Access to genes and genomes with Ensembl

Introduction and Worked Example

January 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

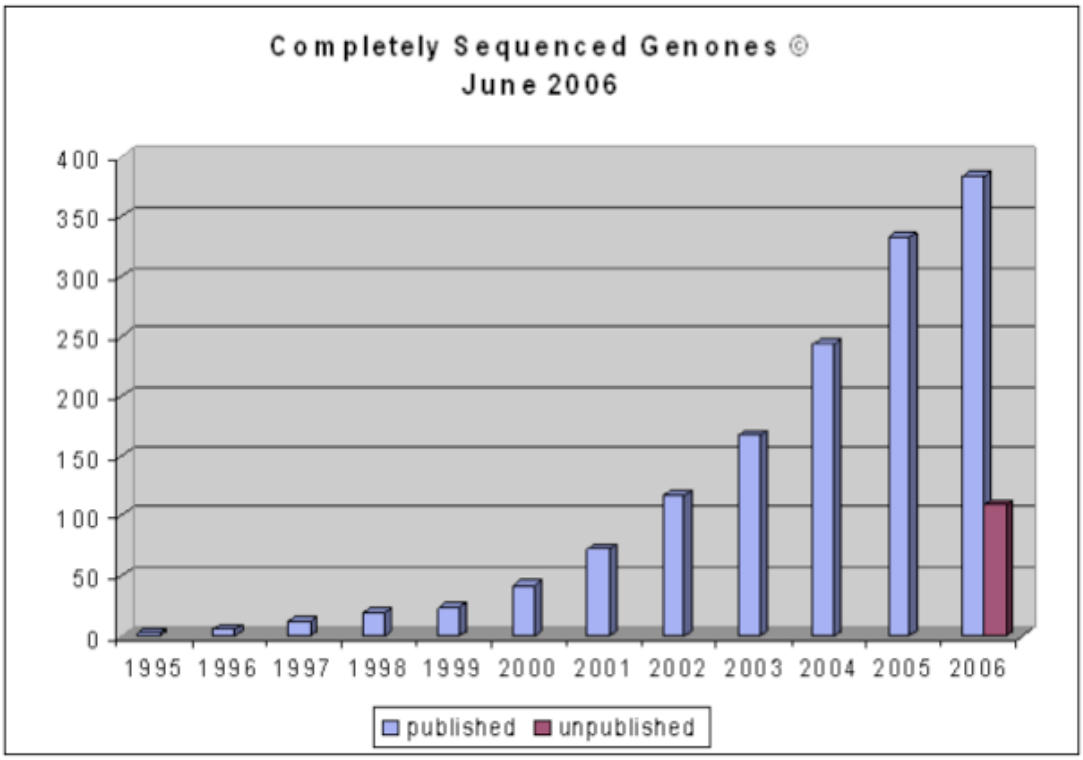


Figure 1. Completely sequenced genomes as of june 2006 (figure taken from http://www.genomesonline.org).

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

Ensembl Human

Ensembl release 42

Go

e.g. AL138722.15.1.44776, ENSG00000196160

HOME · BLAST · BIOMART · SITEMAP HELP

genome

Your Ensembl

- Show account · Log out
- Save bookmark

Healthchecks

- Health checks
- Old Health checks

Help & Documentation

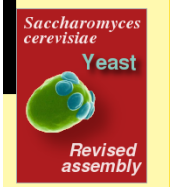
- Table of Contents
- Helpdesk
- About Ensembl
- Downloading data
- Displaying your own data
- Ensembl software

Select a species

- Mammals
- Other chordates
- Other eukaryotes

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page



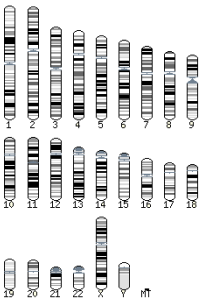
Search Ensembl *Homo sapiens*

Search: Go

e.g. **chromosome X** or **14:10000..200000** or **BRCA2**

Karyotype

Click on a chromosome for a closer view



Jump directly to sequence position

Chromosome: -- or region

From (bp):

To (bp): Go

Ensembl headlines: Release 42 (December 2006)

- **New - User accounts** (*all species*)
- **New species - Duck-billed Platypus** (*Ornithorhynchus anatinus*)
- **New Dog assembly and genebuild** (*Canis familiaris*)
- **New Chicken assembly and genebuild** (*Gallus gallus*)
- **New Human Ensembl-Vega** (*Homo sapiens*)

More news...

Go to your account to customise this news panel

About the Human genome

Assembly

This release is based on the NCBI 36 assembly of the human genome [November 2005]. The data consists of a reference assembly of the complete genome plus the Celera WGS and a number of alternative assemblies of individual haplotypic chromosomes or regions.

Full list of assemblies

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- Nature 431, 931 - 945 (21 October 2004)
- WT Sanger Institute Press Release

Annotation

The human genome sequence is now considered sufficiently stable that the three major genome browsers have come together to produce a common set of gene IDs for their annotations. This Consensus CDS ID set has been encorporated into the Ensembl database alongside the existing identifiers.

- More information about the CCDS project.

The ENCODE (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome.

- More information about the ENCODE resources at E

Statistics

Assembly:	NCBI 36, Oct 2005
Genebuild:	Ensembl, Aug 2006
Database version:	42.36d
Known genes:	21,774
Novel genes:	1,036
Pseudogenes:	1,069
RNA genes:	3,976
Genscan gene predictions:	69,185
Gene exons:	270,661
Gene transcripts:	44,676
Base Pairs*:	3,253,037,807
Golden Path Length:**	3,093,120,360
Most common InterPro domains:	Top 40 Top 500

* Total number of base pairs = sum of lengths of DNA table

** Reference assembly (Golden path) length = sum of non-redundant top level seq regions

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.

A ‘Vega’ gene
(a consortium external to Ensembl)

STEP 4:
Click on ‘ENSG00000130427’

Ensembl release 42 - Dec 2006

HOME · BLAST · BIOMART · SITEMAP HELP

e.g. AL138722.15.1.44776, ENSG00000196160

Your Ensembl

- Show account · Log out
- Save bookmark
- Save configuration as...

Feature type

- Gene (4)
  - Homo sapiens (4)

Species

- Homo sapiens (4)
  - Gene (4)

Exalead Help

To exclude a category click on the "−":

To restrict to a category click on the name of the category.

To reset a category click on the "R" or it's name.

Healthchecks

- Health checks
- Old Health checks

Ensembl Archive

- View previous release of page in Archive!
- Stable Archive! link for this page

Ensembl text search

EPO Gene [Search]

Your query matched 4 entries in the search database

Vega protein_coding Gene: OTTHUMG00000023044 (HGNC Symbol: EPO) [ContigView]
Vega protein_coding gene: OTTHUMG00000023044 has 1 transcript: OTTHUMT00000059365 and associated peptide: OTTHUMP00000024662
erythropoietin
The gene has the following external identifiers mapped to it:
CCDS: CCDS5705, CCDS5705.1
Entrez Gene: 2056, EPO
HGNC Symbol: EPO, 3415
MIM: 133170
RefSeq DNA: NM_000799
UniProtKB/Swiss-Prot: P01588
Vega gene: EPO, OTTHUMG00000023044
Vega transcript: EPO-001, OTTHUMT00000059365
Vega translation: OTTHUMP00000024662

Source: el41; **Feature type**: Gene; Homo sapiens; **Species**: Homo sapiens; G

Ensembl protein_coding Gene: ENSG00000130427 (HGNC Symbol: EPO) [ContigView]
Ensembl protein_coding gene: ENSG00000130427 has 1 transcript: ENST00000252723 and associated peptide: ENSP00000252723
Erythropoietin precursor (Epoietin). [Source:Uniprot/SWISSPROT;Acc:P01588]
The gene has the following external identifiers mapped to it:
Affymx Microarray Focus: 207257_at
Affymx Microarray HCG110: 1023_at
Affymx Microarray HuGeneFL: X02158_rna1_at
Affymx Microarray U133: 207257_at, 217254_s_at
Affymx Microarray U95: 1023_at
Agilent CGH: A_14_P113914
Agilent Probe: A_23_P145669, A_23_P145664
CCDS: CCDS5705, CCDS5705.1
EMBL: AF053356, S65458, BC093628, AF202307, AF202313, X02157, AF202306, AF202314, AF202308, X02158, M11319, AF202312, AF202309, AF202310, AC009488, BC111937, AF202311
Entrez Gene: 2056
GE Healthcare/Amersham Codelink WGA: GE79554
GO: GO:0005615, GO:0007165, GO:0043249, GO:0005128, GO:0007267, GO:0006950, GO:0005179, GO:0001666, GO:0008015, GO:0007275, GO:0005576
HGNC Symbol: EPO, 3415
Illumina: GI_4503598
IPI: IPI00307226.3, IPI00307226
MIM gene: 133170
PDB: 1CN4, 1EER, 1BUY
Protein ID: AAC78791.1, CAA26095, CAA26094.1, AAA52400.1, AAF23132.1, AAF23133.1, AAF17572.1, AAP22357.1, AAI11938, AAF17572, AAI11938.1, AAC78791, AAF23132, AAD13964, CAA26094, AAP22357, AAA52400, AAF23134.1, AAF23134, CAA26095.1, AAH93628.1, AAD13964.1, AAF23133, AAH93628
RefSeq DNA: NM_000799.2, NM_000799
RefSeq peptide: NP_000790.2, NP_000790
UniGene: Hs.2303
UniProtKB/Swiss-Prot: EPO_HUMAN, Q549U2, Q9UDZ0, Q9UHA0, P01588, Q9UEZ5
UniProtKB/TrEMBL: Q2M2L6_HUMAN, Q2M2L6

Source: el41; **Feature type**: Gene; Homo sapiens; **Species**: Homo sapiens; Gene;

Ensembl protein_coding Gene: ENSG00000121053 (HGNC Symbol: EPX) [ContigView]
Ensembl protein_coding gene: ENSG00000121053 has 1 transcript: ENST00000225371 and associated peptide: ENSP00000225371
Eosinophil peroxidase precursor (EC 1.11.1.7) (EPO) [Contains: Eosinophil peroxidase light chain; Eosinophil peroxidase heavy chain]. [Source:Uniprot/SWISSPROT;Acc:P11678]
The gene has the following external identifiers mapped to it:
Affymx Microarray Focus: 214627_at
Affymx Microarray HuGeneFL: X14346_at
Affymx Microarray U133: Hs.46295.0.S2_3p_at, 214627_at
Affymx Microarray U95: 34587_at
Agilent CGH: A_14_P104496
Agilent Probe: A_23_P89192
CCDS: CCDS11602, CCDS11602.1
EMBL: M29911, M29908, X14346, M29907, M29910, M29909, M29906, M29912, M29913, M29905, M29904, DQ054598
Entrez Gene: 8288
GE Healthcare/Amersham Codelink WGA: GE523720
GO: GO:0005615, GO:0005509, GO:0005506, GO:0004601, GO:0042744, GO:0016491
HGNC Symbol: 3423, EPX
Illumina: GI_4503594
IPI: IPI00006690.1, IPI00006690
MIM disease: 261500
MIM gene: 131399
Protein ID: AAY43126, AAA58458, AAY43126.1, CAA32530.1, CAA32530, AAA58458.1
RefSeq DNA: NM_000502, NM_000502.2
RefSeq peptide: NP_000493.1, NP_000493
UniGene: Hs.279259
UniProtKB/Swiss-Prot: Q4TVP3, P11678, PERE_HUMAN

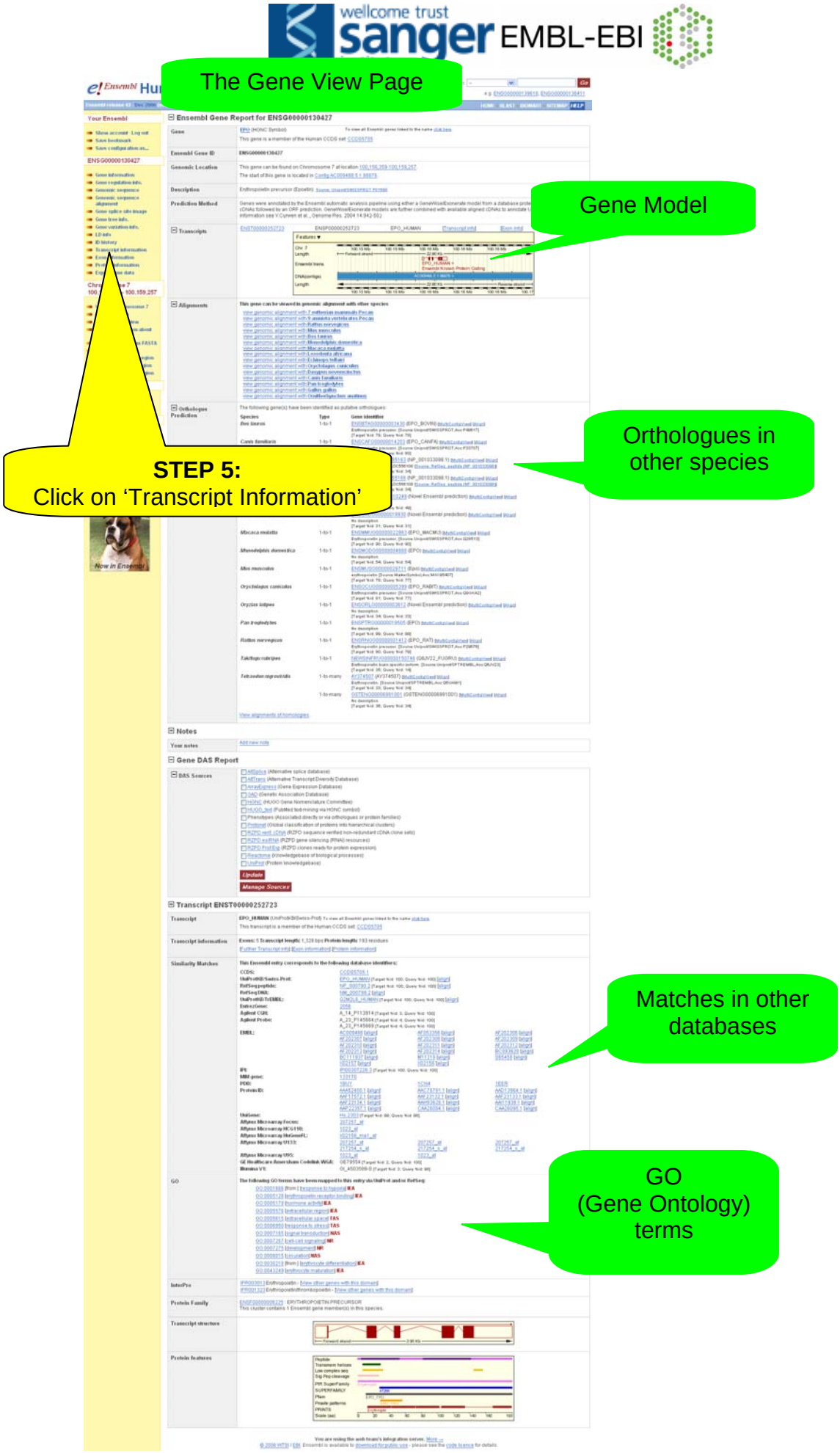
Source: el41; **Feature type**: Gene; Homo sapiens; **Species**: Homo sapiens; Gene;

Ensembl protein_coding Gene: ENSG00000187266 (HGNC Symbol: EPOR) [ContigView]
Ensembl protein_coding gene: ENSG00000187266 has 1 transcript: ENST00000222139 and associated peptide: ENSP00000222139
Erythropoietin receptor precursor (EPO-R). [Source:Uniprot/SWISSPROT;Acc:P19235]
The gene has the following external identifiers mapped to it:
Affymx Microarray Focus: 396_f_at, 37986_at
Affymx Microarray HCG110: 1087_at, 396_f_at
Affymx Microarray HuGeneFL: M60459_at
Affymx Microarray U133: 209962_at, 4876922C_3p_at, 209963_s_at, 396_f_at, 37986_at, 215054_at
Affymx Microarray U95: 1087_at, 396_f_at, 37986_at
Agilent CGH: A_14_P129163
Agilent Probe: A_23_P381954, A_23_P367899, A_23_P101494
CCDS: CCDS12260.1, CCDS12260
EMBL: X57282, BC112153, M77244, M34986, M60459, S45332, M76595
Entrez Gene: 2057
GE Healthcare/Amersham Codelink WGA: GE58209
GO: GO:0016020, GO:0005515, GO:0007165, GO:0005887, GO:0004900, GO:0016021
HGNC Symbol: 3416, EPOR
Illumina: GI_4557561
IPI: IPI00401741, IPI00017476.1, IPI00401741.1, IPI00017476
MIM disease: 133100
MIM gene: 133171
PDB: 1EBA, 1EBP, 1CN4, 1ERN, 1EER
Protein ID: AAB23271.1, AAA52401.1, AAA52393, AAA52401, AAA52392, CAA40550, AAI12154.1, AAA52403, AAA52393.1, AAA52392.1, AAB23271, AAA52403.1, CAA40550.1, AAI12154
RefSeq DNA: NM_000121.2, NM_000121
RefSeq peptide: NP_000112.1, NP_000112
UniprotKB/SpliceVariant: P19235-3
UniProtKB/Swiss-Prot: P19235, Q15443, EPOR_HUMAN
UniProtKB/TrEMBL: Q2M205, Q2M205_HUMAN

Source: el41; **Feature type**: Gene; Homo sapiens; **Species**: Homo sapiens; Gene;

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.


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

wellcome trust sanger institute EMBL-EBI

e! 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 data

Chromosome 7
100,156,359 - 100,15

- View of Chr
- Graphical v
- Graphical o
- Export info region
- Export seq
- Export EMBL
- 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 **Swiss-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	CCCGGAGCCGGACCGGGGCCACCGCGCCCGCTCTGCTCCGACACCGCGCCCCCTGGACAG CCGCCCTCTCCTCCAGGCCCGTGGGGCTGGCCTGCACCGCCGAGCTTCCCGGGATGAGG GCCCCCGGTGTGGTCACCGGCGCGCCCCAGGTCGCTGAGGGACCCGCCAGGCGCGGA GATGGGGGTGCACGAATGTCCTGCCTGGCTGTGGCTTCTCCTGTCCCTGCTGTCGCTCCC TCTGGGCCTCCCAGTCCTGGGCGCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGA GAGGTACCTCTTGGAGGCCAAGGAGGCCGAGAATATCACGACGGGCTGTGCTGAACACTG CAGCTTGAATGAGAATATCACTGTCCCAGACACCAAAGTTAATTTCTATGCCTGGAAGAG GATGGAGGTCGGGCAGCAGGCCGTAGAAGTCTGGCAGGGCCTGGCCCTGCTGTCGGAAGC TGTCCTGCGGGGCCAGGCCCTGTTGGTCAACTCTTCCCAGCCGTGGGAGCCCCTGCAGCT GCATGTGGATAAAGCCGTCAGTGGCTTCGCAGCCTCACCACTCTGCTTCGGGCTCTGGG AGCCCAGAAGGGAGCCATCTCCCCTCCAATGCGGCTCAGCTGCTCCACTCCGAACAAT CACTGCTGACACTTTCCGCAAACTCTTCCGAGTCTACTCCAATTTCCTCCGGGGAAAGCT GAAGCTGTACACAGGGGAGGCCTGCAGGACAGGGGACAGATGACCAGGTGTGTCCACCTG GGCATATCCACCACCTCCCTCACCAACATTGCTTGTGCCACACCCTCCCCCGCCACTCCT GAACCCCGTCGAGGGGCTCTCAGCTCAGCGCCAGCCTGTCCCATGGACATCCAGTGCCA GCAATGACATCTCAGGGGCCAGAGGAACTGTCCAGAGAGCAACTCTGAGATCTAAGGATG TCACAGGGCCAACTTGAGGGCCCAGAGCAGGAAGCATTCAGAGAGCAGCTTTAAACTCAG GGACAGAGCCATGCTGGGAAGACGCCTGAGCTCACTCGGCACCCTGCAAAATTTGATGCC AGGACACGCTTTGGAGGCGATTTACCTGTTTCGCACCTACCATCAGGGACAGGATGACC TGGATAACTTAGGTGGCAAGCTGTGACTTCTCCAGGTCTCACGGGCATGGGCACTCCCTT GGTGGCAAGAGCCCCCTTGACACCGGGGTGGTGGGAACCATGAAGACAGGATGGGGGCTG GCCTCTGGCTCTCATGGGGTCCAAGTTTTGTGTATTCTTCAACCTCATTGACAAGAACTG AAACCACC **Show the following features:** Exons **Number residues:** No Refresh	

You are using the web team's integration server. More →

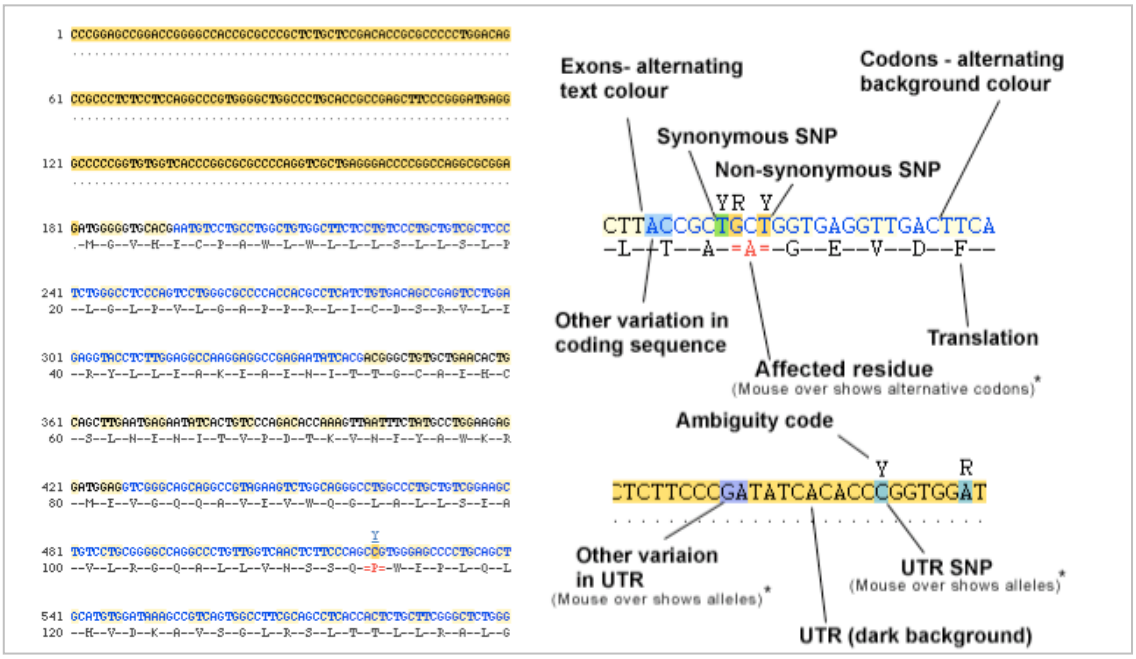
© 2006 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

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 gctactttgcgggaactcagcaacccaggcatctctgagtctccgcccaagac ccccagggggagtgtccgggagcccagcctttcccagatagcacgctccgccc agggtgcgcaaccggctgcactcccctcccgcgacccagggcccgggagcagc acccacacgcacgtctgcagcagccccgctcacgccccggcgagcctcaac ctgcccctgctctgaccccgggtggcccctaccccctggcgaccccctcacgcacacagcct ctcccccaccccccacccgcgcacgcacacatgcagatAacagcccccgaccccccggccaga gccgcagagtccctgggccaccccggccgctcgctgcgctgcgccgcaccgcgctgtcct
1	ENSE00001130431	7	1	100,156,359	100,156,552	-	1	194	CCCGGAGCCGGACCGGGGCCACCGCGCCCCGCTCTGCTCCGACACCGCGCCCCCTGGACAG CCGCCCTCTCCTCCAGGCCCGTGGGGCTGGCCCTGCACCGCCGAGCTTCCCGGGATGAGG GCCCCCGGTGTGGTCACCCGGCGCGCCCCAGGTCGCTGAGGGACCCCGGCCAGGCGCGGA GATGGGGGTGCACG
	Intron 1-2	7	1	100,156,553	100,157,1			564	gtgagtactcgcgggctgggcgctcccgcccgcccgggtccctgtttgagcggggattta gcgccccggctattggccaggaggtggctgggttcaaggaccggcgacttgtcaaggacc ccggaagggggagggggtgggggcagcctccacgtgccagcggggactttggggagtcct tggggatggcaaaaacctgacctgtgaaggggacacagtttggggttgaggggaagaag gtttgggggttctgctgtgccagtggagaggaagctgataagctgataacctgggcgctg gagccaccacttatctgccagaggggaagcctctgtcacaccaggattgaagtttggccg gagaagtggatgctggtagctgggggtggggtgtgcacacggcagcaggattgaatgaag gccagggaggcagcacctgagtgcttgcatggttggggacaggaaggacgagctggggca gagacgtggggatgaaggaagctgtccttccacagccaccccttctcccctcccgcctgac tctcagcctggctatctgttctag
2	ENSE00001144077	7	1	100,157,117	100,157,262	1	0	146	AATGTCCTGCCTGGCTGTGGCTTCTCCTGTCCCTGCTGTCGCTCCCTCTGGGCCTCCCAG TCCTGGGCGCCCCCACCACGCCTCATCTGTGACAGCCGAGTCCTGGAGAGGTACCTCTTGG AGGCCAAGGAGGCCGAGAATATCACG
	Intron 2-3	7	1	100,157,263	100,157,520			258	gtgagacccctccccagcacattccacagaactcacgctcagggcttcagggaactcct cccagatccaggaacctggcacttggtttggggtggagttgggaagctagacactgcccc cctacataagaataagtctggtggccccaaaccatacctggaaactaggcaaggagcaaa gccagcagatcctacggcctgtgggccagggccagagccttcagggacccttgactcccc gggctgtgtgcatttcag
3	ENSE00001130423	7					0	87	ACGGGCTGTGCTGAACACTGCAGCTTGAATGAGAATATCACTGTCCCAGACACCAAAGTT AATTTCTATGCCTGGAAGAGGATGGAG
	Intron 3-4	7						615	gtgagttcctttttttttttttccttctttggagaatctcatttgcgagcctgatt ttggatgaaagggagaatgatcgagggaaaggtaaaatggagcagcagagatgaggctgc ctgggcgcagaggctcacgtctataatcccaggctgagatggccgagatgggagaattgc ttgagccctggagtttcagaccaacctaggcagcatagtgagatcccccatctctacaaa catttaaaaaaattagtcaggtgaggtggtgcatggtggtagtcccagatatttggaagg ctgaggcgggaggatcgctttgagcccaggaatttgaggctgcagtgagctgtgatcacac cactgcactccagcctcagtgacagagtgaggccctgtctcaaaaaagaaaagaaaaag aaaaataatgagggctgtatggaatacattcattattcattcactcactcactcactcac tcattcattcattcattcattcaacaagtcttattgcatacttctgtttgctcagcttg 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								aatatgactcttggcttttctgttttctgggaacctccaaatcccctggctctgtcccac tcctggcagcagtgcagcaggtccaggtccgggaaacgaggggtggagggggctgggccc tacgtgctgtctcacacagcctgtctgacctctcgaccctaccgggcctgaggccacaag ctctgcctacgctggtcaataaggtggctccattcaaggcctcaccgcagtaaggcagct gccaaccctgcccagggcaaggctgcagtgcgctgagattgtcatcaaggagagggaggc cagaggacgggtccttttgggagttttgggggctggtaacagctgccaaccctggtcact ggccctgttaatttctgcctcttctttggtctcttcggctgatgccacactcccagcat cagccttccgcctcaatgcacatcctcgctaagtccagctgcctcctttactcacctggg gactgtcaccatgaatactc.....

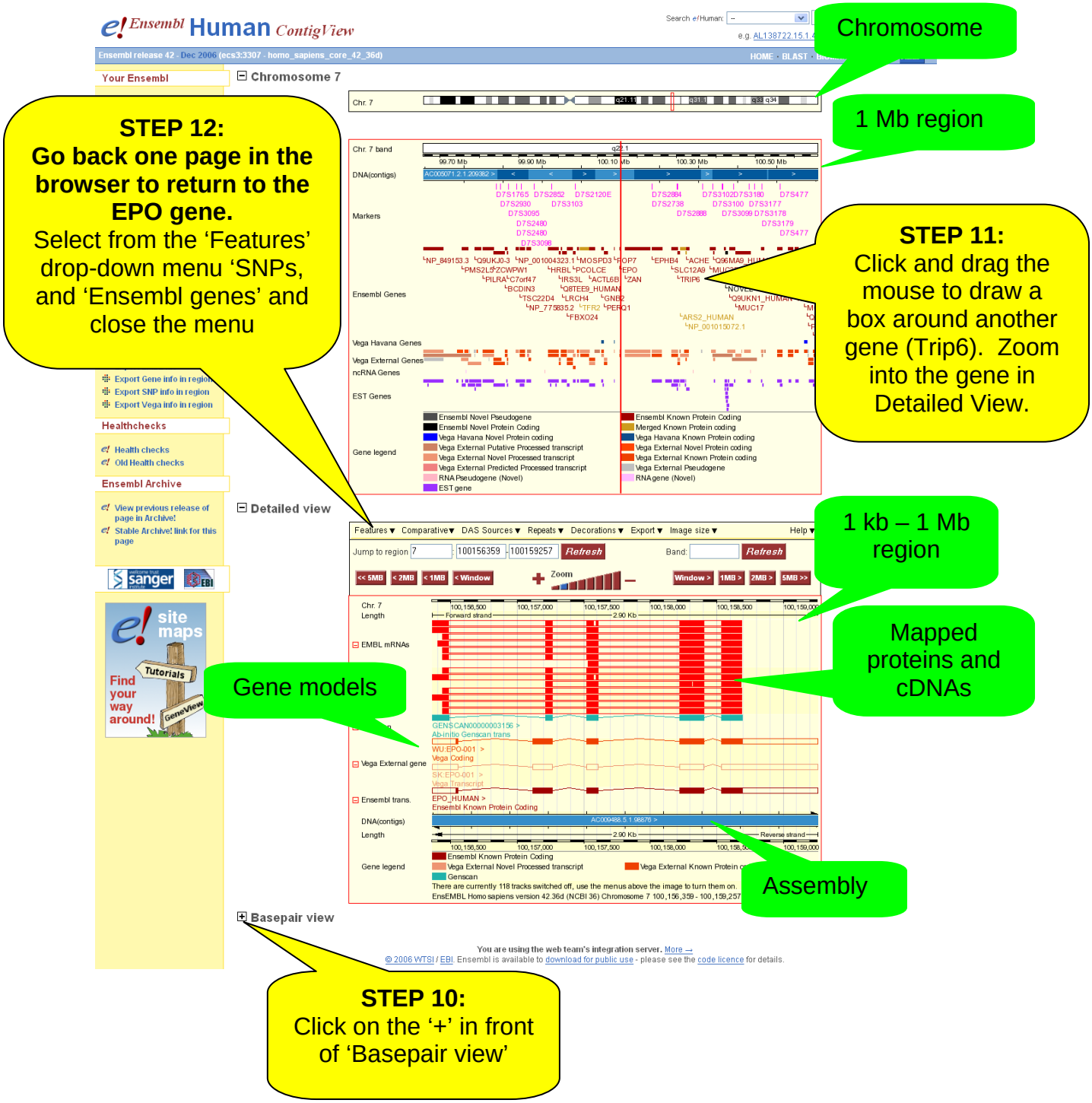
Flank (green)

Intron (blue)

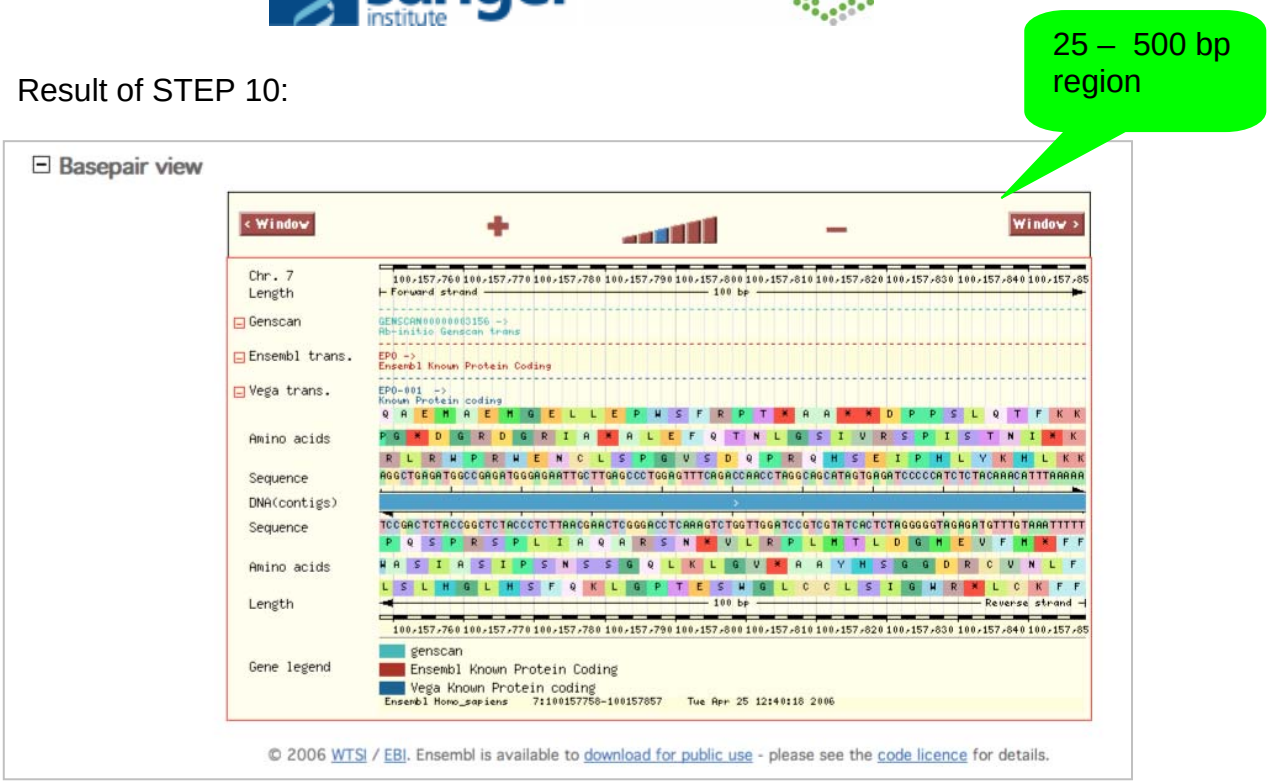
Coding sequence (black)

UTR (purple)

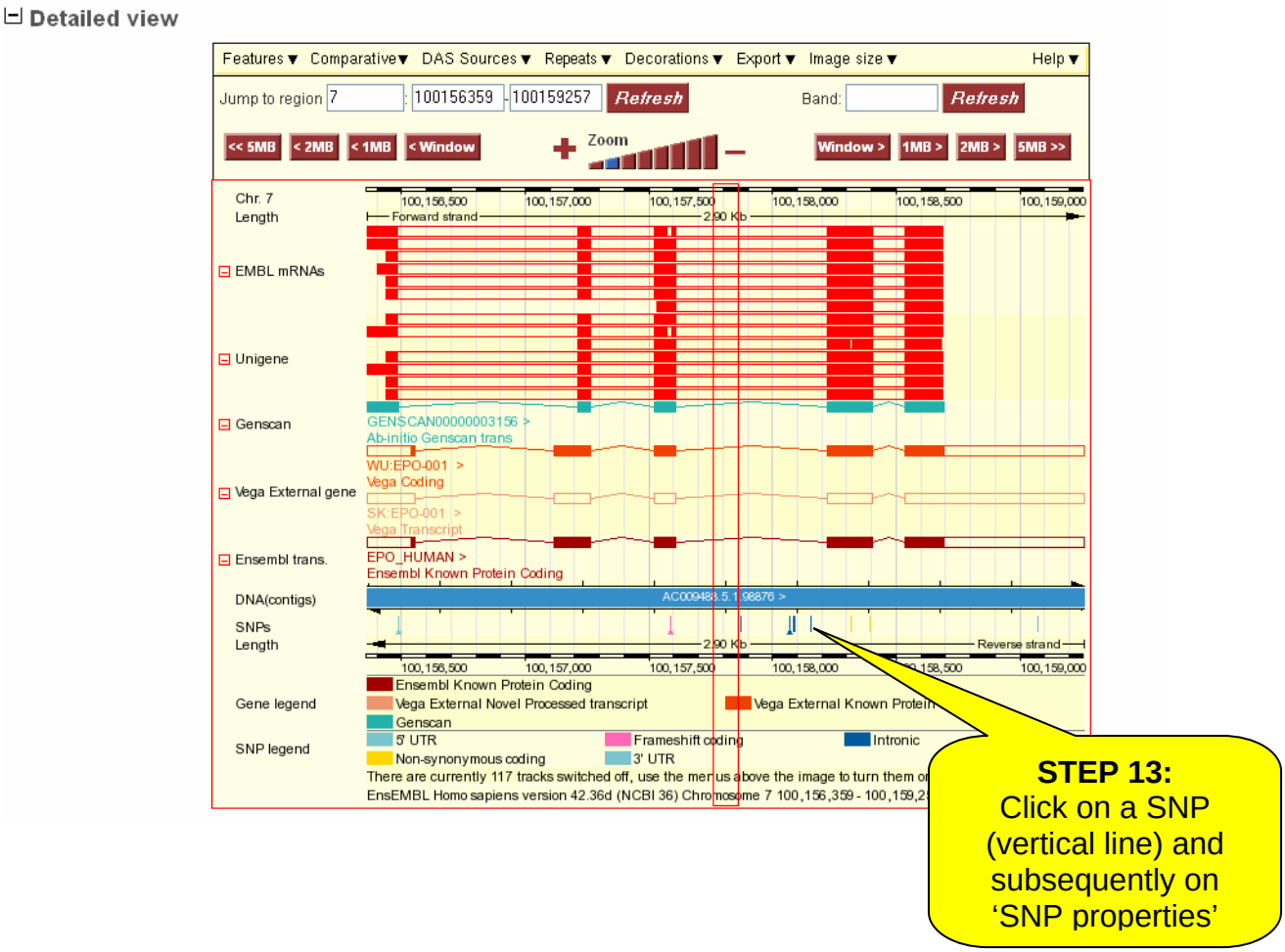
Result of STEP 9:

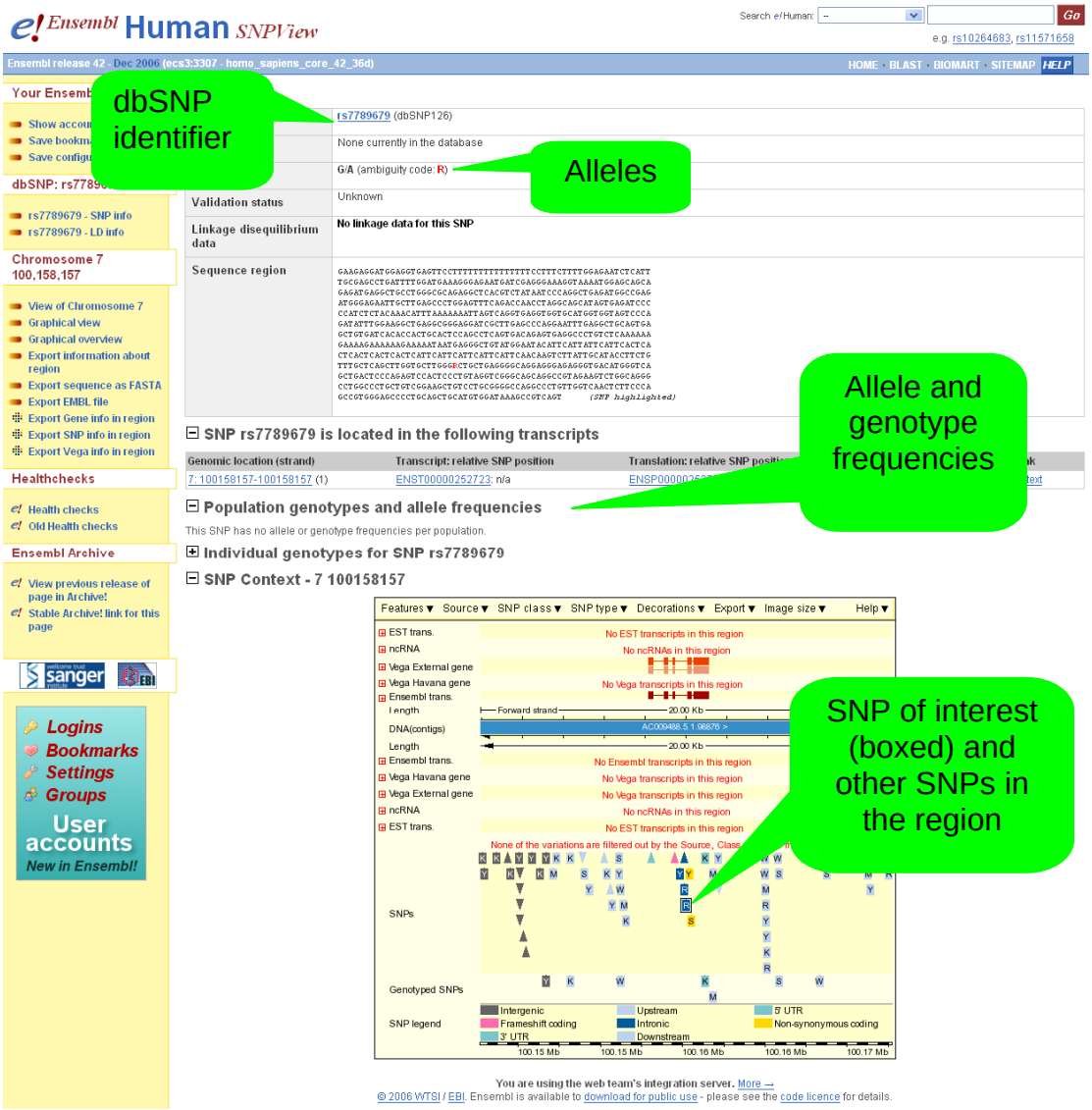


Result of STEP 10:



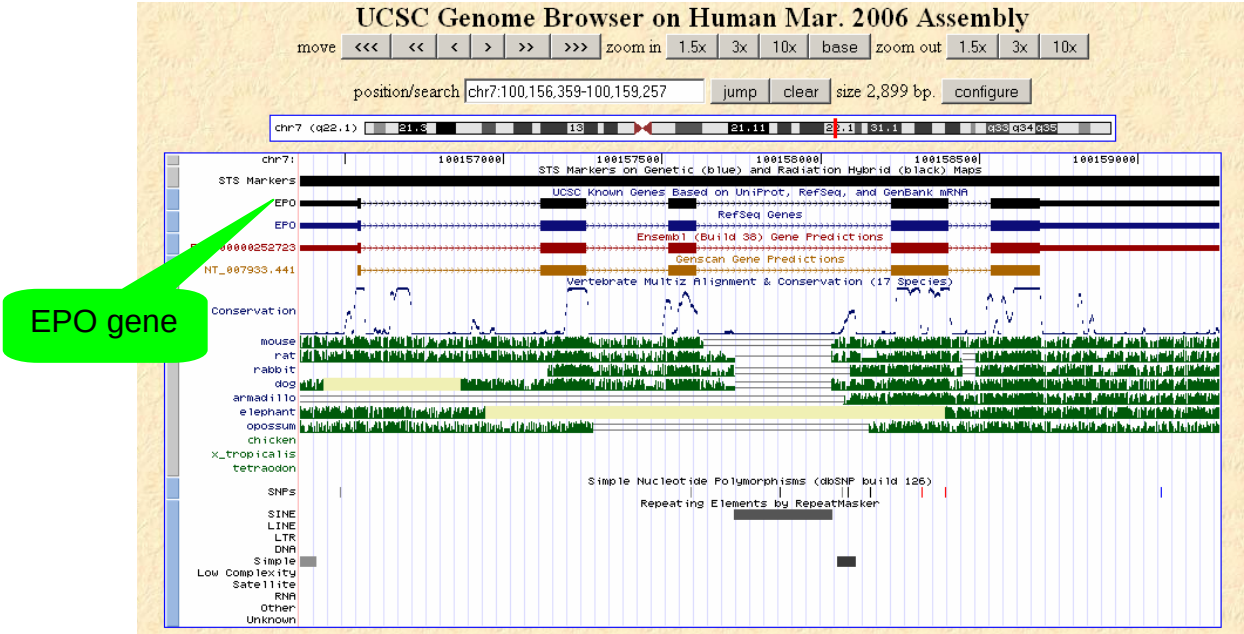
Result of STEP 12:





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.

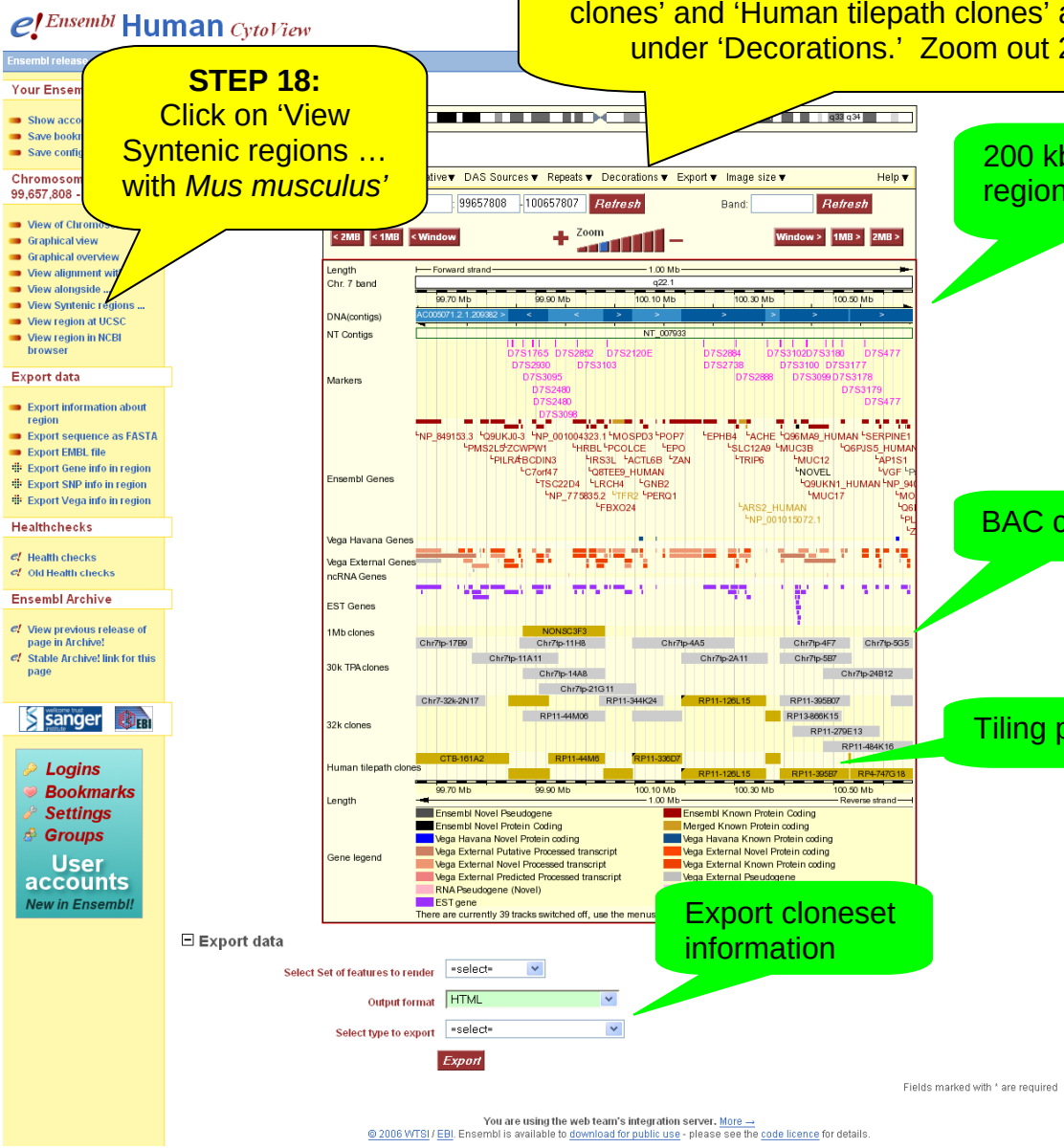
STEP 18:
Click on 'View Syntenic regions … with *Mus musculus*'

200 kb – 50 Mb region

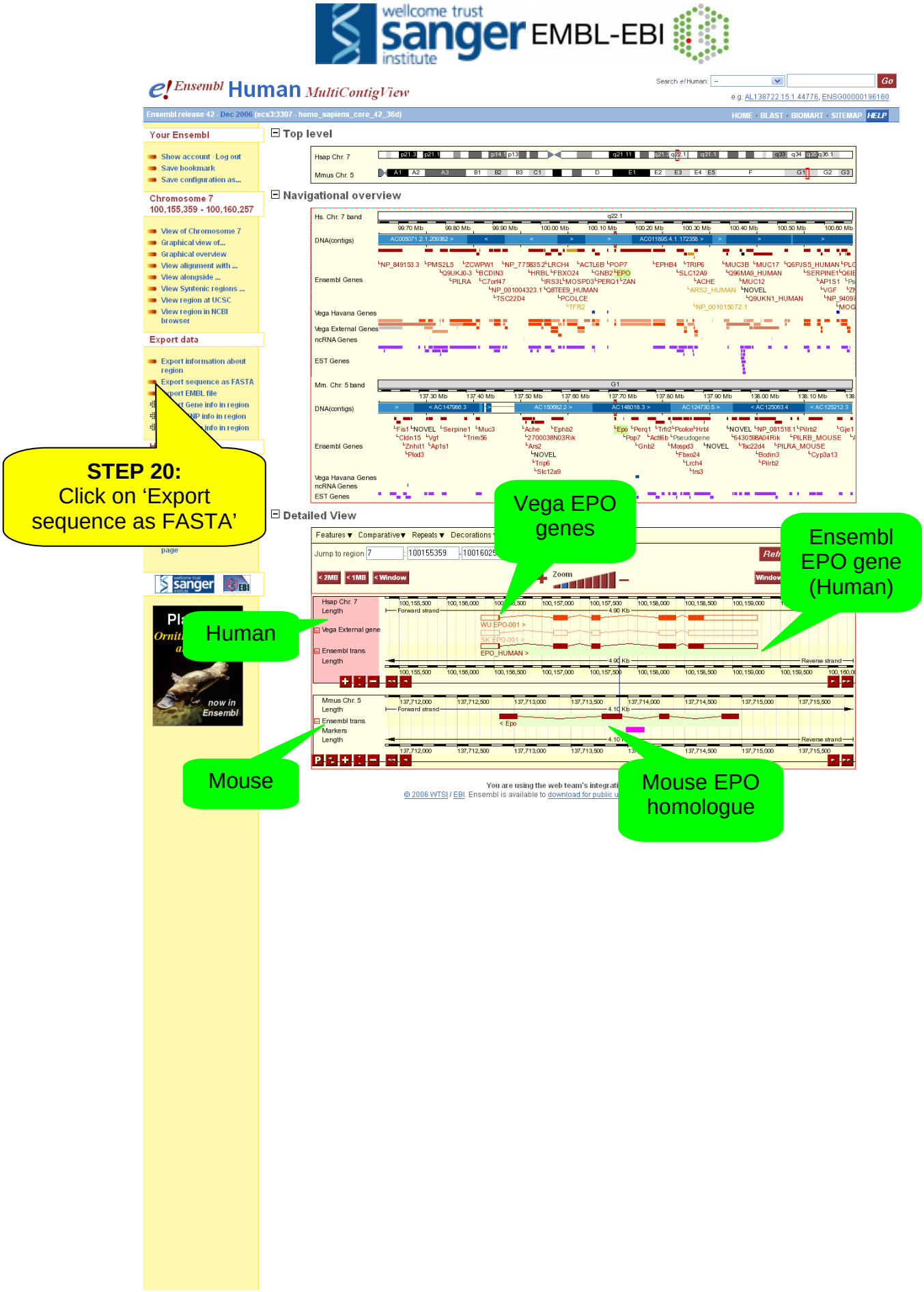
BAC clones

Tiling path clones

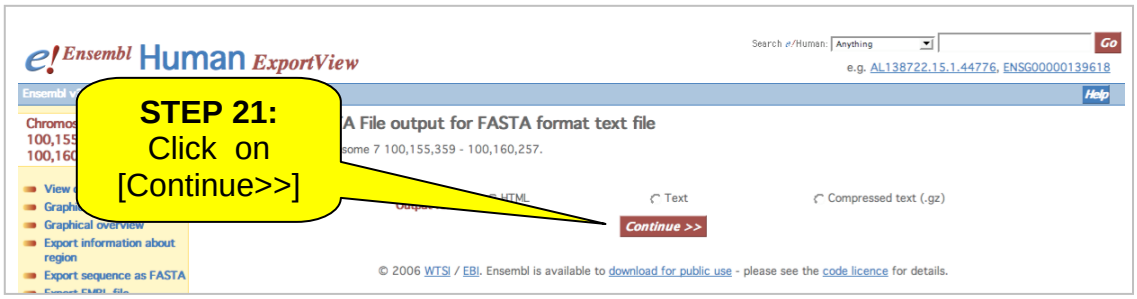
Export cloneset information



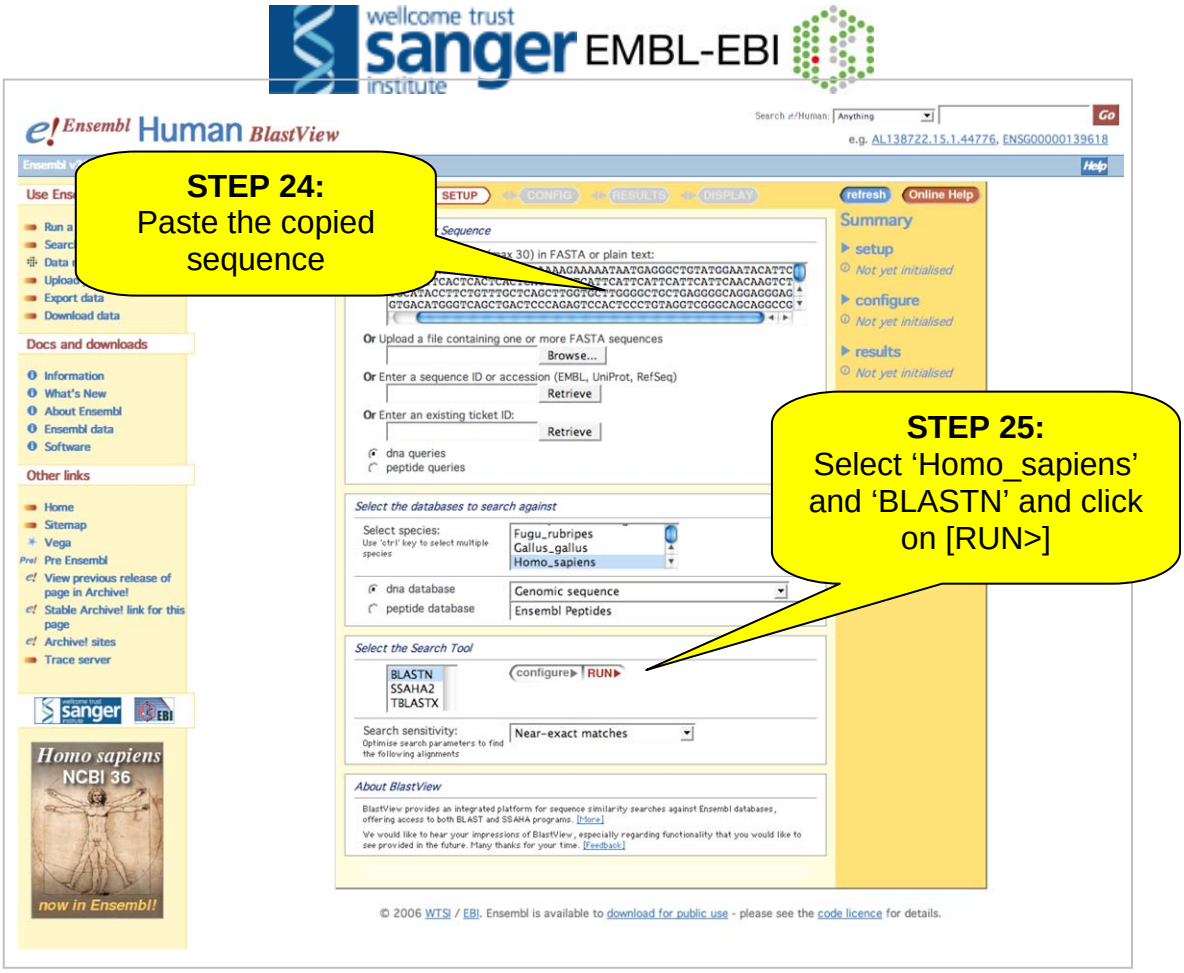

Human genes
Mouse chromosomes
Mouse homologues
Syntenic block
Human chromosome
STEP 19:
Click on
[MultiContigView]
Homo sapiens chromosome 7
Homology Matches
Homo sapiens Genes
Mus musculus Homologues
Navigate Homology
Change Chromosome
Your Ensembl
Chromosome 7
Healthchecks
Ensembl Archive
Guinea Pig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

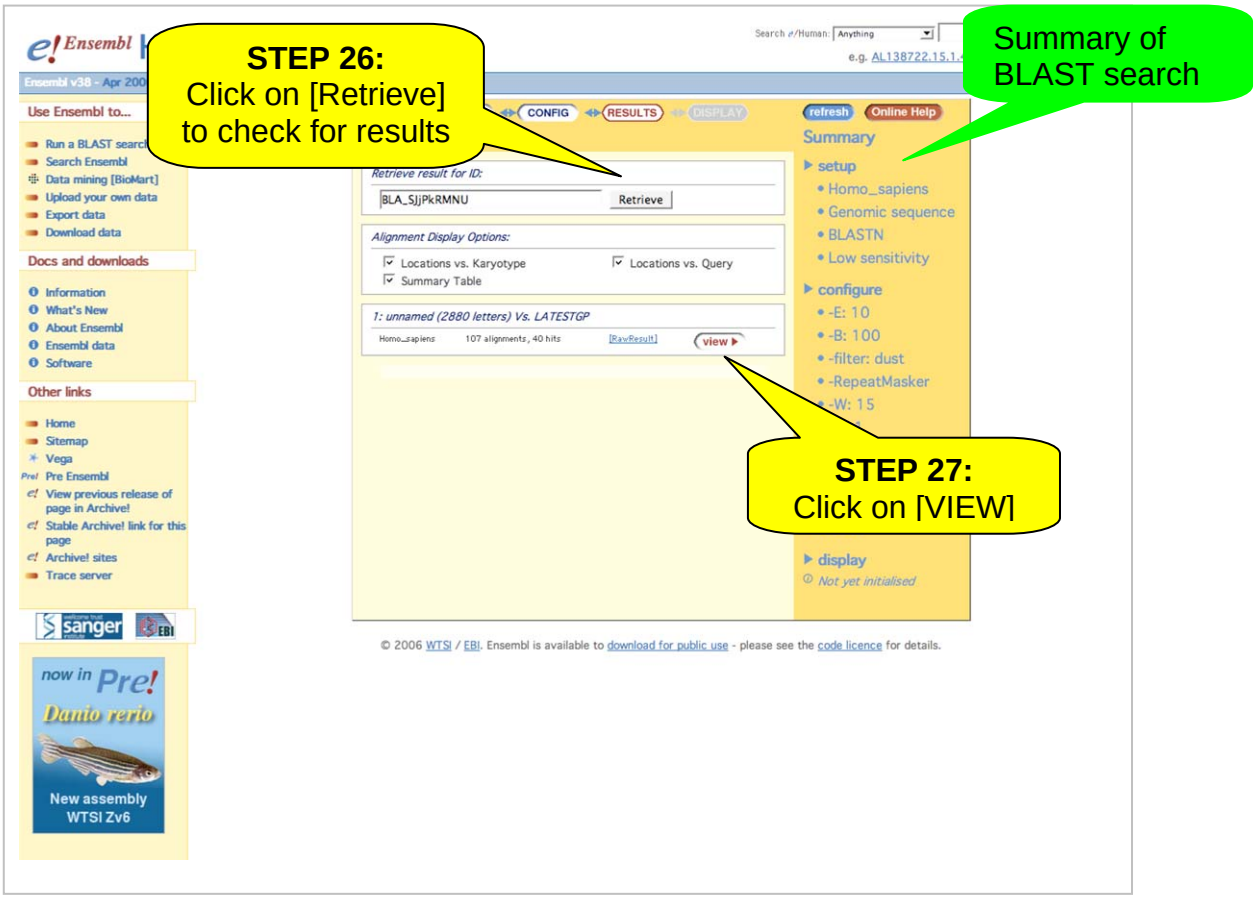


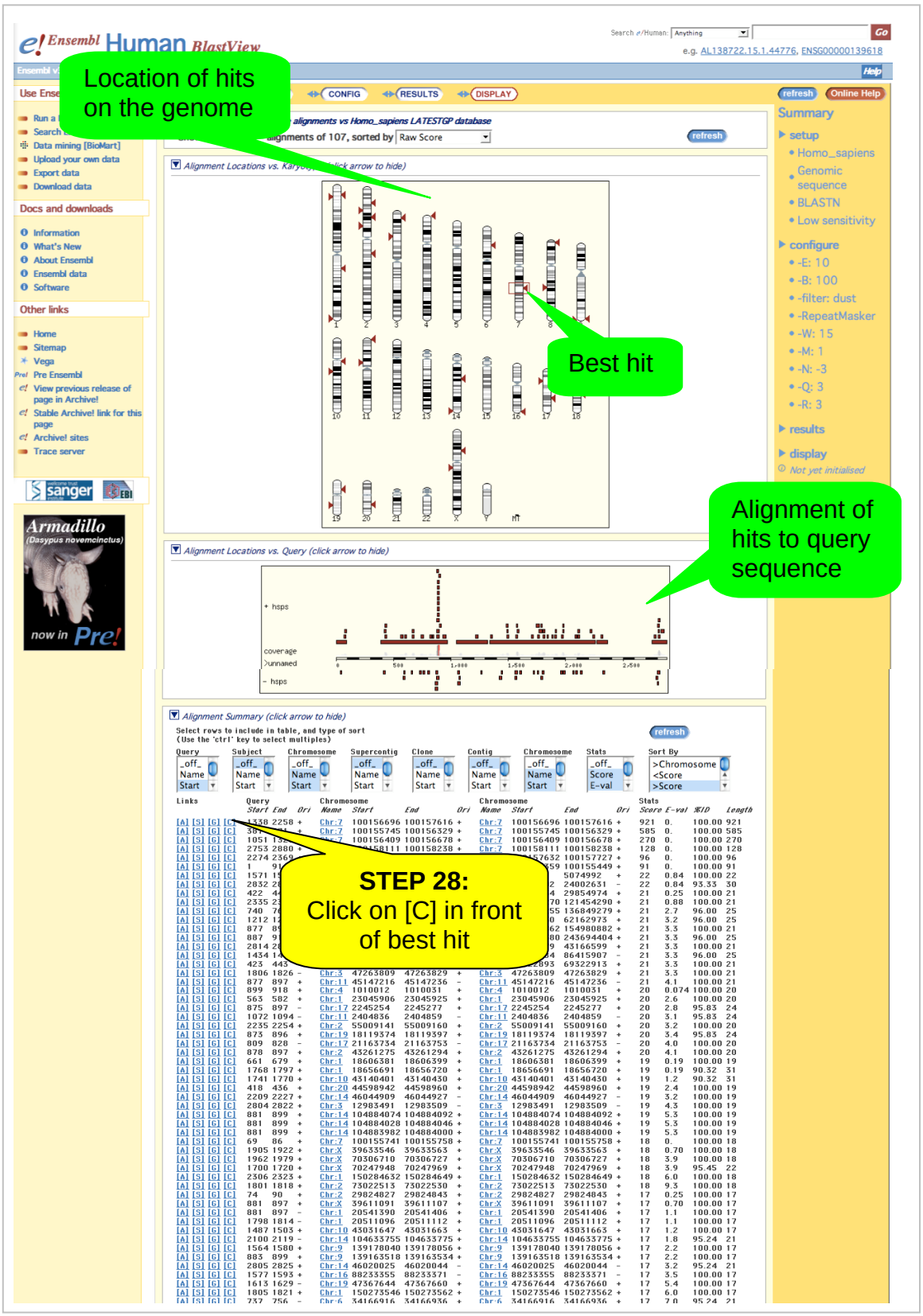
STEP 20:
Click on 'Export sequence as FASTA'


Ensembl Human ExportView
Search e/Human: Anything
Go
e.g. AL138722.15.1.44776, ENSG00000139618
Help
A File output for FASTA format text file
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>>]

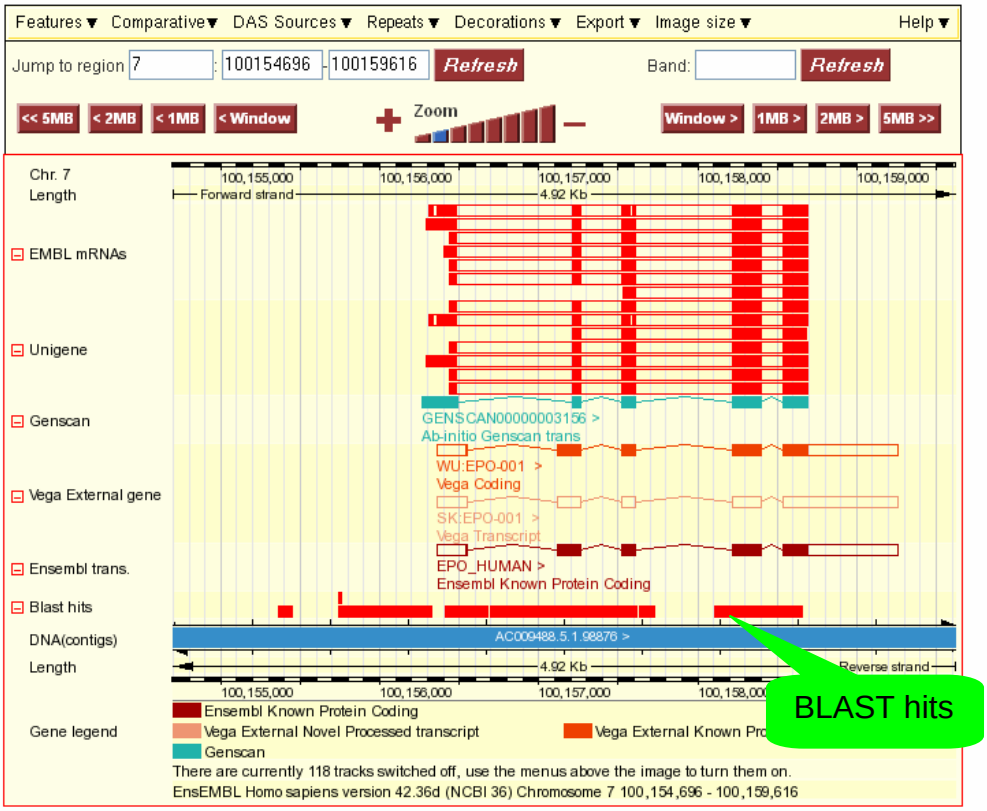

Ensembl Human ExportView
Search e/Human:
Go
e.g. AL138722.15.1.44776, ENSG00000196160
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Show account · Log
Save bookmark
Save configuration as...
Chromosome 7
100,155,359 - 100,160,257
View of Chromosome 7
Graphical view
Graphical overview
View alignment 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
Select the databases to search against
Select species:
Fugu_rubripes
Gallus_gallus
Homo_sapiens
Select the Search Tool
BLASTN
SSAHA2
TBLASTX
configure
RUN
Search sensitivity:
Near-exact matches
Summary
setup
configure
results


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
view
Summary
setup
Homo_sapiens
Genomic sequence
BLASTN
Low sensitivity
configure
-E: 10
-B: 100
-filter: dust
-RepeatMasker
-W: 15
display


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…



END of the Worked Example

EXERCISES and ANSWERS

(NOTE: please use release 42 (from the archive site) as the answers may have changed in newer releases. Thank you.)

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 'Similarity Match' 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.