: D459-D465
http://nar.oupjournals.org/cgi/content/abstract/33/suppl_1/D459

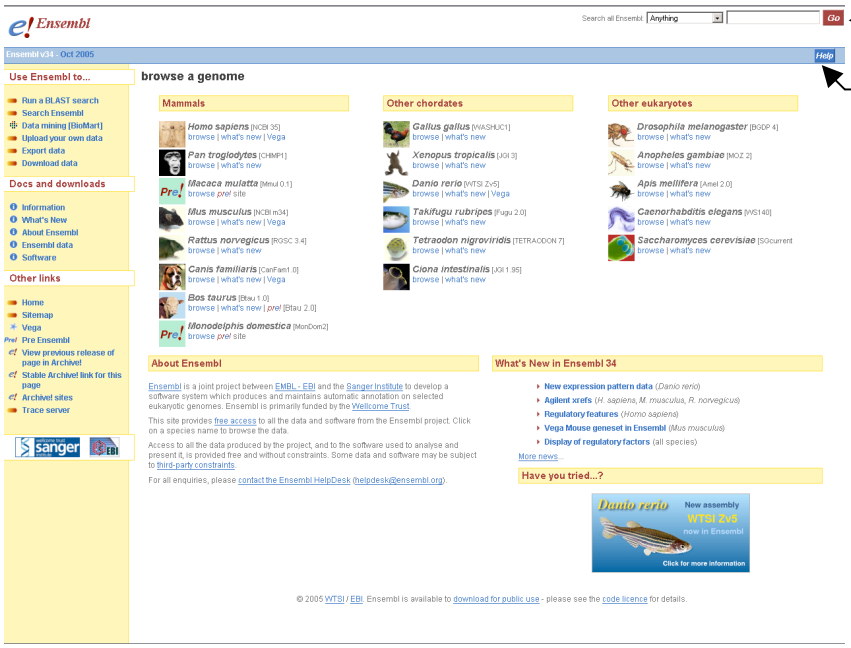
Karolchik, D *et al.* (2003)
The UCSC Genome Browser Database.
Nucleic Acids Res **31**: 51-54
http://nar.oupjournals.org/cgi/content/full/31/1/51

Imanishi, T *et al.* (2004)
Integrative Annotation of 21,037 Human Genes Validated by Full-Length cDNA Clones.
PLoS Biology **2**: 0856-0875
http://biology.plosjournals.org/perlserv/?request=get-document&doi=10.1371/journal.pbio.0020160

Dombrowski, SM and Maglott, D (2002)
Using the Map Viewer to Explore Genomes
In: The NCBI Handbook: 20-1-20-25
http://www.ncbi.nlm.nih.gov:80/books/bv.fcgi?rid=handbook.chapter.1562

How to access genomic information using Ensembl (worked example)

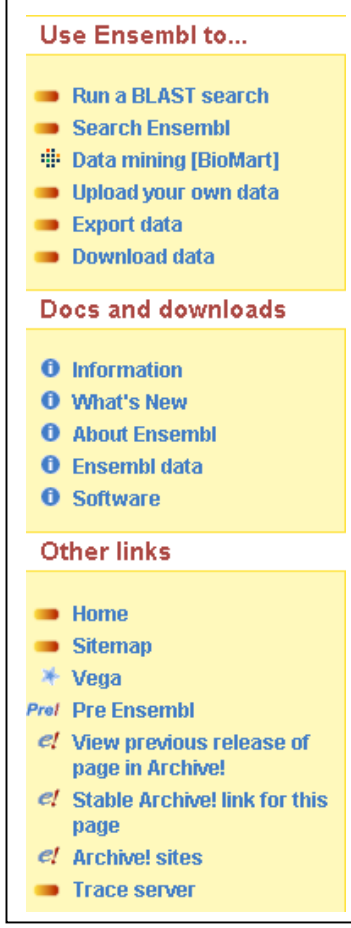
STEP 1:
Load **Ensembl**:
http://www.ensembl.org



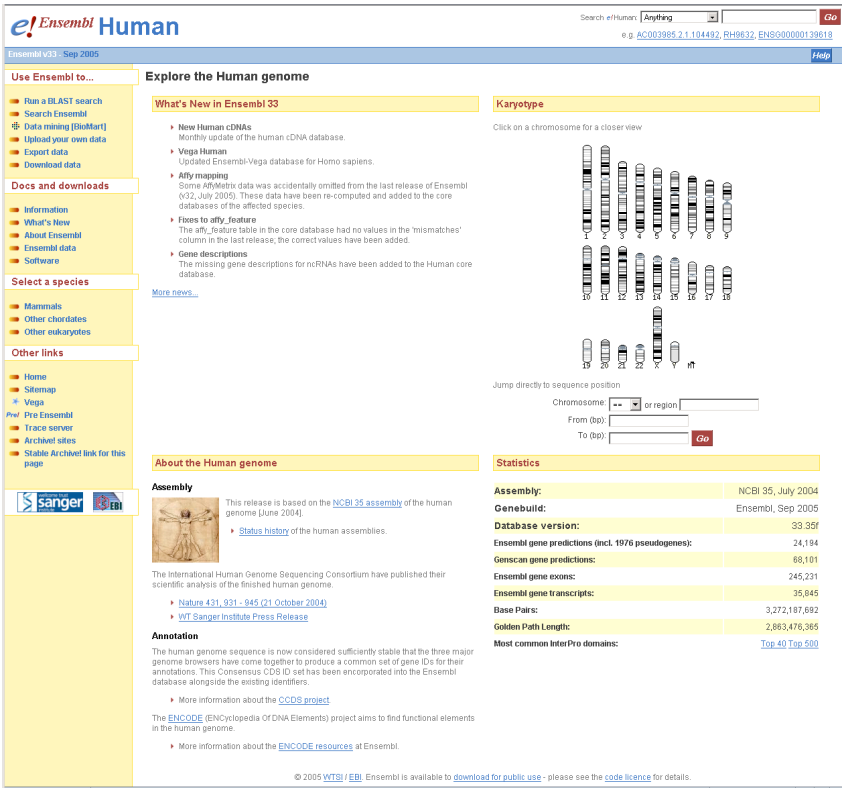
Multi-species text searching

Help

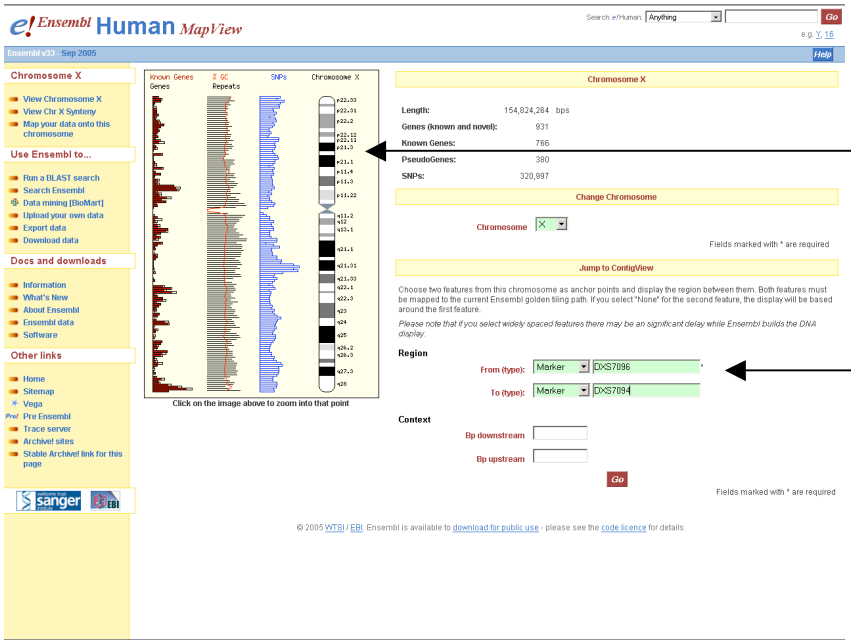
Sidebar navigation:



STEP 2:
Click [**browse**] under Homo sapiens



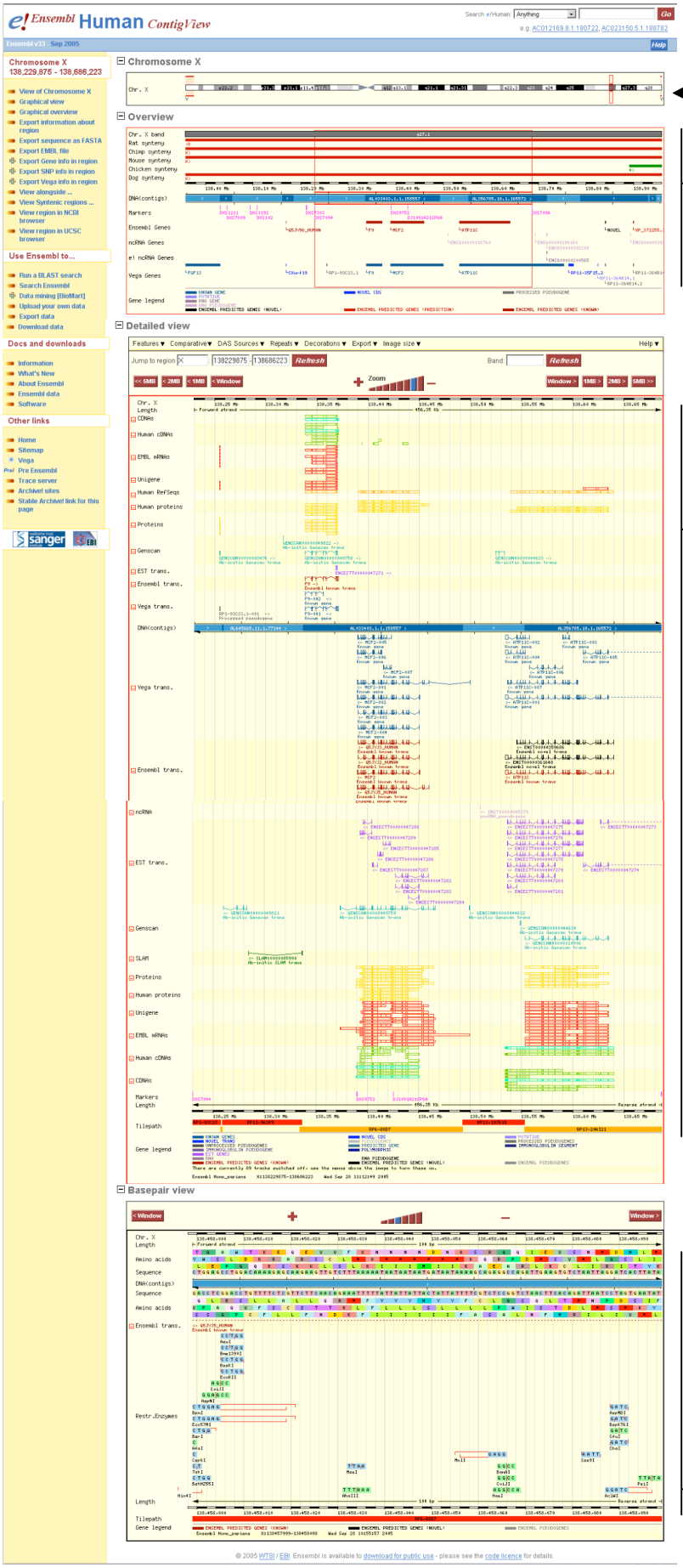
STEP 3:
Go to **MapView** by clicking on chromosome *X* in the Karyotype panel



Click to display that region

Display a region between any two features

STEP 4:
Go to **ContigView** choosing to display the region between *Markers* DXS7096 and DXS7094



Entire **Chromosome**

Overview shows a region of 1 Mb

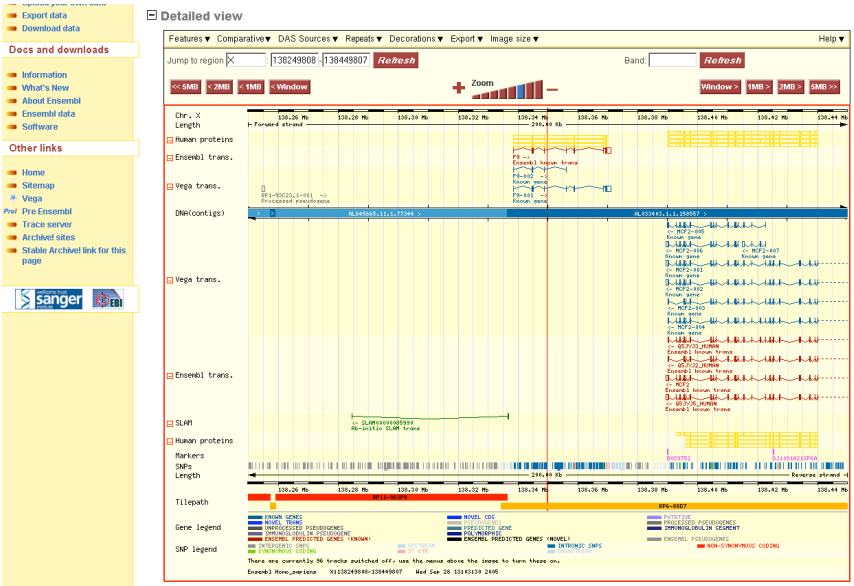
STEP 5:
Click on the F9 gene in Overview to center **Detailed view** on this gene and zoom in using the [+] button

Detailed view shows a region of 1 kb – 1 Mb

STEP 6:
Customise **Detailed view** using the *Features* drop-down menu:

- [x] SNPs
- [ ] Genotyped SNPs
- [x] tRNA
- [ ] CpG Islands
- [ ] Eponine Regions
- [ ] First EF
- [x] Markers
- [x] Vega genes
- [x] Ensembl genes
- [x] ncRNA genes
- [x] e! ncRNA genes
- [ ] EST genes
- [ ] Genscan
- [x] SLAM
- [ ] Proteins
- [x] Human proteins
- [ ] Human RefSeqs
- [ ] Other proteins
- [ ] Unigene
- [ ] EMBL mRNAs
- [ ] Human cDNAs
- [ ] CDNAs
- [ ] Human EST (ex.)
- [ ] AFFY HG-U110

Basepair view shows a region of 25-500 bp

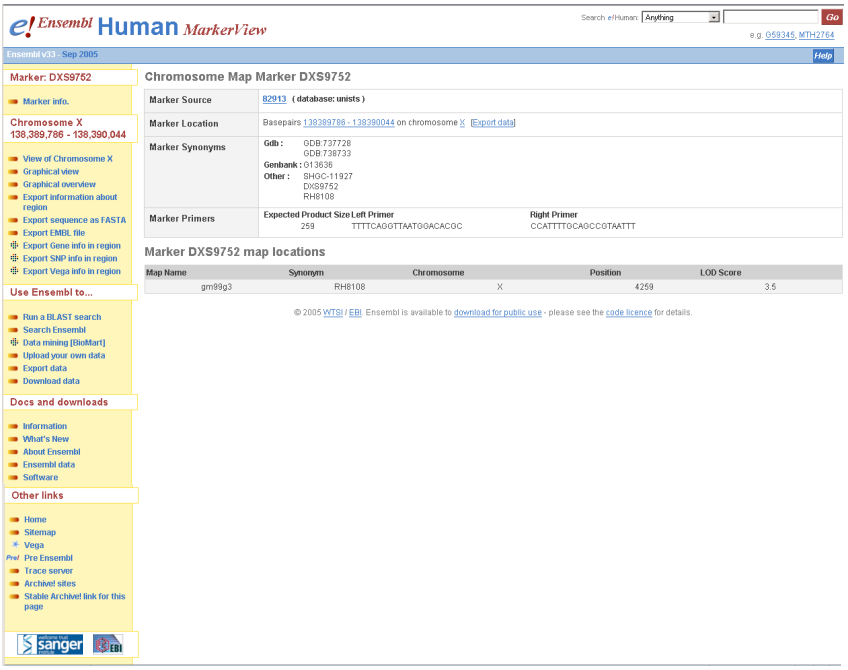


Features on forward (+) strand

Assembly

Features on reverse (-) strand

STEP 7:
Go to **MarkerView** by clicking on a marker in the Markers track and subsequently on *Marker info* in the pop-up menu



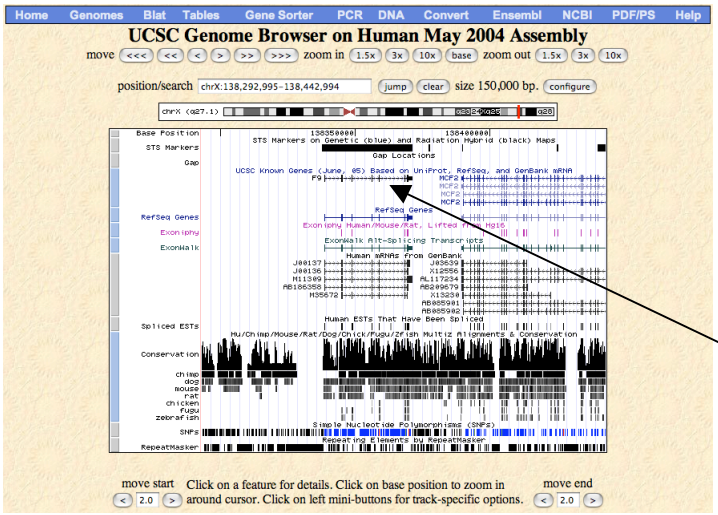
STEP 8:
Go back to **ContigView** via the sidebar:

Graphical view

Center again on the F9 gene and zoom out

View this region in the UCSC browser (which will open in a new browser window!)

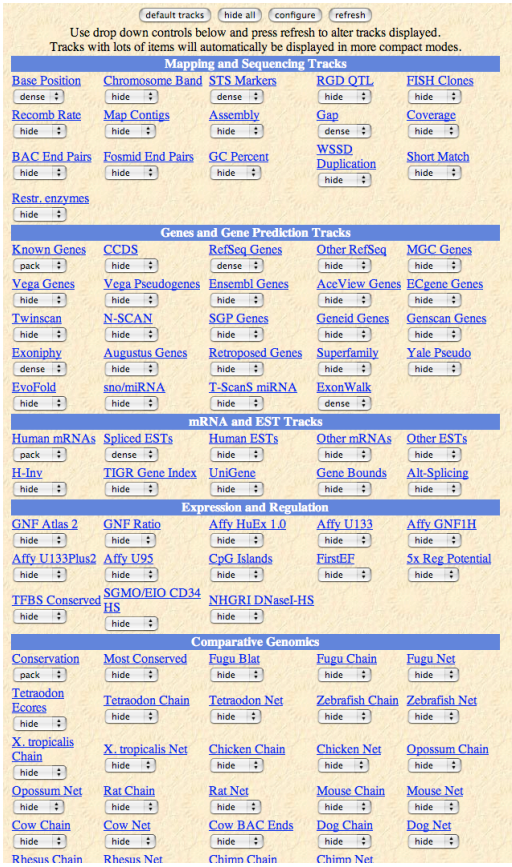
View region in UCSC browser



Navigation controls

Main track information panel

STEP 9:
Click on the F9 gene for all the information on that gene and links to other sources.



Track controls: hide, show, collapse, etc. numerous different tracks

Home Genomes Genome Browser Blat Tables Gene Sorter PCR FAQ Help

Human Gene F9 Description and Page Index

Description: coagulation factor IX
Alternate Gene Symbols: Factor IX, FIX, J00137, M11309
CCDS: CCDS14666.1
Representative Refseq: NM_000133 **Protein:** P00740 (aka FA9_HUMAN)
RefSeq Summary: This gene encodes vitamin K-dependent coagulation factor IX that circulates in the blood as an inactive zymogen. This factor is converted to an active form by factor XIa, which excises the activation peptide and thus generates a heavy chain and a light chain held together by one or more disulfide bonds. The role of this activated factor IX in the blood coagulation cascade is to activate factor X to its active form through interactions with Ca+2 ions, membrane phospholipids, and factor VIII. Alterations of this gene, including point mutations, insertions and deletions, cause factor IX deficiency, which is a recessive X-linked disorder, also called hemophilia B or Christmas disease.

Page Index	Quick Links	UniProt Comments	Sequence	Microarray	RNA Structure
Protein Structure	Other Species	GO Annotations	mRNA Descriptions	Pathways	Methods

Gene description

Quick Links to Tools and Databases

Genome Browser	Gene Sorter	Proteome Browser	Table Schema	UniProt	Entrez Gene
PubMed	OMIM	GeneLynx	GeneCards	HGNC	CGAP
Reactome	Stanford SOURCE	ExonPrimer	Ensembl	Jackson Labs	

External links

Comments and Description Text from UniProt (Swiss-Prot/TrEMBL)

ID: FA9_HUMAN
DESCRIPTION: Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor) (Plasma thromboplastin component) (PTC).
FUNCTION: Factor IX is a vitamin K-dependent plasma protein that participates in the intrinsic pathway of blood coagulation by converting factor X to its active form in the presence of Ca(2+) ions, phospholipids, and factor VIIIa.
CATALYTIC ACTIVITY: Selective cleavage of Arg-I-Ile bond in factor X to form factor Xa.
SUBUNIT: Heterodimer of a light chain and a heavy chain; disulfide-linked.
SUBCELLULAR LOCATION: Secreted.
TISSUE SPECIFICITY: Synthesized primarily in the liver and secreted in plasma.
DOMAIN: Calcium binds to the gamma-carboxyglutamic acid (Gla) residues and, with stronger affinity, to another site, beyond the Gla domain.
PTM: Activated by factor XIa, which excises the activation peptide.
DISEASE: Defects in F9 are the cause of recessive X-linked hemophilia B (HEMB) [MIM:306900]; also known as Christmas disease.
DISEASE: Mutations in position 43 (Oxford-3, San Dimas) and 46 (Cambridge) prevents cleavage of the propeptide, mutation in position 93 (Alabama) probably fails to bind to cell membranes, mutation in position 191 (Chapel-Hill) or in position 226 (Nagoya OR Hilo) prevent cleavage of the activation peptide.
PHARMACEUTICAL: Available under the names BeneFix (Baxter and American Home Products). Used to treat hemophilia B.
MISCELLANEOUS: In 1952, one of the earliest researchers of the disease, Dr. R.G. Macfarlane used the patient's surname, Christmas, to refer to the disease and also to refer to the clotting factor which he called the 'Christmas Factor' At the time Stephen Christmas was a 5-year-old boy. He died in 1993 at the age of 46 from acquired immunodeficiency syndrome contracted through treatment with blood products.
SIMILARITY: Belongs to the peptidase S1 family.
SIMILARITY: Contains 2 EGF-like domains.
SIMILARITY: Contains 1 Gla (gamma-carboxy-glutamate) domain.
DATABASE: NAME=HAEMB; NOTE=Hemophilia B database; FTP="ftp://ftp.ebi.ac.uk/pub/databases/haemb/".
DATABASE: NAME=BeneFix; NOTE=Clinical information on BeneFix; WWW="http://www.wyeth.com/products/benefix.asp".

Uniprot description

Microarray Expression Data

Expression ratio colors: red high/green low Submit

GNF Expression Atlas 2 Data from U133A and GNF1H Chips

fetal brain, whole brain, temporal lobe, parietal lobe, occipital lobe, prefrontal cortex, cingulate cortex, cerebellum, cerebellum peduncles, amygdala, hypothalamus, thalamus, subthalamic nucleus, caudate nucleus, globus pallidus, olfactory bulb, pons, medulla oblongata, superior cervical ganglion, trigeminal ganglion, ciliary ganglion, spinal cord, dorsal root ganglion, thymus, tonsil, lymph node, bone marrow, BM-CD71+ early erythroid, BM-CD33+ myeloid, BM-CD105+ endothelial, BM-CD34+, whole blood — Ratios / Absolute

PB-BDCA4+ dentritic cells, PB-CD14+ monocytes, PB-CD56+ NK Cells, PB-CD4+ T cells, PB-CD8+ T cells, PB-CD19+ B cells, leukemia lymphoblastic(molt4), 721 B lymphoblasts, leukemia promyelocytic(hl60), lymphoma Burkitts Raji, lymphoma Burkitts Daudi, leukemia chronic myelogenous(k562), colorectal adenocarcinoma, appendix, skin, adipocyte, fetal thyroid, thyroid, pituitary gland, adrenal gland, adrenal cortex, prostate, salivary gland, pancreas, pancreatic islets, atrioventricular node, heart, cardiac myocytes, skeletal muscle, tongue, smooth muscle, uterus, uterus corpus, trachea — Ratios / Absolute

bronchial epithelial cells, fetal lung, lung, kidney, fetal liver, liver, placenta, ovary, testis, testis Leydig cell, testis germ cell, testis interstitial, testis seminiferous tubule — Ratios / Absolute

Microarray data

Normal Human Tissue cDNA Microarrays

mRNA Secondary Structure of 3' and 5' UTRs

Region	Fold Energy	Bases	Energy/Base	Display As		
5' UTR	-1.50	29	-0.052	Picture	PostScript	Text
3' UTR	-365.20	1387	-0.263	Picture	PostScript	Text

The RNAfold program from the Vienna RNA Package is used to perform the secondary structure predictions and folding calculations. The estimated folding energy is in kcal/mol. The more negative the energy, the more secondary structure the RNA is likely to have.

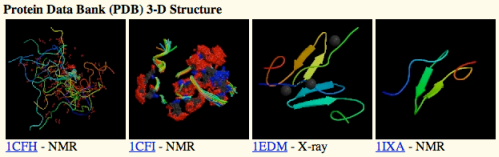
UTR info

Protein Domain and Structure Information

InterPro Domains: Graphical view of domain structure
IPR000152 - Aspartic acid and asparagine hydroxylation site
IPR000742 - EGF-like, subtype 2
IPR001881 - EGF-like calcium-binding
IPR001438 - Type II EGF-like signature
IPR006209 - EGF-like
IPR002383 - Coagulation factor, Gla region
IPR009003 - Peptidase, trypsin-like serine and cysteine proteases
IPR001254 - Peptidase S1 and S6, chymotrypsin/Hap
IPR001314 - Peptidase S1A, chymotrypsin
IPR000294 - Vitamin K-dependent carboxylation/gamma-carboxyglutamic region

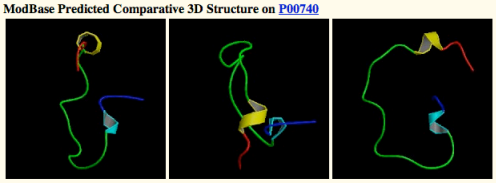
Pfam Domains:
PF00008 - EGF-like domain
PF00594 - Vitamin K-dependent carboxylation/gamma-carboxyglutamic (GLA) domain
PF00089 - Trypsin

Protein Data Bank (PDB) 3-D Structure



1CFH - NMR 1CFI - NMR 1EDM - X-ray 1IXA - NMR
To conserve bandwidth, only the images from the first 4 structures are shown.
1MGX - NMR 1RFN - X-ray

ModBase Predicted Comparative 3D Structure on P00740



Front Top Side

Protein domain and structure information

Homologous Genes in Other Species (BLASTP Best Hit)

Mouse	Rat	Zebrafish	D. melanogaster	C. elegans	S. cerevisiae
Genome Browser	Genome Browser	Genome Browser	Genome Browser	Genome Browser	No homolog
Gene Details	Gene Details		Gene Details	Gene Details	
Gene Sorter	Gene Sorter		Gene Sorter	Gene Sorter	
Jackson Lab	RGD	Ensembl	FlyBase	WormBase	
Protein Sequence	Protein Sequence		Protein Sequence	Protein Sequence	
Alignment	Alignment		Alignment	Alignment	

Homology/orthology

Gene Ontology (GO) Annotations with Structured Vocabulary

Molecular Function:
GO:0003803 coagulation factor IXa activity
GO:0004252 serine-type endopeptidase activity
GO:0004263 chymotrypsin activity
GO:0004295 trypsin activity
GO:0005509 calcium ion binding
GO:0008233 peptidase activity
GO:0016787 hydrolase activity

Biological Process:
GO:0006508 proteolysis and peptidolysis
GO:0007596 blood coagulation

Cellular Component:
GO:0005576 extracellular region
GO:0005615 extracellular space

Gene Ontology

Descriptions from all associated mRNAs.

J00137 - Human factor IX (Christmas factor) mRNA, complete cds.
J00136 - Human factor IX (Christmas factor) mRNA, complete coding sequence.
M11309 - Human coagulation factor IX mRNA, complete cds.
M35672 - Human coagulation factor IX mRNA, partial cds.
AB186358 - Homo sapiens F9 mRNA for coagulation factor IX, complete cds.

Matching mRNAs

Biochemical and Signaling Pathways

KEGG - Kyoto Encyclopedia of Genes and Genomes
hsa04610 - Complement and coagulation cascades - Homo sapiens

BioCarta from NCI Cancer Genome Anatomy Project
h_intrinsicPathway - Intrinsic Prothrombin Activation Pathway
h_plateletAppPathway - Platelet Amyloid Precursor Protein Pathway

Reactome (by CSHL, EBI, and GO)

Protein P00740 (Reactome details) participates in the following top level event(s):

109582 Hemostasis
163841 Post-translational modification of proteins

Pathways

STEP 10:
Go back and zoom in so that not more thyan 30kb is displayed, then click on anything in *Conservation* area to see cross-species genomic alignments

```
Alignment 10 of 189 in window, score 140455.0
Human       May 2004   (hg17)      chrX:138356013-138356172, strand +, size 160
Chimp       Nov. 2003  (panTro1)   chrX:143051699-143051858, strand +, size 160
Mouse       May 2004   (mm5)       chrX:51668588-51668758, strand +, size 171
Rat         June 2003  (rn3)       chrX:145623292-145623461, strand +, size 170
Dog         July 2004  (canFam1)   chrX:112565350-112565510, strand +, size 161
Chicken     Feb. 2004  (galGal2)   chr4:4805791-4805952, strand +, size 162
Fugu        Aug. 2002  (fr1)       chrUn:179830911-179831074, strand +, size 164

hg17.chrX     gctgtt--ac-tgtctatttt-----------gcttcttttagATGTAAC------ATGTAACATTAAGA
panTro1.chrX  gctgtt--ac-tgtctatttt-----------gcttcttttagATGTAAC------ATGTAACATTAAGA
mm5.chrX      gccatt--ac-ggttttttttttcctttctttctttttagATGCAAC------GTGTAACATTAAAA
rn3.chrX      gccattagat-tactttctt---ttcttttctttctttagATGCAAC------ATGTAACATTAAAA
canFam1.chrX  gctgtt--ac-tgttgatttt----------gcttctctttcagATGTAAC------ATGCAACATTAAGA
galGal2.chr4  gctttc--ag-catctcattt---------tgatttctttcagACTCTAC------TTGTGCTACGAAAA
fr1.chrUn     gatgcc--acacgtgactctt-----------gtcttttccagAGGTGTCCAAACAGTGTTCCGTCAACA
              *       *          **             * **  ***     *       **       ** *

hg17.chrX     ATGGCAGATGCGAGCAGTTTTGTA-AAAATAGTGCTGATAACAAGGTGGTTTGCTCCTGTACTGAGGGAT
panTro1.chrX  ATGGCAGATGCGAGCAGTTTTGTA-AAAATAGTGCTGATAACAAGGTGGTTTGCTCCTGTACTGAGGGAT
mm5.chrX      ATGGCAGGTGCAAGCAGTTTTGTA-AAAACAGTCCTGATAACAAGGTAATTTGTTCCTGCACTGAGGGAT
rn3.chrX      ATGGCAGGTGTAAGCAGTTTTGTA-AAAACAGCCCTGATAACAAGATAATTTGTTCCTGCACTGAGGGAT
canFam1.chrX  ATGGCAGATGCAAGCAGTTTTGTA-AATTGGGCCCCGATAACAAGGTGGTTTGTTCCTGTACTACGGGAT
galGal2.chr4  ATGGGGGCTGCGAGCACTTCTGCA-GGCATGACACACCACAGAAGGCTGTGTGCTCCTGCGCCCAGTGGCT
fr1.chrUn     ACGGCGGATGTTCCCATTTCTGTGTGATGCAG----GGGGACATAAGTGTGTGTCACTGTGCAGTTGGGC
              * **  * **    ** ** **                  * *      * **   ***  *    **

hg17.chrX     ATCGACTTGCAGAAAACCAGAAGTCCTGTGAACCAGCAGgt
panTro1.chrX  ATCGACTTGCAGAAAACCAGAAGTCCTGTGAACCAGCAGgt
mm5.chrX      ACCAACTTGCAGAAGACCAGAAGTCCTGTGAACCAACAGgt
rn3.chrX      ACCAACTTGCAGAAGACCAGAAGTCCTGTGAACCAGCAGgt
canFam1.chrX  ACCAACTTGCGGAAGACCAAAGGTCCTGTGAACCAGCAGgt
galGal2.chr4  ACAAGCTTCATGAAGACGGAAAGTCCTGTAAACCTGCAGgt
fr1.chrUn     ACCGATTGGGCCTGGACAAGAAGTCCTGCGAGCCAACAGgt
              *     *         **   * ******  * **  *****
```

Alignment of conserved sequences with colored exons.

STEP 11:
Go back and click on a matching mRNA for information on that sequence

Home Genomes Genome Browser Blat Tables Gene Sorter PCR FAQ Help

J00136

Information on mRNA J00136

Description: Human factor IX (Christmas factor) mRNA, complete coding sequence.
Gene: F9
Product: factor IX
Author: Choo,K.H., Gould,K.G., Rees,D.J. and Brownlee,G.G.
Organism: Homo sapiens
Tissue: liver
Development stage: n/a
Cell line: n/a
Sex: n/a
Library: n/a
Clone: n/a
CDS: 1..1389
Date: 1996-03-30
Version: 1
GeneLynx J00136
Stanford SOURCE: J00136
mRNA sequence: J00136

mRNA/Genomic Alignments

SIZE	IDENTITY	CHROMOSOME	STRAND	START	END	QUERY	START	END	TOTAL
1434	99.6%	X	+	138338444	138369798	J00136	1	1437	1437

View table schema

Data last updated: 2005-10-02

Description

The mRNA track shows alignments between human mRNAs in GenBank and the genome.

Matching mRNA data

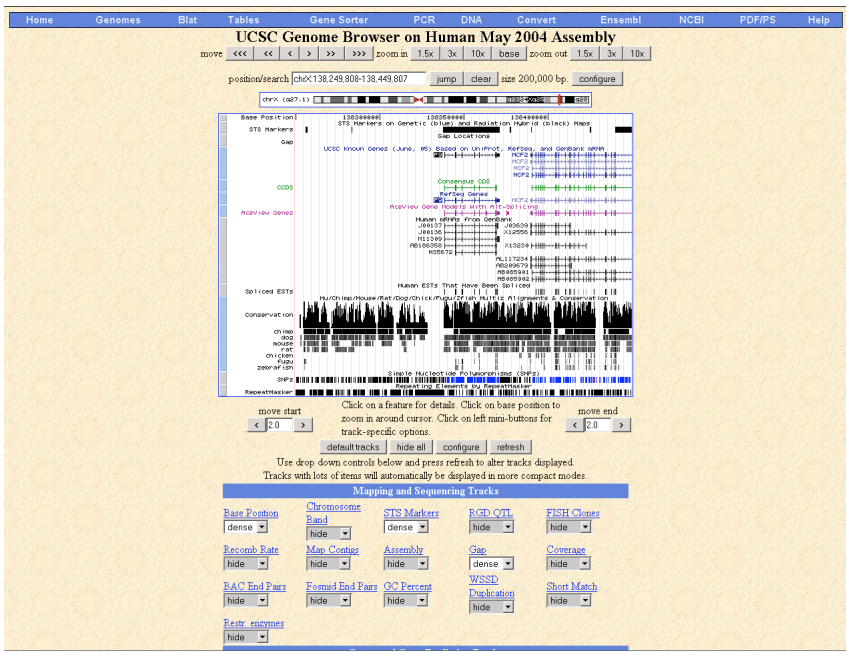
STEP 12:
Click on link to see alignments

```
atgagatctt taacattgcc aattaggtca gtggtcccaa gtagtcactt  138369143
agaaaatctg tgtatgtgaa atactgtttg tgacttaaaa tgaaatttat  138369193
ttttaatagG TGAACATAAT ATTGAGCAGA CAGAACATAC AGAGCAAAAG  138369243
CGAAATGTGA TTCGAATTAT TCCTCACCAC AACTACAATG CAGCTATTAA  138369293
TAAGTACAAC CATGACATTG CCCTTCTGGA ACTGACCAA CCCTTAGTGC  138369343
TAAACAGCTA CCTTACACCT ATTTGCATTG CTGACAAGGA ATACACCAAC  138369393
ATCTTCCTCA AATTTGGATC TGGCTATGTA AGTGGCTGGG gAAGAGTCTT  138369443
CCACAAAGGG AGATCAGCTT TAGTTCTTCA GTACCTTAGA GTTCCACTTG  138369493
TTGACCGAGC CACATGTCTT CGATCTACAA AGTTCACCAT CTATAACAAC  138369543
ATGTTCTGTG CTGGCTTCCA TGAAGGAGGT AGAGATTCAT GTCAAGGAGA  138369593
TAGTGGGGGA CCCCATGTTA CTGAAGTGGA AGGGACCAGT TTCTTAACTG  138369643
GAATTATTAG CTGGGGTGAA GAGTGTGCAA TGAAAGGCAA ATATGGAATA  138369693
TATACCAAGG TATCCCGGTA TGTCAACTGG ATTAAGGAAA AAACAAAGCT  138369743
CACTTAATGA AAGATGGATT TCCAAGGTTA ATTCATTGGA ATTGAAAATT  138369793
AACAGggcct ctcactaact aatcactttc ccatcttttg ttagatttga  138369843
atatatacat tctatgatca ttgctttttc tctttacagg ggagaatttc  138369893
atatt
```

Side by Side Alignment

```
000000001 atgcagcgcgtgaacatgatcatggcagaatcaccaagcctcatcaccat 000000050
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
138338444 atgcagcgcgtgaacatgatcatggcagaatcaccaggcctcatcaccat 138338493

000000051 ctgccttttaggatatctactcagtgctgaatgtacag 000000088
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
138338494 ctgccttttaggatatctactcagtgctgaatgtacag 138338531

000000089 tttttcttgatcatgaaaacgccaacaaaattctgaatcggccaaagagg 000000138
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
138344689 tttttcttgatcatgaaaacgccaacaaaattctgaatcggccaaagagg 138344738

000000139 tataattcaggtaaattggaagagtttgttcaagggaaccttgagagaga 000000188
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
138344739 tataattcaggtaaattggaagagtttgttcaagggaaccttgagagaga 138344788

000000189 atgtatggaagaaaagtgtagttttgaagaaccacgagaagtttttgaaa 000000238
>>>>>>>>> ||||||||||||||||||||||||||||||||||||||||||||||||||| >>>>>>>>>
138344789 atgtatggaagaaaagtgtagttttgaagaagcacgagaagtttttgaaa 138344838

000000239 acactgaaa 000000247
>>>>>>>>> ||||||||| >>>>>>>>>
138344839 acactgaaa 138344847
```

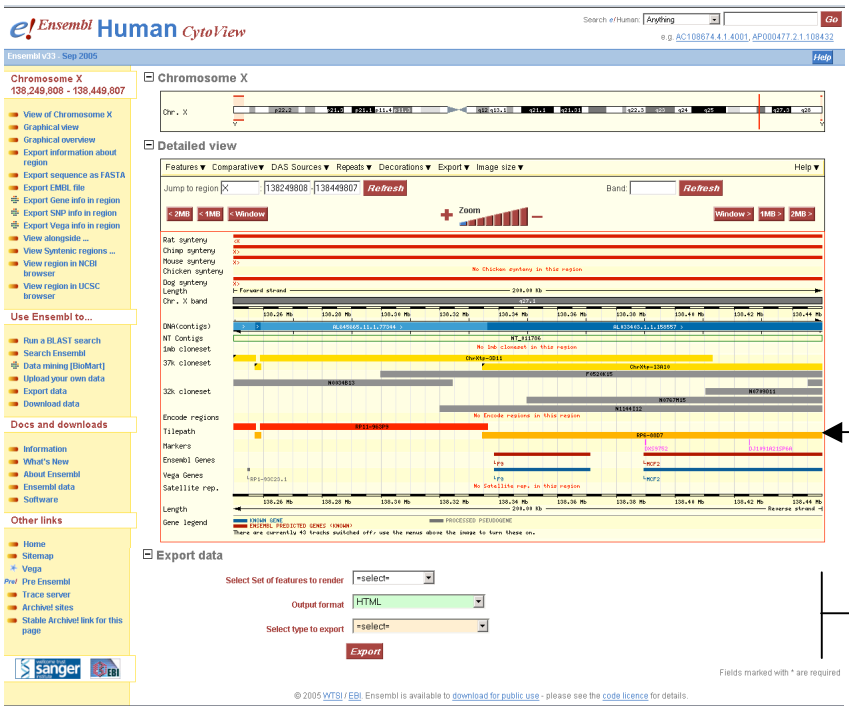
Matching mRNA alignment with color coded exon, intron and UTR sequence in genomic context



STEP 13:
Close the UCSC browser window

Go to **CytoView**:

Graphical overview



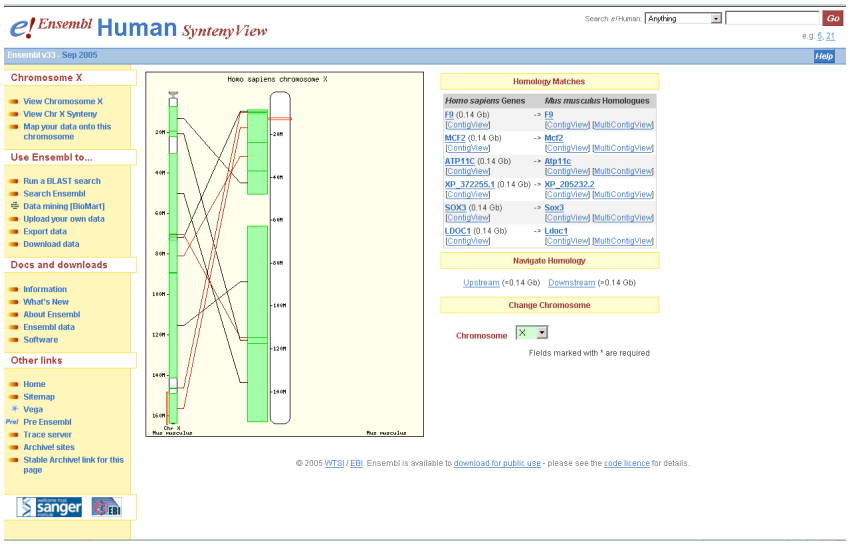
Tiling path

Export clones and features

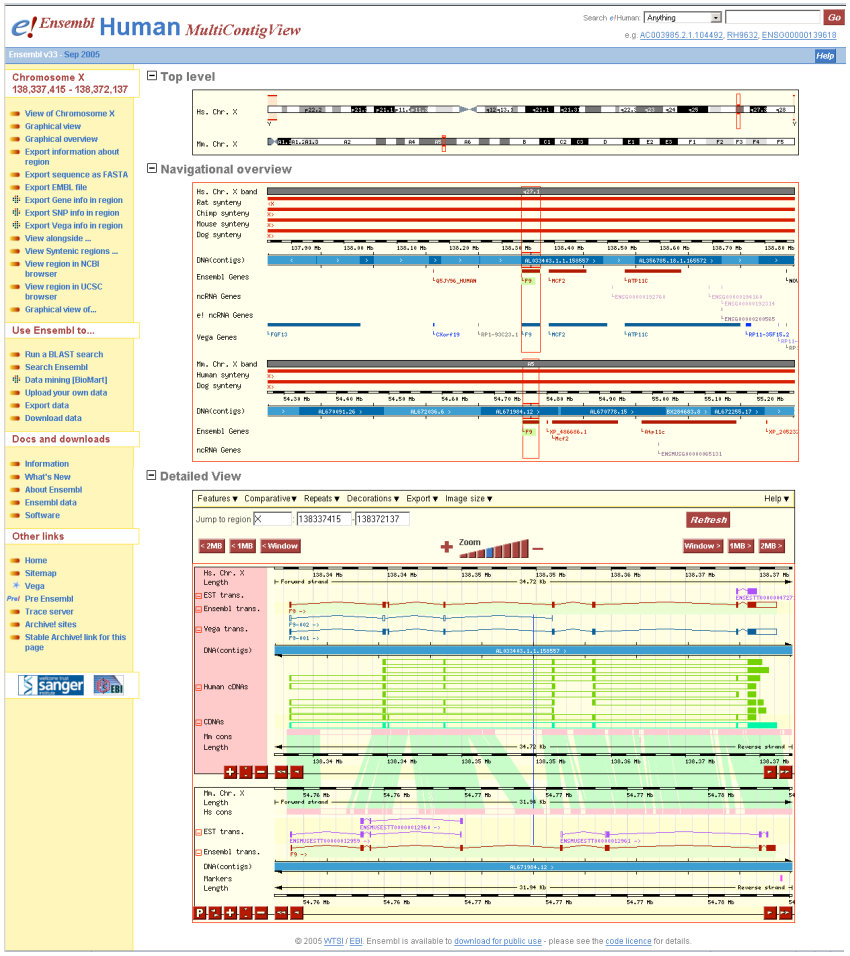
STEP 14:
Go to mouse **SyntenyView**:

View Syntenic regions ...

... with *Mus musculus*



STEP 15:
Go to human-mouse **MultiContigView** for the F9 gene using the [**MultiContigView**] link in the Homology Matches table



STEP 16:
Go back to **ContigView**:

Graphical view

Go to **GeneView** by clicking on the F9 gene in the Ensembl trans. track in **Detailed View** and subsequently in the pop-up menu on *Gene: ENSG00000101981*

Ensembl Human GeneView

Ensembl v33 - Sep 2005

ENSG00000101981

- Gene information
- Gene splice site image
- Gene variation info.
- Genomic sequence
- Export data
- Transcript information
- Exon information
- Peptide information

Chromosome X
138,338,415 - 138,371,137

- View of Chromosome X
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining (BioMart)
- Upload your own data
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- Trace server
- Archive sites
- Stable Archive link for this page

Ensembl Gene Report for ENSG00000101981

Gene	F9 (HGNC Symbol ID) (to view all Ensembl genes linked to the name click here) This gene is a member of the human CCDS set: CCDS14666
Ensembl Gene ID	ENSG00000101981
Genomic Location	This gene can be found on Chromosome X at location: 138,338,415-138,371,137 This start of this gene is located in Contig AL033403.1.1.158557.
Description	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor) (Plasma thromboplastin component) (PTC). Source: Uniprot/SWISSPROT P00740
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 ORF prediction or from Genscan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.

Transcripts: ENST00000218099 — ENSP00000218099 — F9 [Transcript info] [Exon info] [Peptide info]

Features ▼

Chr. X	136.33 Mb	136.34 Mb	136.35 Mb	136.36 Mb	136.37 Mb	136
Length	⊢ Forward strand		52.72 Kb			
Ensembl trans.		F9 (+) Ensembl known transc				
DNA(contigs)						
Length	136.33 Mb	136.34 Mb	136.35 Mb	52.72 Kb	136.37 Mb	136
				⊢ Reverse strand ⊣		

Orthologue Prediction — The following gene(s) have been identified as putative orthologues by reciprocal BLAST analysis:

Species	Type	dNdS	Gene identifier
Canis familiaris	UBRH	0.23373	ENSCAFG00000018998 (FA9_CANFA) [MultiContigView] [Align] Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor). [Source:Uniprot/SWISSPROT;Acc:P19540]
Bos taurus	UBRH	0.28398	ENSBTAG00000004083 (Q95ME8_BOVIN) [MultiContigView] [Align] Coagulation factor IX (Fragment). [Source:Uniprot/SPTREMBL;Acc:Q95ME8]
Danio rerio	UBRH		ENSDARG00000010097 (f9) [MultiContigView] [Align] coagulation factor IX [Source: RefSeq_peptide (NP_878288)]
Mus musculus	UBRH	0.18656	ENSMUSG00000031138 (F9) [MultiContigView] [Align] coagulation factor IX. [Source:MarkerSymbol;Acc:MGI:88389+]
Fugu rubripes	UBRH		SINFRUG00000142924 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Gallus gallus	UBRH		ENSGAL00000006513 (FA9_CHICK) [MultiContigView] [Align] Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor). [Source:Uniprot/SWISSPROT;Acc:Q8O4X8]
Xenopus tropicalis	UBRH		ENSXETG00000003472 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Rattus norvegicus	UBRH	0.25220	ENSRNOG00000003430 (Novel Ensembl prediction) [MultiContigView] [Align] No description
Tetraodon nigroviridis	UBRH		GSTENG00012000001 (GSTENG00012000001) [MultiContigView] [Align] No description
	RHS		GSTENG00031251001 (GSTENG00031251001) [MultiContigView] [Align] No description

View alignments of homologies.

UBRH = (U)nique (B)est (R)eciprocal (H)it
MBRH = one of (M)any (B)est (R)eciprocal (H)its
RHS = Reciprocal Hit based on Synteny around BRH
DWGA = Derived from Whole Genome Alignment

Disease Matches	This Ensembl entry corresponds to the following OMIM disease identifiers: **Hemophilia B (3)** • [Omim ID: 306900] - View disease information

Gene DAS Report

DAS Sources

- AltSplice (Alternative splice database)
- AltTrans (Alternative Transcript Diversity Database)
- ArrayExpress (Gene Expression Database)
- GAD (Genetic Association Database)
- HGNC (HUGO Gene Nomenclature Committee)
- HUGO_text (PubMed text-mining via HUGO symbol)
- Phenotypes (Associated directly or via orthologues or protein families)
- Protonet (Global classification of proteins into hierarchical clusters)
- Reactome (Knowledgebase of biological processes)
- UniProt (Protein knowledgebase)

Manage Sources

Transcript ENST00000218099

Transcript	F9 (HGNC Symbol ID) (to view all Ensembl genes linked to the name click here) This transcript is a member of the human CCDS set: CCDS14666
Transcript information	**Exons:** 8 **Transcript length:** 2,802 bps **Protein length:** 461 residues [Further Transcript info] [Exon information] [Protein information]
Similarity Matches	**This Ensembl entry corresponds to the following database identifiers:**

HGNC Symbol:	F9			
CCDS:	CCDS14666.1			
UniProt/Swiss-Prot:	FA9_HUMAN (Target %id: 100; Query %id: 100) [align]			
RefSeq peptide:	NP_000124.1 (Target %id: 100; Query %id: 100) [align]			
RefSeq DNA:	NM_000133.2 (Target %id: 99; Query %id: 99) [align]			
UniProt/TrEMBL:	Q5JYJ8_HUMAN (Target %id: 100; Query %id: 100) [align] Q7Z7N0_HUMAN (Target %id: 9; Query %id: 97) [align] Q7Z7N1_HUMAN (Target %id: 14; Query %id: 88) [align] Q96XR9_HUMAN (Target %id: 11; Query %id: 100) [align]			
EntrezGene:	2158			
EMBL:	AF536327 [align]	AL033403 [align]	AY226143 [align]	AY289424 [align]
	AY289425 [align]	J00136 [align]	J00137 [align]	K02046 [align]
	K02049 [align]	K02051 [align]	K02052 [align]	K02053 [align]
	K02402 [align]	M11309 [align]	M19063 [align]	M35672 [align]
	S66752 [align]	S66834 [align]		
IPI:	IPI00296176.2 (Target %id: 100; Query %id: 100)			
MIM:	134510	134540	306900	
PDB:	1CFH	1CFI	1EDM	1KA
	1MQX	1RFN		
Protein ID:	AAA51981.1 [align]	AAA52023.1 [align]	AAA52456.1 [align]	AAA52763.1 [align]
	AAA56822.1 [align]	AAA98726.1 [align]	AAB28588.1 [align]	AAB29758.1 [align]
	AAB59820.1 [align]	AAM98180.1 [align]	AAO73937.1 [align]	AAP34372.1 [align]
	AAP34373.1 [align]	CAA42103.1 [align]		
UniGene:	Hs.522798 (Target %id: 99; Query %id: 99)			
Affymx Microarray Focus:	207218_at			
Affymx Microarray HuGeneFL:	K02402_at			
Affymx Microarray U133:	207218_at	207218_at	207218_at	g10518507_3p_at
Affymx Microarray U95:	35556_at			

GO — **The following GO terms have been mapped to this entry via UniProt:**

- GO:0003803 [coagulation factor IXa activity] TAS
- GO:0004283 [chymotrypsin activity] IEA
- GO:0004295 [trypsin activity] IEA
- GO:0005509 [calcium ion binding] IEA
- GO:0005615 [extracellular space] IEA
- GO:0006508 [proteolysis and peptidolysis] IEA
- GO:0007596 [blood coagulation] TAS
- GO:0008233 [peptidase activity] IEA

InterPro

- IPR001438 Type II EGF-like signature - [View other genes with this domain]
- IPR001314 Peptidase S1A, chymotrypsin - [View other genes with this domain]
- IPR000294 Vitamin K-dependent carboxylation/gamma-carboxyglutamic region - [View other genes with this domain]
- IPR001881 EGF-like calcium-binding - [View other genes with this domain]
- IPR000742 EGF-like, subtype 2 - [View other genes with this domain]
- IPR002383 Coagulation factor, Gla region - [View other genes with this domain]
- IPR000152 Aspartic acid and asparagine hydroxylation site - [View other genes with this domain]
- IPR006209 EGF-like - [View other genes with this domain]
- IPR001254 Peptidase S1 and S6, chymotrypsin/Hap - [View other genes with this domain]

Protein Family	ENSF00000000622 : COAGULATION FACTOR This cluster contains 5 Ensembl gene member(s)
Transcript structure	⊢ Forward strand — 52.72 Kb
Protein features	Prints: GLA_blood, EGF_1, Peptidase_S1A Profile: P550034, EGF_2, EGF_like, Peptidase_S1_S6 Prosite: VitK_dep_GLA, EGF_Ca_bd, Asp_hydroxyl_S Pfam: EGF_like, Peptidase_S1_S6, VitK_dep_GLA Signal peptide, Peptide Scale (aa) 0 40 80 120 160 200 240 280 320 360 400 461

Genomic location

Transcript structure

Homologous genes in other species

Links to other databases

GO annotation

Protein domains

Find related proteins

Protein structure

STEP 17:
Go to **TransView**:

Transcript information
ENST00000218099

Ensembl Human TransView

Search e/Human: Anything Go

e.g. ENST00000333656, ENST00000328225

Ensembl v33 - Sep 2005 Help

ENST00000218099

- Gene info.
- Gene splice site image
- Gene variation info.
- Genomic sequence
- Exon info.
- Transcript info.
- Export data
- Peptide info.

Chromosome X
138,338,415 - 138,371,137

- View of Chromosome X
- Graphical view
- Graphical overview
- Export information about region
- Export sequence as FASTA
- Export EMBL file
- Export Gene info in region
- Export SNP info in region
- Export Vega info in region

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload your own data
- Export data
- Download data

Docs and downloads

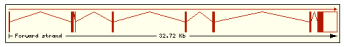
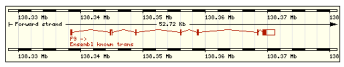
- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- Trace server
- Archive! sites
- Stable Archive! link for this page

sanger EBI

Ensembl Transcript Report

Transcript	F9 (HGNC Symbol ID) (to view all Ensembl genes linked to the name click here) This transcript is a member of the human CCDS set: CCDS14666
Ensembl Transcript ID	ENST00000218099
Transcript information	**Exons:** 8 **Transcript length:** 2,802 bps **Translation length:** 461 residues This transcript is a product of gene: ENSG00000101981
Genomic Location	This transcript can be found on Chromosome X at location: 138,338,415-138,371,137 This start of this transcript is located in Contig AL033403.1.1.158557.
Description	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor) (Plasma thromboplastin component) (PTC). Source: Uniprot/SWISSPROT P00740
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 ORF prediction or from Genscan exons supported by protein, cDNA and EST evidence. GeneWise models are further combined with available aligned cDNAs to annotate UTRs.
Similarity Matches	**This Ensembl entry corresponds to the following database identifiers:** **HGNC Symbol:** F9 **CCDS:** CCDS14666.1 **UniProt/Swiss-Prot:** F4S_HUMAN (Target %id: 100; Query %id: 100) [align] **RefSeq peptide:** NP_000124.1 (Target %id: 100; Query %id: 100) [align] **RefSeq DNA:** NM_000133.2 (Target %id: 99; Query %id: 99) [align] **UniProt/TrEMBL:** Q5JVJ8_HUMAN (Target %id: 100; Query %id: 100) [align] Q7Z7N0_HUMAN (Target %id: 9; Query %id: 97) [align] Q7Z7N1_HUMAN (Target %id: 14; Query %id: 98) [align] Q86XR9_HUMAN (Target %id: 11; Query %id: 100) [align] **EntrezGene:** 2158 **EMBL:** AF536327 [align] AL033403 [align] AY226143 [align] AY269424 [align] AY269425 [align] J00136 [align] J00137 [align] K02049 [align] K02049 [align] K02051 [align] K02052 [align] K02053 [align] K02402 [align] M11309 [align] M19063 [align] M35672 [align] S66752 [align] S68634 [align] **IPI:** IPI00296176.2 (Target %id: 100; Query %id: 100) **MIM:** 134510 134540 306900 **PDB:** 1CFH 1CFI 1EDM 1IXA 1M0X 1RFN **Protein ID:** AAA51981.1 [align] AAA52023.1 [align] AAA52456.1 [align] AAA52763.1 [align] AAA56822.1 [align] AAA98726.1 [align] AAB28588.1 [align] AAB29758.1 [align] AAB59620.1 [align] AAM96189.1 [align] AAO73937.1 [align] AAP34372.1 [align] AAP34373.1 [align] CAA42103.1 [align] **UniGene:** Hs.522798 (Target %id: 99; Query %id: 99) **Affymx Microarray Focus:** 207218_at **Affymx Microarray HuGeneFL:** K02402_at **Affymx Microarray U133:** 207218_at 207218_at 207218_at g10518507_3p_at **Affymx Microarray U95:** 35556_at
GO	**The following GO terms have been mapped to this entry via UniProt:** GO:0003803 [coagulation factor IXa activity] TAS GO:0004263 [chymotrypsin activity] IEA GO:0004295 [trypsin activity] IEA GO:0005509 [calcium ion binding] IEA GO:0005615 [extracellular space] IEA GO:0006508 [proteolysis and peptidolysis] IEA GO:0007596 [blood coagulation] TAS GO:0008233 [peptidase activity] IEA
InterPro	IPR001436 Type II EGF-like signature - [View other genes with this domain] IPR001314 Peptidase S1A, chymotrypsin - [View other genes with this domain] IPR000294 Vitamin K-dependent carboxylation/gamma-carboxyglutamic region - [View other genes with this domain] IPR001881 EGF-like calcium-binding - [View other genes with this domain] IPR000742 EGF-like, subtype 2 - [View other genes with this domain] IPR002383 Coagulation factor, Gla region - [View other genes with this domain] IPR000152 Aspartic acid and asparagine hydroxylation site - [View other genes with this domain] IPR006209 EGF-like - [View other genes with this domain] IPR001254 Peptidase S1 and S6, chymotrypsin/Hap - [View other genes with this domain]
Protein Family	ENSF00000000622 : COAGULATION FACTOR This cluster contains 5 Ensembl gene member(s)
Transcript structure	
Transcript neighbourhood	
Transcript sequence	1 ACCACTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAAT 61 CACCAGGCCTCATCACCATCTGCCTTTTAGGATATCTACTCAGTGCTGAATGTACAGTTT 121 TTCTTGATCATGAAAACGCCAACAAAATTCTGAATCCGGCCAAAGAGGTATAATTCAGGTA 181 AATTGGAAGAGTTTGTTCAAGGGAACCTTGAGAGAGAATGTATGGAATGAAAAGTGTAGTT 241 TTGAAGAAGCACGAGAAGTTTTGAAAACACTGAAAGAACAACAGTGAATTTGGAAGCAGT 301 ATGTTGATGGAGATCAGTGTGAGTCCAATCCATGTTTAAATGGCGGCAGTTGCAAGGATG 361 ACATTAATTCCTATGAATGTTGGTGTCCCTTTGGATTTGAAGGAAAGAACTGTGAATTAG 421 ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGCTGATA 481 ACAAGGTGGTTGCTCCTGTACTGAGGGATATCCGACTTGCAGAAAACCAGAAGTCCTGTG 541 AACCAGCAGTGCCATTTCCATGTGGAAGAGTTTCTGTTTCATCAAACTTCTAAGCTCACCC 601 GTGCTGAGACTGTTTTCCTGATGTGGACTATGTAAATTCTACTGAAGCTGAAACCATTT 661 TGGATAACATCACTCAAAGCACCCAATCATTTAATGACTTCACTCGGCTTGTTGGTGGAG 721 AAGATGCCAAACCAGGTCAATTCCCTTGGCAGGTTGTTTTGAATGGTAAATTTGATGCAT 781 TCTGTGGAGGCTCTATCGTTAATGAAAAATGGATTGTAACTGCTGCCCACTGTGTTGAAA 841 CTGGTGTTAAAATTACAGTTGTCGCAGGTGAACATAAATATTGAGGAGACAGAACATACAG 901 AGCAAAAGCCGAAATGTGATTCGATTATTCCTCACCACAACTACAATGCAGCTATTAATA 961 AGTACAACCATGACATTGCCCTTCTGGAACTGGACGAACCCTTATGTCTAAACAGCTACG 1021 TTACACCTATTTGCATTGCTGACAAGGAATACACGAACATCTTCCTCAAATTTGGATCTG 1081 GCTATGTAAGTGGCTGGCGGAAGACTCTTCCACAAAAGGGAATCAGCTTTAGTTTCTCAGT 1141 ACCTTAGAGTTCCACTTGTTGACCGAGCCACATGTCTTCGATCTACAAGTTCACCATCT 1201 ATAACAACATGTTCTGTGCTGGCTTCCATGAAGGAGGTAGAGATTCATGTCAAGGAGATA 1261 GTCGGCGACCCCATCGTACTGAAGTGCGAAGGCGACCAGTTTCTTAACTGCAATTATTAGCT 1321 GGGGTGAAGAGTGTGCAATGAAAGGCAAATATGGAATATATACCAAGGTATCCCGGTATG 1381 TCAACTGGATTAAGGAAAAACAAAGCTCACTTAATGAAAGATGGATTTCCAAGGTTAAT 1441 TCATTGGAATTGAAAATTAACAGCGCCTCTCACTAACTAATCACTTTCCCATCTTTTCTT 1501 AGATTTGAATATATACATTCTATGATCATTGCTTTTTCTCTTTACAGGGGAGAATTTCAT 1561 ATTTTACCTGAGCAAATTGATTAGAAAATGGAACCACTAGAGGAATATAATGTGTTAGGA 1621 AATTACATCATTTCTAAGGGCCCAGCCCCTCACAAAATTCTGAAGTTAATTCTCCACT 1681 CTGTCCATCAGATACTATGGTTCTCCACTATGGCAACTAACTCACTCAATTTTCCCTCCT 1741 TAGCAGCATTCCATCTTCCCGATCTTCTTTGCTTCTCCAACCAAAACATCAATGTTTATT 1801 AGTTCTGTATACAGTACAGGAGTCTTGGTCTACTCTATCACAGCGCCAGTACCACACTCA 1861 TGAAGAAAGAACACAGGAGTAGCTGAGAGGCTAAAACTCATCAAAAACATACTCCTTTT 1921 CCTCTACCCTATTCCTCAATCTTTTACCTTTTCCAAATCCCCAATCCCCAAATCAGTTTTT 1981 CTCTTCTTACTCCCTCTCTCCCTTTACCCTCCATGGTCGTTAAAGGAGAGATGGGGAG 2041 CATCATTCTGTTATACTTCTGTACACAGTTATACATGTCTATCAAACCCAGACTTGCTTC 2101 CGTAGTGGAGACTTGCTTTTCAGAACATAGGGATGAAGTAAGTGCCTGAAAAGTTTGGG 2161 GGAAAAGTTTCTTTCAGAGAGTTAAGTTATTTATATATATATATATATATATATATATA 2221 TAATATACAATATATAGTGTGTGTGTATGCGTGTGTGTAGACACACACGCATAC 2281 ACACATATAATGGAAGCAATAAGCCATTCTAAGAGCTTGTATGGTTATGGAGGCTCGACT 2341 AGGCATTTTCACGAAGGCAAGAATTGGCATATCATGTGAATTAAAAAGCTGAACATTGA 2401 CCCAGACATATTGTACTCTTTCTAAAAATAATAATAATAATGCTAACAGAAAGAAGAGAA 2461 CCGTTCGTTTGCAATCTACAGCTAGTAGAGACTTTGAGCGAAGAATTCAACAGTGTGTCTT 2521 CAGCAGTGTTCAGAGCCAAGCAAGAGTTGAAGTTGCCTAGACCAGAGAGAGATAATATC 2581 ATGTCTCCTTTAACTAGCATACCCCGAAGTGGAGAAGGGTGCAGCAGGCTCAAAGGCATA 2641 AGTCATTCCAATCAGCCAACTAAGTTGTCCTTTTCTGGTTTCCTGTTCACCATGGAACAT 2701 TTTGATTATAGTTAATCCTTATCTTCTAGAGAGTTTGCTGACCAACTGACGT 2761 ATGTTTCCCTTTGTGAATTAATAAACTGGTGTTCTGGTTCAT Show the following features: Exons Number residues: Yes Refresh

© 2005 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

Transcript sequence

Customise transcript sequence

STEP 18:
Go to **ExonView**:

Exon info.

Ensembl Human ExonView

Ensembl v33 - Sep 2005

ENST00000218099

Gene info.
Gene splice site image
Gene variation info.
Genomic sequence
Exon info.
Transcript info.
Export data
Peptide info.

Chromosome X
138,338,415 - 138,371,137

View of Chromosome X
Graphical view
Graphical overview
Export information about region
Export sequence as FASTA
Export EMBL file
Export Gene info in region
Export SNP info in region
Export Vega info in region

Use Ensembl to...

Run a BLAST search
Search Ensembl
Data mining [BioMart]
Upload your own data
Export data
Download data

Docs and downloads

Information
What's New
About Ensembl
Ensembl data
Software

Other links

Home
Sitemap
Vega
Pre Ensembl
Trace server
Archive sites
Stable Archive! link for this page

Ensembl Exon Report

Transcript	F9 (HGNC Symbol ID)(to view all Ensembl genes linked to the name click here) This transcript is a member of the human CCDS set: CCDS14666
Ensembl Transcript ID	ENST00000218099
Transcript information	Exons: 8 Transcript length: 2,802 bps Translation length: 461 residues This transcript is a product of gene: ENSG00000101981
Genomic Location	This transcript can be found on Chromosome X at location: 138,338,415-138,371,137 This start of this transcript is located in Contig AL033403.1.1.158557.
Description	Coagulation factor IX precursor (EC 3.4.21.22) (Christmas factor) (Plasma thromboplastin component) (PTC). Source: UniprotSWISSPROT P00740
Rendering options	Flanking sequence at either end of transcript Intron base pairs to show at splice sites Show full intronic sequence Show exons only Go

Exon Information

No.	Exon / Intron	Chr	Strand	Start	End	Length	Sequence
	5' upstream sequence						.....ccagaagtaaatacagctcagcttgtactttggtacaactaatcgacctt
1	ENSE00000677288	X	1	138,338,415	138,338,531	117 bp	ACCACTTTTCACAATCTGCTAGCAAAGGTTATGCAGCGCGTGAACATGATCATGGCAGAAT CACCAGGCCTCATCACCATCTGCCTTTTAGGATATCTACTCAGTGCTGAATGTACAG
	Intron 1-2	X	1	138,338,532	138,344,688	6,157 bp	gtttgtttccttttttaaaatacat.....aaagaattattcttttacatttcag
2	ENSE00000677287	X	1	138,344,689	138,344,852	164 bp	TTTTTCTTGATCATGAAAACGCCAACAAAATTCTGAATCGGCCAAAGAGGTATAATTCAG GTAAATTGGAAGAGTTTGTTCAAGGGAACCTTGAGAGAGAATGTATGGAAGAAAAGTGTA GTTTTGAAGAAGCACGAGAAGTTTTTGAAAACACTGAAAGAACA
	Intron 2-3	X	1	138,344,853	138,345,040	188 bp	gtgagtatttccacataatacccctt.....atttgtcttcttttttattcttttatag
3	ENSE00001173335	X	1	138,345,041	138,345,065	25 bp	ACTGAATTTTGGAAGCAGTATGTTG
	Intron 3-4	X	1	138,345,066	138,348,754	3,689 bp	gtaagcaattcattttatcctctag.....ttcaatttcttaacctatctcaaag
4	ENSE00001173327	X	1	138,348,755	138,348,868	114 bp	ATGGAGATCAGTGTGAGTCCAATCCATGTTTTAAATGGCGGCAGTTGCAAGGATGACATTA ATTCCTATGAATGTTGGTGTCCTTTGGATTTGAAGGAAAGAACTGTGAATTAG
	Intron 4-5	X	1	138,348,869	138,356,041	7,173 bp	gtaagtaactattttttgaatactc.....ttactgtctattttgctctcttttag
5	ENSE00001173320	X	1	138,356,042	138,356,170	129 bp	ATGTAACATGTAACATTAAGAATGGCAGATGCGAGCAGTTTTGTAAAAATAGTGCTGATA ACAAGGTGGTTTGCTCCTGTACTGAGGGATATCGACTTGCAGAAAACCAGAGATGCCTGTG AACCAGCAG
	Intron 5-6	X	1	138,356,171	138,358,740	2,570 bp	gtcataatctgaataagatttttta.....actaatttttcttctatttttctag
6	ENSE00001173315	X	1	138,358,741	138,358,943	203 bp	TGCCATTTCCATGTGGAAGAGTTTCTGTTTCACAAACTTCTAAGCTCACCCGTGCTGAGA CTGTTTTTCCTGATGTGGACTATGTAAATTCTACTGAAGCTGAAACCATTTTGGATAACA TCACTCAAACGACCCAATCATTTAATGACTTCACTCGGGTTGTTGGTGGAGAAGATGCCA AACCAGCTCAATTTCCCTTGGCAG
	Intron 6-7	X	1	138,358,944	138,368,419	9,476 bp	gtactttatactgatggtgtgtcaa.....aagagtcttaatttgttttcacag
7	ENSE00001173311	X	1	138,368,420	138,368,534	115 bp	GTTGTTTTTGAATGGTAAAGTTGATGCATTCTGTGGAGGCTCTATCGTTAATGAAAAATGG ATTGTAACTGCTGCCCACTGTGTTGAAACTGGTGTTAAAATTACAGTTGTCGCAG
	Intron 7-8	X	1	138,368,535	138,369,202	668 bp	gtaaatacacagaaagaataatast.....ttaaaatgaaatttattttttaatag
8	ENSE00001029145	X	1	138,369,203	138,371,137	1,935 bp	GTGAACATAATATTGAGGAGACAGAACATACAGAGCAAAAGCGAAATGTGATTCGAATTA TTCCTCACCAACAACTACAATGCAGCTATTAATAAGTACAACCATGACATTGCCGCTTCTGG AACTGGACGAACCCTTAGTGCTAAACAGCTACGTTACACCTATTTGCATTGCTGACAAGG AATACACGAACATCTTCCTCAAATTTGGATCTGGCTATGTAAGTGGCTGGGGAAGAGTCT TCCACAAAGGGAGATCAGCTTTAGTTCTTCAGTACCTTAGAGTTCCACTTGTTGACCGAG CCACATCTCTTGGATCTACAAAGTTCACCATCTATAACAACATGTTCTGTGCTGCTTCC ATGAAGGAGGTAGAGATTCATGTCAAGGAGATAGTGGGGGACCCCATGTTACTGAAGTGG AAGGGACCAGTTTCTTAACTGGAATTATTAGCTGGGGTGAAGAGTGTGCAATGAAAGGCA AATATGGAATATATACCAAGGTATCCCGGTATGTCAATCGGATTAAGGAAAAAACAAAGC TCACTTAATCAAAAGATGGATTTCCAAGGTTAATTCATTGGAATTGAAAATTAACAGGGCC TCTCACTAACTAATCACTTTCCCATCTTTTGTTAGATTTGAATATATACATTCTATGATC ATTGCTTTTTCTCTTTACAGGGGAGAATTTCATATTTTACCTGAGCAAATTGATTAGAAA ATGGAACCACTAGAGGAATATAATGTGTTAGGAAATTACAGTCATTTCTAAGGGCCCAGC CCTTGACAAAATTGTGAAGTTAAATTCTCCACTCTGTCCATCAGATACTATGGTTCTCCA CTATGGCAACTAACTCACTCAATTTTCCCTCCTTAGCAGCATTCCATCTTCCCGATCTTC TTTGCTTCTCCAACCAAAACATCAATGTTTATTAGTTCTGTATACAGTACAGGATCTTTG GTCTACTCTATCACAAGCCACGTACCACACTCATGAAGAAAGAACACAGGAGTAGCTGAG AGGCTAAAACTCATCAAAAACACTACTCCTTTTCCTCTACCCTATTCCTCAATCTTTTAC CTTTTCCAAATCCCAATCCCCAAATCAGTTTTTCTCTTTCTTACTCCCTCTCTCCCTTTT ACCCTCATGGTCGTTAAAGGAGAGATGGGGAGCATCATTCTGTTATACTTCTGTACACA GTTATACATGTCTATCAAACCCAGACTTGCTTCCGTAGTGGAGACTTGCTTTTCAGAACA TAGGGATGAAGTAAGGTGCCTGAAAAGTTTGGGGGAAAAGTTTCTTTCAGAGAGTTAAGT TATTTTATATATATAATATATATATAAAATATATAATATACAATATAAATATATATGTGTG TGTCTATGCCTGTTGTTACACACACACGCATACACACATATAATGGAGCAATAAGCCAT TCTAAGAGCTTGTATGGTTATGGAGGTCTGACTAGGCATGATTTCACGAAGGCAAGATTG GCATATCATTGTAACTAAAAAAGCTGACATTGACCCAGACATATTGTACTCTTTTCTAAAA ATAATAATAATAATGCTAACACAAAGAAGAAGAACCGTTCGTTTGCAATCTACAGCTAGTA GAGACTTTGAGGAAGAATTCAACAGTGTGTCTTCAGCAGTGTTCAGAGCCAAGCAAGAAG TTGAAGTTGCCTAGACCAGAGGGACATAAGTATCATGTCTCCTTTAACTAGCATACCCCGA AGTGGAGAAGGGTGCAGCAGGCTCAAAGGCATAAGTTCATTCAATCAGCCAACTAAGTTG TCCTTTTCTGGTTCGTGTTCACCATGAAACATTTTGATATAGTTAATCCTTCTATCTTT GAATCTTCTAGAGAGTTGCTGACCAACTGACGTATGTTTCCCTTTGTGAATTAATAAACT GGTGTTCTGGTTCAT
	3' downstream sequence						accttggctttttgtggattccattgatgtgaatcagtcaccctgtattt.....

Support Evidence

The supporting evidence below consists of the sequence matches on which the exon predictions were based and are sorted by alignment score.

There are a large number of supporting evidence hits for this transcript. Only the top ten 10 hits have been shown. Click to view all 32 supporting evidence hits.

© 2005 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

Customise exon-intron-flanking sequence information

Exon and flanking sequences:

- flanking sequence
- UTR
- exon
- intron

Supporting evidence used to build the gene

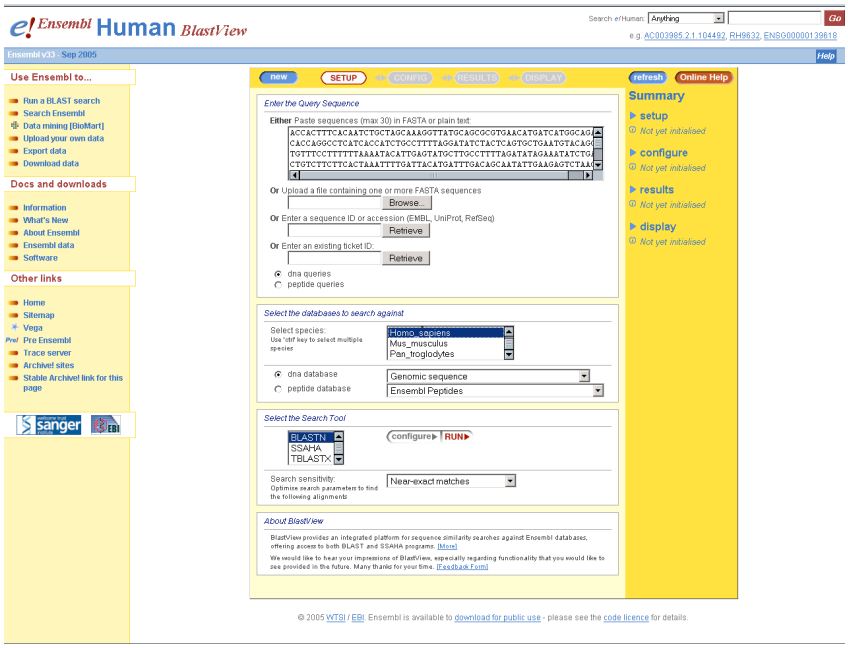
STEP 19:
Export the genomic sequence of the F9 gene in FASTA format:

Export sequence as FASTA

STEP 20:
Copy part of the sequence

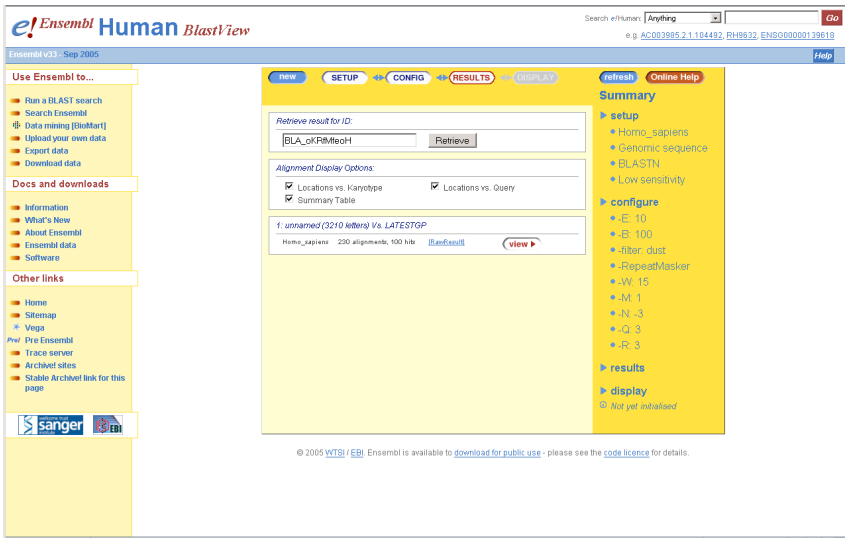
STEP 21:
Go to **Ensembl BLAST**:

Run a BLAST search



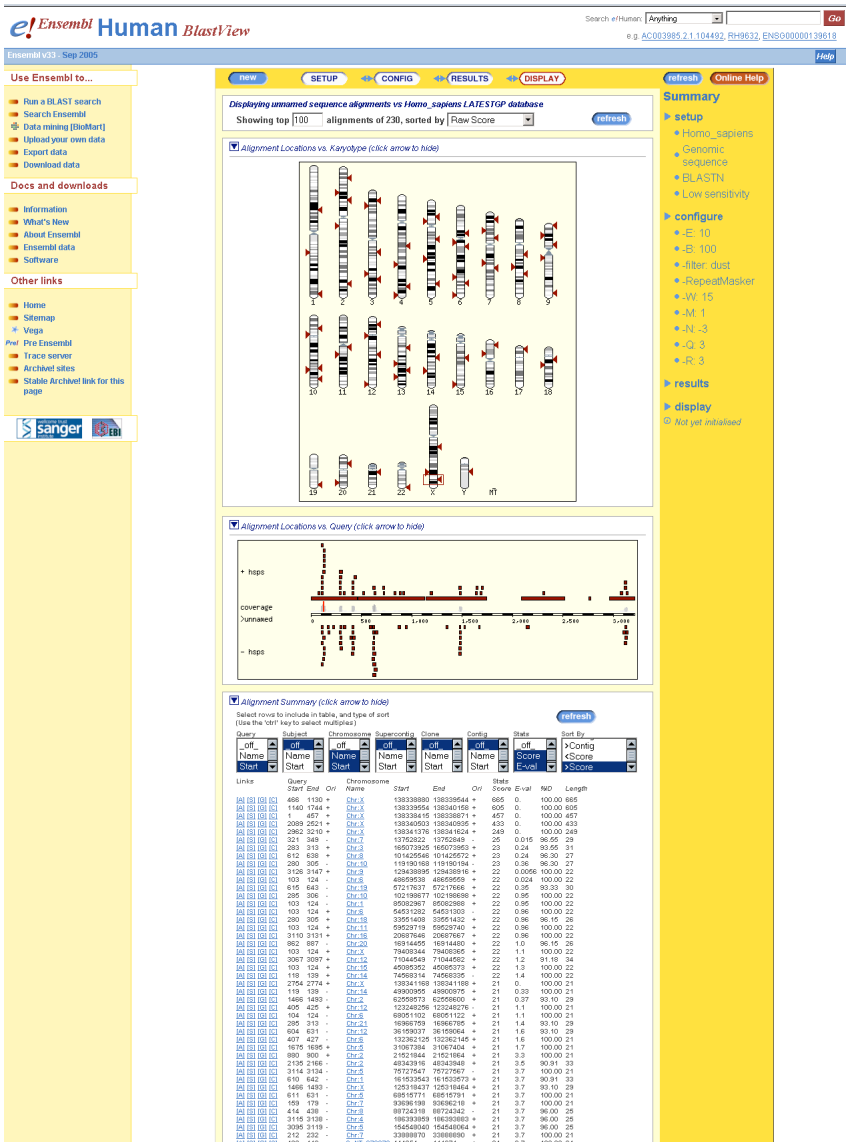
STEP 22:
Paste the sequence in the input box, select *Homo sapiens* as the species and *BLASTN* as the search tool

STEP 23:
Click [**RUN**]

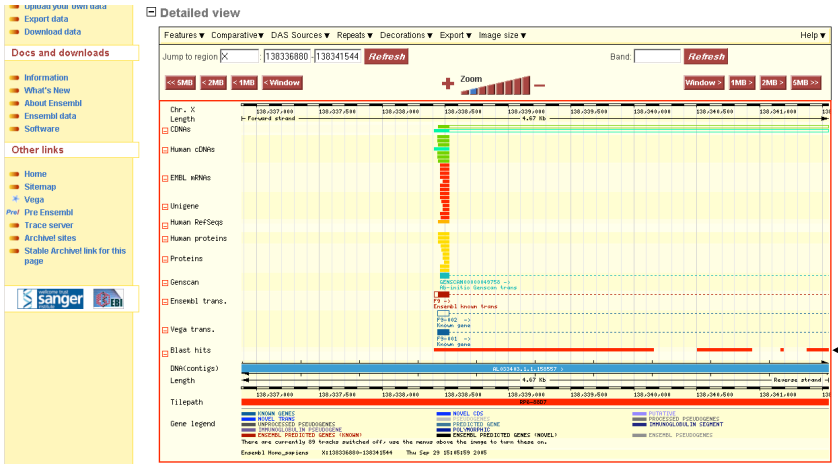


STEP 24:
Click [**Retrieve**] to see whether the query is completed; if so, the [**VIEW**] button is shown

STEP 25:
Click [**VIEW**]



STEP 26:
Click [**C**] in front of the best hit to view this hit in its genomic context in **ContigView**



BLAST hits

Summary

- Genome browsers render the plain sequence more accessible.
- Ensembl is a system, which provides automatic genome annotation, yet strongly based on experimental evidence from cDNA and protein sequences in public sequence databases.
- Ensembl heavily links to data sets from other species, as well as to external resources.

Exercises

1. Exploring features related to a gene

Find the gene report for the human HRAS gene.
Hint: use the search/find boxes to do a text search.

How many transcripts are predicted for this gene? What is the size of the longest predicted mRNA? How many exons does it have?

With what diseases is HRAS associated? What is its cellular function?
Hint: follow some of the links in the 'Similarity Matches' section of GeneView.

How many amino acids does the HRAS transcript code for? Which InterPro domains does the protein product contain?

Find the GO section of GeneView and follow the links to explore the 'Gene ontology' terms (describing gene and protein function) in Ensembl GOView.

In which chromosomal band, on which clone and contig in the genomic sequence assembly is HRAS located? In ContigView, look at regions that are conserved with mouse (turn on 'Mm blastz' and 'Mm high cons bz' tracks).

Is there a putative mouse orthologue? If so, where is it in the mouse genome?

2. Exploring a region

Display the region between markers D12S764 and D12S1871 in ContigView.
Hint: start by clicking chromosome 12 on the human homepage.

How many clones/contigs are used to make this portion of the assembly?

Centre the display around the TENC1 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 closest marker to the TENC1 gene? How many synonyms does this marker have?

Zoom in on the TENC1 gene. Identify a coding SNP and look at the corresponding SNPView page.
Hint: turn on the SNP track if necessary.

Zoom out again, and export an Ensembl gene list for the region. Include Ensembl gene id, known gene external id, known gene external database name, gene description, InterPro short description.
Hint: use the left-hand menu to get to MartView.

3. 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 mouse chromosome 2 between 18.7 Mb and 19.1 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 EntrezGene, UniProt/Swiss-Prot or the Jackson Lab's Mouse Genome Informatics site.
Hint: the 'Marker Symbol' in the 'Similarity Matches' section links to MGI.

Can you find out anything about the possible functions of the novel genes?
Hint: Try looking at homologs in other species and at other members of protein families.

Exercises answers

1. Exploring features related to a gene

Enter the gene name into the text search box and follow the link to GeneView (don't follow links to HRAS-like suppressors).

The HRAS gene (ENSG00000174775) has 2 predicted transcripts. The longer transcript is ENST00000311175. The length of the transcript is 1113 bp, 7 exons (look in Transcript section of GeneView).

Check the 'Disease Matches' section, and follow the links to OMIM and EntrezGene or UniProt/Swiss-Prot. HRAS is associated with bladder cancer. It is a protein that binds GDP/GTP and has intrinsic GTPase activity.

The transcript codes for 189 aa. Check the 'InterPro' section in GeneView. The domains include IPR001806 (Ras GTPase), IPR001230 (Prenyl group, CAAX box, attachment site), IPR001687 (ATP/GTP-binding site motif A (P-loop)) and IPR006689 (ARF/SAR superfamily).

Clicking on a GO identifier gives you a GOView page showing the position of that term in the GO structure (note the number of Ensembl genes mapped to each term). Click Help to find out more about GOView.

Going back to GeneView, follow one of the links in 'Genomic location' to get a ContigView display. HRAS is located on 11p15.5. It is located on contig AC137894.5 (see GeneView - Genomic Location, or point to the contig in ContigView). If you follow the link to the EMBL entry (or if you turn on the 'Tile path' track from the Decorations menu of ContigView) you can see this is clone RP13-46H24. Turn on the display of regions conserved with mouse using the 'Comparative' menu in ContigView.

Back in GeneView, ENSMUSG00000025499 (Hras1) is named in the 'Orthologue Prediction' section. Click on it to go to its GeneView page in the Mouse Ensembl site. Follow the link to mouse GeneView to find that it is on mouse chromosome 7.

2. Exploring a region

Start on the human homepage and click on chromosome 12. Enter the markers in MapView.

The region includes sequence from 4 different contigs (one is quite small) – you can see this better if you turn on the tile path track.

The genes above the contig track are transcribed in the forward orientation; the genes below are in the reverse orientation.
Get marker information by pointing your mouse to the marker at the right of TENC1. It is D12S2110 and 6 synonyms are listed.

SNPs can be turned on using the 'Features' menu. Coding SNPs are shown in red (non-synonymous) and green (synonymous). Note the 'SNP Context' display in SNPView.

The left-hand menu has an option 'Export Gene info in region' that takes you to BioMart. Enter the required data (the chromosomal coordinates and Ensembl Gene ID are already selected for you).

3. Exploring the mouse genome with Ensembl

No fixed answers.

Manual Genome Annotation: the Vertebrate Genome Annotation Browser VEGA

STEP 1:
Load **Vega**:
http://vega.sanger.ac.uk

STEP 2:
Select human genome annotation

Vega Genome Browser

About Vega

The Vertebrate Genome Annotation (VEGA) database is a central repository for high quality, frequently updated, manual annotation of vertebrate finished genome sequence.

Details of the projects for each species are available through the homepages for human, mouse, zebrafish and dog.

The website is built upon code from the Ensembl project.

Documentation and Help

Vega contains context sensitive help that can be accessed by clicking on Help on the title bar, or

Help

The help documentation also contains a Vega FAQ.

Use the sitemap to find out what's available in Vega.

Use Vega

Species	Last Updated
Human	12 May 2005
Mouse	3 Aug 2005
Zebrafish	25 Apr 2005
Dog	14 Feb 2005

What's new

Update to Vega Mouse
3rd Aug 2005

Additional annotation for mouse, and comparative analysis of human haplotypes and mouse strains ... more

Update to Vega Human
12th May 2005

This release of Vega contains manual annotation of five additional human chromosomes (1,16,18,19 and Y) plus an extra Chr 6 MHC region haplotype (QBL). ... more

Feedback

We are keen to recieve extra information on annotation, and to hear your comments, problems, and suggestions on Vega. Please send us feedback.

STEP 3:
Search for gene symbol NFIB (Nuclear Factor IB) under *All* or *Gene*

Ideograms of annotated chromosomes or regions

Vega Human TextView

Home Dog Human Mouse Zebrafish BLAST Export Data Search Feedback Help

Search Homo sapiens All indexes for: nfib

Display up to 20 results in standard format

[Help with AV querying] Lookup Help

1 matches in the *Homo sapiens* Gene index [first 5 matches shown]:

1. Vega Gene: OTTHUMG00000021027
Vega gene OTTHUMG00000021027 has 8 transcripts: OTTHUMT00000055475, OTTHUMT00000055470, OTTHUMT00000055474, OTTHUMT00000055471, OTTHUMT00000055469, OTTHUMT00000055472, OTTHUMT00000055468, OTTHUMT00000055473
Description: nuclear factor IVB
The gene has the following external identifiers mapped to it:
EntrezGene: 4781
HUGO: 7785, NFIB
MIM: 600729
RefSeq_dna: NM_005596
Uniprot/SWISSPROT: O00712
Vega_gene: , OTTHUMG00000021027, NFIB
http://vega.sanger.ac.uk/Homo_sapiens/geneview?gene=OTTHUMG00000021027&db=core

STEP 4:
Follow the link to the Vega **GeneView** page

7 matches in the *Homo sapiens* Peptide index [first 5 matches shown]:

1. Vega Peptide: OTTHUMP00000022768
Vega peptide OTTHUMP00000022768 is a product of Vega gene OTTHUMG00000021027 [transcript OTTHUMT00000055471, NFIB-004]
http://vega.sanger.ac.uk/Homo_sapiens/protview?peptide=OTTHUMP00000022768&db=core

2. Vega Peptide: OTTHUMP00000022769
Vega peptide OTTHUMP00000022769 is a product of Vega gene OTTHUMG00000021027 [transcript OTTHUMT00000055472, NFIB-005]
http://vega.sanger.ac.uk/Homo_sapiens/protview?peptide=OTTHUMP00000022769&db=core

3. Vega Peptide: OTTHUMP00000022770
Vega peptide OTTHUMP00000022770 is a product of Vega gene OTTHUMG00000021027 [transcript OTTHUMT00000055473, NFIB-006]
http://vega.sanger.ac.uk/Homo_sapiens/protview?peptide=OTTHUMP00000022770&db=core

4. Vega Peptide: OTTHUMP00000022771
Vega peptide OTTHUMP00000022771 is a product of Vega gene OTTHUMG00000021027 [transcript OTTHUMT00000055475, NFIB-008]
http://vega.sanger.ac.uk/Homo_sapiens/protview?peptide=OTTHUMP00000022771&db=core

5. Vega Peptide: OTTHUMP00000022765
Vega peptide OTTHUMP00000022765 is a product of Vega gene OTTHUMG00000021027 [transcript OTTHUMT00000055468, NFIB-001]
http://vega.sanger.ac.uk/Homo_sapiens/protview?peptide=OTTHUMP00000022765&db=core

[View more matches....]

Manually curated peptides

8 matches in the *Homo sapiens* Transcript index [first 5 matches shown]:

1. Vega Transcript: OTTHUMT00000055468
Description: nuclear factor IVB
This transcript has the following external identifiers mapped to it:
CCDS: CCDS6474, CCDS6474.1
Vega_transcript: OTTHUMT00000055468, NFIB-001
Vega_translation: OTTHUMP00000022765
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000055468&db=core

2. Vega Transcript: OTTHUMT00000055469
Description: nuclear factor IVB
This transcript has the following external identifiers mapped to it:
Vega_transcript: NFIB-002, OTTHUMT00000055469
Vega_translation: OTTHUMP00000022766
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000055469&db=core

3. Vega Transcript: OTTHUMT00000055470
Description: nuclear factor IVB
This transcript has the following external identifiers mapped to it:
Vega_transcript: NFIB-003, OTTHUMT00000055470
Vega_translation: OTTHUMP00000022767
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000055470&db=core

4. Vega Transcript: OTTHUMT00000055471
Description: nuclear factor IVB
This transcript has the following external identifiers mapped to it:
Vega_transcript: NFIB-004, OTTHUMT00000055471
Vega_translation: OTTHUMP00000022768
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000055471&db=core

5. Vega Transcript: OTTHUMT00000055472
Description: nuclear factor IVB
This transcript has the following external identifiers mapped to it:
Vega_transcript: OTTHUMT00000055472, NFIB-005
Vega_translation: OTTHUMP00000022769
http://vega.sanger.ac.uk/Homo_sapiens/transview?transcript=OTTHUMT00000055472&db=core

[View more matches....]

Manually curated transcripts

empowered Help Desk / Suggestions

Vega Human GeneView

Curated Locus Report

Curated Locus	NFIB (HGNO ID) (to view all Vega genes linked to the name click here) Member of Human CCDS set
Locus ID	OTTHUMG00000021027 [View in Ensembl]
Version	2
Date	Gene last modified on 15/09/2004 (Created on 11/12/2003)
Type	Known [Definition]
Genomic Location	**View gene in genomic location:** 14071847 - 14304581 bp (14.1 Mb) on chromosome 9 **This gene is located in sequence:** AL136366.24.1.120282
Description	nuclear factor IVB
Author	This locus was annotated by Havana <vega@sanger.ac.uk>
Database Matches	**CCDS:** CCDS6474.1 **EntrezGene:** 4781 **HUGO:** NFIB **MIM:** 600728 **RefSeq dna:** NM_005596 **Uniprot/SWISSPROT:** O00712
Sequence Markup	View genomic sequence for this gene with exons highlighted
Export Data	Export gene data in EMBL, GenBank or FASTA
Curated Transcripts	1: NFIB-001 (OTTHUMT00000055469) [Transcript information] [Exon information & supporting evidence] [Protein information] 2: NFIB-002 (OTTHUMT00000055468) [Transcript information] [Exon information & supporting evidence] [Protein information] 3: NFIB-003 (OTTHUMT00000055470) [Transcript information] [Exon information & supporting evidence] [Protein information] 4: NFIB-004 (OTTHUMT00000055471) [Transcript information] [Exon information & supporting evidence] [Protein information] 5: NFIB-005 (OTTHUMT00000055472) [Transcript information] [Exon information & supporting evidence] [Protein information] 6: NFIB-006 (OTTHUMT00000055473) [Transcript information] [Exon information & supporting evidence] [Protein information] 7: NFIB-007 (OTTHUMT00000055474) [Transcript information] [Exon information & supporting evidence] [No translation] 8: NFIB-008 (OTTHUMT00000055475) [Transcript information] [Exon information & supporting evidence] [Protein information]
DAS Sources	GAD (Genetic Association Database) HUGO text (PubMed text-mining via HUGO symbol) Protonet (Global classification of proteins into hierarchical clusters) Reactome (Knowledgebase of biological processes) UniProt (Protein knowledgebase) Manage Sources

Transcript/Translation Summary

NFIB-004	**Stable ID:** OTTHUMT00000055471 **Version:** 2 **Class:** Coding **Exons:** 12 **Transcript length:** 2869 bp **Translation length:** 561 residues [Transcript information] [Exon information & supporting evidence] [Protein information]
InterPro	IPR004863 MH1 domain - [View other Vega genes with this domain] IPR000694 Proline-rich region - [View other Vega genes with this domain] IPR000647 CTF/NF-I family - [View other Vega genes with this domain] IPR002965 Proline-rich extension - [View other Vega genes with this domain]
Transcript Structure	
Protein Features	
Remarks	nuclear factor I/B (NFIB2, NFIB3, NFI-RED) nuclear factor I/B (NFIB2, NRFIB3, NFI-RED)
NFIB-005	**Stable ID:** OTTHUMT00000055472 **Version:** 2 **Class:** Coding **Exons:** 11 **Transcript length:** 2381 bp **Translation length:** 494 residues [Transcript information] [Exon information & supporting evidence] [Protein information]
InterPro	IPR004863 MH1 domain - [View other Vega genes with this domain] IPR000694 Proline-rich region - [View other Vega genes with this domain] IPR000647 CTF/NF-I family - [View other Vega genes with this domain] IPR002965 Proline-rich extension - [View other Vega genes with this domain]
Transcript Structure	
Protein Features	
Remarks	nuclear factor I/B (NFIB2, NFIB3, NFI-RED) nuclear factor I/B (NFIB2, NRFIB3, NFI-RED)
NFIB-006	**Stable ID:** OTTHUMT00000055473 **Version:** 2 **Class:** Coding **Exons:** 3 **Transcript length:** 1274 bp **Translation length:** 200 residues [Transcript information] [Exon information & supporting evidence] [Protein information]
InterPro	IPR004863 MH1 domain - [View other Vega genes with this domain] IPR000647 CTF/NF-I family - [View other Vega genes with this domain]
Transcript Structure	
Protein Features	
Remarks	nuclear factor I/B (NFIB2, NRFIB3, NFI-RED) proetin translation at 3' end extended compared to that published in EM:U70862.1
NFIB-007	**Stable ID:** OTTHUMT00000055474 **Version:** 2 **Class:** Transcript **Exons:** 2 **Transcript length:** 591 bp **Transcript does not translate** [Transcript information] [Exon information & supporting evidence] [No translation]
Transcript Structure	
Remarks	nuclear factor I/B (NFIB2, NRFIB3, NFI-RED)
NFIB-008	**Stable ID:** OTTHUMT00000055475 **Version:** 2 **Class:** Coding **Exons:** 8 **Transcript length:** 840 bp **Translation length:** 168 residues [Transcript information] [Exon information & supporting evidence] [Protein information]
InterPro	IPR000694 Proline-rich region - [View other Vega genes with this domain] IPR000647 CTF/NF-I family - [View other Vega genes with this domain] IPR002965 Proline-rich extension - [View other Vega genes with this domain]
Transcript Structure	
Protein Features	
Remarks	nuclear factor I/B (NFIB2, NFIB3, NFI-RED)

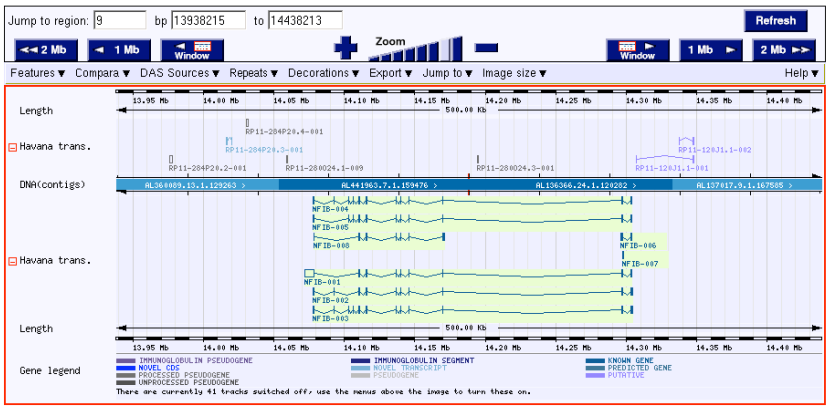
STEP 5:
Follow the genomic location link to the Vega **ContigView** page

List of manually curated transcripts and their corresponding transcript models

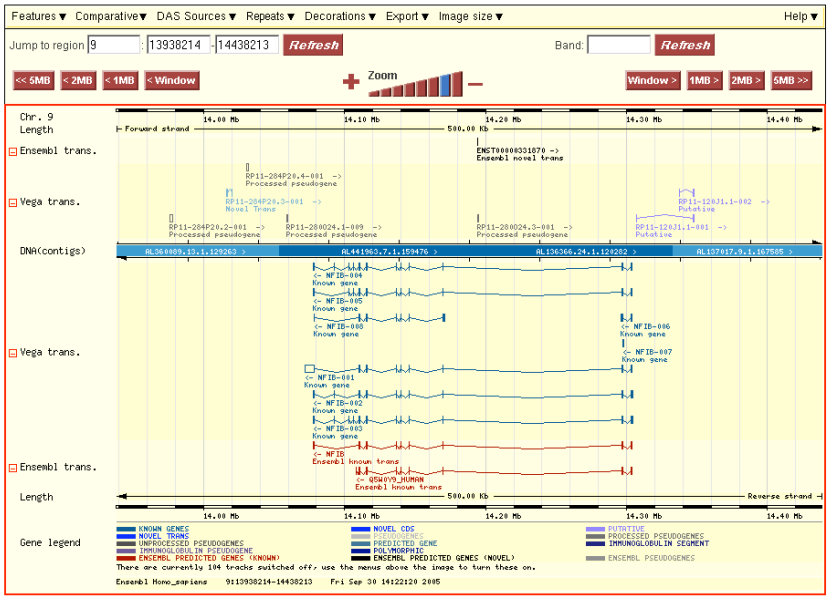
Transcript panels

Transcripts model

Protein model and protein domains



STEP 6:
Compare the manually annotated region in Vega with the automatic annotation provided by Ensembl



Ensembl predicts 2 transcripts models for the NFIB gene, Vega predicts 8

Ensembl predicts a protein coding gene where Vega predicts a pseudogene

Vega predicts transcripts that Ensembl doesn't predict

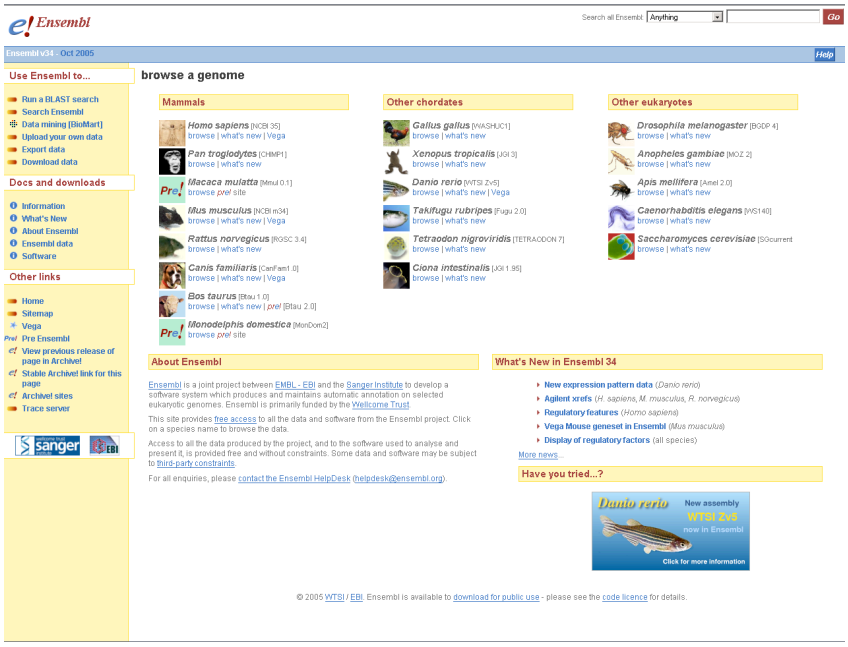
Summary

- Vega is a browser displaying manually annotated vertebrate genomes.
- Manual annotation is labour intensive but generally of higher accuracy than automatic annotation.
- Manual annotation is thus considered the 'gold standard'.

Data mining in Ensembl with BioMart

The gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.
Which other genes related to human diseases locate to the same band?

STEP 1:
Load **Ensembl**:
http://www.ensembl.org



STEP 2:
Go to **BioMart**:

Data mining [BioMart]

STEP 3:
On the **START** page select the datasets *Ensembl* and *Homo sapiens genes (NCBI35)*

STEP 4:
Proceed to the **FILTER** page with **[next>]**

STEP 5:
In the *REGION* section select *Chromosome X* and *Band Start q28* and *Band End q28*

STEP 6:
In the *GENE* section select *Disease genes* and *Only*

Filter by GO terms

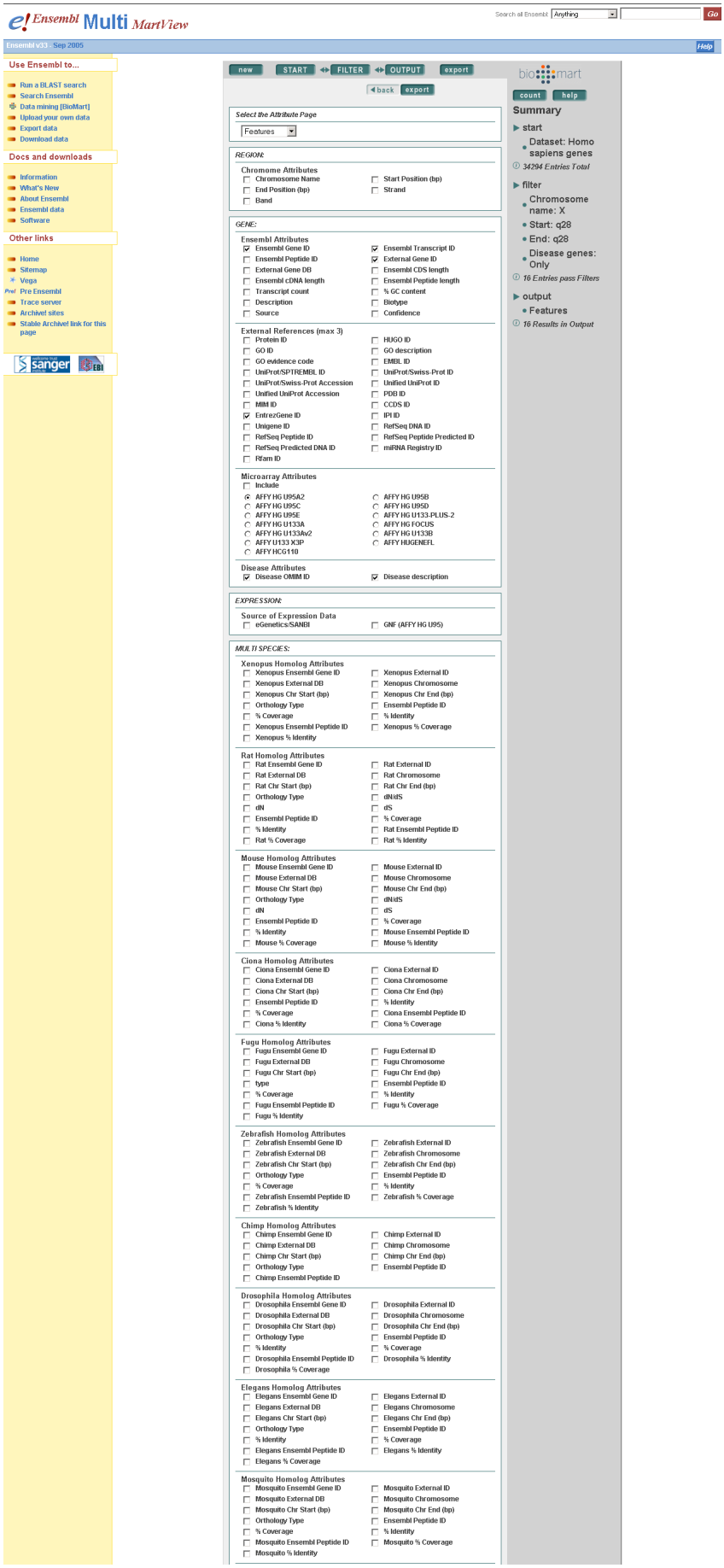
Filter by expression pattern

Filter by ortholog association

Filter by protein domain

Filter by SNP association

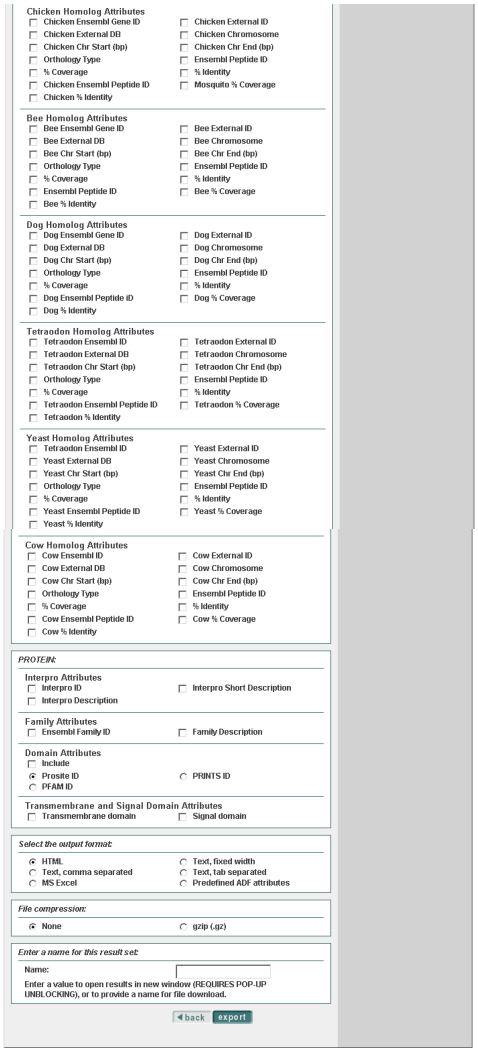
STEP 7:
Proceed to the **OUTPUT** page with **[next>]**



STEP 8:
On the **OUTPUT** page select the *Features* Attribute Page

Ensembl Gene ID is selected default

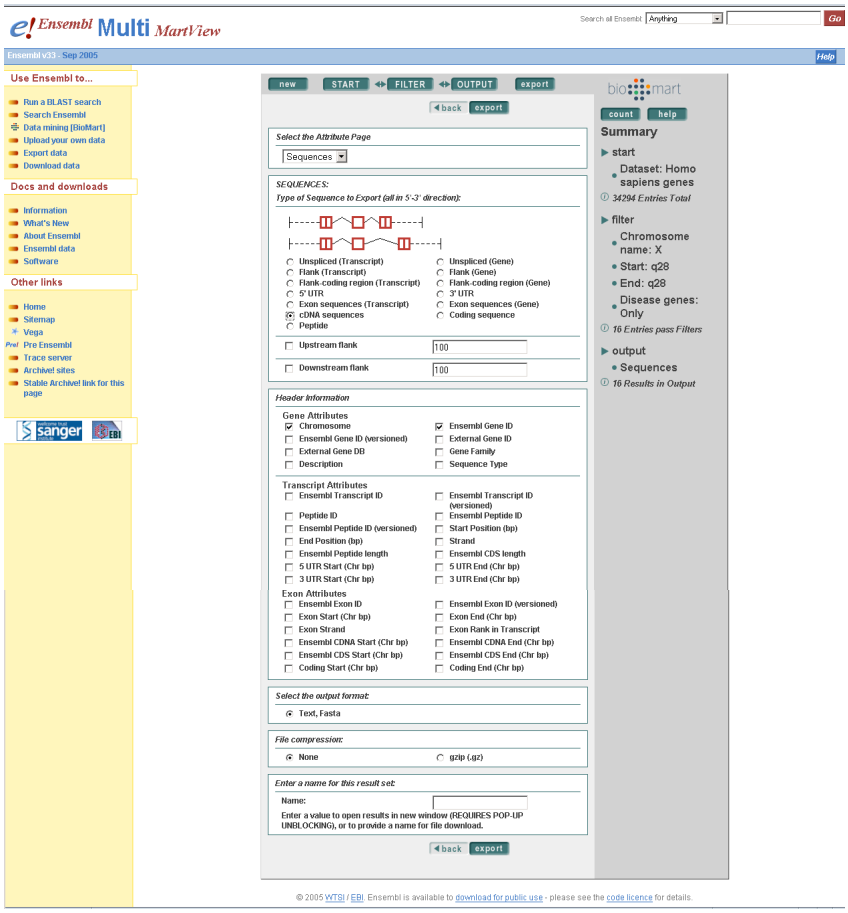
Also select *Ensembl Transcript ID*, *External Gene ID*, *EntrezGene ID*, *Disease OMIM ID* and *Disease description*



STEP 9:
Select *HTML* as output format

STEP 10:
Click [**export**]

Ensembl Gene ID	Ensembl Transcript ID	External Gene ID	EntrezGene ID	Disease OMIM ID	Disease description
ENSG00000155966	ENST00000342251	AFF2		309548	Mental retardation, X-linked, FRAXE type (3)
ENSG00000155966	ENST00000286437	AFF2	2334	309548	Mental retardation, X-linked, FRAXE type (3)
ENSG00000010404	ENST00000361190	IDS	3423	309900	Mucopolysaccharidosis II (3)
ENSG00000010404	ENST00000304701	IDS		309900	Mucopolysaccharidosis II (3)
ENSG00000010404	ENST00000340855	IDS	3423	309900	Mucopolysaccharidosis II (3)
ENSG00000171100	ENST00000306167	MTM1	4534	310400	Myotubular myopathy, X-linked (3)
ENSG00000101986	ENST00000218104	ABCD1	215	300100	Adrenoleukodystrophy (3)
ENSG00000101986	ENST00000218104	ABCD1	215	300100	Adrenomyeloneuropathy (3)
ENSG00000198910	ENST00000361699	L1CAM	3897	308840	Hydrocephalus due to aqueductal stenosis, 307000 (3)
ENSG00000198910	ENST00000361699	L1CAM	3897	308840	MASA syndrome, 303350 (3)
ENSG00000198910	ENST00000361699	L1CAM	3897	308840	Spastic paraplegia, 312900 (3)
ENSG00000198910	ENST00000361981	L1CAM	3897	308840	Hydrocephalus due to aqueductal stenosis, 307000 (3)
ENSG00000198910	ENST00000361981	L1CAM	3897	308840	MASA syndrome, 303350 (3)
ENSG00000198910	ENST00000361981	L1CAM	3897	308840	Spastic paraplegia, 312900 (3)
ENSG00000126895	ENST00000337474	AVPR2	554	304800	Diabetes insipidus, nephrogenic (3)
ENSG00000126895	ENST00000358927	AVPR2		304800	Diabetes insipidus, nephrogenic (3)
ENSG00000169057	ENST00000303391	MECP2	4204	312750	Rett syndrome (2) (?)
ENSG00000102076	ENST00000218195	OPN1LW	5956	303700	Colorblindness, blue monochromatic (3)
ENSG00000102076	ENST00000218195	OPN1LW	5956	303900	Colorblindness, protan (3)
ENSG00000102076	ENST00000218195	OPN1LW	5956	600140	Rubenstein-Taybi syndrome, 180849 (3)
ENSG00000147380	ENST00000276348	OPN1MW	2652	303700	Colorblindness, blue monochromatic (3)
ENSG00000147380	ENST00000276348	OPN1MW	2652	303800	Colorblindness, deutan (3)
ENSG00000166160	ENST00000276343	OPN1MW	2652	303700	Colorblindness, blue monochromatic (3)
ENSG00000166160	ENST00000276343	OPN1MW	2652	303800	Colorblindness, deutan (3)
ENSG00000196924	ENST00000356744	FLNA		311300	Otopalatodigital syndrome, type I (2)
ENSG00000196924	ENST00000360319	FLNA	2316	311300	Otopalatodigital syndrome, type I (2)
ENSG00000102119	ENST00000218240	EMD	2010	310300	Emery-Dreifuss muscular dystrophy (3)
ENSG00000102125	ENST00000335196	TAZ		302060	Barth syndrome (3)
ENSG00000102125	ENST00000335196	TAZ		302060	Cardiomyopathy, X-linked dilated, 300069 (3)
ENSG00000102125	ENST00000335196	TAZ		302060	Endocardial fibroelastosis 2 (2)



STEP 11:
Use the browser BACK button to go back to the **OUTPUT** page

Select the *Sequences* Attribute Page.

Select *cDNA sequences*

STEP 12:
Click [**export**]

```
>X|ENSG00000155966
CGCCGCCTGTGCAGCCGCTGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCTGCCGCCCCGGCTG
CCGCGCCGCGCCGCTGCCTCTGCCCCGGCCGCCCCCGCCGCCGCTGCCGCCGCCGGCCCGCAGCCAGCCAGGCGGGCGGC
CCAGCCCGCCTGAGCCCGCAGCGGCTGCCGCCGCAGCGTCGGGTCGCTGGGTGCGCGGGCTACCGCGGACCGAGCGGACC
CGAGTGGGCGACCAGGCGCTTGCCCGCCCAGTGCCACTGCCGCCGCTTCCTCGCCGGAGCACAGGACCAGACACCTCCAG
CGCCCGCTGCTGCTGCCGATGCGGCCCGGACACTTTTAGCTGGGCGGGAGGGCTGGAGAGCCGGGGGCCGCCGAGAACCG
CCAGCGAGCTGTGCCGAGAGCCGCGCCGACCCGCTGCGATCAGGGACAGGCGCCCGCCCGCCGCCGCCGCCTGGCCGCTA
TGGATCTATTCGACTTTTTCAGAGACTGGGACTTGGAGCAGCAGTGTCACTATGAACAAGACCGTAGTGCACTTAAAAAA
AGGGAATGGGAGCGGAGGAATCAAGAAGTCCAGCAAGAAGACGATCTCTTTTCTTCAGGCTTTGATCTTTTTGGGGAGCC
ATACAAGGTAGCTGAATATACAAACAAAGGTGATGCACTTGCCAACCGAGTCCAGAACACGCTTGGAAACTATGATGAAA
TGAAGAATTTGCTAACTAACCATTCTAATCAGAATCACCTAGTGGGAATTCCAAAGAATTCTGTGCCCCAGAATCCCAAC
AACAAAAATGAACCAAGCTTTTTTCCAGAACAAAAGAACAGAATAATTCCACCTCACCAGGATAATACCCATCCTTCAGC
ACCAATGCCTCCACCTTCTGTTGTGATACTGAATTCAACTCTAATACACAGCAACAGAAAATCAAAACCTGAGTGGTCAC
GTGATAGTCATAACCCTAGCACTGTACTGGCAAGCCAGGCCAGTGGTCAGCCAAACAAGATGCAGACTTTGACACAGGAC
CAGTCTCAAGCCAAACTGGAAGACTTCTTTGTCTACCCAGCTGAACAGCCCCAGATTGGAGAAGTTGAAGAGTCAAACCC
ATCTGCAAAGGAAGACAGTAACCCTAATTCTAGTGGAGAAGATGCTTTCAAAGAAATCTTTCAATCCAATTCACCGGAAG
AATCTGAATTCGCCGTGCAAGCGCCTGGGTCTCCCCTAGTGGCTTCCTCTTTATTAGCTCCTAGCAGTGGCCTTTCAGTT
CAAAACTTCCCACCAGGGCTTTACTGCAAAACAAGCATGGGGCAGCAAAAGCCAACTGCATACGTCAGACCCATGGATGG
CCAGGACCAGGCACCGGACATCTCACCAACACTGAAACCTTCAATTGAATTTGAGAACAGCTTTGGGAATCTGTCATTTG
GAACACTCTTGGATGGAAAACCCAGTGCAGCCAGTTCAAAGACTAAACTGCCAAAGTTCACCATCCTCCAAACAAGTGAA
GTAAGCCTTCCCAGTGATCCAAGCTGTGTTGAAGAAATCTTGCGGGAGATGACCCATTCCTGGCCTACTCCTCTCACTTC
CATGCATACTGCTGGACACTCTGAGCAGAGCACCTTTTCCATCCCAGGACAGGAATCGCAGCATCTGACCCCAGGATTCA
CCTTACAAAAGTGGAATGACCCAACCACCAGAGCTTCTACAAAGTCAGTGTCTTTCAAATCGATGCTTGAGGATGACCTG
AAGCTGAGCAGTGATGAAGATGACCTTGAGCCTGTGAAGACCTTGACCACTCAGTGCACTGCCACTGAGCTCTACCAGGC
TGTTGAAAAGGCAAAACCTAGGAATAATCCTGTGAACCCACCCTTGGCCACTCCCCAGCCCCCACCTGCAGTGCAAGCCA
```

Summary

- BioMart is a highly flexible tool for data mining.
- Queries are defined applying convenient filters in just three steps:
  the 'START', 'FILTER' and 'OUTPUT' phases.
- Genomic regions, Gene identifiers, Gene Ontology terms, and many other sources of information can serve as filters.
- BioMart heavily links data sets within Ensembl and provides links to external resources.

Exercises

1. Retrieve all SNPs for 'novel' human G-protein coupled receptor genes (GPCRs – IPR000276) on chromosome 2.

Choose 'Data mining [BioMart]' from the left-hand menu.

Choose the database and the species for your query on the START page:

- Select 'Ensembl 34' as we are looking for a gene list (NOT a SNP list!).
- Select 'Homo sapiens genes (NCBI35)'.

Click next. Note that the Summary to the right is updated. It indicates the total number of Ensembl genes present in the database.

Select your genes of interest using the FILTER page:

- Select 'Chromosome 2'.
- Select 'Known Genes' and tick 'Excluded' to retrieve 'novel' genes only.
- Click count at the top of the Summary panel and note that the number of genes fulfilling these criteria is given under filter.
- In the 'Protein' section, select 'InterPro ID(s)' and enter 'IPR000276'
- Click count again and note that the number of genes is updated. Click next.

Select the OUTPUT for your gene list:

- Choose the 'SNPs' Attribute page.
- In the 'Gene' section 'Ensembl Gene ID' is selected by default – also select 'Ensembl Transcript ID' and 'Ensembl Peptide length'.
- In the 'Gene associated SNPs' section 'Reference ID' is selected by default – also select 'Allele', 'Peptide location (aa)', 'Location in Gene (coding etc)', 'Synonymous Status' and 'Peptide Shift'.
- Choose 'MS Excel' as output format.
- Click export.

Note that the output for this query gives you one row for each SNP, and if there are alternative transcripts then SNP data is given for each. This means that a particular SNP may appear more than once. Find the coding SNPs, and note that you have information about the effect of the SNP, and its location within the protein. How many coding SNPs are there in each transcript? Of these, how may affect the amino acid sequence?

2. Retrieve all mouse homologues of human disease genes containing transmembrane domains located between 1p22.1 and 1q22.

Hint: Start with a human gene list. Include "Disease OMIM ID" and "Disease description" in the OUTPUT section.

3. Retrieve the gene structure of the following mouse genes: ENSMUSG00000042351 ENSMUSG00000022393

Hint: Use the 'Ensembl Gene ID(s)' option under 'ID list limit' on the FILTER page. Select on the OUTPUT, 'Structures' Attribute page 'Ensembl Gene ID', 'Ensembl Transcript ID', 'Ensembl Exon ID', 'Exon Start' and 'Exon'. Export the gene structure in HTML format. Then take the link from the Ensembl genes to GeneView in order to confirm the gene structure.

4. Retrieve the sequences 5kb upstream of all human 'known' genes between D1S2806 and D1S464.

Hint: Use 'Flank (Gene)' under the sequence type options on the OUTPUT, 'Sequences' Attribute page. 'Known' genes in Ensembl are Ensembl gene predictions that could be mapped to external database entries (e.g. UniProt/Swiss-Prot) with a high similarity score.

5. Retrieve all human SNPs that have a TSC ID, from chromosome 6 between 15 Mb and 15.2 Mb, with 200 bases flanking sequence.

Hint: This time you want to start with a 'SNP' list.

6. Design your own query!

Exercises answers

Warning: if there has recently been a new Ensembl release based on an updated gene set and/or an updated SNP set, these numbers may have changed.

1. You should find four novel genes on chromosome 2 with this InterPro domain. The result set has five transcripts and a total of about 2371 rows of output. All transcripts have one or more coding SNPs, most of which are non-synonymous (synonymous status = no).

2. You should find five genes. One of these is linked to two OMIM IDs and five Disease descriptions, while the other four genes are linked to one OMIM ID and one Disease description.

3. -

4. You should find 26 genes.

5. You should find 69 SNPs.

6. -

Notes

Notes

Notes