Access to genes and genomes with Ensembl

Introduction and Worked Example

February 2007

CONTENTS

Introduction

Ensembl is one of the world's primary resources for genomic research, a resource through which scientists can access the human genome as well as the genomes of other model organisms. Because of the complexity of the genome and the many different ways in which scientists want to use it, Ensembl has to provide many levels of access with a high degree of flexibility. Through the Ensembl website a wet-lab researcher with a simple web browser can for example perform BLAST searches against chromosomal DNA, download a genomic sequence or search for all members of a given protein family. But Ensembl is also an all-round software and database system that can be installed locally to serve the needs of a genomic centre or a bioinformatics division in a pharmaceutical company enabling complex data mining of the genome or large-scale sequence annotation.

The need for automatic annotation

Recent years have seen the release of huge amounts of sequence data from genome sequencing centres (figure 1). However, this raw sequence data is most valuable to the

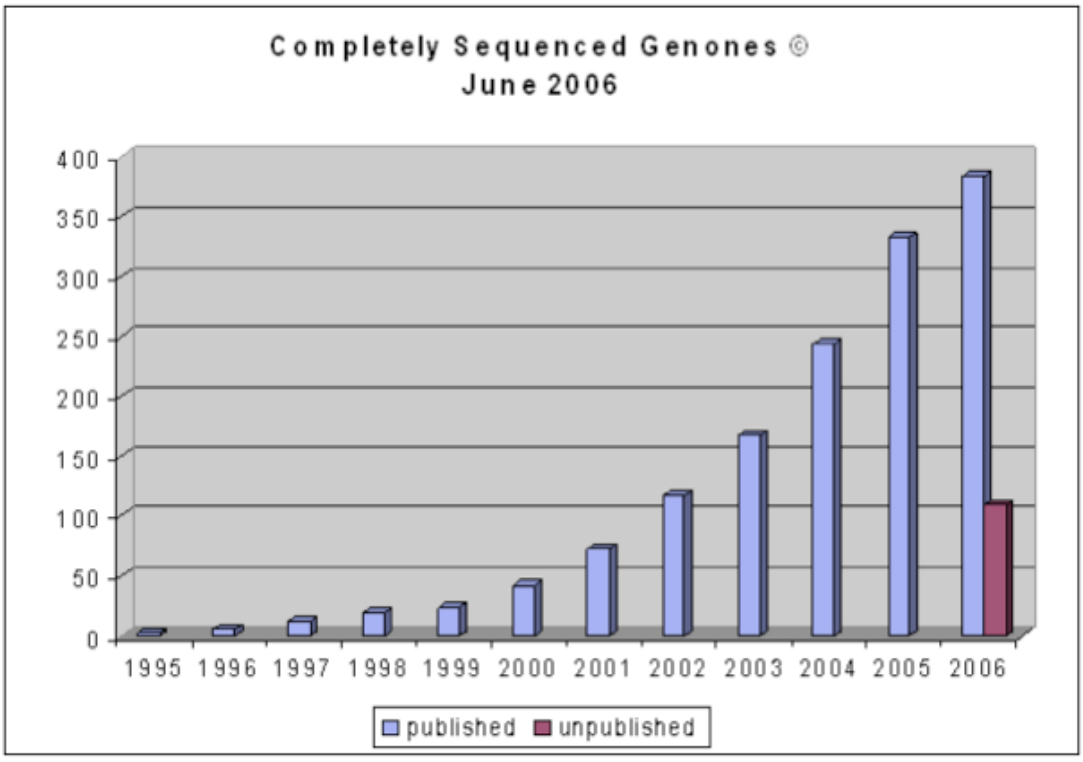


Figure 1. Completely sequenced genomes as of june 2006 (figure taken from http://www.genomesonline.org).

laboratory biologist when provided along with quality annotation of the genomic sequence. This information can be the starting point for planning experiments, interpreting Single Nucleotide Polymorphisms, inferring the function of gene products, predicting regulatory sites for gene expression and so on. The currently agreed 'gold standard' for the annotation of eukaryotic genomes is annotation made by a human being. This so-called “manual annotation” is based on information derived from sequence homology searches, the results of various *ab initio* gene prediction methods and literature searches. Annotation of large genomes (such as mouse and human) that meet this standard is slow and labour intensive, taking large teams of annotators years to complete. As a result, the annotation can almost never be entirely up-to-date and free of inconsistencies (as the annotation process usually begins before the sequencing process is complete). Hence, an automated annotation system is desirable since it is a relatively rapid process that allows frequent updates to accommodate new data. To meet this need, we produced the Ensembl annotation system by observing how annotators build gene structures and condensing this process into a set of rules.

The start of Ensembl

Ensembl's genesis was in response to the acceleration of the public effort to sequence the human genome in 1999. At that point it was clear that if annotation of the draft sequence was to be available in a timely fashion it would have to be automatically generated and that new software systems would be needed to handle genome data sets that were much larger, much more fragmented and much more rapidly changing than anything previous dealt with.

Ensembl was conceived in three parts: as a scalable way of storing and retrieving genomic data; as a web site for genome display; and as an automatic annotation method based around a set of heuristics. It was initially written for the draft human genome, which was sequenced clone-by-clone but has also been successfully used for whole genome shotgun assemblies. The storage and display parts of Ensembl are used for all the genomes currently present in Ensembl, while the automatic gene annotation has been run for most of the genomes with the exception of Takifugu, Tetraodon, Fruitfly, *C. elegans* and Yeast.

Over the past few years Ensembl has grown into a large scale enterprise, with substantial computing resources enabling it to process and provide live database access to currently more than 25 different genomes (figure 2) and a bimonthly update frequency to its website. It has a large community of users in both industry and academia, using it as a base for their individual organisation’s experimental and computational genome based investigations, some of which maintain their own local installations.

Ensembl is a collaboration between the European Bioinformatics Institute (EBI) and the Wellcome Trust Sanger Institute, both located on the Wellcome Trust Genome Campus in Hinxton, Cambridge, UK. Ensembl is funded

principally by the Wellcome Trust, with additional funding from the European Molecular Biology Laboratory (EMBL), the National Institutes of Health – National Institute of Allergy and Infectious Disease (NIH-NIAID) and the Biotechnology and Biological Sciences Research Council (BBSRC).

The Ensembl software and database system

As a software/database system Ensembl can be best described as a hybrid of a scripting programming language (Perl) and a relational database (MySQL, pronounced “My Ess Que Ell”).).
Ensembl Perl software inherits from a tradition of biological object-design developed through BioPerl (http://www.bioperl.org/). This means that developers at Ensembl aimed at creating reusable pieces of software that would faithfully describe biological entities such as gene, transcript, protein, genomic clone or chromosome. Rules of usage and design of Ensembl and BioPerl objects can be best learned while using them, browsing their code and through a bit of trial-and-error. There is a comprehensive BioPerl tutorial available at the BioPerl website.
The Ensembl database is based on a relational database called MySQL. SQL in MySQL stands for ‘Structured Query Language’, a universal database programming language shared by many relational databases. Because MySQL is available free of charge for non-commercial developers, every academic centre can install its own local copy of MySQL (MySQL server) and download Ensembl data from the Ensembl ftp site. Simple queries of the database can be handled using the SQL language (see appendix), but for complex queries demanded by most biological analyses the Ensembl MySQL server is best accessed using Ensembl Perl objects.

The Ensembl annotation pipeline

The Ensembl analysis and annotation pipeline is based on a rule set of heuristics that a human annotator would use. All Ensembl gene predictions are based on experimental evidence, which is imported via manually curated UniProt/Swiss-Prot, partially manually curated NCBI RefSeq and automatically annotated UniProt/TrEMBL records. Untranslated regions (UTRs) are annotated to the extent supported by EMBL mRNA records. As there is no guarantee that UTR sequences in EMBL records are complete there is similarly no guarantee that the Ensembl genome analysis and annotation pipeline has enough biological evidence to predict complete UTR regions. For a limited number of species regulatory regions are annotated, but this annotation isn’t very extensive yet as the set of well-characterised promoters is still small and there is currently no algorithm yielding reliable results on a genomic scale.

The Ensembl website

Ensembl provides easy access to genomic information with a number of visualisation tools. The Ensembl website gives you for example the possibility to directly download data, whether it is a DNA sequence of a genomic contig

you are trying to identify novel genes in, or positions of SNPs in a gene you are working on. The key Ensembl web pages are called Views (e.g. GeneView, ContigView and SNPView), and will all be introduced appropriately later on. An updated version of the website is released bimonthly. Old versions are for at least two years accessible on the 'Archive!' website. Apart from that the 'Pre!' website provides displays of genomes that are still in the process of being annotated. There is also an ftp site to download large amounts of data from the Ensembl database, as well as the data-mining tool BioMart, that allows rapid retrieval of information from the databases. Finally, Ensembl BLAST offers the possibility to perform sequence searches against genomes and Ensembl gene and peptide sets.

Further reading

Hubbard, T.J.P. *et al.*
Ensembl 2007
Nucleic Acids Res. 2007 *(Database Issue)*

Birney, E. *et al*.
Ensembl 2006.
Nucleic Acids Res. 2006 Jan 34:D556-D561 (2006)

Hubbard, T. *et al.*
Ensembl 2005.
Nucleic Acids Res. 2005 33 D447-D453 (2005)

Birney, E. *et al*. *
An Overview of Ensembl.
Genome Research 14(5): 925-928 (2004)

Kasprzyk, A. *et al*.
EnsMart: a generic system for fast and flexible access to biological data.
Genome Research (2004) 14:1, 160-9.

Ashurst, J. L. et *al.*
The Vertebrate Genome Annotation (Vega) database.
Nucl. Acids Res. 33:D459-D465 (2005)

* This paper was part of the may 2004 issue of Genome Research which included an Ensembl special covering detailed aspects of the Ensembl web site, the underlying scalable database system for storing genome sequence and annotation information, as well as the automated genome analysis and annotation pipeline.

SPECIES		ASSEMBLY		GENEBUILD	
Mammals					
Human	*Homo sapiens*	NCBI 36	oct 2005	Ensembl	jul 2006
Chimpanzee	*Pan troglodytes*	PanTro 2.1	mar 2006	Ensembl	mar 2005
Rhesus macaque	*Macaca mulatta*	MMUL 1	feb 2006	Ensembl	aug 2006
Bushbaby*	*Otolemur garnettii*	BUSHBABY1			
Mouse	*Mus musculus*	NCBI m36	dec 2005	Ensembl	apr 2006
Rat	*Rattus norvegicus*	RGSC 3.4	dec 2004	Ensembl	feb 2006
Rabbit	*Oryctolagus cuniculus*	RABBIT	may 2005	Ensembl	aug 2006
Dog	*Canis familaiaris*	CanFam 1.0	jul 2004	Ensembl	nov 2004
Cat*	*Felis catus*	CAT			
Cow	*Bos taurus*	Btau 2.0	mar 2005	Ensembl	dec 2005
Pig**	*Sus scrofa*				
Shrew*	*Sorex araneus*	sorAra1			
Hedgehog*	*Erinaceus europaeus*	eriEur1			
Microbat*	*Myotis lugigfugus*	MICROBAT1			
Armadillo	*Dasypus novemcinctus*	ARMA	may 2005	Ensembl	aug 2006
Elephant	*Loxodonta africana*	BROAD E1	may 2005	Ensembl	aug 2006
Lesser hedgehog tenrec	*Echinops telfairi*	TENREC	may 2005	Ensembl	aug 2006
Opossum	*Monodelphis domestica*	MonDom 4.0	jan 2006	Ensembl	feb 2006
Platypus*	*Ornithorhynchus anatinus*	OANA 5			
Other species					
Chicken	*Gallus gallus*	WASHUC 1	mar 2004	Ensembl	dec 2005
X. tropicalis	*Xenopus tropicalis*	JGI 4.1	aug 2005	Ensembl	nov 2005
Zebrafish	*Danio rerio*	Zv 6	mar 2006	Ensembl	aug 2006
Fugu	*Takifugu rubripes*	FUGU 4.0	jun 2005	IMCB/JGI	may 2005
Tetraodon	*Tetraodon nigroviridis*	TETRAODON 7	apr 2003	Genoscope	sep 2004
Stickleback	*Gasterosteus aculeatus*	BROAD S1	feb 2006	Ensembl	aug 2006
Medaka	*Oryzias latipes*	HdrR 1	oct 2005	Ensembl	may 2006
C. intestinalis	*Ciona intestinalis*	JG 12	mar 2005	Ensembl	feb 2006
C. savignyi	*Ciona savignyi*	CSAV 2.0	oct 2005	Ensembl	apr 2006
Fruitfly	*Drosophila melanogaster*	BDGP 4	jul 2005	FlyBase	mar 2006
Anopheles	*Anopheles gambiae*	AgamP 3	feb 2006	VectorBase	oct 2005
Aedes	*Aedes aegypti*	AaegL 1	aug 2005	VectorBase	jun 2006
C. elegans	*Caenorhabditis elegans*	WS 150	nov 2005	WormBase	nov 2005
S. cerevisiae	*Saccharomyces cerevisiae*	SGD 1	nov 2005	SGD	nov 2005

Figure 2 – Species in Ensembl, including name and date of their genome assembly and source and date of the genebuild. * = currently only available on the Pre! website, ** = only clone information available.

WORKED EXAMPLE – A walk through the main pages of the Ensembl browser, using the EPO (Erythropoietin precursor) gene as an example.

STEP 1:
Load Ensembl
www.ensembl.org

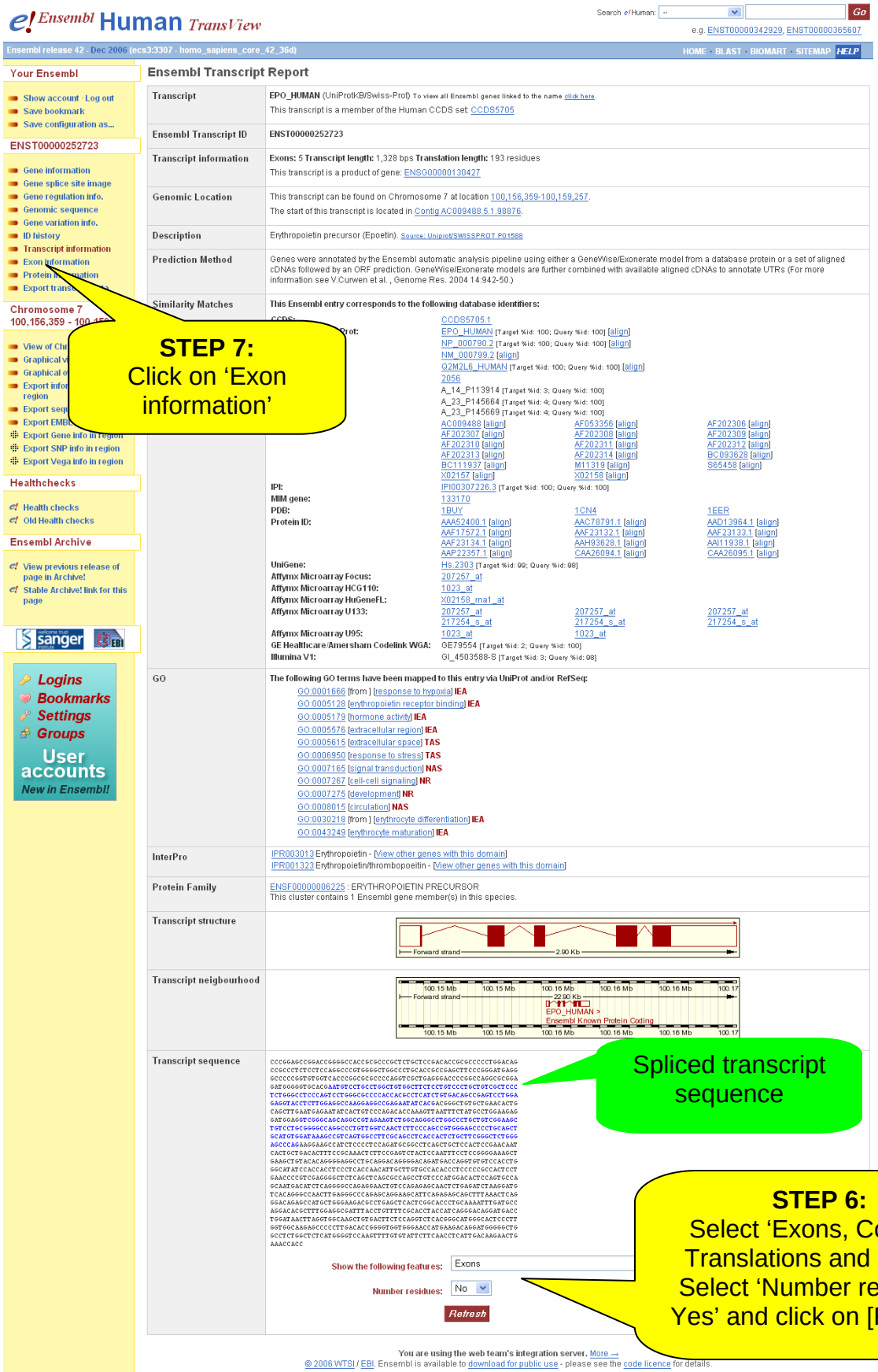
STEP 3:
Type in ‘EPO Gene’.
Click ‘Go’.

Karyotype

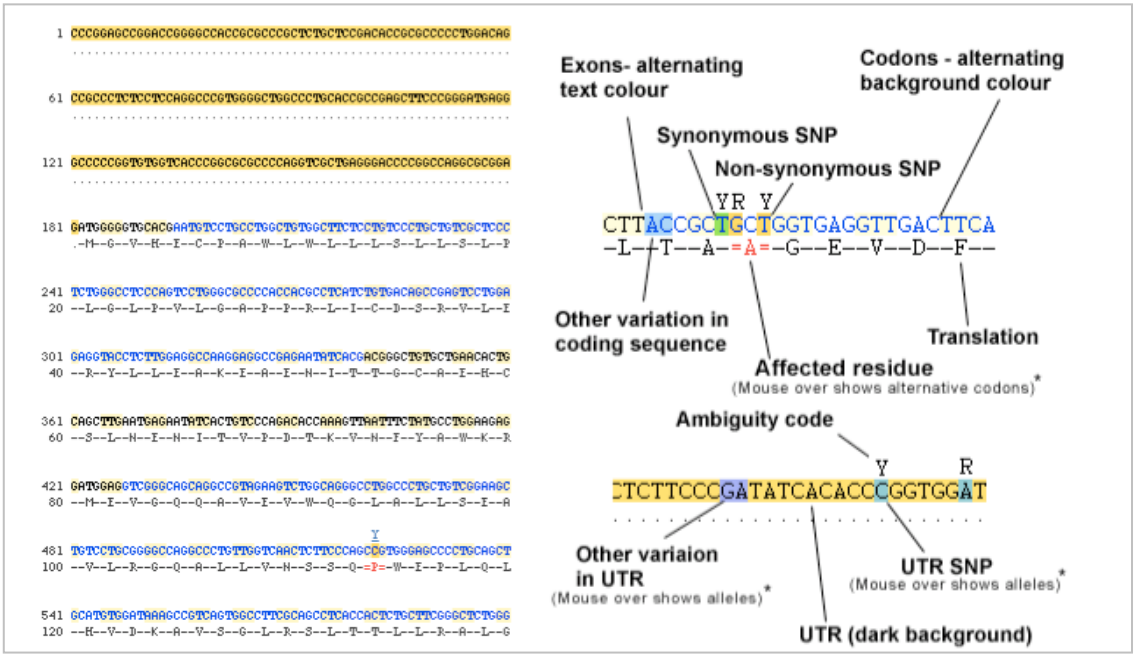
Source and version of assembly and genebuild


A ‘Vega’ gene
(a consortium external to Ensembl)
STEP 4:
Click on ‘ENSG00000130427’
Ensembl text search
EPO Gene
Search
Your query matched 4 entries in the search database
Vega protein_coding Gene: OTTHUMG00000023044 (HGNC Symbol: EPO) [ContigView]
Ensembl protein_coding Gene: ENSG00000130427 (HGNC Symbol: EPO) [ContigView]
Ensembl protein_coding Gene: ENSG00000121053 (HGNC Symbol: EPX) [ContigView]
Ensembl protein_coding Gene: ENSG00000187266 (HGNC Symbol: EPOR) [ContigView]
Your Ensembl
Feature type
Species
Exalead Help
Healthchecks
Ensembl Archive
Dog CanFam 2.0
Now in Ensembl


The Gene View Page
Gene Model
Orthologues in other species
STEP 5:
Click on 'Transcript Information'
Matches in other databases
GO (Gene Ontology) terms

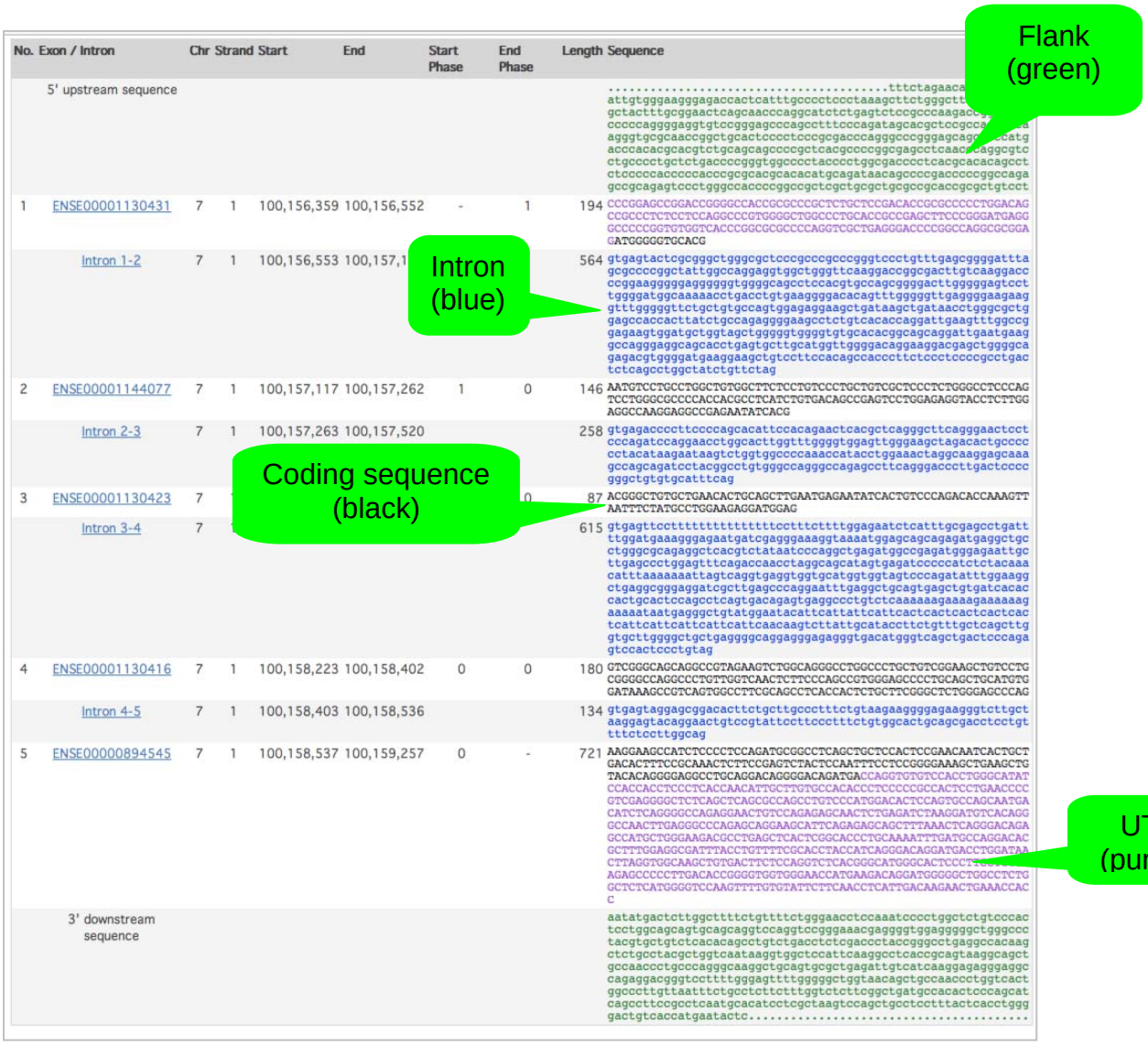

Ensembl Human TransView
Search e!Human:
Go
e.g. ENST00000342929, ENST00000365607
Ensembl release 42 - Dec 2006 (ecs3:3307 - homo_sapiens_core_42_36d)
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Show account · Log out
Save bookmark
Save configuration as...
ENST00000252723
Gene information
Gene splice site image
Gene regulation info.
Genomic sequence
Gene variation info.
ID history
Transcript information
Exon information
Protein information
Export transcript
Chromosome 7
100,156,359 - 100,159,257
View of Chromosome
Graphical view
Graphical overview
Export information for region
Export sequence for region
Export EMBL for region
Export Gene info in region
Export SNP info in region
Export Vega info in region
Healthchecks
Health checks
Old Health checks
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Logins
Bookmarks
Settings
Groups
User accounts
New in Ensembl!
Ensembl Transcript Report
Transcript
EPO_HUMAN (UniProtKB/Swiss-Prot) To view all Ensembl genes linked to the name click here.
This transcript is a member of the Human CCDS set: CCDS5705
Ensembl Transcript ID
ENST00000252723
Transcript information
Exons: 5 Transcript length: 1,328 bps Translation length: 193 residues
This transcript is a product of gene: ENSG00000130427
Genomic Location
This transcript can be found on Chromosome 7 at location 100,156,359-100,159,257.
The start of this transcript is located in Contig AC009488.5.1.98876.
Description
Erythropoietin precursor (Epoetin). Source: UniprotSWISSPROT P01588
Prediction Method
Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned cDNAs followed by an ORF prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate UTRs (For more information see V.Curwen et al. , Genome Res. 2004 14:942-50.)
Similarity Matches
This Ensembl entry corresponds to the following database identifiers:
CCDS: CCDS5705.1
Prot: EPO_HUMAN [Target %id: 100; Query %id: 100] [align]
NP_000790.2 [Target %id: 100; Query %id: 100] [align]
NM_000799.2 [align]
Q2M2L6_HUMAN [Target %id: 100; Query %id: 100] [align]
2056
A_14_P113914 [Target %id: 3; Query %id: 100]
A_23_P145664 [Target %id: 4; Query %id: 100]
A_23_P145669 [Target %id: 4; Query %id: 100]
AC009488 [align] AF053356 [align] AF202306 [align]
AF202307 [align] AF202308 [align] AF202309 [align]
AF202310 [align] AF202311 [align] AF202312 [align]
AF202313 [align] AF202314 [align] BC093628 [align]
BC111937 [align] M11319 [align] S65458 [align]
X02157 [align] X02158 [align]
IPI: IPI00307226.3 [Target %id: 100; Query %id: 100]
MIM gene: 133170
PDB: 1BUY 1CN4 1EER
Protein ID: AAA52400.1 [align] AAC78791.1 [align] AAD13964.1 [align]
AAF17572.1 [align] AAF23132.1 [align] AAF23133.1 [align]
AAF23134.1 [align] AAH93628.1 [align] AAI11938.1 [align]
AAP22357.1 [align] CAA26094.1 [align] CAA26095.1 [align]
UniGene: Hs.2303 [Target %id: 99; Query %id: 98]
Affymx Microarray Focus: 207257_at
Affymx Microarray HCG110: 1023_at
Affymx Microarray HuGeneFL: X02158_rna1_at
Affymx Microarray U133: 207257_at 207257_at 207257_at
217254_s_at 217254_s_at 217254_s_at
Affymx Microarray U95: 1023_at 1023_at
GE Healthcare/Amersham Codelink WGA: GE79554 [Target %id: 2; Query %id: 100]
Illumina V1: GI_4503588-S [Target %id: 3; Query %id: 98]
GO
The following GO terms have been mapped to this entry via UniProt and/or RefSeq:
GO:0001666 [from] [response to hypoxia] IEA
GO:0005128 [erythropoietin receptor binding] IEA
GO:0005179 [hormone activity] IEA
GO:0005576 [extracellular region] IEA
GO:0005615 [extracellular space] TAS
GO:0006950 [response to stress] TAS
GO:0007165 [signal transduction] NAS
GO:0007267 [cell-cell signaling] NR
GO:0007275 [development] NR
GO:0008015 [circulation] NAS
GO:0030218 [from] [erythrocyte differentiation] IEA
GO:0043249 [erythrocyte maturation] IEA
InterPro
IPR003013 Erythropoietin - [View other genes with this domain]
IPR001323 Erythropoietin/thrombopoietin - [View other genes with this domain]
Protein Family
ENSF00000006225 : ERYTHROPOIETIN PRECURSOR
This cluster contains 1 Ensembl gene member(s) in this species.
Transcript structure
Forward strand
2.90 Kb
Transcript neigbourhood
100.15 Mb 100.15 Mb 100.16 Mb 100.16 Mb 100.16 Mb 100.17
Forward strand
22.90 Kb
EPO ; HUMAN >
Ensembl Known Protein Coding
Transcript sequence
Show the following features:
Exons
Number residues:
No
Refresh
You are using the web team's integration server. More →
© 2006 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.
STEP 7:
Click on 'Exon information'
Spliced transcript sequence
STEP 6:
Select 'Exons, Codons, Translations and SNPs'.
Select 'Number residues: Yes' and click on [Refresh]

Result of STEP 6:



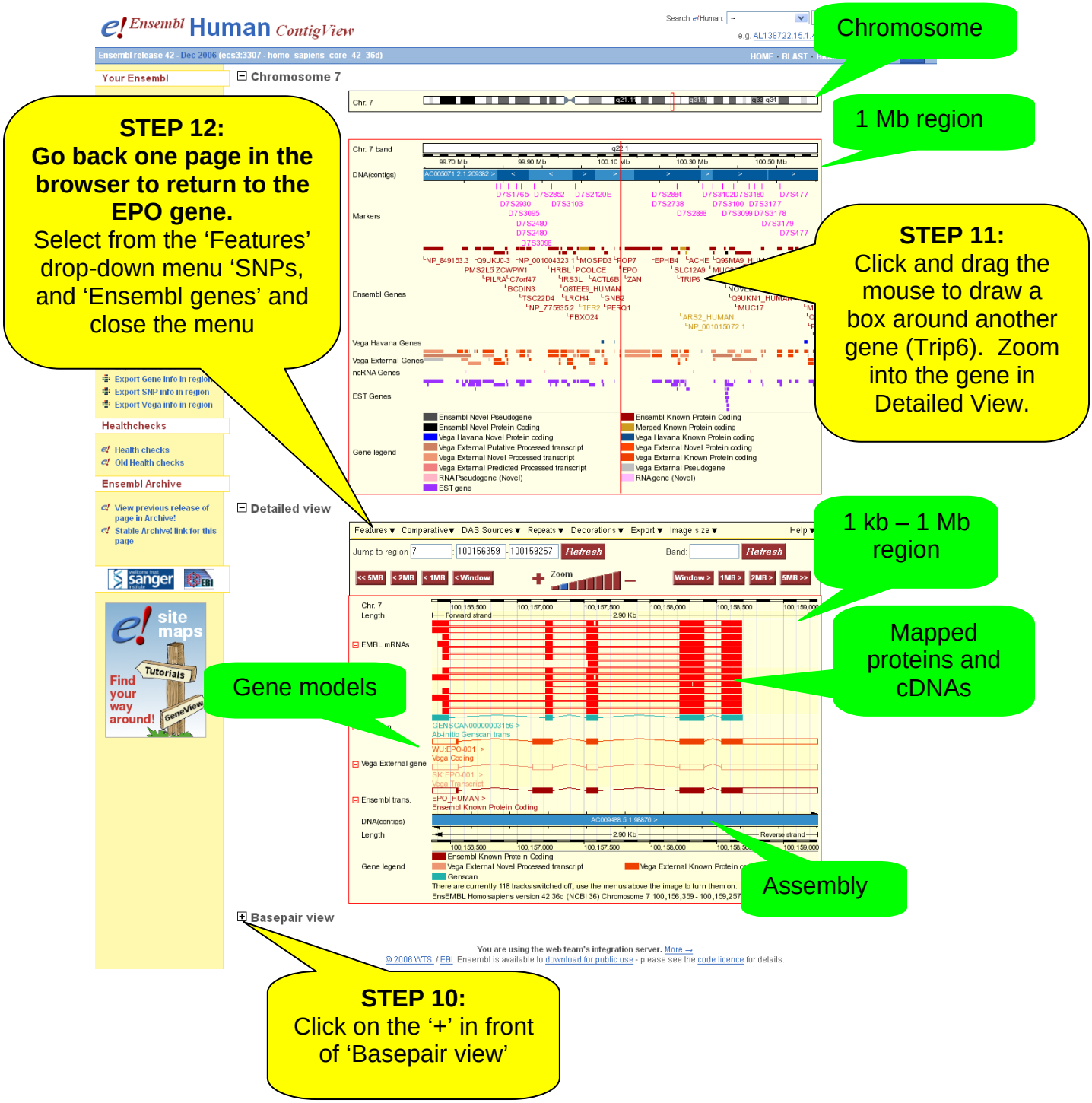
Result of STEP 7:

Result of STEP 8:



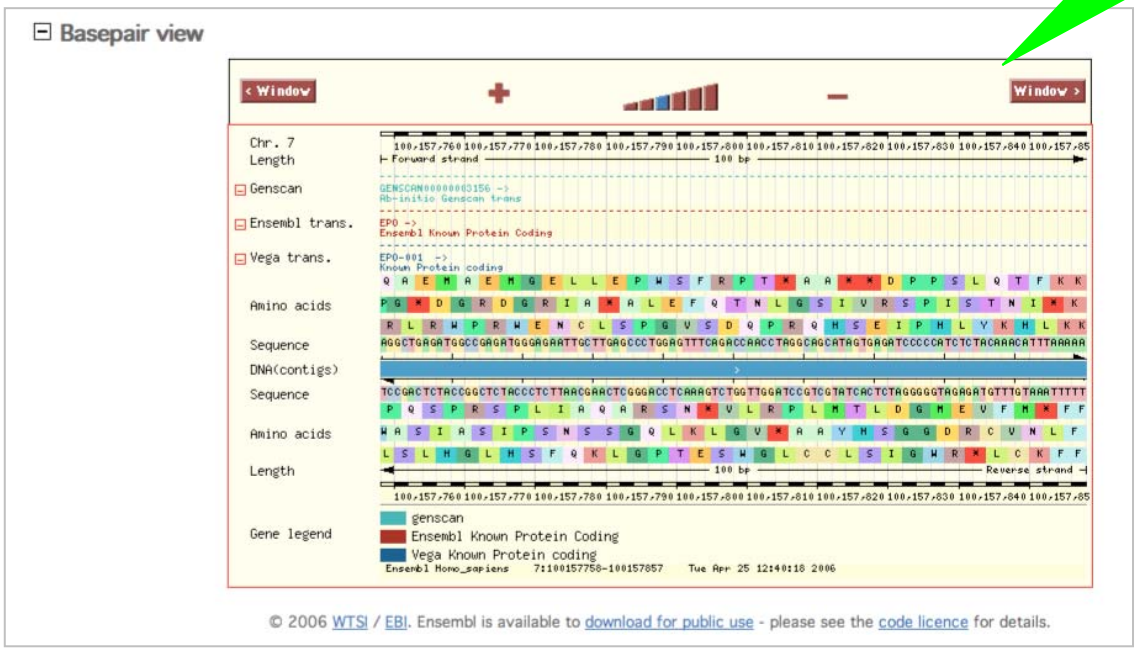
No.	Exon / Intron	Chr	Strand	Start	End	Start Phase	End Phase	Length	Sequence
	5' upstream sequence								.....tttctagaaca attgtgggaagggagaccactcatttgcccctccctaaagcttctgggctt gctacttgcggaactcagcaaccaggcatctctgagtctccgcccaagacc ccccaggggaggtgtccgggagccagcctttccagatagcacgctccgcca agggtgcgcaaccggctgcactccctcccgcgacccagggcccgggagcagc acccacacgcacgtctgcagcagcccgctcacgccccggcgagcctcaac ctgccctgctctgaccccgggtggccccctaccctggcgaccctcacgcacacagcct ctccccacccccacccgcgcacgcacacatgcagataacagccccgaccccccggccaga gccgcagagtccctgggcccaccccggccgctcgctgcgctgcgccgcaccgcgctgtcct
1	ENSE00001130431	7	1	100,156,359	100,156,552	-	1	194	CCCGGAGCCCGACCGGGGCCACCGCCGCCCCCTCTGCTCCGACACCGCGCCCCCTGGACAG CCGCCCTCTCCTCCAGGCCCGTGGGGCTGCCCCTGCACCGCCGAGCTTCCCCGGATGAGG GCCCCCGGTGTGGTCACCCGGCGCGCCCCAGGTCGCTGAGGGACCCCGGCCAGGCGCGGA GATGGGGGTGCACG
	Intron 1-2	7	1	100,156,553	100,157,1			564	gtgagtactcgcgggctgggcgctcccgcccgcccgggtccctgtttgagcggggattta gcgccccggctattggccaggaggtggctgggttcaaggaccggcgacttgtcaaggacc ccggaagggggaggggggtgggggcagcctccacgtgccagcggggactttgggggagtcct tggggatggcaaaaacctgacctgtgaaggggacacagtttgggggttgagggggaagaag gtttgggggttctgctgtgccagtggagaggaagctgataagctgataacctgggcgctg gagccaccacttatctgccagaggggaagcctctgtcacaccaggattgaagtttggccg gagaagtggatgctggtagctgggggtgggggtgtgcacacggcagcaggattgaatgaag gccagggaggcagcacctgagtgcttgcatggttgggggacaggaaggacgagctgggggca gagacgtgggggatgaaggaagctgtccttccacagccacccttctccctcccgcctgac tctcagcctggctatctgttctag
2	ENSE00001144077	7	1	100,157,117	100,157,262	1	0	146	AATGTCCTGCCTGGCTGTGGCTTCTCCTGTCCCTGCTGTCGCTCCCTCTGGGCCTCCCAG TCCTGGGCCCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGAGAGGTACCTCTTGG AGGCCAAGGAGGCCGAGAATATCACG
	Intron 2-3	7	1	100,157,263	100,157,520			258	gtgagacccttccccagcacattccacagaactcacgctcagggcttcagggaactcct cccagatccaggaacctggcacttggtttgggggtggagttgggaagctagacactgcccc cctacataagaataagtctggtggccccaaaccatacctggaaactaggcaaggagcaaa gccagcagatcctacggcctgtgggccagggccagagccttcagggacccttgactcccc gggctgtgtgcatttcag
3	ENSE00001130423	7					0	87	ACGGGCTGTGCTGAACACTGCAGCTTGAATGAGAATATCACTGTCCCAGACACCAAAGTT AATTTCTATGCCTGGAAGAGGATGGAG
	Intron 3-4	7						615	gtgagttccttttttttttttttccttttctttggagaatctcatttgcgagcctgatt ttggatgaaagggagaatgatcgagggaaaggtaaaatggagcagcagagatgaggctgc ctgggcgcagaggctcacgtctataatcccaggctgagatggcgagatgggagaattgc ttgagccctggagtttcagaccaacctaggcagcatagtgagatccccatctctacaaa catttaaaaaaattagtcaggtgaggtggtgcatggtggtagtcccagatatttggaagg ctgaggcgggaggatcgcttgagcccaggaatttgaggctgcagtgagctgtgatcacac cactgcactccagcctcagtgacagagtgaggccctgtctcaaaaaagaaaagaaaaag aaaaataatgaggggctgtatggaatacattcattattcattcactcactcactcactcac tcattcattcattcattcaacaagtcttattgcatacctctgtttgctcagcttg gtgcttggggctgctgaggggcaggagggagagggtgacatgggtcagctgactcccaga gtccactccctgtag
4	ENSE00001130416	7	1	100,158,223	100,158,402	0	0	180	GTCGGGCAGCAGGCCGTAGAAGTCTGGCAGGGCCTGGCCCTGCTGTCGGAAGCTGTCCTG CGGGGCCAGGCCCTGTTGGTCAACTCTTCCCAGCCGTGGGAGCCCCTGCAGCTGCATGTG GATAAAGCCGTCAGTGGCCTTCGCAGCCTCACCACTCTGCTTCGGGCTCTGGGAGCCCAG
	Intron 4-5	7	1	100,158,403	100,158,536			134	gtgagtaggagcggacacttctgcttgccctttctgtaagaaggggagaagggtcttgct aaggagtacaggaactgtccgtattccttccctttctgtggcactgcagcgacctcctgt tttctccttggcag
5	ENSE00000894545	7	1	100,158,537	100,159,257	0	-	721	AAGGAAGCCATCTCCCCTCCAGATGCGGCCTCAGCTGCTCCACTCCGAACAATCACTGCT GACACTTTCCGCAAACTCTTCCGAGTCTACTCCAATTTCCTCCGGGGAAAGCTGAAGCTG TACACAGGGGAGGCCTGCAGGACAGGGGACAGATGACCAGGTGTGTCCACCTGGGCATAT CCACCACCTCCCTCACCAACATTGCTTGTGCCACACCCTCCCCCGCCACTCCTGAACCCC GTCGAGGGGCTCTCAGCTCAGCGCCAGCCTGTCCCATGGACACTCCAGTGCCAGCAATGA CATCTCAGGGGCCAGAGGAACTGTCCAGAGAGCAACTCTGAGATCTAAGGATGTCACAGG GCCAACTTGAGGGCCCAGAGCAGGAAGCATTCAGAGAGCAGCTTTAAACTCAGGGACAGA GCCATGCTGGGAAGACGCCTGAGCTCACTCGGCACCCTGCAAAATTTGATGCCAGGACAC GCTTTGGAGGCGATTTACCTGTTTTCGCACCTACCATCAGGGACAGGATGACCTGGATAA CTTAGGTGGCAAGCTGTGACTTCTCCAGGTCTCACGGGCATGGGCACTCCCT AGAGCCCCCTTGACACCGGGGTGGTGGGAACCATGAAGACAGGATGGGGGCTGGCCTCTG GCTCTCATGGGGTCCAAGTTTTGTGTATTCTTCAACCTCATTGACAAGAACTGAAACCAC C
	3' downstream sequence								aatatgactcttggcttttctgttttctgggaacctccaaatccccctggctctgtcccac tcctggcagcagtgcagcaggtccaggtccgggaaacgagggggtggaggggggctgggccc tacgtgctgtctcacacagcctgtctgacctctcgaccctaccgggcctgaggccacaag ctctgcctacgctggtcaataaggtggctccattcaaggcctcaccgcagtaaggcagct gccaacctgcccagggcaaggctgcagtgcgctgagattgtcatcaaggagagggaggc cagaggacgggtcctttgggagttttgggggctggtaacagctgccaaccctggtcact ggcccttgttaatttctgcctcttctttggtctcttcggctgatgccacactcccagcat cagccttccgcctcaatgcacatcctcgctaagtccagctgcctccttactcacctggg gactgtcaccatgaatactc.....

Result of STEP 9:

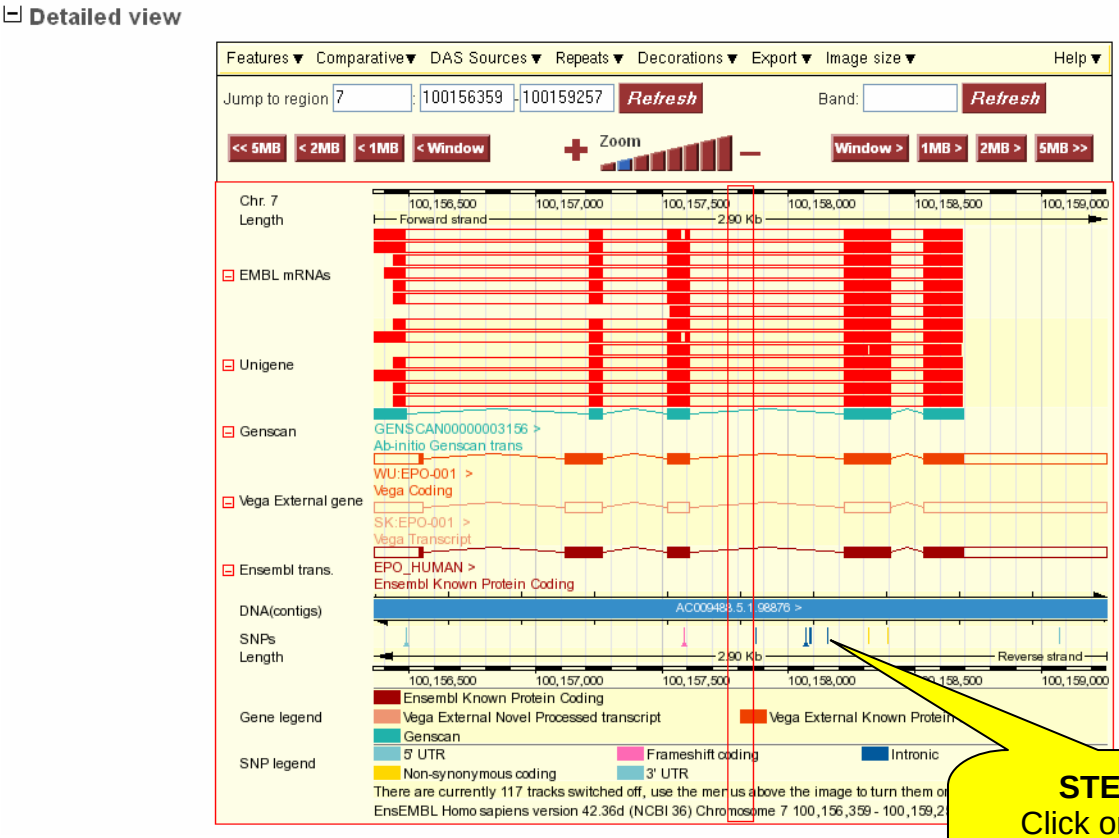


Result of STEP 10:

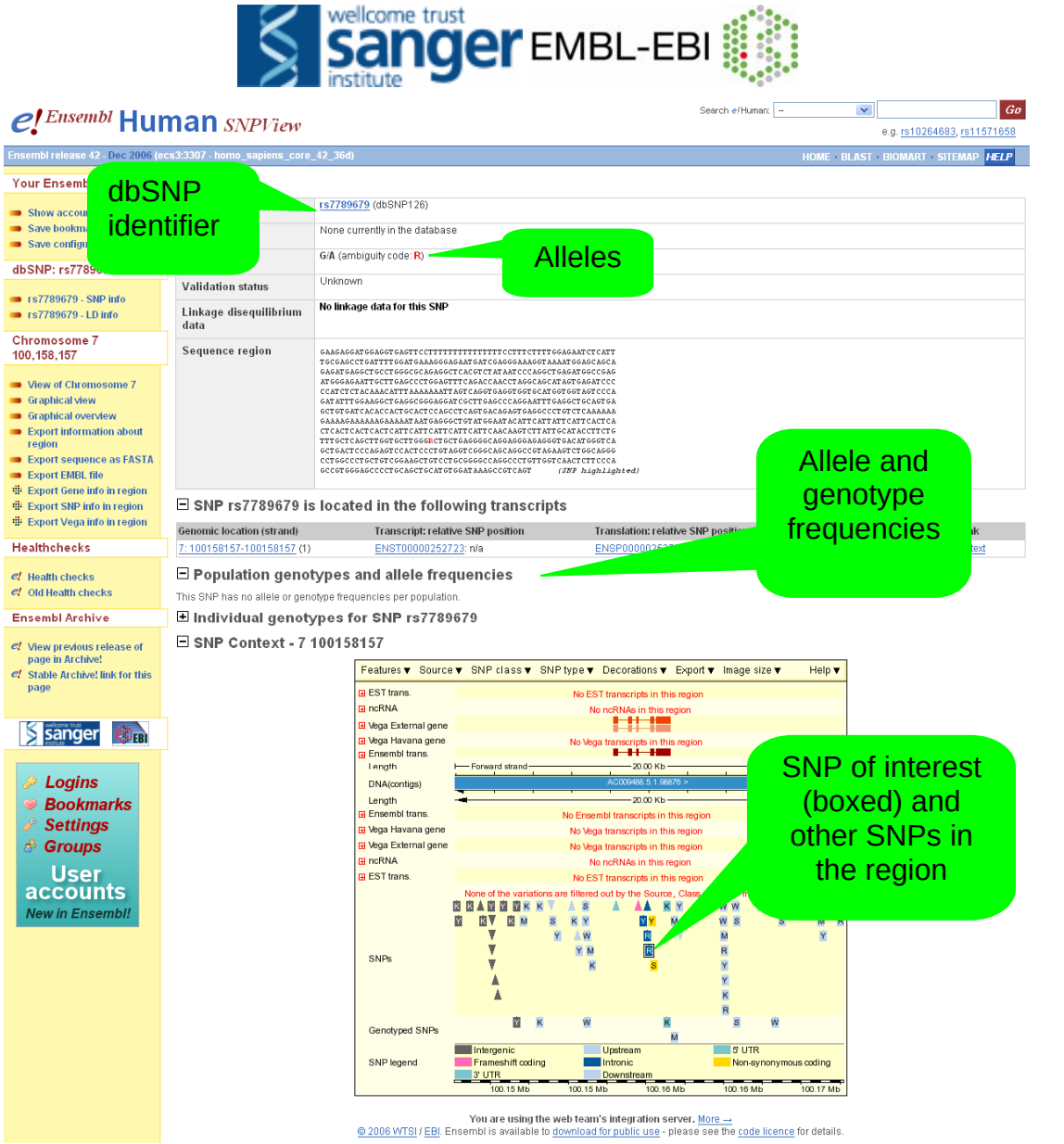
25 – 500 bp region



Result of STEP 12:

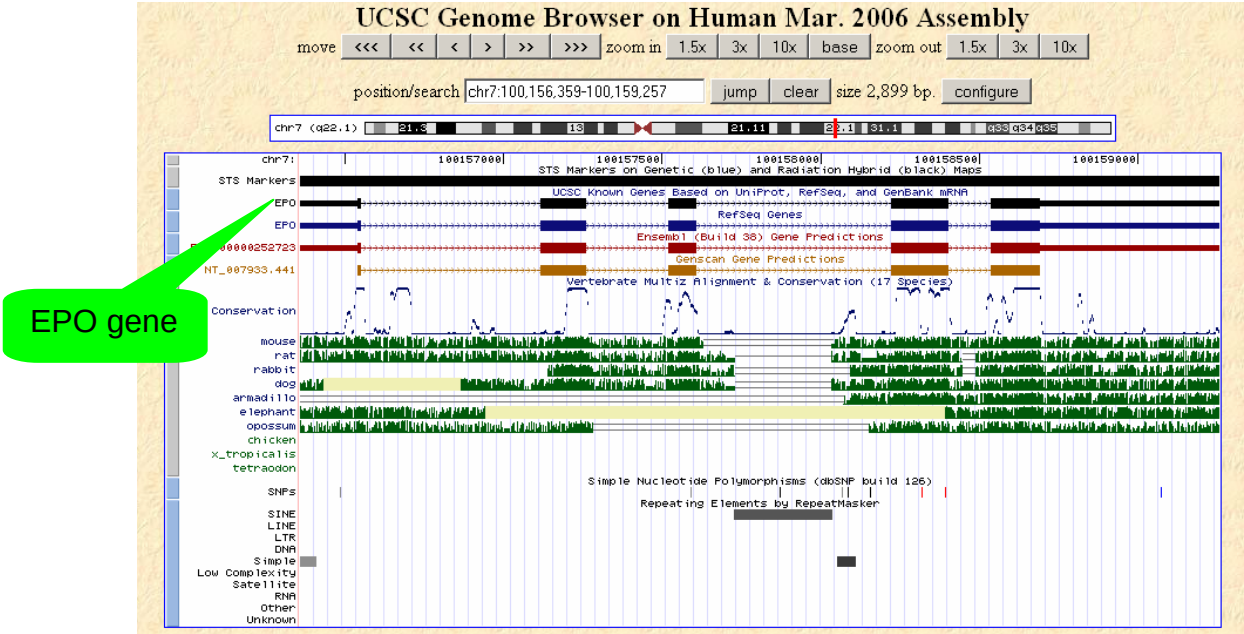


STEP 13:
Click on a SNP (vertical line) and subsequently on 'SNP properties'



STEP 14:
Go back to ContigView with the back button of the internet browser.

STEP 15:
To see the same chromosomal region in the UCSC genome browser, click on ‘Show in UCSC browser’ on the left of the page. A new window will open.

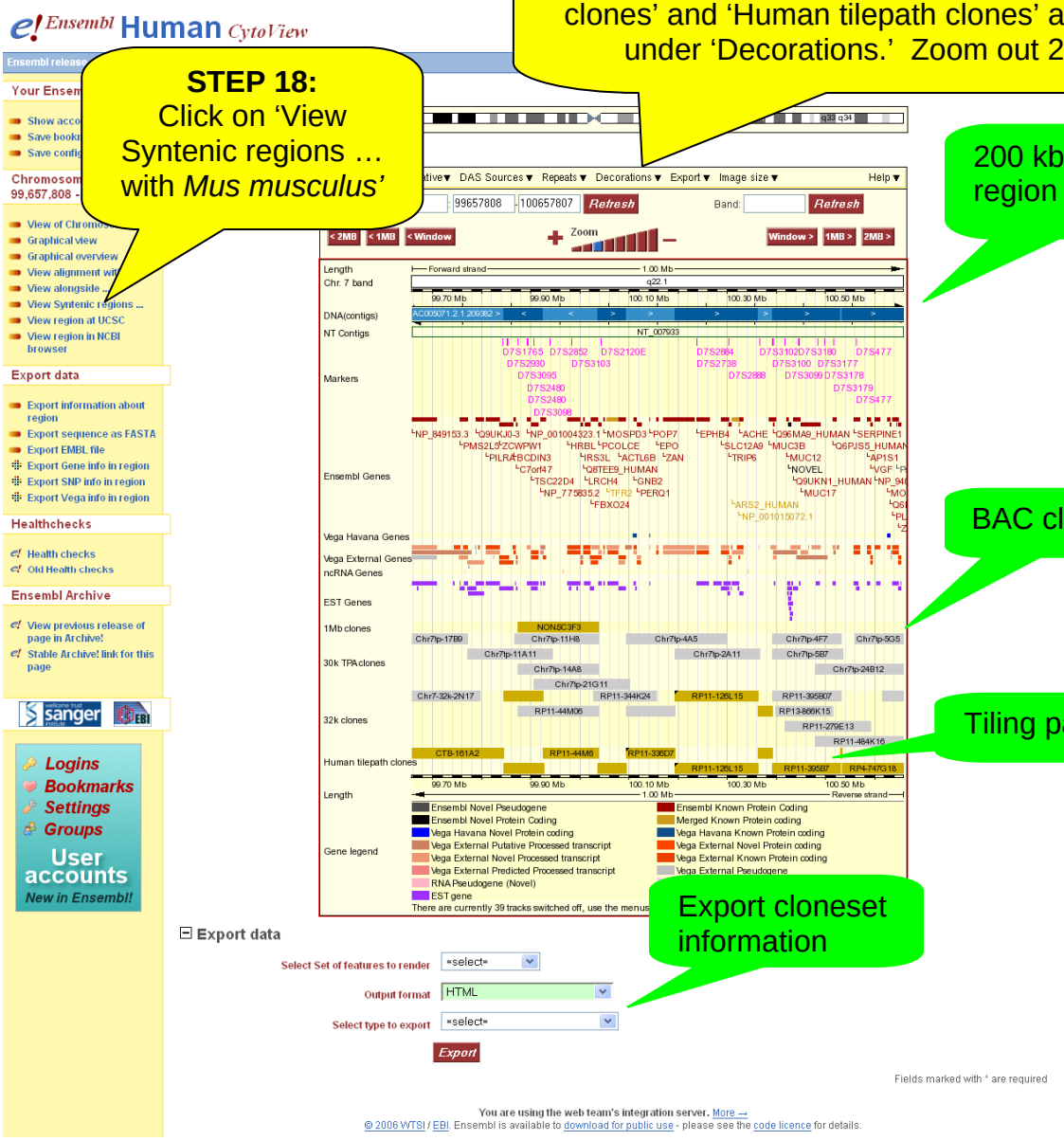


STEP 16:
Once you see the EPO gene and close this window. (You can turn on ‘Ensembl genes’ by changing ‘hide’ to ‘full’)

Click on ‘Graphical Overview’ on the left hand of the ContigView page to reach CytoView.

STEP 17:
Make sure ‘1Mb clones, ‘30k TPA clones’, ‘32k clones’ and ‘Human tilepath clones’ are selected under ‘Decorations.’ Zoom out 2 steps.

STEP 18:
Click on ‘View Syntenic regions … with *Mus musculus’*

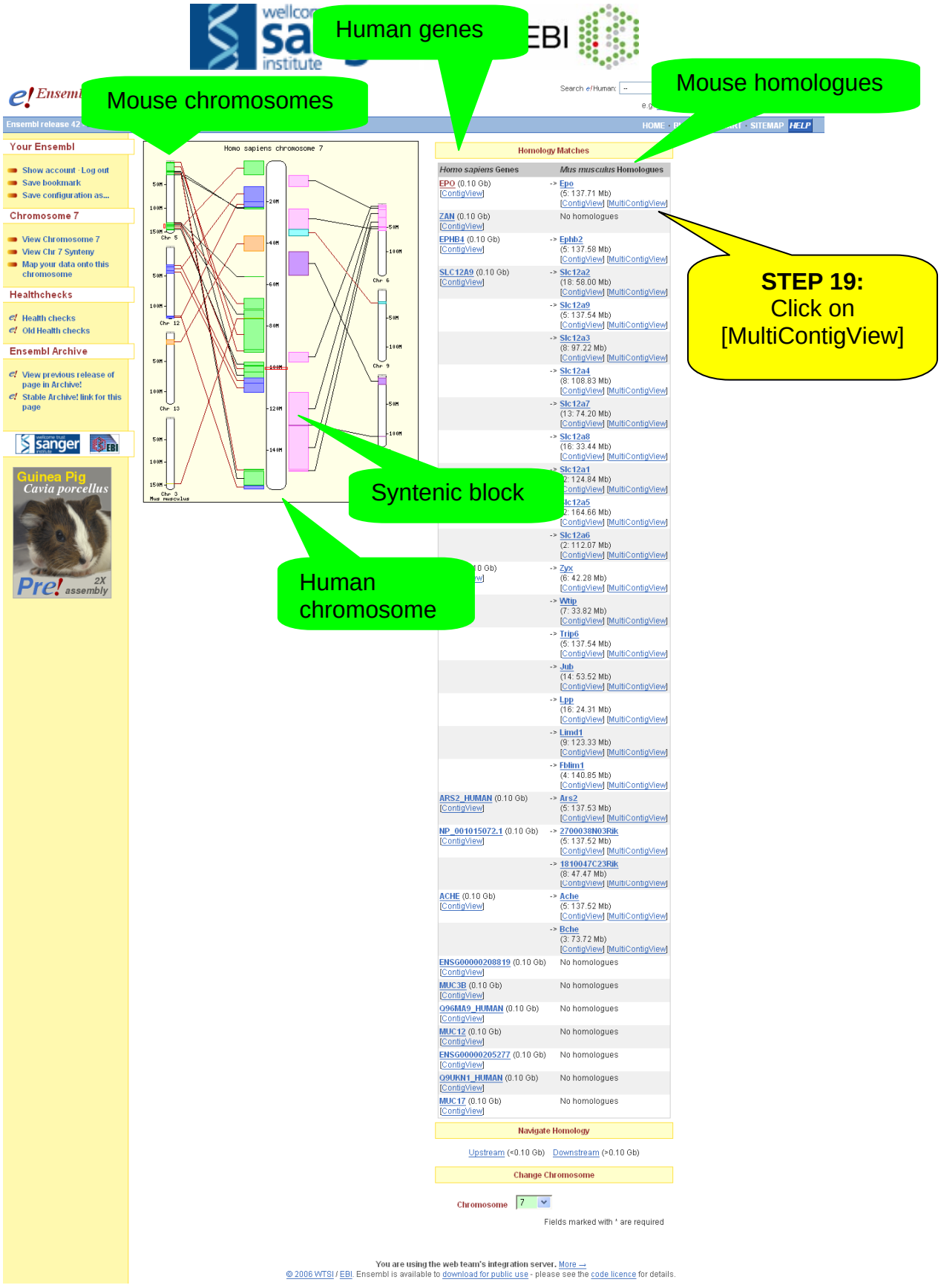


200 kb – 50 Mb region

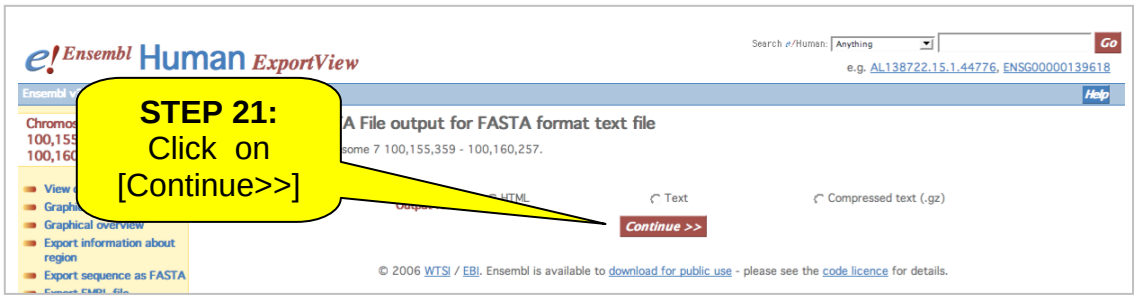
BAC clones

Tiling path clones

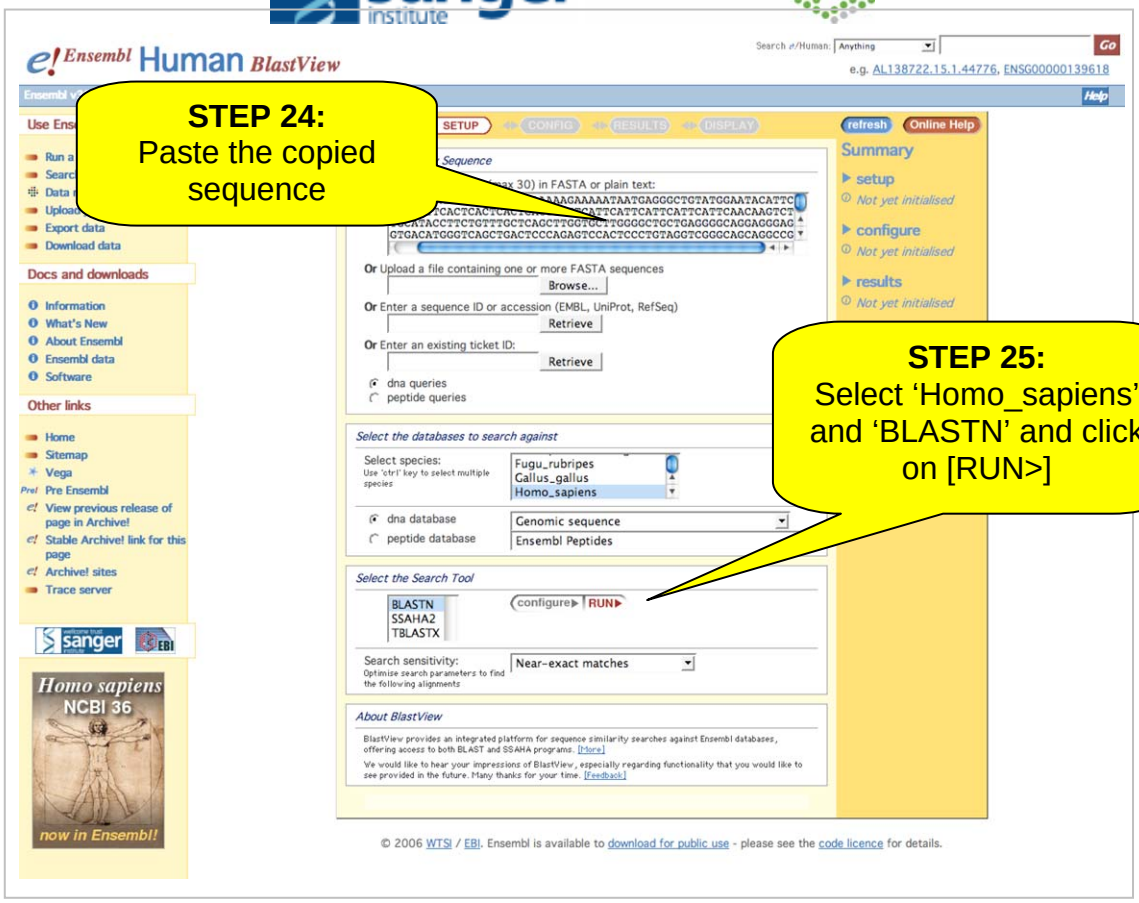
Export cloneset information

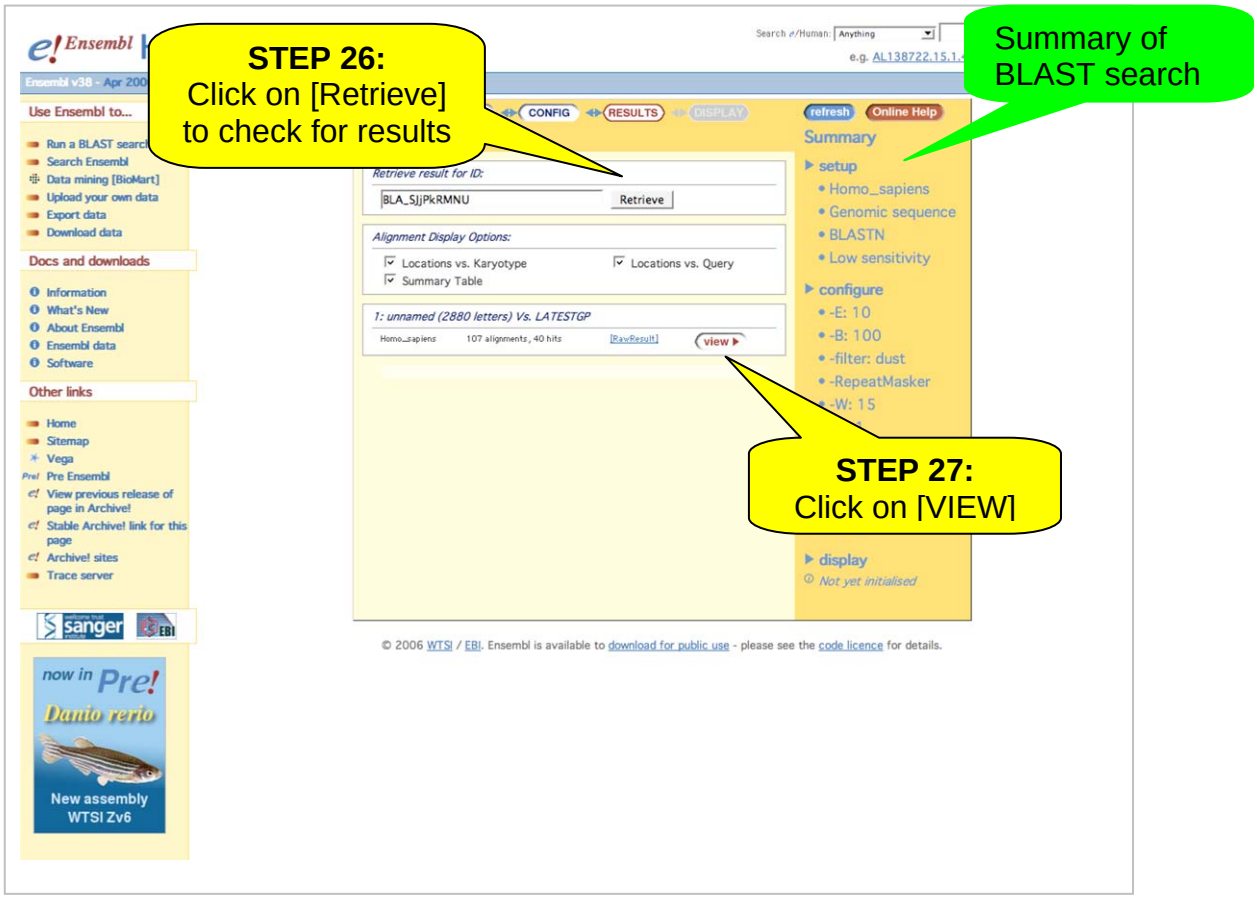

Human genes
Mouse homologues
Mouse chromosomes
Syntenic block
Human chromosome
STEP 19:
Click on
[MultiContigView]
Homo sapiens chromosome 7
Homology Matches
Homo sapiens Genes
Mus musculus Homologues
Navigate Homology
Change Chromosome
Your Ensembl
Chromosome 7
Healthchecks
Ensembl Archive
Guinea Pig
Cavia porcellus
Pre! 2X assembly

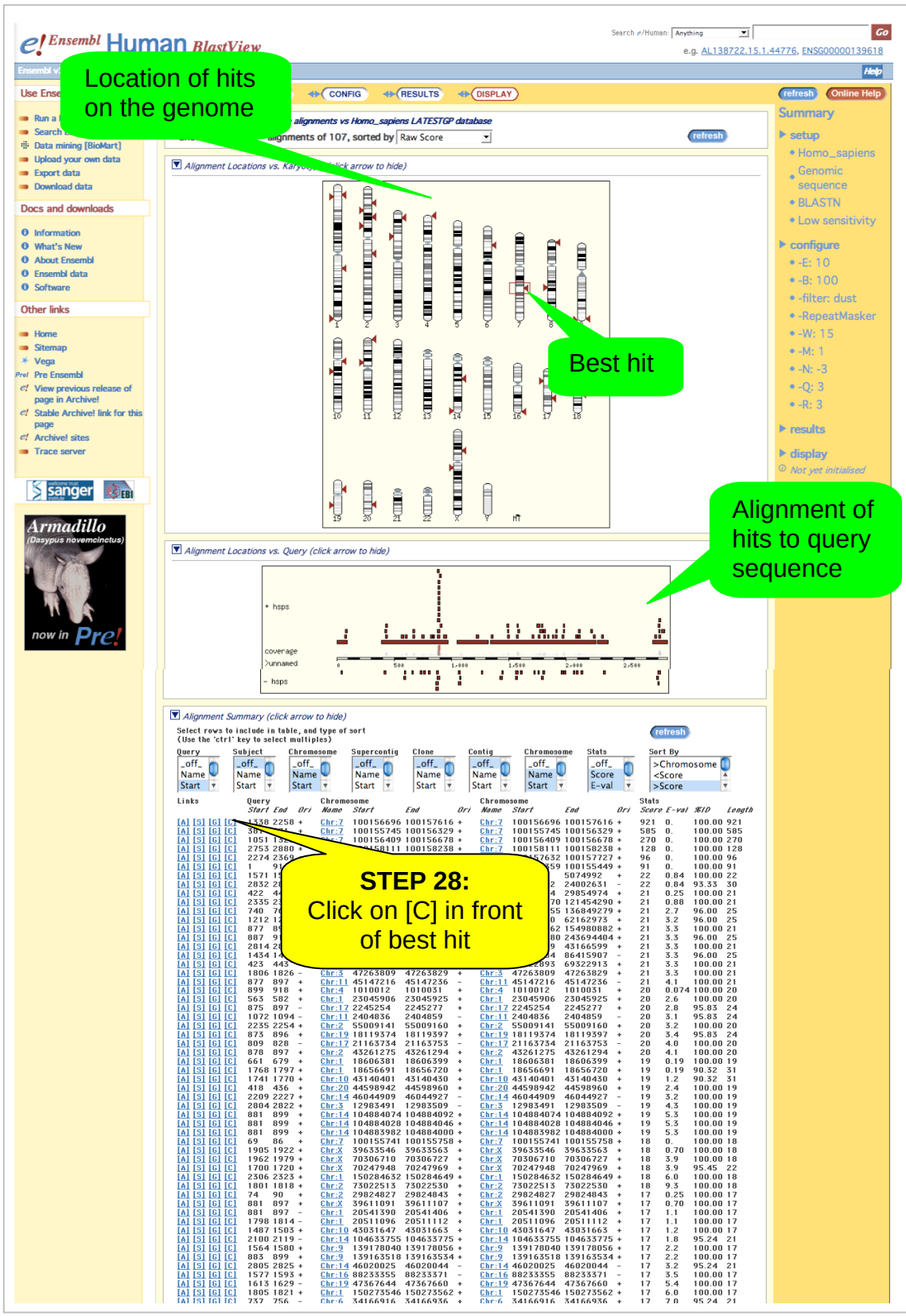

Ensembl Human MultiContigView
Ensembl release 42 - Dec 2006 (ecs3:3307 - homo_sapiens_core_42_36d)
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Show account · Log out
Save bookmark
Save configuration as...
Chromosome 7
100,155,359 - 100,160,257
View of Chromosome 7
Graphical view of...
Graphical overview
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region in NCBI browser
Export data
Export information about region
Export sequence as FASTA
Export EMBL file
Top level
Hsap Chr. 7
Mmus Chr. 5
Navigational overview
Hs. Chr. 7 band
DNA(contigs)
Ensembl Genes
Vega Havana Genes
Vega External Genes
ncRNA Genes
EST Genes
Mm. Chr. 5 band
Detailed View
Features
Comparative
Repeats
Decorations
Jump to region
< 2MB
< 1MB
< Window
Zoom
Hsap Chr. 7
Length
Forward strand
Vega External gene
Ensembl trans.
Reverse strand
Mmus Chr. 5
Markers
STEP 20:
Click on 'Export sequence as FASTA'
Vega EPO genes
Ensembl EPO gene (Human)
Human
Mouse
Mouse EPO homologue
© 2006 WTSI / EBI. Ensembl is available to


Ensembl Human ExportView
Search e!Human: Anything Go
e.g. AL138722.15.1.44776, ENSG00000139618
Help
A File output for FASTA format text file
some 7 100,155,359 - 100,160,257.
Text
Compressed text (.gz)
Continue >>
Graphical overview
Export information about region
Export sequence as FASTA
© 2006 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.
STEP 21:
Click on [Continue>>]


Ensembl Human ExportView
Search e!Human: Go
e.g. AL138722.15.1.44776, ENSG00000196160
Ensembl release 42
HOME · BLAST · BIOMART · SITEMAP · HELP
Your Ensembl
Show account · Log
Save bookmark
Save configuration as...
Chromosome 7
100,155,359 - 100,160,257
View of Chromosome 7
Graphical view
Graphical overview
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region in NCBI browser
Export data
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
Healthchecks
Health checks
Old Health checks
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Saccharomyces cerevisiae
Yeast
Revised assembly
STEP 22:
Select and copy a part of the sequence
STEP 23:
Click on 'BLAST'
You are using the web team's integration server. More →
© 2006 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

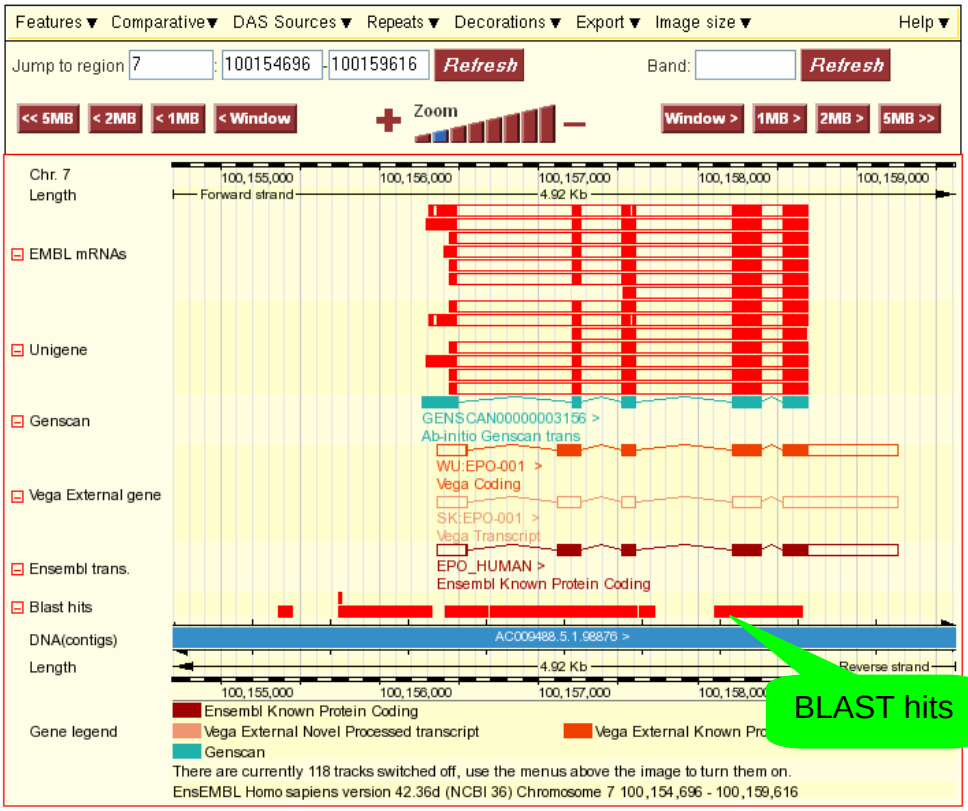

STEP 24:
Paste the copied sequence
STEP 25:
Select 'Homo_sapiens' and 'BLASTN' and click on [RUN>]
Ensembl Human BlastView
Sequence
Or Upload a file containing one or more FASTA sequences
Browse...
Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)
Retrieve
Or Enter an existing ticket ID:
Retrieve
dna queries
peptide queries
Select the databases to search against
Select species:
Fugu_rubripes
Gallus_gallus
Homo_sapiens
dna database
Genomic sequence
peptide database
Ensembl Peptides
Select the Search Tool
BLASTN
SSAHA2
TBLASTX
configure
RUN
Search sensitivity:
Near-exact matches
About BlastView
Summary
setup
configure
results
Homo sapiens
NCBI 36
now in Ensembl!


STEP 26:
Click on [Retrieve] to check for results
Summary of BLAST search
STEP 27:
Click on [VIEW]
Retrieve result for ID:
BLA_SJjPkRMNU
Retrieve
Alignment Display Options:
Locations vs. Karyotype
Locations vs. Query
Summary Table
1: unnamed (2880 letters) Vs. LATESTGP
Homo_sapiens
107 alignments, 40 hits
[RawResult]
view
Summary
setup
Homo_sapiens
Genomic sequence
BLASTN
Low sensitivity
configure
-E: 10
-B: 100
-filter: dust
-RepeatMasker
-W: 15
display
now in Pre!
Danio rerio
New assembly
WTSI Zv6


Location of hits on the genome
Best hit
Alignment of hits to query sequence
STEP 28:
Click on [C] in front of best hit

Back in the contigview page…

Detailed view



END of the Worked Example

EXERCISES and ANSWERS

Note: the answers to these exercises may change upon new releases of Ensembl. If you use these exercises in the future, please use the appropriate archive site.

1. Exploring features related to a gene

(a) Search for the human TAC1 gene by typing 'human TAC1 gene' in the search window.

(b) How many transcripts are predicted for this gene? What is the size of the longest predicted mRNA? How many exons does it have? How many amino acids does it code for?

(c) Look up 'Ensembl genes' in the glossary:
Follow the 'Help and Documentation' link, 'HelpDesk section' to this url:
http://www.ensembl.org/Homo_sapiens/glossaryview

Follow some of the links in the 'Similarity Matches' section of GeneView. What is a possible function of TAC1?

(d) Which InterPro domains does the protein product contain?

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

(f) In which chromosomal band and on which clone and contig in the genomic sequence assembly is the TAC1 gene located?

(g) Go back to GeneView by clicking on 'TAC1' in the Overview panel and following the link for the gene. Is there a putative mouse orthologue? If so, where is it in the mouse genome?

2. Exploring a region

(a) Display the region between markers D12S764 and D12S1871 in ContigView. Start on the human homepage, and click on chromosome 12.

(b) How many contigs are used to make this portion of the assembly? View the human tile path clones. Do they correspond to the assembly?

(c) What is the closest marker to the TENC1 gene? How many synonyms does this marker have?

(e) Zoom in (towards the '+') three steps on the zoom triangle and turn on the SNP track. Identify an intronic SNP and look at the corresponding SNPView page.

3. Exploring the zebrafish (*Danio rerio*) genome with Ensembl

(a) Bring up a ContigView display of zebrafish (*Danio rerio*) chromosome 1 between 64.0 Mb and 64.5 Mb.

(b) 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 ZFIN site.

(c) Can you find out anything about the possible functions of one of the novel genes? For this, try looking at homologues in other species, at other members of protein families and InterPro domains.

Answers (Browsing Ensembl)

1. Exploring features related to a gene

(a) A 'Vega' gene and 'Ensembl' gene will be shown. VEGA (Verterbrate Genome Annotation) is a consortium of manual curators for certain chromosomes in human, mouse, zebrafish, pig and dog. However, we would like to explore the 'Ensembl Gene: ENSG00000006128'. To ascertain it is indeed the TAC1 gene check that the HGNC symbol (the 'official' gene name given by the HUGO Gene Nomenclature Committee) is 'TAC1'. Click on the 'Ensembl Gene: ENSG00000006128' link to go to the GeneView page for this gene.

(b) The TAC1 gene (ENSG00000006128) has 3 predicted transcripts, ENST00000319273, ENST00000346867 and ENST00000350485. Scroll down to the 'Transcript' sections for more information about these transcripts. The longest transcript is ENST00000319273. The length of this transcript is 1060 bp. It has 7 exons and codes for 129 aa.

(c) The TAC1 gene is Protachykinin 1 precursor. Follow the links to MIM and EntrezGene or UniProt/Swiss-Prot in the 'Similarity Matches' section to learn more. Choose 'UniProt' under 'DAS Sources' to see references in the literature (click 'Update' after making the selection). Also the GO (Gene Ontology) and InterPro sections can give you clues about the biological and molecular function of the TAC1 protein. Tachikinins are neuropeptides. These hormones are thought to function as neurotransmitters which interact with nerve receptors and smooth muscle cells. They are known to induce behavioral responses and function as vasodilators and secretagogues.

(d) Check the 'InterPro' section in GeneView. The domains include IPR013055 (Tachykinin/Neurokinin like), IPR002040 (Tachykinin/Neurokinin), IPR008215 (Tachykinin) and IPR008216 (Protachykinin).

(e) Clicking on a GO identifier gives you a GOView page (loading of the page can take a while) 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.

(f) Go back to GeneView and click the 'Graphical View' link in the side menu to go to ContigView. In the 'Overview' panel you can see that TAC1 is located on band 7q21.3 ('Chr.7 band' track). In the 'Detailed view' panel you can see that it is located on contig AC004140.2.1.74918 ('DNA(contigs)' track). If you click on the contig and follow the link to the EMBL source (or if you turn on the 'Human tile path clones' track from the 'Decorations' menu of ContigView) you can see that this sequence is derived from clone RP5-841B21.

(g) In GeneView, ENSMUSG00000061762 (Tac1) is named in the 'Orthologue Prediction' section. Click on it to go to its GeneView page to find that it is located on mouse chromosome 6.

2. Exploring a region

(a) Start on the homepage for human and click on chromosome 12 to go MapView. In the 'Jump to ContigView' section choose 'From (type): Marker D12S764 To (type): Marker D12S1871' and click [Go]. This leads you to ContigView.

(b) The displayed region in the Overview panel is larger than the area between the two markers. The red line or small box is drawn over the first marker (D12S764). Zoom into the region between the two markers by drawing a box with the mouse around it.

The region will be displayed below, in 'Detailed View'. This region includes sequence from 4 different contigs (one is quite small), displayed in light blue and dark blue in the 'DNA(contigs)' track. To see also the 4 clones that make up this region, select the 'Human tilepath clones' track from the 'Decorations' menu. Clones are shown in gold and pink. Portions of the 'Tile path clones' were used to form the assembly and correspond to 'contigs'. The clones overlap each other whereas the contigs don't.

(c) Marker D12S2110 is closest to the TENC1 gene. Click on the marker (i.e. on the vertical bar representing it, not on its name) and follow the link 'Marker info' to the MarkerView page. There are 2 synonyms listed.

(d) SNPs can be turned on using the 'Features' menu. Coding SNPs are shown in yellow (non-synonymous) and green (synonymous), intronic SNPs are dark blue. Click on a SNP. Be careful to click exactly on the vertical bar representing the SNP, otherwise you will get the wrong pop-up menu. Follow the link 'SNP properties' to the SNPView page. Note the 'SNP Context' display in SNPView.

3. Exploring the zebrafish (*Danio rerio*) genome with Ensembl

(a) Start on the homepage for zebrafish (*Danio rerio*). In the 'Karyotype' section choose 'Chromosome: 1', 'From (type): Base pair: 64000000 To (type) Base pair: 64500000' and click [Go]. This leads you to ContigView for a larger region. Type in the base pairs in Detailed View (change the second number to 64500000).

(b) On the 'Overview' panel of ContigView Ensembl known and novel genes are displayed in the 'Ensembl Genes' track in reddish brown and black, respectively. Known genes are Ensembl gene predictions that match species specific entries in the UniProt and/or RefSeq database, while novel genes map back to entries from other species. There are 12 known and 9 novel genes. Click on one of the known genes to go to its GeneView page and explore the links in the 'Similarity Matches' section.

(c) To find out more about the possible function of a novel gene there are many options. Click on the gene in ContigView to go to its GeneView page. If the gene has orthologues in other species (shown in the 'Orthologue Prediction' section) and these orthologues are better characterized than the novel zebrafish gene this can give a clue about the possible function of this gene. If the gene belongs to a family (shown in the 'Protein Family' section) other family members may provide information. InterPro domains (shown in the 'InterPro' section) may also provide clues.

BIOMART

Exercises

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

Note: As this is the first exercise we walk you this time through BioMart step-by-step (but of course you can also try to do this exercise without our help!)

Go to 'www.biomart.org' and click on 'central server.'

Click on **Dataset**. Choose the database and the species for your query as follows:
- Select 'Ensembl 42' as we are looking for a gene list (NOT a SNP list!).
- Select 'Homo sapiens genes (NCBI36)'.

Click on '**Filters**' at the left. Filter this dataset to select your genes of interest as follows:
- Expand the 'REGION' section at the right by clicking on the '+'. Select 'Chromosome 2'. Click [count] at the top of the panel and note the number of Ensembl genes in *Homo sapiens* chromosome 2.
- In the 'GENE' section, select 'Status: NOVEL'.
- In the 'PROTEIN' section, select the second 'Limit to genes with these family or domain IDs' option. Select 'InterPro ID(s)' and enter 'IPR000276' in the box. Click [count] again and note that the number of genes is updated.

Click on '**Attributes**' (at the left). Select the output for your gene list as follows:
- Select the 'SNPs' Attribute Page.
- In the 'Gene' section 'Ensembl Gene ID' and 'Ensembl Transcript ID' are selected by default – also select 'Ensembl Peptide length'.
- In the 'Gene associated SNPs' section select 'Reference ID', 'Allele', 'Peptide location (aa)', 'Location in Gene (coding etc)', 'Synonymous Status' and 'Peptide Shift'.

Click on '**Results**' (at the top) to obtain the first 10 rows of your table. To obtain the entire table in excel, change the option at the top (display maximum rows) from 'HTML' to 'XLS' and 'Export all results to File'. Click 'Go'.

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 many affect the amino acid sequence? Are there any indels (insertions/deletions)?

2. Retrieve the sequences of the exons of the human MEFV gene in FASTA format.

3. Retrieve the gene structure (i.e. start and end coordinates of exons) of the mouse gene ENSMUSG00000042351.

4. Retrieve all human disease genes containing transmembrane domains located between p11.2 and q22.

5. The file http://www.ebi.ac.uk/~xose/Affy_exercise.txt contains a list of probeset IDs from a microarray experiment using the Affymetrix array HG-U133 Plus 2.0 (human). Retrieve the 500 bp upstream of the transcripts matching these probeset IDs.

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

7. Retrieve all human SNPs that have an ID from The SNP Consortium (TSC), from chromosome 6 between 15 Mb and 15.2 Mb, with 200 bases flanking sequence.

8. We will walk you through this advanced exercise using the comparative features of BioMart.

Retrieve the mouse homologues of *Homo sapiens* genes CASP1, CASP2, CASP3, and CASP4.

Dataset: 'Ensembl Homology' and the default options (human and mouse).
Secondary Dataset (click on 'Dataset' at the bottom left).
Linked: *Homo sapiens* genes

Filters (in the secondary dataset): **GENE:** 'ID list limit HGNC Symbol(s)'.
Enter the human HGNC (HUGO) symbols in the box: CASP1, CASP2, CASP3, and CASP4

Return to the first Dataset (panel in upper left) and choose Attributes:
Under 'Features' select 'Mus musculus gene' and select 'Ensembl gene ID' and 'Description'.
Results display the mouse homologues of the human CASP genes.

9. Design your own query!

Answers (BioMart)

Warning: if there has recently been a new Ensembl release based on an updated gene set and/or an updated SNP set, these answers may have changed - please check with one of the instructors.

1. You should find three novel genes on chromosome 2 with this InterPro domain. The result set has three transcripts and a total of 303 rows of output. All transcripts have one or more coding SNPs ('Location in Gene' is 'coding'), most of which are non-synonymous ('Synonymous status' is 'no') and thus affect the amino acid sequence of the encoded peptide. There are various indels ('Allele' is '-/' or '/-').

2. Click '**NEW**' to begin a new query.

Choose under **Dataset** 'Ensembl 42' and 'Homo sapiens genes (NCBI36)'.

Select under **Filters**: (the '**GENE**' section) 'ID list limit HGNC Symbol(s)' and enter 'MEFV'.

Choose the **Attributes** Page '**Sequences**'. Under the 'SEQUENCES' panel, select 'Exon sequences (Gene)'. Under 'Header Information' add 'Ensembl Exon ID' to the default options.

Click '**Results**' at the top.

You should find 10 exon sequences.

3. Dataset: 'Ensembl 42' and 'Mus musculus genes (NCBIM36)'.
Filters: GENE 'ID list limit Ensembl Gene ID(s)': enter the mouse gene ID.
Attributes 'Structures': select in the **EXON** panel: 'Ensembl Exon ID', 'Exon Start' and 'Exon End'.
Click '**Results**'.

You should find 8 exons. Take the link from the Ensembl Gene ID in your output back to the GeneView page to confirm the BioMart data with the gene structure displayed on this page.

4. Dataset: 'Ensembl 42' and 'Homo sapiens genes (NCBI36)'.
Filters: Region 'Chromosome 1', 'Band Start p11.2, 'Band End q22'
Gene: 'with Disease Association Only' (look under 'ID LIST FILTERS')
Protein: 'Transmembrane domains Only'.

Choose under **'Attributes'** the 'Features' menu and select 'GO ID' and 'GO description' along with the default options ('Ensembl Gene ID' and 'Transcript ID').

Results should show 5 Ensembl genes (multiple transcripts and GO terms).

5. Dataset: 'Ensembl 42' and 'Homo sapiens genes (NCBI36)'.
Filters: GENE: 'ID list limit' : Affy hg u133 plus 2 ID(s) and enter the list of probeset IDs.
Attributes: 'Sequences' select 'Flank(Transcript)', 'Upstream flank 500'. In the header, apart from the already default selected options, select 'Ensembl Transcript ID'.

You should find upstream sequences for the transcripts of 24 genes (Hint: click 'count' to see the number of genes!)

6. Dataset: 'Ensembl 42' and 'Homo sapiens genes (NCBI36)'.
Filters: REGION 'Marker' : Start D1S2806' End D1S464'
GENE: 'Status: KNOWN'.
Attributes 'Sequences' and select, apart from the already default selected options, 'Flank (Gene)' and 'Upstream flank 5000'.

You should find sequences for 25 genes.

When you choose the option 'Flank(Gene)' you will see only one upstream sequence per gene in the output. In the case where a gene has multiple transcripts, the upstream sequence of the transcript that extends the furthest at the 5' end is shown. If you want to export the upstream sequences for each transcript you should choose the option 'Flank (Transcript)'.

'Known' genes are Ensembl gene predictions that could be matched to same-species external database entries (e.g. UniProt/SwissProt) with a high similarity score (i.e. with BLAST or a similar sequence identity-matching program)

7. Database: 'ENSEMBL 42 VARIATION' and **Dataset:** 'Homo sapiens SNPs (dbSNP126;HGVbase 15; TSC 1; affy GeneChip Mapping Array)'.
Filters: REGION: 'Chromosome 6', 'Base pair Start 15000000', 'Base pair End 15200000'
GENERAL SNP FILTERS: ID list filters: 'SNPs with TSC ID(s) Only'.
Attributes 'Sequences': SEQUENCES : 'SNP sequences', 'Upstream flank 200', 'Downstream flank 200'.

You should find 69 SNPs.

COMPARATIVE GENOMICS

Exercises exploring homology and family information.

1. Main exercise: Explore a protein family in human, mouse and rat, identify putative orthologues, and explore regions of conserved synteny.

(a) Find the **GeneView** page for human SNX5.

(b) Examine the protein family.

-Take the link to the associated Protein Family.

Q1: How many human Ensembl genes produce peptides in this family?

Q2: Are they all 'known' genes?

Q3: Are there peptides in the same family for mouse (*Mus musculus*), rat (*Rattus norvegicus*) and zebrafish (*Danio rerio*)? (How many?)

Q4: What about invertebrate species?

Click on one of the rat peptides to go to rat **ProtView**.
From there take the link to the corresponding rat **FamilyView**.

Q5: How many rat Ensembl genes are part of this family? Does this number differ from the number of peptides belonging to this family you found before? Why?

Find your way to mouse **FamilyView**, and follow the link to mouse Snx5 (**GeneView**).

Have a look at the section 'Orthologue Prediction'. Follow the link to human SNX5, which takes you back to where you started.

(c) Examine the genomic context of the human and mouse genes.

From human SNX5 **GeneView**, follow the link 'Graphical view' to **ContigView.**

Q6: In which chromosomal region is the human gene located?

Customise the 'Detailed view' display of **ContigView**; select only 'Ensembl Genes' (from the 'Features' menu), 'Mouse BLASTz (net)' and 'Rat (BLASTz (net)' (from the 'Comparative' menu) and deselect all other options.
Have a look at the mouse and rat conserved regions in relation to the human Ensembl transcript. Note that there is correspondence with exons, but note also that this is not perfect. Zoom in to examine in more detail.

The conserved regions are probably showing “ungrouped” (a red ‘+’ shows to the left of the track label). Click on the red ‘+’ to the left of the ‘Mm blastz’ track: **ContigView** will reload, a red ‘-’ replaces the ‘+’ and the hits are now “grouped”. Note that clicking on the track produces a pop-up with details of and a link to that region in mouse. Click again on the red ‘-’ to the left of the ‘Mm blastz’ track to ungroup the conserved regions again.

Click on a mouse match in this track, and take the link ‘Dotter’ to **DotterView**. Note the dots on the diagonals where exons align. Zoom in to examine a smaller region.

Go back to human **ContigView**, click on a mouse match, and this time take the link ‘Jump to Mus musculus’.

This takes you to the corresponding display in mouse **ContigView**.

Zoom and/or customise the **ContigView** display to focus on the mouse Snx5 transcript, and turn on the human matches track if necessary. Compare the amount of sequence showing as matched (the same threshold Blast score is used).

(d) Examine the synteny blocks that these homologous genes are part of.

Take the ‘Graphical overview’ link to mouse **CytoView**. Zoom out several steps to see a large region and select all option from the ‘Comparative’ menu. Note the coloured blocks indicating regions where gene order is conserved in human and other species (‘synteny blocks’). Click on a synteny block and see the information and links.

Take the ‘View Syntenic Regions with Homo sapiens’ link to **SyntenyView**. Note that the large chromosome in the middle is the mouse chromosome – as you have come from a mouse page. The red box shows the region you have come from. The smaller chromosomes at the sides are the human chromosomes that have blocks where gene order is conserved with the mouse chromosome.

To the right is a list of genes found in this region, together with their homologues (putative orthologues). You can scroll along this list by using the ‘Upstream’ and ‘Downstream’ links at the bottom. The synteny information may increase your confidence that the two genes are real orthologues!

For more details on the functionality of **SyntenyView**, consult the associated Help page – click the blue [Help] button in the top right corner.

Additional exercises:

2. Compare the BRCA2 gene in human and mouse.
Find the human BRCA2 gene.
Identify its orthologue in mouse.
Display the human gene in **ContigView**, and examine the 'mouse matches' track for the region around the human gene.
Compare the two genes with respect to length and number of exons using **MultiContigView** and **AlignSliceView**. You can reach these pages from **ContigView** using the 'View alongside .' and 'View alignment with …' links, respectively.
View the regions of conserved synteny that include the genes using **SyntenyView**.
Have a look at the **GeneTreeView** page for human BRCA2.

3. Protein Family Exercise

Have a look at the protein family ENSF00000000779

Q7: How many genes in the family are there in human/mouse/rat, and what are their chromosomal locations?

View the Ensembl members of the family with **JalView** (click on the button from the 'FamilyView' page.
Scroll right to see the region with good alignments.

Try exporting the alignments in CLUSTALW format (use the 'Export alignments') link on the left of the FamilyView page.

4. Browse the human-mouse synteny blocks
Have a look at a number of human and mouse chromosomes in **SyntenyView** in order to get some idea as to the size, orientation and distribution of synteny blocks.
Hint: Entry points to SyntenyView include the MapView display of a chromosome and 'View Syntenic regions …' links from ContigView or CytoView pages.
Look at the synteny blocks in **CytoView** displays.
Export both the human and mouse sequence of a synteny block.
Hint: Display them in ContigView, take the 'Export sequence as FASTA' link.

5. Choose any gene of interest to you, and try to identify an orthologue in another species. Confirm whether they are part of a synteny block. If you have time, take the sequence of a transcript from one species (cut and paste from **TransView**) and try a BLAST search in the other species. How do the results compare?

6. Similarly you may have a region of interest. Check whether it is part of a synteny block and which genes it contains in human and mouse.

Answers (Comparative Genomics)

Main Exercise (1)

Q1: In **FamilyView** for Family ID ENSF00000002022 (Sorting Nexins) you may see there are 3 genes producing peptides in this family.

Q2: They are all 3 "Known" protein-encoding Ensembl genes. (Click on gene ID, name or genome location links to see this under "location of Ensembl genes…" OR click on the red arrows in the window showing the chromosomes).

Q3: Yes. There are 4 in *Mus musculus* (mouse), 6 in *Rattus norvegicus* (rat) and 2 in *Danio rerio* (zebrafish). (If you forget the species name, go to the homepage.)

Q4: There are mostly vertebrates in Ensembl. Of the invertebrate species (yeast, *C. elegans*, mosquito (2 species), fly and sea squirt (2 species)) there are homologues in:
mosquito (2), sea squirt (10), fly (1), *C.elegans* (2) as of the Ensembl 41 release.

Q5: 4 genes encode peptides that are part of the family. Genes can have more than one transcript (leading to 6 peptides in the sorting nexin family).

Q6: Chromosome 20, band p11.23

Additional Exercises (2)

Q7: *Hint: you can search Ensembl for the protein family ID, and go to FamilyView from species-specific geneview pages.*
Human: 3 genes
Mouse: 4 genes
Rat: 3 genes

EVALUATING GENES AND TRANSCRIPTS

(The 'GeneBuild')

Exercises examining supporting evidence for Ensembl genes.

Examine the evidence for the GPT gene.

1. Display the human GPT gene in GeneView.

Enter GPT into the text search box at the top of any human Ensembl page. Take the link to the **GeneView** page for the Ensembl Gene: ENSG00000167701.

Q1: What are the external database sources for the gene name and for the description?

Scroll down the **GeneView** page and have a look at the predicted exon structure. Note the 5' and 3'UTRs (untranslated regions).

2. Examine the supporting evidence for the GPT gene in ExonView.

Click on 'Exon information' in the left-hand menu to go to **ExonView**.
The bottom section of **ExonView** shows the supporting evidence that was used during the Ensembl transcript building process.

Q2: Which databases did the entries come from? Which support the UTRs?

Click on a supporting evidence entry to see the alignments of the Ensembl predicted transcript against the supporting evidence.

Click on the 'Gene information' link in the left-hand menu to return to the **GeneView** 'Gene Report' for GPT.

3. Examine homology evidence for the GPT gene in ContigView.

Click on the link ''Graphical View' in the left-hand menu. This takes you to **ContigView** displaying only the region encompassing the gene. Zoom out by clicking on the '-' button next to the Zoom triangle.

Look at the homology evidence tracks, e.g. Human proteins, Unigene, Human cDNAs. Note that the track labels are links to the help page. Clicking on the evidence brings up a pop-up menu with a link to the external database entry. Try some links.

Condense the Unigene and Protein track as well as the Overview section by clicking on the '-' boxes.

Under ‘Features’ select ‘EST(ex.)’ and ‘EMBL mRNAs’. Under ‘Decorations’ select ‘Show empty tracks’. This will help you remember which tracks you have selected. Examine the evidence for the GPT gene in the new evidence tracks.

4. Compare transcript predictions made by other methods.

In the ‘Features’ menu, turn off most of the evidence and make sure Vega Genes and Genscans are turned on.

Look at the Genscan track. Zoom out further.

Q3: How does the Genscan prediction differ from the Ensembl prediction? Note that this track shows *ab initio* Genscan predictions, not relying on supporting evidence.

Use the ‘DAS sources’ menu to turn on tracks showing transcript predictions from other groups (e.g. NCBI Gnomon). Compare and contrast!

5. Look at the ‘Similarity Matches’ section.

Go to **TransView** by clicking on the Ensembl transcript ALAT1_HUMAN and taking the direct ‘Transcr.’ link from the pop-up menu.

‘Known’ Ensembl transcripts like this one (shown in red in **ContigView**) have been successfully mapped to external database entries (note that this mapping is done *after* the genes have been built). These entries are given in the ‘Similarity Matches’ section of **TransView** (repeated in the ‘Transcript’ section of **GeneView**). Have a look at the types of databases linked out to.

Q4: What do the Target and Query % ids indicate? Check the online Help pages.

Answers (Evaluating Genes and Transcripts).

Q1: Name: HUGO Gene Nomenclature Committee (HGNC). Description: Uniprot/Swiss-Prot.

Q2: They are from the EMBL (BC018207.1), UniProt/Swiss-Prot (P24298) and RefSeq (NP_005300.1 and NM_005309.1) databases (click on the ID number to go to the original database entry. You may have to scroll down to find the correct entry.) The boxes represent the exons, the darker green they are, the better the supporting evidence is. UTR’s are found in exons 1, 2 and 12. Only cDNA sequences can support UTRs. One EMBL entry supports the 5’ (and 3’) UTR. Why are the boxes for exon 2 and 12 green for the peptide sequences NP_005300.1 and P24298? Because these exons also contains coding sequence (not only UTRs). (How do we know where the UTRs are? … check the above sequence, the UTR sequences are shown in purple!)

Q3: The Genscan transcript prediction mistakenly spans both the GPT and the PPP1R16A genes. It also shows an exon not present in the Ensembl prediction for GPT and doesn't predict UTRs.

Q4: In TransView, under Similarity Matches, Target %ID indicates the percentage of the Ensembl prediction matching the external sequence database and Query %ID is the percentage of the external database sequence matching the Ensembl prediction! Can you find this in the help pages?

VARIATIONS

Exercises exploring SNPs in Ensembl.

1. From the **ContigView** page displaying Human chr7: 116722634 - 116822633 select a non-synonymous coding SNP. (Hint… zoom down one step in the 'Detailed View' panel to view SNPs.)

2. In the article "Screening of the delta-F508 mutation and analysis of two single nucleotide polymorphisms of the CFTR gene in a sample of the general population of Valparaiso, Chile" by L.A. Vera et al." (Rev Med Chil 2005, 133:767-775.) the delta F508 mutation and the SNPs M470V and T854T are studied. Can you find this in-del and these two SNPs in Ensembl?

3. Retrieve all the validated SNPs associated with the human CFTR gene. How many of these SNPs are coding?

Answers (Variations)

Warning: if there has recently been a new Ensembl release based on an updated gene set and/or an updated SNP set, these answers may have changed - please check with one of the instructors.

1. To get to this page, click on the 'human' icon on the Ensembl homepage. Enter the chromosome number and region in base pairs under the karyotype. This should take you to **ContigView**. The 'Detailed View' panel is zoomed out too far to display SNPs, so zoom in one step using the zoom triangle. You may have to turn on the 'SNPs' track from the 'Features' drop-down menu on top of 'Detailed View'.

SNPs should now be shown as horizontal lines of different colours along the chromosome. The SNP legend is shown at the bottom of the panel (if not, turn on SNP legend the 'Decorations' roll-down menu.) There are not many non-synonymous coding SNPs (shown in yellow) in this region!. You can get information about a SNP by clicking on it (such as SNP ID, in this example a non-synonymous coding SNP in this region is: rs28513898), and you can also follow the link 'SNP properties' to its SNPView page.

2. Perform a text search for 'CFTR' in human. This should lead you to ENSG00000001626. Go to the **GeneView** page and then the **TransView** page by clicking on 'transcript information' on the left hand navigation column. By selecting the options 'Exons, Codons, Translations and SNPs' and 'Number residues: yes' you can display the SNPs in the transcript sequence. Alleles and alternative codons are shown by pointing your mouse over the nucleotide and amino acid residues, respectively.

Go back to the **GeneView** page, and continue on to the **GeneSNPView** page for this gene by clicking 'Gene variation info.' in the side menu.

Both the in-del and the two SNPs are displayed in the 'SNPs and variations' figure and the 'Variations and consequences' table. In the figure, M470V is shown as 'V/M' in yellow and delta F508 as 'F/-', also in yellow (as these are both non-synonymous coding SNPs). T854T is shown as 'T' in green, as it is a synonymous coding SNP. Note that you can use the 'SNP class' and 'SNP type' drop-down menus in the figure to configure both figure and table.

3. Go to www.biomart.org and click on the 'central server' link. Select Ensembl Gene 42 and Homo sapiens genes. In **Filters**, in the GENE section, enter either the HGNC symbol (CFTR) or Ensembl Gene ID (ENSG00000001626) under 'ID list limit'. Also, select 'Associated with validated SNPs' only under 'SNP'.
Go to '**Attributes**' and select the SNPs page. Under 'GENE ASSOCIATED SNPs' select the options 'Reference ID', 'Allele' and 'Location in Gene (coding etc)'. Click results and export as an excel file. This gene contains 935 validated SNPs of which 81 are coding.

The SNPs can also be viewed in the Ensembl browser. Starting from the GeneView page for CFTR, click on 'Gene variation info' on the left. Selecting only 'Non-synonymous' and 'Synonymous' SNPs under the 'SNP type' roll-down menu will show only the coding SNPs in the diagram and table below.