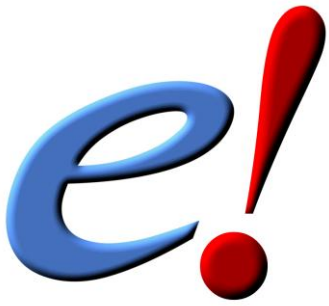
Access to genes and genomes with Ensembl



Course Manual

Nov, 2008

CONTENTS

I) 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

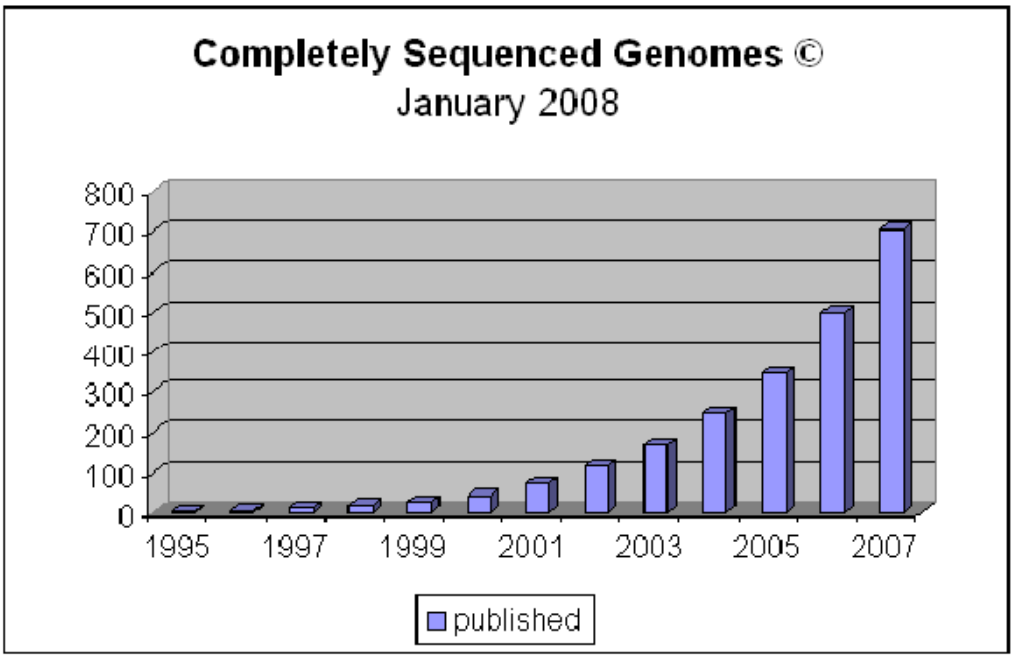


Figure 1. Completely sequenced genomes as of Jan, 2008 (figure taken from http://www.genomesonline.org).

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, and BLAST/BLAT sequence searching and alignment.

Further reading

Flicek, P. *et. al*
Ensembl 2008
Nucleic Acids Res. Jan 2008; 36: D707 - D714

Spudich, G., Fernández-Suárez, X. M., and Birney, E.
Genome Browsing with Ensembl: a practical overview
Brief Funct Genomic Proteomic, 2007 Sept; 6: 202-219

Fernández Suárez X. M. and Schuster M.
Using the Ensembl Genome Server to Browse Genomic Sequence Data. *Current Protocols in Bioinformatics*, UNIT 1.15, January 2007.

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*.[1]
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)

* Additional references can be found here:
http://www.ensembl.org/info/about/publications.html

[1] 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: Mammals		ASSEMBLY		GENEBUILD	
Armadillo	*Dasypus novemcinctus*	ARMA	May-05	Ensembl	Aug-06
Bushbaby	*Otolemur garnettii*	otoGar1	May-06	Ensembl	Feb-07
Cat	*Felis catus*	CAT	Feb-07	Ensembl	Jun-06
Chimpanzee	*Pan troglodytes*	PanTro 2.1	Mar-06	Ensembl	May-07
Cow	*Bos taurus*	Btau 3.1	Feb-07	Ensembl	Sep-06
Dog	*Canis familaiaris*	CanFam 2.0	May-06	Ensembl	Dec-06
Elephant	*Loxodonta africana*	BROAD E1	May-05	Ensembl	Aug-06
Guinea pig	*Cavia porcellus*	cavPor2	Feb-07	Ensembl	Oct-06
Hedgehog	*Erinaceus europaeus*	eriEur1	Feb-07	Ensembl	Oct-06
Horse	*Equus caballus*	Equus1	Sep-07	Ensembl	Sep-07
Human	*Homo sapiens*	NCBI 36	Oct-05	Ensembl	Sep-07
Lesser hedgehog tenrec	*Echinops telfairi*	TENREC	May-05	Ensembl	Aug-06
Microbat	*Myotis lugigfugus*	myoLuc1	Mar-06	Ensembl	Jan-07
Mouse	*Mus musculus*	NCBI m37	Apr-07	Ensembl	Sep-07
Mouse Lemur	*Microcebus murinus*	micMur1	Jun-07	Ensembl	Jul-07
Opossum	*Monodelphis domestica*	MonDom 5.0	Oct-06	Ensembl	Feb-07
Orangutan	*Pongo pygmaeus*	PPYG2	Sep-07	Ensembl	Oct-07
Pig*	*Sus scrofa*	Sscrofa1			
Pika	*Ochotona princeps*	OchPri2.0	Jun-07	Ensembl	Jul-07
Platypus	*Ornithorhynchus anatinus*	OANA 5	Dec-05	Ensembl	Jan-07
Rabbit	*Oryctolagus cuniculus*	RABBIT	May-05	Ensembl	Aug-06
Rat	*Rattus norvegicus*	RGSC 3.4	Dec-04	Ensembl	Feb-06
Rhesus macaque	*Macaca mulatta*	MMUL 1	Feb-06	Ensembl	Aug-06
Shrew	*Sorex araneus*	sorAra1	Oct-05	Ensembl	Apr-07
Squirrel	*Spermophilus tridecemlineatus*	speTri1	Jun-06	Ensembl	Oct-06
Tree shrew	*Tupaia belangeri*	tupBel1	Feb-07	Ensembl	Oct-06
Species: Other					
Aedes	*Aedes aegypti*	AaegL 1	Oct-05	VectorBase	Jun-06
Anole Lizard*	*Anolis carolinensis*	AnoCar1.0	Feb-07		
Anopheles	*Anopheles gambiae*	AgamP 3	Feb-06	VectorBase	Jun-07
C. elegans	*Caenorhabditis elegans*	WS 180	Sep-07	WormBase	Sep-07
Chicken	*Gallus gallus*	WASHUC 2	May-06	Ensembl	Aug-06
C. intestinalis	*Ciona intestinalis*	JGI 2	Mar-05	Ensembl	Feb-06
C. savignyi	*Ciona savignyi*	CSAV 2.0	Oct-05	Ensembl	Apr-06
Fruitfly	*Drosophila melanogaster*	BDGP 5.4	Nov-07	FlyBase	Oct-07
Lamprey*	*Petromyzon marinus*	PMAL3			
Medaka	*Oryzias latipes*	HdrR 1	Oct-05	Ensembl	May-06
S. cerevisiae	*Saccharomyces cerevisiae*	SGD 1.01	Dec-06	SGD	Dec-06
Stickleback	*Gasterosteus aculeatus*	BROAD S1	Feb-06	Ensembl	Jun-06
Takifugu	*Takifugu rubripes*	FUGU 4.0	Jun-05	Ensembl	Nov-07
Tetraodon	*Tetraodon nigroviridis*	TETRAODON 7	Apr-03	Genoscope	Sep-04
X. tropicalis	*Xenopus tropicalis*	JGI 4.1	Aug-05	Ensembl	Nov-05
Zebrafish	*Danio rerio*	Zv 7	Apr-07	Ensembl	Jun-07

Table 1 – Species, assemblies and gene sets in Ensembl
(*currently available in Pre! website pre.ensembl.org)

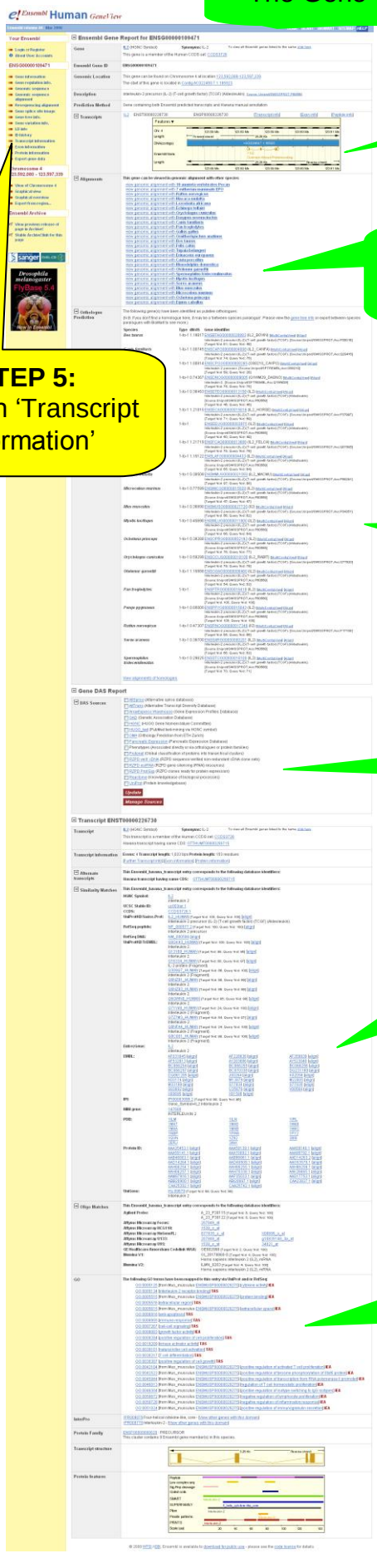
II) WORKED EXAMPLE – A walk through the main pages of the Ensembl browser, using the IL2 (Interleukin-2 precursor) gene as an example.

STEP 1:
Load Ensembl
www.ensembl.org

STEP 3:
Type in ‘IL2 Gene’.
Click ‘Go’.


Ensembl Human SearchView
Search>>
Ensembl release 49 - Mar 2008
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Feature type
Gene (6)
Homo sapiens (6)
Species
Homo sapiens (6)
Gene (6)
Exalead Help
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Pufferfish
Takifugu rubripes
4.0
Full Ensembl genebuild
Ensembl text search
IL2 gene
Search
Your query matched 6 entries in the search database
Vega protein_coding Gene: OTTHUMG00000133075 (HGNC Symbol: IL2) [ContigView]
Vega protein_coding gene OTTHUMG00000133075 has 2 transcripts:
15 exons: ENSE00000673455, ENSE00000673456, ENSE00000673457,
ENSE00001107670, ENSE00001107672, ENSE00001107679, ENSE00001107684,
interleukin 2
The gene has the following external identifiers mapped to it:
CCDS: CCDS3726, CCDS3726.1
Ensembl Human Gene: ENSG00000109471
Ensembl transcript sharing CDS with Havana: ENST00000226730
Entrez Gene: 3558, IL2
HGNC Symbol: IL2: 6001
MIM gene: 147680
RefSeq DNA: NM_000586
Sequence Publications: 3260003
UniProtKB/Swiss-Prot: P60568
Vega gene: OTTHUMG00000133075, IL2
Vega transcript: OTTHUMT00000316449, OTTHUMT00000256715, IL2-001, IL2-002
Vega translation: 104039, OTTHUMP00000164090
Source: e49; Feature type: Gene; Homo sapiens; Gene;
Ensembl protein_coding Gene: ENSG00000109471 (HGNC Symbol: IL2) [ContigView]
Ensembl protein_coding gene ENSG00000109471 has 1 transcript: ENST00000226730, associated peptide: ENSP00000226730 and 4 exons: ENSE00000935278, ENSE00000935280, ENSE00001138256, ENSE00001293064
Interleukin-2 precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin). [Source:Uniprot/SWISSPROT;Acc:P60568]
The gene has the following external identifiers mapped to it:
Affymx Microarray Focus: 207849_at
Affymx Microarray HCG110: 1538_s_at
Affymx Microarray HuGene FL: S77835_s_at, X00695_s_at
Affymx Microarray U133: 207849_at, g10835148_3p_at
Affymx Microarray U95: 34021_at, 1538_s_at
Agilent Probe: A_23_P30115, A_23_P30122
CCDS: CCDS3726, CCDS3726.1
EMBL: U25676, BC070338, K03174, AF031845, S82692, AF532913, BC066257, V00564, DQ231169, DQ861285, X01586, BC066256, X00695, S77835, J00264, AY283686, AF359939, K02056, BC066254, AF228636, S77834, BC066255, M22005, M13879, AY523040, M33199
Entrez Gene: lymphokine, IL-2, TCGF, 3558, IL2
GE Healthcare/Amersham Codelink WGA: GE902998
GO: GO:0008083, GO:0005615, GO:0050672, GO:0030307, GO:0050728, GO:0051024, GO:0005515, GO:0005576, GO:0005134, GO:0042104, GO:0019209, GO:0005125, GO:0006916, GO:0007267, GO:0030101, GO:0042523, GO:0048304, GO:0045944, GO:0006955, GO:0030217, GO:0046013, GO:0008284
Havana transcript having same CDS: OTTHUMT00000256715, IL2-001
HGNC Symbol: IL-2, IL2: 6001
Illumina V1: GI_28178860-S
Illumina V2: ILMN_9283
IPI: IPI00003099.2, IPI00003099
Done
A 'VEGA' gene (VEGA is a team of manual curators external to Ensembl)
STEP 4:
Click on 'ENSG00000109471'

The Gene View Page



Gene Model

Whole Genome Alignments

Orthologues in other species

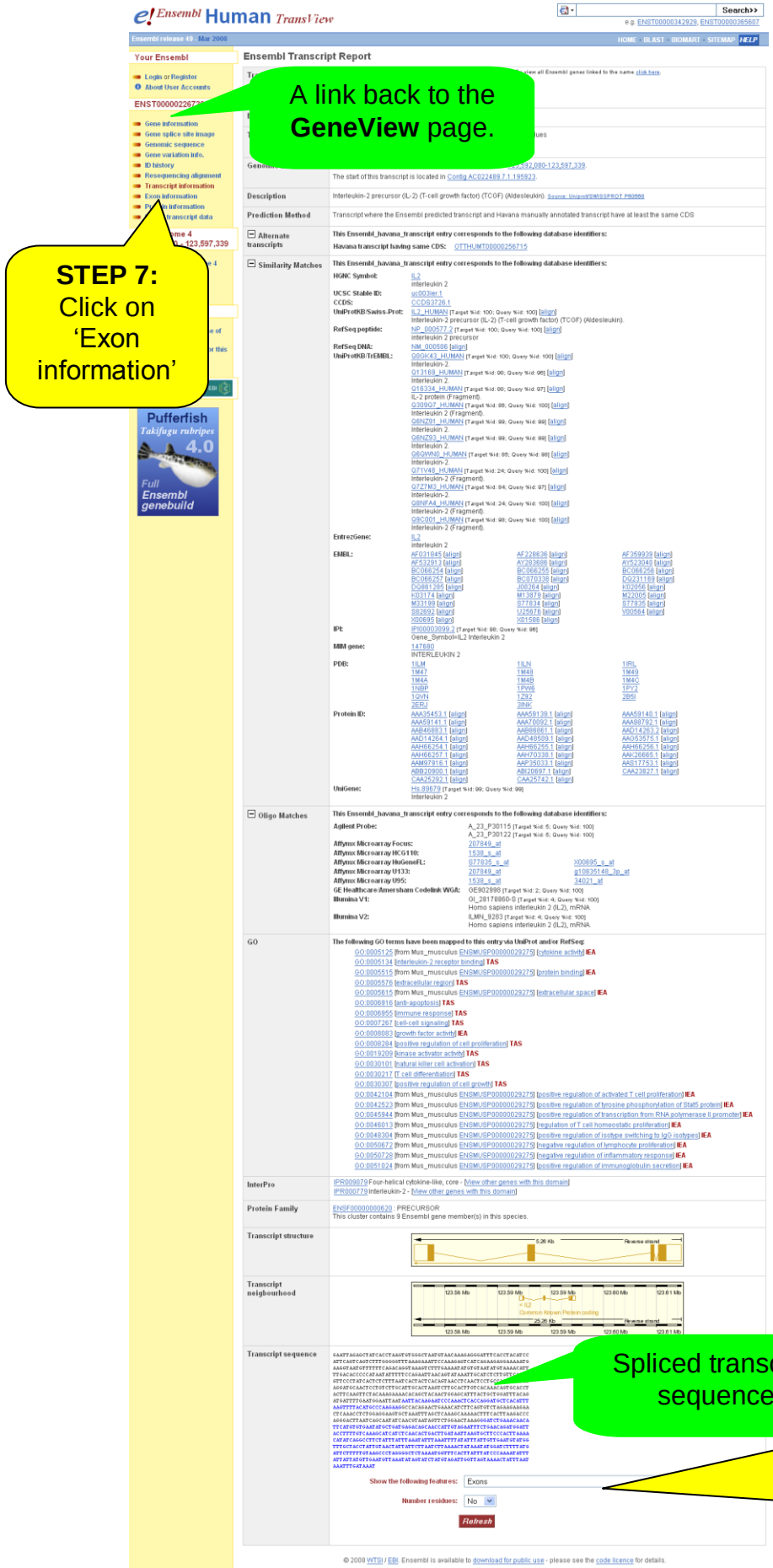
External Information (DAS)

IDs in other databases

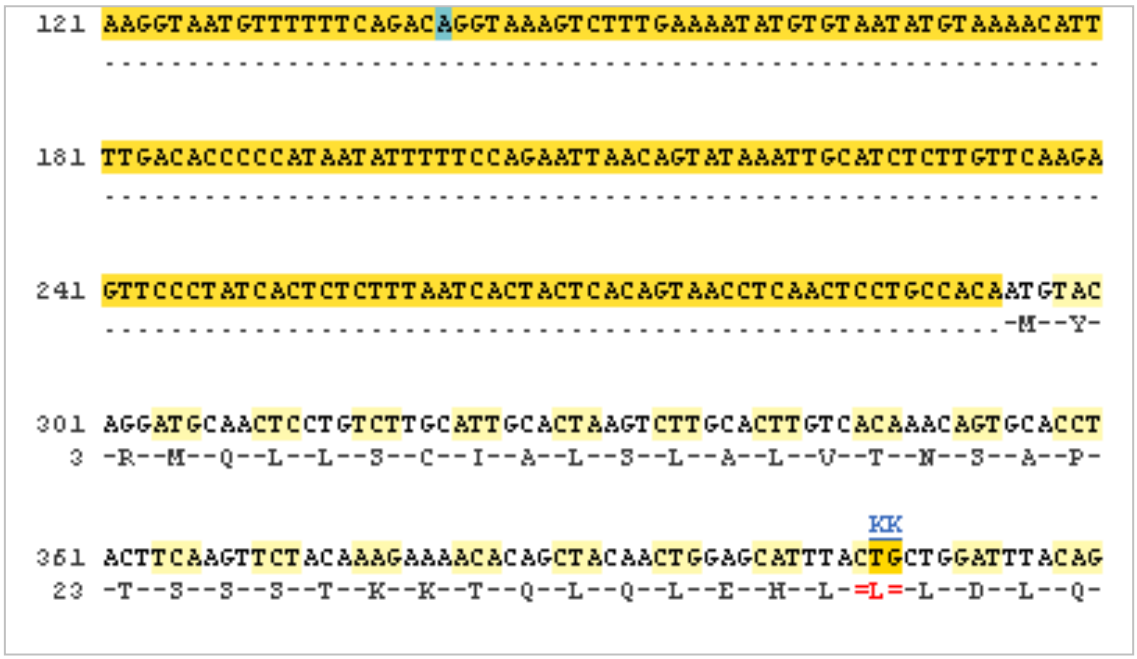
GO (Gene Ontology) terms

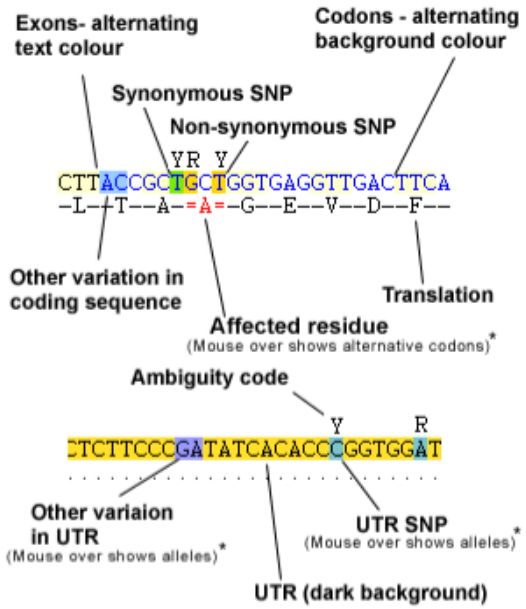
Protein domains

STEP 5:
Click on 'Transcript Information'

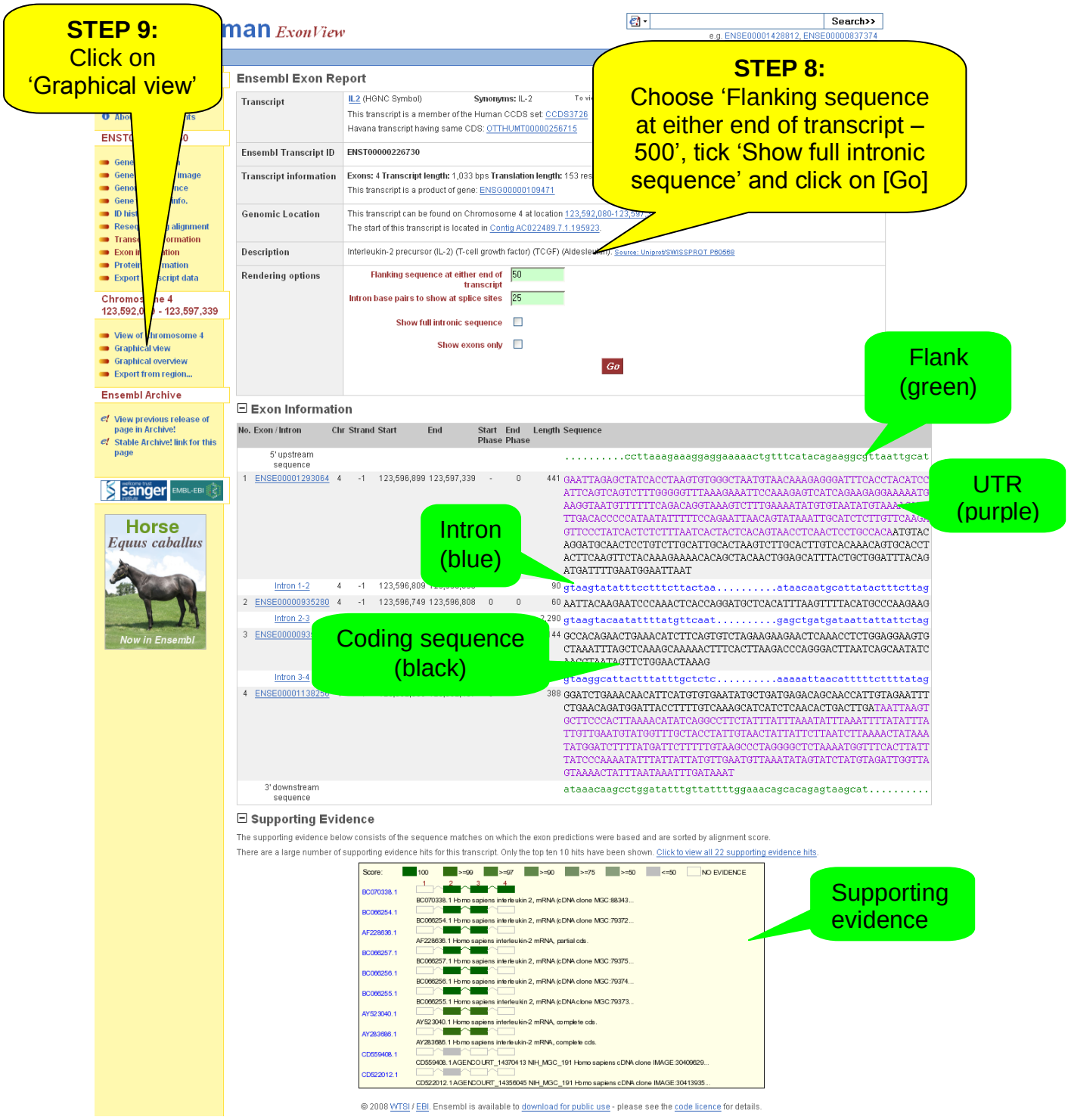

Ensembl Human TransView
Search>>
e.g. ENST00000342929, ENST00000365607
Ensembl release 49 - Mar 2008
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Gene information
Gene splice site image
Genomic sequence
Gene variation info.
ID history
Resequencing alignment
Transcript information
Exon information
Transcript data
Ensembl Transcript Report
A link back to the GeneView page.
STEP 7: Click on 'Exon information'
Transcript report: each transcript for a gene has a 'TransView' page.
Description
Interleukin-2 precursor (IL-2) (T-cell growth factor) (TCGF) (Aldesleukin).
Prediction Method
Transcript where the Ensembl predicted transcript and Havana manually annotated transcript have at least the same CDS
Alternate transcripts
Similarity Matches
Oligo Matches
GO
InterPro
Protein Family
Transcript structure
Transcript neigbourhood
Transcript sequence
Spliced transcript sequence
Show the following features:
Exons
Number residues:
No
Refresh
STEP 6: Select 'Exons, Codons, Translations and SNPs'. Select 'Number residues: Yes' and click on [Refresh]
Pufferfish Takifugu rubripes 4.0 Full Ensembl genebuild
© 2008 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

Result of STEP 6:



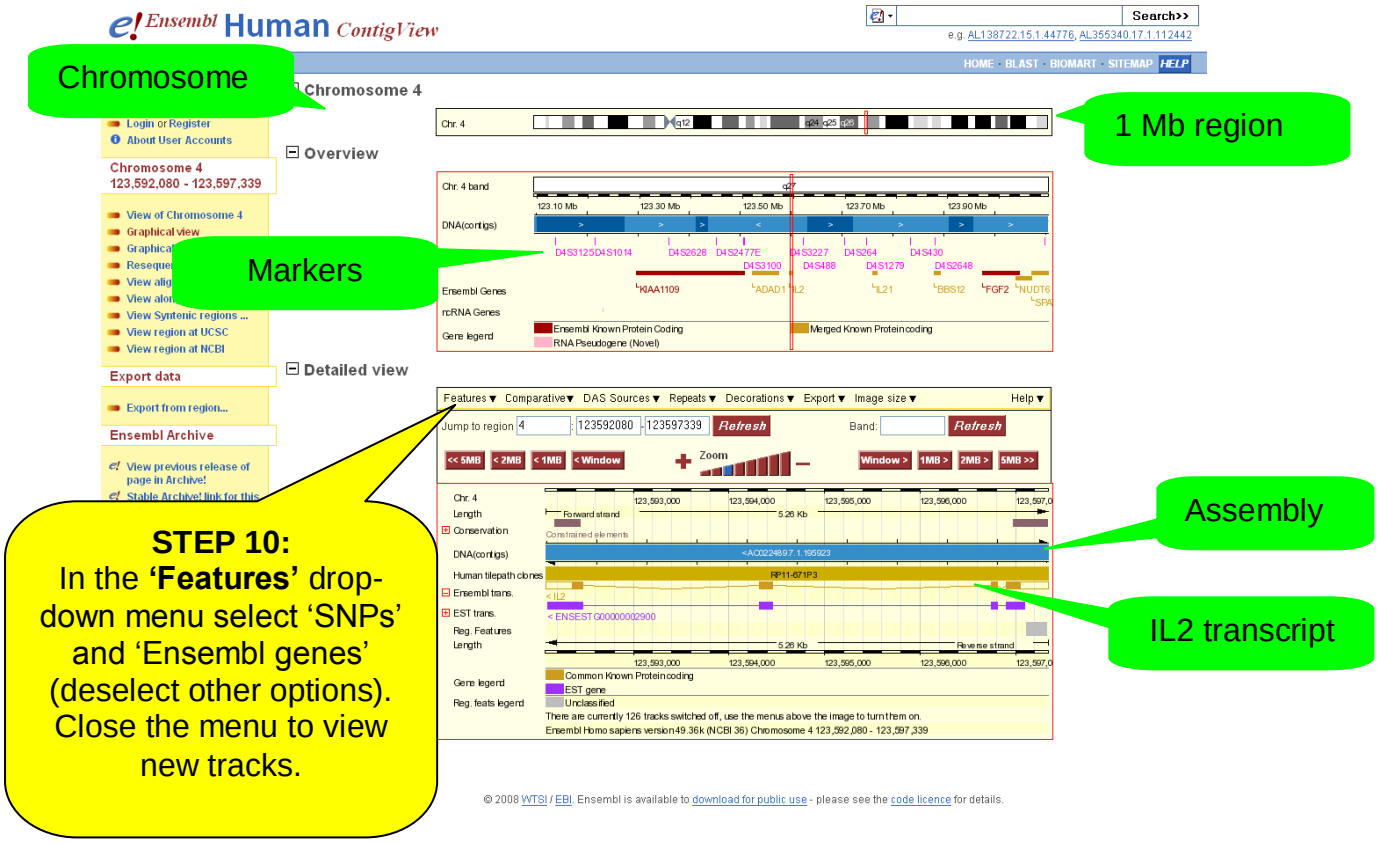


Result of STEP 7:

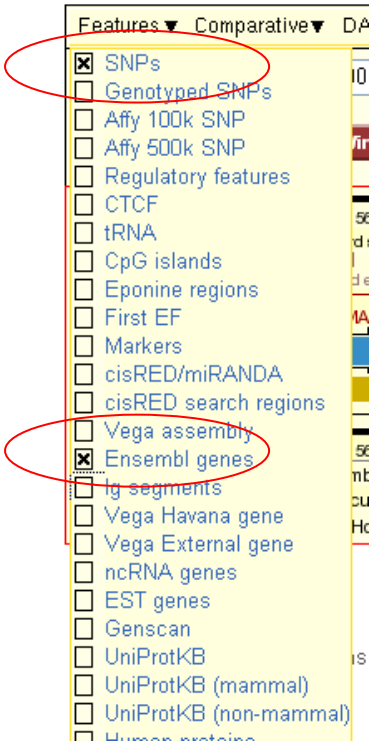


Result of STEP 8:

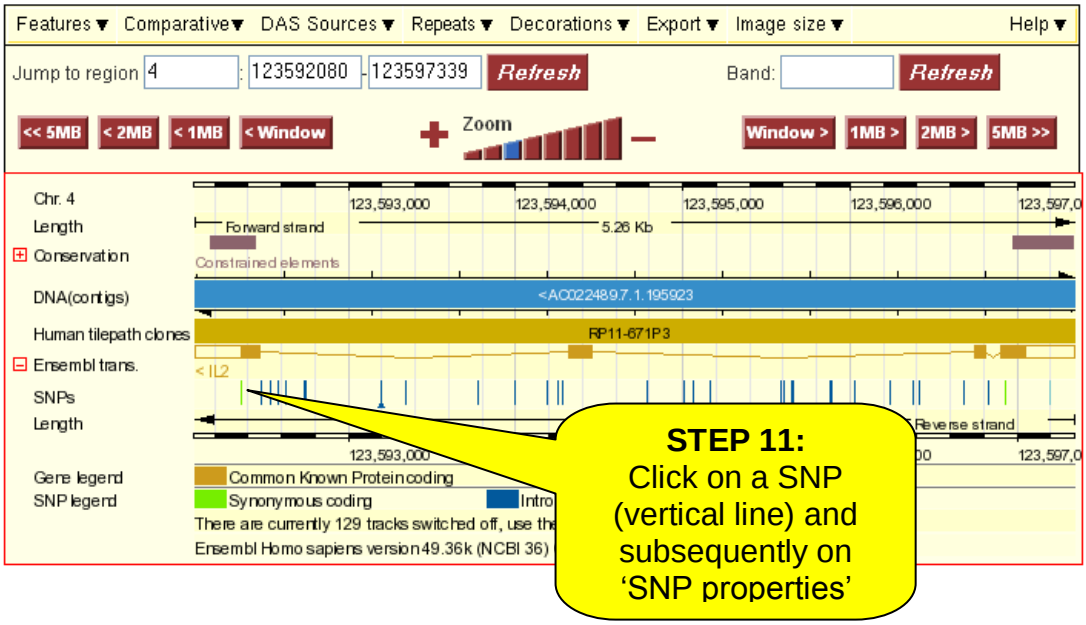
Result of STEP 9: ContigView



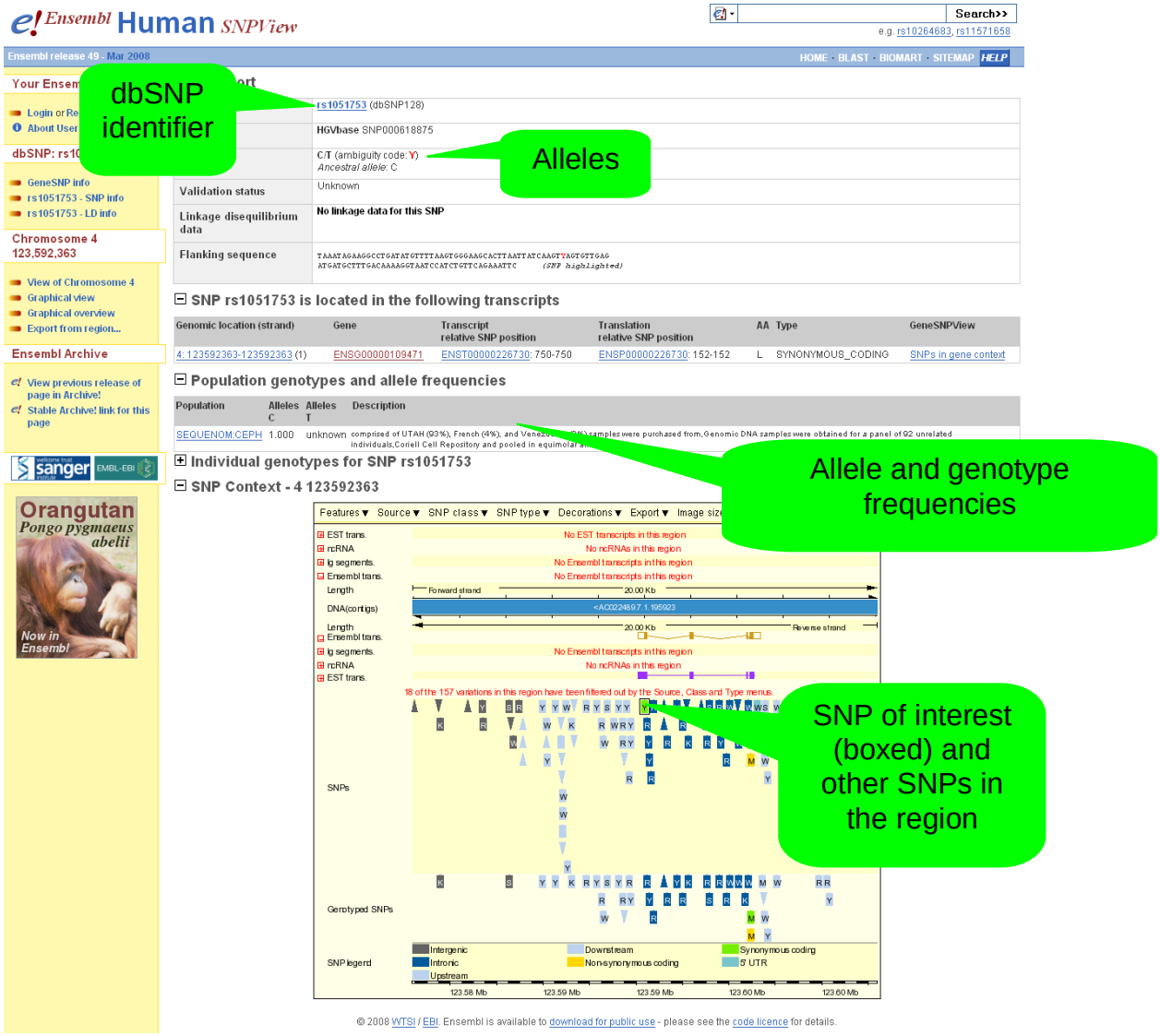
(STEP 10): Deselect all other options from the features menu and select 'SNPs' and 'Ensembl genes'.



RESULT OF STEP 10

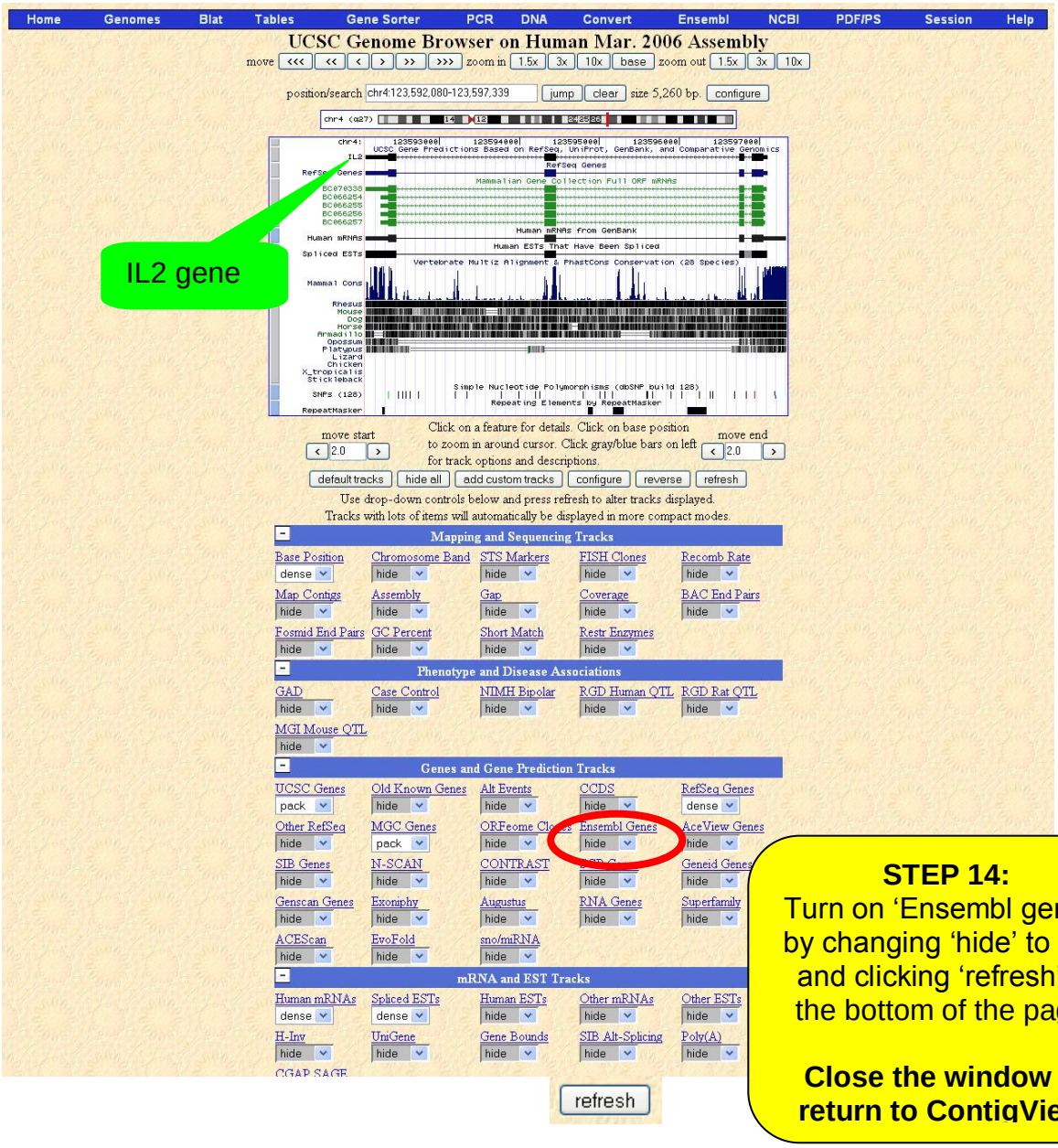


STEP 11:
Click on a SNP (vertical line) and subsequently on 'SNP properties'



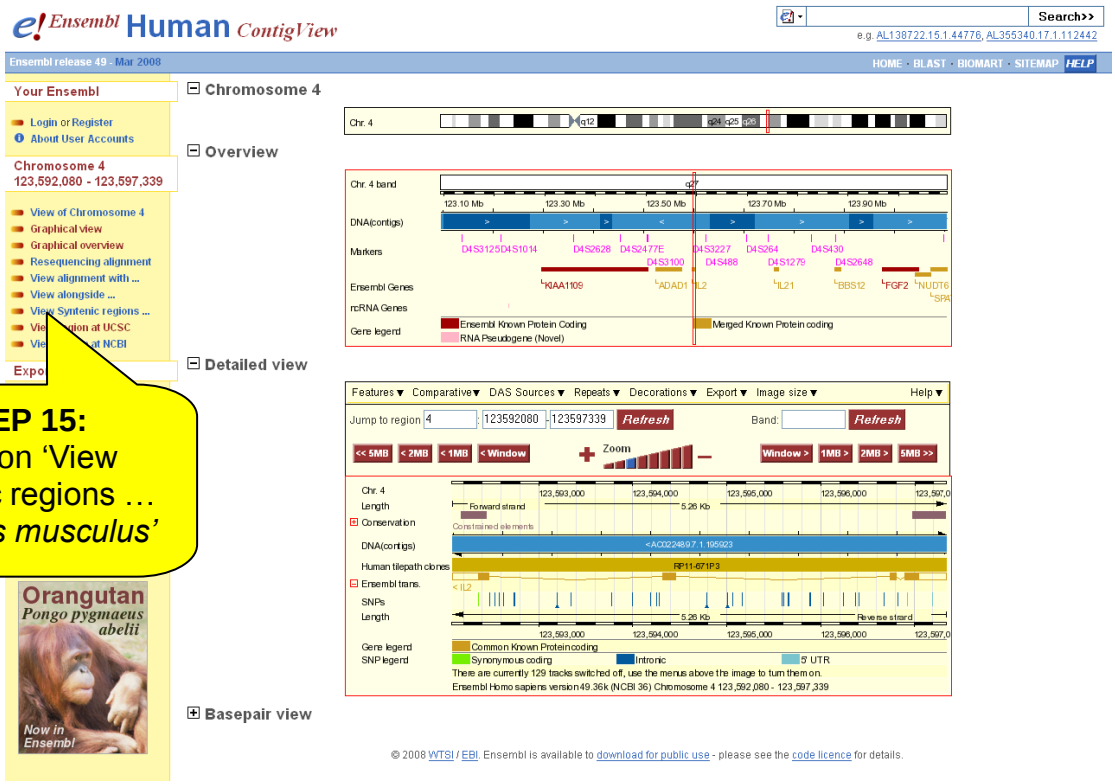
STEP 12:
Go back to ContigView with the back button of the internet browser.

STEP 13:
To see the same chromosomal region in the UCSC genome browser, click on 'View region at UCSC' on the left of the page. A new window will open.

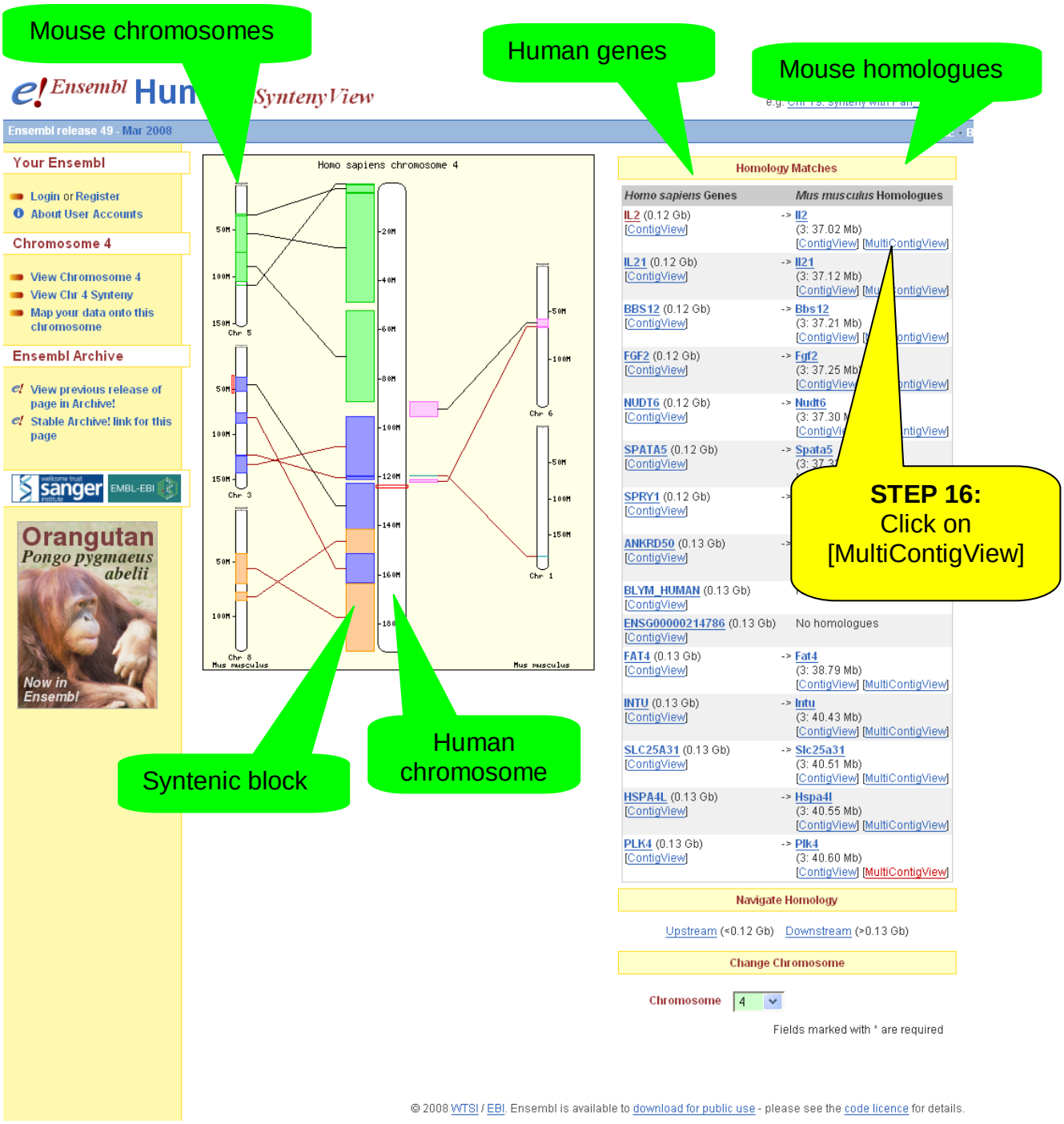


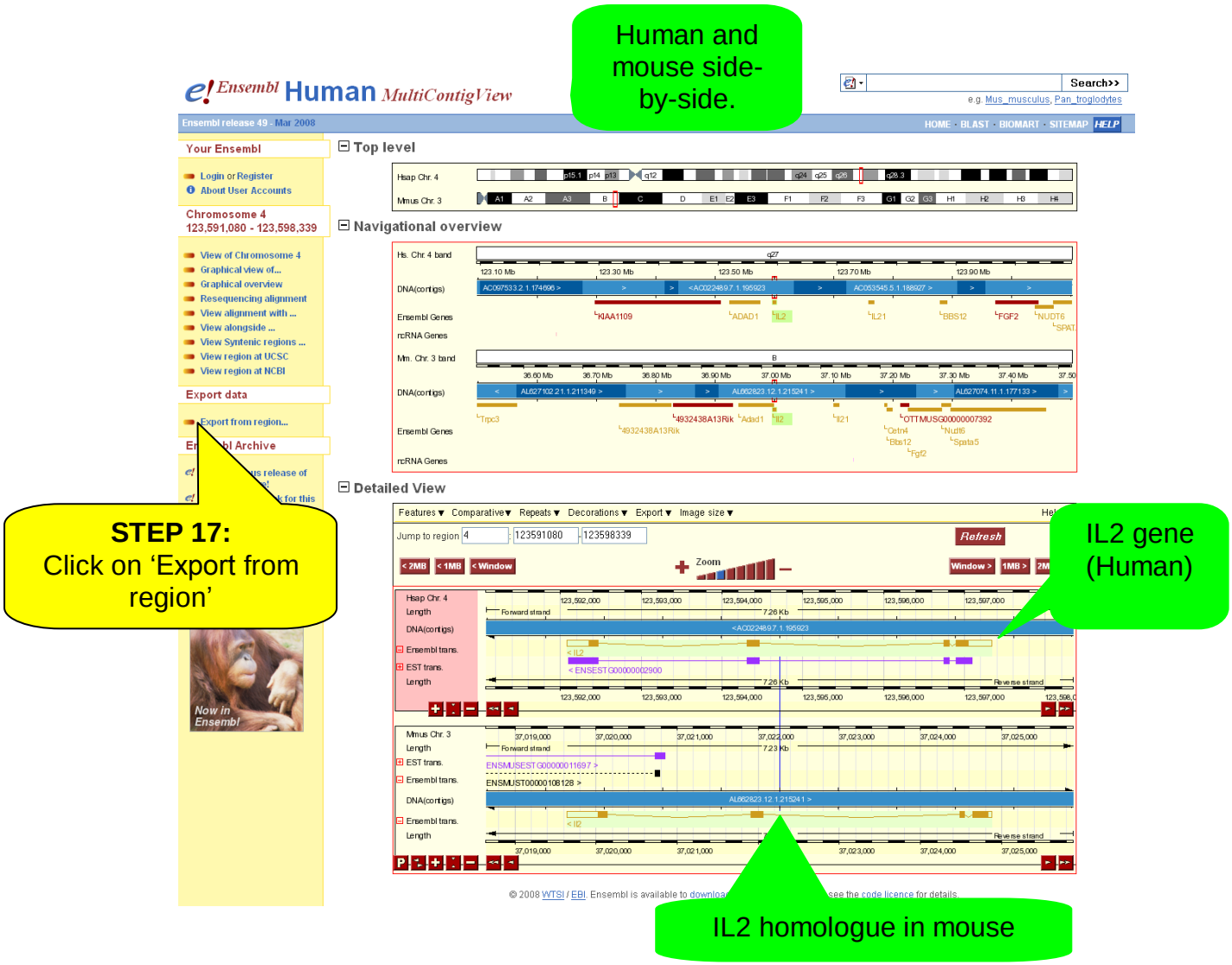
STEP 14:
Turn on 'Ensembl genes' by changing 'hide' to 'full' and clicking 'refresh' at the bottom of the page.

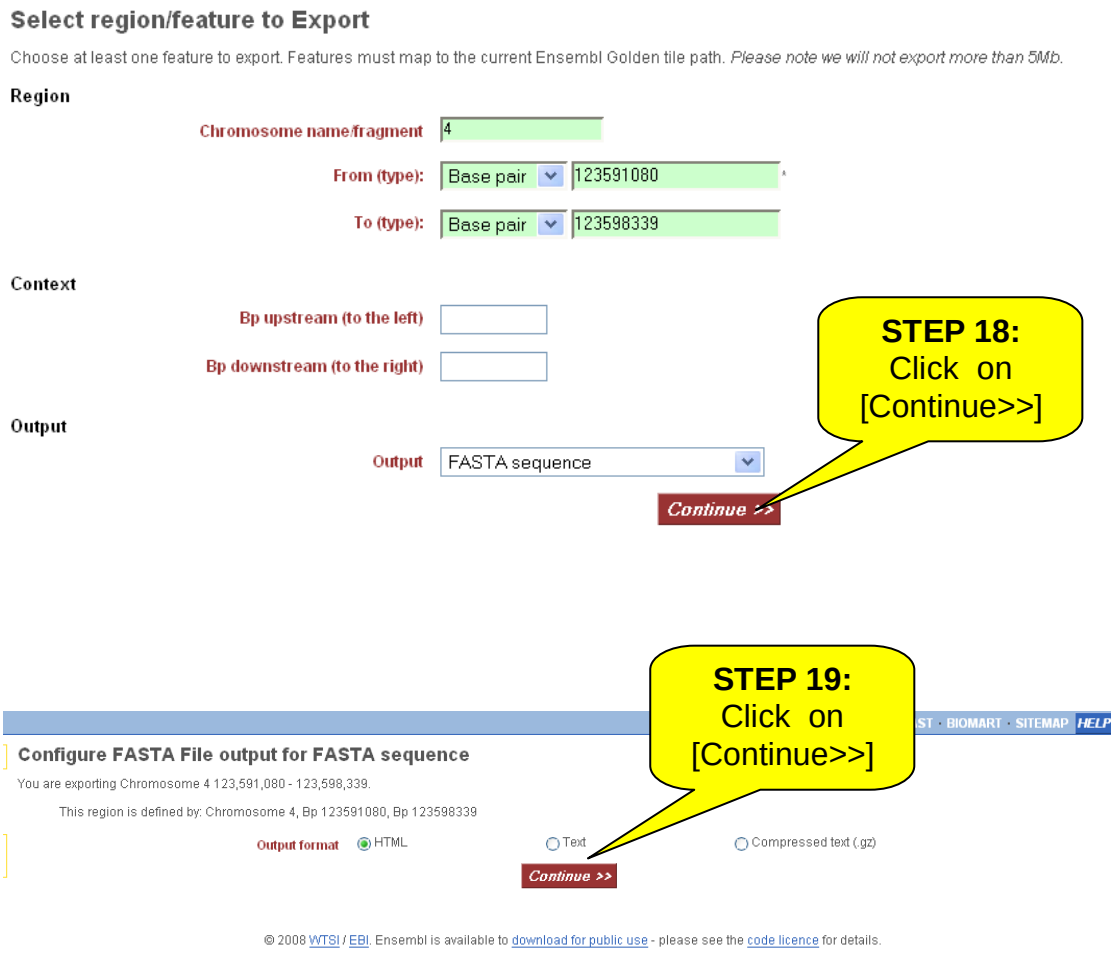
Close the window to return to ContigView.



STEP 15:
Click on 'View Syntenic regions … with *Mus musculus*'

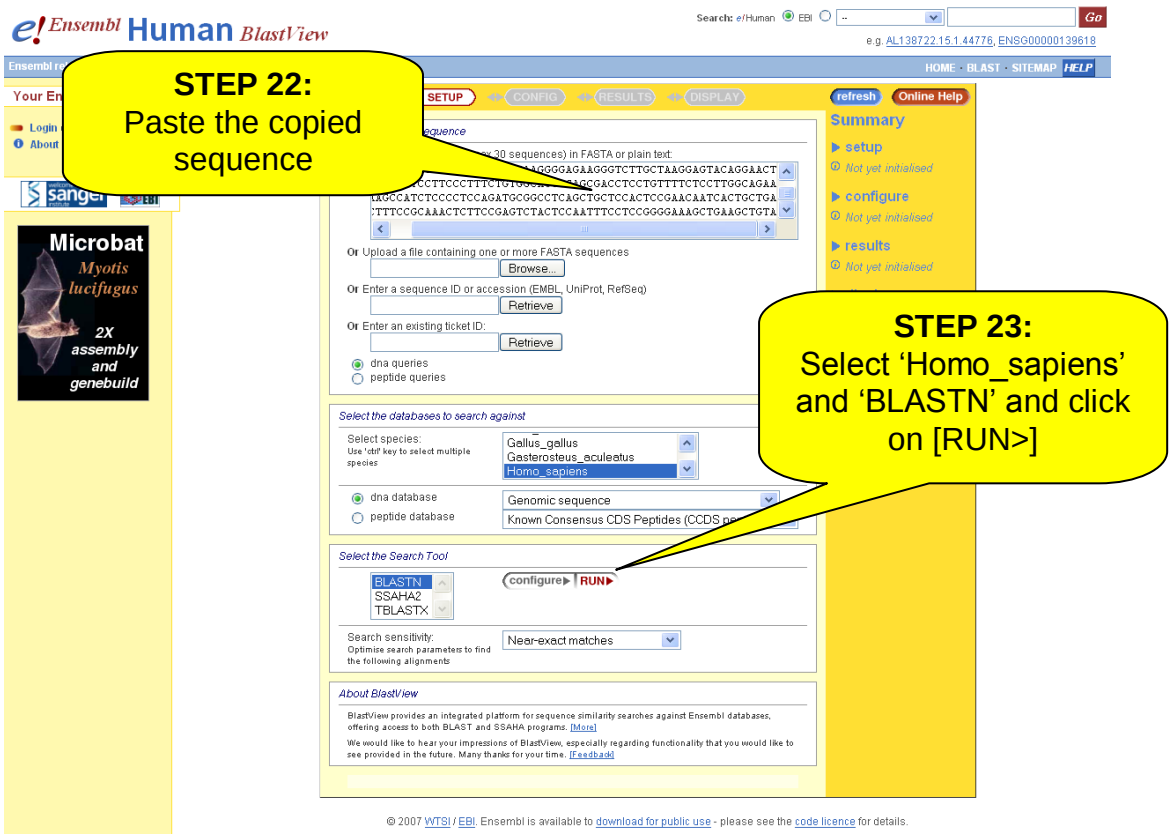

Mouse chromosomes
Human genes
Mouse homologues
Ensembl Hum SyntenyView
Ensembl release 49 - Mar 2008
Your Ensembl
Login or Register
About User Accounts
Chromosome 4
View Chromosome 4
View Chr 4 Synteny
Map your data onto this chromosome
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Homo sapiens chromosome 4
Homology Matches
Homo sapiens Genes
Mus musculus Homologues
IL2 (0.12 Gb) [ContigView]
-> Il2 (3: 37.02 Mb) [ContigView] [MultiContigView]
IL21 (0.12 Gb) [ContigView]
-> Il21 (3: 37.12 Mb) [ContigView] [MultiContigView]
BBS12 (0.12 Gb) [ContigView]
-> Bbs12 (3: 37.21 Mb) [ContigView]
FGF2 (0.12 Gb) [ContigView]
-> Fgf2 (3: 37.25 Mb)
NUDT6 (0.12 Gb) [ContigView]
-> Nudt6 (3: 37.30
SPATA5 (0.12 Gb) [ContigView]
-> Spata5
SPRY1 (0.12 Gb) [ContigView]
ANKRD50 (0.13 Gb) [ContigView]
BLYM_HUMAN (0.13 Gb) [ContigView]
ENSG00000214786 (0.13 Gb) [ContigView]
No homologues
FAT4 (0.13 Gb) [ContigView]
-> Fat4 (3: 38.79 Mb) [ContigView] [MultiContigView]
INTU (0.13 Gb) [ContigView]
-> Intu (3: 40.43 Mb) [ContigView] [MultiContigView]
SLC25A31 (0.13 Gb) [ContigView]
-> Slc25a31 (3: 40.51 Mb) [ContigView] [MultiContigView]
HSPA4L (0.13 Gb) [ContigView]
-> Hspa4l (3: 40.55 Mb) [ContigView] [MultiContigView]
PLK4 (0.13 Gb) [ContigView]
-> Plk4 (3: 40.60 Mb) [ContigView] [MultiContigView]
STEP 16:
Click on
[MultiContigView]
Navigate Homology
Upstream (<0.12 Gb) Downstream (>0.13 Gb)
Change Chromosome
Chromosome 4
Fields marked with * are required
Orangutan
Pongo pygmaeus abelii
Now in Ensembl
Syntenic block
Human chromosome
© 2008 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.


Human and mouse side-by-side.
Ensembl Human MultiContigView
Search>>
e.g. Mus_musculus, Pan_troglodytes
Ensembl release 49 - Mar 2008
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Chromosome 4
123,591,080 - 123,598,339
View of Chromosome 4
Graphical view of...
Graphical overview
Resequencing alignment
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region at NCBI
Export data
Export from region...
Top level
Navigational overview
Detailed View
Refresh
Zoom
IL2 gene (Human)
IL2 homologue in mouse
STEP 17:
Click on 'Export from region'
Now in Ensembl

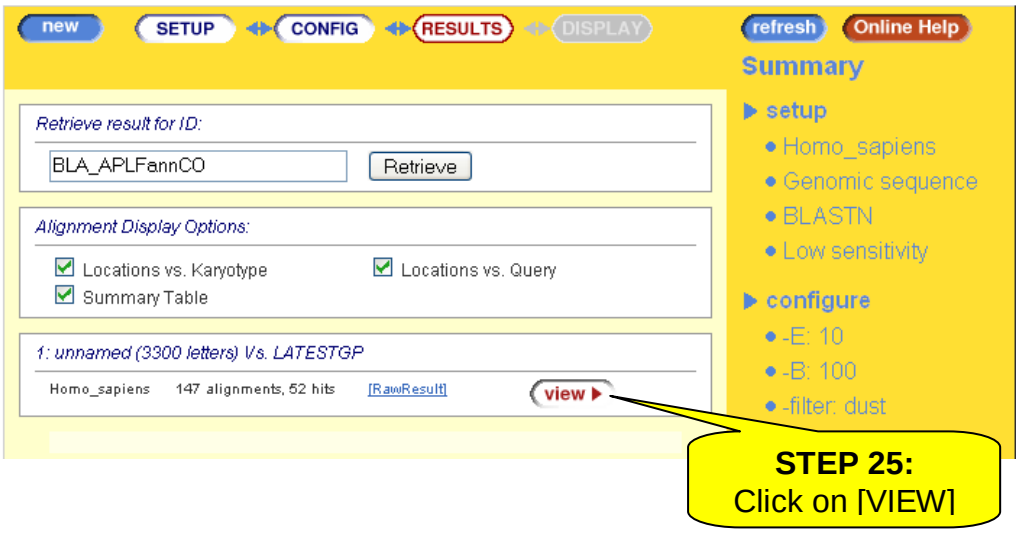

Select region/feature to Export
Choose at least one feature to export. Features must map to the current Ensembl Golden tile path. Please note we will not export more than 5Mb.
Region
Chromosome name/fragment
4
From (type):
Base pair
123591080
To (type):
Base pair
123598339
Context
Bp upstream (to the left)
Bp downstream (to the right)
Output
Output
FASTA sequence
Continue >>
STEP 18:
Click on
[Continue>>]

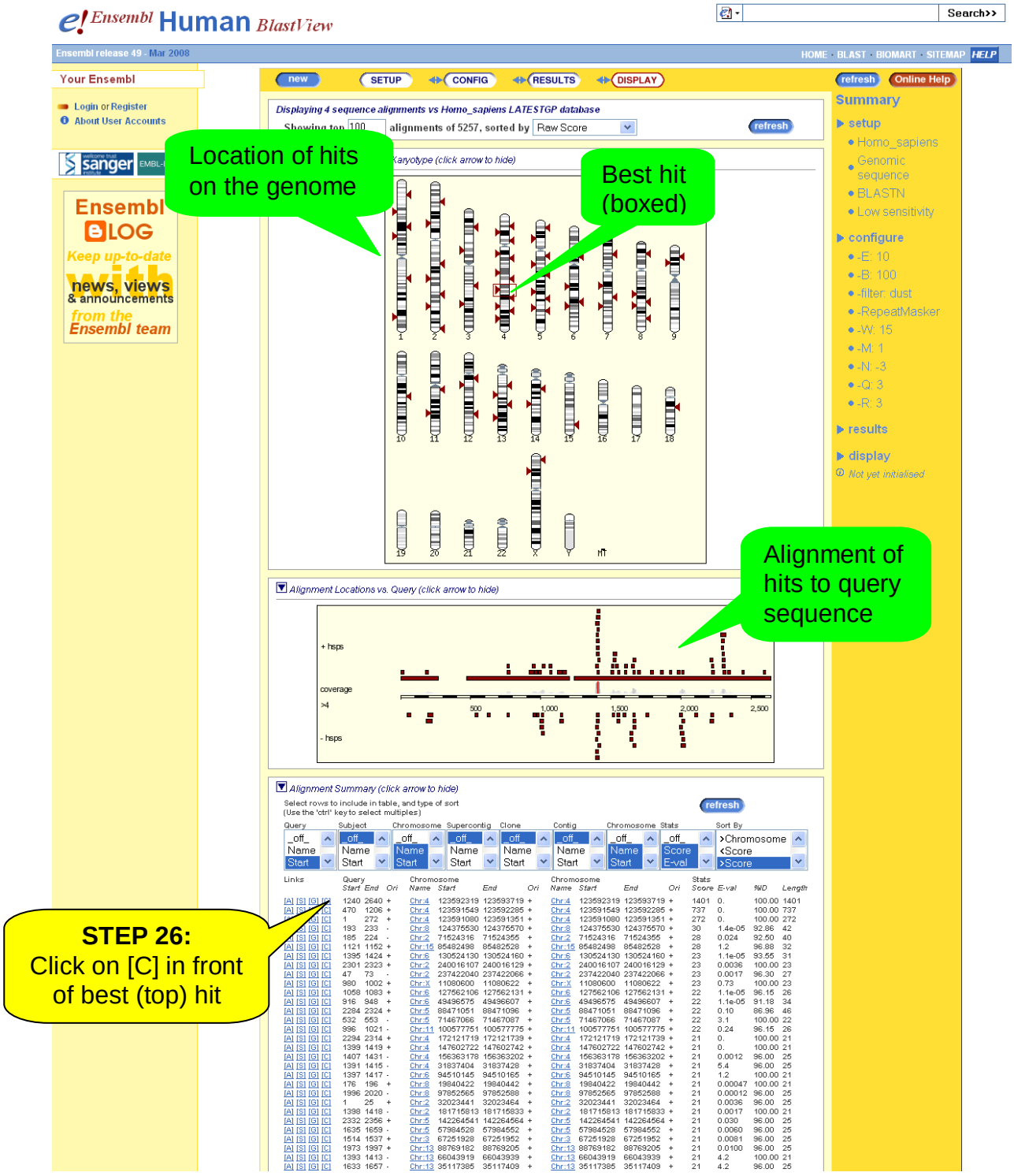
STEP 19:
Click on
[Continue>>]
BIOMART · SITEMAP HELP
Configure FASTA File output for FASTA sequence
You are exporting Chromosome 4 123,591,080 - 123,598,339.
This region is defined by: Chromosome 4, Bp 123591080, Bp 123598339
Output format
HTML
Text
Compressed text (.gz)
Continue >>
© 2008 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

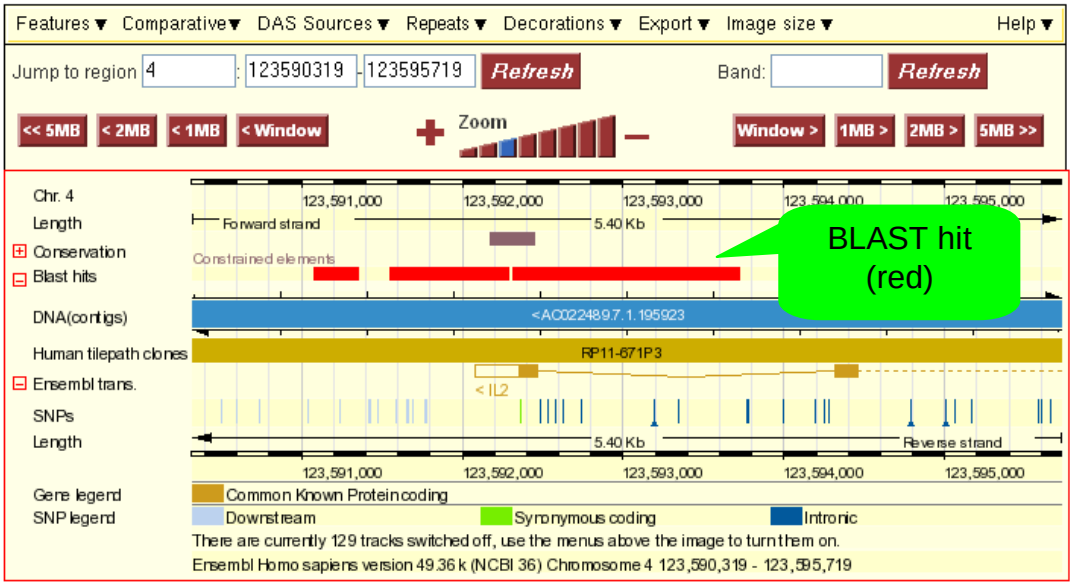

STEP 20:
Select and copy a part of the sequence
STEP 21:
Click on 'BLAST'
Ensembl Human
Search>>
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Chromosome 4
123,591,080 - 123,598,339
View of Chromosome 4
Graphical view
Graphical overview
Resequencing alignment
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region at NCBI
Export data
Export from region...
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Horse
Equus caballus
Now in Ensembl


STEP 22:
Paste the copied sequence
STEP 23:
Select 'Homo_sapiens' and 'BLASTN' and click on [RUN>]
Ensembl Human BlastView
HOME · BLAST · SITEMAP HELP
SETUP
CONFIG
RESULTS
DISPLAY
refresh
Online Help
Summary
setup
Not yet initialised
configure
Not yet initialised
results
Not yet initialised
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:
Use 'ctrl' key to select multiple species
Gallus_gallus
Gasterosteus_aculeatus
Homo_sapiens
dna database
peptide database
Genomic sequence
Select the Search Tool
BLASTN
SSAHA2
TBLASTX
configure
RUN
Search sensitivity:
Optimise search parameters to find the following alignments
Near-exact matches
About BlastView
Microbat
Myotis lucifugus
2X assembly and genebuild


Human BlastView
Ensembl release 44 - Apr 2007
Your Ensembl
Login or Register
About User Accounts
Bushbaby
Otolemur garnettii
2X assembly and genebuild
new
SETUP
CONFIG
RESULTS
DISPLAY
refresh
Online Help
Summary
setup
Homo_sapiens
Genomic sequence
BLASTN
Low sensitivity
Retrieve result for ID:
BLA_APLFannCO
Retrieve
Retrieving Results
1: unnamed (3300 letters) Vs. LATESTGP
Homo_sapiens
Job Queued
Summary of BLAST search
STEP 24:
Click on'Retrieve'


new
SETUP
CONFIG
RESULTS
DISPLAY
refresh
Online Help
Summary
setup
Homo_sapiens
Genomic sequence
BLASTN
Low sensitivity
configure
-E: 10
-B: 100
-filter: dust
Retrieve result for ID:
BLA_APLFannCO
Retrieve
Alignment Display Options:
Locations vs. Karyotype
Locations vs. Query
Summary Table
1: unnamed (3300 letters) Vs. LATESTGP
Homo_sapiens
147 alignments, 52 hits
[RawResult]
view
STEP 25:
Click on [VIEW]


Ensembl Human BlastView
Search>>
Ensembl release 49 - Mar 2008
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Ensembl BLOG
Keep up-to-date with news, views & announcements from the Ensembl team
new
SETUP
CONFIG
RESULTS
DISPLAY
refresh
Online Help
Summary
setup
Homo_sapiens
Genomic sequence
BLASTN
Low sensitivity
configure
-E: 10
-B: 100
-filter: dust
-RepeatMasker
-W: 15
-M: 1
-N: -3
-Q: 3
-R: 3
results
display
Not yet initialised
Displaying 4 sequence alignments vs Homo_sapiens LATESTGP database
Showing top 100 alignments of 5257, sorted by Raw Score
refresh
Location of hits on the genome
Best hit (boxed)
Karyotype (click arrow to hide)
Alignment Locations vs. Query (click arrow to hide)
Alignment of hits to query sequence
+ hsps
coverage
>4
- hsps
Alignment Summary (click arrow to hide)
Select rows to include in table, and type of sort
(Use the 'ctrl' key to select multiples)
refresh
Query
Subject
Chromosome
Supercontig
Clone
Contig
Chromosome
Stats
Sort By
>Chromosome
<Score
>Score
Links
Query
Chromosome
Chromosome
Stats
STEP 26:
Click on [C] in front of best (top) hit
Back in the contigview page…



END of the Worked Example

EXERCISES and ANSWERS

Note: the answers to these exercises correspond to current version (50) of Ensembl. If you use these exercises at a later date, please use the archive site for version 50.

III) BROWSING ENSEMBL

These exercises address using the browser to determine a variety of gene-relevant information such as transcript number and size, protein domains, functional classes and sequence.

1. Exploring features related to a gene

Exercise 1 begins with the TAC1 (tachykinin precursor 1) gene and moves into the browser from the main GeneView page.

(a) Open the home page of Ensembl (www.ensembl.org). This is the current version. 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) 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?

(h) Go to the Ensembl main page. Look up 'Ensembl genes' in the glossary by following the 'Using Ensembl' link in the 'Help and Documentation' at the left. Follow the link: 'Browse the provided data' on the website.
http://www.ensembl.org/Homo_sapiens/glossaryview

2. Exploring a region

Exercise 2 begins with a search for a specific chromosomal region, rather than one gene.

(a) Click on the large 'e!' at the top left of the screen to start a new search. Go to the human homepage, and click on chromosome 12. From 'MapView', choose to display the region between markers D12S764 and D12S1871 (in the 'Jump to ContigView' section).

(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) Click on a marker. What are other names for the marker? Is there an expected product size (what does this mean?)

(d) In ContigView, zoom in three steps on the zoom triangle/ladder of 'Detailed view' (towards the '+') and turn on the SNP track. Identify an intronic SNP and look at the corresponding SNPView page.

Answers (Browsing Ensembl)

1. Exploring features related to a gene

(a) Click on the identifier ENSG00000006128 from the search results. 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'. You should now be in the GeneView page.

(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. See this information in the heading of each 'Transcript section' of the GeneView page. The length of ENST00000319273 is 1060 bp. It has 7 exons and codes for 129 aa.

(c) The TAC1 gene encodes for Protachykinin 1 precursor. Follow the links to MIM and EntrezGene or UniProt/Swiss-Prot in the 'Similarity Matches' section to learn more. 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 showing the position of that term in the hierarchical 'GO tree' (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) and clone RP5-841B21 ('Human tilepath clones' track).

(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 (band A1).

2. Exploring a region

(a) 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 region between the two markers will be displayed 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. Clones are shown in gold and pink. Portions of the 'Tile path clones' sequences were used to form the assembly and correspond to 'contigs'. The clones overlap each other whereas the contigs don't.

(c) Click on a marker name (shown in pink at the top) and follow the link 'Marker info' to the MarkerView page. Other names in the UniSTS database will be at the top. Synonyms (or names in other databases) are listed further down. The expected product size is the calculated size of the fragment (in base pairs) if both primers are used against the genome.

(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.

IV) Data mining in Ensembl with BioMart
Worked Example

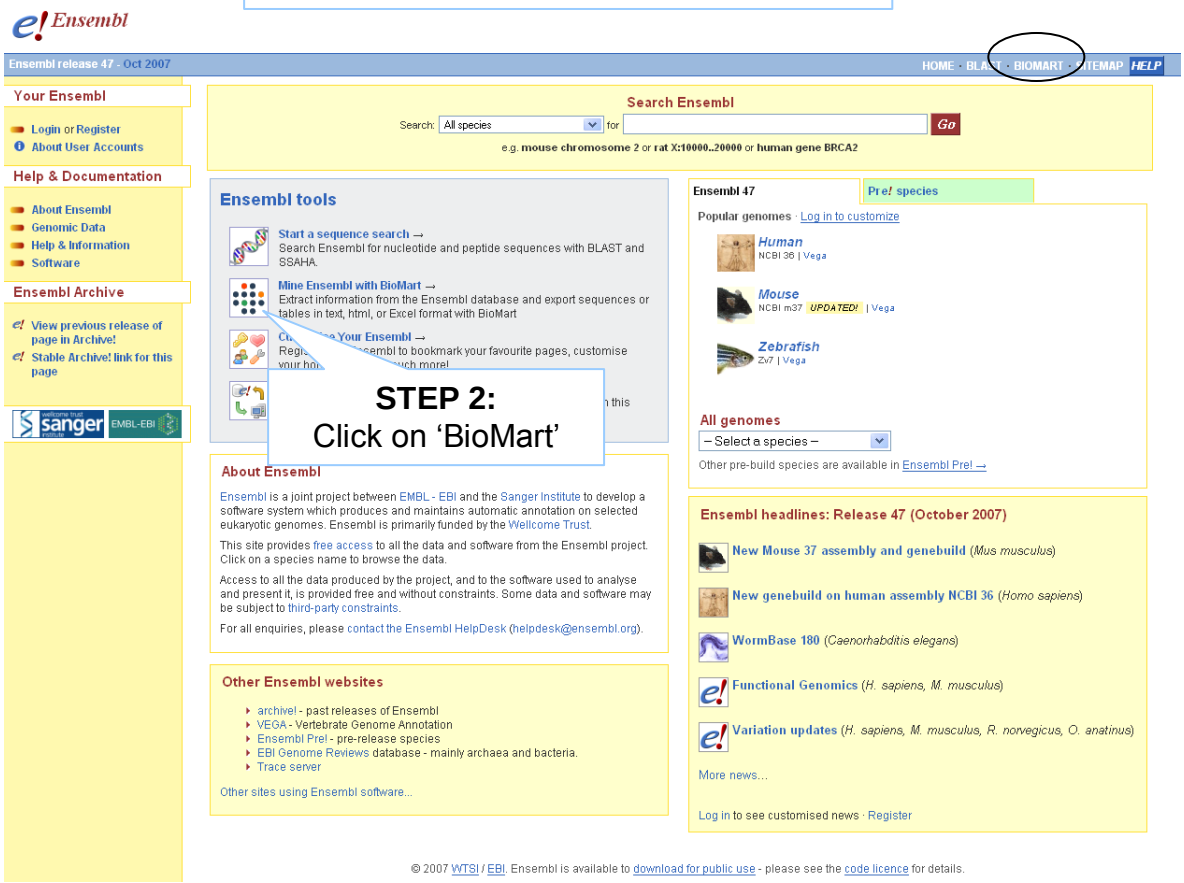
The human gene encoding Glucose-6-phosphate dehydrogenase (G6PD) is located on chromosome X in cytogenetic band q28.

Which other genes related to human diseases locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs?

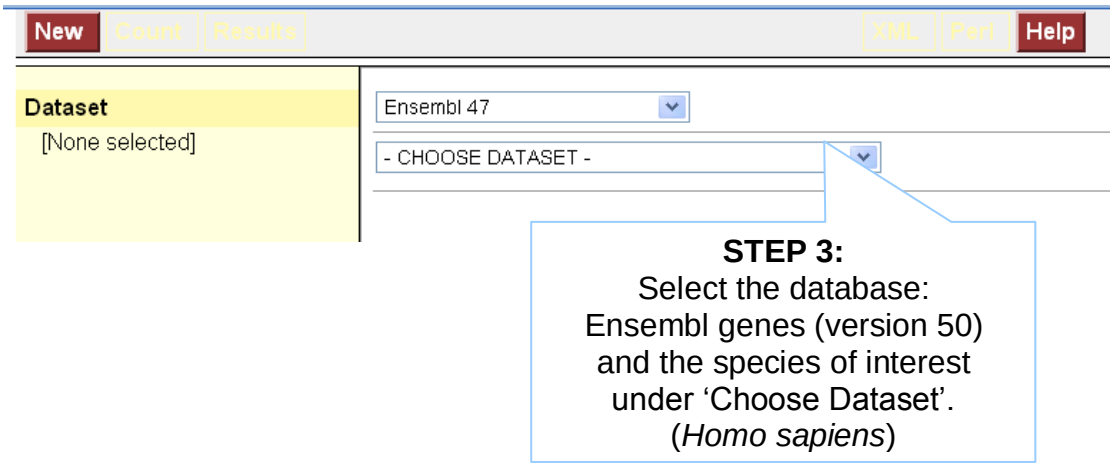
What are their cDNA sequences?

Follow the worked example below to answer these questions.

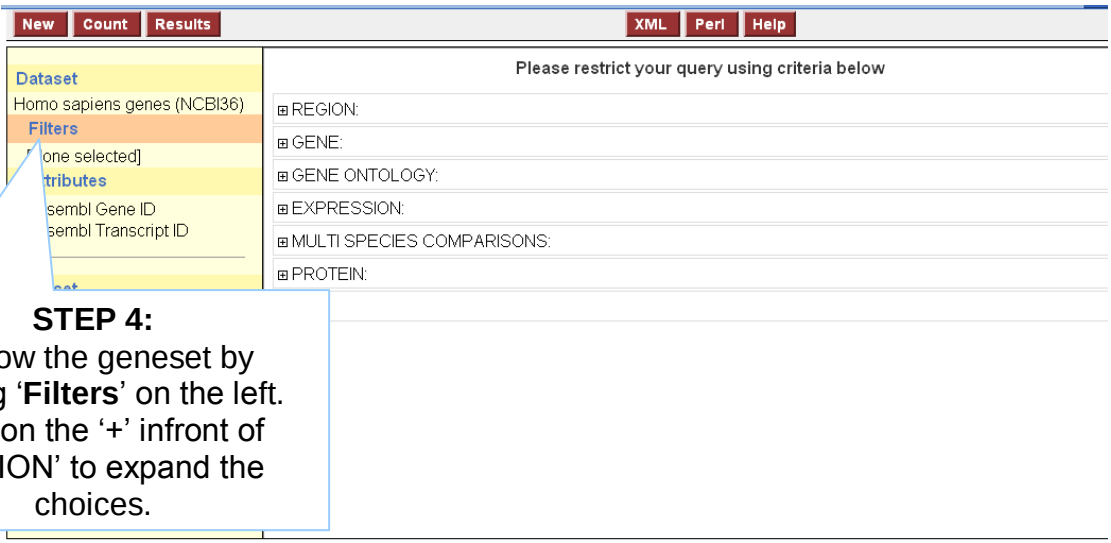
STEP 1:
Go to the Ensembl main page
www.ensembl.org



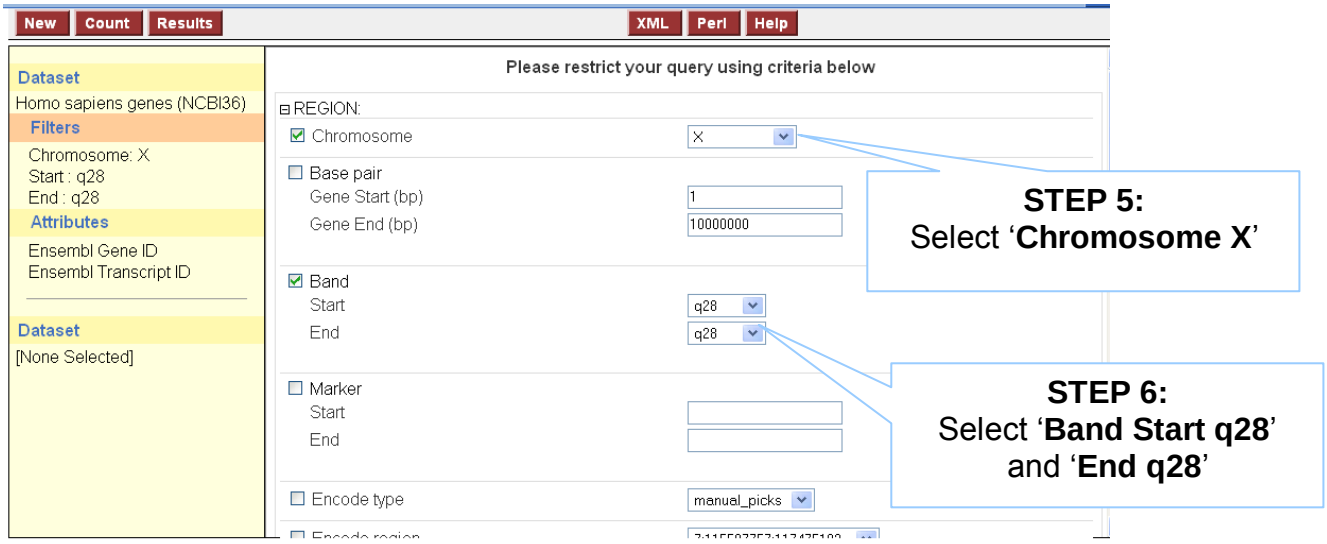
STEP 2:
Click on 'BioMart'

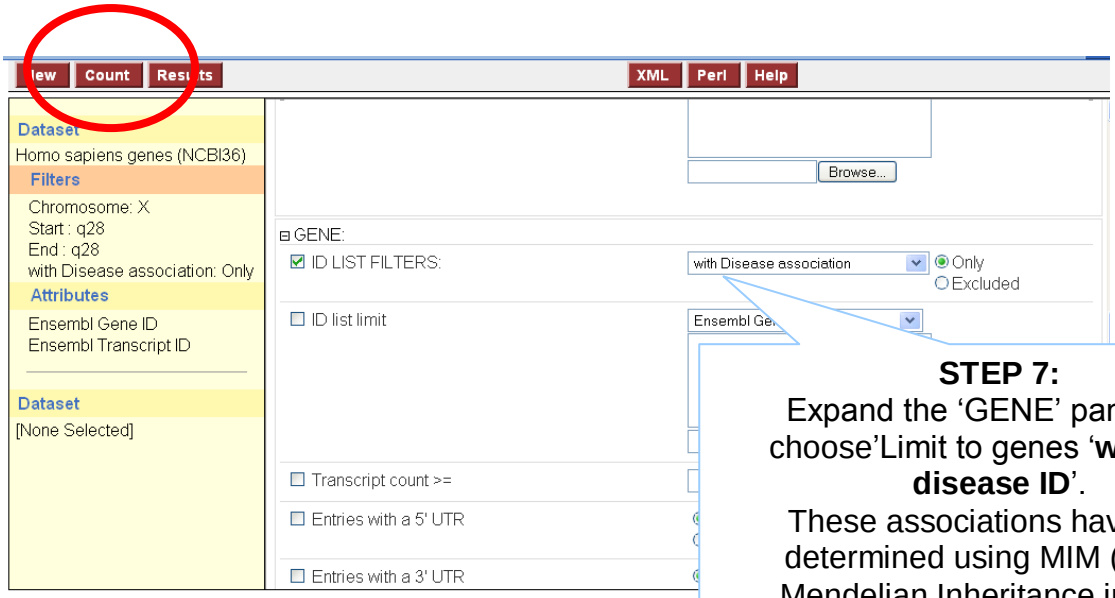


STEP 3:
Select the database: Ensembl genes (version 50) and the species of interest under 'Choose Dataset'. (*Homo sapiens*)



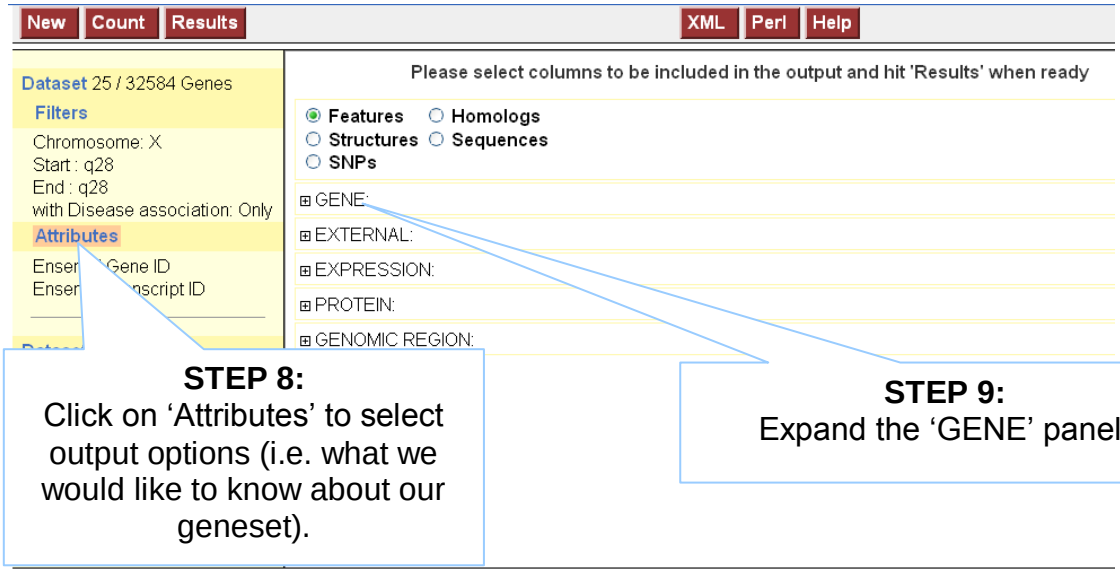
STEP 4:
Narrow the geneset by clicking '**Filters**' on the left. Click on the '+' infront of 'REGION' to expand the choices.


New
Count
Results
XML
Perl
Help
Dataset
Homo sapiens genes (NCBI36)
Filters
Chromosome: X
Start : q28
End : q28
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]
Please restrict your query using criteria below
REGION:
Chromosome
X
Base pair
Gene Start (bp)
1
Gene End (bp)
10000000
Band
Start
q28
End
q28
Marker
Start
End
Encode type
manual_picks
STEP 5:
Select 'Chromosome X'
STEP 6:
Select 'Band Start q28' and 'End q28'



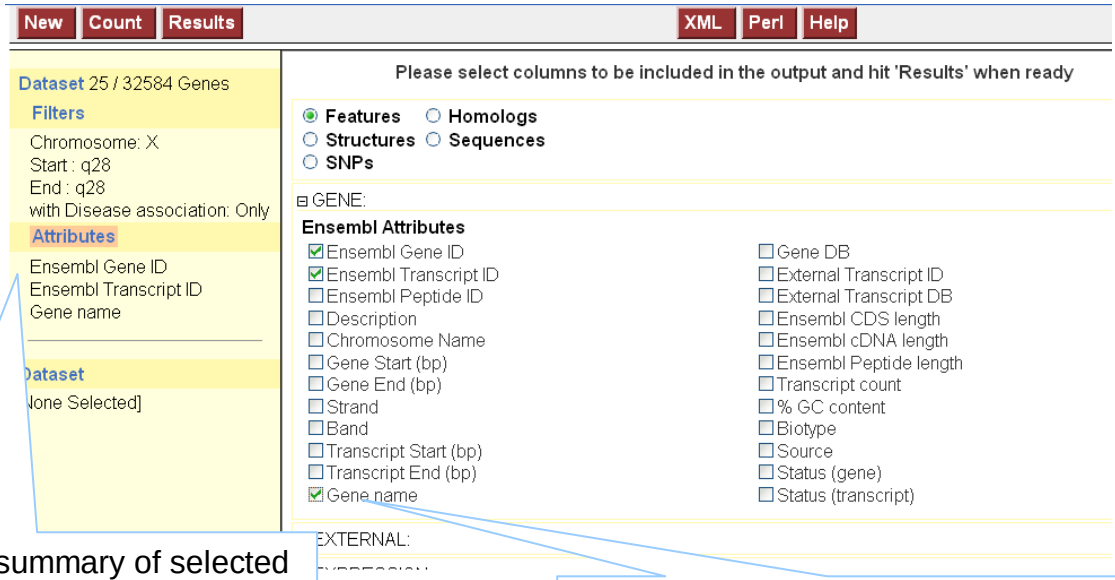
STEP 7:
Expand the 'GENE' panel and choose'Limit to genes '**with MIM disease ID**'.
These associations have been determined using MIM (Online Mendelian Inheritance in Man).
www.ncbi.nlm.nih.gov/omim/

The filters have determined our gene set. Click **'Count'** (at the top) to see how many genes have passed these filters.



STEP 8:
Click on 'Attributes' to select output options (i.e. what we would like to know about our geneset).

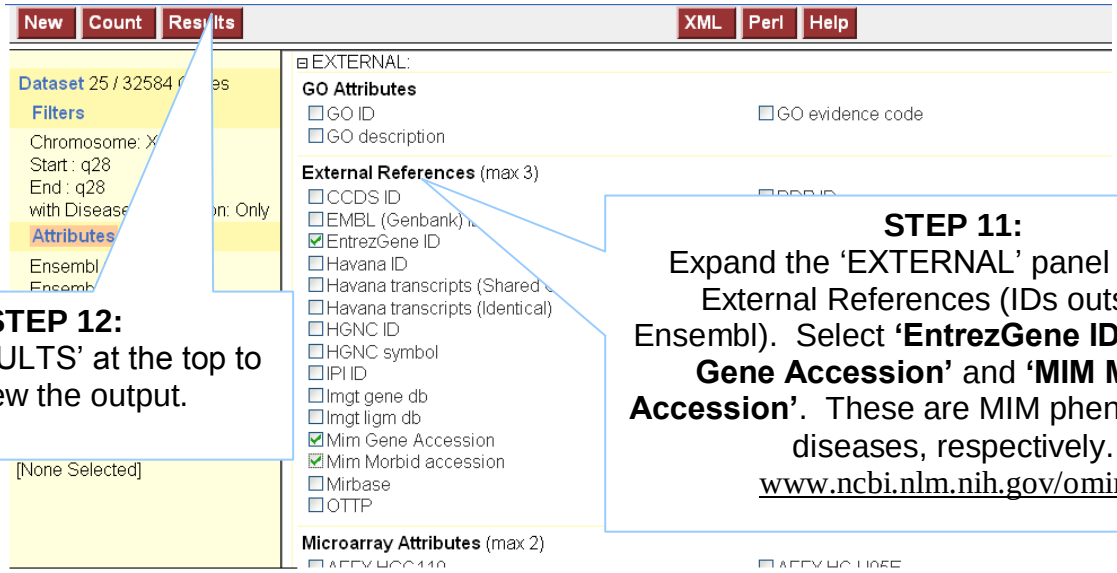
STEP 9:
Expand the 'GENE' panel.



Note the summary of selected options.

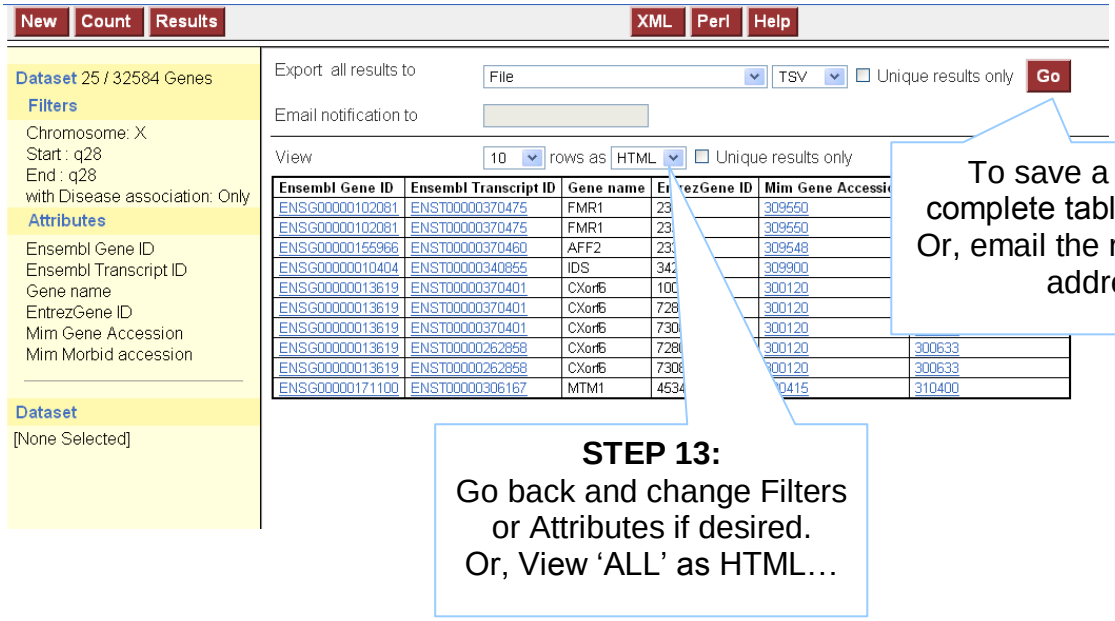
The order of attributes determines the order of columns in the result table.

STEP 10:
Select, along with the default options, '**Associated Gene name**' (this shows the gene symbol from HGNC).



STEP 12:
Click 'RESULTS' at the top to preview the output.

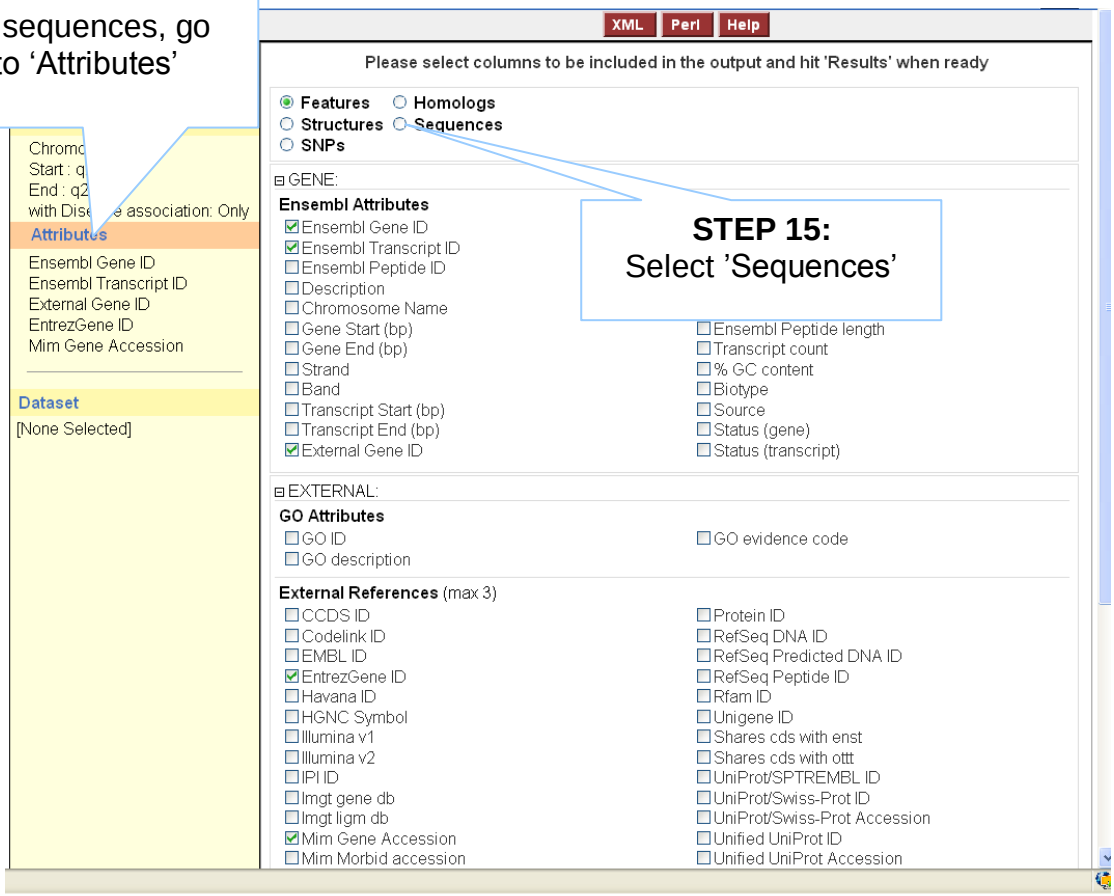
STEP 11:
Expand the 'EXTERNAL' panel to select External References (IDs outside of Ensembl). Select **'EntrezGene ID'** and **'Mim Gene Accession'** and **'MIM Morbid Accession'**. These are MIM phenotypes and diseases, respectively.
www.ncbi.nlm.nih.gov/omim/



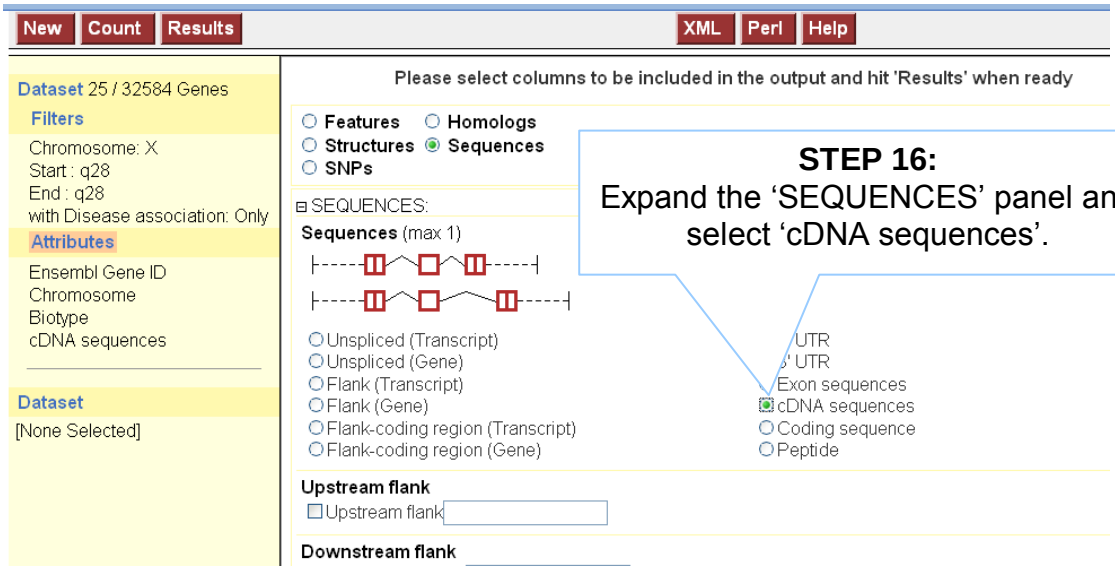
Result Table 1

Ensembl Gene ID	Ensembl Transcript ID	Gene name	EntrezGene ID	Mim Gene Accession	Mim Morbid accession
ENSG00000102081	ENST00000370475	FMR1	2332	309550	300623
ENSG00000102081	ENST00000370475	FMR1	2332	309550	300624
ENSG00000155966	ENST00000370460	AFF2	2334	309548	309548
ENSG00000010404	ENST00000340855	IDS	3423	309900	309900
ENSG00000013619	ENST00000370401	CXorf6	10046	300120	300633
ENSG00000013619	ENST00000370401	CXorf6	728030	300120	300633
ENSG00000013619	ENST00000370401	CXorf6	730818	300120	300633
ENSG00000013619	ENST00000262858	CXorf6	728030	300120	300633
ENSG00000013619	ENST00000262858	CXorf6	730818	300120	300633
ENSG00000171100	ENST00000306167	MTM1	4534	300415	310400
ENSG00000147383	ENST00000370274	NSDHL	50814	300275	308050
ENSG00000130821	ENST00000253122	SLC6A8	6535	300036	300352
ENSG00000130821	ENST00000253122	SLC6A8	731026	300036	300352
ENSG00000185825	ENST00000345046	BCAP31	10134	300398	300475
ENSG00000185825	ENST00000370133	BCAP31	10134	300398	300475
ENSG00000101986	ENST00000218104	ABCD1	215	300371	300100
ENSG00000101986	ENST00000218104	ABCD1	642762	300371	300100
ENSG00000101986	ENST00000218104	ABCD1	215	300371	300475
ENSG00000101986	ENST00000218104	ABCD1	642762	300371	300475
ENSG00000198910	ENST00000370058	L1CAM	3897	308840	142623
ENSG00000198910	ENST00000370058	L1CAM	3897	308840	303350
ENSG00000198910	ENST00000370058	L1CAM	3897	308840	304100
ENSG00000198910	ENST00000370058	L1CAM	3897	308840	307000
ENSG00000198910	ENST00000370060	L1CAM	3897	308840	142623

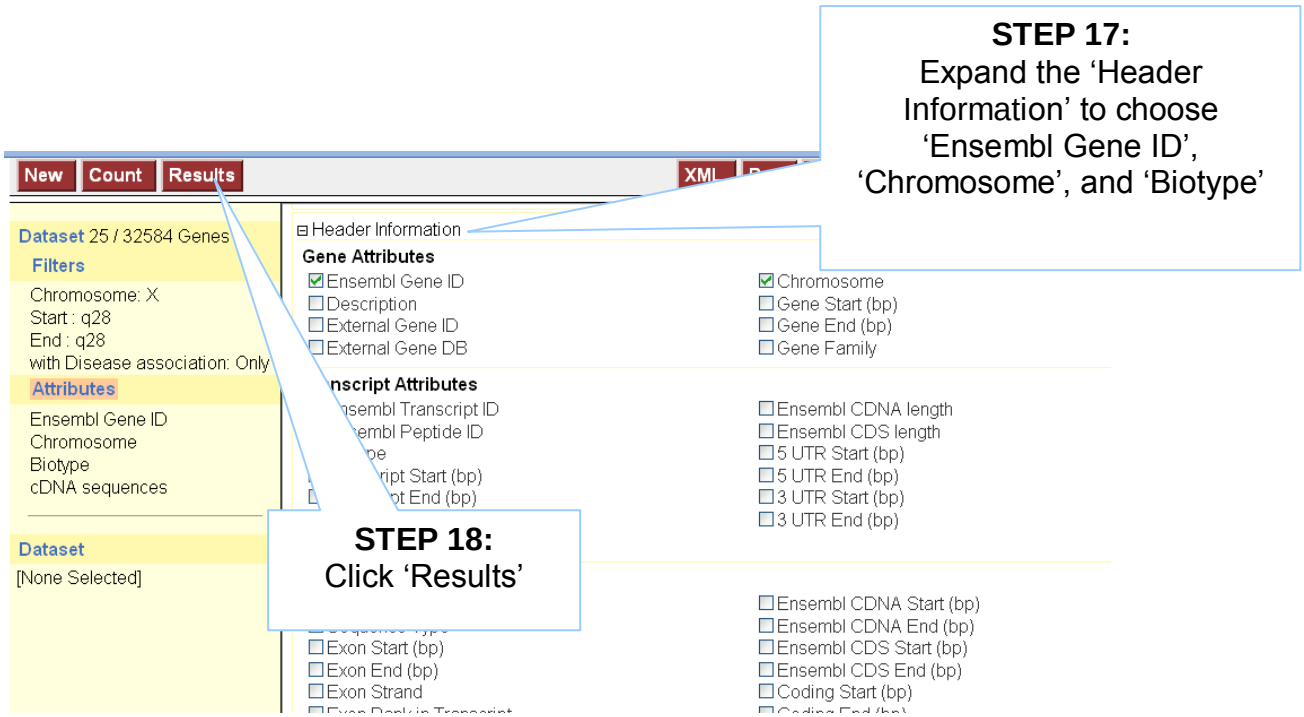
STEP 14:
To view sequences, go back to ‘Attributes’

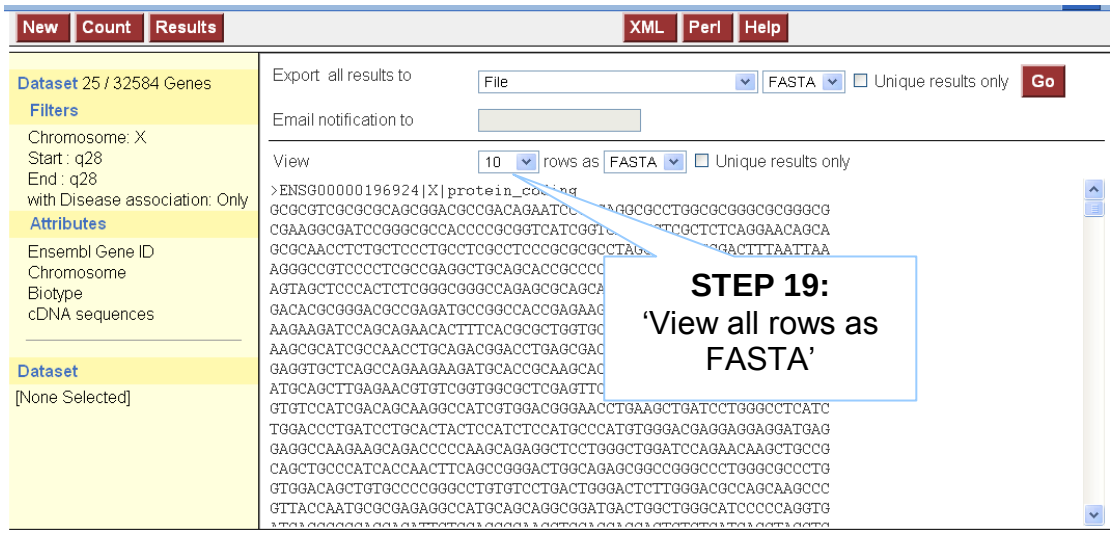


STEP 15:
Select ’Sequences’



STEP 16:
Expand the ‘SEQUENCES’ panel and select ‘cDNA sequences’.


STEP 17:
Expand the 'Header Information' to choose 'Ensembl Gene ID', 'Chromosome', and 'Biotype'
New
Count
Results
XML
Dataset 25 / 32584 Genes
Filters
Chromosome: X
Start : q28
End : q28
with Disease association: Only
Attributes
Ensembl Gene ID
Chromosome
Biotype
cDNA sequences
Dataset
[None Selected]
Header Information
Gene Attributes
Ensembl Gene ID
Description
External Gene ID
External Gene DB
Chromosome
Gene Start (bp)
Gene End (bp)
Gene Family
Ensembl CDNA length
Ensembl CDS length
5 UTR Start (bp)
5 UTR End (bp)
3 UTR Start (bp)
3 UTR End (bp)
Ensembl CDNA Start (bp)
Ensembl CDNA End (bp)
Ensembl CDS Start (bp)
Ensembl CDS End (bp)
Exon Start (bp)
Exon End (bp)
Exon Strand
Coding Start (bp)
STEP 18:
Click 'Results'


New
Count
Results
XML
Perl
Help
Dataset 25 / 32584 Genes
Filters
Chromosome: X
Start : q28
End : q28
with Disease association: Only
Attributes
Ensembl Gene ID
Chromosome
Biotype
cDNA sequences
Dataset
[None Selected]
Export all results to
File
FASTA
Unique results only
Go
Email notification to
View
10
rows as
FASTA
Unique results only
STEP 19:
'View all rows as FASTA'

RESULTS

Header: chromosome, Ensembl Gene ID, Biotype

```
>ENSG00000196924|X|protein_coding
GCGCGTCGCGCGCAGCGGACGCCGACAGAATCCCCGAGGCGCCTGGCGCGGGCGCGGGCG
CGAAGGCGATCCGGGCGCCACCCCGCGGTCATCGGTCACCGGTCGCTCTCAGGAACAGCA
GCGCAACCTCTGCTCCCTGCCTCGCCTCCCGCGCGCCTAGGTGCCTGCGACTTTAATTAA
AGGGCCGTCCCCTCGCCGAGGCTGCAGCACCGCCCCCCCGGCTTCTCGCGCCTCAAAATG
AGTAGCTCCCACTCTCGGGCGGGCCAGAGCGCAGCAGGCGCGGCTCCGGGCGGCGGCGTC
GACACGCGGGACGCCGAGATGCCGGCCACCGAGAAGGACCTGGCGGAGGACGCGCCGTGG
AAGAAGATCCAGCAGAACACTTTCACGCGCTGGTGCAACGAGCACCTGAAGTGCGTGAGC
AAGCGCATCGCCAACCTGCAGACGGACCTGAGCGACGGGCTGCGGCTTATCGCGCTGTTG
GAGGTGCTCAGCCAGAAGAAGATGCACCGCAAGCACAACCAGCGGCCCACTTTCCGCCAA
ATGCAGCTTGAGAACGTGTCGGTGGCGCTCGAGTTCCTGGACCGCGAGAC
GTGTCCATCGACAGCAAGGCCATCGTGGACGGGAACCTGAAGCTGATCCT
TGGACCCTGATCCTGCACTACTCCATCTCCATGCCCATGTGGGACGAGGA
GAGGCCAAGAAGCAGACCCCCAAGCAGAGGCTCCTGGGCTGGATCCAGAACAAGCTGCCG
CAGCTGCCCATCACCAACTTCAGCCGGGACTGGCAGAGCGGCCGGGCCCTGGGCGCCCTG
GTGGACAGCTGTGCCCCGGGCCTGTGTCCTGACTGGGACTCTTGGGACGCCAGCAAGCCC
GTTACCAATGCGCGAGAGGCCATGCAGCAGGCGGATGACTGGCTGGGCATCCCCCAGGTG
...
GACAAGGGGGAGTACACACTGGTGGTCAAATGGGGGGACGAGCACATCCCAGGCAGCCCC
TACCGCGTTGTGGTGCCCTGAGTCTGGGGCCCGTGCCAGCCGGCAGCCCCCAAGCCTGCC
CCGCTACCCAAGCAGCCCCGCCCTCTTCCCCTCAACCCCGGCCCAGGCCGCCCTGGCCGC
CCGCCTGTCACTGCAGCCGCCCCTGCCCTGTGCCGTGCTGCGCTCACCTGCCTCCCCAGC
CAGCCGCTGACCTCTCGGCTTTCACTTGGGCAGAGGGAGCCA
>ENSG00000013619|X|protein_coding
GGCGCGGAGCCGGGCGGTGGGAATGGAGCGAGCAGATTGAGGCCGCCACTGCAGCGCCGC
CAGCATGAACTTGGCCGCGAGCTGAAGCGGCCGGCGGCGGGCGGGCGCGC
CGCTAGCCAGGGGGTGATCTGCAACAGGGCTGGGTTCCTTGGCGGCCGCC
CTGCCCGCGCCGGCGGCCGCCCGCCCCTCGGACACTGCCCCCGCCGCCGC
CAGCACGCCCTGTGTCTAGGTCGTTTGGGAAACGCCTTGGAGAGTCAAGAATAAATTTGC
AGGTCAAACAATGGATGACTGGAAAAGTCGGCTTGTAATCAAGAGCATGCTTCCCCATTT
CGCCATGGTGGGAAATCGTCAGGAGCCCAGAAAGCTCCAGGAATCGGGAAAGAAGCCCTC
GTGGATGGAGGAAGAAGATTTATCTTTTCTCTACAAGAGCAGCCCAGGAAGAAAGCATCA
GGGAACTGTTAAGAGGAGACAAGAAGAAGACCACTTCCAGTTTCCAGACATGGCTGATGG
GGGCTACCCTAATAAAATTAAGAGGCCTTGCCTTGAAGATGTCACCCTTGCAATGGGCCC
AGGTGCTCATCCTAGTACTGCTTGTGCAGAACTGCAGGTCCCTCCATTGACAATAAATCC
TAGCCCTGCGGCTATGGGAGTGGCTGGCCAGTCATTACTGCTGGAGAATAACCCTATGAA
TGGCAACATCATGGGCTCACCATTTGTAGTACCACAGACTACAGAAGTGGGACTGAAAGG
GCCCACTGTTCCTTACTATGAGAAAATCAACAGCGTGCCGGCTGTAGACCAGGAGCTTCA
AGAGCTGCTAGAGGAGCTCACCAAAATTCAAGACCCTTCTCCAAATGAGCTAGATCTTGA
GAAGATACTGGGGACGAAGCCAGAAGAGCCACTGGTTTTAGATCATCCCCAGGCAACCCT
AAGCACAACTCCCAAGCCTTCGGTTCAGATGTCACACTTGGAGAGCCTGGCTTCCAGCAA
GGAGTTTGCTTCTAGTTGCAGCCAAGTTACTGGCATGTCACTTCAGATCCCATCCTCCTC
CACAGGGATCAGCTATTCGATTCCTTCCACCAGTAAGCAGATAGTGTCACCGAGTTCTTC
AATGGCACAGTCCAAGAGCCAGGTCCAGGCCATGCTCCCTGTCGCTCTGCCCCCCTTACC
AGTGCCTCAGTGGCATCACGCCCACCAGCTGAAGGCGTTGGCAGCCAGCAAGCAGGGGTC
TGCTACAAAGCAGCAAGGGCCCACCCCCAGTTGGTCTGGTCTGCCTCCTCCAGGACTCTC
TCCACCTTACCGCCCAGTGCCATCACCACACCCACCACCGCTGCCACTGCCACCACCACC
```

cDNA 1

cDNA 2

V) **BIOMART**

These exercises have been designed to familiarise you with different questions you can answer with this tool, and the types of data you can retrieve with BioMart.

1. Retrieve all SNPs for 'novel' human G-protein coupled receptor genes (GPCRs – use the InterPro domain ID: 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!)

Start a new BioMart session by clicking 'New', or go back to the Ensembl homepage and click on 'Mine Ensembl with Biomart' under 'Ensembl tools'.

Choose the **database** and the **dataset** for your query as follows:
- Select 'Ensembl 50'
- 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 on *Homo sapiens* chromosome 2.
- In the 'GENE' section, select 'Status (gene)' '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 now 3.

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 Protein.
- In the 'GENE ASSOCIATED SNPs' section 'Reference ID' is selected. Also select 'Allele', 'Protein location (aa)', 'Location in Gene' and 'Protein Allele'.

Note: Clicking on count now will not show an altered number. Attribute selections should not affect the count (i.e. the number of genes that have passed the filters).

Click on '**Results**' (at the top) to obtain the first 10 rows of your table. To obtain the entire table select 'View all rows as HTML' or export a file by clicking 'Go'. Check the box 'Unique results only', otherwise you can end up with redundant rows!!

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.

2. Click 'New' to start a new query. Retrieve the gene structure (i.e. start and end coordinates of exons) of the mouse gene ENSMUSG00000042351.

3. Retrieve all human genes that are located between p11.2 and q22 (these are bands on chromosome 1) and that have a MIM disease ID.

4. The file at 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). Enter the probe list at this url into BioMart to retrieve the 500 bp upstream of the transcripts matching these probeset IDs.

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

6. Retrieve sequence (including reference ID in the header) of all human SNPs on chromosome 6, band p25.3, that have an ID from Watson's genome. Export both the SNP sequence and 200 bases flanking (adjacent to) it.

7. Retrieve the mouse homologues of *Homo sapiens* genes CASP1, CASP2, CASP3, and CASP4 (these are the HGNC symbols for the genes).

8. Design your own query!

Answers (BioMart)

1. You should find ***three*** novel genes on chromosome 2 with this InterPro domain. The result set has three transcripts (one for each gene) and a total of 375 rows of output (to see this, change the option from TSV to XLS under 'Export all results', select the option 'Unique results only' and click 'Go', then open in Excel so you don't have to count the rows manually). The three genes/transcripts have two, eight and one coding SNPs ('Location in Gene' is 'coding'), respectively. Most of these are non-synonymous and thus affect the amino acid sequence of the encoded protein. One allele is a stop codon (*)- can you find it?

2. Click New. Select: Database and dataset: 'Ensembl 50' and '*Mus musculus* genes (NCBIM37)'.

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***. Follow 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.

3. Database and dataset: 'Ensembl 50' and '*Homo sapiens* genes (NCBI36)'.

Filters: REGION: 'Chromosome 1', 'Band Start p11.2, 'Band End q22', **GENE**: 'Limit to genes … with MIM disease ID(s) Only'.

Attributes: Features: **EXTERNAL:** select 'MIM Morbid Description' along with the default options ('Ensembl Gene ID' and 'Transcript ID').

Results should show ***18 Ensembl genes*** (some of which with multiple transcripts).

4. Database and dataset: 'Ensembl 50' 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: SEQUENCES: select 'Flank (Transcript)', 'Upstream flank 500'. In 'Header Information', apart from the already default selected options, select 'Ensembl Transcript ID'.

You should find upstream sequences for the transcripts of ***26 genes*** (Hint: click 'count' to see the number of genes!)

5. Database and dataset: 'Ensembl 50' and '*Homo sapiens* genes (NCBI36)'.

Filters: REGION: 'Marker' Start D1S2806 End D1S464, **GENE:** 'Status: KNOWN'.

Attributes: Sequences: select, apart from the already default selected options, 'Flank (Gene)' and 'Upstream flank 5000'.

You should find sequences for ***41 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)

6. Database: 'SNP' and **dataset:** '*Homo sapiens* SNPs

(dbSNP127;HGVbase 15; TSC 1; ENSEMBL, Affy 100K and 500K arrays)

Filters: REGION: 'Chromosome 6', 'Band p25.3', **GENERAL SNP FILTERS:** SNP source: 'with Variation synonym ensemblwatson ID(s)' Only.

Attributes Sequences: SEQUENCES: 'SNP sequences', 'Upstream flank 200', 'Downstream flank 200'. **SNP: SNP attributes**: default is 'Reference ID'.

You should find ***2473 SNPs***.

7. Database: 'Ensembl 50' **Dataset:** *Homo sapiens* genes (NCBI36)

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

Attributes: Under **Homologs**, select in the **MOUSE ORTHOLOGS** panel 'Mouse Ensembl Gene ID'. Also, scroll back up to the **GENE** panel. Along with the default option 'Ensembl gene ID', select 'Associated Gene Name' and deselect 'Ensembl Transcript ID' (these attributes are for the starting dataset… i.e. Human).

Results displays the mouse orthologues of the 4 human CASP genes.

VI) EVALUATING GENES AND TRANSCRIPTS

(The 'GeneBuild')

Main Exercise: Examine the evidence for the FOXH1 gene.

These exercises focus on the FOXH1 gene to demonstrate how the underlying protein and mRNA used to build an Ensembl gene can be seen. Is it a well-determined gene…?

1. Display the human FOXH1 gene in GeneView.

Enter FOXH1 into the text search box at the top of any human Ensembl page. Take the link to the **GeneView** page for the Ensembl Gene: ENSG00000160973.

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

2. Examine the supporting evidence for the FOXH1 gene in ExonView.

Scroll down the **GeneView** page and have a look at the predicted exon structures. Note the 5' and 3'UTRs (untranslated regions). Compare the two transcripts using this view.

Click on 'Exon information' in the left-hand menu to go to **ExonView** for one of the transcripts. 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?

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

*Optional: Click on the 'Gene information' link in the left-hand menu to return to the **GeneView** 'Gene Report' for FOXH1.*

3. Examine other evidence for the FOXH1 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 other protein/mRNA tracks, e.g. Human proteins, Unigene (for all species), Human cDNAs, and EMBL mRNAs (again across species). These are proteins/mRNAs that align to the genome in this area. Note that the track labels are links to the help page. Clicking on a block drawn in the new tracks brings up a pop-up menu with a link to the database entry. Try some links.

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

Under 'Decorations' select 'Show empty tracks'. This will help you remember which tracks you have selected. Examine the evidence for the FOXH1 gene in the new evidence tracks. Are there proteins and mRNAs that align to the Ensembl prediction?

4. Compare transcript predictions made by other methods.

In the 'Features' menu, turn off most of the evidence and make sure Ensembl genes, Vega Havana 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.

Q4: What is a Vega Havana transcript?

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 FOXH1 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 such as UniProt or NCBI EntrezGene entries *for the same species*. (Note that this mapping is done *after* the genes have been built). Novel transcripts may match to mRNA and protein information in other databases, for an alternate species. Thus a novel transcript is novel for that species.

Matches to IDs in other databases are shown 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.

Q5: What do the Target and Query % ids indicate? Check the online Help pages.

Answers (Evaluating Genes and Transcripts).

A1: Name: HUGO Gene Nomenclature Committee (HGNC). Description: Uniprot/Swiss-Prot.

A2: 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. Click on a green box for the alignments with the transcript.

A3: The Genscan transcript prediction shows exons not present in the Ensembl transcripts for FOXH1 and doesn't predict UTRs. Genscan is an *ab initio* predictor, a program run on the sequence alone, without using protein and mRNA evidence. It has the tendency to overpredict exons.

A4: Havana is a subgroup of VEGA, the Vertebrate Genome Annotation consortium. Havana transcripts are manually curated (determined on a case-by-case basis) and are merged with Ensembl predictions if they match up exactly (leading to golden transcripts). For more, go to these links:

http://www.sanger.ac.uk/HGP/havana/
http://vega.sanger.ac.uk/index.html

A5: 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?

VII) COMPARATIVE GENOMICS

1. Investigating a protein family.

This exercise focuses on protein families and orthologies, using AIM1, a cancer-related gene, as an example.

Find the **GeneView** page for human AIM1, a gene absent in a type of melanoma. Beware… there is a VEGA gene as well as an Ensembl gene! Let's work with the Ensembl gene for now.

Q1: What is the MIM ID for the AIM1 transcript? Hint… find this in 'Similarity Matches' for one of the transcripts. Click on the MIM ID…

Examine the protein family by taking the link to the associated Protein Family (the **FamilyView** page).

Q2: How many human Ensembl genes produce peptides in this family?

Q3: Are they all 'known' genes?

Q4: Are there peptides in the same family for mouse (*Mus musculus*), rat (*Rattus norvegicus*) and zebrafish (*Danio rerio*)?

Q5: Families are calculated using Ensembl peptides along with UniProt peptides. Can you find on this page where the UniProt peptides in this family are listed?

2. Genomic alignments and conserved regions

Find the **ContigView** page for human SNX5**.**

Q6: In which band is the human SNX5 gene located?

The 'constrained elements' track shows the most highly conserved regions calculated from the 10-way multi-species alignment.

Q7: Do any of these regions fall outside of exons?

Expand the 'Conservation' track by clicking on the '+' in front of it. This shows a score for each basepair based on how conserved it is across the alignments. Clusters of peaks correspond to the 'Constrained elements blocks' shown above.

Collapse the track if desired. Choose the pairwise alignment with mouse from the Comparative menu. Selecting the '+' in front of the track will expand the track so that all alignments on the same mouse chromosome and strand will be clustered together.

Q8: On which chromosome(s) and strand(s) are the mouse alignments for this region?

Turn on the frog (*X. tropicalis*) pairwise alignment with human, using the Comparative roll-down menu.

Q9: Are there fewer or more conserved regions between human and frog when compared with human and mouse? Would you expect an SNX5 frog gene in this region?

Click on a frog alignment. 'Jump to Xenopus tropicalis' to view ContigView for frog. Can you see the snx5 gene there?

3. Examine the syntenic regions

Syntenic regions in Ensembl are calculated from the pairwise alignments (whole genome alignments between two species) and correspond to long regions (100 kb or more) that have high sequence similarity. Gene order is expected to be conserved in these regions. Two pages are available to view these regions: **SyntenyView** and **CytoView**.

To go to SyntenyView, where two species may be compared, start by finding the **ContigView** page for human Myosin VI.

Click on 'View syntenic regions with… *Mus musculus*', a link at the left hand side of the page.

Q10: How many mouse chromosomes show syntenic regions with human chromosome 6?

Q11: Is there a mouse homologue to human Myosin VI? Can you see this without leaving the **SyntenyView** page?

Gene order is conserved in human and mouse for this syntenic region. Comparing homologous genes across both species can show where gene loss may have occurred.

Click on '**MultiContigView**' next to 'Myo6' in mouse to reach a comparison of the human and mouse chromosomes in this region.

Q12: What does the vertical blue line connecting the lowest panels show?

Q13: The Myosin VI genes are golden in human and mouse. What does the colour signify?

Syntenic regions may also be viewed in the **CytoView** page. Take the 'Graphical overview' link at the left to go to **CytoView**. Turn on syntenic regions with mouse and opossum using the 'Comparative' roll-down menu.

Click on the red and green lines in the 'Mouse' and 'Opossum' tracks to see which base pairs and chromosome strand are syntenic to human in this region.

The 'View alongside…' link at the left will take you back to **MultiContigView**. Instead, take the 'View alignment with…' link and choose '*Mus musculus*'. This will bring you to the '**AlignSliceView**' page, where again, human and mouse chromosomes can be compared side-by-side. However, as **MultiContigView** shows the chromosomes as they are, **AlignSliceView** shows the alignment, and the genomic assembly may be altered (gaps introduced) to fit the alignment in this view. Read the 'Help' page to understand the view.

Select the '12 amniota vertebrates' in the 'Comparative' roll-down menu.

Q14: Is there a gene in this aligned region across the twelve species?

Answers (Comparative Genomics)

Main Exercise (1)

A1: Find this in the GeneView page. Make sure the ID at the top is ENSG00000112297. The MIM ID (corresponding to the entry on the 'Online Mendelian Inheritance in Man database) is 601797, found in the 'Similarity Matches' section of the **GeneView** page. (See the report for the second, AIM1, transcript).

A2: Family IDs change with every release… in 50, it is fam50v00000001399. 3 human genes..

A3: They are all "Known" protein-encoding Ensembl genes (all genes have a name; none are 'novel').

A4: Yes. Scroll down the 'FamilyView' page to find the peptide lists for other species.

A5: In the 'Other peptides…' section, just under the gene names and karyotype, the UniProt identifiers clustered into this family are listed. 4 mouse, 4 rat, and 4 zfish proteins.

A6: Band p11.23, chromosome 20 *(hint: go to **ContigView**)*

A7: Yes. This track is turned on by default- if not, switch it on in the 'Comparative' roll-down menu under 'Constrained elements'. Looking at the visual display in Detailed view, most of the conserved blocks line up with exons (if you are unsure, zoom in.) Clicking on a conserved element shows its precise location across the genomes aligned. There are some conserved

elements near the 3' end of the gene that fall outside of exons (especially in the long intron of the SNX5 transcript). Could these be important regulatory regions? *Hint… turn on ncRNA (non-coding RNAs) in the Features menu.*

A8: All are on chromosome 2, forward strand. In the expanded format, i.e. a (-) is found next to the track, all alignments appear grouped together on one line, indicating they are on the same chromosome and strand. Click on a block to find this is chromosome 2, forward strand.

A9: Fewer. You would expect a gene, because the conserved regions with frog line up to the human SNX5 exons, supporting a hypothesis in which those exons have been conserved across time in frog as well as in human. Thus, they are probably part of an expressed protein.

A10: 7 chromosomes in mouse are shown in **SyntenyView**. These are the chromosomes that share syntenic blocks with human chromosome 6.

A11: The homologues are listed at the right of the page. Mouse Myo6 is a homologue of human Myo6. It is on chromosome 9 at position 80.01 Mb in mouse.)

A12: The vertical blue line in the lowest, 'Detailed View' panel connects homologous genes across, in this case, human and mouse. Remember homologues are determined using the longest translation of a gene only.

A13: A golden gene has at least one transcript that is identical to (and merged with) a transcript determined by manual curation (the Havana team).

A14: Yes, a gene is present in these aligned regions for all species. Some exons align well, indicating they could be homologues.

VIII) VARIATIONS

These exercises focus on SNPs (Single Nucleotide Polymorphisms) in Ensembl. Most of these are downloaded from dbSNP and placed along the sequence assembly using flanking sequence to the SNP to match it.

1. Display all SNPs for a region.

*This exercises focuses on SNPs in the **ContigView** page.*

Go to the **ContigView** page displaying Human chromosome 7: bp 116124438-116129149. Turn on SNPs.

Q1: Are there non-synonymous coding SNPs in the gene? *(Remember, non-synonymous SNPs change the amino acid sequence).*

You can see information about a SNP by clicking on it. Click on a non-synonymous SNP and follow the link 'SNP properties' to its **SNPView** page.

In the **SNPView** page, click on ' SNPs in gene context' at the right of the second panel. You will be taken to the **GeneSNPView** page, where you can see all the SNPs for the MET gene. Under 'SNP type' roll-down menu in the **GeneSNPView** page, select only 'non-synonymous'.

2. Display SNPs linked to a disease within the transcript sequence.

Starting with disease-causing SNPs, how can we view them within the sequence? Hint… focus on the TransView page.

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 SNPs leading to peptide alleles M470V and T854T are studied.

Q2: Find these two SNPs in the **TransView** and **GeneSNPView** pages. Note, the positions refer to numbers in the amino acid sequence.

3. Display all SNPs for one gene.

Retrieve all the validated SNPs associated with the human CFTR gene as follows. Display only validated SNPs in the '**GeneSNPView**' page used in question 2.

Now use BioMart to ask how many of these validated SNPs are coding.

Q3: How many validated SNPs are in the coding region of CFTR?

Answers (Variations)

A1: 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**. 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 vertical lines of different colours along the chromosome. The SNP legend is shown at the bottom of the panel. This region contains one exon from the 'MET' gene. There are a few non-synonymous coding SNPs (shown in yellow) in this region.

A2: 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. At the bottom of the transcript sequence are panels for customisation of the display. 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. M470V is shown as a red 'V' at position 470 in the peptide sequence (the second line of numbering in the sequence).

Continue on to the **GeneSNPView** page for this gene by clicking 'Gene variation info.' in the side menu.

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 (as it is a non-synonymous coding SNPs). T854T is shown as 'T' in green, as it is a synonymous coding SNP. The diagram only shows bp positions, so it could be easier to find the two allelic changes in the table. Note that you can use the 'SNP class' and 'SNP type' drop-down menus in the figure to configure (simplify) both figure and table by choosing only to see coding SNPs.

A3: Validated SNPs have been checked after submission to determine if they are SNPs or mutations. The validation status comes from NCBI, and explanation to the validation terms can be found in the 'Help' to SNP-based pages.

To view only validated SNPs, configure the 'Validation' roll-down menu in the **GeneSNPView** page by deselecting 'no information' and closing the menu. Now only validated SNPs are shown. How many are coding? Deselect, under the 'SNP type' roll-down menu in this page, upstream, downstream, UTR, intronic and intergenic SNPs. Close the menu. The display and table will be simplified. Count **14** coding SNPs that have been validated.

Go to BioMart. Select the SNP database and *Homo sapiens* SNPs. In **Filters**, expand the 'GENERAL SNP FILTERS' section. Select 'SNPs that have been validated' (Only). In the 'GENE ASSOCIATED SNP FILTERS' section, enter the Ensembl Gene ID ENSG00000001626. Select 'Ensembl Gene Location' as 'Coding Only'.

'Count' gives you **14** validated, coding SNPs.

IX) TYING IT TOGETHER – CASE STUDIES (using Ensembl comprehensively)

Consider the following case studies. The answers to these questions tie together pages and tools on the Ensembl site.

1) I work on non-coding RNA genes such as snoRNA (small nucleolar RNA) and miRNA. Are there any on mouse chromosome 1 (of any type), and where are they? How many are there (use BioMart). Could I view them graphically?

2) Many regulatory factors appear in the 5'UTR, and I am interested in searching for them using motif-scanning tools. How would I obtain the 5'UTR or upstream region in the IRX4 gene in human? What can I learn about the UTR and upstream regions using **ContigView**?

3) *(Using DAS):* We have determined two SNPs in the lab not found in Ensembl. We have not yet submitted to dbSNP, but can we just display these SNPs on our computer in Ensembl? Here is some information about the SNPs:

Names: 1 and 2
Positions: both on chromosome 3 (mouse)
base pairs: 75305500 and 75450001
Type and subtype are both 'a' (internal lab naming convention)
Phase is 0, Score is 100.

Hint: start from ContigView for mouse chromosome 3: 75300000 - 75500000

4) I would like to export the gene structure of zfish 'pp2ca2' in GFF format (I will use a program that requires data in this format).
(Find out more about the GFF format here)
http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml

Case Studies: Answers

1) Using BioMart to find all the RNA genes will take 2 queries as the filters only allow a subset of ncRNA types to be selected. There are **192** total RNA genes on chromosome 1. You need only select the filers, then click 'count' to see how many genes there are.

From the result table, follow the link to one of these genes into the Ensembl browser (for example, ENSMUSG00000070103). Alternatively, go to '**ContigView**' for a region on chromosome 1and select, in the 'features' menu, 'ncRNA' to visualise the Ensembl ncRNAs in the chromosome. Also, select 'rFAM' to select RNAs predicted by this program. You can also select tRNA, miRNA, and other options.

2) **Step 1: Exporting UTR sequence.** Export the UTR information by clicking 'Export gene data' at the left of the **GeneView**, **TransView**, or other page and then selecting '5' or 3' UTR' in the next window to export the sequence.

Alternatively, use **BioMart** to export the 5' and 3' UTRs.

Step 2: Visual display and comparison of promoter elements. ***What can you learn about the upstream regions of this gene in ContigView? Use the features menu and conserved tracks.***

In **ContigView** (the 'graphical view' link) select 'CTCF' for CTCF binding sites (click the blue 'Help' if you are unsure of what this is), CpG islands, and motifs in the CisRED/miRANDA database' for sequences associated with regulatory regions. For genes that have no UTR annotated, it can be helpful to select ''Eponine' (a program that predicts transcription start sites) under the 'Features' menu. Also, in the Comparative menu, select alignments such as '10 amniota vertebrates'. Constrained elements show the highest sequence similarity for the alignment.

3) ***HINT: Upload as a DAS source from ContigView.***
Go to **ContigView** encompassing this region in mouse, then select **'DAS Sources'** from the roll-down menu above the detailed-view panel and **'Manage Sources'**. Next, click on '**Upload your data**' in the new window (at the left). Read the instructions by clicking on the link at the top of the page (especially the 'formatted correctly' link.) Enter in your email and a NON-SECURE password, then paste your data according to the format described in the link. (Note: columns must be separated by tabs, NOT SPACES! This will require making the data columns in another program such as Notepad).

Once the data is entered, click 'Next' and select **ContigView** (note you can select more than one page). Click 'Next' again, name the track if you'd like, and click 'Finish'. Once you reach the DAS sources list (end page) 'Close window' to refresh **ContigView**… a new track should be displayed, along with SNPs 1 and 2!

Here is the correct format for the SNP information given in the question:
http://www.ebi.ac.uk/~gspudich/workshop_presentations/snp_example.txt

Note: you can deselect your track under the 'DAS sources' menu of ContigView.

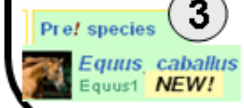
4) To export in GFF format: Select 'export gene data' at the left of the **GeneView** page and select GFF under 'Output format'. Or, export the chromosome and base pair start and stop of the exons using the 'structures' attribute page in **BioMart**. Select GFF format under 'Display rows as GFF' and click 'Go' to export the file.

Ensembl Did you know?

Focus

1 **Genome sequence assemblies** determined by sequencing institutes are incorporated into **Ensembl**. View the assembly used for a species on the main page.

2 **The Ensembl gene set** is based on mRNA and protein evidence and is aligned to the genomic sequence assembly in the 'genebuild'. Use **'ExonView'** to view this information.



3 **Ensembl focuses on vertebrates** however other model organisms are included such as yeast, fly, and soil worm. View newly available genomic sequences in the ***Pre!*** site.

Browsing

4 **A new version** of **Ensembl** is released every two months. View older versions in the ***Archive!*** site. New gene sets are determined when a new sequence assembly is available.

ContigView and **GeneView** can be customised using the 'features' menu. Select options to visualise transcripts, variations, regulatory regions, etc. Use the DAS menu to show data in current databases external to **Ensembl**.

5

*View a short video about **ContigView** here!*

http://www.ensembl.org/common/Workshops_Online

Tools

6 Export SNP info in region

Use BioMart to get tables of genes and annotation in Microsoft Excel, HTML, or txt format. You can also export sequences with this Web-based tool!

www.biomart.org

7 **BLAST and SSAHA** are alignment programs- use them to match a sequence on any genome.

8 **View variations** for a gene or sequence in **GeneSNPView** or **ContigView.**

We calculate phylogenetic gene trees, alignments, and homologies. **9**

HELP **10** Send any questions or comments to

helpdesk@ensembl.org