

Access to genes and genomes with Ensembl



Course Manual
Course on Comparative Genomics,
Europe, 2011

Esembl version e60 and EnsemblGenomes version e7
<http://www.ensembl.org/>
<http://www.ensemblgenomes.org/>

1	Introduction.....	3
2	Walking through the website.....	7
2.1	Tying it together – The human LDLR gene	22
2.2	EXERCISES and ANSWERS	23
2.2.1	BROWSING ENSEMBL (exercises).....	23
2.2.2	BROWSING ENSEMBL (answers).....	24
2.2.3	USER UPLOAD (exercises and answers)	27
2.2.4	Variations and functional genomics (optional exercises)	30
2.2.5	Variations and functional genomics (answers)	31
2.2.6	Ensembl Plants (exercises optional).....	33
2.2.7	Ensembl Plants (answers).....	36
2.2.8	Ensembl Bacteria (exercises optional)	38
	Ensembl Bacteria (answers).....	40
2.2.9	Sea squirt and Diatoms (exercises optional).....	42
2.2.10	Sea squirt and Diatoms (answers)	45
3	BioMart.....	47
3.1	Mining data- worked example.....	47
3.1.1	BIOMART (exercises)	55
3.1.2	BIOMART (answers)	57
4	Exercises for day II	61
4.1.1	COMPARATIVE GENOMICS (exercises)	61
4.1.2	COMPARATIVE GENOMICS (answers).....	63
4.1.3	GENEBUILD (exercises).....	65
4.1.4	GENEBUILD (answers)	66
5	Quick Guide to Databases and Projects	68

1 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.

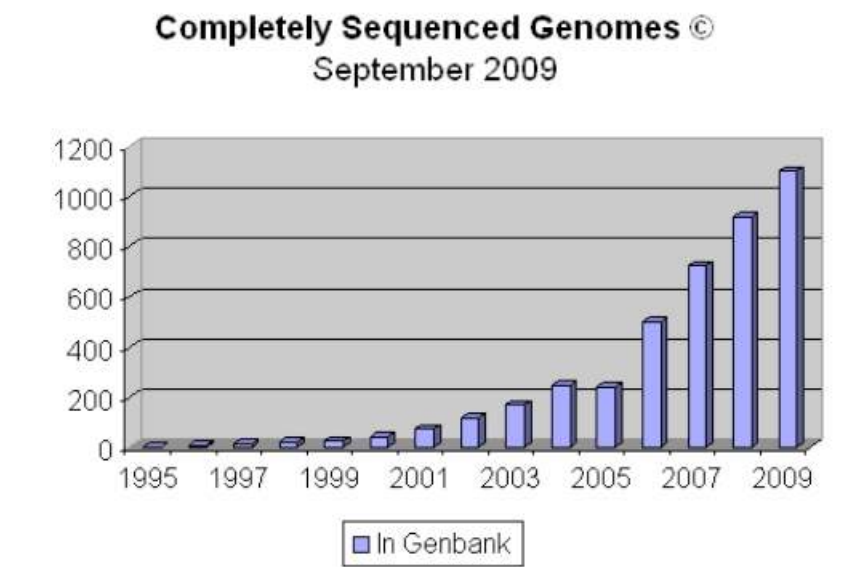


Figure 1. Completely sequenced genomes as of January, 2009 (figure taken from <http://www.genomesonline.org>).

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

Flicek, P. *et al.*

Ensembl 2011

Nucleic Acids Res. Advanced Access (*Database Issue*)

<http://nar.oxfordjournals.org/content/early/2010/11/02/nar.gkq1064.full>

Ensembl Methods Series

<http://www.biomedcentral.com/series/ENSEMBL2010>

Xosé M. Fernández-Suárez and Michael K. Schuster

Using the Ensembl Genome Server to Browse Genomic Sequence Data.

UNIT 1.15 in *Current Protocols in Bioinformatics*, Jun 2010.

Giulietta M Spudich and Xosé M Fernández-Suárez

Touring Ensembl: A practical guide to genome browsing

BMC Genomics 2010, 11:295 (11 May 2010)

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

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

Birney, E. *et al.*¹

An Overview of Ensembl.

Genome Research 14(5): 925-928 (2004)

Ashurst, J. L. *et al.*

The Vertebrate Genome Annotation (Vega) database.

Nucl. Acids Res. 33:D459-D465 (2005)

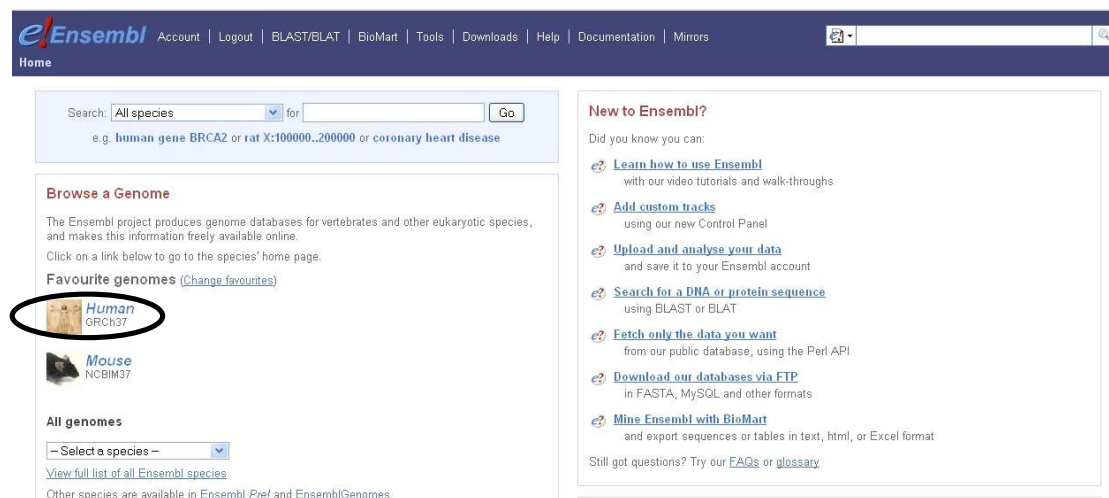
¹ This paper was part of the May 2004 issue of *Genome Research* which included an Ensembl special covering detailed aspects of the Ensembl web site, the underlying scalable database system for storing genome sequence and annotation information, as well as the automated genome analysis and annotation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

2 Walking through the website

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

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

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



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

About this species

- Description
 - Genome Statistics
 - Assembly and Genebu
 - Top 40 InterPro hits
 - Top 500 InterPro hits
 - What's New
 - Sample entry points
 - Karyotype
 - Location (AL032821.2)
 - Gene (BRCA2)
 - Transcript (FOXP2-203)
- Configure this page
- Manage your data
- Export data
- Bookmark this page

Search Ensembl Human

Search for: Go
e.g. gene BRCA2 or AL032821.2.1.143563 or muscular dystrophy

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.
- config 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 CCDS 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

[New Search](#)

- Configure this page
- Manage your data
- Export data
- Bookmark this page

Results Summary [Result in Detail >](#)

By Species

Species	Count
Total	282
Homo sapiens	282

[Gene \(282\)](#)

By Feature type

Feature type	Count
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 302 entries in the search database. Viewing hits 1-10

1 2 3 4 ... 28 29 30 31

[Ensembl protein_coding Gene: ENSG00000163914 \(HGNC Symbol: **RHO**\)](#) [\[Region in detail\]](#)

Description: **rho**dopsin [Source:HGNC Symbol;Acc:10012]

Source: e59; Feature type: Gene; Homo sapiens;

[Havana protein_coding Gene: OTTHUMG00000159542 \(**RH**\)](#)

Description: **rho**dopsin

Source: e59; Feature type: Gene; Homo sapiens; Species: Homo sapiens; Gene;

[Ensembl protein_coding Gene: ENSG0000004777 \(HGNC Symbol: ARHGAP33\)](#) [\[Region in detail\]](#)

Description: **Rho** GTPase activating protein 33 [Source:HGNC Symbol;Acc:23085]

Source: e59; Feature type: Gene; Homo sapiens; Species: Homo sapiens; Gene;

[Ensembl protein_coding Gene: ENSG0000006607 \(HGNC Symbol: FARP2\)](#) [\[Region in detail\]](#)

Description: FERM, **Rho**GEF and pleckstrin domain protein 2 [Source:HGNC Symbol;Acc:16460]

Source: e59; Feature type: Gene; Homo sapiens; Species: Homo sapiens; Gene;

**Click
ENSG00000163914**

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

Human (GRCh37) Location: 3:129,247,483-129,254,012 Gene: RHO

Gene-based displays

- Gene summary
- Splice variants (2)
- Supporting evidence
- Sequence
- External references (2)
- Regulation
- Comparative Genomics
 - Genomic alignments (51)
 - Gene Tree (image)
 - Gene Tree (text)
 - Gene Tree (alignment)
 - Orthologues (106)
 - Paralogues (12)
 - 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

Gene: RHO (ENSG00000163914)

rhodopsin [Source:HGNC Symbol;Acc:10012]

Location [Chromosome 3: 129,247,483-129,254,012](#) forward strand.

Transcripts There are 2 transcripts in this gene

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RHO-001	ENST00000296271	2592	ENSP00000296271	348	Protein coding	CCDS3063
RHO-002	ENST00000511172	454	ENSP00000427467	152	Protein coding	-

Gene summary [help](#) [Splice variants »](#)

Name [RHO](#) (HGNC Symbol)

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 Gene containing both Ensembl genebuild transcripts and [Havana](#) manual curation, see [article](#)

Alternative genes This gene corresponds to the following database identifiers:
Havana gene: [OTTHUMG00000159542](#) [view all locations](#)

Transcripts for the nearby IFT122 gene

Blue bar is the genome

RHO transcripts click for info

Configuring the display
Tip: use the "Configure this page" link on the left to show additional data in this region.

Ensembl release 59 - Jul 2010 © WTSI / EBI [About Ensembl](#) | [Contact Us](#) | [Help](#)
[Permanent link](#) - [View in archive site](#)

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 (2)
- Supporting evidence
- Sequence**
- External references (2)
- Regulation
- Comparative Genomics
 - Genomic alignments (51)
 - Gene Tree (image)
 - Gene Tree (text)

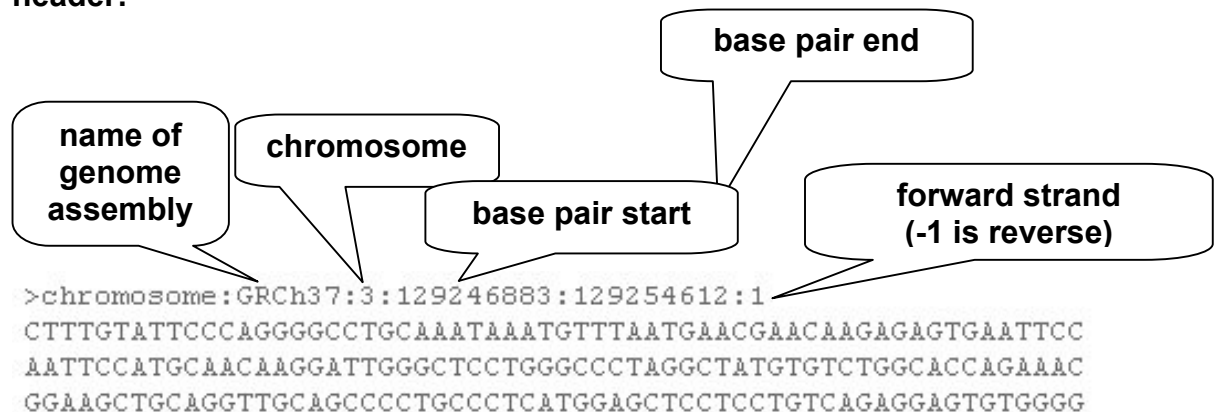
Click Sequence

```
>chromosome:GRCh37:3:129246883:129254612:1
CTTTGTATTCCCAGGGGCCTGCAAAATAAATGTTTAATGAACGAACAAGAGAGTGAATTCC
AATTCCATGCAACAAGGATTGGGCTCCTGGGCCCTAGGCTATGTGTCTGGCACCAGAAAC
GGAAGCTGCAGGTTGCAGCCCTGCCCTCATGGAGCTCCTCCTGTCAGAGGAGTGTGGGG
ACTGGATGACTCCAGAGGTAACCTTGTGGGGGAACGAACAGGTAAGGGGCTGTGTGACGAG
ATGAGAGACTGGGAGAATAAACAGAAAGTCTCTAGCTGTCCAGAGGACATAGCACAGAG
GCCCCATGGTCCCTATTTCAAAACCCAGGCCACCAGACTGAGCTGGGACCTTGGGAGACCA
AGTCATGCAGAAAGTTAGGGGACCTTCTCCTCCCTTTTCTGGATCCTGAGTACCTCTCCT
CCTGACCTCAGGCTTCTCCTAGTGTACCTTGGCCCTCTTAGAAGCCAATTAGGCC
TCAGTTTCTGCAGCGGGGATTAATATGATTATGAACACCCCAATCTCCAGATGCTGAT
TCAGCCAGGAGCTTAGGAGGGGGAGGTCACCTTATAAGGGTCTGGGGGGTTCAGAACCA
GAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCTTCGACGATTCTTGGGTGG
GAGCAGCCACGGGTGACGCACAAGGGCCACAGCCATGAATGGCACAGAAGGCCCTAATT
CTACGTGCCCTTCTCCAATGCGACGGGTGTGGTACGACGCCCTTCGAGTACCACAGTA
CTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCTACATGTTTCTGCTGATCGT
GCTGGGCTTCCCATCAACTTCTCAGCTCTACGTCACCGTCCAGCACAAGAAATGCG
CACGCTCTCAACTACATCCTGCTCAACCTAGCCGTGGCTGACCTCTTCATGGTCCTAGG
TGGCTTACCAGCACCTCTACACCTCTGTCATGGATACTTCGTCTTCGGGCCACAGG
ATGCAATTGGAGGGCTTCTTGGCACCTGGGCGGTATGAGCCGGGTGTGGGTGGGGTG
TGCAGGAGCCCGGAGCATGGAGGGGTCTGGGAGAGTCCCGGGCTTGGCGGTGTGGCTG
AGAGGCCTTCTCCTTCTCCTGTCTCAATGTTATCCAAAGCCCTCATATATTCAGTC
```

Upstream
sequence

Exon
sequence

The sequence is shown in FASTA format. Take a look at the FASTA header:



Exons are highlighted within the genomic sequence. Variations can be added with the [Configure this page](#) link found at the left. Click on it now.

Configure page | Custom Data | Your account

Configure view
- Display options

Display options

5' Flanking sequence (upstream): 600 * (Maximum of 1000000)

3' Flanking sequence (downstream): 600 * (Maximum of 1000000)

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

Filter variations by population: All Warning: This could cause the page to take a long time to load

Minor allele frequency for population filter: 0.1 (Maximum of 0.1)

Line numbering: Relative to this sequence

Fields marked with * are required

Display variations

Turn on line numbers

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

Exons: ENSG00000163914 exons | Ensembl exons in this region

Variations: 3' UTR | 5' UTR | Downstream | Essential splice site | HGMD mutation | Intergenic | Intronic | Non-synonymous coding | Splice site SNP | Upstream

```
>chromosome: GRCh37:3:129246883:129254612:1
1 CTTTGATTCCAGGGGCTGCAATAAATGTTTAAATGAACGAACAAGAGAGTGAAATTCC 60
61 AATTCATGCAACAAAGGATTGGGCTCCTGGGCCCTAGGCTATGTGTCTGGCACCAGAAAC 120
121 GGAAGCTGCAGGTTGCAGCCCTGCCCTCATGGAGCTCCTCCTGTCAGAGGAGTGTGGGG 180
181 ACTGGATGACTCCAGAGGTAACCTGTGGGGGAACAAACAGGTAAGGGGCTGTGTGACGAG 240
241 ATGAGAGACTGGGAGAATAAACAGAAAGTCTCTAGCTGTCCAGAGGACATAGCACAGAG 300
301 GCCCATGGTCCCTATTTCAAACCCAGGCCACCAGACTGAGCTGGGACCTTGGGACAGACA 360
361 AGTCATGCAGAAAGTTAGGGGACCTTCTCCTCCCTTTCTCGATCCTGAGTACCTCTCCT 420
421 CCCTGACCTCAGGCTTCTCCTAGTGTACCTTGGGCCCTCTTAGAAGCCAATTAGGCC 480
481 TCAGTTTCTGCAGCGGGGATTAAATGATTATGAACACCCCAATCTCCAGATGCTGAT 540
541 TCAGCCAGGAGCTTAGGAGGGGAGGTCACCTTTATAAGGGTCTGGGGGGTCAAGACCCA 600
601 GAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTTCAGCATTCTTGGGTGG 660
661 GAGCAGCCCGGGTCAAGCCCAAGGGCCACAGCCATGAATGGCAAGAGGCCCTTAACCT 720
721 CTACGTGCCCTTCTCCAATGCGAGGGGTGTGGTACGCAGCCCTTCGAGTACCCACAGTA 780
781 CTACGTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCTACAGTTCTGCTGATCGT 840
841 GCTGGCTTCCSCATCAACTTCTCAGGCTCTACGTACCCGTCAGCACAGAAGCTGCG 900
901 CACGCCCTCTCAACTACATCTGCTCAACCTAGCCGTGGCTGACCTCTTCATGGTCTAGG 960
961 TGGCTTCCAGCAGCCCTCAACCTCTCTGATGGATACCTGCTCTTGGGGCCACAGG 1020
1021 ATGCAATTGGAGGGCTTCTTGGCCACCTGGGGTGTATGAGCCGGGTGTGGGTGGGGTG 1080
1081 TGCAGGAGCCCGGAGCATGGAGGGGTCTGGGAGAGTCCCGGGCTTGGCGGTGGTGGCTG 1140
1141 AGAGCCCTTCTCCCTTCTCTGTCCTGTAATGTTATCCAAAGCCCTCATATATTCAGTC 1200
```

215: rs72987932 G/A;

644: rs2269736 G/A;

669: rs7984 A/G; 705: CM930648 HGMD MUTATION;

738: CM930648 HGMD MUTATION;

813: CM941249 HGMD MUTATION;

845: CM930652 HGMD MUTATION;

954: CM910329 HGMD MUTATION;

963: CM930654 HGMD MUTATION;

1022: CM045432 HGMD MUTATION;

Link to variation information

Variations in the sequence are highlighted, and represented by the IUPAC code. The code “R” 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 6 primates, EPO. The EPO pipeline refers to the programs behind the whole genome alignments - click the [help](#) button for more.

Click  at the left. Turn on Conservation regions to “All conserved regions” in the menu. Also, click “6 primates EPO” at the left of the configuration panel. Deselect “Ancestral sequence”. Close the menu.

```

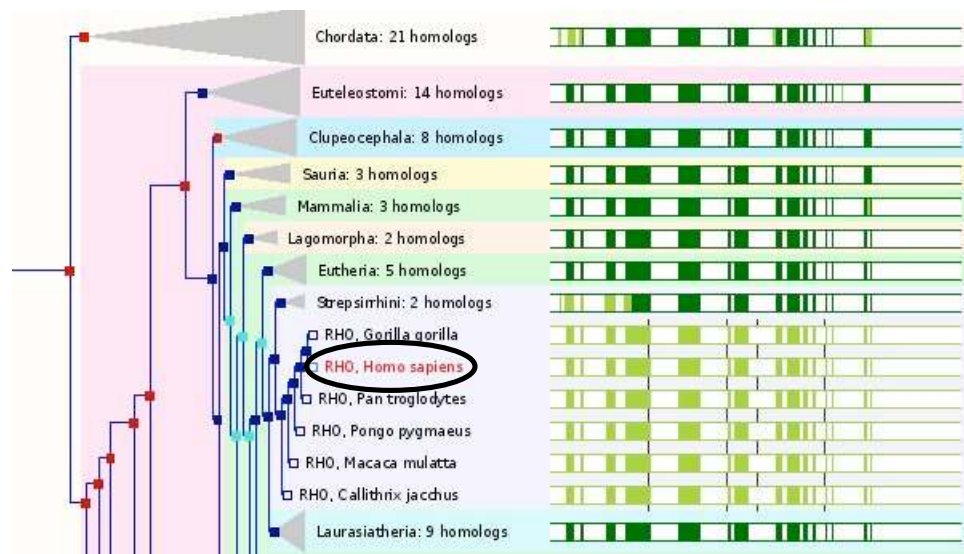
Homo sapiens      CCTGGATCCTGAGTACCTCTCCTCCCTGACCTCAGGCTTCCTCCTAGTGTACCTTGGCCCTCTTGAAGCCAATTAGGCC
Pan troglodytes   CCTGGATCCTGAGTACCTCTCCTCCCTGACCTCAGGCTTCCTCCTAGTGTACCTTGGCCCTCTTGAAGCCAATTAGGCC
Gorilla gorilla   CCTGGATCCTGAGTACCTCTCCTCCCTGACCTCAGGCTTCCTCCTAGTGTACCTTGGCCCTCTTGAAGCCAATTAGGCC
Pongo pygmaeus    CCTGGGTCCTGAGTACCTCTCCTCCCGACCTCAGGCTTCCTCCTGGTGTACCTTGGCCCTCTTGAAGCCAATTAGGCC
Macaca mulatta    CCAGGGTCCTGAGTACCTCTCCTCCCTGACCTCAGGCTTCCTCCTGGTGTACCTTGGCCCTCTTGAAGCCAATTAGGCC
Callithrix jacchus CCAAGGTCCTGAGTACCTCCCTCCCTGACCTCAGGCTTCCTCCTGGTGTACCTTGGGACCTTGGGAAAGCCAATTAGGCC

Homo sapiens      CCCCCAATCTCCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGGAGGTCACCTTTATAAGGGTCTG-----GGGGGG
Pan troglodytes   CCCCCAATCTCCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGGAGGTCACCTTTATAAGGGTCTG-----GGGGGG
Gorilla gorilla   CCCCCAATCTCCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGGAGGTCACCTTTATAAGGGTCTG-----GGGGGG
Pongo pygmaeus    CCCCCAATCTCCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGGAGGTCACCTTTATAAGGGTCTG-----GGGGGG
Macaca mulatta    CCCCCAATCTCCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGGAGGTCACCTTTATAAGGGTCTG-----GGGGGG
Callithrix jacchus CC-CCAACTCTCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGGAGGTCACCTTTATAAGGGTCTGTGGGGGTGGGGGGGG

Homo sapiens      CTGAGCTCAGGCCTTCGCAGCATTCTTGGGTGGGAGCAGCCACGGGTACGCCACAAGGGCCACAGCCATGAATGGCACAGAAG
Pan troglodytes   CTGAGCTCAGGCCTTCGCAGCATTCTTGGGTGGGAGCAGCCGTGGGTACGCCCAAGGGCCACAGCCATGAATGGCACAGAAG
Gorilla gorilla   CTGAGCTCAGGCCTTCGCAGCATTCTTGGGTGGGAGCAGCCGCGGGTCAGCCACAAGGGCCACAGCCATGAATGGCACAGAAG
Pongo pygmaeus    CTGAGCTCAGGCCTTCGCAGCATTCTTGGGTGGGAGCAGCCGCGGGTCAGCCACAAGGGCCACAGCCATGAATGGCACAGAAG
Macaca mulatta    CTGAGCTCAGGCCTTCGCAGCATTCTTGGGTGGGAGCAGCCGCGGGTCAGCCACAAGGGCCACAGCCATGAATGGCACAGAAG
Callithrix jacchus CTGAGCTCAGGCCTTTCGCAGCATTCTTGGGTGGGAGCAGCCGTGGGGCAACCACAAGGGCCACAGTCATGAATGGCACTGAAG
  
```

Exons are highlighted in red, conserved 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 used to determine orthologues and paralogues.



Click on any node (square) to reveal the taxonomic level, or to collapse or expand a subtree.

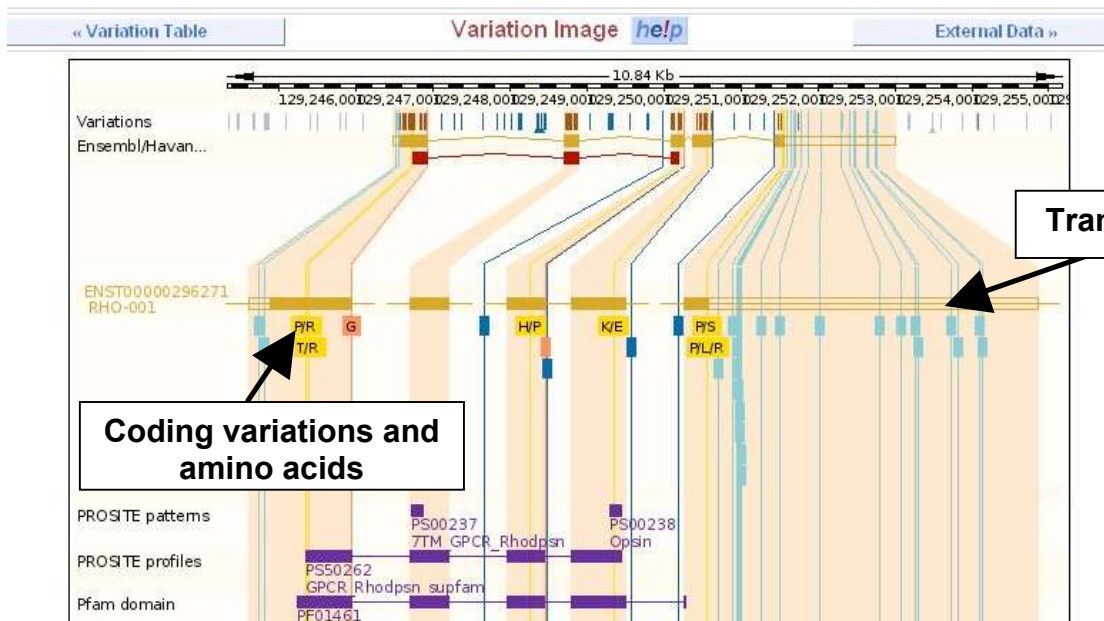
Click the [Orthologues](#) link at the left of this page to view homologues detected in this tree.

« Gene Tree (image)			Orthologues help				Paralogues »	
The following gene(s) have been identified as putative orthologues:								
Species	Type	dN/dS	Ensembl identifier	Location	Target %id	Query %id	External ref.	
Alpaca (<i>Vicugna pacos</i>)	1-to-1	n/a	ENSVPAG00000003334 Multi-species view Alignment Gene Tree (image)	GeneScaffold_3012:4867-8438:1	77	77	RHO rhodopsin [Source:HGNC Symbol;Acc:10012]	
Anole Lizard (<i>Anolis carolinensis</i>)	1-to-1	n/a	ENSACAG000000014258 Multi-species view Alignment Gene Tree (image)	scaffold_163:632250-640401:-1	81	82	OPSD_ANOCA Rhodopsin [Source: UniProtKB/Swiss-Prot, acc: P41591]	
Anole Lizard (<i>Anolis carolinensis</i>)	Possible ortholog	n/a	ENSACAG000000005789 Multi-species view Alignment Gene Tree (image)	scaffold_24:1173306-1186146:1	39	33	Novel Ensembl prediction No description	
Anole Lizard (<i>Anolis carolinensis</i>)	Possible ortholog	n/a	ENSACAG000000006891 Multi-species view Alignment Gene Tree (image)	scaffold_44:1171109-1177064:-1	38	28	Novel Ensembl prediction No description	
Anole Lizard (<i>Anolis carolinensis</i>)	Possible ortholog	n/a	ENSACAG000000008548 Multi-species view Alignment Gene Tree (image)	scaffold_1370:18504-28260:1	44	28	Q9W6K3 _ANOCA P opsin [Source: UniProtKB/TrEMBL, acc: Q9W6K3]	

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

First click on **Variation table** at the left. Show all the non-synonymous coding variations for this gene, by clicking "show".

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:

Ensembl Account | Logout | BLAST/BLAT | BioMart | Tools | Downloads | Help | Documentation | Mirrors

Human (GRCh37) Location: 3:129,247,483-129,254,012 Gene: RHO Variation: rs28933395

Variation displays

- Summary
- Gene/Transcript (2)
- Population genetics
- Individual genotypes (1)
- Context
- Linked variations
- Phenotype Data (2)
- Physiologic Context (4)
- External Data
- Configure this page
- Manage your data
- Export data

Variation: rs28933395

Variation class: SNP (rs28933395 source dbSNP 131 - Variants (including SNPs and indels) imported from dbSNP)

Synonyms: HGMD-PUBLIC CMS20608 Uniprot VAR_004778

Present in: Clinical/LSDb variations from dbSNP, ENSEMBL-Venter

Alleles: C/G (Ambiguity code: S) Ancestral allele: C

Location: This feature maps to 3:129247734 (forward strand) | View in location tab

Variation summary help

Validation status: Proven by previous (Feature tested and validated by a non-computational method).

HGVs names: ENST00000296271.3 c.158C>G ENSP00000296271.3 p.Pro53Arg

Linkage disequilibrium data: No linkage data for this variant

Flanking Sequence (dbSNP):

```

TTGGCCCTCTTAGAGCCAAATTAGGCCCTCAGTTTCTGCAGCGGGGATTATATGATTA
TGAACACCCCAATCTCCAGATGCTGATTACGCCAGGAGCTTAGGAGGGGAGGTCACT
TTATAAGGGTCTGGGGGGTTCAGAACCCAGAGTCTCCAGCTGGAGCCCTGAGTGGCTGA
GCTCAGGCCCTTCGACGACTTCTTGGGTGGGAGCAGCCACGGGTACGCCACAAGGGCCACA
GCCATGAATGGCACAGAGGCCCTAATTCTACGTGCCCTTCTCCAATGCGACGGGTGTG
TACGCGACCTCTTCCAAATGACCAAGTACCTGCTGAGCCATGGGAGTTCTCCATG
CTGGCGCTTACATGTTCTTCTGCTGATGCTGCTGGGCTTCCGCTGCTGCTCAGGCTC
TACGTACCGCTCCAGCACAAGAGCTGCGCAGCGCTCTCAAGAGCTGCTGCTCAACTA
GCCGTGGCTGACCTCTTCTAGGTGCTAGGTGCTTACACAGCAAGCTGCTGCTGCTG
CATGGATATCTGCTCTGGGCCCCAGGATGCAATTTGGAGGGCTTCTTTCAGCCCTG
GGCGGTATGAGCGGGGTGTGGGTGGGTGTGACAGGAGCCGGGAGCATGGAGGGGTGG
GAGAGTCCCGGGCTTGGCGGTGTGGGTGAGAGGCCCTTCTCCCTTCTCCTGCTGCTC
TGTTATCCAAAGCCCTCATATATTACGTCAACAAACACCATTCATGGTATAGCCGGG
GCTGTTTGTGACGGGCTGGCA
(Variant highlighted)
AGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTCCGACGATCTTGGGTG
AGCAGCCACGGGTGACCCACAAGGGCCACAGCCATGAATGGCACAGAGGCCCTAATTCT
TACGTGCTCTTCCAAATGACCAAGTACCTGCTGAGCCATGGGAGTTCTCCATG
TACCTGGCTGAGCAGTGGCAGTCTCTCATGCTGGCGCTCATGTTCTGCTGATCGTG
CTGGGCTTCCSCTACCACTCTCAGCTCTACGCTCAGCCGTCAGCACAAGAGCTGCGC
ACGCTCTCAACTACATCTGCTCAACCTAGCCGTGGCTGAGCTCTTCTGCTGCTAGGT
GGCTTACACGACCTCTACACCTCTCTGATGATGATCTGCTGCTTGGGGCCACAGGA
TGCAATTGGAGGGCTTCTTGGACCTTGGCGGTGAAATTGGCTGCTGCTGCTTGGTG
GTCTTGGCCATCGAGCGGTAC
(Variant highlighted in red, differences highlighted in blue)

```

Ensembl release 60 - Nov 2010 © WTSI / EBI

About Ensembl | Contact Us | Help

Permanent link - View in archive site

Links to associated phenotypes

Variation of interest

Go back to the [Gene tab](#). Now, let's focus more closely on a transcript (spliced mRNA). Select the longer transcript from the table (ENST00000296271). This will lead to the Transcript summary display.

Human (GRCh37) Location: 3:129,247,483-129,254,012 Gene: RHO Transcript: RHO-001 Variation: rs28933395

Transcript-based displays

- Transcript summary
- Supporting evidence (13)
- Sequence
 - Exons (5)
 - cDNA
 - Protein
- External References
 - General identifiers (32)
 - Oligo probes (27)
 - Gene ontology (37)
- Genetic Variation
 - Population comparison
 - Comparison image
- Protein Information
 - Protein summary
 - Domains & features (34)
 - Variations (6)
- External Data
- Personal annotation
- ID History
 - Transcript history
 - Protein history
- Configure this page
- Manage your data
- Export data
- Bookmark this page

Transcript: RHO-001 (ENST00000296271)

rhodopsin [Source:HGNC Symbol;Acc:10012]

Location: Chromosome 3: 129,247,483-129,254,012 forward strand.

Gene: This transcript is a product of gene ENSG00000163914 - There are 2 transcripts in this gene

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RHO-001	ENST00000296271	2592	ENSP00000296271	348	Protein coding	CCDS3063
RHO-002	ENST00000511172	454	ENSP00000427467	152	Protein coding	-

Transcript summary help

Supporting evidence »

Export Image

Statistics: Exons: 5 Transcript length: 2,592 bps Translation length: 348 residues

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

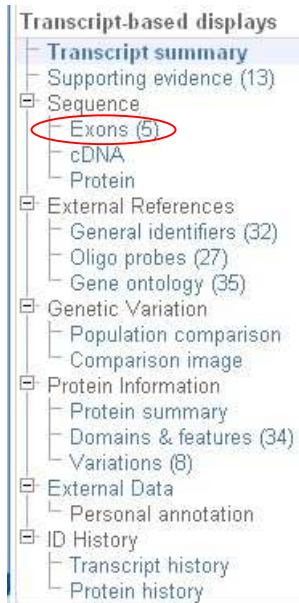
Type: Known protein coding

Prediction Method: Transcript where the Ensembl genebuild transcript and the Vega manual annotation have the same sequence, for every base pair. See [article](#).

Alternative transcripts: This transcript corresponds to the following database identifiers:

Transcript having exact match between ENSEMBL and HAVANA: [OTTHUMT00000356101](#) [view all locations]

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



Choose the **Exons** option first, which highlights exon sequences. (exons, introns and flanking sequence are shown).

« Supporting evidence » **Exons** [help](#) « cDNA sequence »

Show **All** entries Show/hide columns Search:

No.	Exon / Intron	Start	End	Start Phase	End Phase	Length	Sequence
1	5' upstream sequence	129,247,483	129,247,937	-	1	455	gcttaggagggggaggtcactttataa
1	ENSE000001079597	129,247,483	129,247,937	-	1	455	GAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTAGGAGCAGCCACGGGTCAGCCACAGGGCCACAGCCATGCTACGTGCCCTTCTCCAATGCGACGGGTGTGGTACGCCTACCTGGCTGAGCCATGGCAGTCTCTGAGCCGCTGGGCTTCCCATCAACTTCTCAGCCAGCGTCACGCCTCTCAACTACATCCTGCTCAACCAGCCGCTGGCTTACCAGCACCCTCTACACCTCTCTGCATGGAATGCAATTTGGAGGGCTTCTTTGCCACCCCTGGGCG
2	Intron 1-2	129,247,938	129,249,718			1,781	gtatgagccgggtgtgggtgggtg.....tgGTGAAATTGCCCTGTGGCTCTTGGTGGCTTGGCCATAGCCCATGAGCAACTTCCGCTTGGGGAGAACCATGGGTCATGGCGCTGGCTGGCGCCGACCCCACTCGC
2	ENSE000001152211	129,249,719	129,249,719				

Green: flanking sequence

Purple: UTR

Blue: introns

Black: coding sequence

Configure this page Manage your data Export data Bookmark this page Download view as RTF

Click on the **Help** button (circled in red) for page specific help. A link to the glossary, FAQs, and videos 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). If you would like to export this view, including the colours, click **Download view as RTF**. A “Rich Text Format” document will be generated that can be opened in Word.

Now click the **cDNA** link to see the spliced transcript sequence.

Home > Human [GRCh37]
Location: 3:129,247,483-129,254,012 Gene: RHO Transcript: RHO-001 Variation: rs28933994

Transcript: RHO-001 (ENST00000296271)
rhodopsin [Source:HGNC Symbol;Acc:10012]

Location [Chromosome 3: 129,247,483-129,254,012](#) forward strand.
Gene This transcript is a product of gene [ENSG00000163914](#) - There are 2 transcripts in this gene

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RHO-001	ENST00000296271	2592	ENSP00000296271	348	Protein coding	CCDS3063
RHO-002	ENST00000511172	454	ENSP00000427467	152	Protein coding	-

Exons [cDNA sequence](#) [help](#)

Key

Codons [Alternating codons](#) [Alternating codons](#)

Exons [Alternating exons](#) [Alternating exons](#)

Variations [3' UTR](#) [5' UTR](#) [Non-synonymous coding](#)

Other features [UTR](#)

```

1  GAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTTCRCAGCATTCTTGGGTGG
.....
61  GAGCAGCCRCGGGTACGCCACAAGGGCCACAGCCATGAATGGCACAGAAGGCCCTAACTT
.....ATGAATGGCACAGAAGGCCCTAACTT
.....-M--N--G--T--E--G--P--N--F

121 CTAGTGCCCTTCTCCAAATGCGACGGGTGTGGTACGCAGCCCCCTTCGAGTACCCACAGTA
27  CTACGTGCCCTTCTCCAAATGCGACGGGTGTGGTACGCAGCCCCCTTCGAGTACCCACAGTA
9  --Y--V--P--F--S--N--A--T--G--V--V--R--S--P--F--E--Y--P--Q--Y

181 CTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCCTACATGTTTCTGCTGATCGT
87  CTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCCCTACATGTTTCTGCTGATCGT
29  --Y--L--A--E--P--W--Q--F--S--M--L--A--Y--M--F--L--L--I--V

241 GCTGGGCTTCCSCATCAACTTCTCTCAAGCTCTACGTCACCGTCCAGCACAAAGAGCTGCG
147 GCTGGGCTTCCCATCAACTTCTCTCAAGCTCTACGTCACCGTCCAGCACAAAGAGCTGCG
49  --L--G--F--P--I--N--F--L--T--L--Y--V--T--V--Q--H--K--R--L--R

```

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.

Sequence variants are 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.

Transcript: RHO-001 (ENST00000296271)

Description: rhodopsin [Source:HGNC Symbol;Acc:10012]
 Location: Chromosome 3: 129,247,483-129,254,012 forward strand.
 Gene: This transcript is a product of gene [ENSG00000163914](#) - There are 2 transcripts in this gene

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RHO-001	ENST00000296271	2592	ENSP00000296271	348	Protein coding	CCDS3063
RHO-002	ENST00000511172	454	ENSP00000427467	152	Protein coding	-

« Protein sequence » **General identifiers** [help](#) « Oligo probes »

This transcript corresponds to the following database identifiers:

External database type	Database identifier
CCDS	CCDS3063.1 [view all locations]
EntrezGene	RHO rhodopsin [view all locations]
European Nucleotide Archive (was EMBL)	AB065668 [align] [view all locations] BC112104 [align] [view all locations] BC112106 [align] [view all locations] B537381 [align] [view all locations] S81166 [align] [view all locations] S81169 [align] [view all locations] U16824 [align] [view all locations] U49742 [align] [view all locations]
HGNC (curated)	RHO-001 [view all locations]
HGNC Symbol	RHO rhodopsin [view all locations]

This page can be ordered by “External database type” by clicking on the arrow (circled in red in the above figure) next to the title.

Other views include microarray probes and gene ontology terms from the GO consortium (www.geneontology.org).

Click on [Protein summary](#) to view mapped domains and signatures.

Name	Transcript ID	Length (bp)	Protein ID	Length (aa)	Biotype	CCDS
RHO-201	ENST00000296271	2758	ENSP00000296271	348	Protein coding	CCDS3063

« Comparison image » **Protein summary** [help](#) « Domains & features »

Protein
 Transmembrane...
 Low complexity...
 Superfamily do...
 Prints domain

Pfam domain
 PROSITE profiles
 PROSITE patterns
 Variations
 Scale bar

Ensembl protein

Signatures mapped to the sequence

Clicking on [Domains & features](#) shows a table of protein signatures.

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

Ensembl Account | Logout | BLAST/BLAT | BioMart | Tools | Downloads | Help | Documentation | Mirrors

Human (GRCh37) Location: 3:129,247,483-129,254,012 Gene: RHO Transcript: RHO-001

Location-based displays

- Whole genome
- Chromosome summary
- Region overview
- Region in detail**
 - Comparative Genomics
 - Alignments (image) (52)
 - Alignments (text) (52)
 - Multi-species view (48)
 - Synteny (15)
 - Genetic Variation
 - Resequencing (2)
 - Linkage Data
 - Markers
 - Other genome browsers
 - UCSC
 - NCBI
 - Vega

Configure this page | Manage your data | Export data | Bookmark this page

Region overview **Region in detail** help

RHO and neighbouring genes

Chromosome bands
Contigs
Ensembl/Havana g...
ncRNA gene
ncRNA pseudogene
Gene Legend
Ensembl Homo sapiens version 60.37e (GRCh37) Chromosome 3: 128,750,748 - 129,750,747
processed transcript
pseudogene
RNA gene

Location: 3:129247483-129254012 Go
Gene: Go

Change the focus to another gene or region

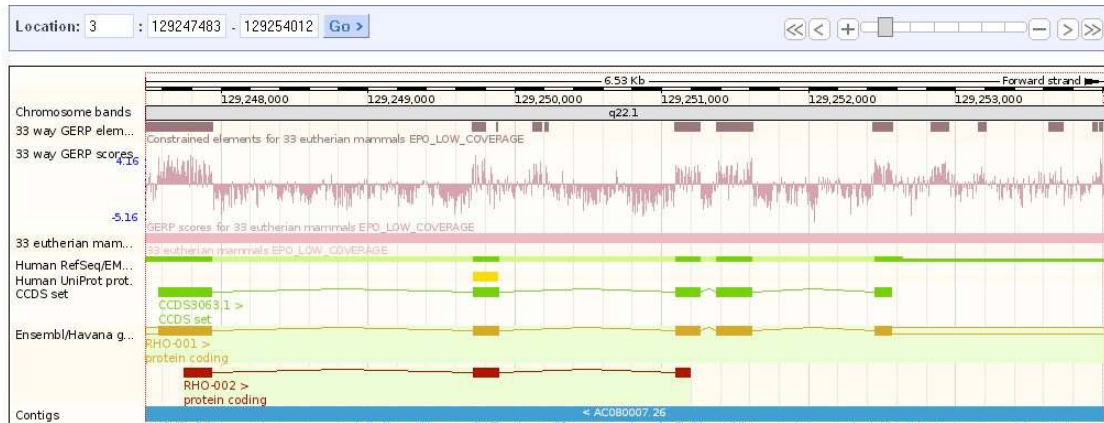
RHO transcripts

Chromosome bands
Human RefSeq/EM...
CCDS set
Ensembl/Havana g...
Contigs
Reg. Feats
Gene Legend
Reg. Features Lege...
There are currently 300 tracks turned off. To change the tracks you are displaying, use the "Configure this page" link on the left.

Ensembl release 60 - Nov 2010 © WTG / EBI
About Ensembl | Contact Us | Help
Permanent link - View in archive site

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 **Sequence variants (all sources)** to the display. Also, choose to view the **Human UniProt prot.** track in normal, expanded form by choosing the **labels** option. Click on the **"Multiple alignments"** menu, and choose the three tracks for the **34 eutherian mammals** (including **"Conservation score"** and **"Constrained elements"**). Close the menu.



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

```
>3 dna:chromosome chromosome:GRCh37:3:129247482:129254177:1
ACAGTCATCCAGCTGGAGCCCTGAGTGGCTGAGCTCAGGCCCTTCGCAGCATTCTTGGGTG
GGAGCAGCCACGGGTGAGCCACAAGGGCCACAGCCATGAATGGCACAGAAAGCCCTAACT
TCTACGTGCCCTTCTCCAATGCCAGGGGTGTGGTACGCAGCCCTTCGAGTACCCACAGT
ACTACCTGGCTGAGCCATGGCAGTTCTCCATGCTGGCCGCTACATGTTTCTGCTGATCG
TGCTGGGCTTCCCATCAACTTCTCAGCTCTACGTACCCGTCCAGCACAAAGAGCTGC
GCACGGCTCTCAACTACATCCTGCTCAACCTAGCCGTGGCTGACCTCTTATGGTCTAG
GTGGCTTCCACAGCACCCTCTACACCTCTCTGCATGGATACTTCTGCTTCCGGGCCACAG
GATGCAATTTGGAGGGCTTCTTTGCCACCCTGGGCGGTATGAGCCGGGTGTGGGTGGGT
GTGCAGGAGCCCGGAGCATGGAGGGGTCTGGGAGAGTCCCGGGCTTGGCGGTGGTGGCT
GAGAGGGCTTCTCCCTTCTCCTGTCTGTCAATGTTATCCAAAGCCCTCATATATTCAGT
CAACAAACACCATTCATGGTGATAGCCGGGTGCTGTTTGTGAGGGCTGGCACTGAACA
CTGCCCTTGATCTTATTTGGAGCAATATGCGCTTGTCTAATTTACAGCAAGAAAAGTGA
CTGAGGCTCAAGAAAGTCAAGCGCCCTGCTGGGCGTCAACAGGGACGGGTGCAGAGTT
GAGTTGGAAGCCCGCATCTATCTCGGGCCATGTTTGCAGCACCAGCCCTCTGTTTCCCTT
GGAGCAGCTGTGTGAGTCAGACCCAGGCTGGGCACTGAGGGAGAGCTGGGCAAGCCAGA
CCCCCTCTCTCTGGGGGCCAAGCTCAGGGTGGGAAGTGGATTTCATTCTCCAGTCAT
TGGGTCTTCCCTGTGCTGGGCAATGGGCTCGGTCCCTCTGGCATCCTCTGCTCCCTC
TCAGCCCTGTCTCAGGTGCCCTCCAGCCTCCCTGCCGCTTCCAAGTCTCCTGGTGT
TGAGAACCCGCAAGCAGCCGCTCTGAAGCAGTTCCTTTTGTCTTGAATAATGTCTTGCA
TTTAACAGGAAAACAGATGGGGTGTGTCAGGGATAACAGATCCCACTTAACAGAGAGGAA
AACTGAGGCAGGGAGAGGGGAAGAGACTCATTAGGGATGTGCCAGGCAGCAACAAGAG
CCTAGGTCTCCTGGCTGTGATCCAGGAATATCTCTGCTGAGATGCAGGAGGAGACGCTAG
```

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



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

Important Notice
We now used Blat 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 dna: chromosome chromosome:GRCh37:3:129247482:129254177
AGAGTCATCCAGCTGAGCCCTGAGTGGCTGAGCTCAGGCTTCGCGCATTCTTGG
GGAGCAGCCACAGGTCAGCCACAAAGGCTCAGCCATGAATGGCAGAAAGCCCTA
TCTACGTGCCCTTCTCCAATGCGAGGGGTGTGTACGAGCCCTTCGA9TACCAC.
```

 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:

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 impressions of BlastView, especially regarding functionality that you would like to see provided in the future. Many thanks for your time. [Feedback](#)

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)

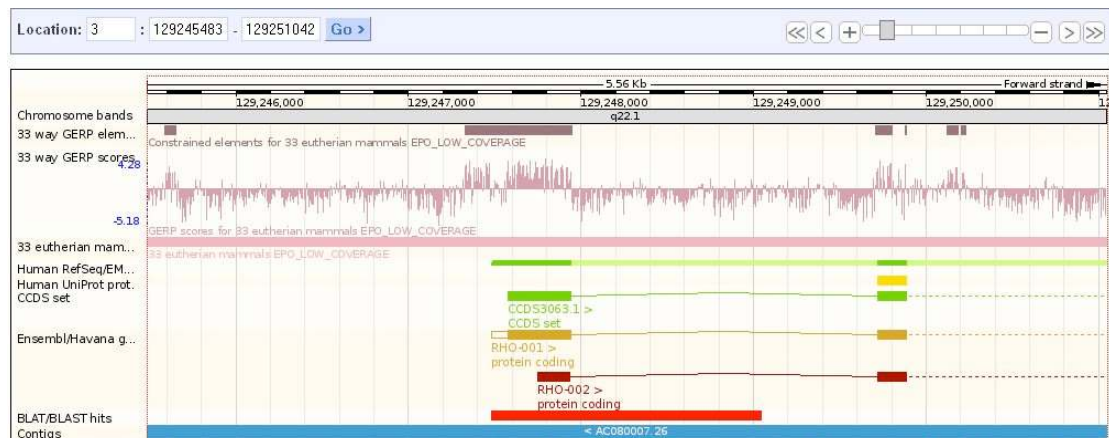
Alignment Locations vs. Query (click arrow to hide)

Alignment Summary (click arrow to hide)
 Selected rows to include in table, and type of sort (Use the ctrl key to select multiples)

Query	Subject	Chromosome	Supercontig	Clone	Contig	Chromosome	Chromosome	Chromosome	Stats	Sort By
Name	Name	Name	Name	Name	Name	Name	Name	Name	Score	>Chromosome
Start	Start	Start	Start	Start	Start	Start	Start	Start	E-val	<Score
[A]	[S]	[G]	[C]							
2382 +	Chr:3	129247482	129248863 +	Chr:3	129247482	129248863 +	Chr:3	129247482	129248863 +	12042 0.0e+00 100.00 2382

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!).

Clicking on [C] should reveal the BLAT hit in “Region in detail”:



You may click on the red bar to see the score, %ID, and other BLAST/BLAT values.

2.1 Tying it together – The human LDLR gene

You are working with a transcript of the human **LDLR** gene that has 18 exons. Start at www.ensembl.org.

- 1) For what is **LDLR** an acronym (what is the expanded name)?
- 2) Go to the transcript tab for the 18-exon transcript.
- 3) Look under **External References** at the left. Find the **General identifiers** section for the transcript.
 - What Swiss-Prot record matches to the Ensembl protein? What NCBI RefSeq protein matches to the Ensembl protein? (This will be an ID starting with NP).
 - Does MIM (Online Mendelian Inheritance in Man) show this protein to be related to any diseases?
- 4) Click **Gene ontology**. What functions are listed for LDLR?
- 5) Click on **cDNA** at the left. Where does the 5' UTR end (which base pair number, in the transcript sequence?) Are there variations that change the protein sequence?
- 6) Click on **population comparison** at the left of the transcript tab.
 - For the non-synonymous SNP in Venter's genome, what is the allele in the reference sequence, and what are the alleles in Venter's sequence?
- 7) Jump to the variation by clicking on the ID (rs11669576) Click on **population genetics** at the left.
 - What genotypes are found in the Yoruba population (CSHL-HAPMAP-YRI)? Do these differ from other populations studied?
- 8) Click **Gene/Transcript** at the left. This lists Ensembl genes and transcripts affected by the variation.
- 9) Click on the gene identifier (**ENSG00000130164**).
- 10) The **Variation Image** should be shown
 - To clarify the image, turn off the variation source "HGMD-PUBLIC" in *configure this page*.

- Non-synonymous and synonymous SNPs are shown in yellow and green, respectively. The possible amino acids at each position are shown.

11) Click **Genomic alignments** at the left. Multispecies alignments can be seen in this view.

- Select “6 primates EPO” in the Alignment menu at the top. Click “Go”.
- Click configure this page. Change Conservation regions from “none” to “all conserved regions” (if they are not already turned on).
- Click the “6 primates EPO” link at the left of the menu.
- Deselect “Ancestral sequence”. Click [✓]
- Blue shading shows identical nucleotides. Can you see the insertion in the *Callithrix jacchus* (Marmoset)?
- Turn on variations using “configure this page” and the “Variations: yes and show links” option.
- Click [✓] to close.
- Can you find rs11669576 in the sequence (hint: search with your browser and look for the link at the right, and yellow variation)

2.2 EXERCISES and ANSWERS

Note: the answers to these exercises correspond to current version (60) of Ensembl. If you use these exercises in the future, after Ensembl is updated, please use the archive site for version 60.

2.2.1 BROWSING ENSEMBL (exercises)

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.

Exercise 1 – Exploring a gene

(a) Find the human F9 (Coagulation factor IX Precursor) gene. On which chromosome and which strand of the genome is this gene located? How many transcripts (splice variants) are there? How many of these transcripts are protein coding?

- (b)** What is the longest Ensembl transcript? How long is the protein it encodes? How many exons does it have? Are any of the exons completely or partially untranslated?
- (c)** Which transcripts have a CCDS record associated with it? What does that mean?
- (d)** Have a look at the external references for ENST00000218099. What is the function of F9?
- (e)** Are there probesets on microarray platforms that can be used to monitor ENST00000218099 expression?
- (f)** Does ENSP00000218099 show peptidase activity?
- (g)** Go to the Protein summary view, if you are not there already. Click on the peptidase signature in the Pfam track. Note the ID number. Now click on the peptidase signature in the SMART track. Follow the link to the InterPro record. Do you see the Pfam ID?
- (h)** How many non-synonymous SNPs have been discovered for the protein encoded by ENST00000218099?
- (i)** Is there a mouse orthologue predicted for the human F9 gene?

Exercise 2 – Exploring a region

- (a)** Go to the region from bp 52,600,000 to 53,300,000 on human chromosome 4. What does the transparent part in the 'Contigs' track represent? (If you are not sure what contigs are, click on Help and then Glossary.)
 - (b)** Zoom in on the SGCB transcript, including a bit of flanking sequence on both sides.
 - (c)** Is there a protein from UniProtKB that aligns to the genome at the same location as SGCB?
 - (d)** Export the genomic sequence of the region you are looking at in FASTA format.
-

2.2.2 BROWSING ENSEMBL (answers)

- 1) **(a)** Go to the Ensembl homepage (<http://www.ensembl.org>).

Select '**Search: Human**' and type '**F9 gene**'

Click [Go].

Click on '**Homo sapiens**' on the page with search results.

Click on '**Gene**'.

Click on '**Ensembl protein_coding Gene: ENSG00000101981 (HGNC Symbol: F9)**'.

The human F9 gene is located on the X chromosome on the forward strand. Ensembl has three transcripts annotated for this gene, ENST00000218099 (F9-001), ENST00000394090 (F9-201), and ENST00000479617 (F9-002). Only F9-001 and F9-201 are protein coding.

(b) Look at the table. The longest transcript is ENST00000218099. The length of this transcript is 2780 base pairs and the length of the encoded protein is 461 amino acids. To see the exons, click on **ENST00000218099**. It has eight exons. Click on '**Sequence - Exons**' in the side menu. The first and last exon are partially untranslated (UTR sequence shown in purple).

Click "Show/hide columns" at the top to deselect columns you don't want (for example, select only "Sequence" to show the sequence alone).

(c) In the table of transcripts, there is a CCDS ID for F9-001 and F9-201. CCDS is the consensus coding sequence set. These coding sequences (CDS) have been agreed upon by Ensembl and NCBI. Since these two transcripts have the same CDS, they differ in the UTR alone.

(d) Click on '**External References - General identifiers**' in the side menu. Explore some of the links (good places to start are usually 'WikiGenes' and 'UniProtKB/Swiss-Prot').

Do the same for '**Gene ontology - Ontology table**'.

Factor IX is a vitamin K-dependent plasma protein that participates in the intrinsic pathway of blood coagulation by converting factor X to its active form in the presence of Ca^{2+} ions, phospholipids, and factor VIIIa (this is the description as found in UniProtKB/Swiss-Prot).

(e) Click on '**External References - Oligo probes**' in the side menu. Probesets from Affymetrix, Aglient, Codelink, Illumina, and Phalanx match to this transcript sequence.

(f) Click on 'ENST00000218099' if you are not already on the

'Transcript: F9-001' tab. Click on '**Protein Information - Protein summary**' in the side menu. And/or click on '**Protein Information - Domains & features**' in the side menu.

Alternatively, click on the link **ENSP00000218099** to jump to the **protein summary**. Peptidase signatures are found in the protein sequence, indicating there may be peptidase activity. Click on the "Peptidase" bars to see that the signature covers a range of amino acids (the different numbers reflect different lengths of the Peptidase signature in different databases, i.e. Superfamily, SMART, Pfam and PROSITE).

(g) The Pfam ID for the Peptidase_S1_S6 signature is PF00089. The InterPro ID is IPR001254. InterPro integrates information from different databases. The SMART, PROSITE, and Pfam signatures have been clustered into one InterPro record. PF00089 is listed in the "Signatures" section of the IPR001254 record. This can also be seen in the "Domains and features" table.

(h) Click on '**Protein Information - Variations**' in the side menu. For the protein encoded by ENST00000218099 three non-synonymous SNPs have been discovered: rs6048, rs1801202 and rs4149751.

(i) Click on the '**Gene: F9**' tab.
Click on '**Orthologues**' in the side menu.
There is one mouse orthologue predicted for human F9, ENSMUSG00000031138.

Answer

2(a) Go to the Ensembl homepage.
Type '**human 4:52600000..53300000**' in the '**Search**' box.
Click [**Go**].

The open part in the 'Contigs' track represents a gap in the genome assembly (although the human genome is called 'finished', there are still gaps!). Note that this region is very close to the centromere of the chromosome.

(b) **Search for SGCB** in the Gene box of the main panel. Zoom out one step in the zoom slide. Alternatively, **draw a box with your mouse** around the **SGCB** transcript. Click on 'Jump to region' in the pop-up menu.

(c) If it is not turned on, turn on the UniProt track as follows:
Click on '**Configure this page**' in the side menu.
Type '**UniProt**' in the 'Search display:' text box.

Select 'Human UniProt prot - Labels.'

Close the menu.

NP_000223.1 aligns to the genome with the same exon structure as SGCB. Q16585 is another UniProt protein that aligns in the same place as one exon of SGCB. Click on it to see it is Beta-sarcoglycan.

(d) Click on 'Export data' in the side menu.

Click [Next>].

Click on 'Text'.

Note that the sequence has a header that provides information about the genome assembly (GRCh37), the chromosome, the start and end coordinates and the strand. For example:

>4 dna:chromosome chromosome:GRCh37:4:52883261-52908260:1

2.2.3 USER UPLOAD (exercises and answers)

Exercise 1 – Attaching a file

Have a look at the following file:

http://www.ebi.ac.uk/~bert/n-scan_genes.txt

It contains annotations for three transcripts of the human HFE gene (ENSG00000010704) generated by the N-SCAN gene structure prediction software, as shown on the UCSC Genome Browser (<http://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=148199581&c=chr22&g=nscanGene>).

The file is in GFF (General Feature Format) format:

<http://genome.ucsc.edu/FAQ/FAQformat>

Attach the file to Ensembl and have a look at the result.

Answer

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Select 'Search: Human' and type 'hfe gene' in the 'for' text box.
- ☞ Click [Go].
- ☞ Click on 'Gene' on the page with search results.
- ☞ Click on 'Homo sapiens'.
- ☞ Click on '[Region in detail]' behind 'Ensembl protein_coding Gene: ENSG00000010704 (HGNC Symbol: HFE)'.

- ☞ Click [Manage your data] in the side menu.
- ☞ Click on 'Attach URL Data'.
- ☞ Enter the URL of the file in the 'File URL' text box.
- ☞ Enter 'N-SCAN genes' in the 'Name for this track' text box.
- ☞ Click [Next>].
- ☞ Click [✓].

A new track named 'N-SCAN genes' should now have been added to the Region in detail page.

To display the names of the transcripts:

- ☞ Click [Configure this page] in the side menu.
- ☞ Select 'N-SCAN genes - Labels'.
- ☞ Click [✓].

Note that, at the moment, the CDS information in the GFF file is not taken into account in Ensembl and thus no distinction between the UTR and CDS of the transcripts can be seen.

Exercise 2 – Uploading data

Have a look at the following file:

http://www.ebi.ac.uk/~bert/medip-chip_cd4.txt

It contains methylation values in the genomic region of the human ICAM3 gene (ENSG00000076662) in CD4 cells (Rakyan *et al.* Genome Res. 2008 Sep;18(9):1518-29).

The file is in BED (Browser Extensible Data) format:

<http://genome.ucsc.edu/FAQ/FAQformat>

Note that only the chrom, chromStart, chromEnd and score have been used in this case. To display the data as a histogram the useScore attribute in the track line is set to 3.

Upload the data to Ensembl and have a look at the result.

MeDIP data for CD4 cells is also available as a DAS track in Region in Detail. Can you find it?

Answer

- 🔗 Go to the Ensembl homepage (<http://www.ensembl.org>).
- 🔗 Select 'Search: Human' and type 'icam3 gene' in the 'for' text box.
- 🔗 Click [Go].
- 🔗 Click on 'Gene' on the page with search results.
- 🔗 Click on 'Homo sapiens'.
- 🔗 Click on '[Region in detail]' behind 'Ensembl protein_coding Gene: ENSG00000076662 (HGNC Symbol: ICAM3)'.
- 🔗 Click [Manage your data] in the side menu.
- 🔗 Click on 'Upload Data'.
- 🔗 Enter 'MeDIP-chip CD4' in the 'Name for this upload (optional)' text box.
- 🔗 Copy/paste the data from the medip-chip_cd4.txt file in the 'Paste file' text box.
- 🔗 Click [Upload].
- 🔗 Click [✓].

A new track named 'MeDIP-chip CD4' should now have been added to the Region in detail page.

- 🔗 Click [Configure this page] in the side menu.
- 🔗 Search the menu for MeDIP
- 🔗 Turn on the track for MeDIP-CD4
- 🔗 Click [✓].

A second track should now appear for the MeDIP-chip data.

Exercise 3 – Creating and uploading an annotation file

Create a small text file containing some annotation in either BED or GFF format (<http://genome.ucsc.edu/FAQ/FAQformat>) and upload it to Ensembl.

Note that BED offers the simplest format, with only three required fields, i.e. chrom, chromStart and chromEnd, so it's probably easiest to start with this.

Answer

- 🔗 Create a text file with your annotation in for example Notepad or TextEdit and save it on your computer.
- 🔗 Go to the region you have annotated in Region in detail.
- 🔗 Click [Manage your data] in the side menu.
- 🔗 Click on 'Upload Data'.

- ☞ Enter the name for your track in the 'Name for this upload (optional)' text box.
- ☞ Click [Browse...] behind 'Upload file:'.
- ☞ Select the text file you just created.
- ☞ Click [Upload].
- ☞ Click [✓].

Your data should now be shown as a new track on Region in detail.

Exercise 5 – Removing custom annotation

Remove your attached and uploaded annotations.

Answer

- ☞ Click [Manage your data] in the side menu.
- ☞ Click for each dataset on 'Delete'.
- ☞ Click [✓].

Your annotations should be removed now.

2.2.4 Variations and functional genomics (optional exercises)

Exercise 1

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

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

Exercise 2

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

- (a) Do both individuals (Venter and Watson) have sequence known at the position of the R620W (C1858T) SNP? Look at “Resequencing coverage” to answer this question.

(b) Does either individual have the minor allele?

Exercise 3

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

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

Exercise 4 Gene regulation: STX7

(a) Find the Location tab, "Region in Detail" page for the STX7 gene. Are there regulatory features in this gene region? If so, where in the gene do they appear?

(b) Click on a grey box at the 5' terminal end of the STX7 transcripts (remember, this is one of the rightmost boxes as the STX7 gene is on the reverse strand). A pop-up box should appear, listing an ID such as "ENSR00000131372" for this feature. Click on the ID to open the Regulation tab.

In what cell lines is the core evidence found?

(c) Click on the gold exon that is shown just under the blue bar. Follow the link to ENSG00000079950. Click on the "Regulation" link at the left of the gene tab.

2.2.5 Variations and functional genomics (answers)

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 symbol: PTPN22)'.

(a) Click on '**Variation Table**' in the side menu.

Select "**Show**" in front of "Non-synonymous coding" variations, in the table.

Two of the four PTPN22 transcripts contain a SNP with amino acid (aa) change W/R and aa co-ordinate 620. The 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 Data**' at the left of the Variation page. The allele A is associated with Type 1 Diabetes, however, T and G are also risk alleles. The A allele is associated with Vitiligo.

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**'.

The reference sequence shows the minor allele (A) at this position. Click on the green box corresponding to this SNP to view the allele.

Neither Venter nor Watson is homozygous for the minor allele (A) of rs2476601, which predisposes one for rheumatoid arthritis.

Watson is heterozygous at the position for rs2476601.

Answer Exercise 3

Go to the Ensembl homepage

Click the BioMart link on the toolbar.

Choose the 'Ensembl Genes 60' database. (*Note, you can also do this query using the 'Ensembl Variation 60' 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 – 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)', and 'Protein Allele'.

Click the [Results] button on the toolbar.

View all rows, or select 'Export all results to file – XLS' (unique results only) and click [Go].

Answer Exercise 4

(a) Search for “**human gene STX7**” from the home page. Click on “Region in detail” from the search results.

Regulatory features from the Ensembl “regulatory build” are based on indicators of open chromatin such as CTCF binding sites, DNase I hypersensitive sites, and transcription factor binding sites. Find them in the “Reg. Feats” track. Click on the “Reg. Feats” track name to jump to an article explaining the underlying data.

Shaded grey boxes indicate regulatory features. There are quite a few of them mapping to the STX7 transcripts, including the 5' end.

(b) At the top of the Regulation tab, a list of cell lines is shown. Core evidence is available for 10 cell lines, including CD4. “Promoter associated” indicates there are promoter-like histone modification patterns for that cell type. The evidence described in the top list are shown graphically in the first panel under “Details by cell line”. These are shown as green blocks.

Below the multi-cell graphic, evidence for each cell line is shown in separate panels.

(c) The regulation page shows elements associated with gene regulation from multiple sources. Multiple “regulatory features” are shown as shaded boxes in the diagram, and are listed in the table below. In addition, miRNA targets from the miRanda project are indicated, and CisRED motifs are shown.

2.2.6 Ensembl Plants (exercises optional)



As detailed in the paper, ‘Regulation of plant innate immunity by three proteins in a complex conserved across the plant and animal kingdoms’ (*Genes Dev.* 2007 21: 1484-1493), the gene products of MOS4, ATCDC5 and PRL1 form a protein complex, which is essential for immunity in *Arabidopsis thaliana* and for plants in general.

The first exercise will focus on using the gene and transcript tabs of the Ensembl Plants genome browser to find out more about the MOS4 gene in *Arabidopsis thaliana*.

N.B. The answers to these questions are valid for Ensembl Plants release 6 – August 2010. Should you use these exercises at a later date, be aware the answers may change due to updates.

1. Adding information to a gene name

- (a) From the Ensembl Genomes homepage (www.ensemblgenomes.org) click on '**Plants**' to take you to the homepage for Ensembl Plants. Under '**Search Ensembl Plants**' type in '**MOS4**'.

Why do you get multiple search results? What are the gene names of the search results and to which species do they belong?
- (b) Click on the gene named '**AT3G18165-TAIR-G (Gene (TAIR): MOS4)**'. How many transcripts are associated with this gene? Where does the gene model come from?
- (c) Click on '**Gene Tree (image)**' under '**Plant Compara**'. According to the tree, does this look like a conserved protein in plants? Which 3 genes seem most closely related to '**MOS4**' and to which organisms do they belong?
- (d) Click on '**Orthologues**' under '**Pan-taxonomic Compara**'. How many orthologues are there for '**MOS4**' and what are their orthologue types? What is the gene name for '**MOS4**' in human? Is the '**MOS4**' protein larger than the human orthologue?
- (e) Click on the '**Transcript**' tab. How many exons and introns does '**MOS4**' have? How large is the transcript? Click on '**Variations**' under '**Protein Information**', and also on the '**Protein**' link under '**Sequence**'. How many SNPs are associated with the protein sequence, and where are they?
- (f) Click on '**Gene ontology**' under '**External references**' to see associated functions and cellular location of this protein. Where is the '**MOS4**' protein located in the cell? What plant structure is associated with the protein? How was the evidence for these terms derived?
- (g) Click on '**Protein summary**' under '**Protein Information**'. What is the largest protein domain and where (in aa) does it fall on the protein? The Interpro record is a good place to find more information about the domain- click on an InterPro link or ID (IPR008409) to get there. Is this domain also found in other species?

The next exercise uses the gene LOC_Os09g22450, found in *Oryza sativa*. LOC_Os09g22450 is the orthologue of EDS1 found in *Arabidopsis thaliana*, and is

thought to play a significant role in disease resistance. You will use the location tab to explore the region around the LOC_Os09g22450 gene.

‘Direct interaction between the *Arabidopsis* disease resistance signaling proteins, EDS1 and PAD4’. The EMBO Journal (2001) 20, 5400 – 5411.

2. Exploring a region

- (a) Go to the species page for *Oryza sativa* and search for ‘**9:13548161-13593345**’ (no spaces), which represents chromosome 9 at the genome co-ordinates 13548161bp to 13593345bp. What chromosome arm does this region fall onto? What contig on the genome assembly does this region fall onto? How many genes are shown in this region?
- (b) Identify the ‘**LOC_Os09g22450.1**’ gene. What strand does it fall on? How many transcripts are associated with this gene?
- (c) Click on ‘**Multi-species view**’ and ‘Select species’ to add the alignment for ‘***Oryza sativa indica* group**’. Close the menu to see an alignment of the regions in both rice genomes. Identify the ‘**LOC_Os09g22450**’ gene orthologue in the ‘***Oryza sativa indica* group**’. What is it called? How many other orthologues can be identified for this region? Add the *Sorghum bicolor* alignment. Can you view the orthologues to *Oryza sativa* LOC_Os09g22450 in another way?
- (d) Click on the ‘**LOC_Os09g22450**’ gene and select the link next to ‘**Location**’, which will take you back to the ‘**Region in detail**’ with the ‘**LOC_Os09g22450**’ gene selected. FGENESH is a prediction program that can be useful for finding new genes, however it tends to overpredict. Turn this track on, along with variations. Is there an FGENESH predicted transcript at this locus? Which exon contains the most variations?
- (e) Using ‘Multi-species view’, show orthologue(s) in *Arabidopsis thaliana* to ‘**LOC_Os09g22450**’. What would the orthologue type be (One to one? One to many?)
- (f) How many exons in ‘**LOC_Os09g22450**’ appear to be conserved between *Oryza sativa* and *Arabidopsis thaliana*?

2.2.7 Ensembl Plants (answers)

1. Adding information to a gene name

- (a) Your search for '**MOS4**' should return 3 results. In this case, multiple queries are returned because the term '**MOS4**' is found in the '**Description**' of two of the search results. '**MOS4**' is described as interacting with the '**ATCDC5**' and '**PRL1**' genes listed for *Arabidopsis thaliana*.
- (b) The '**MOS4**' gene has 1 transcript, as shown on the '**Gene page**' by '**Transcripts**'. You can also see this graphically underneath the '**Gene summary**' using the transcript view. '**Prediction Method**' also under '**Gene summary**' shows that the gene was annotated by The Arabidopsis Information Resource (TAIR). This information can also be found by clicking on a transcript on the transcript view graphic.
- (c) The significant protein alignment lengths (green bars) suggest that this gene is conserved in plants. "**View fully expanded tree**" (the option is under the gene tree) to see that the longest alignment gap at the C-terminal end of the plant proteins are accounted for by the longer sequence found in the *Drosophila melanogaster* protein. The 3 genes most closely related to '**MOS4**' in *A. thaliana* are fgenesh2_kg.3__2010__AT3G18165.1 (*A. lyrata*), POPTR_0015s04220 (*Populus trichocarpa*) and GSVIVG00017780001 (*Vitis vinifera*), in order of relation. There are 12 proteins in this tree (including MOS4).
- (d) There should be 11 orthologues displayed with the type '**1 to 1**', reflecting the proteins represented in the gene tree. The orthologue for '**MOS4**' in humans is '**breast carcinoma amplified sequence 2**' (BCAS2), as shown under the '**External ref.**' column. Click on '**Alignment**' to align the '**MOS4**' gene with '**BCAS2**' to find that the '**MOS4**' protein is 253aa and the human protein is 225aa long.
- (e) The '**MOS4**' transcript consists of 6 exons, 5 introns and is 1118bp in length as shown under the '**Transcript summary**'. 23 variations are associated with the protein, 12 synonymous and 11 non-synonymous. You may need to '**configure this page**' and select '**show variation features**' to display SNPs on the Protein sequence page.
- (f) '**MOS4**' appears to be located in the nucleus and nucleolus according to the GO terms '**GO:0005634**' & '**GO:0005730**'. Clicking on the '**PO:0000013**' link reveals that the protein is associated with the '**Leaf**

or pairs/whorls of leaves borne on the stem'. Terms were derived by IDA (Inferred from Direct Assay), IMP (Inferred from Mutant Phenotype) and IEP (Inferred from Expression Pattern) (click on the "Help" for more about this).

- (g) The longest protein domain is the Pfam domain '**BCAS2**'. Clicking on the domain reveals that it falls from 11-251aa on the '**MOS4**' protein. Click on '**Domains and features**' to reveal a table of domains and other features. Click on '**IPR008409**' under '**InterPro**' to reveal the taxonomic coverage for this domain. You can also find the same information by clicking directly on the domain in '**Protein summary**' and selecting '**View InterPro**'.

2. Exploring a region

- (a) The region, '**9:13548161-13593345**', falls on the q arm of chromosome 9 (q9), as shown in the top panel and in the main panel (underneath '**Region in detail**') on the '**Chromosome bands**' track for gene and transcript graphical views. The main panel views also show that the region falls on the contig '**AP006562**', as represented on the blue bar. 11 genes are found in this region (look in the main panel).
- (b) The '**LOC_Os09g22450.1**' gene falls on the forward strand as it is represented above the blue contig bar on the main panel. The ">" after the gene name indicates the 5' to 3' direction of translation. You can also find this information by clicking directly on the gene (top panel) or transcript (main panel). 1 transcript is associated with the '**LOC_Os09g22450 gene**'.
- (c) In '**Multi-species view**', to identify orthologues click '**Configure this page**' and switch on '**Join genes**' (which can be found under '**Comparative features**'). When you close the menu, notice the blue lines linking the orthologues. The '**LOC_Os09g22450**' orthologue in '**Oryza sativa indica group**' is called 'NP_001063086 (BGIOSIBCE029814)'; 6 additional orthologues can also be identified for this region between the two rice genomes.

Now add the alignment for '**Sorghum bicolor**'. Notice that *S.bicolor* shares one orthologue, C5XC28_SORBI (gene name Sb02g02400), with the two rice genomes for this region.

For another way to identify orthologues, go to the gene page for LOC_Os09g22450 and click on 'Orthologues'.

- (d) Make sure you are in the Location tab, 'Region in Detail' for LOC_Os09g22450. Click on '**Configure this page**', and turn on '**FGENESH prediction**' found under the '**Prediction Transcripts**' menu and '**All variations**' found under '**Variation features**'. After closing the menu, the page should reload. You should see a predicted transcript that shows an open reading frame '**FGENESH00000016111**' (note there are no UTRs). Exon 3 contains the most variation.
- (e) Click on '**Multi-species view**' and select the alignment for Arabidopsis thaliana. This should reveal that '**LOC_Os09g22450**' is the orthologue for EDS1 in A.thaliana and the orthologue type is '**one to many**' as there is a paralogue for EDS1 in A.thaliana, AT3G48080.
- (f) Exons 2 and 3 seem to be conserved between O.sativa **LOC_Os09g22450** and A.thaliana **AT3G48080**, as shown by the blastz matches and the green shading that signifies alignments between 2 or more genomes in '**multi-species view**'.

2.2.8 Ensembl Bacteria (exercises optional)



According to the paper, '**Essential *Bacillus subtilis* genes**'*, 192 genes out of ~4100 are defined as being indispensable for *B.subtilis*. The first exercise will focus on using the Ensembl Browser Gene and Transcript tabs to find out more about the essential *B.subtilis* gene, **groEL**.

*Kobayashi, K et. al, Proc Natl Acad Sci 2003 Apr 15;100(8):4678-83. Epub 2003 Apr 7

1. Adding information to a gene name

- (a) From the Ensembl Genomes homepage (www.ensemblgenomes.org) click on '**Bacteria**' to take you to the homepage for Ensembl Bacteria. Now click on '***Bacillus subtilis***'. How were genes and other annotation for *B.subtilis* derived? Click on '**Assembly and Genebuild**' at the left. What is the current genome sequence named? How many 'known pseudogenes' are there in *B.subtilis*?
- (b) Type in '**groEL**' into the search Ensembl Bacteria box and click on the match for '**EBBACG00000000348 (EMBL-Bank: groEL)**'. On which strand is the groEL gene? What genes are adjacent to **groEL**?

- (c) Click on '**Sequence**' at the left. The FASTA sequence is shown. What does the FASTA header tell you? What is the difference between the red sequence and the black, highlighted sequence?
- (d) Try to export the cDNA sequence on the forward strand using the "**Export data**" button at the left of the page.
- (e) Click on the '**Protein**' tab. What protein domains are found in the EBBACP00000000349 protein (hint: this is the purple bar at the top, in the "Protein" track). Click on the domain bars to see which amino acids they map to. Click '**Domains & features**' at the left. How many InterPro domains map to the Ensembl protein? Do they give you a clue about the protein function?
- (f) Follow the '**General Identifiers**' link at the left. Which UniprotKB/Swiss-Prot entry is associated with this transcript? How well does it match the Ensembl protein? What is the KEGG match (click on it)?
- (g) Click on the '**Gene ontology**' link. Where is the groEL protein located in the cell, and which biological processes is it involved in? How was the evidence for these GO terms derived?

The next exercise uses the '**Location**' tab to explore a region in the pathogenic bacterium *Mycobacterium tuberculosis H37Rv*. The following questions are based around the mbtA-J gene cluster, which plays a role in enabling *M.tuberculosis* to grow in iron limited extracellular spaces, such as the lungs, by producing mycobactins.*

*Quadri, LE et. al, Chem Biol. 1998 Nov;5(11):631-45.

2. Exploring a region

- (a) Go to the species page for *M.tuberculosis H37Rv* and search for '**mbtD**'. Click on the search result for mbtD. You should be in the gene tab- click on the "**Location Chromosome: 2,667,255-2,670,269**" to go to the **Location tab**.
- (b) Click on '**configure this page**' to turn on '**Sequence**' and '**Translated Sequence**' in the 'Sequence' menu. Zoom into the mbtD gene so that sequence is shown.
- (c) Zoom out so you see the genes upstream and downstream of mbtD-1. On which strand are these three genes? Click on **mbtE-1** and follow the link to location. The display should recentre.

- (d) Click on **Chromosome summary** at the left. How long is the genome?
How many tRNA genes are found by Ensembl?
- (e) Search for RNaseP in the above right-hand corner search box.
(Change the Ensembl icon to 'Ensembl search this species'.)
- Click on the hit to go to the gene page for '**EBMYCG00000003999**'.
How was this non-coding RNA (ncRNA) predicted?
- (f) Export the gene sequence and run BLASTN to see if this RNase P
aligns to the *M avium* genome. Is there an RNase P gene already
annotated in *M avium*?

Ensembl Bacteria (answers)

1. Adding information to a gene name

- (a) The annotation for ***B.subtilis*** is derived from sequences submitted to EMBL/Genbank/DDBJ databases, and information in UniProtKB and Gene Ontology Annotation (GOA), as stated under '**Annotation**' on its specific species page.

Clicking on '**Assembly and Genebuild**' on the left navigation column reveals that there are '**68 KNOWN pseudogenes**' for the **EB2** genome.

- (b) Under '**Gene summary**', the '**groEL-1**' transcript is represented above the '**contigs**' track or genome assembly (blue bar), which means the gene is located on the **forward** strand. The direction of translation is shown by the '>'. You can confirm the strand by clicking directly on the transcript to reveal further information about the gene/transcript.

The genes '**groES**', '**ydiM-1**' and the known pseudogene '**ydzT-1**' are in the vicinity of the '**groEL**' gene.

- (c) The FASTA header is
>chromosome:EB2:Chromosome:649634:652468:1. This indicates the EB2 genome assembly, base pairs 649,634 to 652,468, and the forward strand (indicated by the +1). The red sequence is the groEL-1 gene. Other highlighted sequence shows neighbouring genes.
- (d) Click the blue "**Export data**" button. Leave strand as "**Feature strand**" to export the forward strand (as the groEL-1 gene is on the forward strand, that is the exported strand). Deselect everything except for the "**cDNA**" option (including changing "Genomic" (Unmasked) to "Genomic" (None).) Click "**Next**", then "**HTML**". You should now see the cDNA sequence for this gene.

- (e) In the '**Protein summary**' view, six signatures are shown from six different databases (e.g. HMAP, TIGRfam, Superfamily, Pfam ...). Most of these signatures cover the complete or length of, or most of, the protein. Names are, for example, Chaprinin_Cpn60, Cpn60/TCP-1. For more information about each signature you can click on the bar, and follow the link to the initial database that annotated the protein signature.

The '**Domains and features**' view shows three InterPro IDs for the six signatures. InterPro brings together annotation from various databases, in order to find consensus protein signatures. Click on one or more of the InterPro records. The signatures are chaperonins, supporting the function of the groEL protein as having a role in protein folding.

- (f) Looking under the '**Transcript summary**' or clicking directly on the transcript itself, reveals that the transcript is '**1635 bps**' in length.

Clicking on the '**General identifiers**' link reveals that the UniprotKB/Swiss-Prot entry is '**P28598**' for this transcript. KEGG is the "Kyoto Encyclopedia of Genes and Genomes". The Ensembl GroEL protein (EBBACP00000000349) matches to KEGG entry BSU06030, a record for the GroEL chaperonin.

- (g) Click on the '**Gene ontology**' link. According to evidence derived from '**Inferred from Electronic Annotation**' (IEA), the protein is located in the '**cytoplasm**' and is involved in '**Response to stress**' and '**Protein refolding**' biological processes.

2. Exploring a region

- (a) You should be in Region in Detail, centred on the mbtD gene.
- (b) Use the zoom bar to display a smaller region (e.g. 100 bp).
- (c) Zooming out shows the mbtE and mbtC genes are on either side of mbtD. All three of these genes are on the reverse strand. Click on the '**mbtJ**' gene and click on '**Location**' to recentre the display on the gene.
- (d) The genome is 4,411,532 base pairs long. There are 45 tRNA genes in Ensembl.
- (e) Some types of non-coding RNAs (ncRNAs), in this case RNase P, are predicted using sequences from RFAM (<http://rfam.sanger.ac.uk/>). See this in the gene tab, in the **gene summary** view.

- (f) Export the sequence using the '**Export data**' button at the left. Under 'Options for FASTA sequence', deselect 'cDNA' and change Genomic from 'None' to 'Unmasked'. Click 'Next' and 'HTML'. Copy the sequence, and then click 'BLAST' at the top of the page.

In the 'BLAST' view, paste the sequence in the box. Make sure 'dna queries' is selected, and the species is *m_avium*. Choose 'BLASTN' and click 'Run'.

The job may take awhile to run. Every few minutes, click on 'Retrieve' to see if the results are available. When you see the 'View' button, click on it.

The highest scoring hit is listed first. The score is over 1,000 and the E-value is almost zero, indicating this is a statistically relevant match. The % identity is nearly 89%.

To view an alignment between the *M.tuberculosis* H37Rv RNase P and *M. avium* genome, click on the [A] in the result table. To view the match in the Ensembl browser, click on [C].

There is an *RNaseP* gene already annotated for *M. avium*. (EBMYCG00000014164)

2.2.9 Sea squirt and Diatoms (exercises optional)

Exercise 1 – The Sea squirt genomes

The genome of *Cionaintestinalis* was published in Science in 2002. (*Dehalet al*, The Draft Genome of *Ciona intestinalis*: Insights into Chordate and Vertebrate Origins. *Science* 2002, **298**:2157-2167) The *Ciona intestinalis* genome is the smallest of any experimentally manipulable chordate, and thus provides a good system for exploring vertebrate evolutionary origins.

The exercises below will focus on finding more information about Sea squirt using www.ensembl.org

Genome

- (a) How many different species of Sea squirt are in Ensembl?
- (b) How were the sequences generated?
- (c) What is the length of *Ciona intestinalis*' genome?

Gene specific information

Move to the *Ciona intestinalis* species specific page.
Go to the Transcript tab on chromosome 1q:7770100-7771867.

- (d) What Ensembl gene can you see in the region? To what "gene type" does it belong?
- (e) What is the molecular function and cellular localization of this transcript?
- (f) What domains are annotated for this transcript?

- (g) Can you find out or deduce from this information to which group this protein belongs?
- (h) Is a similar gene present in *Ciona savignyi*? How similar is the sequence?

Exercise 2 –Diatoms

Diatoms are unicellular photosynthetic algae found both in freshwater and marine ecosystems. Two complete genomes of recently diverged diatoms have been sequenced providing an interesting insight into their evolution. Furthermore these sequences play an important role in a better understanding of endosymbiotic gene transfer.

References

- (1) Armbrust *et al.* (2004). The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science* **306**, 79-86.
- (2) Bowler *et al.* (2008). The Phaeodactylum genome reveals the evolutionary history of diatom genomes. *Nature* **456** (7219): 239–244.

Use www.ensemblgenomes.org

Genome

- (a) How many diatoms can you find in Ensembl Genomes? What are their names?
- (b) How many chromosomes have been revealed during the mapping of *Phaeodactylum tricornutum*?
- (c) Compare the genomes of *Thalassiosira pseudomonana* to *Phaeodactylum tricornutum*. What are the main differences?

Gene specific information

Move to the *Phaeodactylum tricornutum* species specific page and search for “ferredoxin--NADP reductase”. Select “Phatr23717”.

- (d) What is the enzyme classification (EC) identifier for this protein?
- (e) How many exons does this gene have? Are they both found in *Thalassiosira Pseudomonana*?
- (f) Can you find orthologs of this gene in both plants and animal kingdoms?

The Diatom Expressed Sequence Tag (EST) Database contains information from 16 cDNA libraries derived from *Phaeodactylum tricornutum* cells grown in different nutrient and stress regimes. In the next part we will explore these datasets * using Ensembl.

*Maheswari *et al.* (2005). The Diatom EST Database. *Nucleic Acids Research***33**, 344–347.

(g) In order to attach EST data go to **Location** tab. Click **Configure your page** and select **EST alignments**. Is this ferredoxin—NADP reductase expressed in iron limited cells?

Exercise 3 – Exploring a human gene

Use www.ensembl.org

(a) Find the human F9 (Coagulation factor IX Precursor) gene. On which chromosome and which strand of the genome is this gene located? How many transcripts (splice variants) are there? How many of these transcripts are protein coding?

(b) What is the longest Ensembl transcript? How long is the protein it encodes? How many exons does it have? Are any of the exons completely or partially untranslated?

(c) Which transcripts have a CCDS record associated with it? What does that mean?

(d) Have a look at the external references for ENST00000218099. What is the function of F9?

(e) Are there probesets on microarray platforms that can be used to monitor ENST00000218099 expression?

(f) Does ENSP00000218099 show peptidase activity?

(g) Go to the Protein summary view, if you are not there already. Click on the peptidase signature in the Pfam track. Note the ID number. Now click on the peptidase signature in the SMART track. Follow the link to the InterPro record. Do you see the Pfam ID?

(h) How many non-synonymous SNPs have been discovered for the protein encoded by ENST00000218099?

(i) Is there a mouse orthologue predicted for the human F9 gene?

Exercise 4 – Exploring a human region

Use www.ensembl.org

(a) Go to the region from bp 52,600,000 to 53,300,000 on human

chromosome 4. What does the transparent part in the 'Contigs' track represent? (If you are not sure what contigs are, click on Help and then Glossary.)

(b) Zoom in on the SGCB transcript, including a bit of flanking sequence on both sides.

(c) Is there a protein from UniProtKB that aligns to the genome at the same location as SGCB?

(d) Export the genomic sequence of the region you are looking at in FASTA format.

2.2.10 Sea squirt and Diatoms (answers)

Answer 1 - The Sea squirt genomes

Use www.ensembl.org

(a) Two. *Ciona intestinalis* and *Ciona savignyi*.

(b) The genomes were both shotgun-sequenced. You can find this on the species specific **Description** page.

(c) 173,499,994 base pairs. You can find this information for any species in the **Assembly and Genebuild** page.

(d) The Ensembl identifier is ENSCING00000001363 and the gene type is classified as "known protein coding". You can see this by clicking on the transcript.

(e) Follow the link to ENSCINT00000002640 to open transcript tab. Click **Ontology Table**. The protein is involved in nucleic acid binding and transcription regulator activity. It is localized in the nucleus.

(f) Click **Domains and features**. Here you can see that Homeobox and Antennapedia domains are annotated in addition to HTH_lambdarepressr.

(g) The presence of a homeobox domain suggests that the protein belongs to HOX transcription regulators. Indeed the name of this gene is HOX 5. You can see it in the top left corner. If you want to find out more on homeodomains click on the InterPro **Homeobox** link "**IPR001356**" and read through the **Abstract**

(h) There is an ortholog ENSCSAVG00000004156 in *Ciona savignyi*. You can see this in the Gene tab. Click **Gene Tree (image) or Orthologues**. Alternatively, Click on **Location** and **Multi Species** view. Here you can see

the alignment with *Ciona savignyi*. The sequences are more than 60% identical.

Answer 2 - The Diatoms

Use <http://protists.ensembl.org>

- (a) Two. *Thalassiosira pseudomonana* and *Phaeodactylum tricornutum*.
- (b) 33. You can find it in a **Genome tab** under **Karyotype**.
- (c) When you look at **Assembly and Genebuild** you can see that the genome of *Phaeodactylum tricornutum* is shorter than the genome of *Thalassiosira pseudomonana* and contains fewer protein coding genes.
- (d) EC 1.18.1.2.
- (e) It contains two exons. In comparison in *Thalassiosira pseudomonana* the first exon is missing. To see this click **Genomic alignments** under “Protist compara”. In the Alignment view select *Thalassiosira pseudomonana*.
- (f) Yes. In order to see **Orthologs** or a **Gene Tree** for all species you need to look under “Pan-taxonomic Compara”.
- (g) Yes. Two ESTs are shown in the “Iron limited EST” track.

3 BioMart

3.1 Mining data- worked example

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

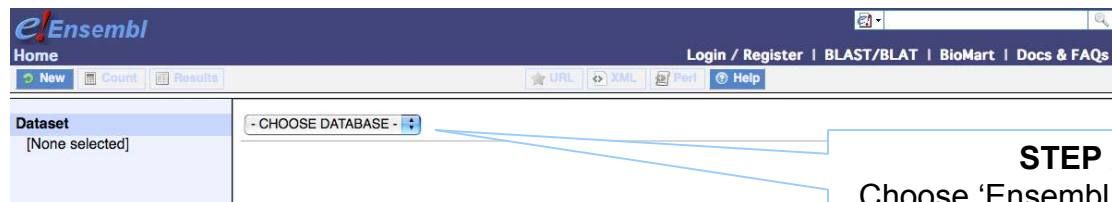
Which other genes with consensus coding sequences assigned by the CCDS project locate to the same band? What are their Ensembl Gene IDs and Entrez Gene IDs?

What are their cDNA sequences?

Follow the worked example below to answer these questions.

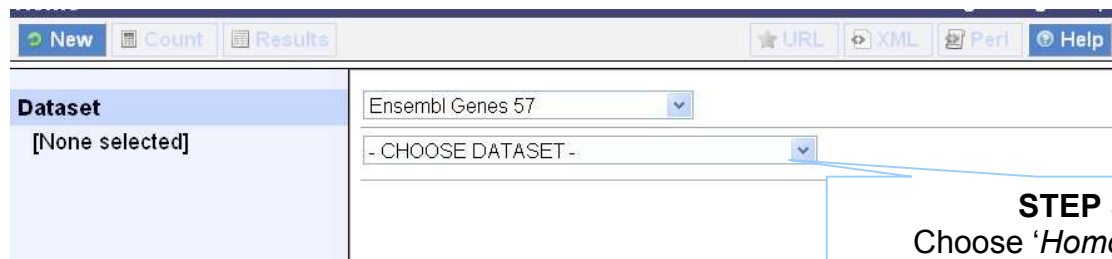
Step 1: Either click on 'BioMart' in the top header of an Ensembl page, or go to <http://www.biomart.org/> and click on the 'MartView' tab.

NOTE: These answers were determined using Ensembl 60



The screenshot shows the Ensembl BioMart interface. The 'Dataset' section on the left indicates '[None selected]'. The main area has a dropdown menu labeled '- CHOOSE DATABASE -'. A blue arrow points from this dropdown to a text box on the right.

STEP 2:
Choose 'Ensembl Genes 60 as the primary database.



The screenshot shows the Ensembl BioMart interface with 'Ensembl Genes 57' selected in the 'Dataset' dropdown. Below it, there is another dropdown menu labeled '- CHOOSE DATASET -'. A blue arrow points from this dropdown to a text box on the right.

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

Ensembl
Home
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New Count Results URL XML Perl Help

Dataset: Homo sapiens genes (NCBI36)
Filters: [None selected]
Attributes: Ensembl Gene ID, Ensembl Transcript ID
Dataset: [None Selected]

Please restrict your query using criteria below

REGION: [collapsed]
GENE: [collapsed]
EXPRESSION: [collapsed]
MULTI SPECIES COMPARISONS: [collapsed]
PROTEIN DOMAINS: [collapsed]
VARIATIONS: [collapsed]

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

Ensembl
Home
Login / Register | BLAST/BLAT | BioMart | Docs & FAQs

New Count Results URL XML Perl Help

Dataset: Homo sapiens genes (NCBI36)
Filters: Chromosome: X, Band Start: q28, Band End: q28
Attributes: Ensembl Gene ID, Ensembl Transcript ID
Dataset: [None Selected]

Please restrict your query using criteria below

REGION: [expanded]
☒ Chromosome: X
☐ Base pair: Gene Start (bp): 1, Gene End (bp): 10000000
☒ Band: Band Start: q28, Band End: q28
☐ Marker: Marker Start: , Marker End:
☐ Encode type: manual_picks
☐ Encode region: 5:131284314:132284313
☐ Multiple Chromosomal Regions (Chr:Start:End:Strand):
 Chromosome Regions:
 Browse...

GENE: [expanded]
EXPRESSION: [expanded]
MULTI SPECIES COMPARISONS: [expanded]
PROTEIN DOMAINS: [expanded]

STEP 5:
Select 'Chromosome X'

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

STEP 7:
Expand the 'GENE' panel.

GENE:

☒ Limit to genes ... with CCDS ID(s)
☒ Only
☐ Excluded

☐ ID list limit Ensembl Gene ID(s)
 Browse...

☐ Transcript count >=

☐ Gene type
 IG_C_gene
 IG_D_gene
 IG_J_gene
 IG_V_gene
 miRNA

STEP 8:

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

New Count Results

Dataset 110 / 51737 Genes
 Homo sapiens genes (GRCh37)

Filters

Chromosome: X
 Band Start: q28
 Band End: q28
 with CCDS ID(s): Only

Attributes

REGION:
 GENE:

STEP 9:

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

The 'Count' shows 108 out of 52,580 total human genes passed the filters.

New Count Results URL XML Perl Help

Dataset 110 / 51737 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

Please select columns to be included in the output and hit 'R'

☒ Features ☐ Homologs
☐ Structures ☐ Variations
☐ Transcript Event ☐ Sequences

GENE:
 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** **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

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:
☐ EXTERNAL:
☐ EXPRESSION:
☐ PROTEIN DOMAINS:

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

☐ Variations ☐ Sequences

☐ GENE:
☐ EXTERNAL:
☐ EXPRESSION:
☐ PROTEIN DOMAINS:

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 13:
Expand the 'EXTERNAL' panel.

External References (max 3)

☐ Clone based Ensembl gene name
☐ Clone based Ensembl transcript name
☐ Clone based VEGA gene name
☐ Clone based VEGA transcript name
☐ CCDS ID
☐ EMBL (Genbank) ID
☐ Ensembl Human gene
☒ **EntrezGene ID**
☐ VEGA transcript ID(s) (OTTT)
☐ Ensembl transcript (where OTTT shares)
☐ HAVANA transcript (where ENST shares CDS with)
☐ 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 VDJ 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
☐ 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', which are diseases/phenotypes from NCBI's OMIM project.

New

Count

Results

URL

XML

Perl

Help

Dataset 108 / 52580 Genes

Homo sapiens genes (GRCh37.p2)

Filters

Chromosome: X

Band Start : q28

Band End : q28

with CCDS ID(s):

Attributes

Ensembl Gene ID

Ensembl Transcript ID

Associated Gene Name

Entrez Gene ID

MIM Morbid Accession

MIM Morbid Description

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	Entrez Gene ID	MIM Morbid Accession	MIM Morbid Description
ENSG00000102119	ENST00000369842	EMD	2010	310300	EMERY-DREIFUSS MUSCULAR DYSTROPHY, 1
25	RPL10	6134	209850	AUTISM	
46	RPL10	6134	209850	AUTISM	
17	RPL10	6134	209850	AUTISM	
35	DNASE1L1	1774			
07	DNASE1L1	1774			
08	DNASE1L1	1774			
ENSG00000013563	ENST00000509585	DNASE1L1	1774		
ENSG00000013563	ENST00000393638	DNASE1L1	1774		
ENSG00000013563	ENST00000369809	DNASE1L1	1774		

STEP 15:

Click 'RESULTS' at the top to preview the output.

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

Dataset 108 / 52580 Genes
Homo sapiens genes (GRCh37.p2)

Export: all results to 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
ENSG00000102119	ENST00000369842	EMD	2010	310300	EMERY-DREIFUSS MUSCULAR DYSTROPHY, 1
ENSG00000147403	ENST000004325	RPL10	6134	209850	AUTISM
ENSG00000147403	ENST000004746	RPL10	6134	209850	AUTISM
ENSG00000147403	ENST000004817	RPL10	6134	209850	AUTISM
ENSG00000013563	ENST000004935	DNASE1L1	1774		
ENSG00000013563	ENST000004907	DNASE1L1	1774		
ENSG00000013563	ENST000004908	DNASE1L1	1774		
ENSG00000013563	ENST000004935	DNASE1L1	1774		
ENSG00000013563	ENST000004907	DNASE1L1	1774		
ENSG00000013563	ENST000004908	DNASE1L1	1774		

To save a file of the complete table, click 'Go'. Or, email the results as a compressed file (best for large result sets).

Ensembl Gene ID	Ensembl Transcript ID	Associated Gene Name	EntrezGene ID	MIM Morbid Accession	MIM Morbid Description
ENSG00000102119	ENST00000369842	EMD	2010	310300	EMERY-DREIFUSS MUSCULAR DYSTROPHY, 1
ENSG00000147403	ENST00000424325	RPL10	6134	209850	AUTISM
ENSG00000147403	ENST00000344746	RPL10	6134	209850	AUTISM
ENSG00000147403	ENST00000369817	RPL10	6134	209850	AUTISM
ENSG00000013563	ENST00000014935	DNASE1L1	1774		
ENSG00000013563	ENST00000369807	DNASE1L1	1774		
ENSG00000013563	ENST00000369808	DNASE1L1	1774		
ENSG00000013563	ENST00000309585	DNASE1L1	1774		
ENSG00000013563	ENST00000393638	DNASE1L1	1774		
ENSG00000013563	ENST00000369809	DNASE1L1	1774		
ENSG00000102125	ENST00000350743	TAZ	6901	300069	CARDIOMYOPATHY, DILATED, 3A
ENSG00000102125	ENST00000350743	TAZ	6901	300183	LEFT VENTRICULAR NONCOMPACTION, X-LINKED
ENSG00000102125	ENST00000350743	TAZ	6901	302060	BARTH SYNDROME
ENSG00000102125	ENST00000299328	TAZ	6901	300069	CARDIOMYOPATHY, DILATED, 3A
ENSG00000102125	ENST00000299328	TAZ	6901	300183	LEFT VENTRICULAR NONCOMPACTION, X-LINKED
ENSG00000102125	ENST00000299328	TAZ	6901	302060	BARTH SYNDROME
ENSG00000102125	ENST00000351413	TAZ	6901	300069	CARDIOMYOPATHY, DILATED, 3A

Result Table 1

Ensembl Home Login / Register | BLAST/BLAT

[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

Protein
Ensembl Gene
Ensembl Transc

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, click on 'Attributes'

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

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
cDNA sequences

Dataset
[None Selected]

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

☐ Features ☐ Transcript Event
☐ Structures ☐ Homologs
☐ Variations ☒ Sequences

☐ SEQUENCES:

Sequences (max 1)

-----☐-----
-----☐-----

☐ Unspliced (Transcript)
☐ Unspliced (Gene)
☐ Flank (Transcript)
☐ Flank (Gene)
☐ Flank-coding region (Transcript)
☐ Flank-coding region (Gene)

☐ 5' UTR
☐ 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'.

☐ Header Information

Gene Information
☒ Ensembl Gene Name
☐ Description
☒ Associated Gene Name
☐ Associated Gene ID

Transcript Information
☐ CDS Length
☐ CDS Start
☐ CDS End
☐ 5' UTR Start
☐ 5' UTR End
☐ 3' UTR Start

Exon Information
☐ Ensembl Exon ID
☐ Exon Chr Start (bp)
☐ Exon Chr End (bp)

☒ Chromosome Name
☐ Gene Start (bp)
☐ Gene End (bp)
☐ Ensembl Protein ID (s)

☐ 3' UTR
☒ Ensembl Transcript ID
☐ Ensembl Transcript Name
☐ Strand
☐ Transcript Start (bp)
☐ Transcript End (bp)

☐ Strand
☐ Exon Rank in Transcript
☐ Constitutive Exon

STEP 20:
Expand the 'Header
Information' section.

STEP 21:
Customise the FASTA header.
Choose 'Associated Gene
Name' and 'Chromosome
Name', in the **Gene information**
section.

New **Count** **Results** URL XML Perl Help

Dataset 105 / 49506 Genes
Homo sapiens genes (GRCh37)

Filters
Chromosome: X
Band Start: q28
Band End: q28
with CCDS ID(s): Only

Attributes
Ensembl Gene ID
Ensembl Transcript ID
cDNA sequences
Associated Gene Name
Chromosome Name

Dataset
[None Selected]

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

☐ Features ☐ Trans
☐ Structures ☐ Homo
☐ Variations ☒ Sequ

☐ 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'

[illegible]

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

>ENS000000013619|ENST00000262858|MAMLD1|X
AAGCCCTGTGTCAGTGCCTTTGGGAAACGCCCTGGAGAGTCAAGAAATAAAATTTGCAGGT
CAAAACCATGGATGATCGGAAAAGTCGGCTTGTAATCAAGAGCATGCTTCCCCATTTCCGT
ATGGTGGGAAATCTCTCAGGAGCCCAAGAAAGCTCCAGGAATCTGGGAAAGAGCCCTCTGTGG
ATGGAGGAAGAAGATTATCTTTTCTCTACAAGAGCAGCCAGGAAGAAGCATCAGGGA
ACTGTTAAGAGGAGACAAGAGAAGACCACTTCCAGTTTCCAGACATGGCTGATGGGGGC
TACCTTAATAAAAAATCAAGAGGCTTGCTTGAAGATGTCACTTTCGAATGGGCCCAAGGT
GCTCATCTAGTACTGCTTGTGAGACAAGTGCAGGCTCCCTCCATTGACAATAAAATCCTGAG
CCTGCGGCTATGGGAGTGGCTGGCCAGTCATTACTGCTGGAGAATAACCCATATGAATGGC
AACATCATGTGGGCTCACCAATTTGTAGTACCAACGCTGCAGTACGTAAGTGGGACATGAAGGGCCC
ACTGTTCTTACTATGAGAAAAATCAACAGCGTGCCGGCTGTAGACACGAGGAGTTTCAAGAG
CTGTAGAGGAGCTCACAAAAATCAAGACCTTCTCCAATGAGCTAGATCTTGAGAGAG
ATACTGGGACGAAGCCAGAAGAGCCACTGGTTTTAGATCATCCCCAGGCAACCCTAAGC
ACAACCTCCCAAGCCTTCGGTTCAGATGTACACTTGGAGAGCCTGGCTTCCAGCAAGGAG
TTTGCTTCTAGTTGACGCCAAGTTACTGGCATGTCACTTCAGATCCCATCCTCTCCACA

cDNA 1

ATAGAAAAACCCACCCACTGTCCTGTAACCTTTTCTCAGTGTCCAGACITTTCTGTAATC
ACATTTTAATTGCCACCTCGTATTTCACTTCTACATTTGAAATCTGGCGTCTGTTTCAAG
CCAGTGTGTGTTTTTCTTCTGTAATAAACACCGAGGAAAAAGTG
>ENS00000166008|ENST00000298974|MAGEA9|X
GTGCGCACTGGGGGTGAGAGAGAGGGAGAGGCTCCTTCTGAGGGGCGGCTTGATACCG
GTGGAGGAGCTCAGGAAAGCAGGCAAGCCTTGCTGTGAGACAGTGTCTCTCAGCGAGA
GCAGAGGAGACCCAGGCAGTGTGACGAGTGAAGGTTCTGGGACAGGCTAACCGAGGA
CAGGAGCCCCAAGAGGCCCCAGACAGCACTGACGAAGACCTGCTGTGGGTCTCCATCG
CCCAGCTCCTGCCACGCTCCTGACTGTGCTGCCGTGACCAGAGTCATCATGTCTCTCGAGC
AGAGAGGCTCCGCACTGCAAGCCTGATTGAAGCACTTGAGAGCCCAAGGAGGAGCACTTGGGCC
TGATGGGTGACACAGGAACCCACAGGCGAGGAGGAGCACTACCTCTCTCTGCACAGA
AGGAGGAGGAGGTGTCTGCTGTGGGTCACTAAGTCTCCCCAGAGTCTCTCAGGAGGCG
CTTCTCTCTCATTTTCCGTCTACTACACTTTATGGAGCCAAATTCGATGAGGGCTCCAGCA
GTCAGAAAGAGGAGGAGCAAGCTCCTCGGTGCAAGCCAGCACTCAGCTGAGGATTCATGTTC
AAGAAGCACTGAAATTTGAAGGTGGCTGATGTGTGTTCTCTCTGCTGCACAAATATCGAC
TCAAGGAGCCGGTGCACAAAGGCAGAAATCTGGAGAGCGTCATCAAAAATTCAGAGCGCT

cDNA 2

3.1.1 BIOMART (exercises)

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

1. Retrieve all SNPs for '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 Genes 60'
- Select 'Homo sapiens genes (GRCh37.p2)'.

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 '**GERM LINE 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? Because not all variations are within 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/CNIO_nov2010/coursebook_60_v.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

List all variations on human chromosome 1 between nucleotides 800,000 and 850,000. Export genes associated with these variations (if any).

3.1.2 BIOMART (answers)

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. **HINT**, if BioMart is not clearing, click on the "**BioMart**" link at the top of the page rather than "New".

Start with all the zebrafish Ensembl genes:

Choose the '**ENSEMBL Genes 60**' database.
Choose the '**Danio rerio genes (Zv9)**' dataset.

Now filter for the genes on chromosome 3:

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 1268 / 28491 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 available.

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

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

Select '**View All rows as HTML**' or export all results to a file.

3. Click [New].

Choose the '**ENSEMBL Genes 60**' 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).

HINT: You may have to scroll down the menu to see these.

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 Genes 60**' 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 **31** genes. There should be one gene (ENSG00000006611) with the 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. The gene may not be in the first 10 rows, so view more rows to see it.

5. (a) Click **[New]**.

Choose the '**ENSEMBL Genes 60**' 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 **21** out of the 23 human genes at least one mouse orthologue has been identified. ENSG00000123130 has two mouse orthologues and ENSG00000172716 has three. Two human genes (ENSG00000186594 and ENSG00000130844) have none.

6. Step 1: Choose ‘*Ensembl Genes 60*’ 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 60 genes*’, ‘*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. In this exercise, select “Ensembl Variation 60” as the Database, and “Homo sapiens Variation” as the Dataset. This allows intergenic variations to be chosen. You would also use the Variation Mart if you had a list of variation IDs to input in the filters.

Filter with REGION: Chromosome 1

Base pair start: 800000

end: 850000

Click ‘Count’. Note this counts variations, not genes. You should see 897.

To the **Attributes**, add “Ensembl Gene ID”, “Ensembl Transcript ID”, and “Consequence to transcript” under the GENE ASSOCIATED INFORMATION panel.

4 Exercises for day II

4.1.1 COMPARATIVE GENOMICS (exercises)

Exercise 1 - Orthologues, paralogues and genetrees

Find the human Ensembl SMAD2 gene.

(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 SMAD2?

Retrieve an alignment between SMAD2 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 SMAD2 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) Which paralogue has the most sequence similarity with the red-sensitive opsin? (Click on the Target or Query % id column to order the choices from higher percentage to lower).

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

(d) If you were browsing with the OPN1SW gene, how could you find out it is a “blue-sensitive opsin”?

Exercise 3 – Multi-species Alignments

Find the Ensembl F9 gene for human and go to the **Region in detail** page (in the Location tab).

(a) Turn on the cow and mouse BLASTZ alignment tracks along with the frog (*Xenopus tropicalis*) Translated BLAT alignment track. BLASTZ is used to compare two species on the nucleotide level. In which species do you see the most aligned sequence?

(b) Turn on the three tracks for the multispecies 34 eutherian mammals alignments. These are found in “configure this page”, “Multiple alignments”. You can read more about these conservation scores and constrained elements in the “Comparative Genomics” documentation (under the “Documentation” link at the top of the page). How do these tracks differ from the pairwise alignments you looked at in (a)?

4.1.2 COMPARATIVE GENOMICS (answers)

Answer 1

Under '**Search Ensembl**' type '**human gene SMAD2**' and click **[Go]**. On the page with search results click on 'Ensembl protein_coding Gene: **ENSG00000175387** (HGNC (curated): SMAD2)'.

(a) Click on '**Paralogues**' in the side menu.

There are six within-species paralogues predicted for human SMAD2. The first one (SMAD3) 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 the "Alignment" link for a paralogue.

(b) Click on '**Orthologues**' in the side menu.

Yes, there is an orthologue predicted for human SMAD2 in gorilla: ENSGGOG00000004706 (SMAD2).

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

The gene of interest is in red, and paralogues should be shown in blue.

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

SMAD3 is related by a duplication on the level of the "Euteleostomi (Bony vertebrates)". Click on the common ancestor (red node) to see this.

(d) Click on the duplication node (red square) or speciation node (blue square) of any 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.

ENSG00000147380 (OPN1MW) and ENSG00000166160 (OPN1MW2), the genes encoding the green-sensitive (middle-wave) opsins, have the highest Target %id and Query %id.

(c) Look at the Location of '**ENSG00000166160**', '**ENSG00000147380**' and '**ENSG00000128617**'. These will be listed in the Parologue table.

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 (OPN1SW) 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.

(d) To find gene (protein) functions, go to the transcript tab, "General identifiers" section. In the UniProt and RefSeq descriptions, blue-light sensitivity is listed.

Answer Exercise 3

Under '**Search Ensembl**' type '**human gene F9**' and click **[Go]**.

On the page with search results click on '**Ensembl protein_coding Gene: ENSG00000101981 (HGNC symbol: F9)**'. (*Hint: if you click directly on "Region in Detail" from the search results, you don't have to go through the Gene tab.*)

Click on the **Location tab**.

Click on '**Configure this page**' in the side menu

Click on '**BLASTz/LASTz alignments**', select '**Cow (*Bos Taurus*)**' and '**Mouse (*Mus musculus*)**', click on '**Translated BLAT**' alignments, select '***Xenopus tropicalis***' and close the menu.

(a) Species that are closer to human in evolution show a larger extent of conservation. The cow and mouse alignments show high 'coverage', meaning that alignments were found to a large percent of the human genome sequence. The *Xenopus* (frog) alignment matches to some exons of F9. This reflects the Translated BLAT approach, where the genome is translated and then aligned. Alignments are focused on protein sequences. Of course, the quality of the alignment reflects the quality of the genome sequence.

(b) The pink bar corresponding to the “34 eutherian mammal” alignment shows the full region aligned across 34 mammals. Click on the track name (34 eutherian mammals) to see the mammals used in this alignment. The ‘Conservation score’ (34 way GERP score) is calculated from this multispecies alignment. The score shows how well each nucleotide is conserved across the 34 species. A positive score reflects a high conservation. The individual scores can be seen in the plot. Above the plot are the ‘constrained elements’. These show sequences in the alignment with high GERP scores. (High GERP scores signify conserved regions.) Many of the conserved elements match up to exons, but there are also constrained elements in introns. These could be important functional sequences.

In summary, the blocks in the 34 way GERP elements track are the most conserved parts of the genome, when comparing 34 mammals. Keep in mind, this alignment includes the low-coverage (low quality) genome sequences.

4.1.3 GENEBUILD (exercises)

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 (Havana) annotated for this gene?
- (c)** How many transcripts agree between Ensembl and Havana annotation?

(d) What is the reason that Ensembl hasn't found one of the Havana transcripts?

Exercise 4

An example of what can go wrong

Go to the following page in Ensembl release 46 (of August 2007):

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

(a) What is wrong with this gene? What could be the reason for this?

(b) Has this problem been fixed in Ensembl release 60?

4.1.4 GENEBUILD (answers)

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. **21,077 known** and **521 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 Symbol: GALP)'.

(a) From the HUGO Gene Nomenclature Committee (HGNC). Click on the hyperlinked 'GALP' after 'Name' to see the HUGO gene card.

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

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**' type '**mouse gene Epc1**' and click **[Go]**.

On the page with search results click on 'Ensembl protein_coding Gene: **ENSMUSG00000024240** (MGI Symbol: Epc1)'.

(a) **Four**.

(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** (these are the 'gold' transcripts.)

(d) Click on the Epc-004 transcript in the figure.

Click on 'ENSMUST00000124926' in the pop-up menu.

Click on 'Supporting evidence' in the side menu.

In this case the reason is that the transcript ENSMUST00000124926 is built on EST evidence and short cDNA fragments. As Ensembl doesn't build on just EST evidence, it hasn't annotated this transcript in the automatic pipeline, but displays it as a Havana transcript.

Answer 4

Go to

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

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

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

Under 'Search Ensembl' type 'human gene CTNND1' and click [Go].

On the page with search results click on 'Ensembl protein_coding Gene: ENSG00000198561 (HGNC Symbol: 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 CTNND1 transcripts and the TMX2 (formerly TXNDC14) transcripts are shown.

In Ensembl release 60, CTNND1 and TMX2 are annotated as separate genes.

5 Quick Guide to Databases and Projects

Here is a list of databases and projects you will come across in these exercises. Google any one of these to learn more. Projects include many species, unless otherwise noted.

SEQUENCES

EMBL-Bank, NCBI GenBank, DDBJ – Contain nucleic acid sequences deposited by submitters such as wet-lab biologists and gene sequencing projects. These three databases are synchronised with each other every day, so the same sequences should be found in each.

CCDS – coding sequences that are agreed upon by Ensembl, VEGA-Havana, UCSC, and NCBI. (*Human and mouse*).

NCBI Entrez Gene – NCBI's gene collection

NCBI RefSeq – NCBI's collection of 'reference sequences', includes genomic DNA, transcripts and proteins.

UniProtKB – the "Protein knowledgebase", a comprehensive set of protein sequences. Divided into two parts: Swiss-Prot and TrEMBL

UniProt Swiss-Prot – the manually annotated, reviewed protein sequences in the UniProtKB. High quality.

UniProt TrEMBL – the automatically annotated, unreviewed set of proteins (EMBL-Bank translated). Varying quality.

VEGA – Vertebrate Genome Annotation, a selection of manually-curated genes, transcripts, and proteins. (*Human, Mouse, Zebrafish, Gorilla, Wallaby, Pig, and Dog*).

VEGA-HAVANA – The main contributor to the VEGA project, located at the Wellcome Trust Sanger Institute, Hinxton, UK.

GENE NAMES

HGNC – HUGO Gene Nomenclature Committee, a project assigning a unique and meaningful name and symbol to every human gene. (*Human*).

ZFIN – The Zebrafish Model Organism Database. Gene names are only one part of this project. (*Z-fish*).

PROTEIN SIGNATURES

InterPro – A collection of domains, motifs, and other protein signatures. Protein signature records are extensive, and combine information from individual projects such as UniProt, along with other databases such as SMART, PFAM and PROSITE (explained below).

PFAM – A collection of protein families

PROSITE – A collection of protein domains, families, and functional sites.

SMART – A collection of evolutionarily conserved protein domains.

OTHER PROJECTS

NCBI dbSNP – A collection of sequence polymorphisms; mainly single nucleotide polymorphisms, along with insertion-deletions.

NCBI OMIM – Online Mendelian Inheritance in Man – a resource showing phenotypes and diseases related to genes (*human*).