Access to genes and genomes with Ensembl

Course Manual

October/November 2009 (e56)

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 laboratory biologist when provided along with quality annotation of the genomic sequence.

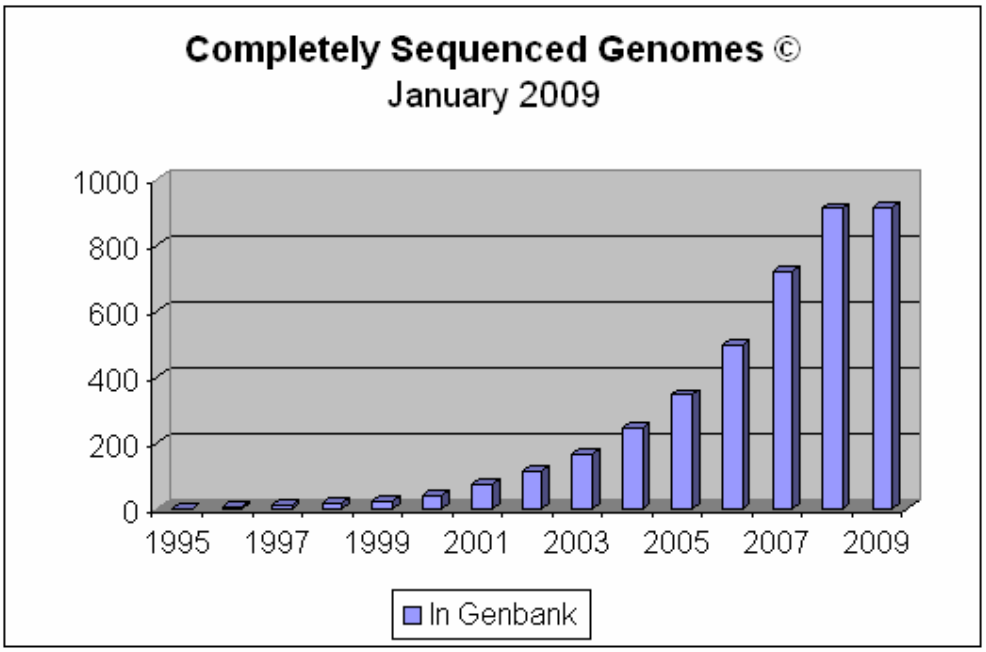


Figure 1. Completely sequenced genomes as of January, 2009 (figure taken from http://www.genomesonline.org).

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 access to genomic information with a number of visualisation tools. The Ensembl website gives you the possibility to directly download data, whether it is the 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 covered in the web-site walk-through. An updated version of the website is released bimonthly. Old versions are accessible on the 'Archive!' website, dating back two years. 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

Hubbard, T.J.P. *et al.*
Ensembl 2009
Nucleic Acids Res. Jan 2009 37: D690-697 *(database issue)*

Vilella, A.J. *et. al*
EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates.
Genome Res. 2009 Feb 19(2):327-35

Smedley, D. *et al*
BioMart – biological queries made easy
BMC Genomics 2009 Jan 14;10:22

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*.[1]
An Overview of Ensembl.
Genome Research 14(5): 925-928 (2004)

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

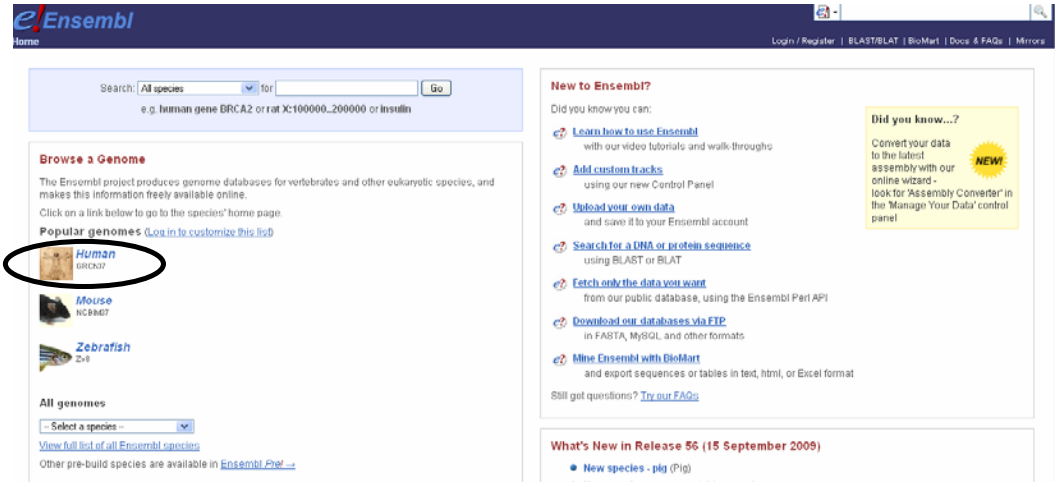
[1] This paper was part of the May 2004 issue of *Genome Research* which included an Ensembl special covering detailed aspects of the Ensembl web site, the underlying scalable database system for storing genome sequence and annotation information, as well as the automated genome analysis and annotation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

II) WALKING THROUGH THE WEBSITE

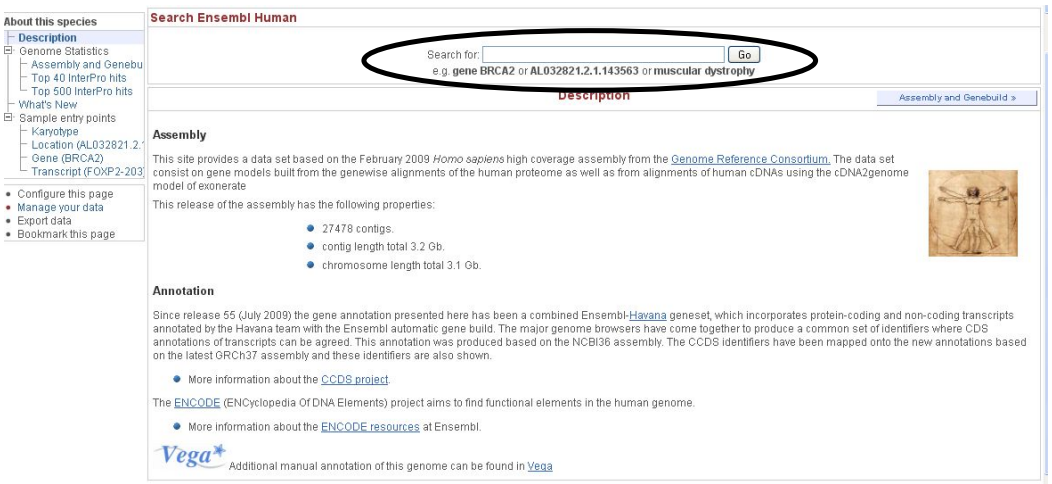
The instructor will guide you through the website using the human *rhodopsin* **(RHO)** gene. The following points will be addressed:

- **The Gene Summary tab and gene-related links:**
  - Are there splice variants?
  - Can I view the genomic sequence with variations?
  - Find orthologues and paralogues
- **The Transcript tab and related links:**
  - What is the protein sequence?
  - What matching proteins and mRNAs are found in other databases?
  - Gene Ontology
- **The Location tab and related links:**
  - What's the conservation track?
  - How do I zoom in and change the gene focus.
  - Un-stacking a track (e.g. human cDNAs)
  - Adding a track (i.e. variations)
- **Exporting a sequence and running BLAT/BLAST**

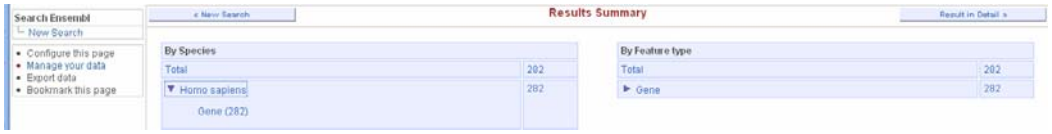
Start by going to ***www.ensembl.org***



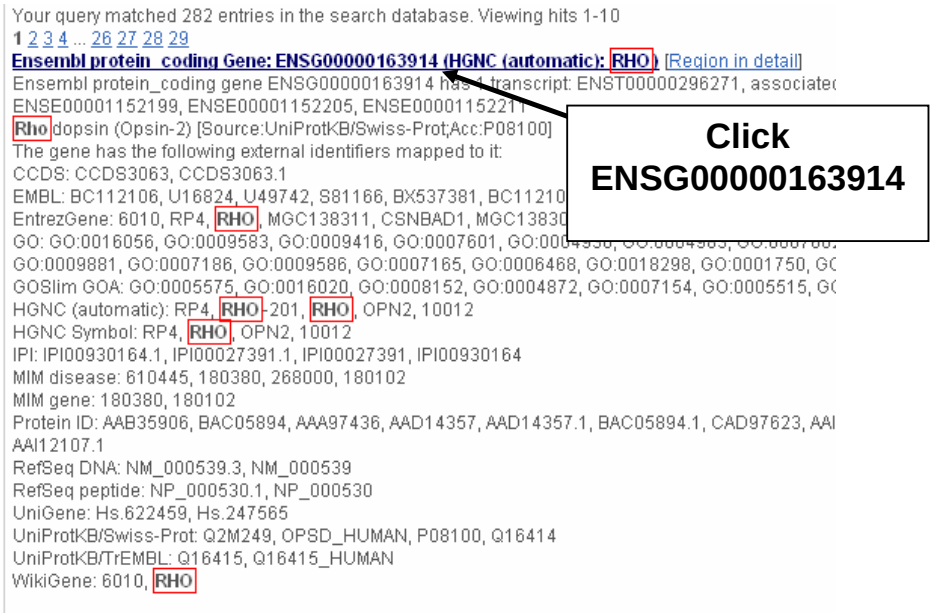
Click on 'Human', or the picture circled above, which brings us to the species home page.



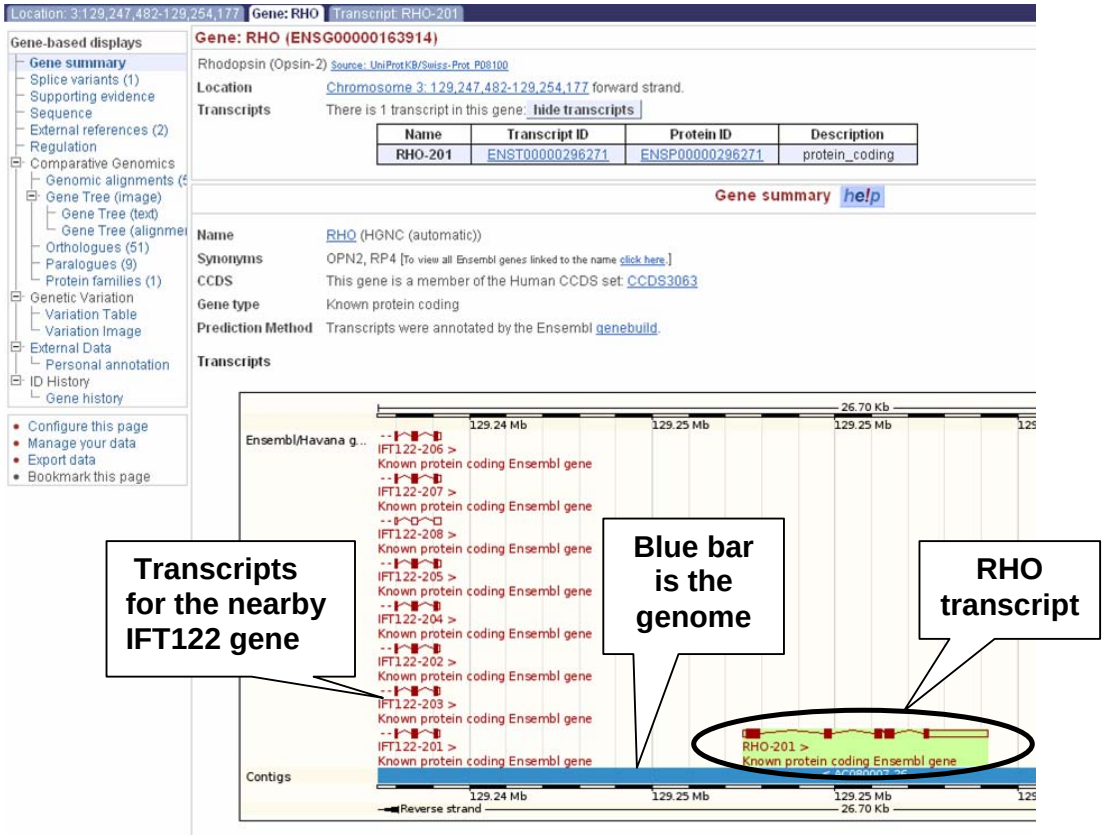
Type 'gene RHO' into the search bar circled above and click the 'Go' button.



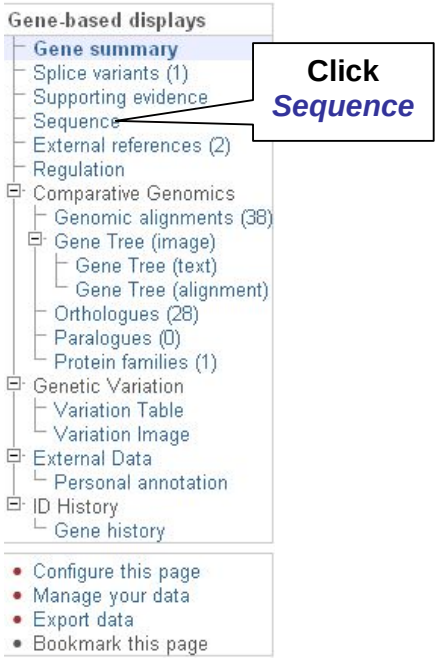
Click the arrow next to Homo sapiens to expand the hits, and click the 'Gene' link when it is revealed.



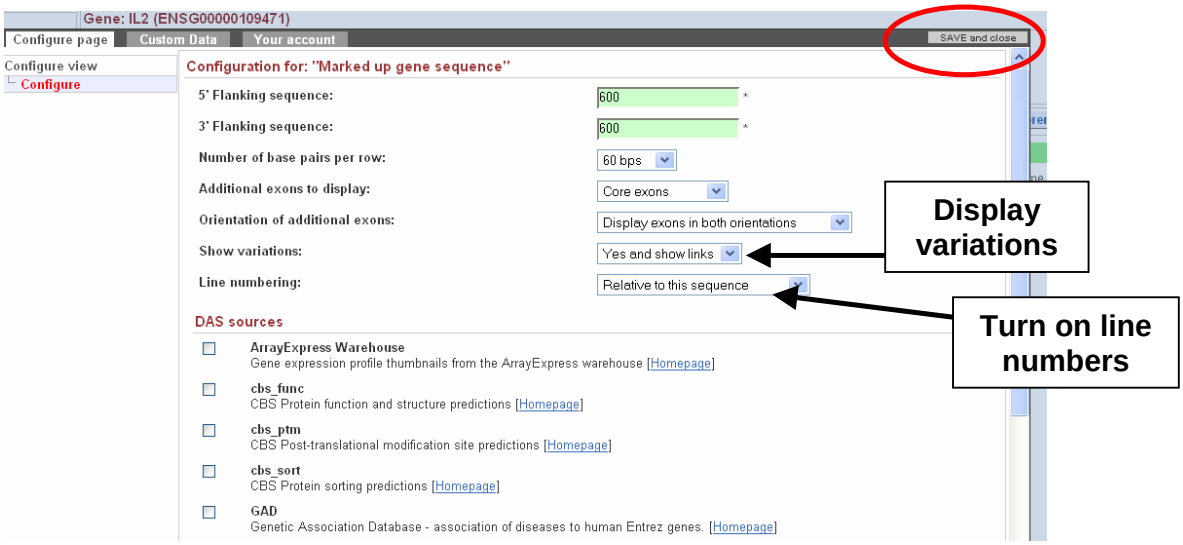
Look through the search results for RHO, the gene symbol. Select the Ensembl gene (i.e. ENSG00000163914), rather than manually curated Havana genes. The following 'Gene' tab should open:



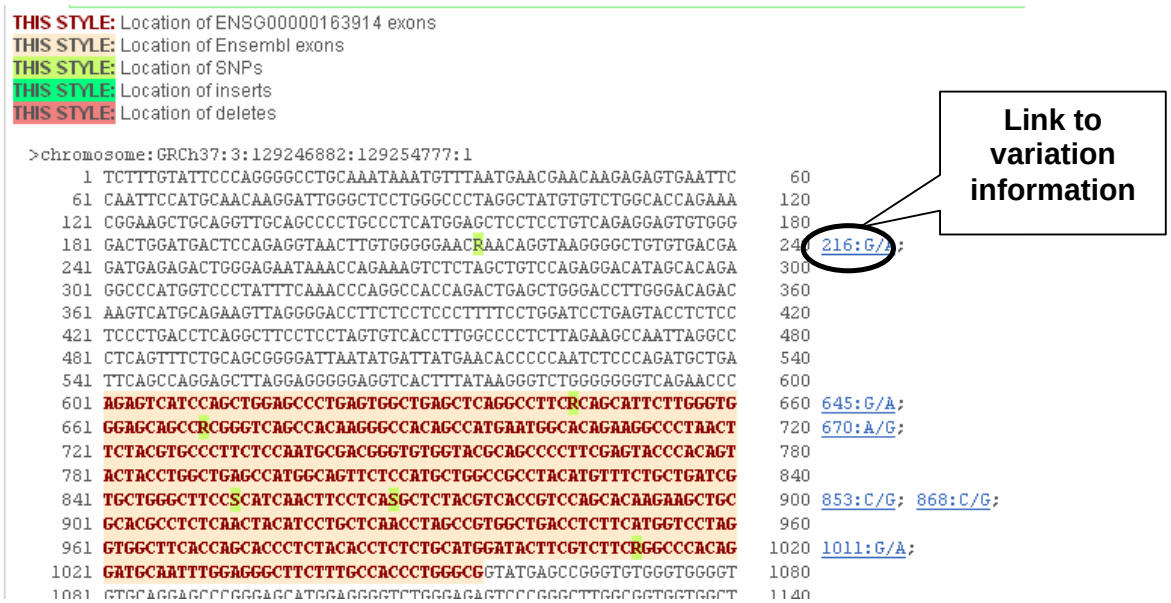
Let's walk through some of the links in the left hand navigation column. How can we view the genomic sequence? Click *Sequence* at the left of the page.



Exons are highlighted within the genomic sequence. Variations can be added with the *Configure this page* link found at the left. Click on *Configure this page* now.



Once you have selected changes (in this example, we display variations and show line numbers) click *Save and Close* at the top right (circled in red, above).



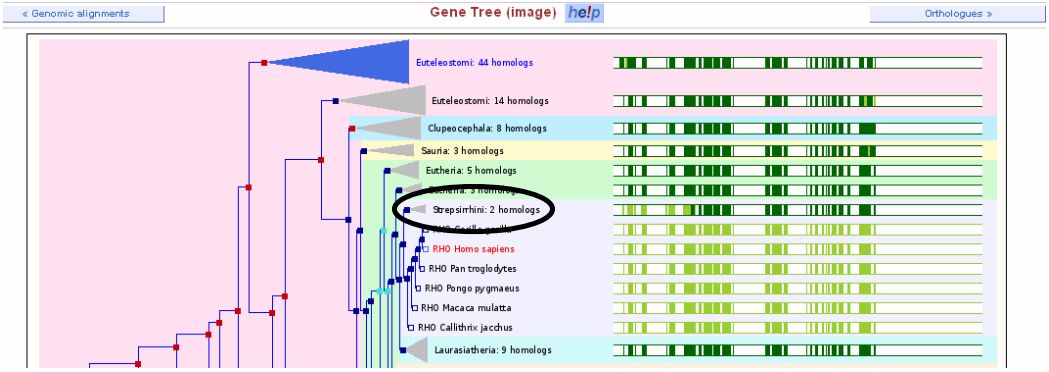
Variations in the sequence are highlighted in green, and represented by the IUPAC code. R in this instance represents alleles A or G. Links to variation pages (one is circled) are shown at the right. Line numbers have been added.

Now let's click on *Genomic alignments*, to see a nucleotide view of the whole genome alignments. Select the 10 eutherian mammals, EPO. The EPO pipeline refers to the programs behind the whole genome alignments - click the '*help*' button for more.

Click *Configure this page* at the left. Turn on conservation regions in the menu.

Exons are highlighted in red, identical nucleotides are highlighted in blue.

Now let's click on *Gene tree (image)*, which will display the current gene in the context of a phylogenetic tree of orthologous and paralogous genes.



Expand the Strepsirrhini subtree (circled above) by clicking on the corresponding node.

Click the *Orthologues* link at the left of this page to view homologues detected in this tree.

« Gene Tree (image)

Orthologues help

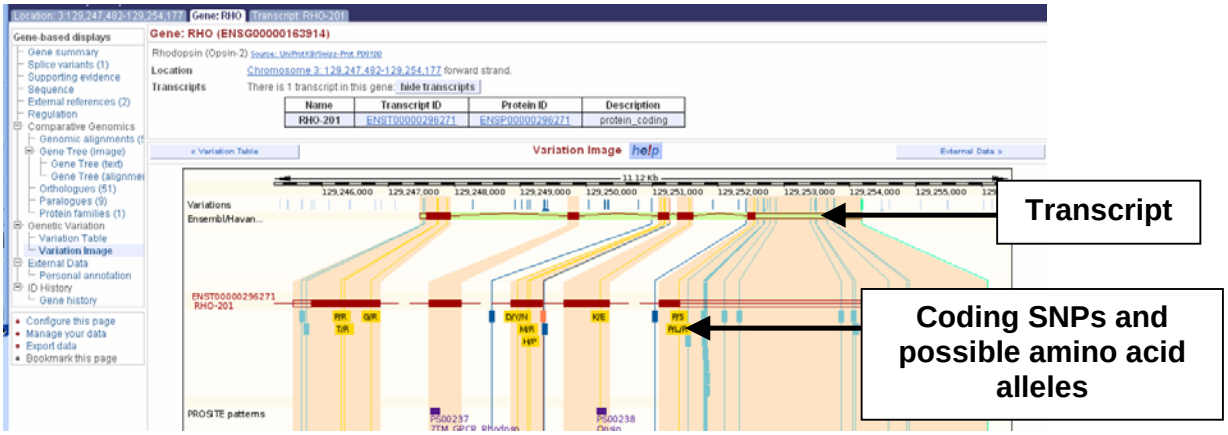
The following gene(s) have been identified as putative orthologues:
(N.B. If you don't find a homologue here, it may be a "between-species paralogue". Please view the gene tree info to see more.)

Species	Type	dN/dS	Ensembl identifier	External ref.
Alpaca (*Vicugna pacos*)	1-to-1	na	ENSVPAG00000003334 Target %id: 77; Query %id: 76 [Multi-species view] [Align]	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: P08100]
Anole Lizard (*Anolis carolinensis*)	1-to-1	na	ENSACAG00000014258 Target %id: 81; Query %id: 82 [Multi-species view] [Align]	OPSD_ANOCA Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: P41591]
Armadillo (*Dasypus novemcinctus*)	1-to-1	na	ENSDNOG00000015133 Target %id: 61; Query %id: 40 [Multi-species view] [Align]	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: P08100]
Bushbaby (*Otolemur garnettii*)	1-to-1	na	ENSOGAG00000010259 Target %id: 78; Query %id: 77 [Multi-species view] [Align]	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: P08100]
Cat (*Felis catus*)	1-to-1	na	ENSFCAG00000000092 Target %id: 96; Query %id: 96 [Multi-species view] [Align]	OPSD_FELCA Rhodopsin. [Source: UniProtKB/Swiss-Prot; acc: Q95KU1]
Chicken (*Gallus gallus*)	1-to-1	na	ENSGALG00000020745 Target %id: 75; Query %id: 50 [Multi-species view] [Align]	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: P08100]
Chimpanzee (*Pan troglodytes*)	1-to-1	0.00000	ENSPTRG00000015379 Target %id: 100; Query %id: 100 [Multi-species view] [Align]	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: P08100]
Cow (*Bos taurus*)	1-to-1	0.03486	ENSBTAG00000001310 Target %id: 92; Query %id: 93 [Multi-species view] [Align]	OPSD_BOVIN Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: P02699]
Dog (*Canis familiaris*)	1-to-1	0.04946	ENSCAFG00000004633 Target %id: 90; Query %id: 95 [Multi-species view] [Align]	OPSD_CANFA Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: P32308]

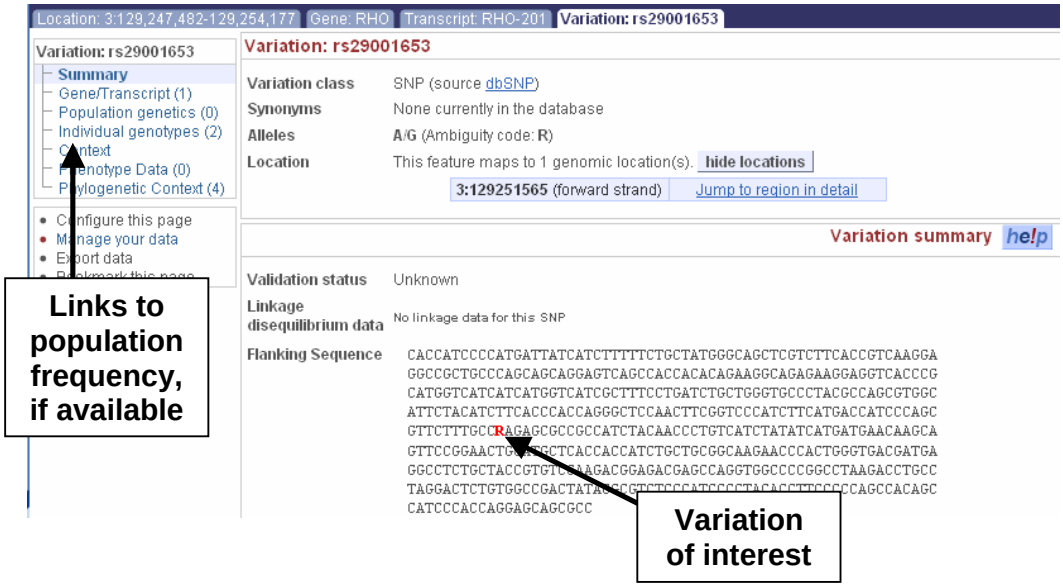
Let's view genetic variation mapped onto all transcripts of a gene.

First click on *Variation table* at the left.

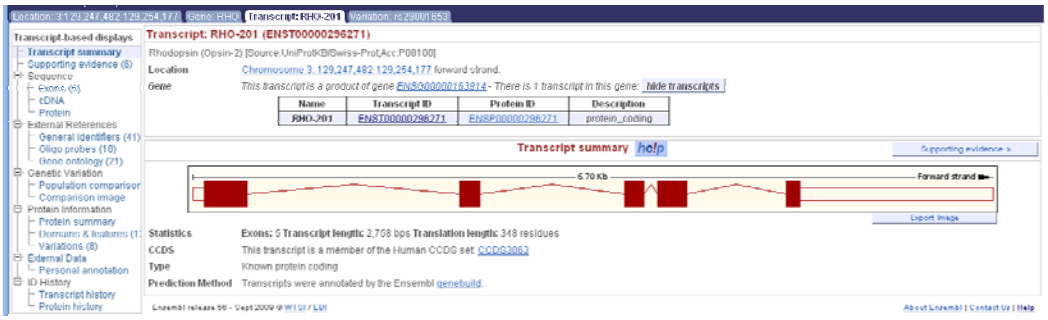
Then click on the *Variation image* (at the left).



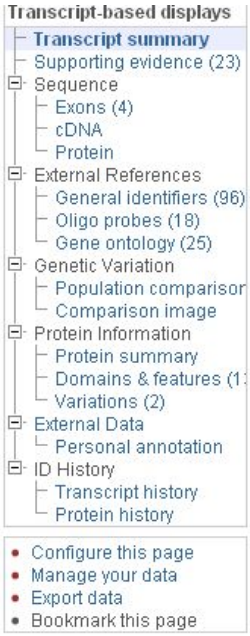
Click any variation, then *Variation properties* to learn more about it. A fourth tab will open:



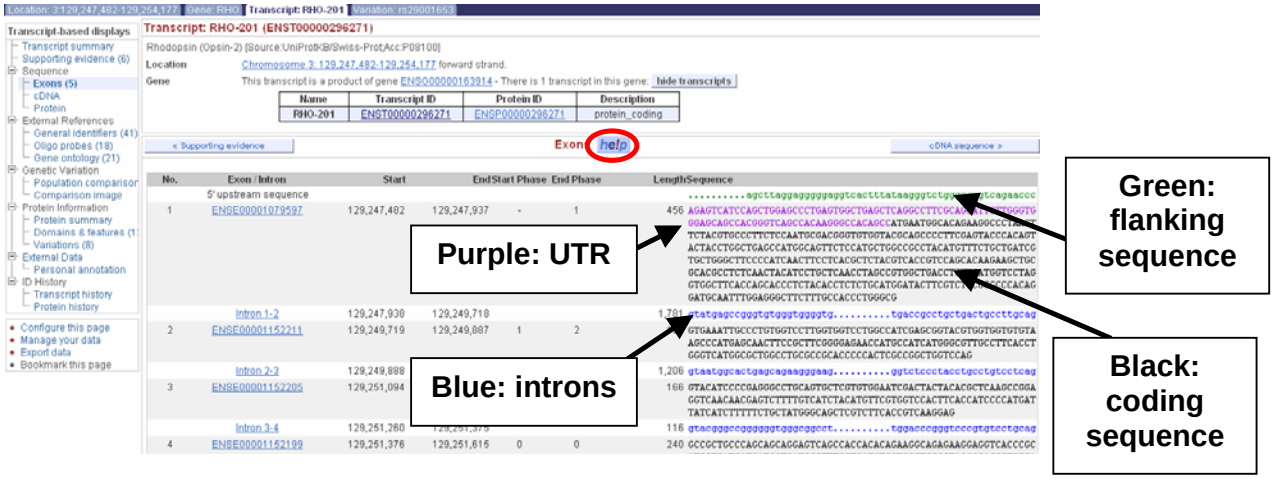
Now, let's focus more closely on the transcript. Select the transcript from the header section by clicking on the *Transcript tab* for RHO. This will lead to the Transcript summary display.



Again, the left hand navigation column provides several options for this particular transcript.



Choose the *Exons* option first, which highlights exon sequences. (The name is misleading … exons are highlighted, but full genomic sequence is shown).

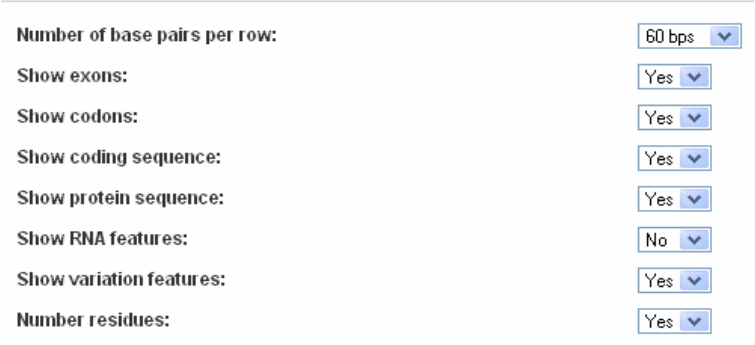


Have you forgotten what the colours mean? No worries- click on the *Help* button (circled in red) and read the help for this page. A link to the glossary is also provided.

You may use the *Configure this page* link to change the display (for example, to show more flanking sequence, or to show full introns).

Click the *cDNA* link to see the spliced transcript sequence. UTR is highlighted in dark yellow, codons are highlighted in light yellow, and exon sequence is shown in black or blue letters to show exon divides.

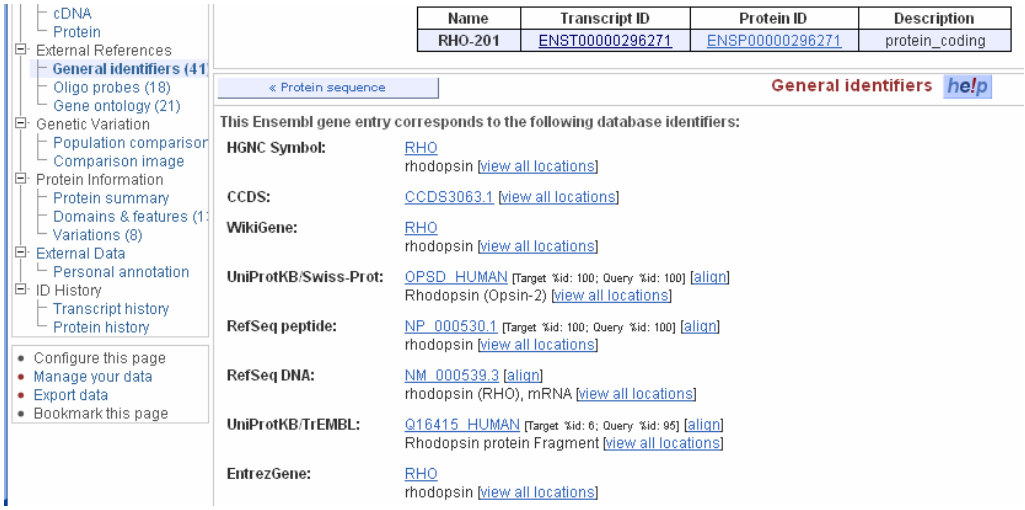
If variations and protein sequence is not shown, turn them on using *Configure this page*. Turn on coding sequence, protein sequence, variation features, and numbering.



Click *Save and close.*

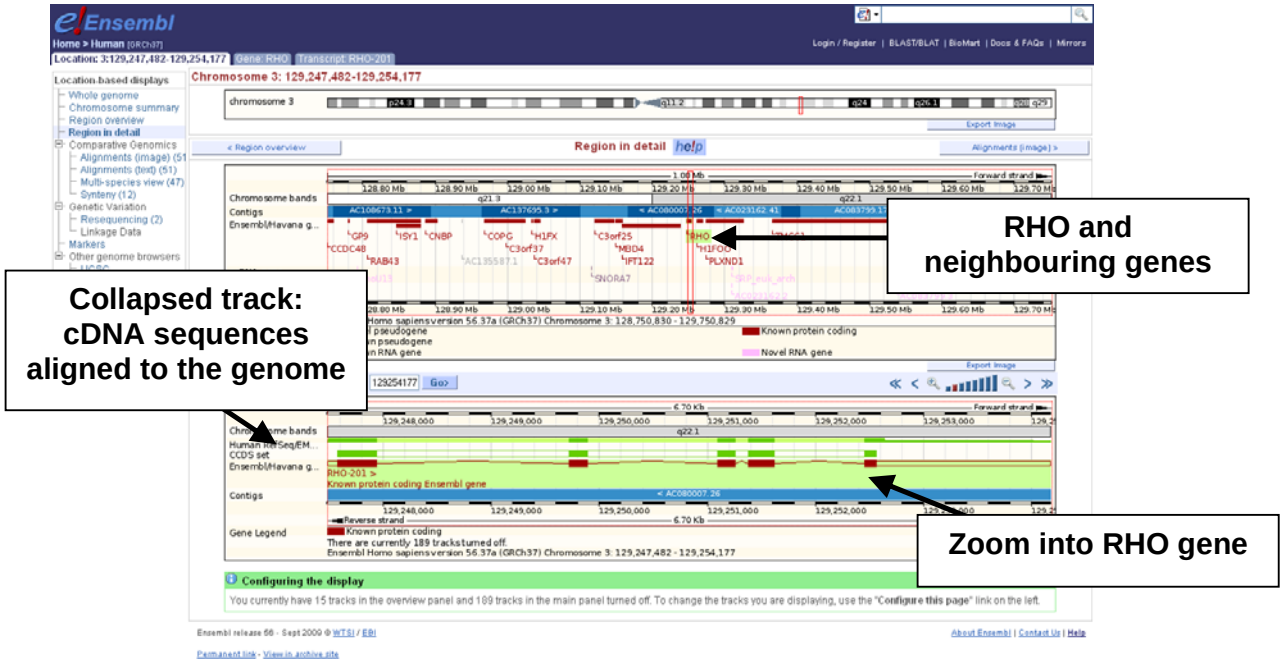
The view should now show variations, represented by highlighted nucleotides, and clickable IUPAC codes above the sequence.

Next, follow the *General identifiers* link at the left, in the *External References* section. The following data display is quite an important one, as it shows sequences in public databases that match to the Ensembl transcript.



Other transcript-specific displays include the cDNA sequence, microarray probes and gene ontology terms from the GO consortium (*www.geneontology.org*).

Let's now view the genomic region in which this gene and its transcript have been annotated by clicking onto the *Location* tab.

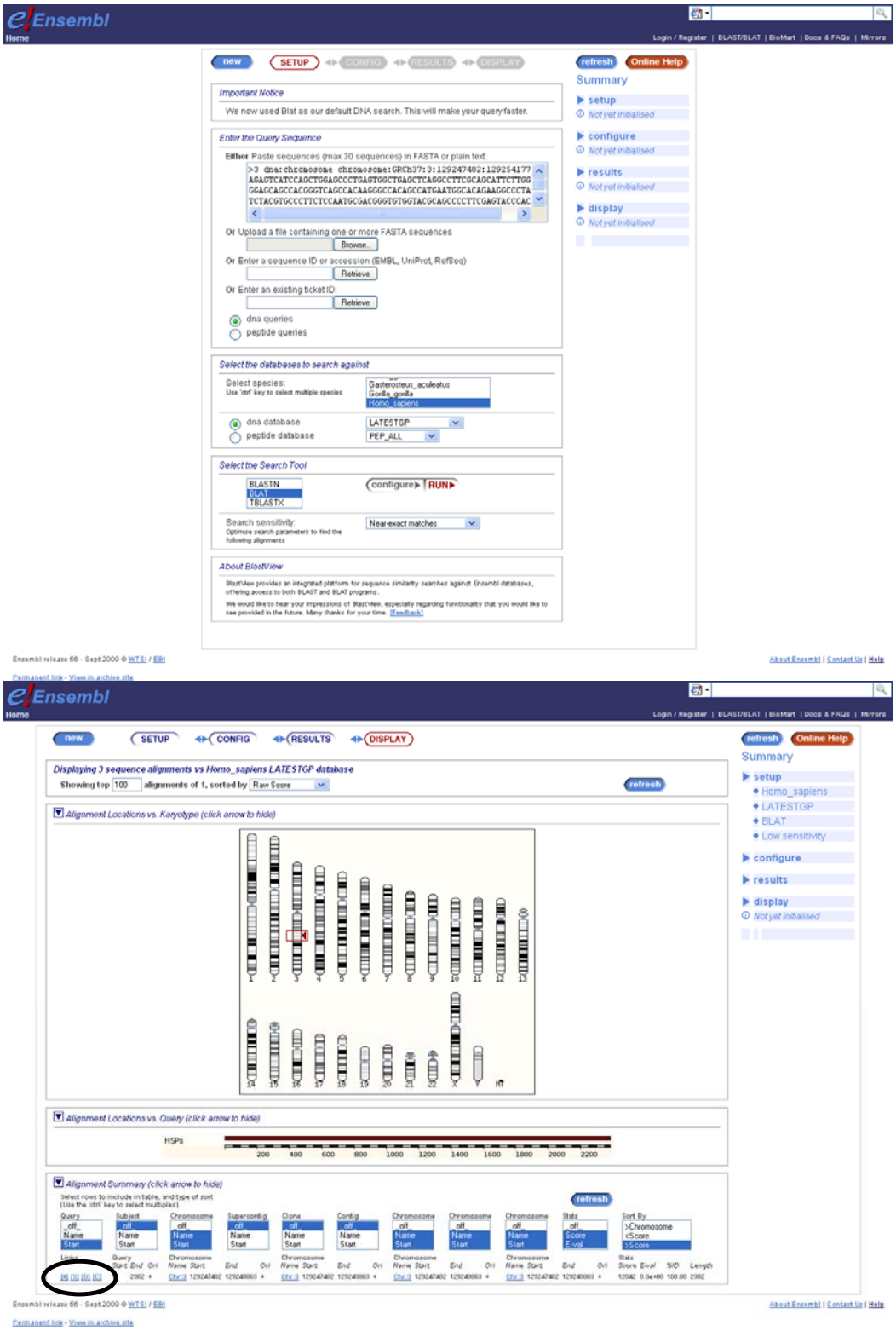


Ensembl *Location* displays are highly configurable. You can switch on additional tracks displaying various feature types that Ensembl annotates in the genome. Use the *Configure this page* link to add *all variations* to the display. Also, in the configuration menu, choose to view the *Human RefSeq/EMBL cDNA* track in normal, expanded form by choosing the *labels* option. Also, find the *Multiple alignments* menu, and then turn on the Conservation score and Constrained elements for 31 eutherian mammals. *Save and close* the menu.

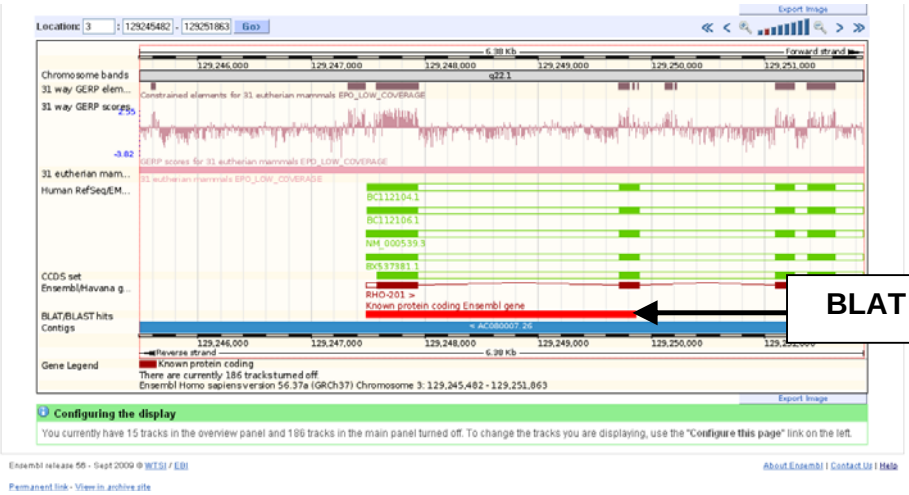
After investigating the *Location display*, we would like to export genomic sequence. Click the *Export location* data option and click *Next.* Now click *HTML*.

```
>3 dna:chromosome chromosome:GRCh37:3:129247482:129254177:1
AGAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCTTCGCAGCATTCTTGGGTG
GGAGCAGCCACGGGTCAGCCACAAGGGCCACAGCCATGAATGGCACAGAAGGCCCTAACT
TCTACGTGCCCTTCTCCAATGCGACGGGTGTGGTACGCAGCCCCTTCGAGTACCCACAGT
ACTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCTACATGTTTCTGCTGATCG
TGCTGGGCTTCCCCATCAACTTCCTCACGCTCTACGTCACCGTCCAGCACAAGAAGCTGC
GCACGCCTCTCAACTACATCCTGCTCAACCTAGCCGTGGCTGACCTCTTCATGGTCCTAG
GTGGCTTCACCAGCACCCTCTACACCTCTCTGCATGGATACTTCGTCTTCGGGCCCACAG
GATGCAATTTGGAGGGCTTCTTTGCCACCCTGGGCGGTATGAGCCGGGTGTGGGTGGGGT
GTGCAGGAGCCCGGGAGCATGGAGGGGTCTGGGAGAGTCCCGGGCTTGGCGGTGGTGGCT
GAGAGGCCTTCTCCCTTCTCCTGTCCTGTCAATGTTATCCAAAGCCCTCATATATTCAGT
CAACAAACACCATTCATGGTGATAGCCGGGCTGCTGTTTGTGCAGGGCTGGCACTGAACA
CTGCCTTGATCTTATTTGGAGCAATATGCGCTTGTCTAATTTCACAGCAAGAAAACTGAG
CTGAGGCTCAAAGAAGTCAAGCGCCCTGCTGGGGCGTCACACAGGGACGGGTGCAGAGTT
GAGTTGGAAGCCCGCATCTATCTCGGGCCATGTTTGCAGCACCAAGCCTCTGTTTCCCTT
GGAGCAGCTGTGCTGAGTCAGACCCAGGCTGGGCACTGAGGGAGAGCTGGGCAAGCCAGA
CCCCTCCTCTCTGGGGGCCCAAGCTCAGGGTGGGAAGTGGATTTTCCATTCTCCAGTCAT
TGGGTCTTCCCTGTGCTGGGCAATGGGCTCGGTCCCCTCTGGCATCCTCTGCCTCCCCTC
TCAGCCCCTGTCCTCAGGTGCCCCTCCAGCCTCCCTGCCGCGTTCCAAGTCTCCTGGTGT
TGAGAACCGCAAGCAGCCGCTCTGAAGCAGTTCCTTTTTGCTTTAGAATAATGTCTTGCA
TTTAACAGGAAAACAGATGGGGTGCTGCAGGGATAACAGATCCCACTTAACAGAGAGGAA
AACTGAGGCAGGGAGAGGGGAAGAGACTCATTTAGGGATGTGGCCAGGCAGCAACAAGAG
CCTAGGTCTCCTGGCTGTGATCCAGGAATATCTCTGCTGAGATGCAGGAGGAGACGCTAG
```

Select the header and a few lines of sequence using Edit/Copy in your browser. Click on the *BLAST/BLAT* link in the bar at the top of the page. Paste the sequence into the appropriate box and select *BLAT* as the search algorithm. Finally, click *Run*.



Follow links (circled above) to an alignment [*A*], the query sequence [*S*], the genome sequence [*G*] and the corresponding Location View [*C*] (for its former name ContigView… or to C (see) the hit!).



Note: you can export the image using the link at the bottom.

END of WORKED EXAMPLE

III) EXERCISES and ANSWERS

Note: the answers to these exercises correspond to current version (56) of Ensembl. If you use these exercises in the future, after Ensembl is updated, please use the archive site for version 56.

http://Sep2009.archive.ensembl.org/index.html

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 Gene Summary page.

(a) Open the home page of Ensembl (www.ensembl.org) (or click on the big 'e!' from the top left corner of any Ensembl page.) Search for the human TAC1 gene by typing 'human gene TAC1' in the search window.

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

(c) Follow some of the links in the 'General Identifiers' section of one of the Transcript tabs. What information can be found about the transcript? View any GO (Gene Ontology) terms.

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

(e) In which chromosomal contig and base pair position in the genomic sequence assembly is the TAC1 gene located?

(f) Is there a putative zebrafish orthologue? If so, where is it in the zebrafish genome?

2. Exploring the Cat Genome (a 2X assembly)

(a) Who sequenced the cat assembly?

(b) View the top 40 InterPro hits, showing common protein domains found in cat genes. What's the first hit?

(c) Search for the gene named OPSD_FELCA. Follow the link showing the gene ID. Is this gene on one or more contigs?

(d) Click on the 'Transcript' tab. View the 'Exons' page. How many exons form this transcript?

(e) On what evidence is this transcript based? What gene and protein IDs from other databases match to the Ensembl transcript? *(Hint… look under 'Supporting evidence' and then 'External References' in the left hand menu.)*

(f) Click the Location tab. When comparing the genome sequence to human, how well are the exons conserved in cat? How about the introns? Draw the BLASTz track onto this page, view the alignments (text) page and also multi-species view.

Answers (Browsing Ensembl)

Answer 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 Gene tab.

(b) The TAC1 gene (ENSG00000006128) has 6 predicted transcripts. Only 4 of these are protein coding. They are: ENST00000319273, ENST00000414974, ENST00000350485, and ENST00000346867. Click on each ENST… identifier for more information about these transcripts. The longest transcript is ENST00000319273. See this information in the heading of each 'Transcript Summary' page. The length of ENST00000319273 is 1239 bp. It has 7 exons and codes for 129 aa.

(c) Click on 'General identifiers' at the left of the transcript page. Click on the HGNC symbol (TAC1) to view the gene name in other databases. Go back to the general identifiers page. Click on the UniProtKB/Swiss-Prot record. Read about the protein, then come back to Ensembl general identifiers and follow the link to MIM.

The GO (Gene Ontology) section can give 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 behavioural responses and function as vasodilators and secretagogues.

(d) Click the 'Domains and Features' link from the transcript tab. The domains include IPR008216 (Protachykinin), IPR013055 (Tachykinin/Neurokinin like), IPR002040 (Tachykinin/Neurokinin), and IPR008215 (Tachykinin). These

also relate to domains in the original database such as Prints, Pfam and ProSite.

(e) Back on the 'Gene' tab and the 'Gene Summary' page, the location is shown as: Chromosome 7: 97,361,220-97,369,878 bp. In the diagram you can see that the gene is located on contig AC004140.2, which marks the position in the genomic assembly. (Not sure what a contig is? *Click on the Help button, then click the link to the Ensembl glossary.*

(f) From the Gene tab, the 'Orthologues' link should show *Danio rerio* ENSDARG00000014490 as an orthologue. Click on it to go to its 'Gene Summary' page to find that it is located on z-fish chromosome 19: 24,483,955-24,487,571.

Answer 2. Exploring genes in Cat (a draft genome)

(a) Select 'Cat' under 'All genomes' to see the genome sequence was determined by the Broad Institute, in conjunction with Agencourt Bioscience. A '2X' assembly is not a thoroughly sequenced genome, and is characterised by gaps, and supercontigs or scaffolds (long stretches of sequence) rather than chromosomes.

(b) Click on 'Top 40 InterPro hits' at the left. This will show a list of common protein domains based on a genome wide study. The first match is to a Proline-rich region. Take the IPR000694 link to read more about it on the InterPro website.

(c) One transcript is shown in the Gene summary tab. The gene is projected from human, and the gene name is OPSD_FELCA, a name taken from the UniProt/Swiss-Prot record. The gene goes across two contigs, numbers 578991 and 578990. Gaps between contigs signify that the order of contigs in the genomic assembly may not be well understood. The 2x (low coverage) cat genome was aligned to the human genome before the annotation build, and because of some lack of genome sequence represented by the gaps, the full OPSD_FELCA gene could not be annotated.

(d) Click the Transcript tab. You might already see there are 5 exons, based on the transcript structure. (Remember, boxes are exons, and lines connecting them are introns). Click on the 'Exons' link under 'Sequence' at the left. The sequence is shown with exons in capital letters, coding sequence in black, and any untranslated region (UTR) on the exons in purple. *(For this example, there are no UTRs.)*

Did you know? Read the Glossary by clicking on 'Help' and following the link at the left.

(e) To see which human protein supported this transcript, click on the 'Supporting evidence' link at the left. All cat genes come from alignments of human coding sequences. In this case, the gene is built on Ensembl Human

Peptide ENSP00000296271. This was the evidence used *at the time of the cat genebuild.*

To see cDNA and protein information currently matching to the Ensembl cat transcript, click on 'General identifiers' in the left hand menu. There are currently 6 matching DNA or protein entries. There is even a good match (99%) to a cat protein in UniProtKB/Swiss-Prot!

(f) Look at a comparison with the human genome by clicking on '*Configure this page*' at the left of the Region in Detail view. Add the human-cat BLASTZ alignment, and SAVE and close the menu. You might see that introns were included in the alignments.

We can also view these alignments by clicking on 'Alignments (text)' at the left. Select the 'Human blastz' alignment at the top of the page. Exons are highlighted in red. Click on 'Configure this page' at the left and turn on 'All conserved regions' in the 'Conservation regions' option. Some of the introns do show conservation with the human sequence.

Finally, click on 'Multi-species view' at the left. This allows a graphical comparison of the human and cat genomes. Aligned regions are shown by pink bars, and connected by green shading. The RHO gene in human is homologous to the cat gene. (Hint: Select 'join genes' in the 'comparative features' menu of the 'configure this page' panel.)

IV) BioMart

Mining data- worked example

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

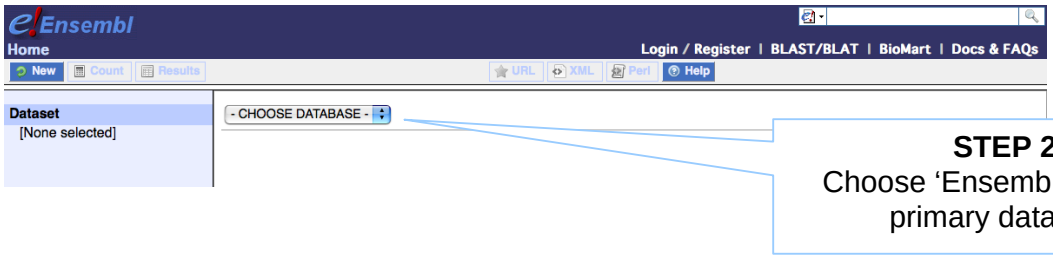
Which other genes with consensus coding sequences assigned by the CCDS project locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs?

What are their cDNA sequences?

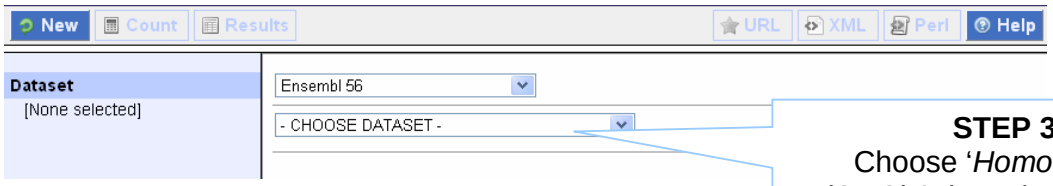
Follow the worked example below to answer these questions.

Step 1: Either click on 'BioMart' in the top right header bar of the Ensembl home page, or go to http://www.biomart.org/ and click on the 'MartView' tab.

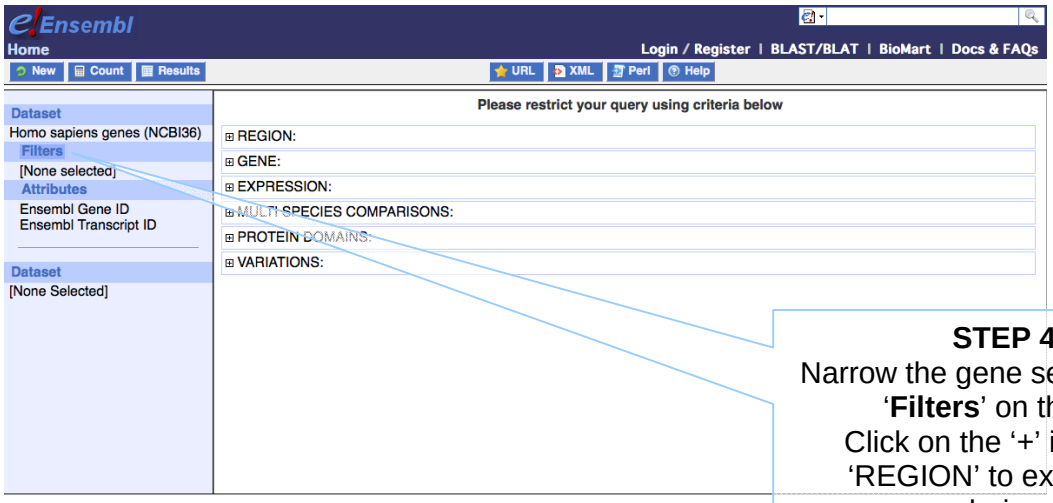
NOTE: These answers were determined using Ensembl 56.



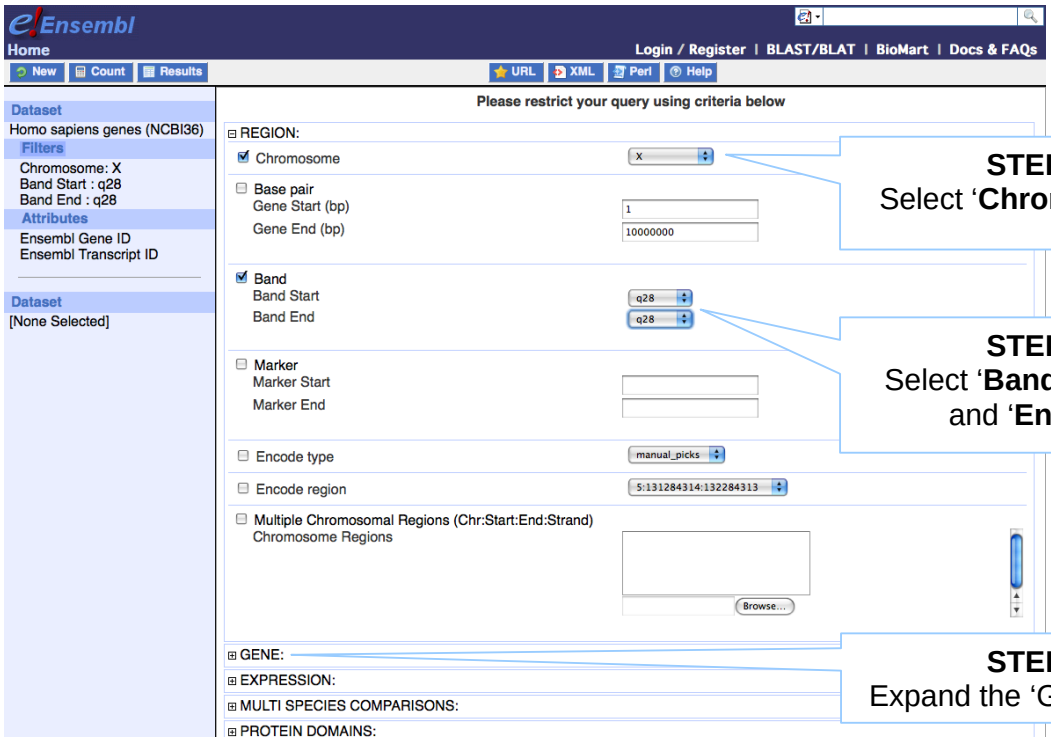
STEP 2:
Choose 'Ensembl 56' as the primary database.



STEP 3:
Choose '*Homo sapiens (GRCh37)*' as the species of interest.



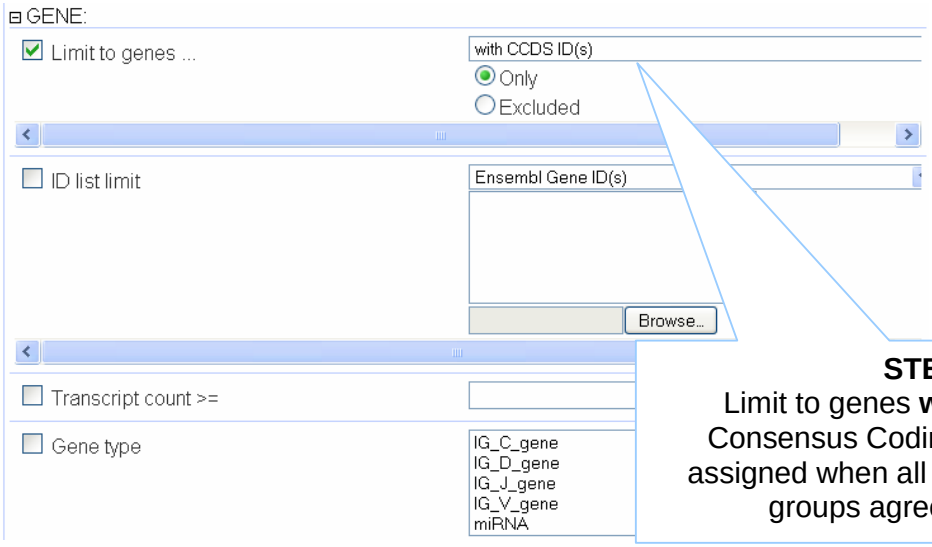
STEP 4:
Narrow the gene set by clicking '**Filters**' on the left.
Click on the '+' in front of 'REGION' to expand the choices.



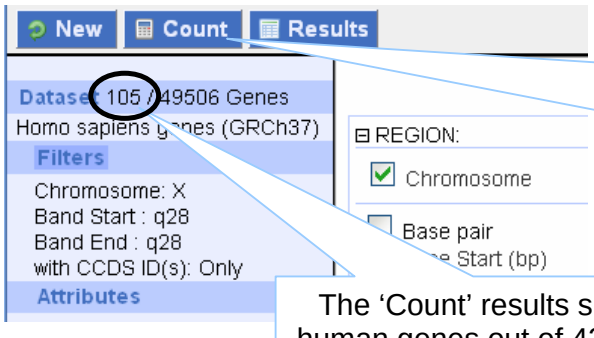
STEP 5:
Select '**Chromosome X**'

STEP 6:
Select '**Band Start q28**' and '**End q28**'

STEP 7:
Expand the 'GENE' panel.

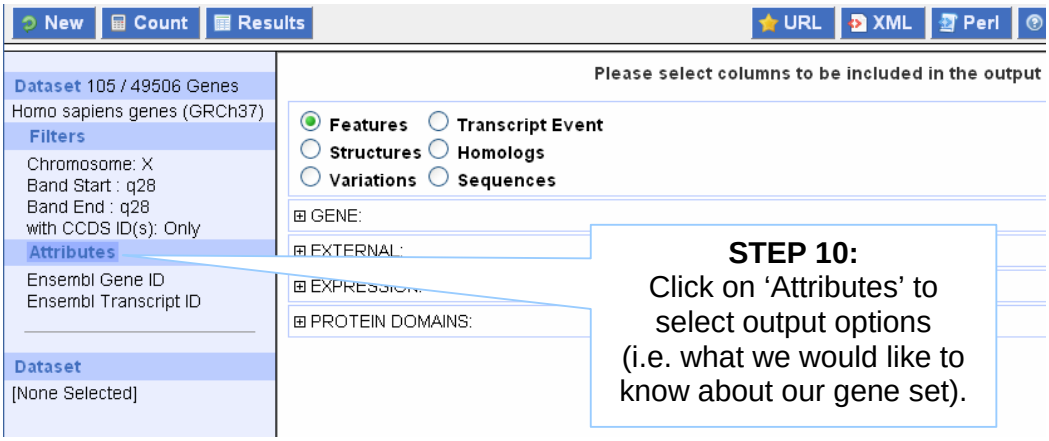


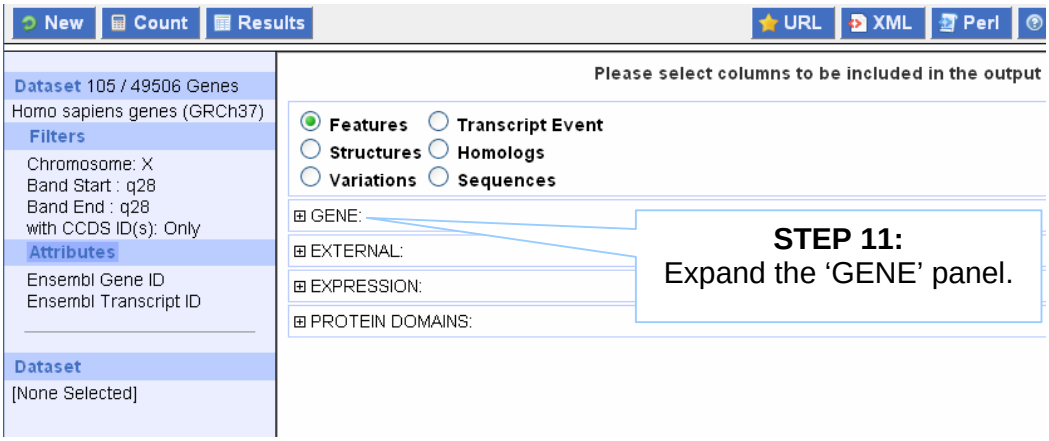
STEP 8:
Limit to genes **with CCDS ID(s)**. Consensus Coding Sequences are assigned when all genome annotation groups agree on a model.

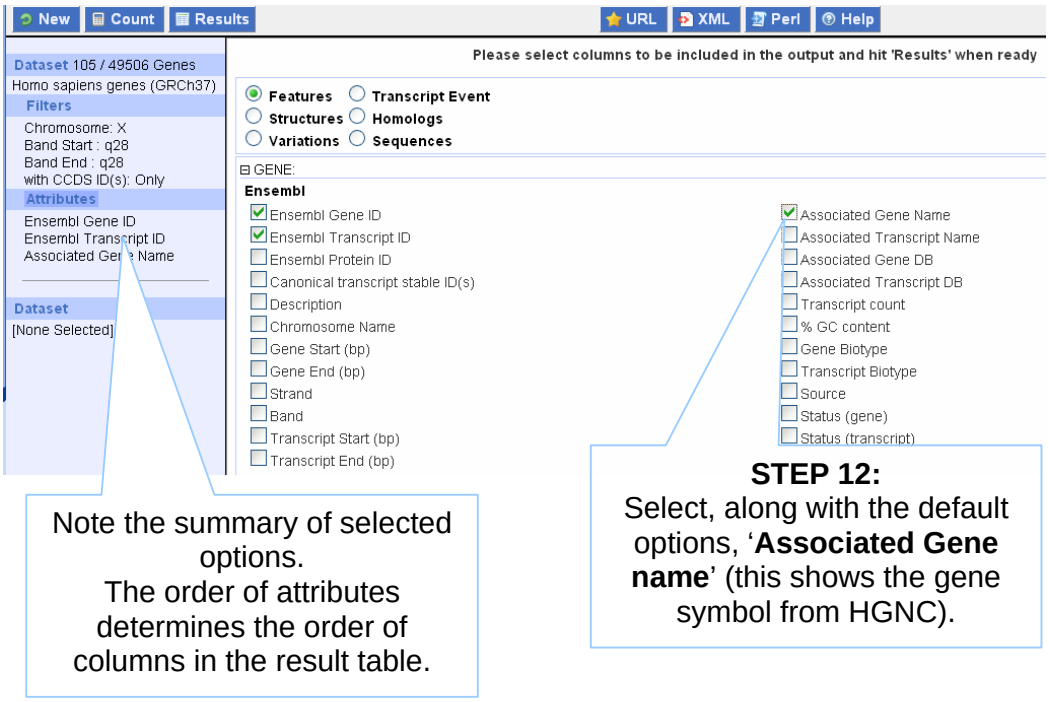


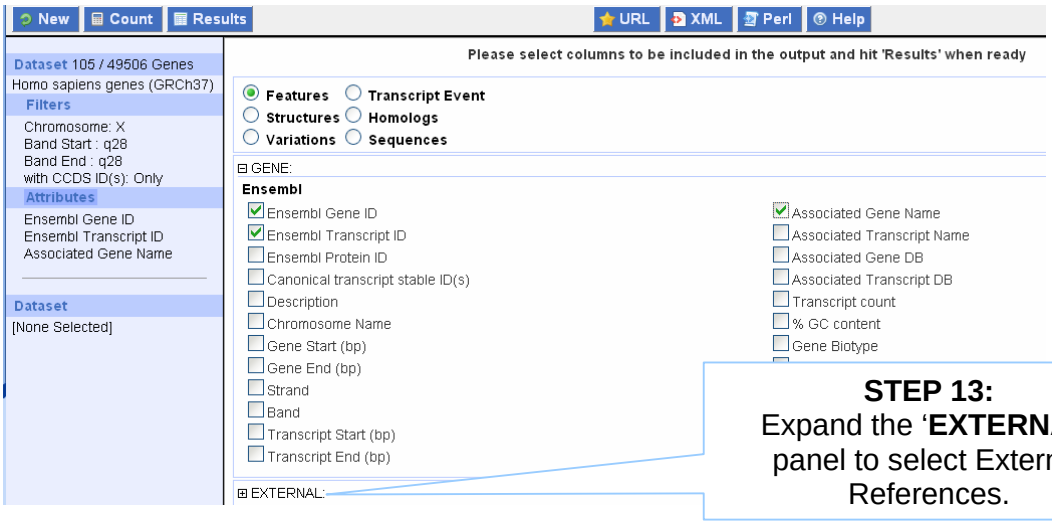
STEP 9:
The filters have determined our gene set.
Click **'Count'** to see how many genes have passed these filters.

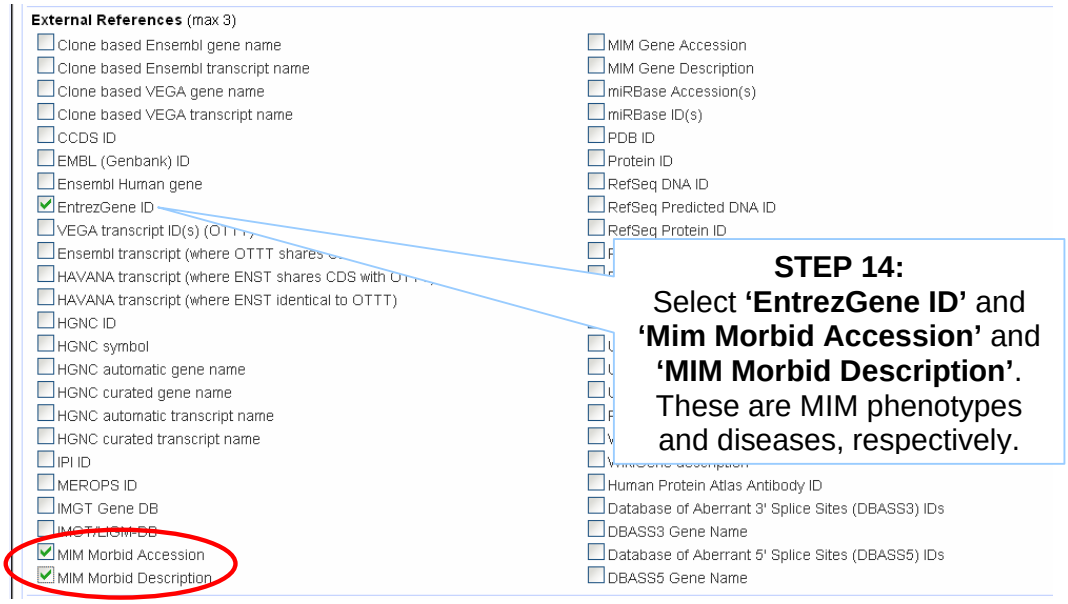
The 'Count' results show 105 human genes out of 42,285 total genes passed the filters.

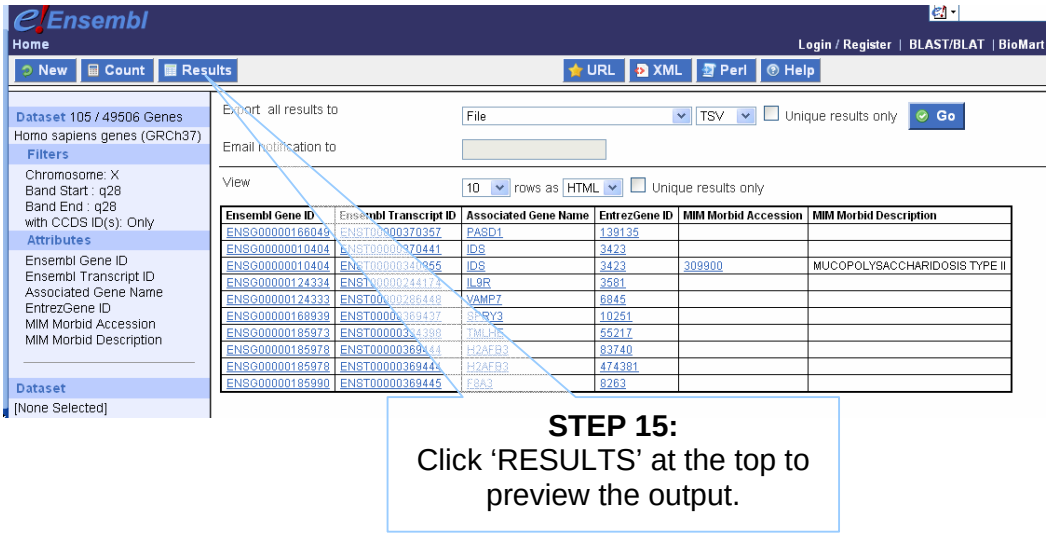

New
Count
Results
URL
XML
Perl
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]
Please select columns to be included in the output
Features
Transcript Event
Structures
Homologs
Variations
Sequences
⊞ GENE:
⊞ EXTERNAL:
⊞ EXPRESSION:
⊞ PROTEIN DOMAINS:
STEP 10:
Click on 'Attributes' to select output options (i.e. what we would like to know about our gene set).

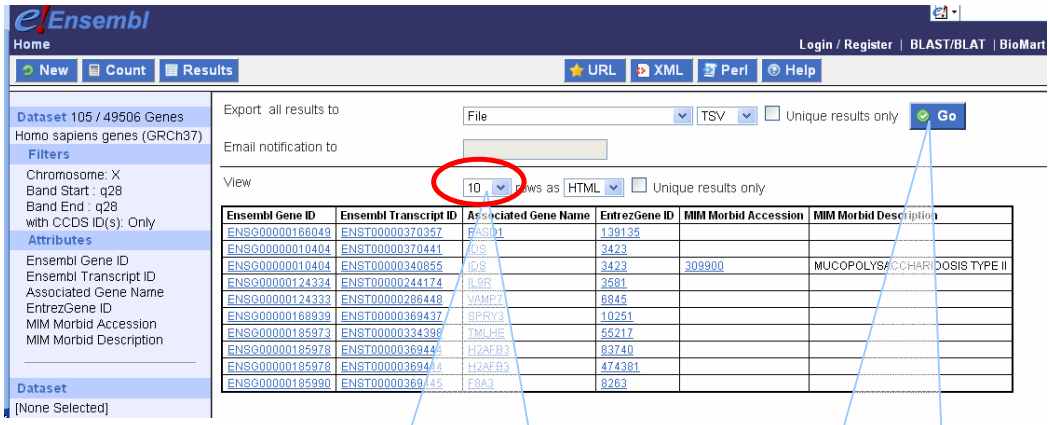

New
Count
Results
URL
XML
Perl
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Dataset
[None Selected]
Please select columns to be included in the output
Features
Transcript Event
Structures
Homologs
Variations
Sequences
⊞ GENE:
⊞ EXTERNAL:
⊞ EXPRESSION:
⊞ PROTEIN DOMAINS:
STEP 11:
Expand the 'GENE' panel.


New
Count
Results
URL
XML
Perl
Help
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name
Dataset
[None Selected]
Please select columns to be included in the output and hit 'Results' when ready
Features
Transcript Event
Structures
Homologs
Variations
Sequences
⊟ GENE:
Ensembl
Ensembl Gene ID
Ensembl Transcript ID
Ensembl Protein ID
Canonical transcript stable ID(s)
Description
Chromosome Name
Gene Start (bp)
Gene End (bp)
Strand
Band
Transcript Start (bp)
Transcript End (bp)
Associated Gene Name
Associated Transcript Name
Associated Gene DB
Associated Transcript DB
Transcript count
% GC content
Gene Biotype
Transcript Biotype
Source
Status (gene)
Status (transcript)
Note the summary of selected options.
The order of attributes determines the order of columns in the result table.
STEP 12:
Select, along with the default options, 'Associated Gene name' (this shows the gene symbol from HGNC).


New
Count
Results
URL
XML
Perl
Help
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name
Dataset
[None Selected]
Please select columns to be included in the output and hit 'Results' when ready
Features
Transcript Event
Structures
Homologs
Variations
Sequences
GENE:
Ensembl
Ensembl Gene ID
Ensembl Transcript ID
Ensembl Protein ID
Canonical transcript stable ID(s)
Description
Chromosome Name
Gene Start (bp)
Gene End (bp)
Strand
Band
Transcript Start (bp)
Transcript End (bp)
Associated Gene Name
Associated Transcript Name
Associated Gene DB
Associated Transcript DB
Transcript count
% GC content
Gene Biotype
EXTERNAL:
STEP 13:
Expand the 'EXTERNAL' panel to select External References.

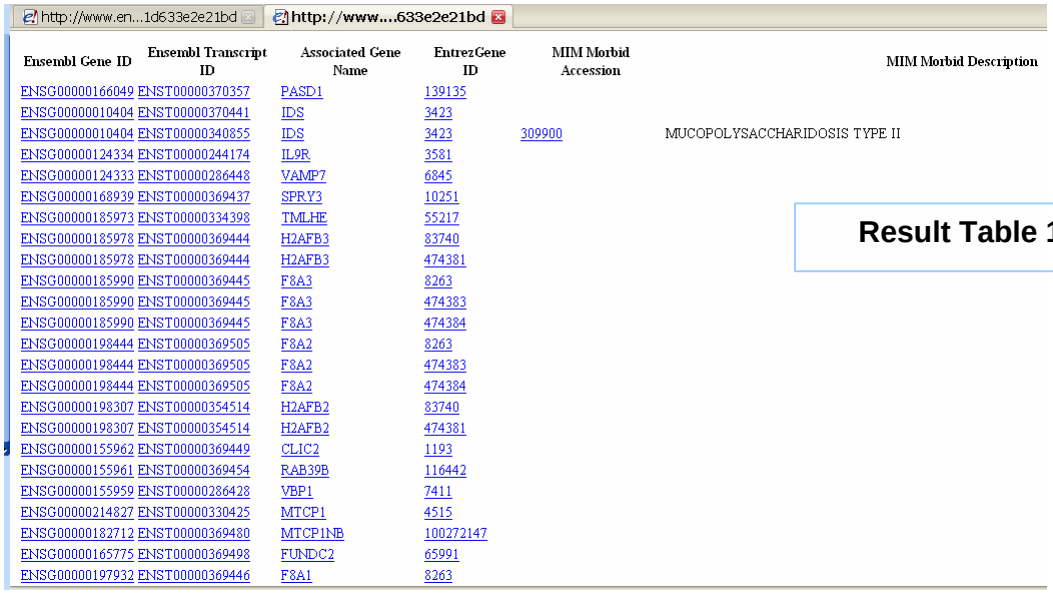

External References (max 3)
Clone based Ensembl gene name
Clone based Ensembl transcript name
Clone based VEGA gene name
Clone based VEGA transcript name
CCDS ID
EMBL (Genbank) ID
Ensembl Human gene
EntrezGene ID
VEGA transcript ID(s) (OTTT)
Ensembl transcript (where OTTT shares
HAVANA transcript (where ENST shares CDS with OTT
HAVANA transcript (where ENST identical to OTTT)
HGNC ID
HGNC symbol
HGNC automatic gene name
HGNC curated gene name
HGNC automatic transcript name
HGNC curated transcript name
IPI ID
MEROPS ID
IMGT Gene DB
IMGT/LIGM-DB
MIM Morbid Accession
MIM Morbid Description
MIM Gene Accession
MIM Gene Description
miRBase Accession(s)
miRBase ID(s)
PDB ID
Protein ID
RefSeq DNA ID
RefSeq Predicted DNA ID
RefSeq Protein ID
Human Protein Atlas Antibody ID
Database of Aberrant 3' Splice Sites (DBASS3) IDs
DBASS3 Gene Name
Database of Aberrant 5' Splice Sites (DBASS5) IDs
DBASS5 Gene Name
STEP 14:
Select 'EntrezGene ID' and 'Mim Morbid Accession' and 'MIM Morbid Description'.
These are MIM phenotypes and diseases, respectively.


Ensembl
Home
Login / Register | BLAST/BLAT | BioMart
New
Count
Results
URL
XML
Perl
Help
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name
EntrezGene ID
MIM Morbid Accession
MIM Morbid Description
Dataset
[None Selected]
Export all results to
File
TSV
Unique results only
Go
Email notification to
View
10
rows as
HTML
Unique results only
Ensembl Gene ID
Ensembl Transcript ID
Associated Gene Name
EntrezGene ID
MIM Morbid Accession
MIM Morbid Description
ENSG00000166049
ENST00000370357
PASD1
139135
ENSG00000010404
ENST00000370441
IDS
3423
ENSG00000010404
ENST00000340955
IDS
3423
309900
MUCOPOLYSACCHARIDOSIS TYPE II
ENSG00000124334
ENST00000244173
IL9R
3581
ENSG00000124333
ENST00000286448
VAMP7
6845
ENSG00000168939
ENST00000369437
SPRY3
10251
ENSG00000185973
ENST00000354398
TMLHE
55217
ENSG00000185978
ENST00000369444
H2AFB3
83740
ENSG00000185978
ENST00000369444
H2AFB3
474381
ENSG00000185990
ENST00000369445
F8A3
8263
STEP 15:
Click 'RESULTS' at the top to preview the output.

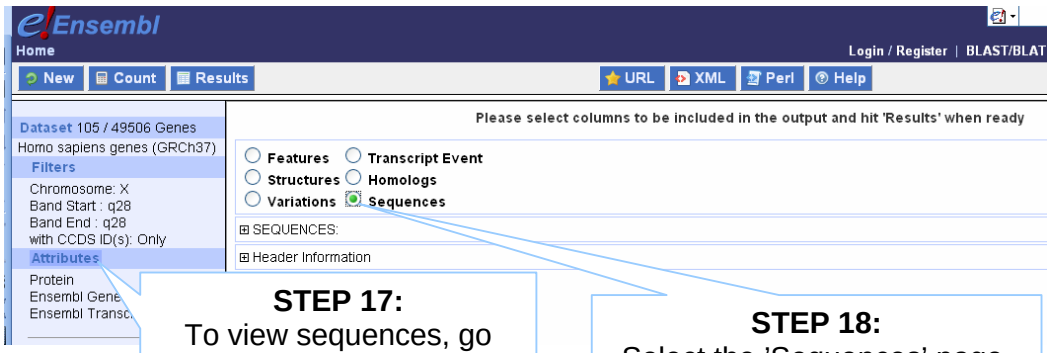


STEP 16:
Go back and change Filters or Attributes if desired.
Or, View ALL rows as HTML…

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

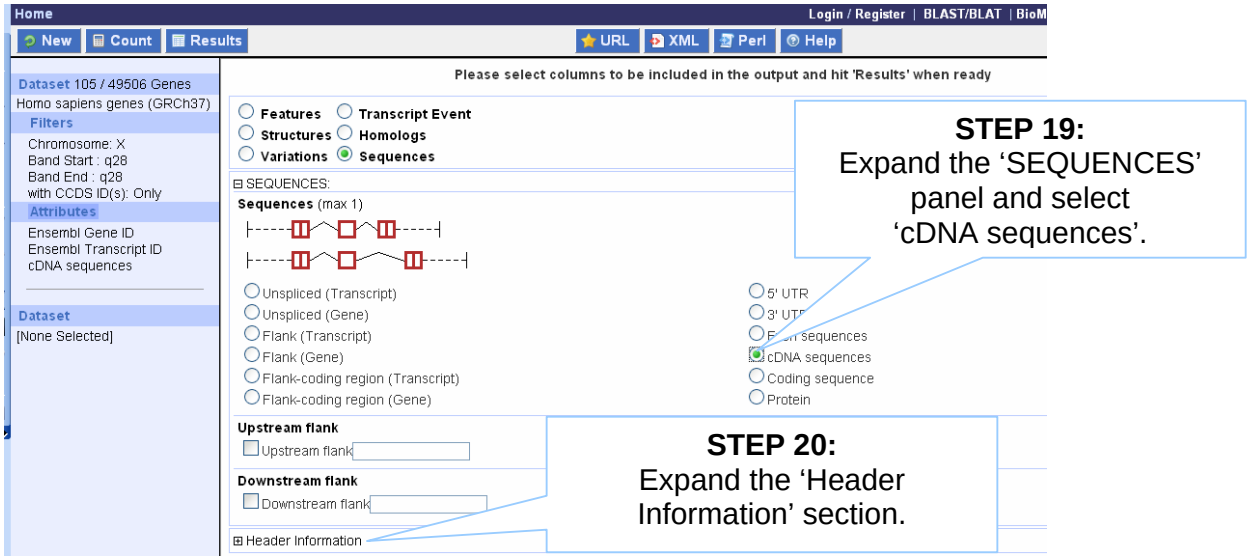


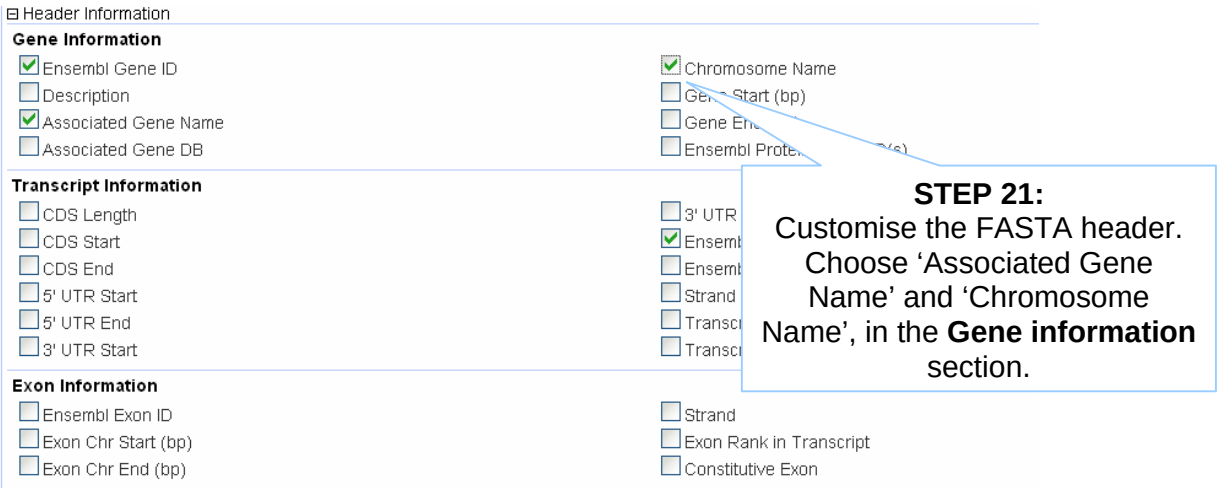
Result Table 1

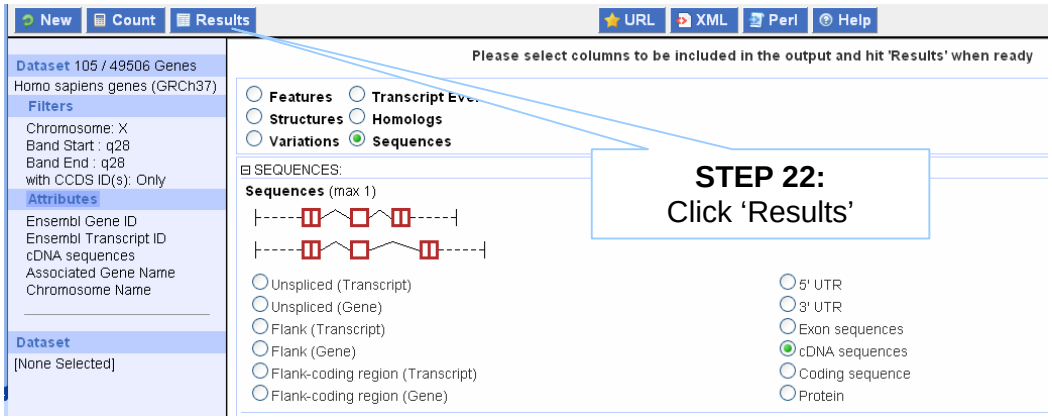


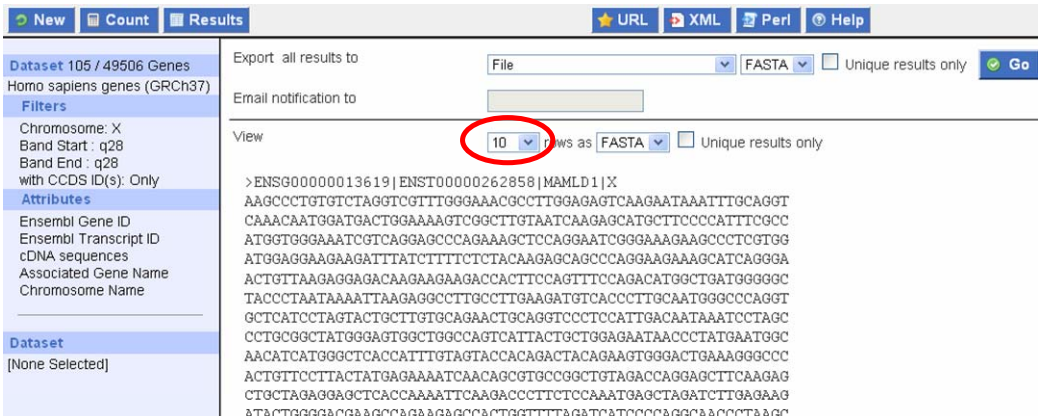
STEP 17:
To view sequences, go back to 'Attributes'

STEP 18:
Select the 'Sequences' page, then expand the 'SEQUENCES' section.


Home
Login / Register | BLAST/BLAT | BioM
New
Count
Results
URL
XML
Perl
Help
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
cDNA sequences
Dataset
[None Selected]
Please select columns to be included in the output and hit 'Results' when ready
Features
Transcript Event
Structures
Homologs
Variations
Sequences
SEQUENCES:
Sequences (max 1)
Unspliced (Transcript)
Unspliced (Gene)
Flank (Transcript)
Flank (Gene)
Flank-coding region (Transcript)
Flank-coding region (Gene)
5' UTR
cDNA sequences
Coding sequence
Protein
Upstream flank
Downstream flank
Header Information
STEP 19:
Expand the 'SEQUENCES' panel and select 'cDNA sequences'.
STEP 20:
Expand the 'Header Information' section.


Header Information
Gene Information
Ensembl Gene ID
Description
Associated Gene Name
Associated Gene DB
Chromosome Name
Transcript Information
CDS Length
CDS Start
CDS End
5' UTR Start
5' UTR End
3' UTR Start
3' UTR
Strand
Exon Information
Ensembl Exon ID
Exon Chr Start (bp)
Exon Chr End (bp)
Strand
Exon Rank in Transcript
Constitutive Exon
STEP 21:
Customise the FASTA header. Choose 'Associated Gene Name' and 'Chromosome Name', in the **Gene information** section.


New
Count
Results
URL
XML
Perl
Help
Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)
Filters
Chromosome: X
Band Start : q28
Band End : q28
with CCDS ID(s): Only
Attributes
Ensembl Gene ID
Ensembl Transcript ID
cDNA sequences
Associated Gene Name
Chromosome Name
Dataset
[None Selected]
Please select columns to be included in the output and hit 'Results' when ready
Features
Structures
Homologs
Variations
Sequences
SEQUENCES:
Sequences (max 1)
Unspliced (Transcript)
Unspliced (Gene)
Flank (Transcript)
Flank (Gene)
Flank-coding region (Transcript)
Flank-coding region (Gene)
5' UTR
3' UTR
Exon sequences
cDNA sequences
Coding sequence
Protein
STEP 22:
Click 'Results'



Again, View ALL rows as FASTA for the full list… (make sure pop-up blocker is off):

>Header: Gene ID, Transcript ID, Gene Name, Chromosome

```
>ENSG00000013619|ENST00000262858|MAMLD1|X
AAGCCCTGTGTCTAGGTCGTTTGGGAAACGCCTTGGAGAGTCAAGAATAAATTTGCAGGT
CAAACAATGGATGACTGGAAAAGTCGGCTTGTAATCAAGAGCATGCTTCCCCATTTCGCC
ATGGTGGGAAATCGTCAGGAGCCCAGAAAGCTCCAGGAATCGGGAAAGAAGCCCTCGTGG
ATGGAGGAAGAAGATTTATCTTTTCTCTACAAGAGCAGCCCAGGAAGAAAGCATCAGGGA
ACTGTTAAGAGGAGACAAGAAGAAGACCACTTCCAGTTTCCAGACATGGCTGATGGGGGC
TACCCTAATAAAATTAAGAGGCCTTGCCTTGAAGATGTCACCCTTGCAATGGGCCCAGGT
GCTCATCCTAGTACTGCTTGTGCAGAACTGCAGGTCCCTCCATTGACAATAAATCCTAGC
CCTGCGGCTATGGGAGTGGCTGGCCAGTCATTACTGCTGGAGAATAACCCTATGAATGGC
AACATCATGGGCTCACCATTTGTAGTACCACAGACTACAGAAGTGGGACTGAAAGGGCCC
ACTGTTCCTTACTATGAGAAAATCAACAGCGTGCCGGCTGTAGACCAGGAGCTTCAAGAG
CTGCTAGAGGAGCTCACCAAAATTCAAGACCCTTCTCCAAATGAGCTAGATCTTGAGAAG
ATACTGGGGACGAAGCCAGAAGAGCCACTGGTTTTAGATCATCCCCAGGCAACCCTAAGC
ACAACTCCCAAGCCTTCGGTTCAGATGTCACACTTGGAGAGCCTGGCTTCCAGCAAGGAG
TTTGCTTCTAGTTGCAGCCAAGTTACTGGCATGTCACTTCAGATCCCATCCTCCTCCACA
GGGATCAGCTATTCGATTCCTTCCACCAGTAAGCAGATAGTGTCACCGAGTTCTTCAATG
GCACAGTCCAAGAGCCAGGTCCAGGCCATGCTCCCTGTCGCTCTGCCCCCCTTACCAGTG
CCTCAGTGGCATCACGCCCACCAGCTGAAGGCGTTGGCAGCCAGCAAGCAGGGGTCTGCT
ACAAAGCAGCAAGGGCCCACCCCCAGTTGGTCTGGTCTGCCTCCTCCAGGACTCTCTCCA
CCTTACCGCCCAGTGCCATCACCACACCCACCACCGCTGCCACTGCCACCACCACCACCC
CCATTCAGCCCCCAGAGCCTCATGGTGTCCTGCATGTCGTCCAATACCTTGTCGGGTAGC
ACTCTCCGAGGCTCTCCCAATGCCTTACTGTCAAGCATGACGTCCAGCAGCAATGCTGCC
CTGGGGCCCGCCATGCCCTATGCTCCTGAGAAGCTCCCCAGCCCTGCTCTCACTCAACAG
CCGCAGTTCGGCCCTCAGAGCTCCATTCTTGCCAACCTCATGTCCTCTACCATCAAAACC
CCTCAAGGACACCTGATGTCTGCTCTGCCTGCCAGCAACCCTGGGCCGTCCCCACCCTAT
CGCCCAGAGAAGCTCTCTAGCCCAGGCTTGCCACAGCAGTCCTTCACCCCACAGTGTTCC
CTGATCCGAAGCCTCACTCCCACCAGTAATCTTCTAAGCCAGCAACAGCAGCAGCAGCAG
.....
ATAGAAAACCCACACCCACTGTCCTGTAAACTTTTCTCAGTGTCCAGACTTTCTGTAATC
ACATTTTAATTGCCACCTCGTATTTCACCTCTACATTTGAAATCTGGCGTCTGTTTCAAG
CCAGTGTGTTTTTTCTTCGTTCTGTAATAAACAGCCAGGAGAAAAGTG
>ENSG00000166008|ENST00000298974|MAGEA9|X
GTGCGCACTGGGGGTCAGAGAGAAGGGAGAGGCCTCCTTCTGAGGGGCGGCTTGATACCG
GTGGAGGAGCTCCAGGAAGCAGGCAGGCCTTGGTCTGAGACAGTGTCCTCAGGTCGCAGA
GCAGAGGAGACCCAGGCAGTGTCAGCAGTGAAGGTTCTCGGGACAGGCTAACCAGGAGGA
CAGGAGCCCCAAGAGGCCCCAGAGCAGCACTGACGAAGACCTGCCTGTGGGTCTCCATCG
CCCAGCTCCTGCCCACGCTCCTGACTGCTGCCCTGACCAGAGTCATCATGTCTCTCGAGC
AGAGGAGTCCGCACTGCAAGCCTGATGAAGACCTTGAAGCCCAAGGAGAGGACTTGGGCC
TGATGGGTGCACAGGAACCCACAGGCGAGGAGGAGGAGACTACCTCCTCCTCTGACAGCA
AGGAGGAGGAGGTGTCTGCTGCTGGGTCATCAAGTCCTCCCCAGAGTCCTCAGGGAGGCG
CTTCCTCCTCCATTTCCGTCTACTACACTTTATGGAGCCAATTCGATGAGGGCTCCAGCA
GTCAAGAAGAGGAAGAGCCAAGCTCCTCGGTCGACCCAGCTCAGCTGGAGTTCATGTTCC
AAGAAGCACTGAAATTGAAGGTGGCTGAGTTGGTTCATTTCCTGCTCCACAAATATCGAG
TCAAGGAGCCGGTCACAAAGGCAGAAATGCTGGAGAGCGTCATCAAAAATTACAAGCGCT
```

cDNA 1

cDNA 2

V) BioMart Exercises and Answers

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 'known' 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 56'
- Select 'Homo sapiens genes (GRCh37)'.

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)' 'KNOWN'.
- In the 'PROTEIN DOMAINS' section, select the '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 **25**.

Click on '**Attributes**' (at the left). Select the output for your gene list as follows:
- Select the 'Variations' Attribute Page.
- In the 'GENE' section 'Ensembl Gene ID' and 'Ensembl Transcript ID' are selected by default – also select 'Ensembl Protein ID'.
- In the 'GENE ASSOCIATED VARIATIONS' section 'Reference ID' is selected. Also select 'Allele', 'Protein location (aa)' 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!!

Why are several columns in the preview table blank? These variations are not in the coding sequence.

Exercise 2

Generate a list of all zebrafish protein-coding genes that are located on chromosome 3. Export gene name, description, Zfin symbol, and InterPro domains.

Exercise 3

For this exercise, it's easier to cut and paste the IDs from the online course booklet. One copy is here:

http://www.ebi.ac.uk/~gspudich/workshop_presentations/coursebook.pdf

BioMart is a very handy tool when you want to convert IDs from different databases. The following is a list of 29 IDs of human proteins from the RefSeq database of NCBI (http://www.ncbi.nlm.nih.gov/projects/RefSeq/):

NP_001218, NP_203125, NP_203124, NP_203126, NP_001007233, NP_150636, NP_150635, NP_001214, NP_150637, NP_150634, NP_150649, NP_001216, NP_116787, NP_001217, NP_127463, NP_001220, NP_004338, NP_004337, NP_116786, NP_036246, NP_116756, NP_116759, NP_001221, NP_203519, NP_001073594, NP_001219, NP_001073593, NP_203520, NP_203522

Generate a list that shows to which Ensembl Gene IDs and to which HGNC symbols these RefSeq IDs correspond.

Exercise 4

In a paper from 1995 Ayyagari *et al.* mapped the human 'Usher Syndrome type I C' to the genomic region between the markers D11S1397 and D11S1310 (Mol. Vis. 1:2, 1995).

Confirm this finding by generating a list of the genes located in this region.

Exercise 5

Forrest *et al.* performed a microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers (Environ Health Perspect. 2005 June; 113(6): 801–807). The microarray used was the Affymetrix U133A/B (also called U133 plus 2) GeneChip. The top 25 up-regulated probe-sets were:

207630_s_at, 221840_at, 219228_at, 204924_at 227613_at, 223454_at, 228962_at, 214696_at, 210732_s_at, 212371_at, 225390_s_at, 227645_at, 226652_at, 221641_s_at, 202055_at, 226743_at, 228393_s_at, 225120_at, 218515_at, 202224_at, 200614_at, 212014_x_at, 223461_at, 209835_x_at, 213315_x_at

(a) Retrieve for the genes corresponding to these probe-sets the Ensembl Gene and Transcript IDs as well as their HGNC symbols (as far as available) and descriptions.

(b) In order to analyse these genes for possible promoter/enhancer elements, retrieve the 2000 bp upstream of the transcripts of these genes.

(c) In order to be able to study these human genes in mouse, identify their mouse orthologues. Also retrieve the genomic coordinates of these orthologues.

Exercise 6

Known dolphin genes match to a protein or mRNA sequence in a public database for dolphin (this is in contrast to 'known by projection' which was based on evidence from another species).

Step 1: For all known dolphin genes in Ensembl, export human homologues.

Step 2: ***Advanced:*** export a list of the human gene IDs alone (select only one attribute, and then select 'Unique results only'.) Do a second query in BioMart with human genes, upload these gene IDs and export gene names!

Exercise 7 – Exploring another database

This query uses the Reactome (http://www.reactome.org) metabolic pathway information. Use MartView at **www.biomart.org.**

Determine in which pathways the gene ENSG00000164305 plays a role.

Exercise 8

Design your own query!

Answers: BIOMART

1. You should find ***25*** known genes on chromosome 2 with this InterPro domain. The result table is quite large; so don't export the entire table if export is going slowly.

2. Click '**NEW**' for a new query.

Start with all the zebrafish Ensembl genes:

Choose the '**ENSEMBL 56**' database.
Choose the '***Danio rerio genes (Zv8)***' dataset.

Now filter for the genes on the 3 chromosome:

Click on **'Filters**' in the left panel.
Expand the '**REGION**' section by clicking on the + box.
Select '**Chromosome 3**'. Make sure the check box in front of the filter is ticked otherwise the filter won't work.

Now filter further for genes that are protein coding:

Expand the '**GENE**' section by clicking on the + box.
Select '**Gene type'** as **'protein_coding**'.
Click the [Count] button on the toolbar.

This should give you 1106 / 27854 Genes.

Specify the attributes to be included in the output (note that a number of attributes will already be default selected):

Click on 'Attributes' in the left panel.
Select the 'Features' attributes page.
Expand the 'GENE' section by clicking on the + box.
Select, in addition to the attributes 'Ensembl Gene ID' and 'Ensembl Transcript ID' that are already default selected, 'Associated Gene Name' and 'Description'.

Expand the 'EXTERNAL' panel to select ZFIN symbols. These will be equal to the Gene Name, when those are available.

Expand the 'PROTEIN' section to add 'InterPro ID' 'InterPro Short Description'.

Click the [Results] button on the toolbar.

If you are happy with how the results look in the preview, output all the results:

Select ‘View All rows as HTML’ or export all results to a file.

3. Click [New].
Choose the ‘**ENSEMBL 56**’ database.
Choose the ‘***Homo sapiens genes (GRCh37)’*** dataset.

Click on **‘Filters’** in the left panel.
Expand the ‘**GENE’** section by clicking on the + box.
Select **‘ID list limit - RefSeq protein ID(s)’** and enter the list of IDs in the text box (either comma separated or as a list).

Click on **‘Attributes’** in the left panel.
Select the **‘Features’** attributes page.
Expand the **‘GENE’** section by clicking on the + box.
Deselect ‘Ensembl Transcript ID’.
Expand the **‘External’** section by clicking on the + box.
Select ‘**HGNC symbol’** and ‘**RefSeq Protein ID’ f**rom the **‘External References’** section.

Click the [Results] button on the toolbar.
Select ‘View All rows as HTML’ or export all results to a file. Tick the box ‘Unique results only’.

Note: BioMart is ‘transcript-centric’, which means that it will give a separate row of output for each transcript of a gene, even if you don’t include the Ensembl Transcript ID in your output. When you don’t want this, use the ‘Unique results only’ option.

Your results should show that the RefSeq IDs map to **10** genes (you can also see this by clicking ‘Count’.

4. Click **[New].**
Choose the **‘ENSEMBL 56’** database.
Choose the ‘***Homo sapiens genes*** (GRCh37)’ dataset.

Click on **‘Filters’** in the left panel.
Expand the **‘REGION’** section by clicking on the + box.
Enter ‘Marker Start: **D11S1397**’ and ‘Marker End: **D11S1310**’.

Click on **‘Attributes’** in the left panel.
Select the **‘Features’** attributes page.
Expand the **‘GENE’** section by clicking on the + box.
Deselect ‘Ensembl Transcript ID’.
Select **‘Associated Gene Name’** and **‘Description’**.

Click the [Results] button on the toolbar.
Select ‘View All rows as HTML’ or export all results to a file. Tick the box ‘Unique results only’.

Your results should show **29** genes. Among these there should be of one (ENSG00000006611) with name 'USH1C' and description 'Harmonin (Usher syndrome type-1C protein) (Autoimmune enteropathy-related antigen AIE-75) (Antigen NY-CO-38/NY-CO-37) (PDZ-73 protein) (Renal carcinoma antigen NY-REN-3). [Source:UniprotKB/SWISSPROT;Acc:Q9Y6N9]'. This suggests that Ayyagari et al. correctly mapped Usher Syndrome type I C to this genomic region.

5. (a) Click **[New].**
Choose the **'ENSEMBL 56'** database.
Choose the 'Homo sapiens genes (GRCh37)' dataset.

Click on **'Filters'** in the left panel.
Expand the '**GENE'** section by clicking on the + box.
Select **'ID list limit - Affy hg u133 plus 2 ID(s)'** and enter the list of probe-set IDs in the text box (either comma separated or as a list).

Click on **'Attributes'** in the left panel.
Select the **'Features'** attributes page.
Expand the **'GENE'** section by clicking on the + box.
Select, in addition to the default selected attributes, '**Description**'.
Expand the '**External'** section by clicking on the + box.
Select **'HGNC symbol'** from the '**External References**' section and '**AFFY HG U133-PLUS-2**' from the '**Microarray Attributes**' section.

Click the **[Results]** button on the toolbar.
Select 'View All rows as HTML' or export all results to a file. Tick the box 'Unique results only'.

Your results should show that the 25 probes map to 23 Ensembl genes.

(b) Don't change Dataset and Filters- simply click on '**Attributes**'.

Select the **'Sequences'** attributes page.
Expand the **'SEQUENCES'** section by clicking on the + box.
Select **'Flank (Transcript)'** and enter **'2000'** in the **'Upstream flank'** text box.
Expand the **'Header information'** section by clicking on the + box.
Select, in addition to the default selected attributes, **'Description'** and **'Associated Gene Name'.**

Note: 'Flank (Transcript)' will give the flanks for all transcripts of a gene with multiple transcripts. 'Flank (Gene)' will give the flanks for the transcript with the outermost 5' or 3' end.

Click the **[Results]** button on the toolbar.

(c) You can leave the Dataset and Filters the same, and go directly to the **'Attributes'** section:

Click on '**Attributes'** in the left panel.
Select the **'Homologs'** attributes page.
Expand the **'GENE'** section by clicking on the + box.
Select **'Associated Gene Name'.**
Deselect 'Ensembl Transcript ID'.
Expand the **'MOUSE ORTHOLOGS'** section by clicking on the + box.
Select '**Mouse Ensembl Gene ID'**, **'Mouse Chromosome', 'Mouse Chr Start (bp)**' and **'Mouse Chr End (bp)'.**

Click the [Results] button on the toolbar.
Check the box 'Unique results only'. Select 'View All rows as HTML' or export all results to a file.

Your results should show that for **23** out of the 23 human genes at least one mouse orthologue has been identified. ENSG00000123130 has two mouse orthologues and ENSG00000172716 has three. Four human genes (ENSG00000197620, ENSG00000186594, ENSG00000130844 and ENSG00000089335) have none.

6. Step 1: Choose *'Ensembl 56'* and *'Tursiops truncates genes (turTru1)'*.
Filters: Expand the *'GENE'* panel and select Status (gene) as *'KNOWN'*.
Attributes: Select *'Human Ensembl Gene ID'* under the *'Homologs'* page.

Step 2: Remove *'Ensembl Gene ID'*, *'Ensembl Transcript ID'*, and *'Ensembl Protein ID'* from the Attributes. Click on *'Unique results only'* and export the file.
Click NEW. Start with *'Ensembl 56'*, '*Homo sapiens genes (GRCh37)'*.
Filters: Expand the GENE panel, and click browse to upload a file into the *'ID List Limit Box'*.
In Attributes, select 'Gene Name'.
Click Results.

7. This query accesses the Reactome BioMart. To do this, click on the 'MartView' tab from www.biomart.org

Choose the **REACTOME** database and **'pathway'** as Dataset.
Click **'Filters'** and enter ENSG00000164305 into the box titled: Limit to pathways containing these IDs. Change the ID type to ENSEMBL gene ID.
Click on **'Attributes'** in the left panel.
Deselect 'Pathway DB_ ID'.
Select 'Pathway Name'.
Click the **[Results]** button on the toolbar.
Select 'View All rows as HTML' or export your results to a file. Tick the box 'Unique results only'.

Your results should show that ENSG00000164305 plays a role in various processes associated with apoptosis (programmed cell death).

VI) EXERCISES GENEBUILD

Exercise 1

(a) From where is the human genome assembly?
(b) How long did it take for Ensembl to perform the last gene build?
(c) How many protein coding genes are there in human? Can you get this same number using BioMart?

Exercise 2

Find the Ensembl GALP (Galanin-like peptide precursor) gene for human.

(a) From what source did Ensembl get the name for this gene? And from where did it get the description 'Galanin-like peptide Precursor'?
(b) On how many pieces of evidence has the transcript of this gene been built?
(c) Why do some pieces of evidence not support the first exon of the transcript?

Exercise 3

Find the Ensembl Epc1 (enhancer of polycomb homolog 1) gene for mouse.

(a) How many transcripts has Ensembl annotated for this gene?
(b) How many transcripts have the manual annotators of Havana annotated for this gene?
(c) How many transcripts agree between Ensembl and Havana annotation?
(d) What is the reason that Ensembl hasn't annotated one of the Havana transcripts?

Exercise 4

An example of what can go wrong ….
Go to the following page in Ensembl release 46 (of August 2007):

http://aug2007.archive.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000198561

(a) What is wrong with this gene? What could be the reason for this?
(b) Has this problem been fixed in Ensembl release 56?

ANSWERS GENEBUILD

Answer 1

Go to http://www.ensembl.org.
Click on the human picture or the word 'Human' next to it.

(a) GRC (the Genome Reference Consortium) hosts the assembly determined from the IHGP (International Human Genome Project).

(b) Click on 'Assembly and Genebuild' in the side menu.
Three months (from March 2009 until May 2009).

(c) Look further down the table. 23,438 known and 183 novel protein coding genes. Get the same number in BioMart by using the Filter: GENE panel:
Gene Type: Protein coding
Status(gene): Known
(click count).
Change status to Novel
(click count).

Answer 2

Go to the Ensembl homepage.
Under 'Search Ensembl' type 'human gene GALP' and click [Go].
On the page with search results click on 'Ensembl protein_coding Gene: ENSG00000197487 (HGNC (automatic): GALP)'.

(a) From the HUGO Gene Nomenclature Committee (HGNC). From UniProtKB/Swiss-Prot record Q9UBC7.

Click on the 'Transcript: GALP-201' tab.
Click on 'Supporting evidence' in the side menu.

(b) Two main pieces of evidence, NM_033106.2, which is a 'known mRNA' in NCBI's RefSeq set, and CCDS12940.1, which is a coding sequence from the CCDS set. To view these records, click on the diagram representing the sequences and follow the link to the ID. Seven other mRNA and protein sequences are drawn below- these contributed or also aligned well to the Ensembl transcript.

(c) The three pieces of protein evidence (NP_001139018, Q9UBC7_1 and Q9UBC7_2) as well as the CCDS evidence (CCDS12940) don't support the first exon of the GALP transcript, because this exon is a completely untranslated region (it is represented by an unfilled box). Thus, protein sequences and coding sequences alone cannot provide any information for this exon.

Answer 3

Go to the Ensembl homepage.
Under 'Search Ensembl' type 'mouse gene Epc1' and click [Go].
On the page with search results click on 'Ensembl protein_coding Gene: ENSMUSG00000024240 (MGI Symbol: Epc1)'.

(a) Three.

(b) Click on ‘Configure this page’ in the side menu.
Click on ‘Other genes’, select ‘Vega Havana gene – Expanded with labels’ and click [SAVE and close].

There are four VEGA-Havana transcripts.

(c) Three (notice that under the Ensembl transcripts, it is written ‘Ensembl/Havana merge’.)

(d) Click on the Epc-004 transcript in the figure.
Click on ‘OTTMUST00000041784’ in the pop-up menu.
Click on ‘Supporting evidence’ in the side menu.

In this case the reason is that the transcript OTTMUST00000041784 is built on one piece of EST evidence (CF730975.1). As Ensembl doesn’t build on just EST evidence, it hasn’t annotated this transcript.

Answer 4

Go to
http://aug2007.archive.ensembl.org/Homo_sapiens/geneview?gene=ENSG00000198561

(a) This is the way Ensembl used to look! The gene has two HGNC symbols associated with it, CTNND1 and TXNDC14. The culprit is one long transcript O60716-27(ENST00000360682) that connects two transcript clusters.

(b) Go to the current Ensembl homepage at www.ensembl.org

Under ‘Search Ensembl’ type ‘human gene CTNND1’ and click [Go].
On the page with search results click on ‘Ensembl protein_coding Gene: ENSG00000198561 (HGNC (automatic): CTNND1)’.
The gene only has one HGNC name, a good indicator of proper annotation.

Click on the ‘Location’ tab.
Zoom out two steps, so both the CTNDD1 transcripts and the TMX2 (formerly TXNDC14) transcripts are shown.

In Ensembl release 56, CTNND1 and TMX2 are annotated as separate genes.

VII) EXERCISES VARIATIONS

Exercise 1

A non-synonymous SNP, R620W (C1858T), in PTPN22 (Tyrosine-protein phosphatase non-receptor type 22) has been identified as a genetic risk factor for a few diseases.

(a) Find the Ensembl page with information for this SNP.
(b) What is the minor allele of this SNP in Caucasians?
(c) Is this minor allele (in (b)) associated with any diseases?

Exercise 2

Find the Genetic Variation - Comparison image page for human PTPN22 (use transcript PTPN22-001).

(a) Do both individuals (Venter and Watson) have sequence coverage at the position of the R620W (C1858T) SNP?
(b) Does either individual have the minor allele?

Exercise 3

Use BioMart to generate an Excel spreadsheet that contains the following information on all SNPs in the transcripts of the human PTPN22 gene: reference ID, alleles (both nucleotides and amino acids), location (both in transcript and in protein) and consequence to the transcript.

Note: you can start with the Ensembl gene database, filter for the PTPN22 gene and then select your attributes from the 'Variations' attributes page.

ANSWERS VARIATIONS

Answer 1

Go to the Ensembl homepage.
Under 'Search Ensembl' type 'human gene PTPN22' and click [Go].
On the page with search results click on 'Ensembl protein_coding Gene: ENSG00000134242 (HGNC (curated): PTPN22)'.

(a) Click on 'Variation Table' in the side menu.
Click on 'Configure this page' in the side menu.
Under 'Select Variation Type', deselect all options except 'Non-synonymous' and click [SAVE and close].

Two of the four PTPN22 transcripts contain a SNP with AA change W/R and AA co-ordinate 620. This SNP, rs2476601, is the one we are looking for.

(b) Click on ‘rs2476601’.
Click on ‘Population genetics’’.

In Caucasians (CSHL-HAPMAP:HapMap-CEU population) the minor allele is A.

(c) Click ‘Phenotype’ at the left of the Variation page. The allele A is associated with Type 1 Diabetes.

Answer Exercise 2

(a) Click on the ‘Gene: PTPN22’ tab.
Click on ‘ENST00000359785’.
Click on ‘Comparison image’ in the side menu.
There is re-sequencing coverage for both Venter and Watson (grey bars).

(b) Click on ‘Configure this page’ in the side menu.
Under ‘Select Variation Type’, deselect all options except ‘Non-synonymous’.

Neither Venter nor Watson is homozygous for the minor allele (A) of rs2476601, which predisposes one for rheumatoid arthritis.
Watson is heterozygous for rs2476601.
Hint… click on the A/G box above to see the rs number.

Answer Exercise 3

Go to the Ensembl homepage
Click the BioMart link on the toolbar.

Choose the ‘Ensembl 56’ database.
Choose the ‘Homo sapiens genes (NCBI36)’ dataset.

Click on ‘Filters’ in the left panel.
Expand the ‘GENE’ section by clicking on the + box.
Select ‘ID list limit – HGNC symbol’ and enter ‘PTPN22’ in the text box.

Click on ‘Attributes’ in the left panel.
Select the ‘Variations’ attributes page.
Expand the ‘GENE ASSOCIATED VARIATIONS’ section by clicking on the + box.
Select, in addition to the attribute ‘Reference ID’ that is already default selected, ‘Allele’, ‘Transcript location (bp)’, ‘Protein location (aa)’, ‘Protein Allele’, and ‘Consequence Type (Transcript Variation)’.

Click the [Results] button on the toolbar.
Select ‘Export all results to file – XLS’ (unique results only) and click [Go].

Open in Excel.

VIII) EXERCISES COMPARATIVE GENOMICS

Exercise 1 - Orthologs, paralogs and genetrees

Find the Ensembl CASP5 (Caspase-5) gene of human.

(a) How many within-species paralogues are predicted for this gene?
Note the Target %id and Query %id. Which paralogue has the most sequence similarity with CASP5?

Retrieve an alignment between CASP5 and one of its paralogues.

(b) Is there an orthologue predicted for this gene in gorilla?

(c) Have a look at the genetree for this gene. Which of the paralogues of CASP5 is due to the most recent duplication event?

(d) Retrieve an alignment between members of any node using Jalview.

Exercise 2 - Rhodopsins

The photoreceptor cells in the retina of the human eye contain a number of different photoreceptors. The rod cells contain rhodopsin, which is responsible for monochromatic vision in the dark. The cone cells all contain one of three types of opsins, which respond to long-wave (red), middle-wave (green) and short-wave (blue) light, respectively, and are responsible for trichromatic color vision (see for instance http://en.wikipedia.org/wiki/Opsin).

(a) Find the gene encoding the red-sensitive opsin.

(b) How many within-species paralogs have been identified for this gene?
Note the 'Target %id' and 'Query %id'. Which paralog has the most sequence similarity with the red-sensitive opsin?

(c) Have a look at the genomic location of the red-, green- and blue-sensitive opsin genes. Does this explain why red-green color blindness is much more prevalent in males than in females (e.g. in the US population 7% vs 0.4%)?

Exercise 3 – The 31-Species Alignment

Find the Ensembl BRCA2 (Breast cancer type 2 susceptibility protein) gene for human and go to the Region in detail page.

(a) Turn on some of the BLASTZ alignment tracks and some of the Translated BLAT alignment tracks. Does the degree of conservation between human and the various other species reflect their evolutionary relationship? Which parts of the BRCA2 gene seem to be the most conserved? Did you expect this?

(b) Turn on the tracks showing the 31 way alignments and constrained elements. These are found in “configure this page”, “Multiple alignments”. Read more about these conservation scores and constrained elements in the comparative genomics documentation (under Docs and FAQs). Do these tracks confirm what you already saw in the tracks with pairwise alignment data?

ANSWERS COMPARATIVE GENOMICS

Answer 1

Under ‘Search Ensembl’ type ‘human gene CASP5’ and click [Go]. On the page with search results click on ‘Ensembl protein_coding Gene: ENSG00000137757 (HGNC (curated): CASP5)’.

(a) Click on ‘Paralogues’ in the side menu.

There are twelve within-species paralogues predicted for human CASP5. The first one has the highest Target %id and Query %id. (Not sure what these are? Click on the Help button, and then ‘Glossary’ in the resulting window.)

Click on [Align] next to the paralogue.

(b) Click on ‘Orthologues’ in the side menu.

Yes, there is an orthologue predicted for human CASP5 in gorilla: ENSGGOG00000015759 (CASP5).

(c) Click on ‘Gene Tree (image)’ in the side menu.
Click on ‘View paralogs of current gene’ under the figure.

Click on the nodes (red squares) for the duplication events that have given rise to the various paralogues.

CASP5 and CASP4 are related by a duplication on the level of the Eutheria (Placental Mammals). Click on the common ancestor (red node) to see this.

(d) Click on the duplication node (red square) or speciation node (blue square) of the sub-tree that you are interested in.
In the pop-up menu click on [Start Jalview].
To edit the alignment display, you can remove sequences using the option Edit > Delete in the menu bar. Note the other available edit options, e.g. Remove Empty Columns.

Answer 2

(a) Go to the Ensembl homepage (http://www.ensembl.org).
Type 'human red-sensitive opsin gene' in the 'Search: for' text box.
Click [Go].
Click on '*Homo sapiens*' on the page with search results.
Click on 'Gene'.
Click on 'Ensembl protein_coding Gene: ENSG00000102076 (HGNC (curated): OPN1LW)'.

'LW' in the gene symbol OPN1LW stands for 'long-wave'.

(b) Click on 'Comparative Genomics - Paralogues' in the side menu.

There have been nine within-species paralogs identified for the human red-sensitive opsin gene. Gene pairs with the highest %ID are listed at the top of the list. ENSG00000166160 (OPN1MW2) and ENSG00000147380 (OPN1MW), the genes encoding the green-sensitive (middle-wave) opsins, have the highest Target %id and Query %id. The green-sensitive opsins have the highest sequence similarity to red-sensitive opsin (Target %id indicates the percentage of the sequence of red-sensitive opsin matching the sequence of the paralog protein. Query %id indicates the percentage of the sequence of the paralog protein matching the sequence of red-sensitive opsin).

(c) Click on 'ENSG00000166160', 'ENSG00000147380' and 'ENSG00000128617'.

The genes for the red and green-sensitive opsins are located next to each other on the X chromosome, while the gene for the blue-sensitive opsin is located on chromosome 7. As females have two X chromosomes a normal gene on one chromosome can often make up for a defective one on the other, whereas males cannot make up for a defective gene. Thus, red-green colour blindness is much more prevalent in males than in females. Variations in the genes for red and green-sensitive opsins can cause subtle differences in colour perception, while tandem rearrangements due to unequal crossing-over between these genes cause more serious defects in colour vision.

Answer Exercise 3

Under 'Search Ensembl' type 'human gene BRCA2' and click [Go].
On the page with search results click on 'Ensembl protein_coding Gene: ENSG00000139618 (HGNC (curated): BRCA2)'.
Click on the Location tab.
Click on 'Configure this page' in the side menu
Click on 'BLASTZ alignments', select some tracks, click on 'Translated BLAT' alignments, select some tracks and click [SAVE and close].

(a) Species that are closer to human in evolution show a larger extent of conservation. Especially the exon sequences of BRCA2 seem to be highly

conserved between the various species, which is what you would expect for a functionally important protein.

(b) The ‘Conservation score’ and ‘Constrained elements 31 way’ tracks largely correspond with the data in the pairwise alignment tracks; the exons of the BRCA2 gene seem to show high conservation.