Access to genes and genomes with Ensembl

Introduction and Worked Example

Dec 2007

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

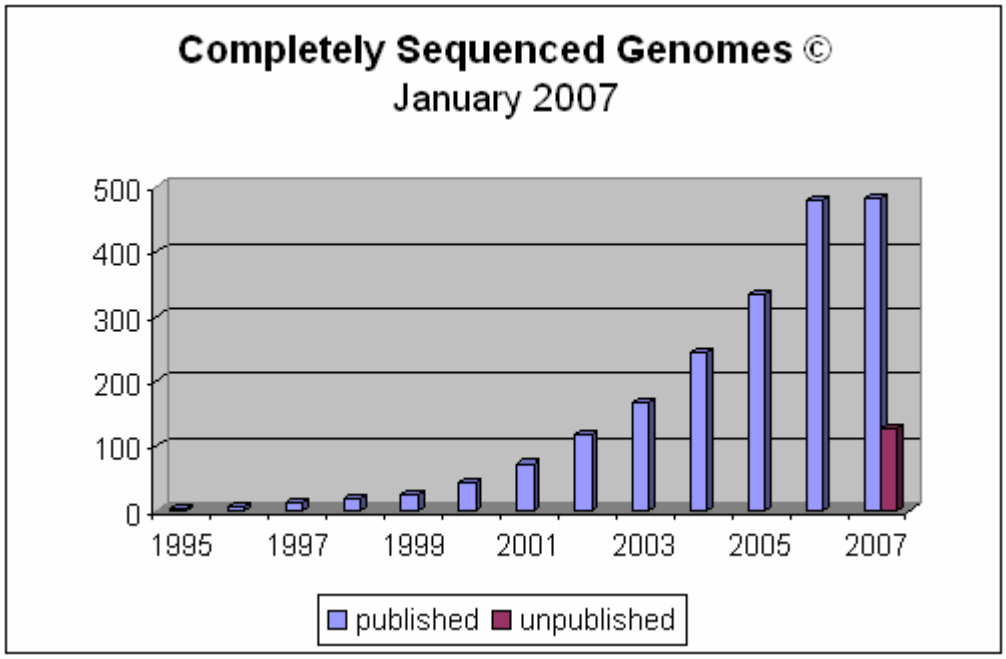


Figure 1. Completely sequenced genomes as of Jan, 2007 (figure taken from http://www.genomesonline.org).

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

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		ASSEMBLY		GENEBUILD	
Mammals					
Human	*Homo sapiens*	NCBI 36	Oct 2005	Ensembl	Sept 2007
Chimpanzee	*Pan troglodytes*	PanTro 2.1	Mar 2006	Ensembl	May 2007
Rhesus macaque	*Macaca mulatta*	MMUL 1	Feb 2006	Ensembl	Aug 2006
Bushbaby	*Otolemur garnettii*	otoGar1	May 2006	Ensembl	Feb 2007
Mouse	*Mus musculus*	NCBI m37	Apr 2007	Ensembl	Sept 2007
Rat	*Rattus norvegicus*	RGSC 3.4	Dec 2004	Ensembl	Feb 2006
Rabbit	*Oryctolagus cuniculus*	RABBIT	May 2005	Ensembl	Aug 2006
Squirrel	*Spermophilus tridecemlineatus*	speTri1	Jun 2006	Ensembl	Oct 2006
Dog	*Canis familaiaris*	CanFam 2.0	May 2006	Ensembl	Dec 2006
Cat	*Felis catus*	CAT	Feb 2007	Ensembl	Jun 2006
Cow	*Bos taurus*	Btau 3.1	Feb 2007	Ensembl	Sep 2006
Pig*	*Sus scrofa*	Sscrofa1			
Shrew	*Sorex araneus*	sorAra1	Oct 2005	Ensembl	Apr 2007
Hedgehog	*Erinaceus europaeus*	eriEur1	Feb 2007	Ensembl	Oct 2006
Guinea pig	*Cavia porcellus*	cavPor2	Feb 2007	Ensembl	Oct 2006
Horse*	*Equus caballus*	Equus1			
Microbat	*Myotis lugigfugus*	myoLuc1	Mar 2006	Ensembl	Jan 2007
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
Tree shrew	*Tupaia belangeri*	tupBel1	Feb 2007	Ensembl	Oct 2006
Opossum	*Monodelphis domestica*	MonDom 5.0	Oct 2006	Ensembl	Feb 2007
Platypus	*Ornithorhynchus anatinus*	OANA 5	Dec 2005	Ensembl	Jan 2007
Other species					
Chicken	*Gallus gallus*	WASHUC 2	May 2006	Ensembl	Aug 2006
X. tropicalis	*Xenopus tropicalis*	JGI 4.1	Aug 2005	Ensembl	Nov 2005
Zebrafish	*Danio rerio*	Zv 7	Apr 2007	Ensembl	Jun 2007
Takiugu	*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	Jun 2006
Medaka	*Oryzias latipes*	HdrR 1	Oct 2005	Ensembl	May 2006
C. intestinalis	*Ciona intestinalis*	JGI 2	Mar 2005	Ensembl	Feb 2006
C. savignyi	*Ciona savignyi*	CSAV 2.0	Oct 2005	Ensembl	Apr 2006
Fruitfly	*Drosophila melanogaster*	BDGP 4.3	Jul 2005	FlyBase	Mar 2006
Anopheles	*Anopheles gambiae*	AgamP 3	Feb 2006	VectorBase	Jun 2007
Aedes	*Aedes aegypti*	AaegL 1	Oct 2005	VectorBase	Jun 2006
C. elegans	*Caenorhabditis elegans*	WS 180	Sep 2007	WormBase	Sep 2007
S. cerevisiae	*Saccharomyces cerevisiae*	SGD 1.01	Dec 2006	SGD	Dec 2006

Table 1 – 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

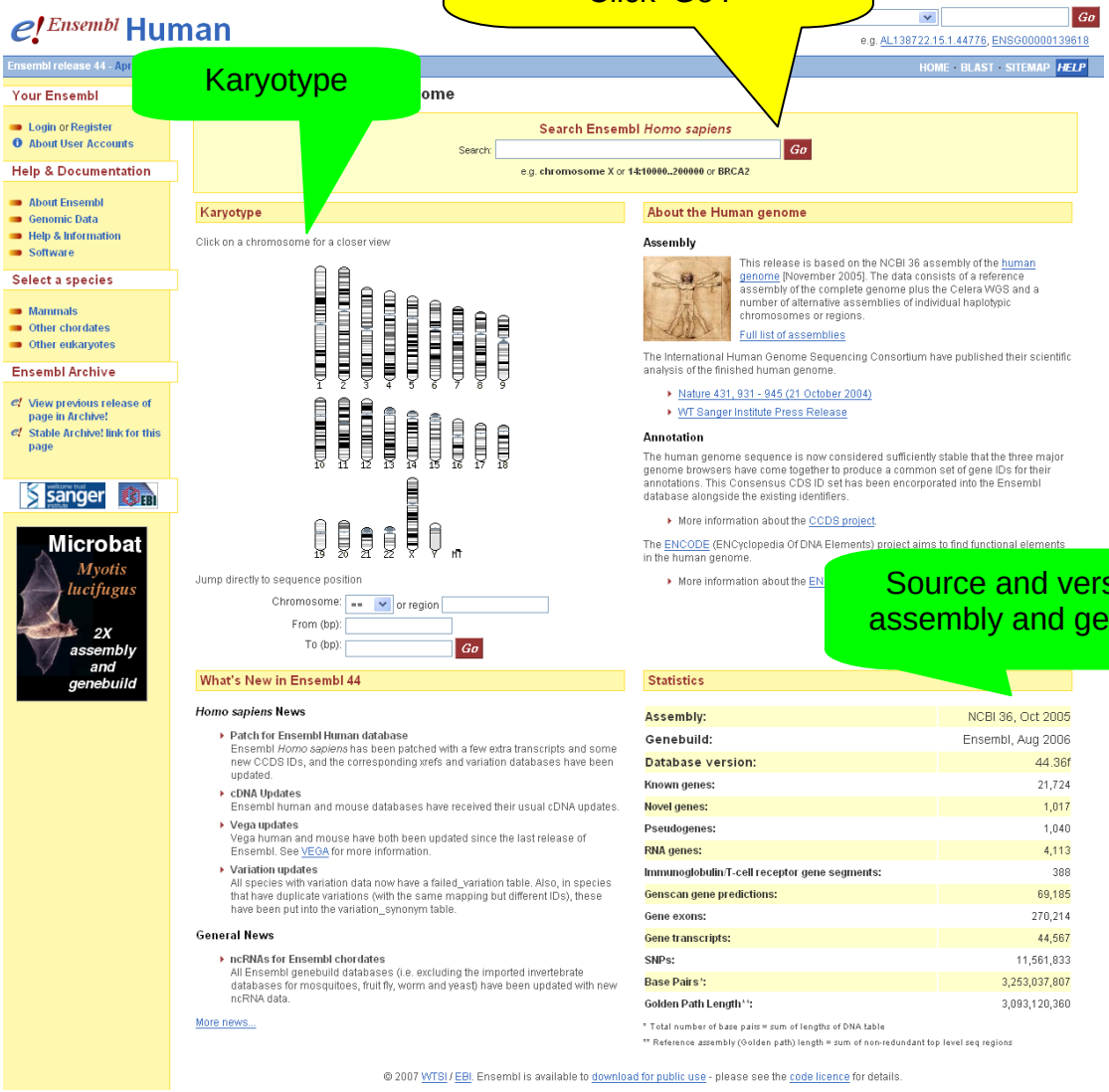
II) WORKED EXAMPLE – A walk through the main pages of the Ensembl browser, using the BRCA2 (Breast cancer 2, early onset) gene as an example.

STEP 1:
Load Ensembl
www.ensembl.org

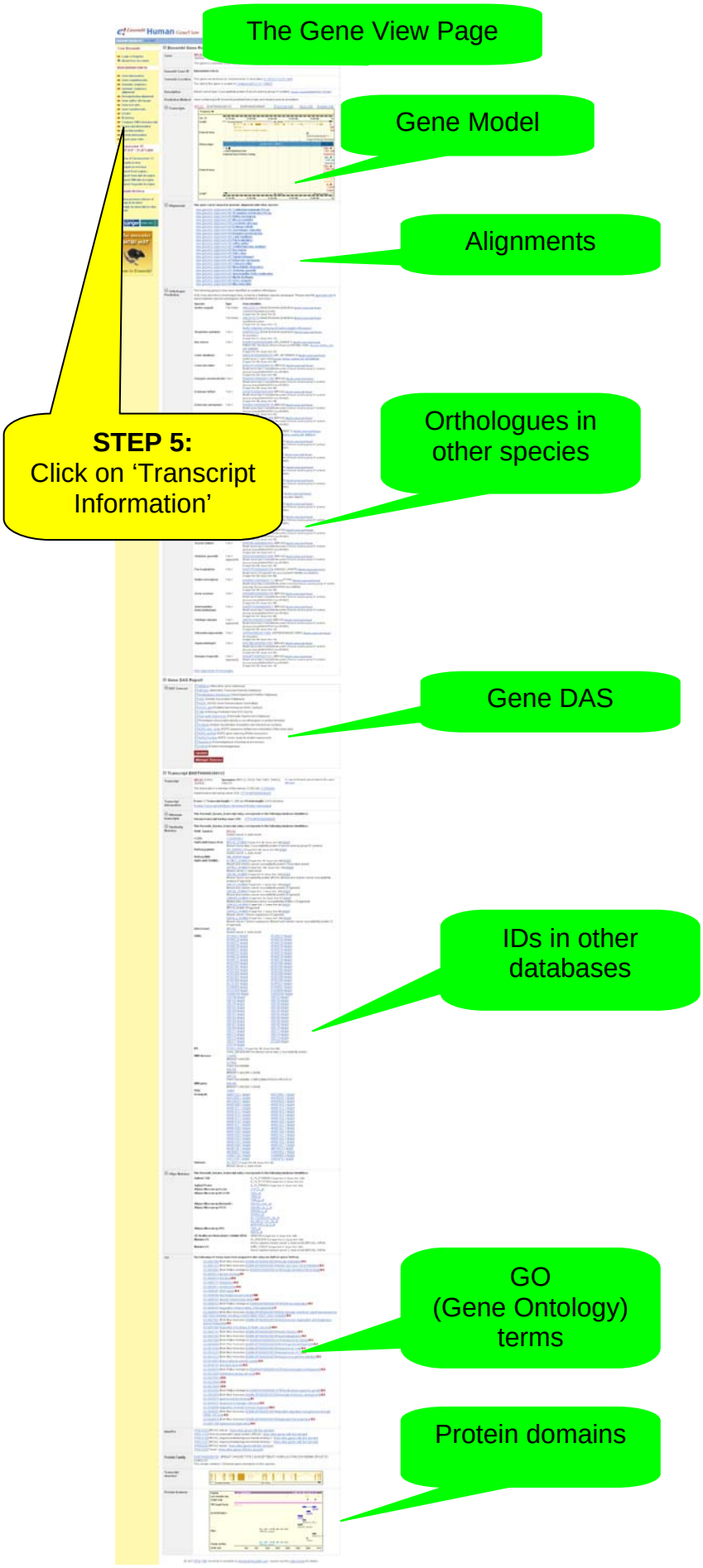
STEP 3:
Type in '**BRCA2 Gene**'.
Click 'Go'.

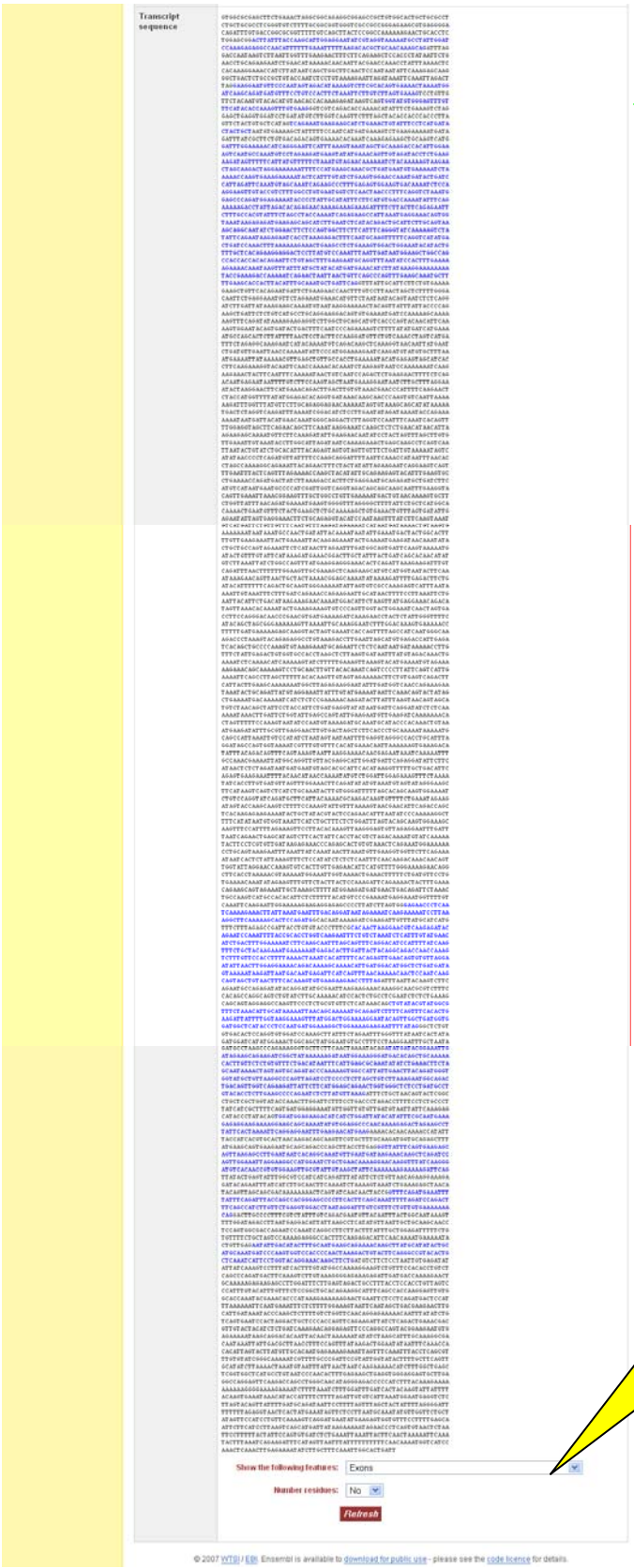
Karyotype

Source and version of assembly and genebuild




Ensembl Human SearchView
Search>>
Ensembl release 47 - Oct 2007
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Feature type
Gene (10)
Homo sapiens (10)
Species
Homo sapiens (10)
Gene (10)
Exalead Help
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Ensembl text search
BRCA2 gene
Search
Your query matched 10 entries in the search database
A 'VEGA' gene (VEGA is a team of manual curators external to Ensembl)
STEP 4:
Click on 'ENSG00000139618'
Caenorhabditis elegans
WormBase 180
now in Ensembl


The Gene View Page
Gene Model
Alignments
Orthologues in other species
STEP 5:
Click on 'Transcript Information'
Gene DAS
IDs in other databases
GO (Gene Ontology) terms
Protein domains

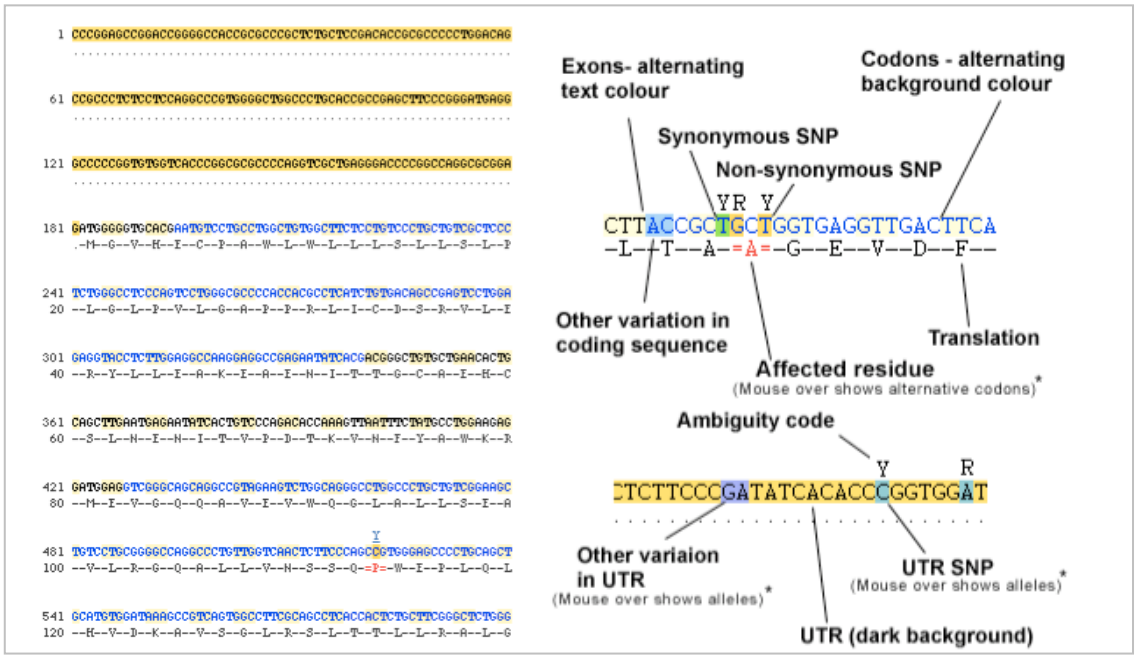


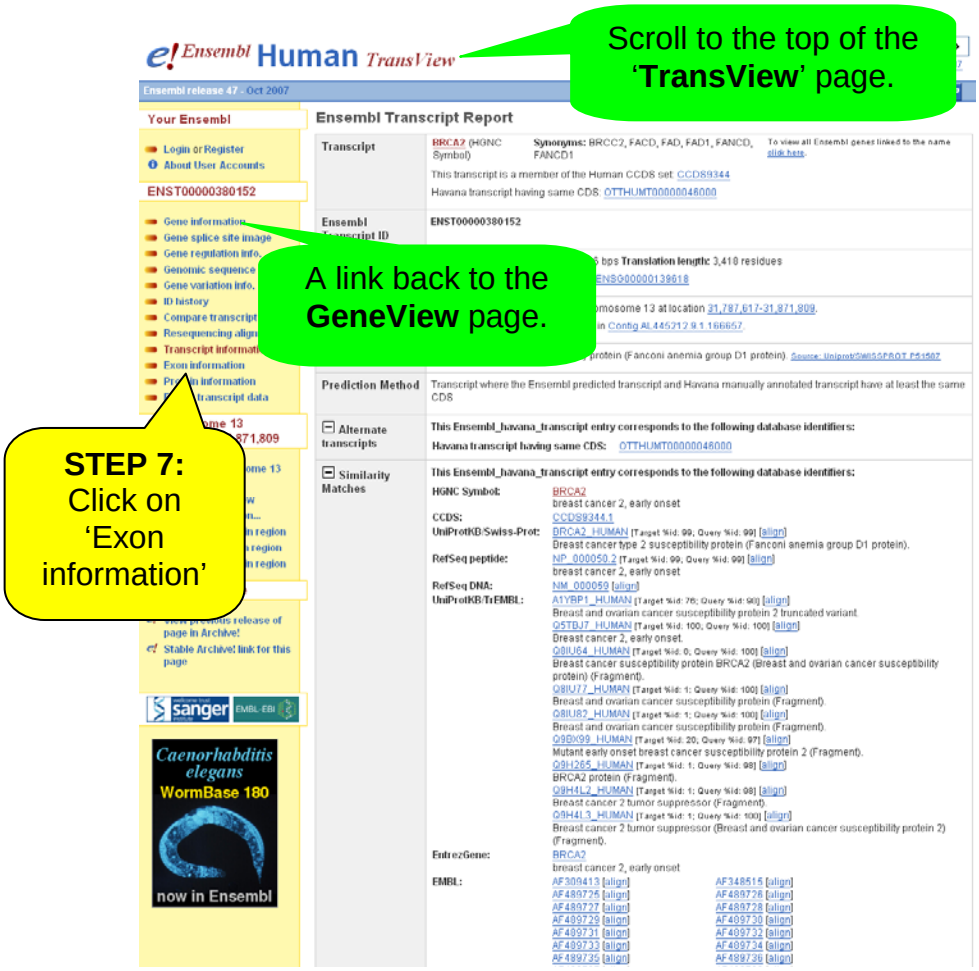
Scroll to the very bottom of the TransView page for the spliced transcript sequence.

Exons are shown in alternating colour.

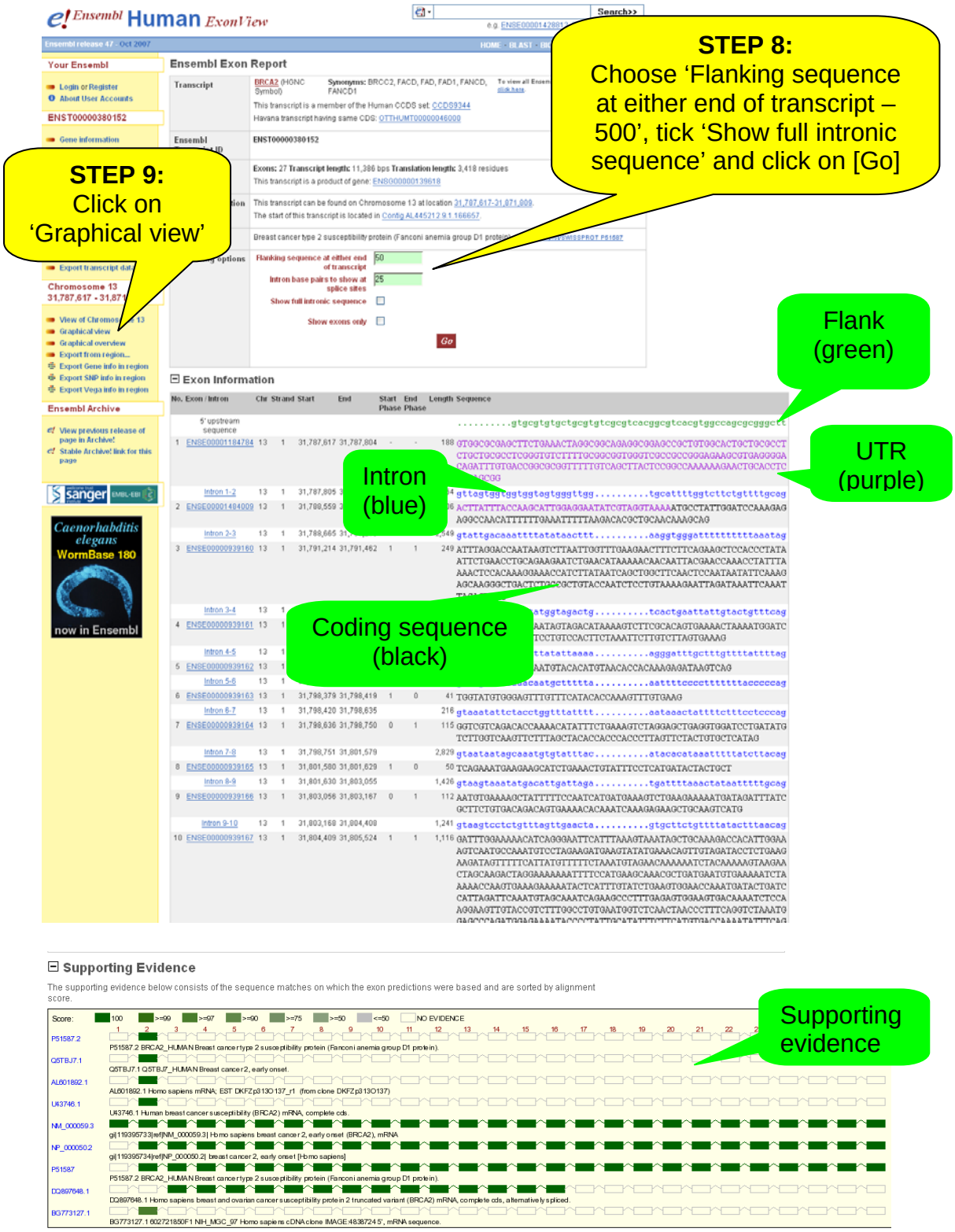
STEP 6:
Select 'Exons, Codons, Translations and SNPs'. Select 'Number residues: Yes' and click on [Refresh]

Result of STEP 6:

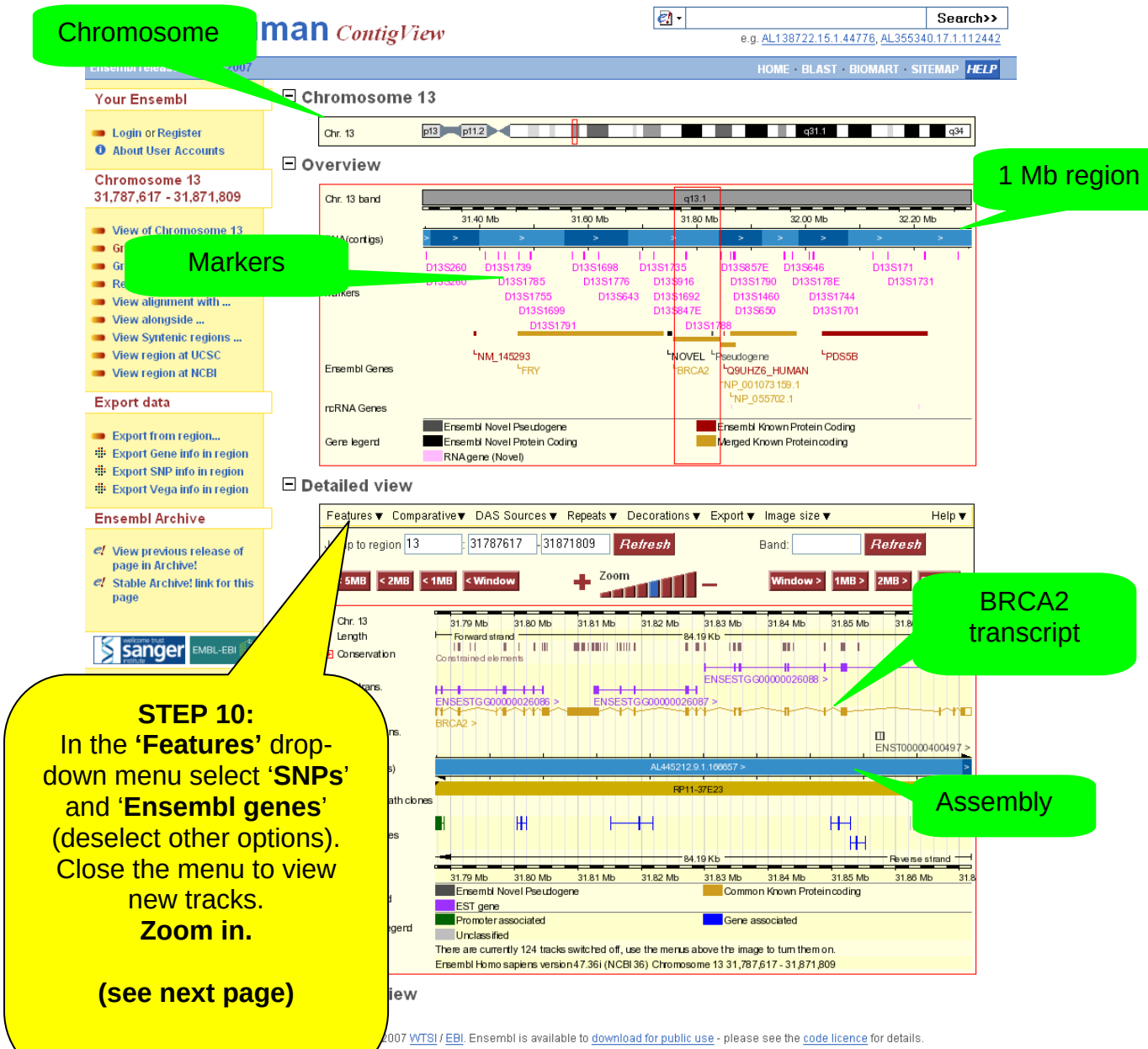




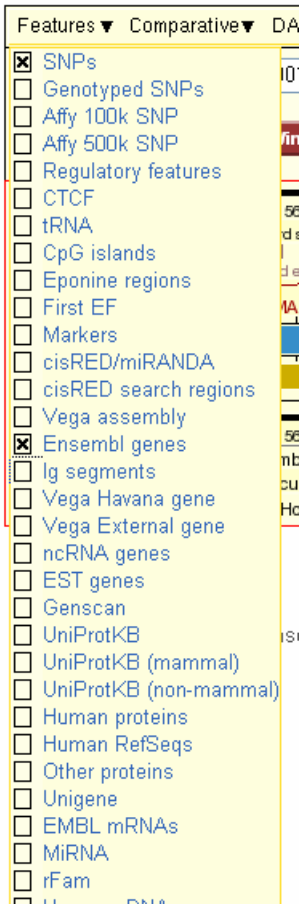
Result of STEP 7:



Result of STEP 9: ContigView

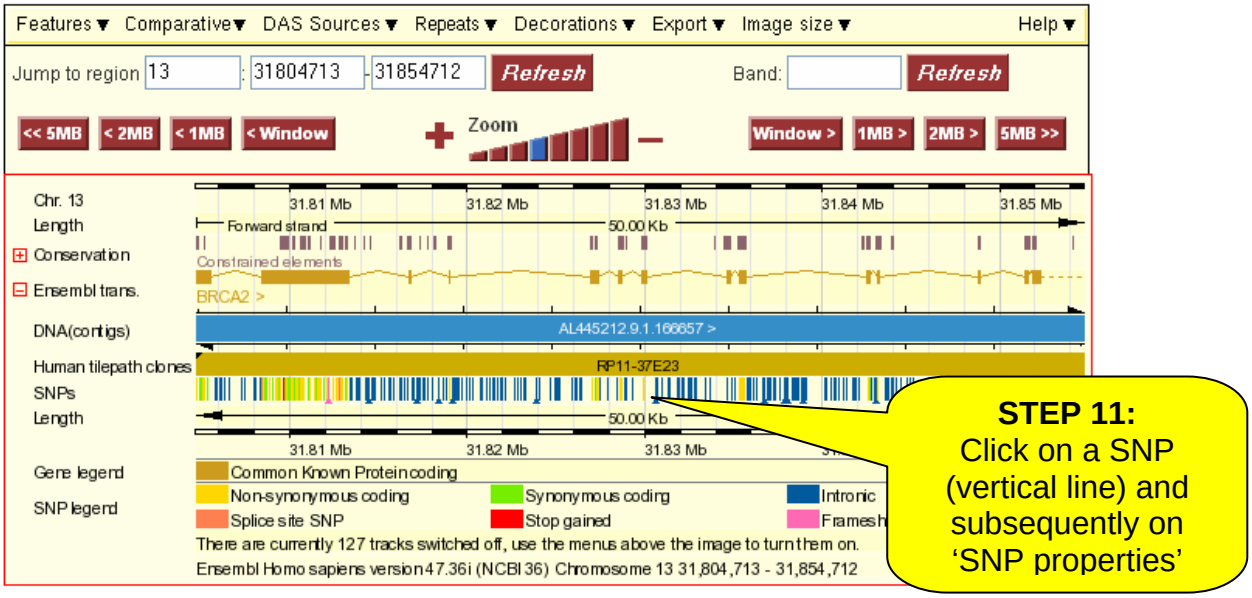


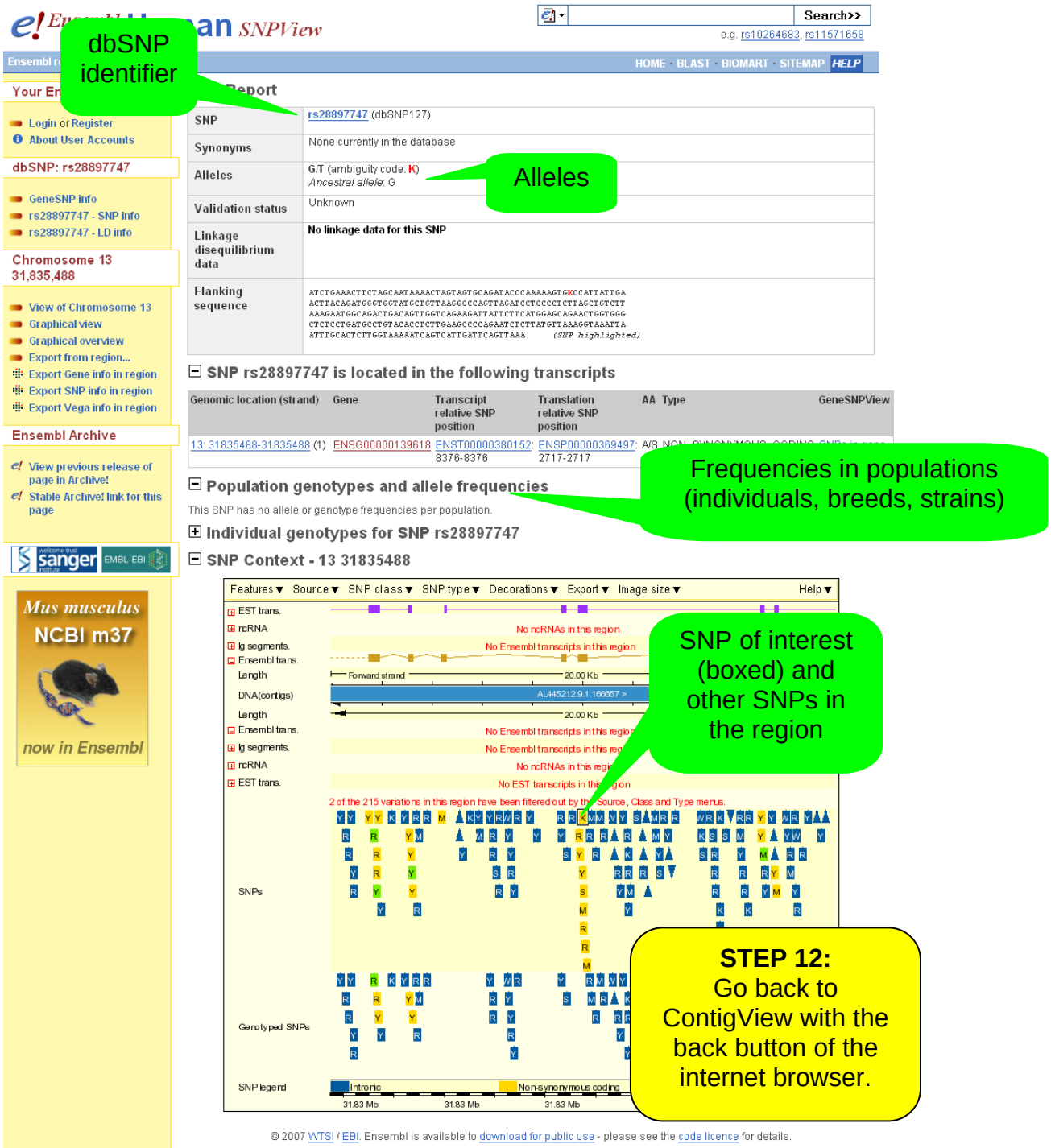
(STEP 10): Deselect all other options from the features menu and select ‘SNPs’ and ‘Ensembl genes’.

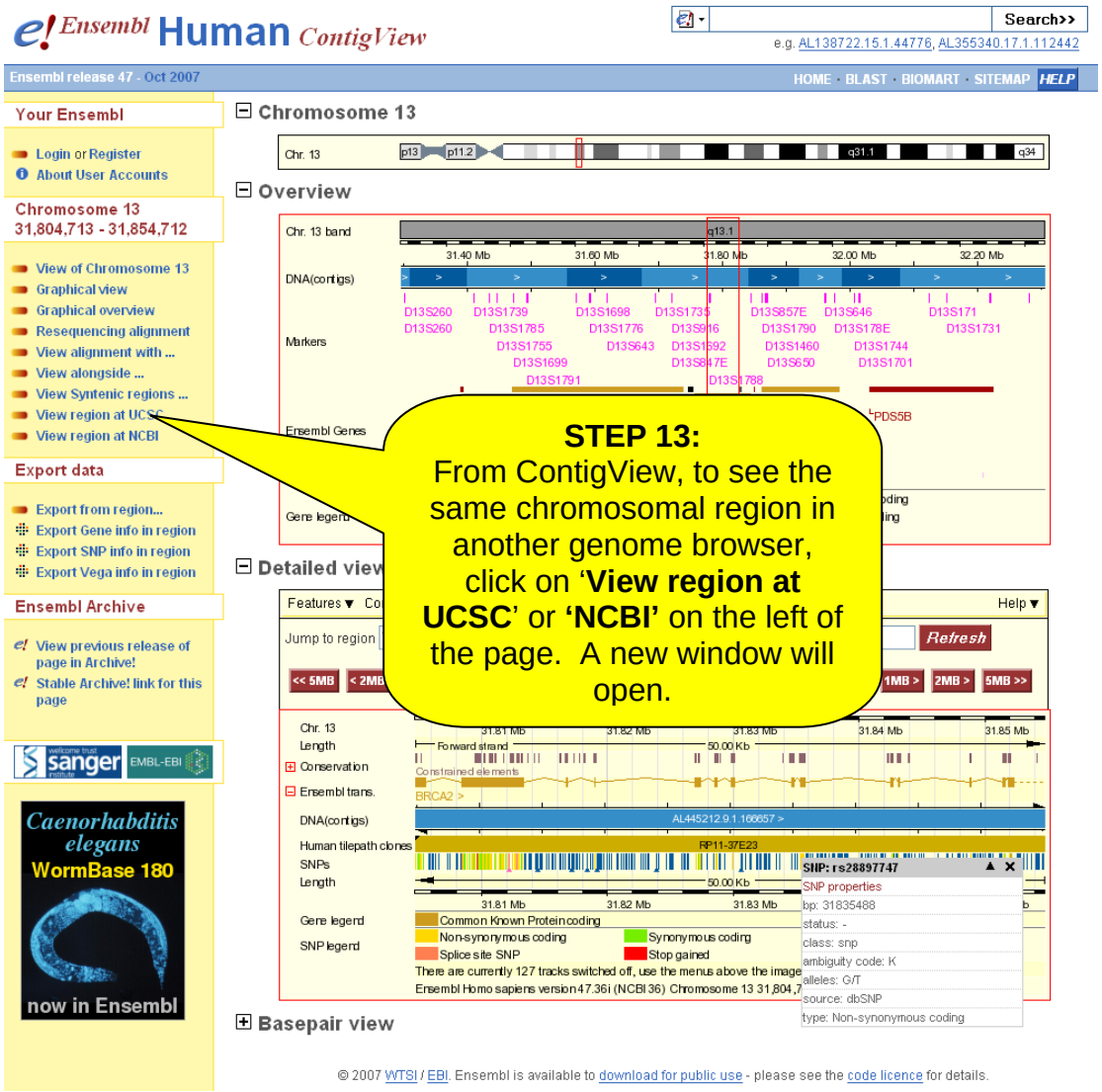


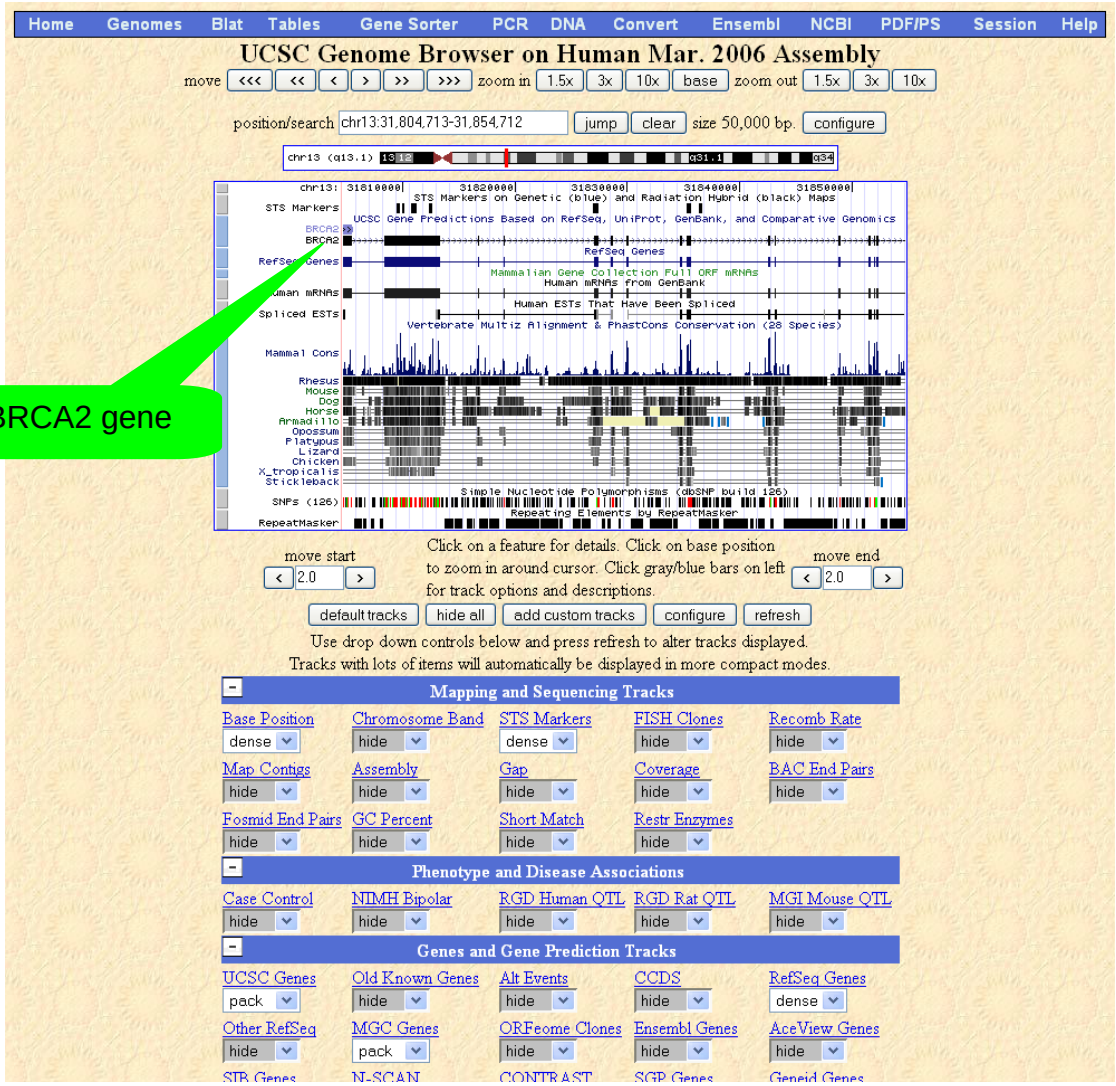
Close the menu and zoom in by clicking on the zoom triangle.

RESULT OF STEP 10:

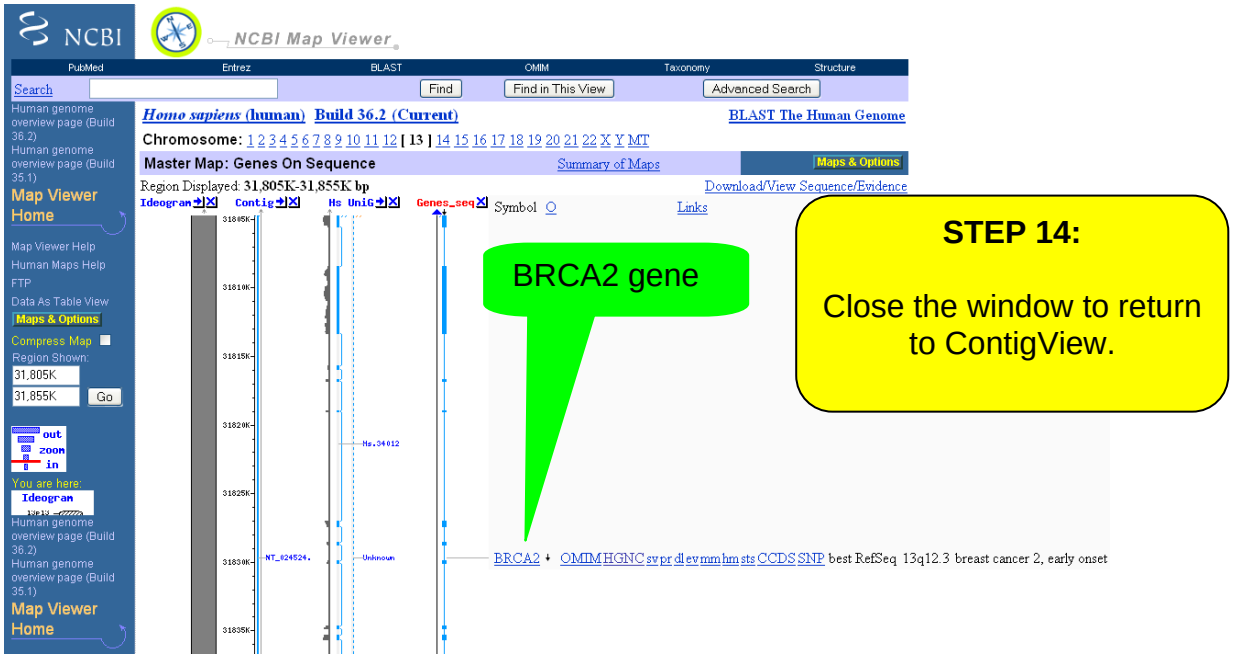



dbSNP identifier
SNPView
Search>>
e.g. rs10264683, rs11571658
HOME · BLAST · BIOMART · SITEMAP HELP
Login or Register
About User Accounts
dbSNP: rs28897747
GeneSNP info
rs28897747 - SNP info
rs28897747 - LD info
Chromosome 13
31,835,488
View of Chromosome 13
Graphical view
Graphical overview
Export from region...
Export Gene info in region
Export SNP info in region
Export Vega info in region
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Mus musculus
NCBI m37
now in Ensembl
SNP
rs28897747 (dbSNP127)
Synonyms
None currently in the database
Alleles
G/T (ambiguity code: K)
Ancestral allele: G
Alleles
Validation status
Unknown
Linkage disequilibrium data
No linkage data for this SNP
Flanking sequence
ATCTGAAACTTCTAGCAATAAAACTAGTAGTGCAGATACCCAAAAAGTGKCCATTATTGA
ACTTACAGATGGGTGGTATGCTGTTAAGGCCCAGTTAGATCCTCCCCTCTTAGCTGTCTT
AAAGAATGGCAGACTGACAGTTGGTCAGAAGATTATTCTTCATGGAGCAGAACTGGTGGG
CTCTCCTGATGCCTGTACACCTCTTGAAGCCCCAGAATCTCTTATGTTAAAGGTAAATTA
ATTTGCACTCTTGGTAAAAATCAGTCATTGATTCAGTTAAA (SNP highlighted)
SNP rs28897747 is located in the following transcripts
Genomic location (strand)
Gene
Transcript relative SNP position
Translation relative SNP position
AA Type
GeneSNPView
13: 31835488-31835488 (1)
ENSG00000139618
ENST00000380152: 8376-8376
ENSP00000369497: 2717-2717
Population genotypes and allele frequencies
This SNP has no allele or genotype frequencies per population.
Frequencies in populations (individuals, breeds, strains)
Individual genotypes for SNP rs28897747
SNP Context - 13 31835488
Features Source SNP class SNP type Decorations Export Image size Help
EST trans.
ncRNA
Ig segments.
Ensembl trans.
Length
Forward strand
20.00 Kb
DNA(contigs)
AL445212.9.1.166657 >
No ncRNAs in this region
No Ensembl transcripts in this region
No EST transcripts in this region
2 of the 215 variations in this region have been filtered out by the Source, Class and Type menus
SNPs
Genotyped SNPs
SNP legend
Intronic
Non-synonymous coding
31.83 Mb
SNP of interest (boxed) and other SNPs in the region
STEP 12:
Go back to ContigView with the back button of the internet browser.
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

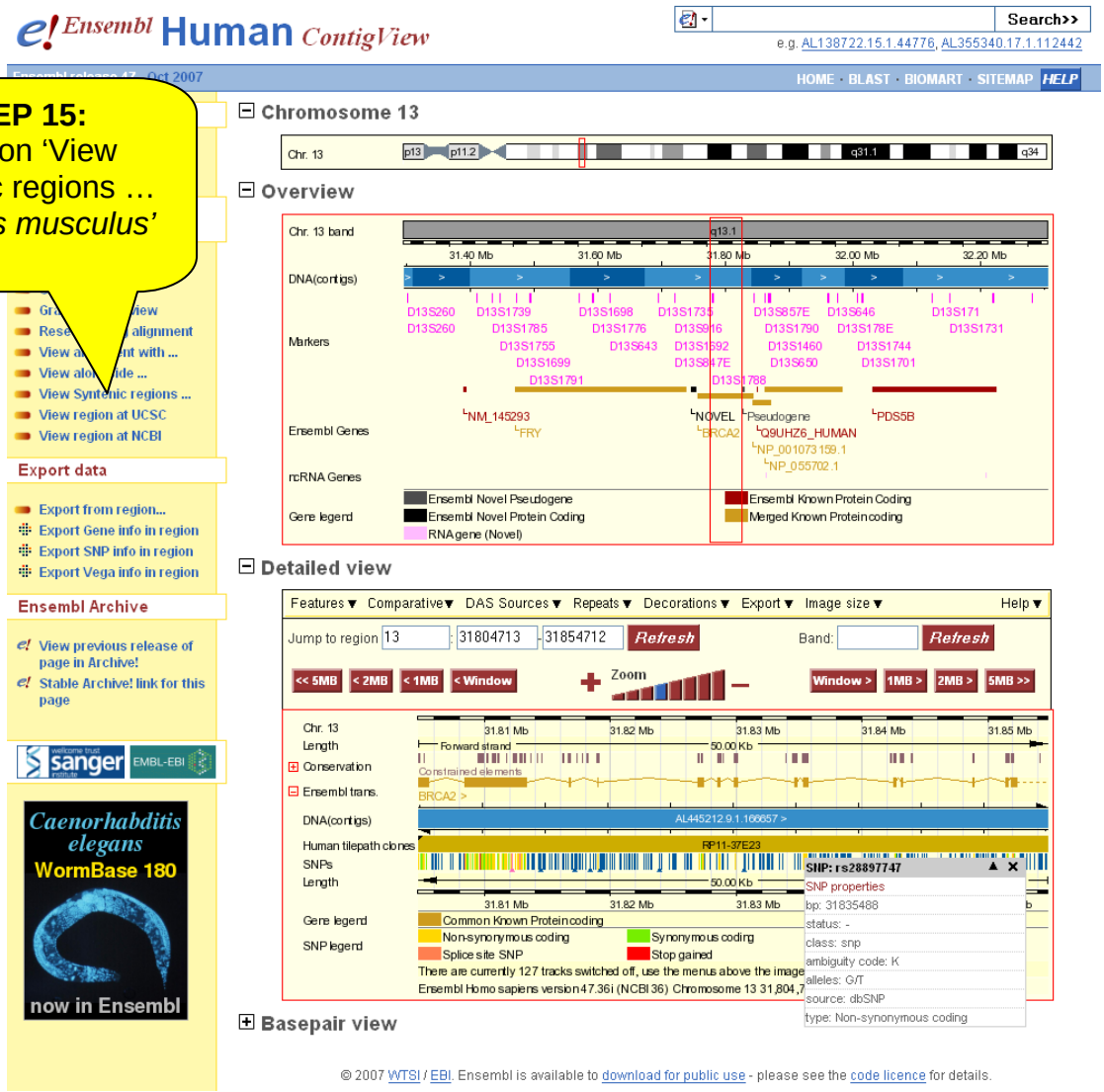

Ensembl Human ContigView
Search>>
e.g. AL138722.15.1.44776, AL355340.17.1.112442
Ensembl release 47 - Oct 2007
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Chromosome 13
31,804,713 - 31,854,712
View of Chromosome 13
Graphical view
Graphical overview
Resequencing alignment
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region at NCBI
Export data
Export from region...
Export Gene info in region
Export SNP info in region
Export Vega info in region
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Caenorhabditis elegans
WormBase 180
now in Ensembl
Chromosome 13
Overview
Detailed view
Basepair view
STEP 13:
From ContigView, to see the same chromosomal region in another genome browser, click on 'View region at UCSC' or 'NCBI' on the left of the page. A new window will open.
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

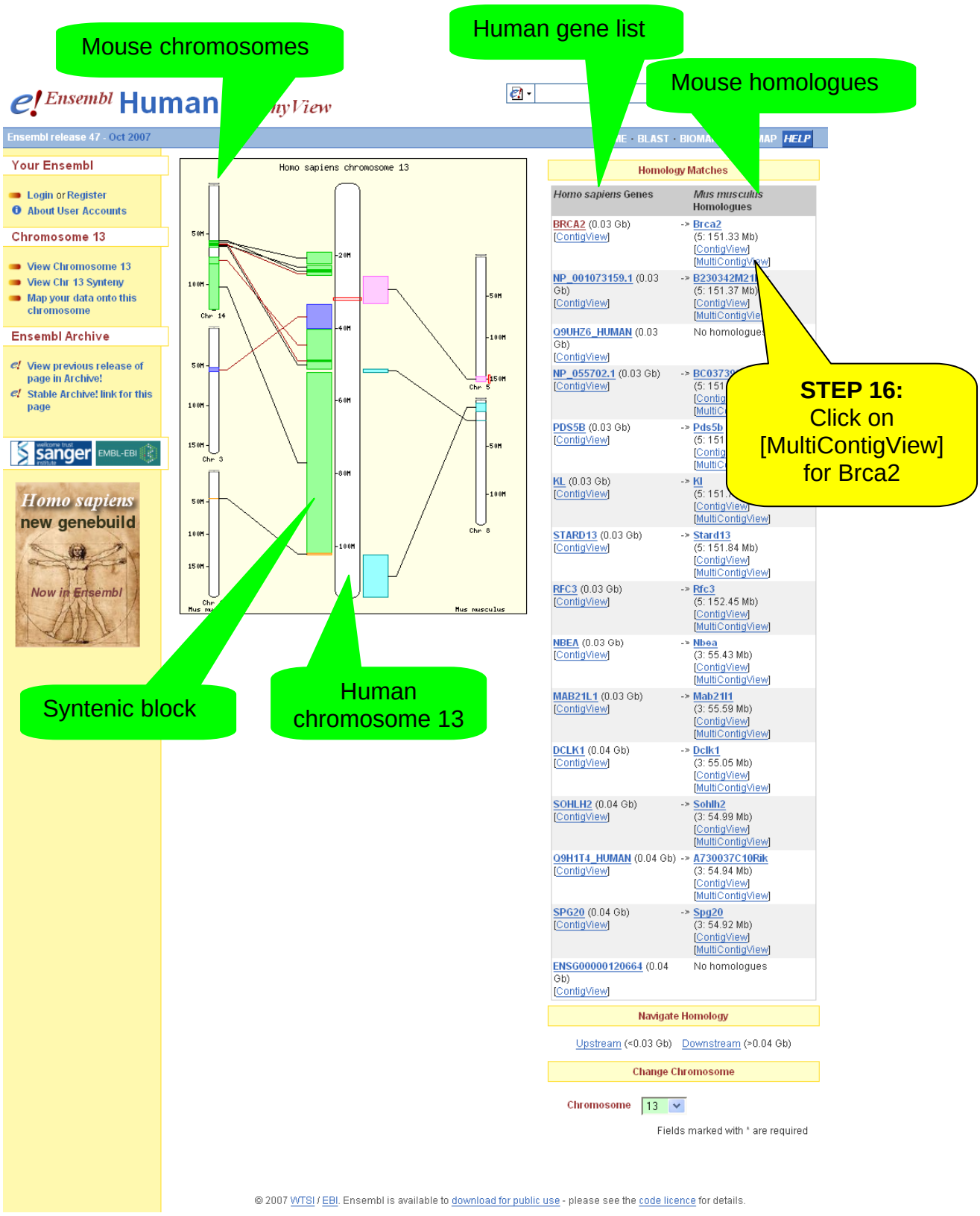

Home Genomes Blat Tables Gene Sorter PCR DNA Convert Ensembl NCBI PDF/PS Session Help
UCSC Genome Browser on Human Mar. 2006 Assembly
BRCA2 gene
Mapping and Sequencing Tracks
Phenotype and Disease Associations
Genes and Gene Prediction Tracks

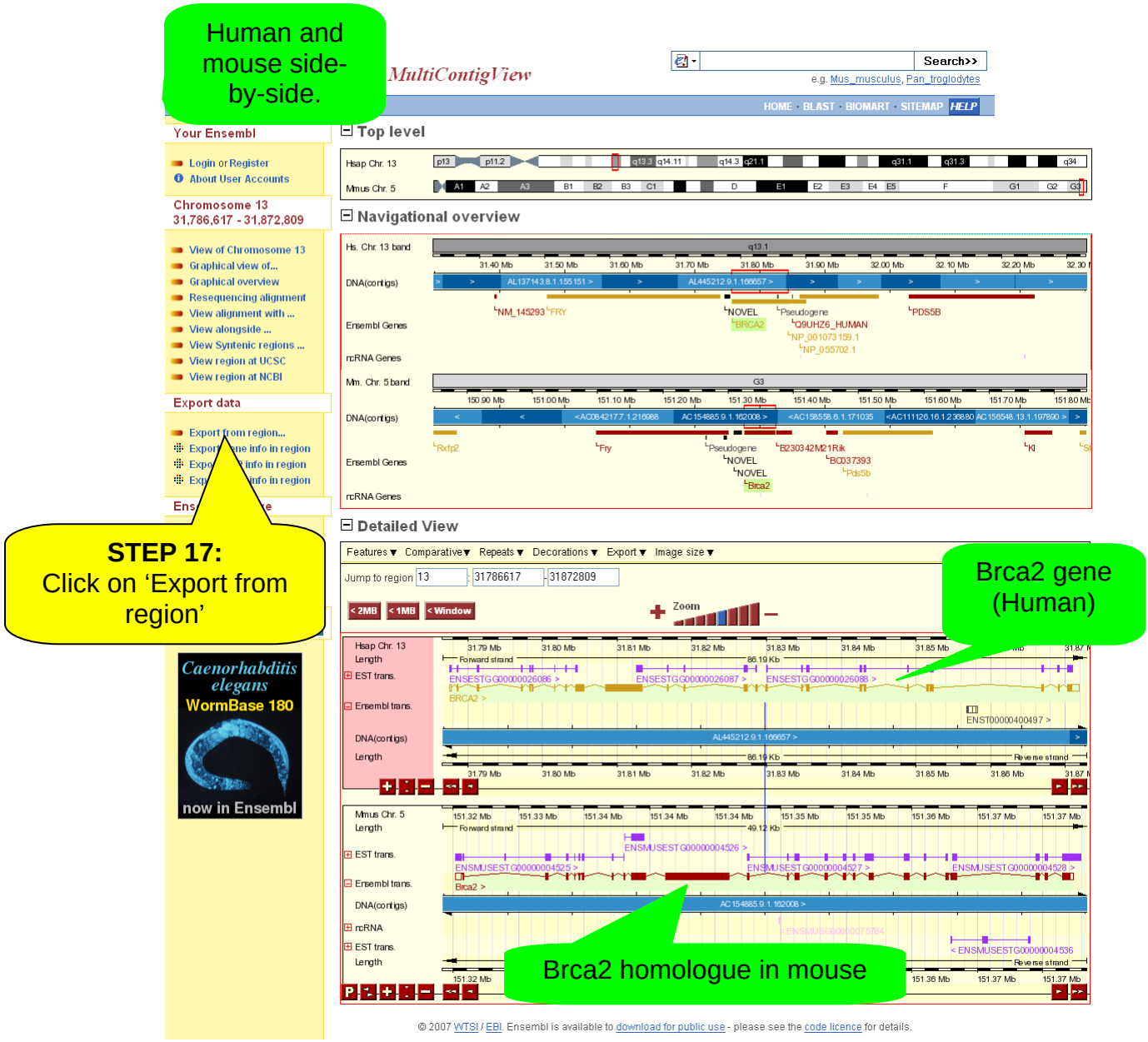
(or, NCBI)

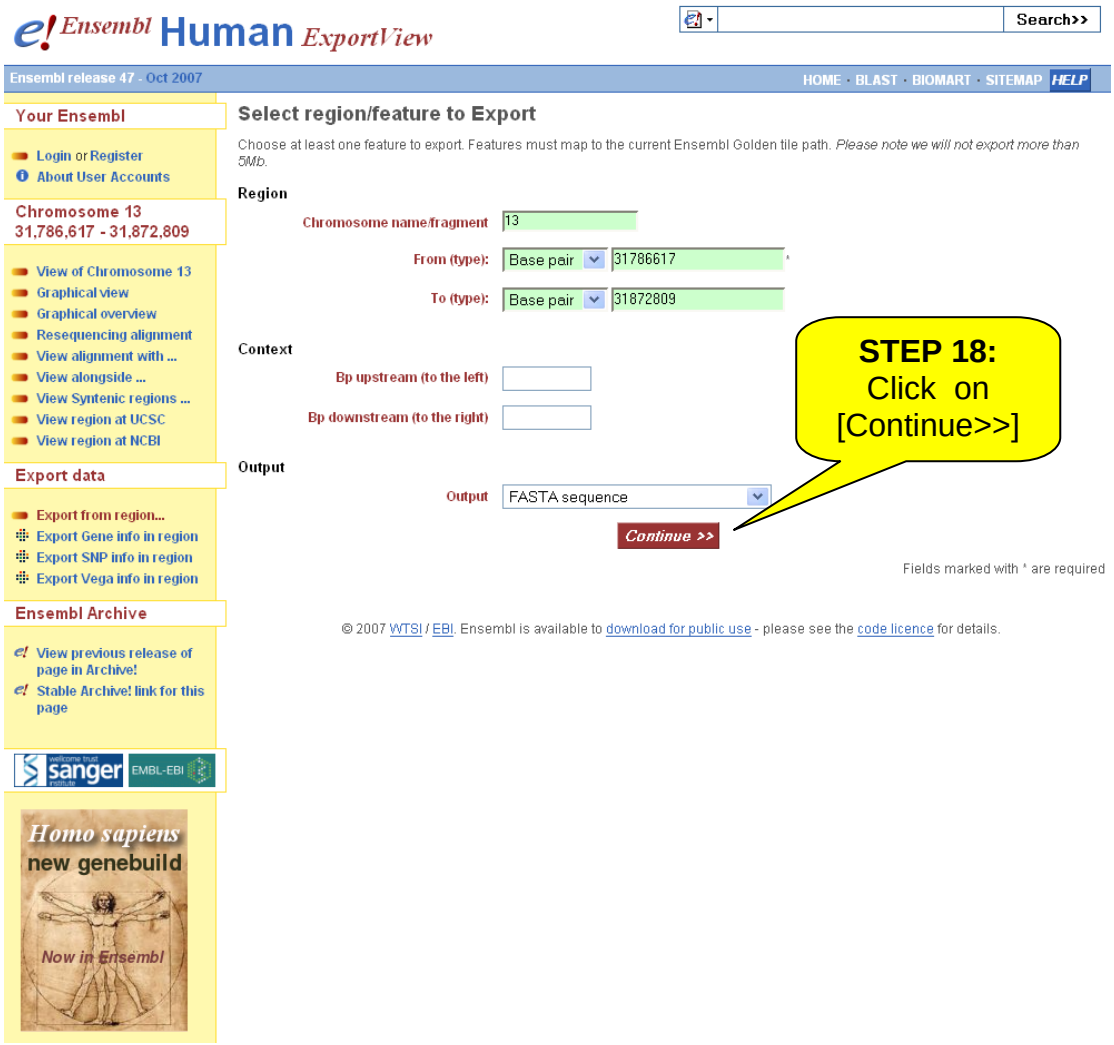

NCBI
NCBI Map Viewer
BRCA2 gene
STEP 14:
Close the window to return to ContigView.

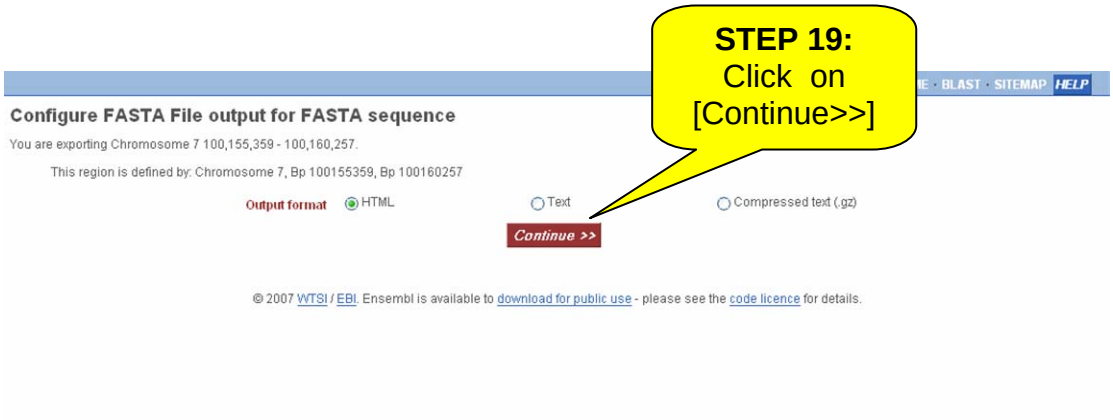
STEP 15:
Click on 'View Syntenic regions … with *Mus musculus'*

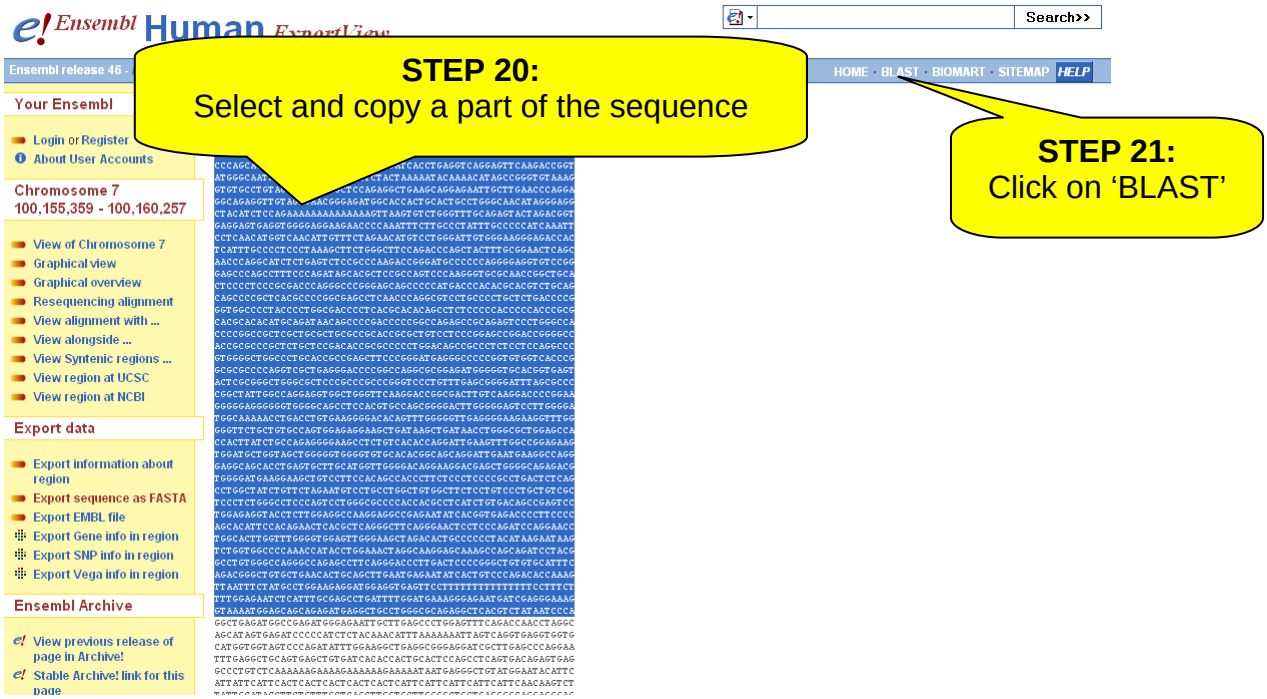


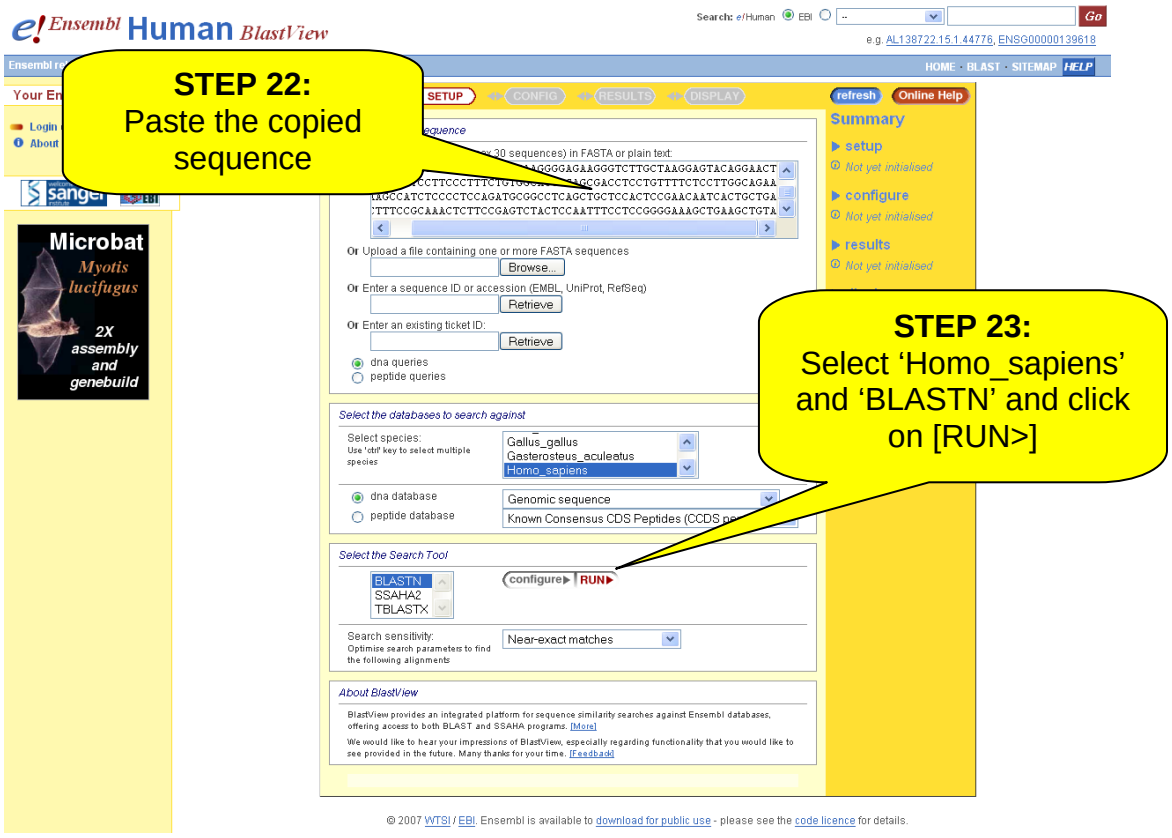

Mouse chromosomes
Human gene list
Mouse homologues
Ensembl Human
Ensembl release 47 - Oct 2007
Your Ensembl
Login or Register
About User Accounts
Chromosome 13
View Chromosome 13
View Chr 13 Synteny
Map your data onto this chromosome
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Homo sapiens chromosome 13
Homology Matches
Homo sapiens Genes
Mus musculus Homologues
STEP 16:
Click on [MultiContigView] for Brca2
Navigate Homology
Upstream (<0.03 Gb) Downstream (>0.04 Gb)
Change Chromosome
Chromosome 13
Fields marked with * are required
Syntenic block
Human chromosome 13

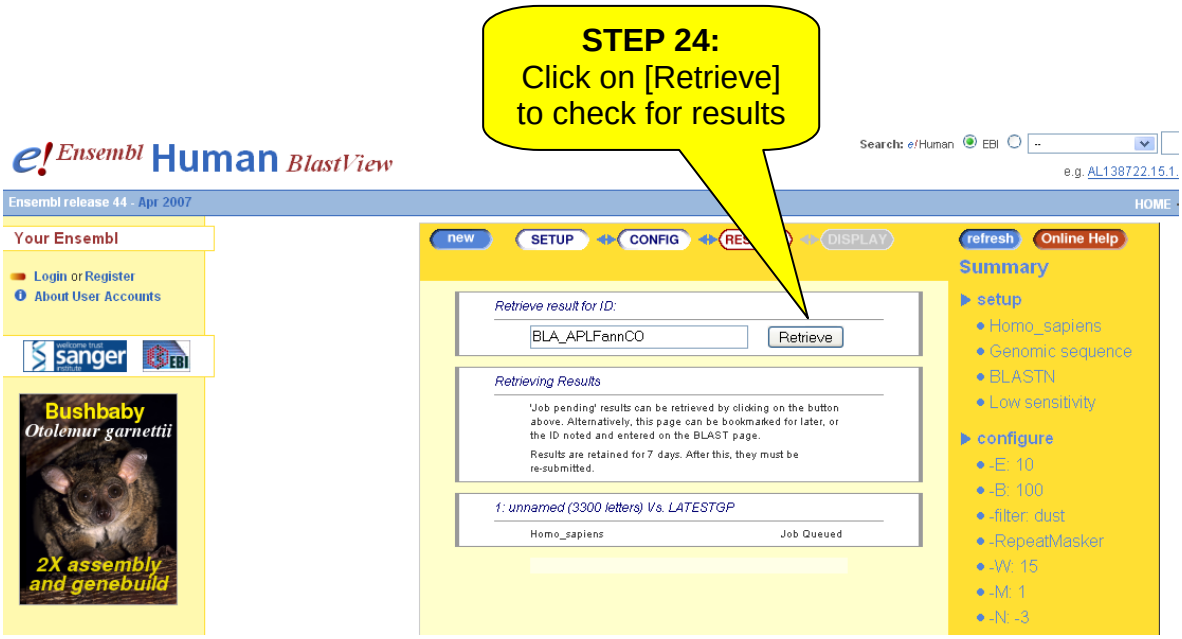

Human and mouse side-by-side.
MultiContigView
Search>>
e.g. Mus_musculus, Pan_troglodytes
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Chromosome 13
31,786,617 - 31,872,809
View of Chromosome 13
Graphical view of...
Graphical overview
Resequencing alignment
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region at NCBI
Export data
Export from region...
Top level
Hsap Chr. 13
Mmus Chr. 5
Navigational overview
Hs. Chr. 13 band
DNA(contigs)
Ensembl Genes
ncRNA Genes
Mm. Chr. 5 band
Detailed View
Features
Comparative
Repeats
Decorations
Export
Image size
Jump to region
13
31786617
31872809
< 2MB
< 1MB
< Window
Zoom
Brca2 gene (Human)
Brca2 homologue in mouse
STEP 17:
Click on 'Export from region'
Caenorhabditis elegans
WormBase 180
now in Ensembl
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.

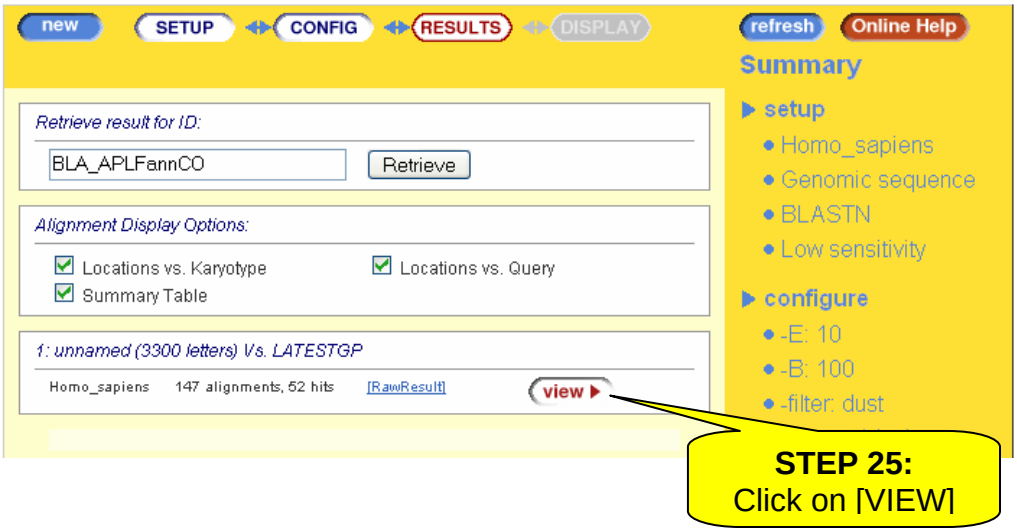

Ensembl Human ExportView
Search>>
Ensembl release 47 - Oct 2007
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Chromosome 13
31,786,617 - 31,872,809
View of Chromosome 13
Graphical view
Graphical overview
Resequencing alignment
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region at NCBI
Export data
Export from region...
Export Gene info in region
Export SNP info in region
Export Vega info in region
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
Homo sapiens
new genebuild
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
13
From (type):
Base pair
31786617
To (type):
Base pair
31872809
Context
Bp upstream (to the left)
Bp downstream (to the right)
Output
Output
FASTA sequence
Continue >>
Fields marked with * are required
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.
STEP 18:
Click on
[Continue>>]

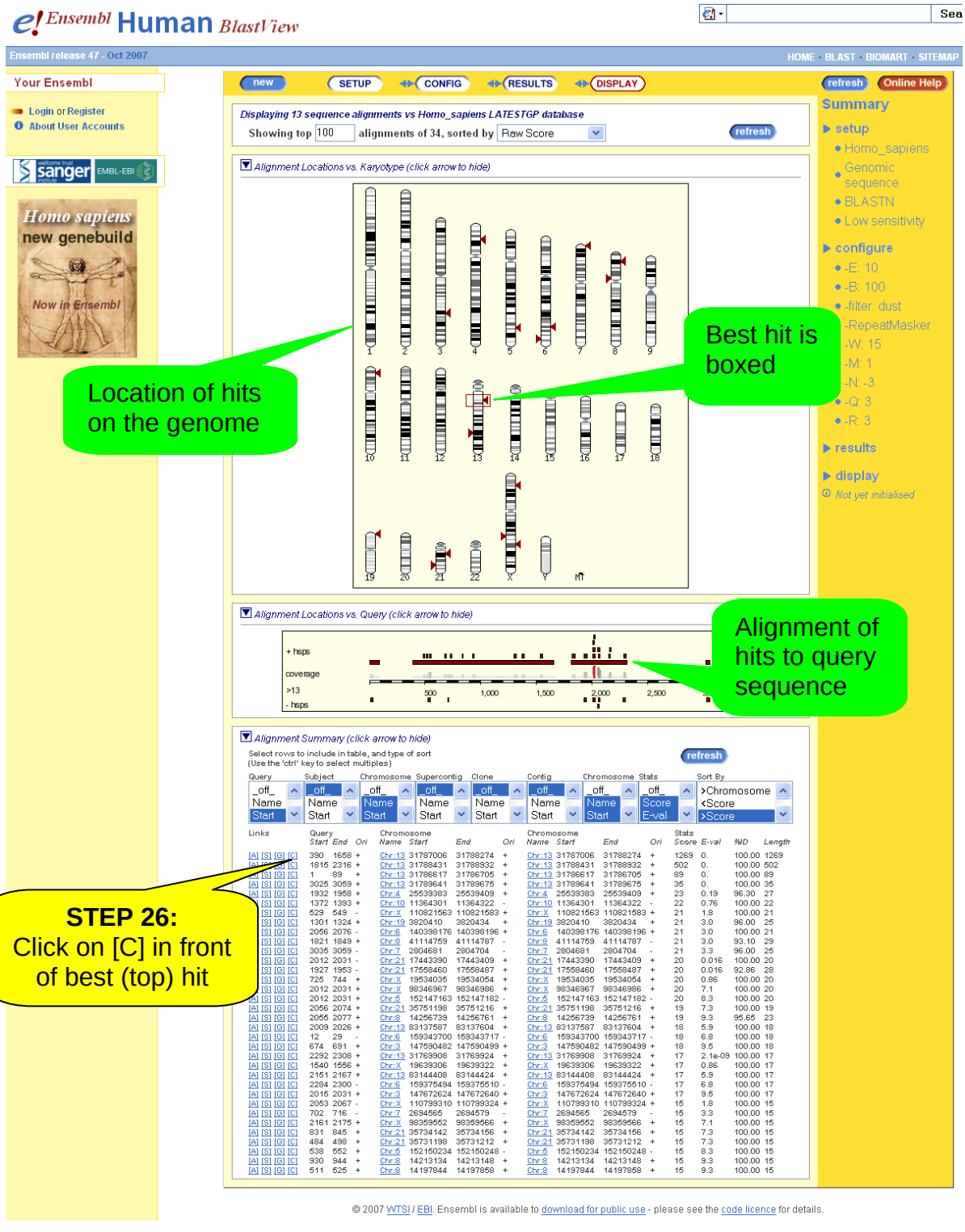

BLAST · SITEMAP HELP
Configure FASTA File output for FASTA sequence
You are exporting Chromosome 7 100,155,359 - 100,160,257.
This region is defined by: Chromosome 7, Bp 100155359, Bp 100160257
Output format
HTML
Text
Compressed text (.gz)
Continue >>
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.
STEP 19:
Click on
[Continue>>]

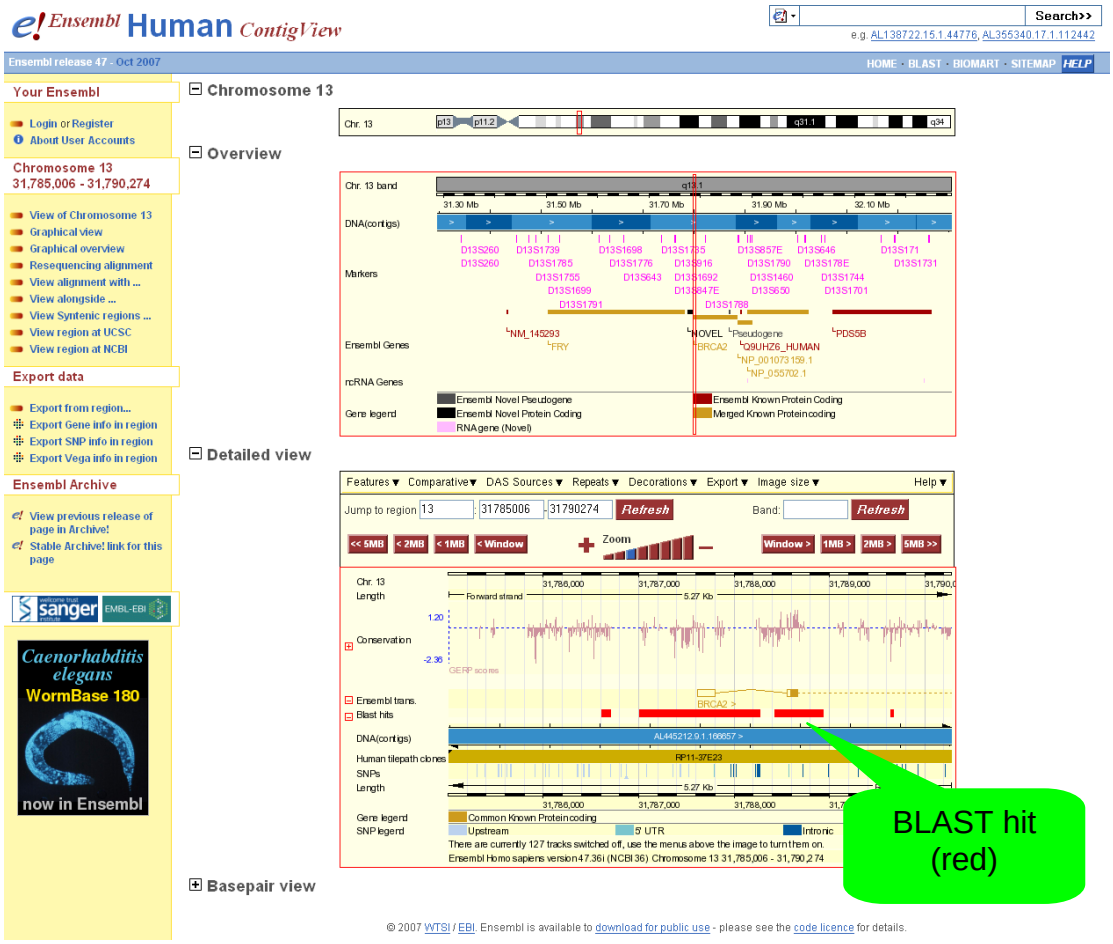

Ensembl Human ExpertView
Search>>
Ensembl release 46 -
HOME · BLAST · BIOMART · SITEMAP HELP
Your Ensembl
Login or Register
About User Accounts
Chromosome 7
100,155,359 - 100,160,257
View of Chromosome 7
Graphical view
Graphical overview
Resequencing alignment
View alignment with ...
View alongside ...
View Syntenic regions ...
View region at UCSC
View region at NCBI
Export data
Export 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
Ensembl Archive
View previous release of page in Archive!
Stable Archive! link for this page
STEP 20:
Select and copy a part of the sequence
STEP 21:
Click on 'BLAST'


Ensembl Human BlastView
Search: e! Human EBI
e.g. AL138722.15.1.44776, ENSG00000139618
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
Known Consensus CDS Peptides (CCDS
Select the Search Tool
BLASTN
SSAHA2
TBLASTX
configure RUN
Search sensitivity:
Optimise search parameters to find the following alignments
Near-exact matches
About BlastView
BlastView provides an integrated platform for sequence similarity searches against Ensembl databases, offering access to both BLAST and SSAHA programs. [More]
We would like to hear your impressions of BlastView, especially regarding functionality that you would like to see provided in the future. Many thanks for your time. [Feedback]
Microbat
Myotis lucifugus
2X assembly and genebuild
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.
STEP 22:
Paste the copied sequence
STEP 23:
Select 'Homo_sapiens' and 'BLASTN' and click on [RUN>]


STEP 24:
Click on [Retrieve]
to check for results
Ensembl Human BlastView
Search: e!Human EBI
e.g. AL138722.15.1
Ensembl release 44 - Apr 2007
HOME
Your Ensembl
Login or Register
About User Accounts
Bushbaby
Otolemur garnettii
2X assembly
and genebuild
new
SETUP
CONFIG
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
Retrieve result for ID:
BLA_APLFannCO
Retrieve
Retrieving Results
'Job pending' results can be retrieved by clicking on the button above. Alternatively, this page can be bookmarked for later, or the ID noted and entered on the BLAST page.
Results are retained for 7 days. After this, they must be re-submitted.
1: unnamed (3300 letters) Vs. LATESTGP
Homo_sapiens
Job Queued


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
Ensembl release 47 - Oct 2007
HOME · BLAST · BIOMART · SITEMAP
Your Ensembl
Login or Register
About User Accounts
Homo sapiens
new genebuild
Now in Ensembl
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 13 sequence alignments vs Homo_sapiens LATESTGP database
Showing top 100 alignments of 34, sorted by Raw Score
refresh
Alignment Locations vs. Karyotype (click arrow to hide)
Location of hits on the genome
Best hit is boxed
Alignment Locations vs. Query (click arrow to hide)
+ hsps
coverage
>13
- hsps
Alignment of hits to query sequence
Alignment Summary (click arrow to hide)
Select rows to include in table, and type of sort
(Use the 'ctrl' key to select multiples)
refresh
STEP 26:
Click on [C] in front of best (top) hit
© 2007 WTSI / EBI. Ensembl is available to download for public use - please see the code licence for details.
Back in the contigview page…



END of the Worked Example

EXERCISES and ANSWERS

Note: the answers to these exercises correspond to version 47 of Ensembl. If you use these exercises at a later date, please use the archive site for version 47.

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 'glossary' link in the 'Help & Information section (on the left) to go to this url:
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.

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

Exercise 3 explores Ensembl genes for a specific region, along with comparisons to mRNA and protein in other databases.

(a) What assembly is currently used for zebrafish?

(b) Bring up a ContigView display of zebrafish (*Danio rerio*) chromosome 1 between base pairs 3900000 and 3910000. Are there any 'known' or 'novel' genes in this region? For one of the known genes, find some information about its function, and look at an entry for it in other databases such as EntrezGene, UniProt/Swiss-Prot or the ZFIN site.

(c) View homologues in other species for a known Zebrafish gene, at other members of protein families and InterPro domains. This can yield information about possible functions.

Answers (Browsing Ensembl)

1. Exploring features related to a gene

(a) Two 'Vega' genes and 'Ensembl' gene will be shown. VEGA (Vertebrate 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 (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. 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) 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.

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

(a) Zv7 is the current assembly (see this in the main page or by clicking on z-fish to go to the species-specific index page).

(b) Start on the homepage for zebrafish (*Danio rerio*). In the 'Karyotype' section choose 'Chromosome: 1', 'From (type): Base pair: 3900000 To (type) Base pair: 3910000' and click [Go].

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. Click on one of the known genes to go to its GeneView page and explore the links in the 'Similarity Matches' section.

(c) Click on the gene in ContigView and follow the Ensembl gene ID to the GeneView page. If the gene has orthologues in other species (shown in the 'Orthologue Prediction' section) and these orthologues are well-understood, possible functions of this gene could be hypothesised. 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.

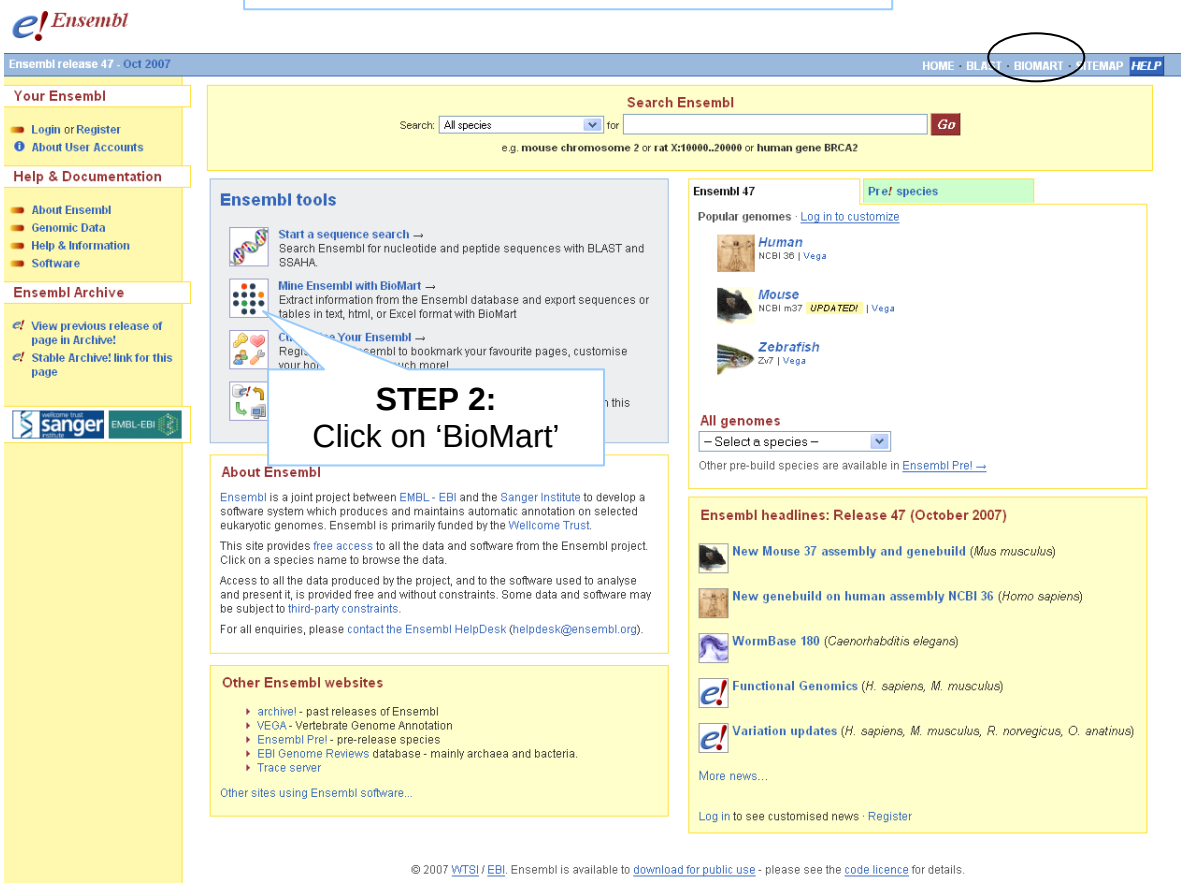
IV) Data mining in Ensembl with BioMart
Worked Example

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

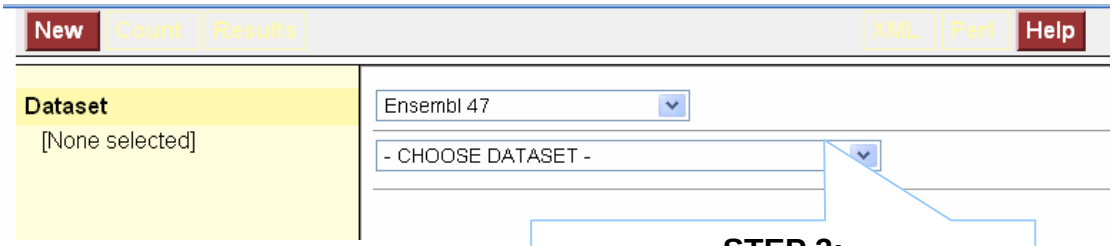
Which other genes related to human diseases locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs? Do they have any functions predicted by Interpro?

What are their cDNA sequences?

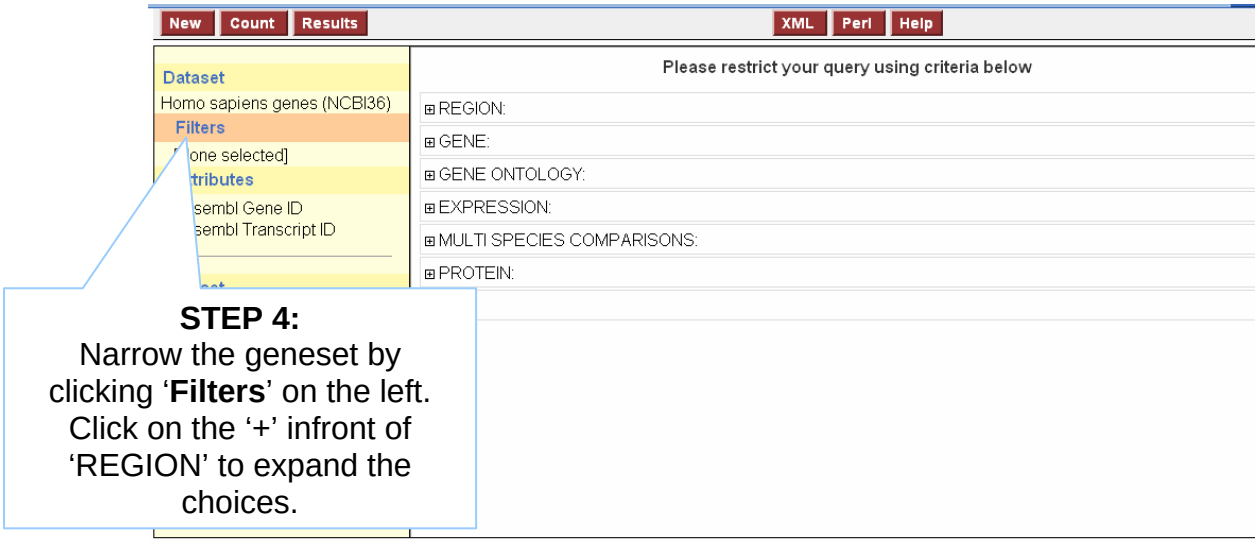
STEP 1:
Go to the Ensembl main page
www.ensembl.org



STEP 2:
Click on 'BioMart'

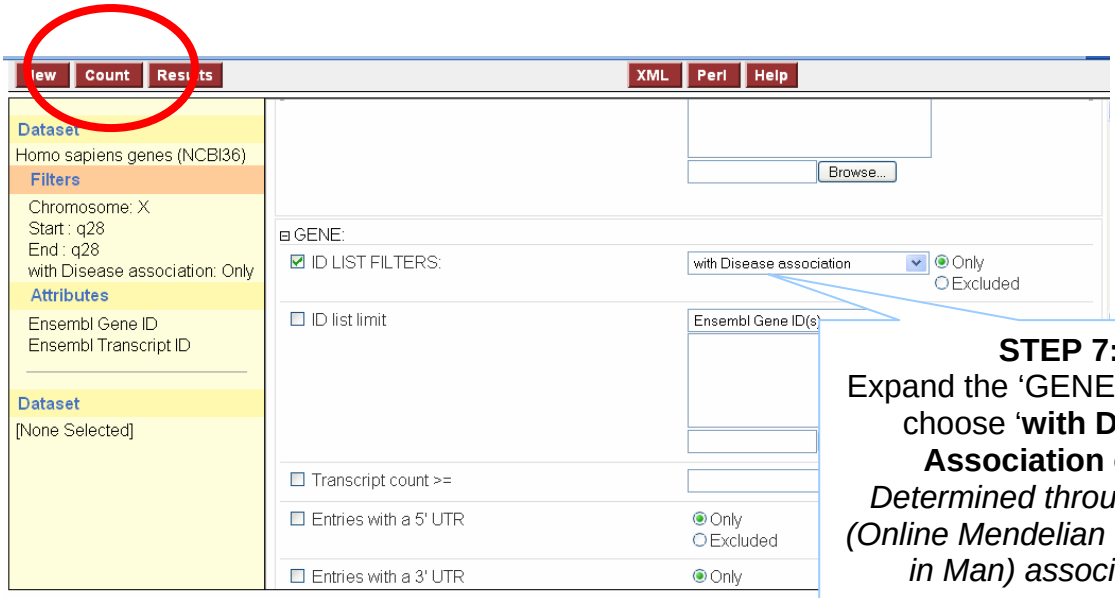


STEP 3:
Select the database: Ensembl genes (version 48) and the species of interest under 'Choose Dataset'. (*Homo sapiens*)



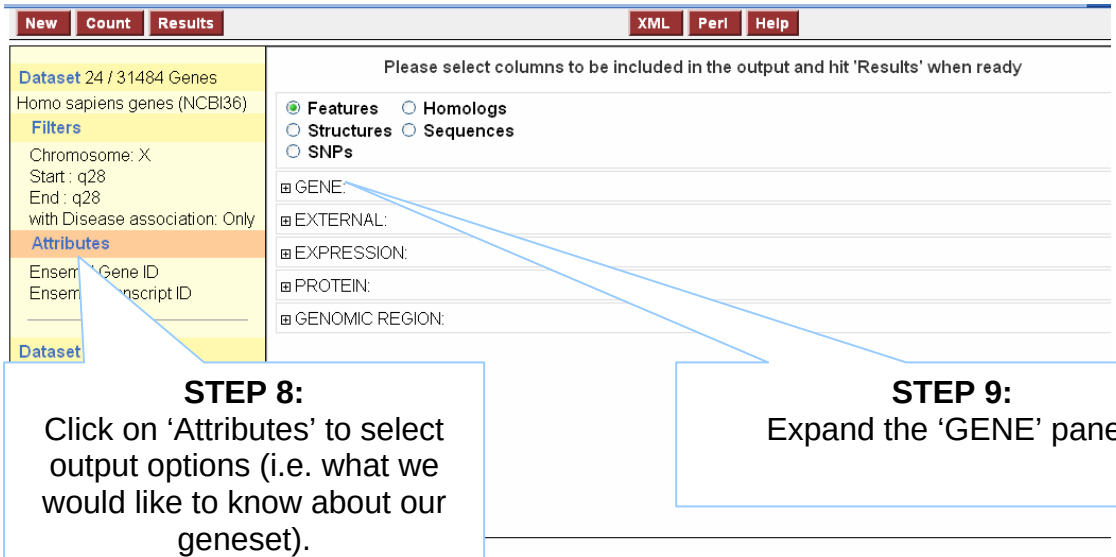
STEP 4:
Narrow the geneset by clicking '**Filters**' on the left. Click on the '+' infront of 'REGION' to expand the choices.


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'



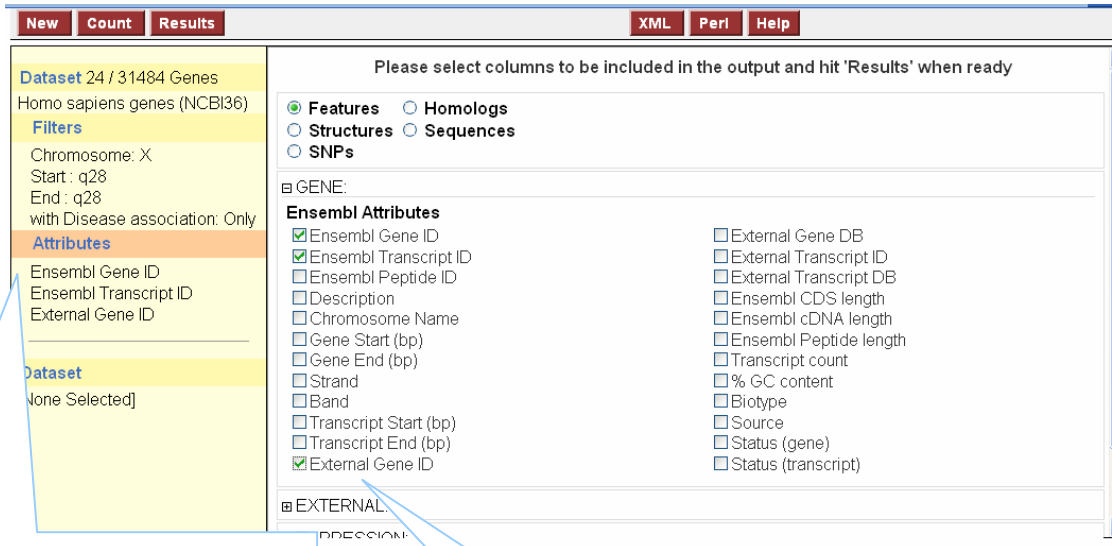
STEP 7:
Expand the 'GENE' panel and choose '**with Disease Association only**'. *Determined through OMIM (Online Mendelian Inheritance in Man) associations.*

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.

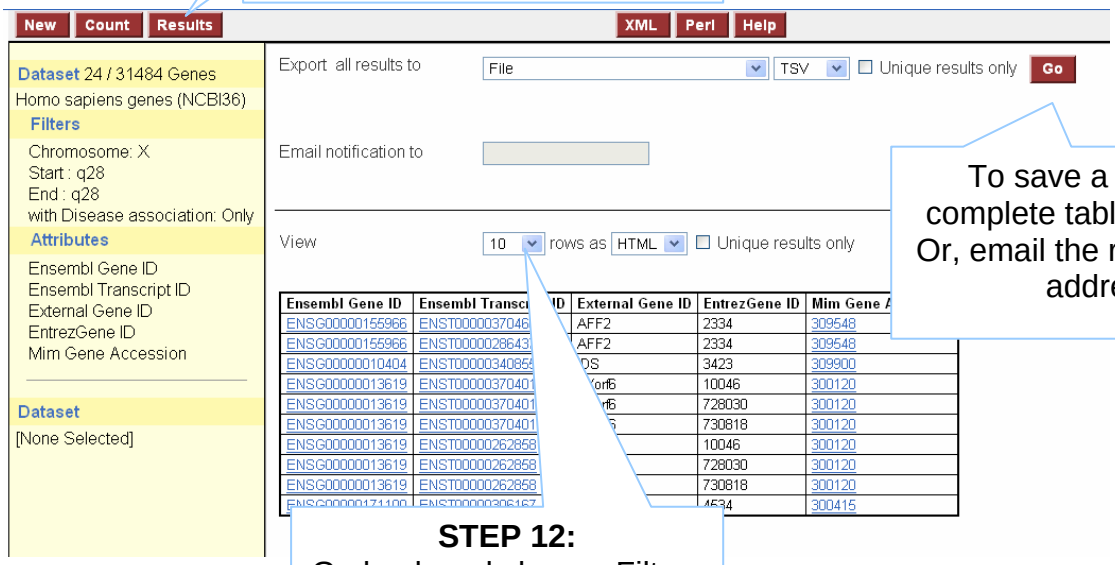


Note the summary of selected options.

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

STEP 10:
Select, along with the default options, '**External Gene ID**' (this shows the gene symbol from HGNC).
Expand the 'EXTERNAL' panel to select '**EntrezGene ID**' and '**Mim Gene Accession**' (this is the ID from OMIM)
www.ncbi.nlm.nih.gov/**omim**/

STEP 11:
Click 'RESULTS' at the top to preview the output.



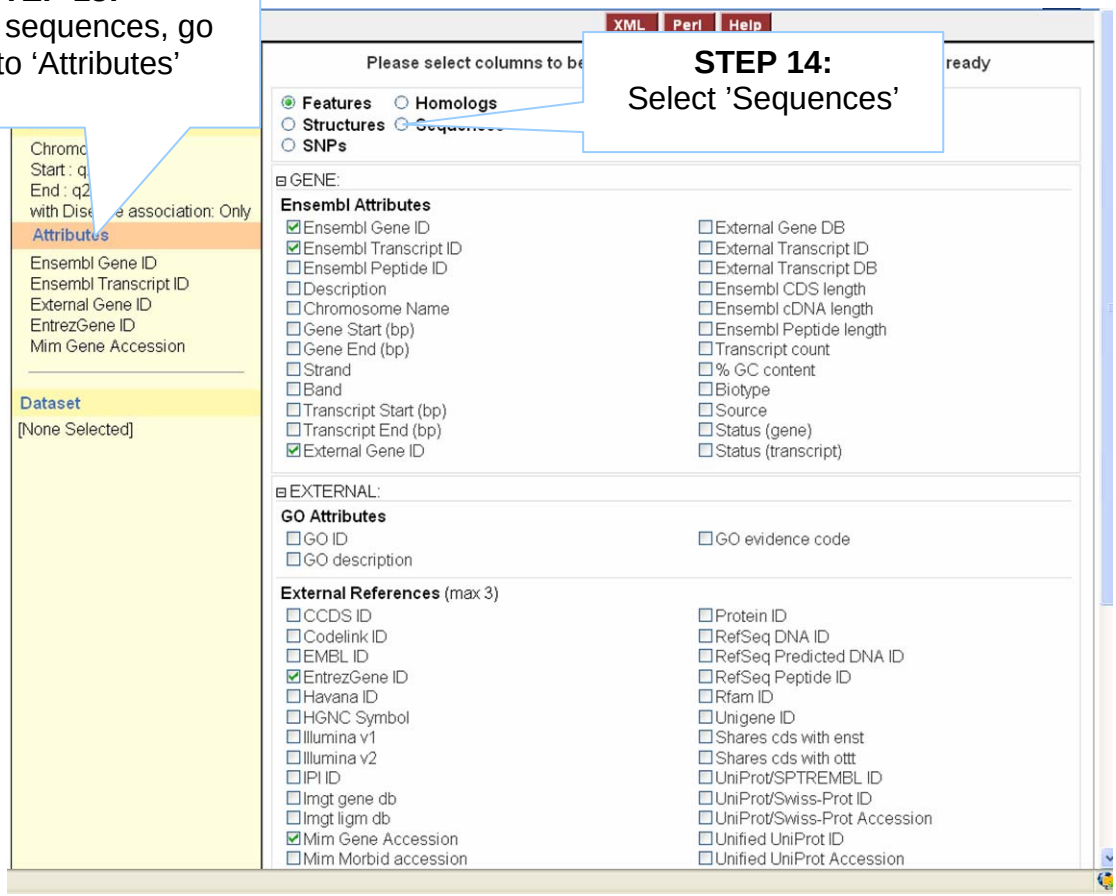
To save a file of the complete table, click 'Go'. Or, email the results to any address.

STEP 12:
Go back and change Filters or Attributes if desired.
Or, View 'ALL' as HTML…

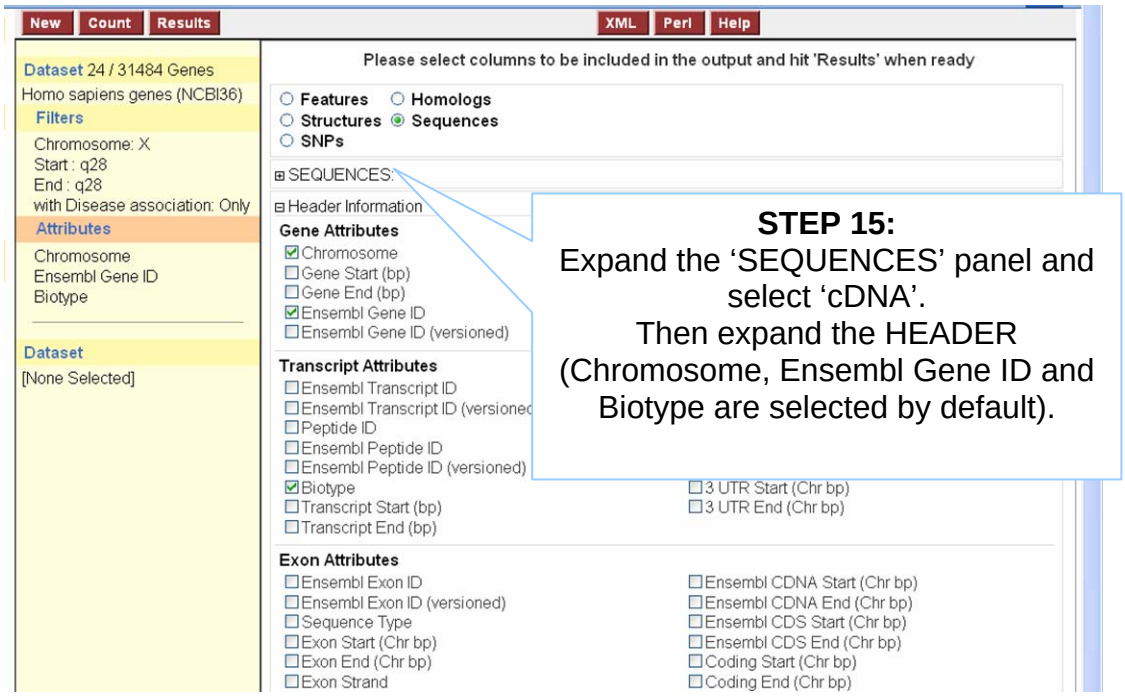
Result Table 1

Ensembl Gene ID	Ensembl Transcript ID	External Gene ID	EntrezGene ID	Mim Gene Accession
ENSG00000155966	ENST00000370460	AFF2	2334	309548
ENSG00000155966	ENST00000286437	AFF2	2334	309548
ENSG00000010404	ENST00000340855	IDS	3423	309900
ENSG00000013619	ENST00000370401	CXorf6	10046	300120
ENSG00000013619	ENST00000370401	CXorf6	728030	300120
ENSG00000013619	ENST00000370401	CXorf6	730818	300120
ENSG00000013619	ENST00000262858	CXorf6	10046	300120
ENSG00000013619	ENST00000262858	CXorf6	728030	300120
ENSG00000013619	ENST00000262858	CXorf6	730818	300120
ENSG00000171100	ENST00000306167	MTM1	4534	300415
ENSG00000147383	ENST00000370274	NSDHL	50814	300275
ENSG00000130821	ENST00000330048	SLC6A8	6535	300036
ENSG00000130821	ENST00000253122	SLC6A8	6535	300036
ENSG00000185825	ENST00000345046	BCAP31	10134	300398
ENSG00000185825	ENST00000370133	BCAP31	10134	300398
ENSG00000101986	ENST00000218104	ABCD1	215	300371
ENSG00000101986	ENST00000218104	ABCD1	642762	300371
ENSG00000198910	ENST00000370060	L1CAM	3897	308840
ENSG00000198910	ENST00000361699	L1CAM	3897	308840
ENSG00000126895	ENST00000358927	AVPR2	554	300538
ENSG00000126895	ENST00000337474	AVPR2	554	300538
ENSG00000169057	ENST00000369964	MECP2	4204	300005
ENSG00000169057	ENST00000303391	MECP2	4204	300005
ENSG00000102076	ENST00000369951	OPN1LW	5956	303900
ENSG00000147380	ENST00000369935	OPN1MW	2652	303800
ENSG00000147380	ENST00000369935	OPN1MW	728458	303800
ENSG00000166160	ENST00000369929	OPN1MW2	2652	303800
ENSG00000166160	ENST00000369929	OPN1MW2	728458	303800
ENSG00000007350	ENST00000369915	TKTL1	8277	300044
ENSG00000196924	ENST00000369850	FLNA		300017
ENSG00000102119	ENST00000369842	EMD	2010	300384
ENSG00000147403	ENST00000369817	RPL10	6134	312173
ENSG00000147403	ENST00000369817	RPL10	647074	312173
ENSG00000147403	ENST00000344746	RPL10	6134	312173
ENSG00000147403	ENST00000344746	RPL10	647074	312173

STEP 13:
To view sequences, go back to ‘Attributes’



STEP 14:
Select ’Sequences’



STEP 15:
Expand the ‘SEQUENCES’ panel and select ‘cDNA’.
Then expand the HEADER (Chromosome, Ensembl Gene ID and Biotype are selected by default).

STEP 16:
Click on ‘Results’.

View all rows as FASTA…

RESULTS

Header: chromosome, Ensembl Gene ID, Biotype

>X|ENSG00000130821|protein_coding

cDNA 1

GCCTCCGCGGGCCCCGGCCGGGGCGGGGGGCGCGGGCCACAGGCCCCTGCTCCGGCCGCC
GCTTGCAGACCGCGGGCGCCGATGTCGCCCGCGCCCCGCTAGGCTGAGCCTCGGGTCGGG
CGAGGAGCCGCCGCAGCCGCCGCCGCCCGAGCCGCGGGCAGGAGCCTCGGGAGCCGCCGC
CGCCGCCGCCGCCGCCCGGCCGGGCCCCGCCGCCGCCCGCGCGCCCCCGGGCCCCCGACA
CACATGAGATTCTTCAGGCTCACTTTCAAGTGCTTCGTGGACTGCTTCTGACTGCGCCGC
CCGCGCCCCGCACCCCGCCGCCCGCCCGCCGCCCCGTCCCCCGGCCCGGCCGCCCCCCGG
CCCCCGGCCGGCCCGCGCCCTCGGGGCCCTCCCCGGTGCCGCCGGTGCCCCCCGCCTGAC
CGCCGCCCCCCGTGAGGCGCCGCGACCCCGGCCCGGCCGTGCGGCCCGCCGAGGCCATGG
CGAAGAAGAGCGCCGAGAACGGCATCTATAGCGTGTCCGGCGACGAGAAGAAGGGCCCCC
TCATCGCGCCCGGGCCCGACGGGGCCCCGGCCAAGGGCGACGGCCCCGTGGGCCTGGGGA
CACCCGGCGGCCGCCTGGCCGTGCCGCCGCGCGAGACCTGGACGCGCCAGATGGACTTCA
TCATGTCGTGCGTGGGCTTCGCCGTGGGCTTGGGCAACGTGTGGCGCTTCCCCTACCTGT
GCTACAAGAACGGCGGAGGTGTGTTCCTTATTCCCTACGTCCTGATCGCCCTGGTTGGAG
GAATCCCCATTTTCTTCTTAGAGATCTCGCTGGGCCAGTTCATGAAGGCCGGCAGCATCA
ATGTCTGGAACATCTGTCCCCTGTTCAAAGGCCTGGGCTACGCCTCCATGGTGATCGTCT
TCTACTGCAACACCTACTACATCATGGTGCTGGCCTGGGGCTTCTATTACCTGGTCAAGT
CCTTTACCACCACGCTGCCCTGGGCCACATGTGGCCACACCTGGAACACTCCCGACTGCG
TGGAGATCTTCCGCCATGAAGACTGTGCCAATGCCAGCCTGGCCAACCTCACCTGTGACC
AGCTTGCTGACCGCCGGTCCCCTGTCATCGAGTTCTGGGAGAACAAAGTCTTGAGGCTGT
CTGGGGGACTGGAGGTGCCAGGGGCCCTCAACTGGGAGGTGACCCTTTGTCTGCTGGCCT
GCTGGGTGCTGGTCTACTTCTGTGTCTGGAAGGGGGTCAAATCCACGGGAAAGATCGTGT
ACTTCACTGCTACATTCCCCTACGTGGTCCTGGTCGTGCTGCTGGTGCGTGGAGTGCTGC
TGCCTGGCGCCCTGGATGGCATCATTTACTATCTCAAGCCTGACTGGTCAAAGCTGGGGT
CCCCTCAGGTGTGGATAGATGCGGGGACCCAGATTTTCTTTTCTTACGCCATTGGCCTGG
GGGCCCTCACAGCCCTGGGCAGCTACAACCGCTTCAACAACAACTGCTACAAGGACGCCA
TCATCCTGGCTCTCATCAACAGTGGGACCAGCTTCTTTGCTGGCTTCGTGGTCTTCTCCA
TCCTGGGCTTCATGGCTGCAGAGCAGGGCGTGCACATCTCCAAGGTGGCAGAGTCAGGGC
CGGGCCTGGCCTTCATCGCCTACCCGCGGGCTGTCACGCTGATGCCAGTGGCCCCACTCT
GGGCTGCCCTGTTCTTCTTCATGCTGTTGCTGCTTGGTCTCGACAGCCAGTTTGTAGGTG
TGGAGGGCTTCATCACCGGCCTCCTCGACCTCCTCCCGGCCTCCTACTACTTCCGTTTCC
AAAGGGAGATCTCTGTGGCCCTCTGTTGTGCCCTCTGCTTTGTCATCGATCTCTCCATGG

>X|ENSG00000155966|protein_coding

cDNA 2

CGCCGCCTGTGCAGCCGCTGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGC
CGCCGCTGCCGCCCCGGCTGCCGCGCCGCGCCGCTGCCTCTGCCCCGGCCGCCCCCGCCG
CCGCTGCCGCCGCCGGCCCGCAGCCAGCCAGGCGGGCGGCCCAGCCCGCCTGAGCCCGCA
GCGGCTGCCGCCGCAGCGTCGGGTCGCTGGGTGCGCGGGCTACCGCGGACCGAGCGGACC
CGAGTGGGCGACCAGGCGCTTGCCCGCCCAGTGCCACTGCCGCCGCTTCCTCGCCGGAGC
ACAGGACCAGACACCTCCAGCGCCCGCTGCTGCTGCCGATGCGGCCCGGACACTTTTAGC
TGGGCGGGAGGGCTGGAGAGCCGGGGGCCGCCGAGAACCGCCAGCGAGCTGTGCCGAGAG
CCGCGCCGACCCGCTGCGATCAGGGACAGGCGCCCGCCCGCCGCCGCCGCCTGGCCGCTA
TGGATCTATTCGACTTTTTCAGAGACTGGGACTTGGAGCAGCAGTGTCACTATGAACAAG
ACCGTAGTGCACTTAAAAAAAGGGAATGGGAGCGGAGGAATCAAGAAGTCCAGCAAGAAG
ACGATCTCTTTTCTTCAGGCTTTGATCTTTTTGGGGAGCCATACAAGGTAGCTGAATATA
CAAACAAAGGTGATGCACTTGCCAACCGAGTCCAGAACACGCTTGGAAACTATGATGAAA
TGAAGAATTTGCTAACTAACCATTCTAATCAGAATCACCTAGTGGGAATTCCAAAGAATT
CTGTGCCCCAGAATCCCAACAACAAAAATGAACCAAGCTTTTTTCCAGAACAAAAGAACA
GAATAATTCCACCTCACCAGGATAATACCCATCCTTCAGCACCAATGCCTCCACCTTCTG
TTGTGATACTGAATTCAACTCTAATACACAGCAACAGAAAATCAAAACCTGAGTGGTCAC
GTGATAGTCATAACCCTAGCACTGTACTGGCAAGCCAGGCCAGTGGTCAGCCAAACAAGA
TGCAGACTTTGACACAGGACCAGTCTCAAGCCAAACTGGAAGACTTCTTTGTCTACCCAG
CTGAACAGCCCCAGATTGGAGAAGTTGAAGAGTCAAACCCATCTGCAAAGGAAGACAGTA

V) BIOMART - Exercises

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 47'
- 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 updated.

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

Click on '**Results**' (at the top) to obtain the first 10 rows of your table. To obtain the entire table select 'View all rows as HTML' or export a file by clicking 'Go'.

Note that the output for this query gives you one row for each SNP, and if there are alternative transcripts then SNP data is given for each. This means that a particular SNP may appear more than once.

Find the coding SNPs, and note that you have information about the effect of the SNP, and its location within the protein. Synonymous status is 'yes' for

silent mutations. Two amino acids will be shown in the 'Peptide Shift' column if there are two alleles on the protein level. The Peptide location (aa), Synonymous Status and Peptide Shift will all be blank if the SNP is not in a coding region.

***Click 'New' to start a fresh topic**

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

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

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

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

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

8. Design your own query!

Answers (BioMart)

1. You should find ***one*** novel gene on chromosome 2 with this InterPro domain. (*Note: there can be more than one gene with one InterPro domain*). The result set has one transcript and a total of 261 rows of output (to see this, change the option from TSV to XLS under 'Export all results' and click 'Go', then open in Excel so you don't have to count the rows manually). The transcript has 8 coding SNPs ('Location in Gene' is 'coding'), most of which are non-synonymous ('Synonymous status' is 'no') and thus affect the amino acid sequence of the encoded peptide. One allele is a stop codon- can you find it?

2. Database and dataset: 'Ensembl 47' and 'Mus musculus genes (NCBIM36)'.

Filters: GENE 'ID list limit Ensembl Gene ID(s)': enter the mouse gene ID.

Attributes 'Structures': select in the **EXON** panel: 'Ensembl Exon ID', 'Exon

Start' and 'Exon End'.

Click '**Results**'.

You should find ***7 exons***. Take the link from the Ensembl Gene ID in your output back to the **GeneView** page to confirm the BioMart data with the gene structure displayed on this page.

3. Database and dataset: 'Ensembl 47' and 'Homo sapiens genes (NCBI36)'.

Filters: REGION 'Chromosome 1', 'Band Start p11.2, 'Band End q22' ,
GENE: 'with Disease Association Only' (look under 'ID LIST FILTERS')

Attributes: Features: select 'GO ID' and 'GO description' along with the default options ('Ensembl Gene ID' and 'Transcript ID').

Results should show ***17 Ensembl genes*** (multiple transcripts and GO terms).

4. Database and dataset: 'Ensembl 47' and 'Homo sapiens genes (NCBI36)'.

Filters: GENE: 'ID list limit': Affy hg u133 plus 2 ID(s) and enter the list of probeset IDs.

Attributes: 'Sequences' select 'Flank (Transcript)', 'Upstream flank 500'. In the header, apart from the already default selected options, select 'Ensembl Transcript ID'.

You should find upstream sequences for the transcripts of ***31 genes*** (Hint: click 'count' to see the number of genes!)

5. Database and dataset: 'Ensembl 47' and 'Homo sapiens genes (NCBI36)'.

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

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

You should find sequences for ***26 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; affy GeneChip Mapping Array)'.

Filters: REGION: 'Chromosome 6', 'Base pair Start 15000000', 'Base pair End 15200000'
GENERAL SNP FILTERS: SNP source: 'SNPs with TSC ID(s) Only'.

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

You should find ***69 SNPs***.

7. Database: 'Ensembl 47' **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' and 'Mouse External ID'. Also select 'Ensembl gene ID' and Transcript ID (default options) and 'Description' in the '**GENE**' panel (these will be for the starting dataset… i.e. Human.)

Results displays the mouse orthologues of the 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?

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 and **ExonView** (see exercise 2).

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

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? Which support the UTRs?

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

Click on the 'Gene information' link in the left-hand menu to return to the **GeneView** 'Gene Report' for 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 Vega External genes and Genscans are turned on.

Look at the Genscan track. Zoom out further.

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

Q4: What is the difference between Vega Havana and Vega External?

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 (note that this mapping is done *after* the genes have been built). These entries are given in the 'Similarity Matches' section of **TransView** (repeated in the 'Transcript' section of **GeneView**). Have a look at the types of databases linked out to.

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

Answers (Evaluating Genes and Transcripts).

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

Q2: 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.

Q3: 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.

Q4: Havana is a subgroup of VEGA. Havana is based at the Sanger, while VEGA is an international consortium of many groups. The Havana track only shows genes built by the subgroup, which 'Vega External' shows the entire Vega set. Only Havana transcripts are merged with Ensembl predictions if they match up (leading to golden transcripts). For more, go to these links:

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

Q5: 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. Main exercise: Genes from mouse, rat, zebrafish and human in the sorting nexin family.

This exercise focuses on protein families and orthologies, using SNX5, a sorting nexin, as an example gene.

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

(b) Examine the protein family.

-Take the link to the associated Protein Family.

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

Q2: Are they all 'known' genes?

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

Q4: What about invertebrate species?

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

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

Find your way to mouse **FamilyView**, and follow the link to mouse Snx5 or Snx6 (**GeneView**).

Have a look at the section 'Orthologue Prediction'. Follow the link to human SNX5, which takes you back to where you started, or SNX6.

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

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

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

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

The conserved regions are probably showing “ungrouped” (a red ‘+’ shows to the left of the track label). This indicates a pink block could be on a different chromosome and/or strand than its neighbour. Click on the red ‘+’ to the left of the ‘Mm blastz’ track: **ContigView** will reload, a red ‘-’ replaces the ‘+’ and the hits are now “grouped” into chromosomes and strands. Note that clicking on the track produces a pop-up with details of and a link to that region in mouse. The menu options change depending on whether the track is grouped or ungrouped.

Make sure the red ‘+’ is shown at the left of the rat ‘Rnor blastz’ track. Click on a rat match in this track, and take the link ‘Dotter’ to **DotterView**. Note the dots on the diagonals where exons align. Zoom in to examine a smaller region.

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

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

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

(d) Examine the synteny blocks.

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

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

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

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

Additional exercises:

2. Compare the BRCA2 gene in human and mouse.

A comparison of a gene involves in breast cancer between two species.

Find the human BRCA2 gene.
Identify its orthologue in mouse.
Display the human gene in **ContigView**, and examine the 'mouse BlastZ' track for the region around the human gene.
Compare the two genes with respect to length and number of exons using **MultiContigView** and **AlignSliceView**. You can reach these pages from **ContigView** using the 'View alongside .' and 'View alignment with …' links, respectively. Note the assembly can be altered (gaps introduced) in **AlignSliceView**, while **MultiContigView** conserves the assemblies and gene distances as is.
View the regions of conserved synteny that include the genes using **SyntenyView**.
Have a look at the **GeneTreeView** page for human BRCA2.

3. Protein Family Exercise

Exploring a protein family and exporting alignments.

Have a look at the protein family ENSF00000000756

Q7: How many genes in the family are there in human?

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

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

4. Browse the human-mouse synteny blocks

Looking at conserved sequence on a larger scale between two species.

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

5. Exploring your own gene of interest

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

6. Exploring your own region of interest

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

Answers (Comparative Genomics)

Main Exercise (1)

Q1: In **FamilyView** for Family ID ENSF00000001822 (Sorting Nexins) you may see there are 3 genes producing peptides in this family.

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

Q3: Yes. There are 4 in *Mus musculus* (mouse), 6 in *Rattus norvegicus* (rat) and 3 in *Danio rerio* (zebrafish).

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

Q5: 4 genes encode peptides that are part of the family (one arrow represents 2 genes very near each other). Genes can have more than one transcript (leading to 6 peptides in the sorting nexin family).

Q6: Chromosome 20, band p11.23 *(hint: go to **ContigView**)*

Additional Exercises (2)

Q7: *Hint: you can search Ensembl for the protein family ID, and go to **FamilyView** from species-specific **ProtView** pages.*
Human: 7 genes

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.

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

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 M470V and T854T are studied. Can you find these two SNPs in Ensembl? Note, the positions refer to numbers in the amino acid sequence.

3. Display all SNPs for one gene.

To display SNPs for a gene, use GeneSNPView or BioMART

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

Answers (Variations)

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

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

non-synonymous coding SNP in this region is: rs28513898), and you can also follow the link ‘SNP properties’ to its **SNPView** page.

From the SNPView page, click on ‘ SNPs in gene context’ at the right of the second panel. Under ‘SNP type’ in the GeneSNPView page, select only ‘non-synonymous’. This displays all the SNPs for a gene that change the amino acid sequence.

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

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. Note that you can use the ‘SNP class’ and ‘SNP type’ drop-down menus in the figure to configure both figure and table.

3. Go to BioMart. Select Ensembl Gene 47 and *Homo sapiens* genes. In **Filters**, in the GENE section, enter either the HGNC symbol (CFTR) or Ensembl Gene ID (ENSG00000001626) under ‘ID list limit’.
Go to ‘**Attributes**’ and select the SNPs page. Under ‘GENE ASSOCIATED SNPs’ select the options ‘Reference ID’, ‘Allele’ and ‘Location in Gene (coding etc)’. Click results and export as an excel file. This gene contains 953 validated SNPs of which 82 are coding.

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

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. Are there any on mouse chromosome 1 (of any type), and where are they? 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 following human genes using Ensembl?

   Gene1: IRX4
   Gene 2: ROCK1

   Export the sequence of ROCK1 and use BLAST, SSAHA, or BLAT to find it again in the human genome.

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) ***Hint: Use BioMart…to find all the RNA genes it will take 2 queries as the filters only allow a subset to be selected. There are 183 total RNA genes on chromosome 1.***

   Use **BioMart** to obtain the RNA genes on chromosome 1, along with type and base pair location.

   From the result table, follow the link to one of these genes into the Ensembl browser (for example, ENSMUSG00000070103). Go to

'**ContigView**' (the link is 'graphical view') and look under the 'features' menu: select '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) **IRX4 and ROCK1:** *Optional: view the UTR in TransView or ExonView (click on transcript information or Exon information to reach these pages)*. Export the UTR information by clicking 'Export gene information' 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/

   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 binding sites, 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 next menu, select Takifugu and Tetraodon 'translated BLAT' tracks to view conserved regions. The conserved region upstream to the gene might be worth exporting to search for regulatory regions.

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'**. 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 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). Click 'Next' and select **ContigView** as the page to display the SNPs (note you can select more than one page). Name the track if you'd like, and once you reach the DAS sources list (end page) refresh contigview… your 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

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