

# Access to genes and genomes with Ensembl



## Course Manual

October/November 2009 (e56)

## **CONTENTS**

I) INTRODUCTION.....	3
II) BROWSING ENSEMBL – Worked example.....	7
III) BROWSING ENSEMBL – Exercises.....	20
Answers.....	21
IV) BIOMART – Worked example.....	23
V) BIOMART – Exercises.....	31
Answers.....	34
VI) EVALUATING GENES AND TRANSCRIPTS (GENEBUILD)	
Exercises.....	38
Answers.....	38
VII) VARIATIONS	
Exercises.....	41
Answers.....	41
VIII) COMPARATIVE GENOMICS	
Exercises.....	43
Answers.....	44

## 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.

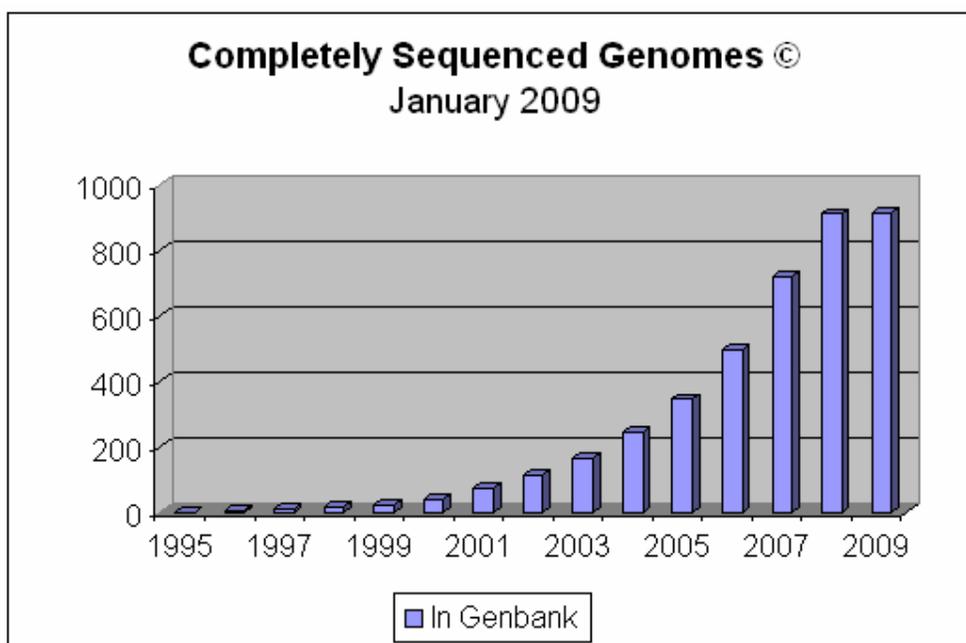


Figure 1. Completely sequenced genomes as of January, 2009 (figure taken from <http://www.genomesonline.org>).

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.*<sup>1</sup>

#### **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)

---

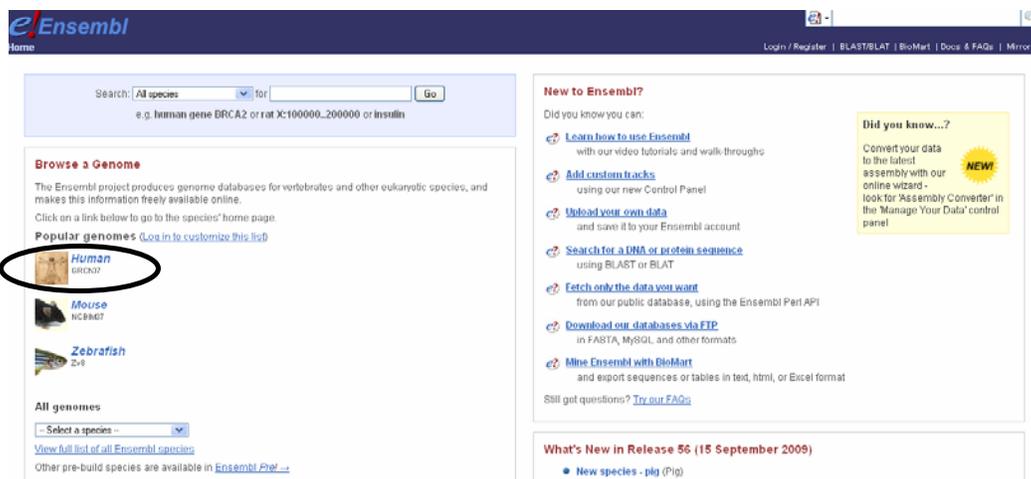
<sup>1</sup> This paper was part of the May 2004 issue of *Genome Research* which included an Ensembl special covering detailed aspects of the Ensembl web site, the underlying scalable database system for storing genome sequence and annotation information, as well as the automated genome analysis and 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](http://www.ensembl.org)**



The screenshot shows the Ensembl website homepage. At the top, there is a search bar with the text "Search: All species for" and a "Go" button. Below the search bar, there is a section titled "Browse a Genome" with the text "The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online. Click on a link below to go to the species' home page." Underneath, there is a section titled "Popular genomes" with a link to "Log in to customize this list". The "Human" link is circled in black. Other links include "Mouse" and "Zebrafish". At the bottom of the "Popular genomes" section, there is a dropdown menu labeled "All genomes" with a "Select a species" button. To the right of the "Browse a Genome" section, there is a "New to Ensembl?" section with several links: "Learn how to use Ensembl", "Add custom tracks", "Upload your own data", "Search for a DNA or protein sequence", "Fetch only the data you want", "Download our databases via FTP", and "Mine Ensembl with BioMart". There is also a "Did you know...?" section with a "NEW" badge and a link to "Convert your data to the latest assembly with our online wizard - look for 'Assembly Converter' in the 'Manage Your Data' control panel". At the bottom right, there is a "What's New in Release 56 (15 September 2009)" section with a link to "New species - plg (Pig)".

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

**About this species**

- Description
  - Genome Statistics
  - Assembly and Genebuil
  - Top 40 InterPro hits
  - Top 500 InterPro hits
- What's New
- Sample entry points
  - Karyotype
  - Location (AL032821.2)
  - Gene (BRCA2)
  - Transcript (FOXP2-203)

**Search Ensembl Human**

Search for:

**Description** [Assembly and Genebuild >](#)

**Assembly**

This site provides a data set based on the February 2009 *Homo sapiens* high coverage assembly from the [Genome Reference Consortium](#). The data set consist on gene models built from the genewise alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate

This release of the assembly has the following properties:

- 27478 contigs.
- contig length total 3.2 Gb.
- chromosome length total 3.1 Gb.

**Annotation**

Since release 55 (July 2009) the gene annotation presented here has been a combined Ensembl-Havana geneset, which incorporates protein-coding and non-coding transcripts annotated by the Havana team with the Ensembl automatic gene build. The major genome browsers have come together to produce a common set of identifiers where CDS annotations of transcripts can be agreed. This annotation was produced based on the NCBI36 assembly. The CDS identifiers have been mapped onto the new annotations based on the latest GRCh37 assembly and these identifiers are also shown.

- More information about the [CCDS project](#).

The [ENCODE](#) (ENCyclopedia OF DNA Elements) project aims to find functional elements in the human genome.

- More information about the [ENCODE resources](#) at Ensembl.

**Vega\*** Additional manual annotation of this genome can be found in [Vega](#)

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

**Search Ensembl**

**Results Summary** [Result in Detail >](#)

By Species	
Total	282
Homo sapiens	282
Gene (282)	

By Feature type	
Total	282
Gene	282

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

Your query matched 282 entries in the search database. Viewing hits 1-10

[1](#) [2](#) [3](#) [4](#) ... [26](#) [27](#) [28](#) [29](#)

**Ensembl protein\_coding Gene: ENSG00000163914 (HGNC (automatic): RHO) [Region in detail]**

Ensembl protein\_coding gene ENSG00000163914 has 1 transcript: ENST00000296271, associated with ENSE00001152199, ENSE00001152205, ENSE00001152211

**Rhodopsin (Opsin-2) [Source:UniProtKB/Swiss-Prot;Acc:P08100]**

The gene has the following external identifiers mapped to it:

CCDS: CCDS3063, CCDS3063.1

EMBL: BC112106, U16824, U49742, S81166, BX537381, BC112106

EntrezGene: 6010, RP4, RHO, MGC138311, CSNBAD1, MGC138300

GO: GO:0016056, GO:0009583, GO:0009416, GO:0007601, GO:0004536, GO:0004503, GO:0007602, GO:0009881, GO:0007186, GO:0009586, GO:0007165, GO:0006468, GO:0018298, GO:0001750, GO:0005622, GO:0005623, GO:0005624, GO:0005625, GO:0005626, GO:0005627, GO:0005628, GO:0005629, GO:0005630, GO:0005631, GO:0005632, GO:0005633, GO:0005634, GO:0005635, GO:0005636, GO:0005637, GO:0005638, GO:0005639, GO:0005640, GO:0005641, GO:0005642, GO:0005643, GO:0005644, GO:0005645, GO:0005646, GO:0005647, GO:0005648, GO:0005649, GO:0005650, GO:0005651, GO:0005652, GO:0005653, GO:0005654, GO:0005655, GO:0005656, GO:0005657, GO:0005658, GO:0005659, GO:0005660, GO:0005661, GO:0005662, GO:0005663, GO:0005664, GO:0005665, GO:0005666, GO:0005667, GO:0005668, GO:0005669, GO:0005670, GO:0005671, GO:0005672, GO:0005673, GO:0005674, GO:0005675, GO:0005676, GO:0005677, GO:0005678, GO:0005679, GO:0005680, GO:0005681, GO:0005682, GO:0005683, GO:0005684, GO:0005685, GO:0005686, GO:0005687, GO:0005688, GO:0005689, GO:0005690, GO:0005691, GO:0005692, GO:0005693, GO:0005694, GO:0005695, GO:0005696, GO:0005697, GO:0005698, GO:0005699, GO:0005700, GO:0005701, GO:0005702, GO:0005703, GO:0005704, GO:0005705, GO:0005706, GO:0005707, GO:0005708, GO:0005709, GO:0005710, GO:0005711, GO:0005712, GO:0005713, GO:0005714, GO:0005715, GO:0005716, GO:0005717, GO:0005718, GO:0005719, GO:0005720, GO:0005721, GO:0005722, GO:0005723, GO:0005724, GO:0005725, GO:0005726, GO:0005727, GO:0005728, GO:0005729, GO:0005730, GO:0005731, GO:0005732, GO:0005733, GO:0005734, GO:0005735, GO:0005736, GO:0005737, GO:0005738, GO:0005739, GO:0005740, GO:0005741, GO:0005742, GO:0005743, GO:0005744, GO:0005745, GO:0005746, GO:0005747, GO:0005748, GO:0005749, GO:0005750, GO:0005751, GO:0005752, GO:0005753, GO:0005754, GO:0005755, GO:0005756, GO:0005757, GO:0005758, GO:0005759, GO:0005760, GO:0005761, GO:0005762, GO:0005763, GO:0005764, GO:0005765, GO:0005766, GO:0005767, GO:0005768, GO:0005769, GO:0005770, GO:0005771, GO:0005772, GO:0005773, GO:0005774, GO:0005775, GO:0005776, GO:0005777, GO:0005778, GO:0005779, GO:0005780, GO:0005781, GO:0005782, GO:0005783, GO:0005784, GO:0005785, GO:0005786, GO:0005787, GO:0005788, GO:0005789, GO:0005790, GO:0005791, GO:0005792, GO:0005793, GO:0005794, GO:0005795, GO:0005796, GO:0005797, GO:0005798, GO:0005799, GO:0005800, GO:0005801, GO:0005802, GO:0005803, GO:0005804, GO:0005805, GO:0005806, GO:0005807, GO:0005808, GO:0005809, GO:0005810, GO:0005811, GO:0005812, GO:0005813, GO:0005814, GO:0005815, GO:0005816, GO:0005817, GO:0005818, GO:0005819, GO:0005820, GO:0005821, GO:0005822, GO:0005823, GO:0005824, GO:0005825, GO:0005826, GO:0005827, GO:0005828, GO:0005829, GO:0005830, GO:0005831, GO:0005832, GO:0005833, GO:0005834, GO:0005835, GO:0005836, GO:0005837, GO:0005838, GO:0005839, GO:0005840, GO:0005841, GO:0005842, GO:0005843, GO:0005844, GO:0005845, GO:0005846, GO:0005847, GO:0005848, GO:0005849, GO:0005850, GO:0005851, GO:0005852, GO:0005853, GO:0005854, GO:0005855, GO:0005856, GO:0005857, GO:0005858, GO:0005859, GO:0005860, GO:0005861, GO:0005862, GO:0005863, GO:0005864, GO:0005865, GO:0005866, GO:0005867, GO:0005868, GO:0005869, GO:0005870, GO:0005871, GO:0005872, GO:0005873, GO:0005874, GO:0005875, GO:0005876, GO:0005877, GO:0005878, GO:0005879, GO:0005880, GO:0005881, GO:0005882, GO:0005883, GO:0005884, GO:0005885, GO:0005886, GO:0005887, GO:0005888, GO:0005889, GO:0005890, GO:0005891, GO:0005892, GO:0005893, GO:0005894, GO:0005895, GO:0005896, GO:0005897, GO:0005898, GO:0005899, GO:0005900, GO:0005901, GO:0005902, GO:0005903, GO:0005904, GO:0005905, GO:0005906, GO:0005907, GO:0005908, GO:0005909, GO:0005910, GO:0005911, GO:0005912, GO:0005913, GO:0005914, GO:0005915, GO:0005916, GO:0005917, GO:0005918, GO:0005919, GO:0005920, GO:0005921, GO:0005922, GO:0005923, GO:0005924, GO:0005925, GO:0005926, GO:0005927, GO:0005928, GO:0005929, GO:0005930, GO:0005931, GO:0005932, GO:0005933, GO:0005934, GO:0005935, GO:0005936, GO:0005937, GO:0005938, GO:0005939, GO:0005940, GO:0005941, GO:0005942, GO:0005943, GO:0005944, GO:0005945, GO:0005946, GO:0005947, GO:0005948, GO:0005949, GO:0005950, GO:0005951, GO:0005952, GO:0005953, GO:0005954, GO:0005955, GO:0005956, GO:0005957, GO:0005958, GO:0005959, GO:0005960, GO:0005961, GO:0005962, GO:0005963, GO:0005964, GO:0005965, GO:0005966, GO:0005967, GO:0005968, GO:0005969, GO:0005970, GO:0005971, GO:0005972, GO:0005973, GO:0005974, GO:0005975, GO:0005976, GO:0005977, GO:0005978, GO:0005979, GO:0005980, GO:0005981, GO:0005982, GO:0005983, GO:0005984, GO:0005985, GO:0005986, GO:0005987, GO:0005988, GO:0005989, GO:0005990, GO:0005991, GO:0005992, GO:0005993, GO:0005994, GO:0005995, GO:0005996, GO:0005997, GO:0005998, GO:0005999, GO:0006000, GO:0006001, GO:0006002, GO:0006003, GO:0006004, GO:0006005, GO:0006006, GO:0006007, GO:0006008, GO:0006009, GO:0006010, GO:0006011, GO:0006012, GO:0006013, GO:0006014, GO:0006015, GO:0006016, GO:0006017, GO:0006018, GO:0006019, GO:0006020, GO:0006021, GO:0006022, GO:0006023, GO:0006024, GO:0006025, GO:0006026, GO:0006027, GO:0006028, GO:0006029, GO:0006030, GO:0006031, GO:0006032, GO:0006033, GO:0006034, GO:0006035, GO:0006036, GO:0006037, GO:0006038, GO:0006039, GO:0006040, GO:0006041, GO:0006042, GO:0006043, GO:0006044, GO:0006045, GO:0006046, GO:0006047, GO:0006048, GO:0006049, GO:0006050, GO:0006051, GO:0006052, GO:0006053, GO:0006054, GO:0006055, GO:0006056, GO:0006057, GO:0006058, GO:0006059, GO:0006060, GO:0006061, GO:0006062, GO:0006063, GO:0006064, GO:0006065, GO:0006066, GO:0006067, GO:0006068, GO:0006069, GO:0006070, GO:0006071, GO:0006072, GO:0006073, GO:0006074, GO:0006075, GO:0006076, GO:0006077, GO:0006078, GO:0006079, GO:0006080, GO:0006081, GO:0006082, GO:0006083, GO:0006084, GO:0006085, GO:0006086, GO:0006087, GO:0006088, GO:0006089, GO:0006090, GO:0006091, GO:0006092, GO:0006093, GO:0006094, GO:0006095, GO:0006096, GO:0006097, GO:0006098, GO:0006099, GO:0006100, GO:0006101, GO:0006102, GO:0006103, GO:0006104, GO:0006105, GO:0006106, GO:0006107, GO:0006108, GO:0006109, GO:0006110, GO:0006111, GO:0006112, GO:0006113, GO:0006114, GO:0006115, GO:0006116, GO:0006117, GO:0006118, GO:0006119, GO:0006120, GO:0006121, GO:0006122, GO:0006123, GO:0006124, GO:0006125, GO:0006126, GO:0006127, GO:0006128, GO:0006129, GO:0006130, GO:0006131, GO:0006132, GO:0006133, GO:0006134, GO:0006135, GO:0006136, GO:0006137, GO:0006138, GO:0006139, GO:0006140, GO:0006141, GO:0006142, GO:0006143, GO:0006144, GO:0006145, GO:0006146, GO:0006147, GO:0006148, GO:0006149, GO:0006150, GO:0006151, GO:0006152, GO:0006153, GO:0006154, GO:0006155, GO:0006156, GO:0006157, GO:0006158, GO:0006159, GO:0006160, GO:0006161, GO:0006162, GO:0006163, GO:0006164, GO:0006165, GO:0006166, GO:0006167, GO:0006168, GO:0006169, GO:0006170, GO:0006171, GO:0006172, GO:0006173, GO:0006174, GO:0006175, GO:0006176, GO:0006177, GO:0006178, GO:0006179, GO:0006180, GO:0006181, GO:0006182, GO:0006183, GO:0006184, GO:0006185, GO:0006186, GO:0006187, GO:0006188, GO:0006189, GO:0006190, GO:0006191, GO:0006192, GO:0006193, GO:0006194, GO:0006195, GO:0006196, GO:0006197, GO:0006198, GO:0006199, GO:0006200, GO:0006201, GO:0006202, GO:0006203, GO:0006204, GO:0006205, GO:0006206, GO:0006207, GO:0006208, GO:0006209, GO:0006210, GO:0006211, GO:0006212, GO:0006213, GO:0006214, GO:0006215, GO:0006216, GO:0006217, GO:0006218, GO:0006219, GO:0006220, GO:0006221, GO:0006222, GO:0006223, GO:0006224, GO:0006225, GO:0006226, GO:0006227, GO:0006228, GO:0006229, GO:0006230, GO:0006231, GO:0006232, GO:0006233, GO:0006234, GO:0006235, GO:0006236, GO:0006237, GO:0006238, GO:0006239, GO:0006240, GO:0006241, GO:0006242, GO:0006243, GO:0006244, GO:0006245, GO:0006246, GO:0006247, GO:0006248, GO:0006249, GO:0006250, GO:0006251, GO:0006252, GO:0006253, GO:0006254, GO:0006255, GO:0006256, GO:0006257, GO:0006258, GO:0006259, GO:0006260, GO:0006261, GO:0006262, GO:0006263, GO:0006264, GO:0006265, GO:0006266, GO:0006267, GO:0006268, GO:0006269, GO:0006270, GO:0006271, GO:0006272, GO:0006273, GO:0006274, GO:0006275, GO:0006276, GO:0006277, GO:0006278, GO:0006279, GO:0006280, GO:0006281, GO:0006282, GO:0006283, GO:0006284, GO:0006285, GO:0006286, GO:0006287, GO:0006288, GO:0006289, GO:0006290, GO:0006291, GO:0006292, GO:0006293, GO:0006294, GO:0006295, GO:0006296, GO:0006297, GO:0006298, GO:0006299, GO:0006300, GO:0006301, GO:0006302, GO:0006303, GO:0006304, GO:0006305, GO:0006306, GO:0006307, GO:0006308, GO:0006309, GO:0006310, GO:0006311, GO:0006312, GO:0006313, GO:0006314, GO:0006315, GO:0006316, GO:0006317, GO:0006318, GO:0006319, GO:0006320, GO:0006321, GO:0006322, GO:0006323, GO:0006324, GO:0006325, GO:0006326, GO:0006327, GO:0006328, GO:0006329, GO:0006330, GO:0006331, GO:0006332, GO:0006333, GO:0006334, GO:0006335, GO:0006336, GO:0006337, GO:0006338, GO:0006339, GO:0006340, GO:0006341, GO:0006342, GO:0006343, GO:0006344, GO:0006345, GO:0006346, GO:0006347, GO:0006348, GO:0006349, GO:0006350, GO:0006351, GO:0006352, GO:0006353, GO:0006354, GO:0006355, GO:0006356, GO:0006357, GO:0006358, GO:0006359, GO:0006360, GO:0006361, GO:0006362, GO:0006363, GO:0006364, GO:0006365, GO:0006366, GO:0006367, GO:0006368, GO:0006369, GO:0006370, GO:0006371, GO:0006372, GO:0006373, GO:0006374, GO:0006375, GO:0006376, GO:0006377, GO:0006378, GO:0006379, GO:0006380, GO:0006381, GO:0006382, GO:0006383, GO:0006384, GO:0006385, GO:0006386, GO:0006387, GO:0006388, GO:0006389, GO:0006390, GO:0006391, GO:0006392, GO:0006393, GO:0006394, GO:0006395, GO:0006396, GO:0006397, GO:0006398, GO:0006399, GO:0006400, GO:0006401, GO:0006402, GO:0006403, GO:0006404, GO:0006405, GO:0006406, GO:0006407, GO:0006408, GO:0006409, GO:0006410, GO:0006411, GO:0006412, GO:0006413, GO:0006414, GO:0006415, GO:0006416, GO:0006417, GO:0006418, GO:0006419, GO:0006420, GO:0006421, GO:0006422, GO:0006423, GO:0006424, GO:0006425, GO:0006426, GO:0006427, GO:0006428, GO:0006429, GO:0006430, GO:0006431, GO:0006432, GO:0006433, GO:0006434, GO:0006435, GO:0006436, GO:0006437, GO:0006438, GO:0006439, GO:0006440, GO:0006441, GO:0006442, GO:0006443, GO:0006444, GO:0006445, GO:0006446, GO:0006447, GO:0006448, GO:0006449, GO:0006450, GO:0006451, GO:0006452, GO:0006453, GO:0006454, GO:0006455, GO:0006456, GO:0006457, GO:0006458, GO:0006459, GO:0006460, GO:0006461, GO:0006462, GO:0006463, GO:0006464, GO:0006465, GO:0006466, GO:0006467, GO:0006468, GO:0006469, GO:0006470, GO:0006471, GO:0006472, GO:0006473, GO:0006474, GO:0006475, GO:0006476, GO:0006477, GO:0006478, GO:0006479, GO:0006480, GO:0006481, GO:0006482, GO:0006483, GO:0006484, GO:0006485, GO:0006486, GO:0006487, GO:0006488, GO:0006489, GO:0006490, GO:0006491, GO:0006492, GO:0006493, GO:0006494, GO:0006495, GO:0006496, GO:0006497, GO:0006498, GO:0006499, GO:0006500, GO:0006501, GO:0006502, GO:0006503, GO:0006504, GO:0006505, GO:0006506, GO:0006507, GO:0006508, GO:0006509, GO:0006510, GO:0006511, GO:0006512, GO:0006513, GO:0006514, GO:0006515, GO:0006516, GO:0006517, GO:0006518, GO:0006519, GO:0006520, GO:0006521, GO:0006522, GO:0006523, GO:0006524, GO:0006525, GO:0006526, GO:0006527, GO:0006528, GO:0006529, GO:0006530, GO:0006531, GO:0006532, GO:0006533, GO:0006534, GO:0006535, GO:0006536, GO:0006537, GO:0006538, GO:0006539, GO:0006540, GO:0006541, GO:0006542, GO:0006543, GO:0006544, GO:0006545, GO:0006546, GO:0006547, GO:0006548, GO:0006549, GO:0006550, GO:0006551, GO:0006552, GO:0006553, GO:0006554, GO:0006555, GO:0006556, GO:0006557, GO:0006558, GO:0006559, GO:0006560, GO:0006561, GO:0006562, GO:0006563, GO:0006564, GO:0006565, GO:0006566, GO:0006567, GO:0006568, GO:0006569, GO:0006570, GO:0006571, GO:0006572, GO:0006573, GO:0006574, GO:0006575, GO:0006576, GO:0006577, GO:0006578, GO:0006579, GO:0006580, GO:0006581, GO:0006582, GO:0006583, GO:0006584, GO:0006585, GO:0006586, GO:0006587, GO:0006588, GO:0006589, GO:0006590, GO:0006591, GO:0006592, GO:0006593, GO:0006594, GO:0006595, GO:0006596, GO:0006597, GO:0006598, GO:0006599, GO:0006600, GO:0006601, GO:0006602, GO:0006603, GO:0006604, GO:0006605, GO:0006606, GO:0006607, GO:0006608, GO:0006609, GO:0006610, GO:0006611, GO:0006612, GO:0006613, GO:0006614, GO:0006615, GO:0006616, GO:0006617, GO:0006618, GO:0006619, GO:0006620, GO:0006621, GO:0006622, GO:0006623, GO:0006624, GO:0006625, GO:0006626, GO:0006627, GO:0006628, GO:0006629, GO:0006630, GO:0006631, GO:0006632, GO:0006633, GO:0006634, GO:0006635, GO:0006636, GO:0006637, GO:0006638, GO:0006639, GO:0006640, GO:0006641, GO:0006642, GO:0006643, GO:0006644, GO:0006645, GO:0006646, GO:0006647, GO:0006648, GO:0006649, GO:0006650, GO:0006651, GO:0006652, GO:0006653, GO:0006654, GO:0006655, GO:0006656, GO:0006657, GO:0006658, GO:0006659, GO:0006660, GO:0006661, GO:0006662, GO:0006663, GO:0006664, GO:0006665, GO:0006666, GO:0006667, GO:0006668, GO:0006669, GO:0006670, GO:0006671, GO:0006672, GO:0006673, GO

Location: 3:129,247,482-129,254,177 Gene: RHO Transcript: RHO-201

**Gene: RHO (ENSG00000163914)**  
 Rhodopsin (Opsin-2) Source: UniProtKB/Swiss-Prot\_P08100

**Location** [Chromosome 3: 129,247,482-129,254,177 forward strand.](#)

**Transcripts** There is 1 transcript in this gene: [hide transcripts](#)

Name	Transcript ID	Protein ID	Description
RHO-201	ENST00000296271	ENSP00000296271	protein_coding

[Gene summary](#) [help](#)

**Name** [RHO](#) (HGNC (automatic))

**Synonyms** OPN2, RP4 [To view all Ensembl genes linked to the name [click here.](#)]

**CCDS** This gene is a member of the Human CCDS set: [CCDS3063](#)

**Gene type** Known protein coding

**Prediction Method** Transcripts were annotated by the Ensembl [genebuild](#).

**Transcripts**

Transcripts for the nearby IFT122 gene

Blue bar is the genome

RHO transcript

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.

**Gene-based displays**

- [Gene summary](#)
- [Splice variants \(1\)](#)
- [Supporting evidence](#)
- [Sequence](#)
- [External references \(2\)](#)
- [Regulation](#)
- Comparative Genomics
  - [Genomic alignments \(38\)](#)
  - Gene Tree (image)
    - [Gene Tree \(text\)](#)
    - [Gene Tree \(alignment\)](#)
  - [Orthologues \(28\)](#)
  - [Paralogues \(0\)](#)
  - [Protein families \(1\)](#)
- Genetic Variation
  - [Variation Table](#)
  - [Variation Image](#)
- External Data
  - [Personal annotation](#)
- ID History
  - [Gene history](#)

• [Configure this page](#)

• [Manage your data](#)

• [Export data](#)

• [Bookmark this page](#)

Click [Sequence](#)

**THIS STYLE:** Location of ENSG00000163914 exons  
**THIS STYLE:** Location of Ensembl exons

```
>chromosome:GRCh37:3:129246882:129254777:1
TCTTTGTATTCCAGGGCCCTGCAAATAAATGTTTAAATGAACGAACAAGAGAGTGAATTTC
CAATTCCATGCAACAAGGATTGGGCTCCTGGGCCCTAGGCTATGTGTCTGGCACCAGAAA
CGGAAGTGCAGGTTGCAGCCCTG6CCCTCATGGAGCTCCTCCTGTGACAGGAGTGTGGG
GACTGGATGACTCCAGAGGTAACCTGTGGGGGAACGAACAGGTAAGGGGCTGTGTGACGA
GATGAGAGACTGGGAGAATAAACAGAAAGTCTCTAGCTGTCCAGAGGACATAGCACAGA
GGCCCATGGTCCCTATTTCAAACCAGGCCACCAGACTGAGCTGGGACCTTGGGACAGAC
AAGTCATGCAGAAAGTTAGGGGACCTTCTCCTCCCTTTCTGATCCTGAGTACCTCTCC
TCCCTGACCTCAGGCTTCCCTCTAGTGTACCTTGGCCCTCTTAGAAGCCAATTAGGCC
CTCAGTTTCTGCAGCGGGGATTAATATGATTATGAACACCCCAATCTCCAGATGCTGA
TTCAGCCAGGAGCTTAGGAGGGGAGGTCACCTTTATAAGGGTCTGGGGGGTTCAGAACCC
AGAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCRGGCCTTCGCRGCTTCTTGGGTG
GGAGCRGCCACGGGTCRGGCCRAGGGCCRAGCCRTGARTGGCRAGAGGGCCCTRACT
TCTACGTGCCCTTCTCCARTGCRAGGGGTGGTTCRAGCCCTTCGAGTACCCRAGT
ACTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCTACATGTTTCTGCTGATCG
TGCTGGGCTTCCCCRTCAACTTCTCAGCTCTCRGCTCRGGTCCRGGCRAGRAGCTGC
GCACGCTCTCAACTACATCCGCTCAACCTAGCCGTTGGCTGACCTCTTCTATGGTCTGAG
GTGGCTTCAACRAGCACCTCTACRCCCTCTGCRGGTACTTCTGCTTGGGGCCRAG
GATGCAATTTGGAGGGCTTCTTTCRCCCTGGGGCGTATGAGCCGGGTGTGGTGGGGT
GTGCAGGAGCCCGGAGCATGGAGGGGCTCTGGGAGAGTCCCGGGCTTGGCGGTGGTGGCT
.....
```

Upstream sequence

Exon sequence

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](#).

Gene: IL2 (ENSG00000109471)

Configure page Custom Data Your account **SAVE and close**

Configuration for: "Marked up gene sequence"

5' Flanking sequence: 600

3' Flanking sequence: 600

Number of base pairs per row: 60 bps

Additional exons to display: Core exons

Orientation of additional exons: Display exons in both orientations

Show variations: Yes and show links

Line numbering: Relative to this sequence

**Display variations**

**Turn on line numbers**

**DAS sources**

- ArrayExpress Warehouse  
Gene expression profile thumbnails from the ArrayExpress warehouse [Homepage]
- cbs\_func  
CBS Protein function and structure predictions [Homepage]
- cbs\_ptm  
CBS Post-translational modification site predictions [Homepage]
- cbs\_sort  
CBS Protein sorting predictions [Homepage]
- GAD  
Genetic Association Database - association of diseases to human Entrez genes [Homepage]

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).

**THIS STYLE:** Location of ENSG00000163914 exons  
**THIS STYLE:** Location of Ensembl exons  
**THIS STYLE:** Location of SNPs  
**THIS STYLE:** Location of inserts  
**THIS STYLE:** Location of deletes

```
>chromosome: GRCh37:3:129246882:129254777:1
1 TCTTTGTATTCCACGGGGCTGCAATAAATGTTAATGAACGAACAAGAGAGTGAATTC 60
61 CAATTCATGCAACAAGGATTGGGCTCCTGGGCCCTAGGCTATGTGTCTGGCACCAGAAA 120
121 CGGAAAGCTCAGAGTTGACAGCCCTGCCTCATGGAGCTCCTCCTGTGACAGGAGTGTGG 180
181 GACTGGAGACTCCAGAGTAACTTGTGGGGAACRACAGGTAAGGGGCTGTGTGACGA 240
241 GATGAGAGACTGGGAGATAAACAGAAAGTCTCTAGCTGTCCAGAGGACATAGCACAGA 300
301 GGCCCATGGTCCCTATTTCAAAACCCAGGGCCACCAGACTGAGCTGGGACCTTGGGACAGAC 360
361 AAGTCATGCGAAGTTAGGGGACCTTCTCCTCCCTTTTCTGGATCCTGAGTACCTCTCC 420
421 TCCTTGACCTCAGGCTTCTCCTAGTGTACCTTGGCCCTCTTAGAAGCCAATTAGGCC 480
481 CTCAGTTTCTGCAGCGGGGATTAATATGATTATGAACACCCCAATCTCCAGATGCTGA 540
541 TTCAGCCAGGAGCTTAGGAGGGGGAGGTCACTTTATAAGGGTCTGGGGGGTTCAGAACCC 600
601 AGAGTCTACAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTTCRCAGCATTCTTGGGTG 660 645:G/A;
GGAGCAGCCRCGGGTCAAGCCACAGGGCCACAGCCRTGATGGCRCAGAGGCCCTACT 720 670:A/G;
TCTACGTGCCCTTCTCCAAATGCGAGGGTGTGGTACGCAGCCCTTCGAGTACCCACAGT 780
ACTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCTACATGTTTCTGCTGATCG 840
781 TGCTGGGCTTCCSCATCACTTCTCAGGCTCAGCTCAGCTCCAGCACAAGAGCTGC 900 853:C/G; 868:C/G;
901 GCAGCCCTCTCACTACATCTGCTCAACCTAGCCGTGGCTGACCTCTTCATGGTCTTAG 960
961 GTGGCTTACCAGCACCCTCTACACCTCTCTGCTGGATCTTCTGTTTCRGGCCACAG 1020 1011:G/A;
1021 GATGCATTTGGAGGGCTCTTTGGCCACCCTGGGCGGTATGAGCCGGGTGTGGGTGGGGT 1080
1081 GTCTAGGAGCCCGGAGATGCAAGGGCTTGGGAGACTTCCGGCTTGGGCTGTGGCT
```

Link to variation information

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.

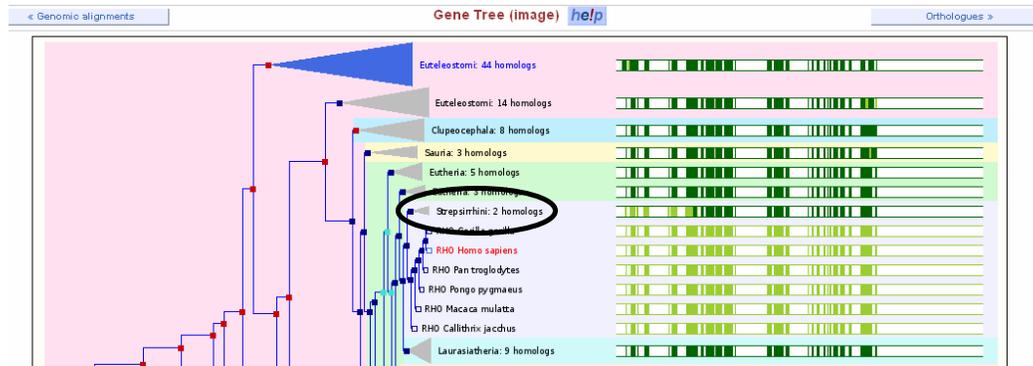
Click [Configure this page](#) at the left. Turn on conservation regions in the menu.

```
Homo_sapiens CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCAGAGTATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Pan_troglodytes CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCAGAGTATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Pongo_pygmaeus CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCAGAGTATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Hacaca_aulatta CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCAGAGTATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Rattus_norvegicus CTGAATCAGCTCTTGGCTTAGGAGGAGAAAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGTGGCTGGAGTGTGCTGTGGAGCCGTCAAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Bos_taurus CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGTGGCTGGAGTGTGCTGTGGAGCCGTCAAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Sus_scrofa CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCCTGAGTGTGCTGTGGAGCCGTCAAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Equus_familiaris CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCCTGAGTGTGCTGTGGAGCCGTCAAGTGGCTGAGCTCAGCCCTTCGACAGCATT
Equus_caballus CTGATTACAGCCAGGAGCTTAGGAGGGGAGGTCACTTTATAAGGGCTCTGGG-G---GGGTACAGACCCTGAGTGTGCTGTGGAGCCGTCAAGTGGCTGAGCTCAGCCCTTCGACAGCATT

Homo_sapiens CTTG-----GCTGAGAC-GC-CAGGGGTACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Pan_troglodytes CTTG-----GCTGAGAC-GC-CGTGGGTACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Pongo_pygmaeus CTTG-----GCTGAGAC-GC-CGCGGGGACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Hacaca_aulatta CTTG-----GCTGAGAC-GA-CGCGGGGACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Rattus_norvegicus CTTGGTCTCTGTCTACGAAGACC-CGTGGGGAGCTGCA-8AGCCGACCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Bos_taurus CTTG-----GCTTGGGACC-GC-CGCGGGGACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Sus_scrofa CTTG-----GACTGAGCC-86CCACAGGGGACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Equus_familiaris CTTA-----GACTGAGCC-86CCACAGGGGACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
Equus_caballus CTTG-----GCTGAGAC-86CCACAGGGGACGCCAAA-866CCACAGCCATGACAGAGAGGCCCTAACTTCTACCTGCGCTTCTCCAAATGCGAGCGGTGTGGTACGACGCC
```

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](#)

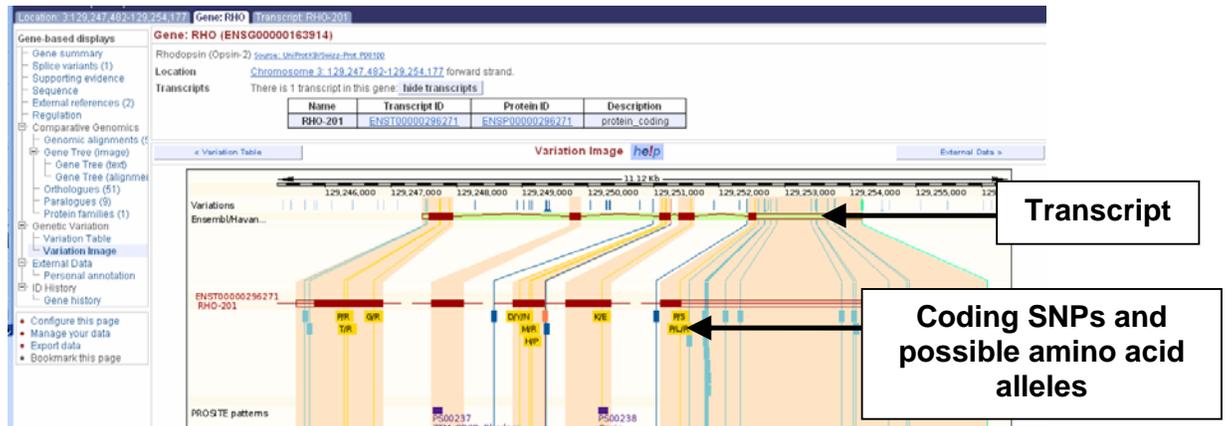
The following gene(s) have been identified as putative orthologues:  
 (N.B. If you don't find a homologue here, it may be a "between-species paralogue". Please view the [gene tree info](#) to see more.)

Species	Type	dN/dS	Ensembl identifier	External ref.
Alpaca ( <i>Vicugna pacos</i> )	1-to-1	na	<a href="#">ENSVFAG0000003334</a> Target 'id: 77; Query 'id: 76 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P08100</a> ]
Anole Lizard ( <i>Anolis carolinensis</i> )	1-to-1	na	<a href="#">ENSACAG00000014258</a> Target 'id: 81; Query 'id: 82 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	OPSD_ANOCA Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P41691</a> ]
Armadillo ( <i>Dasypus novemcinctus</i> )	1-to-1	na	<a href="#">ENSDNQG00000015133</a> Target 'id: 61; Query 'id: 40 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P08100</a> ]
Bushbaby ( <i>Otolemur garnettii</i> )	1-to-1	na	<a href="#">ENSOGAG00000010259</a> Target 'id: 78; Query 'id: 77 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P08100</a> ]
Cat ( <i>Felis catus</i> )	1-to-1	na	<a href="#">ENSFCAG00000000092</a> Target 'id: 96; Query 'id: 96 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	OPSD_FELCA Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: <a href="#">Q951KU1</a> ]
Chicken ( <i>Gallus gallus</i> )	1-to-1	na	<a href="#">ENSGALG00000020745</a> Target 'id: 75; Query 'id: 50 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P08100</a> ]
Chimpanzee ( <i>Pan troglodytes</i> )	1-to-1	0.00000	<a href="#">ENSPTRG00000015379</a> Target 'id: 100; Query 'id: 100 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	RHO Rhodopsin (Opsin-2) [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P08100</a> ]
Cow ( <i>Bos taurus</i> )	1-to-1	0.03488	<a href="#">ENSBTAG0000001310</a> Target 'id: 92; Query 'id: 93 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	OPSD_BOVIN Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P02690</a> ]
Dog ( <i>Canis familiaris</i> )	1-to-1	0.04946	<a href="#">ENSFCAG00000004633</a> Target 'id: 90; Query 'id: 95 <a href="#">[Multi-species view]</a> <a href="#">[Align]</a>	OPSD_CANFA Rhodopsin [Source: UniProtKB/Swiss-Prot; acc: <a href="#">P32308</a> ]

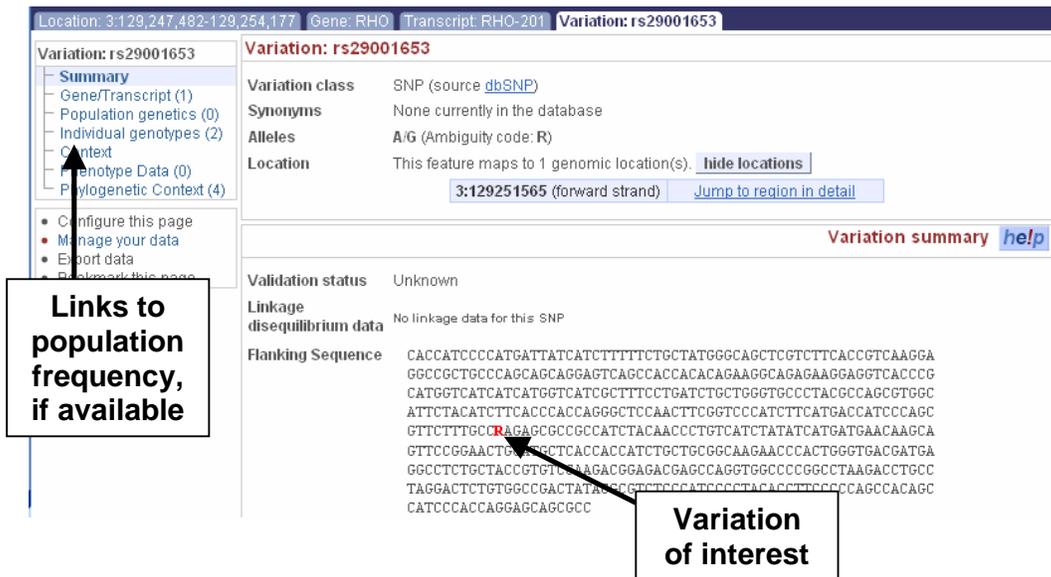
Let's view genetic variation mapped onto all transcripts of a gene.

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.



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.

Number of base pairs per row:	60 bps <input type="button" value="v"/>
Show exons:	Yes <input type="button" value="v"/>
Show codons:	Yes <input type="button" value="v"/>
Show coding sequence:	Yes <input type="button" value="v"/>
Show protein sequence:	Yes <input type="button" value="v"/>
Show RNA features:	No <input type="button" value="v"/>
Show variation features:	Yes <input type="button" value="v"/>
Number residues:	Yes <input type="button" value="v"/>

Click [Save and close](#).

```

1  AGAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTTCRGCAGCATTCTTGGGTG
.....
61  GGAGCAGCCRACGGGTACGCCACAAGGGCCACAGCCATGAATGGCACAGAAGGCCCTAACT
.....ATGAATGGCACAGAAGGCCCTAACT
.....-M--N--G--T--E--G--P--N--

121 TCTACGTGCCCTTCTCCAATGGGACGGGTGTGGTACGCAGCCCTTCGACTACCCACAGT
26 TCTACGTGCCCTTCTCCAATGGGACGGGTGTGGTACGCAGCCCTTCGACTACCCACAGT
9  F--Y--V--P--F--S--N--A--T--G--V--V--R--S--P--F--E--Y--P--Q--

181 ACTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCTACATGTTTCTGCTGATCG
86 ACTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCTACATGTTTCTGCTGATCG
29 Y--Y--L--A--E--P--W--Q--F--S--M--L--A--A--Y--M--F--L--L--I--

241 TGCTGGGCTTCCSCATCAACTTCCTCACSGCTCTACGTCACCGTCCAGCACAGAAGCTGC
146 TGCTGGGCTTCCCATCAACTTCCTCACGCTCTACGTCACCGTCCAGCACAGAAGCTGC
49 V--L--G--F--P--I--N--F--L--T--L--Y--V--T--V--Q--H--K--K--L--

```

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.

Name	Transcript ID	Protein ID	Description
RHO-201	<a href="#">ENST00000296271</a>	<a href="#">ENSP00000296271</a>	protein_coding

[← Protein sequence](#)
[General identifiers](#) [help](#)

This Ensembl gene entry corresponds to the following database identifiers:

**HGNC Symbol:** [RHO](#)  
 rhodopsin [\[view all locations\]](#)

**CCDS:** [CCDS3063.1](#) [\[view all locations\]](#)

**WikiGene:** [RHO](#)  
 rhodopsin [\[view all locations\]](#)

**UniProtKB/Swiss-Prot:** [OPSD\\_HUMAN](#) (Target %id: 100; Query %id: 100) [\[align\]](#)  
 Rhodopsin (Opsin-2) [\[view all locations\]](#)

**RefSeq peptide:** [NP\\_000530.1](#) (Target %id: 100; Query %id: 100) [\[align\]](#)  
 rhodopsin [\[view all locations\]](#)

**RefSeq DNA:** [NM\\_000539.3](#) [\[align\]](#)  
 rhodopsin (RHO), mRNA [\[view all locations\]](#)

**UniProtKB/TrEMBL:** [Q16415\\_HUMAN](#) (Target %id: 6; Query %id: 95) [\[align\]](#)  
 Rhodopsin protein Fragment [\[view all locations\]](#)

**EntrezGene:** [RHO](#)  
 rhodopsin [\[view all locations\]](#)

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

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

**Collapsed track: cDNA sequences aligned to the genome**

**RHO and neighbouring genes**

**Zoom into RHO gene**

**3 Configuring the display**  
 You currently have 15 tracks in the overview panel and 109 tracks in the main panel turned off. To change the tracks you are displaying, use the "Configure this page" link on the left.

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
AGAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTTCGCAGCATTCTTGGGTTG
GGAGCAGCCACGGGTTCAGCCACAAGGGCCACAGCCATGAATGGCACAGAAGGCCCTAACT
TCTACTGTCCCTTCTCCAATGCGACGGGTGTGGTACGCAGCCCTTCGAGTACCCACAGT
ACTACTGTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCTACATGTTTCTGTGATCG
TGCTGGGCTTCCCCATCAACTTCTCAGGCTCTACGTCACCGTCCAGCACAGAAGCTGC
GCAGGCTCTCAACTACATCCTGCTCAACCTAGCCGTGGCTGACCTCTTCATGGTCTTAG
GTGGCTTACACAGACCCTTACACCTCTCTGCATGGATACTTCGTTCCGGGCCACAG
GATGCAATTTGGAGGGCTTCTTTGCCACCCTGGGGGATGAGCCGGGTGTGGGTGGGTT
GTGCAGGAGCCCGGGAGCATGGAGGGTCTGGGAGAGTCCCGGGCTTGGCGTGGTGGCT
GAGAGGCCCTTCTCCTTCTCTGCTCTCAATGTTATCCAAAGCCCTCATATATTCAGT
CAACAAACACCATTCATGGTGATAGCCGGGCTGCTGTTTGTGCAGGGCTGGCACTGAACA
CTGCCCTTGATCTTATTTGGAGCAATATGCGCTTGTCTAATTTACAGCAAGAAAAGTGA
CTGAGGCTCAAGAAAGTCAAGCCCTGCTGGGGGTCACACAGGGACGGGTGCAGAGTT
GAGTTGGAAGCCCGCATCTATCTCGGGCCATGTTTGCAGCACCAAGCCCTCTGTTTCCCTT
GGAGCAGCTGTGCTGAGTCAGACCCAGGCTGGGCACTGAGGGAGAGCTGGGCAAGCCAGA
CCCTCCTCTCTGGGGGCCAAGCTCAGGGTGGGAAGTGGATTTTCCATTCCTCAGTCAT
TGGGCTTCTCCTGTGCTGGGCAATGGGCTCGGTCCCTCTGGCATCCTCTGCCCTCCCTC
TCAGCCCTGTCTCAGGTGCCCTCCAGCCCTCCCTGCCGCTTCCAAAGTCTCCTGGTGT
TGAGAAACCGCAAGCAGCCGCTCTGAAGCAGTTCCCTTTTGTCTTTAGAAATAATGTCTTGCA
TTTAACAGGAAAACAGATGGGGTGTGTCAGGGATAACAGATCCCACTTAACAGAGAGGAA
AACTGAGGCAGGGAGAGGGGAAGAGACTCATTTAGGGATGTGGCCAGGCAGCAACAAGAG
CCTAGGCTCTCTGGCTGTGATCCAGGAATATCTCTGCTGAGATCCAGGAGGAGACGCTAG
```

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*.

**new** **SETUP** ← CONFIG ← RESULTS ← DISPLAY **refresh** **Online Help**

**Important Notice**  
 We now used Dist as our default DNA search. This will make your query faster.

**Enter the Query Sequence**  
 Either Paste sequences (max 30 sequences) in FASTA or plain text  
 >3 chr1:chr3:3086 chr3:3086:158337:3:129247402:129254177  
 AGATTCATCCACTGGAGCCCTGATGTGGCTGACTGAGGCTTCGACGATTTCTTGG  
 GGAGCAGCCAGGGTCAGCCCAAGGGCCACAGCCATGATGGCCAGAGGCCCTA  
 TCTACGTGCCCTTCTCAATGCGAGGGGTGTGTACGAGCCCTTCGATACCCAC

Or Upload a file containing one or more FASTA sequences

Or Enter a sequence ID or accession (EMBL, UniProt, RefSeq)

Or Enter an existing ticket ID:

dna queries  
 peptide queries

**Select the databases to search against**  
 Select species:   
 Use 'ctrl' key to select multiple species:

dna database:   
 peptide database:

**Select the Search Tool**

Search sensitivity:   
 Optimize search parameters to find the following alignments

**About BlastView**  
 BlastView provides an integrated platform for sequence similarity searches against Ensembl databases, offering access to both BLAST and BLAT programs. We would like to hear your expressions or queries, especially regarding functionality that you would like to see provided in the future. Many thanks for your time. [Feedback](#)

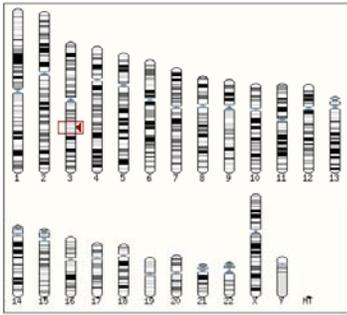
Ensembl release 56 - Sept 2009 © WTSI / EBI

[About Ensembl](#) | [Contact Us](#) | [Help](#)

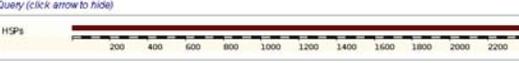
**new** **SETUP** ← CONFIG ← **RESULTS** ← DISPLAY **refresh** **Online Help**

Displaying 3 sequence alignments vs **Homo\_sapiens LATESTGP** database  
 Showing top 100 alignments of 1, sorted by Raw Score

Alignment Locations vs. Karyotype (click arrow to hide)



Alignment Locations vs. Query (click arrow to hide)



Alignment Summary (click arrow to hide)

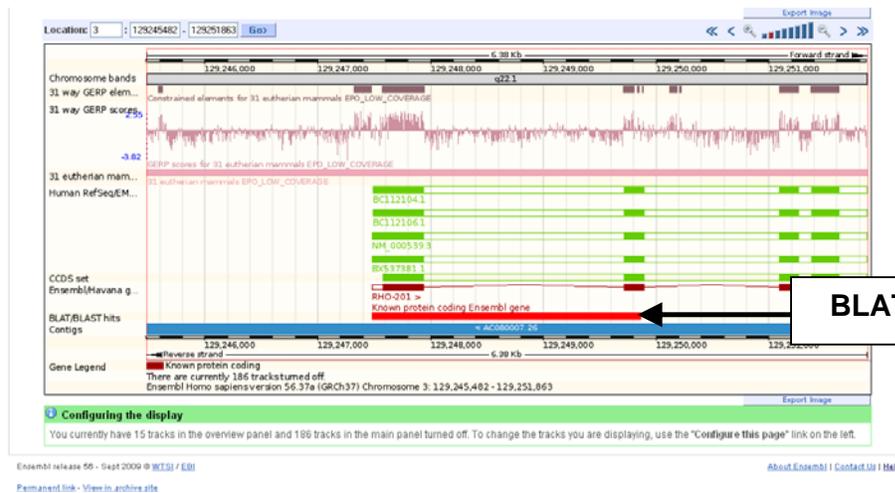
Select rows to include in table, and type of sort (Use the 'ctrl' key to select multiples)

Query	Subject	Chromosome	Supersongid	Clone	Contig	Chromosome	Chromosome	Chromosome	Stats	Sort By												
off																						
Name	Score	>Chromosome																				
Start	E-score	<Score																				
<a href="#">[A]</a>	<a href="#">[S]</a>	<a href="#">[G]</a>	<a href="#">[C]</a>	<a href="#">[L]</a>	<a href="#">[I]</a>	<a href="#">[O]</a>	<a href="#">[U]</a>	<a href="#">[D]</a>	<a href="#">[R]</a>	<a href="#">[E]</a>												
Links	Query	Chromosome	Name	Start	End	Ovl	Chromosome	Name	Start	End	Ovl	Chromosome	Name	Start	End	Ovl	Stats	Score	E-val	%ID	Length	

Ensembl release 56 - Sept 2009 © WTSI / EBI

[About Ensembl](#) | [Contact Us](#) | [Help](#)

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](http://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. Tachikins 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.

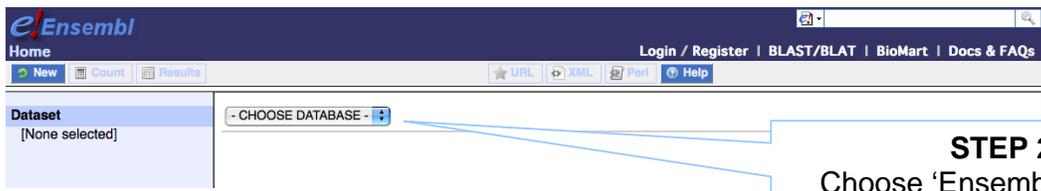
Which other genes with consensus coding sequences assigned by the CCDS project locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs?

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.

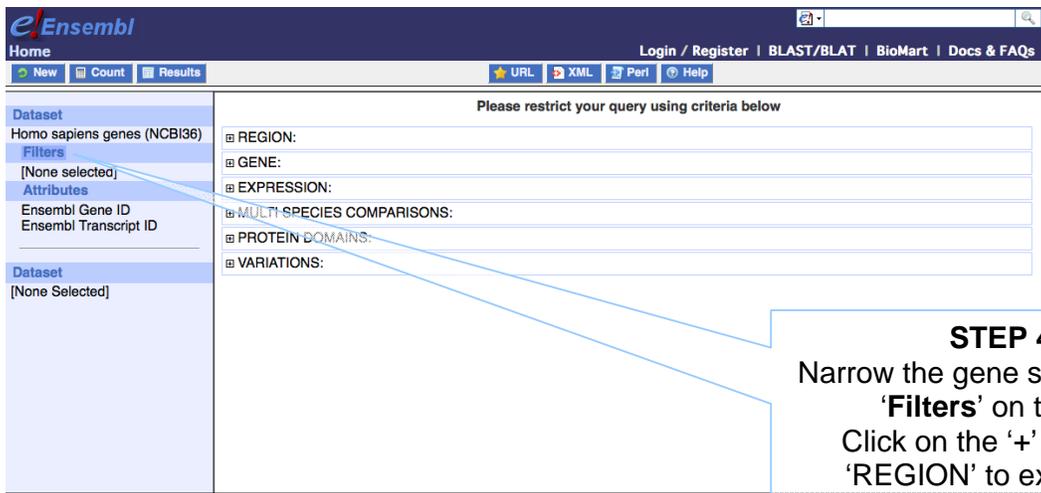
NOTE: These answers were determined using Ensembl 56.



**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)

**STEP 12:**  
 Select, along with the default options, 'Associated Gene name' (this shows the gene symbol from HGNC).

Note the summary of selected options.  
 The order of attributes determines the order of columns in the result table.

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 CDS with OTTT)
- HAVANA transcript (where ENST shares CDS with OTTT)
- 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 LOM 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.

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
ENSG00000186049	ENST00000370357	PASD1	139135		
ENSG00000110404	ENST00000370441	ID8	3423		
ENSG00000110404	ENST00000343955	ID8	3423	309900	MUCOPOLYSACCHARIDOSIS TYPE II
ENSG00000124334	ENST00000244173	IL9R	3581		
ENSG00000124333	ENST00000286448	VAMP7	6845		
ENSG00000168939	ENST00000309437	SPRY2	10251		
ENSG00000185973	ENST00000334399	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.

Export: all results to  TSV  Unique results only

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		
ENSG0000010404	ENST00000370441	IDS	3423		
ENSG0000010404	ENST00000340855	IDS	3423	309900	MUCOPOLYSACCHARIDOSIS TYPE II
ENSG00000124334	ENST00000244174	IL9R	3581		
ENSG00000124333	ENST00000286448	VAMP7	6845		
ENSG00000168939	ENST00000369437	SPRY3	10251		
ENSG00000185973	ENST00000334398	TMLHE	55217		
ENSG00000185978	ENST00000369444	H2AFB3	83740		
ENSG00000185978	ENST00000369444	H2AFB3	474381		
ENSG00000185990	ENST00000369445	F8A3	8263		

**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.

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	EntrezGene ID	MIM Morbid Accession	MIM Morbid Description
<a href="#">ENSG00000166049</a>	<a href="#">ENST00000370357</a>	<a href="#">PASD1</a>	<a href="#">139135</a>		
<a href="#">ENSG0000010404</a>	<a href="#">ENST00000370441</a>	<a href="#">IDS</a>	<a href="#">3423</a>		
<a href="#">ENSG0000010404</a>	<a href="#">ENST00000340855</a>	<a href="#">IDS</a>	<a href="#">3423</a>	<a href="#">309900</a>	<a href="#">MUCOPOLYSACCHARIDOSIS TYPE II</a>
<a href="#">ENSG00000124334</a>	<a href="#">ENST00000244174</a>	<a href="#">IL9R</a>	<a href="#">3581</a>		
<a href="#">ENSG00000124333</a>	<a href="#">ENST00000286448</a>	<a href="#">VAMP7</a>	<a href="#">6845</a>		
<a href="#">ENSG00000168939</a>	<a href="#">ENST00000369437</a>	<a href="#">SPRY3</a>	<a href="#">10251</a>		
<a href="#">ENSG00000185973</a>	<a href="#">ENST00000334398</a>	<a href="#">TMLHE</a>	<a href="#">55217</a>		
<a href="#">ENSG00000185978</a>	<a href="#">ENST00000369444</a>	<a href="#">H2AFB3</a>	<a href="#">83740</a>		
<a href="#">ENSG00000185978</a>	<a href="#">ENST00000369444</a>	<a href="#">H2AFB3</a>	<a href="#">474381</a>		
<a href="#">ENSG00000185990</a>	<a href="#">ENST00000369445</a>	<a href="#">F8A3</a>	<a href="#">8263</a>		
<a href="#">ENSG00000185990</a>	<a href="#">ENST00000369445</a>	<a href="#">F8A3</a>	<a href="#">474383</a>		
<a href="#">ENSG00000185990</a>	<a href="#">ENST00000369445</a>	<a href="#">F8A3</a>	<a href="#">474384</a>		
<a href="#">ENSG00000198444</a>	<a href="#">ENST00000369505</a>	<a href="#">F8A2</a>	<a href="#">8263</a>		
<a href="#">ENSG00000198444</a>	<a href="#">ENST00000369505</a>	<a href="#">F8A2</a>	<a href="#">474383</a>		
<a href="#">ENSG00000198444</a>	<a href="#">ENST00000369505</a>	<a href="#">F8A2</a>	<a href="#">474384</a>		
<a href="#">ENSG00000198307</a>	<a href="#">ENST00000354514</a>	<a href="#">H2AFB2</a>	<a href="#">83740</a>		
<a href="#">ENSG00000198307</a>	<a href="#">ENST00000354514</a>	<a href="#">H2AFB2</a>	<a href="#">474381</a>		
<a href="#">ENSG00000155962</a>	<a href="#">ENST00000369449</a>	<a href="#">CLIC2</a>	<a href="#">1193</a>		
<a href="#">ENSG00000155961</a>	<a href="#">ENST00000369454</a>	<a href="#">RAB39B</a>	<a href="#">116442</a>		
<a href="#">ENSG00000155959</a>	<a href="#">ENST00000286428</a>	<a href="#">VBP1</a>	<a href="#">7411</a>		
<a href="#">ENSG00000214827</a>	<a href="#">ENST00000330425</a>	<a href="#">MTCP1</a>	<a href="#">4515</a>		
<a href="#">ENSG00000182712</a>	<a href="#">ENST00000369480</a>	<a href="#">MTCP1NB</a>	<a href="#">100272147</a>		
<a href="#">ENSG00000165775</a>	<a href="#">ENST00000369498</a>	<a href="#">FUNDC2</a>	<a href="#">65991</a>		
<a href="#">ENSG00000197932</a>	<a href="#">ENST00000369446</a>	<a href="#">F8A1</a>	<a href="#">8263</a>		

**Result Table 1**

Please select columns to be included in the output and hit 'Results' when ready

Features  Transcript Event  
 Structures  Homologs  
 Variations  Sequences

SEQUENCES:  
 Header Information

**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

Please select columns to be included in the output and hit 'Results' when ready

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]

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  
 3' UTR  
 Exon sequences  
 cDNA sequences  
 Coding sequence  
 Protein

Upstream flank  
 Upstream flank

Downstream 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  
 Gene Start (bp)  
 Gene End (bp)  
 Ensembl Protein ID (s)

**Transcript Information**

CDS Length  
 CDS Start  
 CDS End  
 5' UTR Start  
 5' UTR End  
 3' UTR Start

3' UTR  
 Ensembl Transcript ID  
 Ensembl Transcript Name  
 Strand  
 Transcript Orientation  
 Transcript Type

**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

Please select columns to be included in the output and hit 'Results' when ready

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]

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  
 3' UTR  
 Exon sequences  
 cDNA sequences  
 Coding sequence  
 Protein

**STEP 22:**  
 Click 'Results'

New Count Results
URL XML Peri Help

Dataset 105 / 49506 Genes  
 Homo sapiens genes (GRCh37)

Export all results to  FASTA  Unique results only

Email notification to

View  rows as FASTA  Unique results only

```

>ENSG00000013619|ENST00000262858|MAML1|X
AAGCCCTGTGCTAGGTCGTTTGGGAAACGCCTTGGAGAGTCAAGAATAAATTTGCAGGT
CAAACAATGGATGACTGGAAAAGTCGGCTTGTAAATCAAGAGCATGCTTCCCATTTCCGCC
ATGGTGGGAAATCGTCAGGAGCCCGAAGAGCTCCAGGAATCGGGAAAAGAGCCCTCGTGG
ATGGAGGAAGAAGATTTATCTTTTCTCTACAAGAGCAGCCAGGAAGAAAAGCATCAGGGA
ACTGTTAAGAGGAGACAAGAAGAAGACCCTCCAGTTTCCAGACATGGCTGATGGGGGC
TACCCTAATAAAAATTAAGAGGCCCTTGCCTTGAAGATGTCACCCTTGAATGGGCCAGGT
GCTCATCTAGTACTGCTTGTGCAGAACTGCAGGTCCTCCATTGACAATAAATCTTAGC
CCTGGCGCTATGGGAGTGGCTGGCCAGTCACTTACTGCTGGAGAATAACCCATGAAATGGC
AACATCATGGGCTCACCATTGTAGTACCAGACTACAGAAGTGGGACTGAAAAGGGCCC
ACTGTTCCCTTACTATGAGAAAATCAACAGCGTGCCTGGCTGTAGACCAGGAGCTTCAAGAG
CTGCTAGAGGAGCTCACAAAATTCAGAGCCCTTCTCCAAATGAGCTAGATCTTGAGAAG
ATACTGGGACGAAGCCAGAAAGAGCCACTGGTTTTAGATCATCCCAAGCAACCCCTAAGC
ACAACCTCCCAAGCCCTTCGGTTACAGATGTCACACTTGGAGAGCCTGGCTTCCAGCAAGGAG
TTTGTCTTACTGTGAGCCAAAGTTACTGGCATGTCACCTTCAATCCCTCCCTCCACA
GGGATCAGCTATTCCGATTCCTTCCACCACTAAGCAGATAGTGTCAACCGAGTCTTCAATG
GCACAGTCCAAAGAGCCAGGTCAGGCCATGCTCCCTGTGCTCTGCCCCCTTACCAGTG
CCTCAGTGGCATCACGCCACCCAGCTGAAGGCGTTGGCAGCCAGCAAGCAGGGGTCTGCT
ACAAAGCAGCAAGGGCCACCCCAAGTGGTCTGGCTTGCCTCCTCCAGGACTCTCTCCA
CCTTACCCGCCAGTGGCCATCACACACCCACCAACCGCTGCCACTGCCACCAACCAACCC
CCATTCAGCCCCCAGAGCCTCATGGTGTCTGCTGATGTCGTCATACCTTGTGGGTAGC
ACTCTCCGAGGCTCTCCCAATGCTTACTGTCAAGCATGACGCTCCAGCAGCAATGCTGCC
CTGGGCCCCGCTATGCTTCTGCTCTGAGAAAGCTCCCAAGCCCTGCTCTCACTCAACAG
CCGAGTTCGGCCCTCAGAGCTCCATTCTGGCAACCTCATGCTCTACCATCAAAACC
CCTCAAGGACACCTGATGCTGCTTGTGCTGCCAGCAACCCCTGGGCCGTCCCAACCTAT
CGCCAGAGAGCTCTAGCCAGGCTTGGCCAGCAGTCTTCAACCCACAGTGTTC
CTGATCCGAGGCTCACTCCACCACTAATCTTCAAGCCAGCAACAGCAGCAGCAGCAGCAG
.....
ATAGAAAACCCACACCCACTGTCCTGTAACATTTTCTCAGTGTCCAGACTTTCTGTAATC
ACATTTTAAATGGCCACCTCGTATTTTCACTCTACATTTGAAATCTGGCGTCTGTTTCAAG
CCAGTGTGTTTTTTCTTCTGTTCTGTAATAAACAGCCAGGAGAAAAGTG
>ENSG000000166008|ENST00000298974|MAGEA9|X
GTGCGCACTGGGGTTCAGAGAGAAGGGAGAGGCTCCTTCTGAGGGCGGCTTGATACCG
GTGGAGGAGCTCCAGGAAGCAGGCGGCTTGGTCTGAGACAGTGTCTCAGGTCGAGAG
GCAGAGGAGACCAGGCAAGTGTGACAGTGAAGGTTCTCGGGCAGGCTAACAGGAGGA
CAGGAGCCCAAGAGGCCAGAGCAGCACTGACGAAAGACTGCTGTGGGTCTCCATCG
CCAGCTCTGCCCCACGCTCTGACTGCTGCCCTGACAGAGTCACTATGCTCTCTCGAGC
AGAGGAGTCCGCACTGCAAGCCTGATGAAGACCTTGAAGCCCAAGGAGAGGACTTGGGCC
TGATGGGTGCACAGGAACCCACAGGCGAGGAGGAGGACTACCTCCTCCTGACAGCA
AGGAGGAGGAGTGTCTGCTGCTGGTCACTCAAGTCTTCCCAAGTCTCAGGGAGGCG
CTTCTCTCTCAATTTCCGCTACTACACTTTATGGAGCAATTCGATGAGGGCTCCAGCA
GTCAAGCAAGAGGAAGAGCCAAAGCTCTCGGTGACCCAGCTCAGCTGGAGTTTCAATGTTCC
AAGAAACACTGAAAATGAAAGTGGCTGAGTTGGTTTCTGCTCCAAAATATCGAG
TCAAGGAGCGGTCACAAAGGCAGAAATGCTGGAGAGGCTCATCAAAAATTAACAAGCGCT
  
```

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|MAML1|X
AAGCCCTGTGCTAGGTCGTTTGGGAAACGCCTTGGAGAGTCAAGAATAAATTTGCAGGT
CAAACAATGGATGACTGGAAAAGTCGGCTTGTAAATCAAGAGCATGCTTCCCATTTCCGCC
ATGGTGGGAAATCGTCAGGAGCCCGAAGAGCTCCAGGAATCGGGAAAAGAGCCCTCGTGG
ATGGAGGAAGAAGATTTATCTTTTCTCTACAAGAGCAGCCAGGAAGAAAAGCATCAGGGA
ACTGTTAAGAGGAGACAAGAAGAAGACCCTCCAGTTTCCAGACATGGCTGATGGGGGC
TACCCTAATAAAAATTAAGAGGCCCTTGCCTTGAAGATGTCACCCTTGAATGGGCCAGGT
GCTCATCTAGTACTGCTTGTGCAGAACTGCAGGTCCTCCATTGACAATAAATCTTAGC
CCTGGCGCTATGGGAGTGGCTGGCCAGTCACTTACTGCTGGAGAATAACCCATGAAATGGC
AACATCATGGGCTCACCATTGTAGTACCAGACTACAGAAGTGGGACTGAAAAGGGCCC
ACTGTTCCCTTACTATGAGAAAATCAACAGCGTGCCTGGCTGTAGACCAGGAGCTTCAAGAG
CTGCTAGAGGAGCTCACAAAATTCAGAGCCCTTCTCCAAATGAGCTAGATCTTGAGAAG
ATACTGGGACGAAGCCAGAAAGAGCCACTGGTTTTAGATCATCCCAAGCAACCCCTAAGC
ACAACCTCCCAAGCCCTTCGGTTACAGATGTCACACTTGGAGAGCCTGGCTTCCAGCAAGGAG
TTTGTCTTACTGTGAGCCAAAGTTACTGGCATGTCACCTTCAATCCCTCCCTCCACA
GGGATCAGCTATTCCGATTCCTTCCACCACTAAGCAGATAGTGTCAACCGAGTCTTCAATG
GCACAGTCCAAAGAGCCAGGTCAGGCCATGCTCCCTGTGCTCTGCCCCCTTACCAGTG
CCTCAGTGGCATCACGCCACCCAGCTGAAGGCGTTGGCAGCCAGCAAGCAGGGGTCTGCT
ACAAAGCAGCAAGGGCCACCCCAAGTGGTCTGGCTTGCCTCCTCCAGGACTCTCTCCA
CCTTACCCGCCAGTGGCCATCACACACCCACCAACCGCTGCCACTGCCACCAACCAACCC
CCATTCAGCCCCCAGAGCCTCATGGTGTCTGCTGATGTCGTCATACCTTGTGGGTAGC
ACTCTCCGAGGCTCTCCCAATGCTTACTGTCAAGCATGACGCTCCAGCAGCAATGCTGCC
CTGGGCCCCGCTATGCTTCTGCTCTGAGAAAGCTCCCAAGCCCTGCTCTCACTCAACAG
CCGAGTTCGGCCCTCAGAGCTCCATTCTGGCAACCTCATGCTCTACCATCAAAACC
CCTCAAGGACACCTGATGCTGCTTGTGCTGCCAGCAACCCCTGGGCCGTCCCAACCTAT
CGCCAGAGAGCTCTAGCCAGGCTTGGCCAGCAGTCTTCAACCCACAGTGTTC
CTGATCCGAGGCTCACTCCACCACTAATCTTCAAGCCAGCAACAGCAGCAGCAGCAGCAG
.....
ATAGAAAACCCACACCCACTGTCCTGTAACATTTTCTCAGTGTCCAGACTTTCTGTAATC
ACATTTTAAATGGCCACCTCGTATTTTCACTCTACATTTGAAATCTGGCGTCTGTTTCAAG
CCAGTGTGTTTTTTCTTCTGTTCTGTAATAAACAGCCAGGAGAAAAGTG
>ENSG000000166008|ENST00000298974|MAGEA9|X
GTGCGCACTGGGGTTCAGAGAGAAGGGAGAGGCTCCTTCTGAGGGCGGCTTGATACCG
GTGGAGGAGCTCCAGGAAGCAGGCGGCTTGGTCTGAGACAGTGTCTCAGGTCGAGAG
GCAGAGGAGACCAGGCAAGTGTGACAGTGAAGGTTCTCGGGCAGGCTAACAGGAGGA
CAGGAGCCCAAGAGGCCAGAGCAGCACTGACGAAAGACTGCTGTGGGTCTCCATCG
CCAGCTCTGCCCCACGCTCTGACTGCTGCCCTGACAGAGTCACTATGCTCTCTCGAGC
AGAGGAGTCCGCACTGCAAGCCTGATGAAGACCTTGAAGCCCAAGGAGAGGACTTGGGCC
TGATGGGTGCACAGGAACCCACAGGCGAGGAGGAGGACTACCTCCTCCTGACAGCA
AGGAGGAGGAGTGTCTGCTGCTGGTCACTCAAGTCTTCCCAAGTCTCAGGGAGGCG
CTTCTCTCTCAATTTCCGCTACTACACTTTATGGAGCAATTCGATGAGGGCTCCAGCA
GTCAAGCAAGAGGAAGAGCCAAAGCTCTCGGTGACCCAGCTCAGCTGGAGTTTCAATGTTCC
AAGAAACACTGAAAATGAAAGTGGCTGAGTTGGTTTCTGCTCCAAAATATCGAG
TCAAGGAGCGGTCACAAAGGCAGAAATGCTGGAGAGGCTCATCAAAAATTAACAAGCGCT
  
```

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](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](http://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**' from 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 truncatus 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](http://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](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](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](http://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.

- Find the Ensembl page with information for this SNP.
- What is the minor allele of this SNP in Caucasians?
- 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).

- Do both individuals (Venter and Watson) have sequence coverage at the position of the R620W (C1858T) SNP?
- 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)'.

- 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.