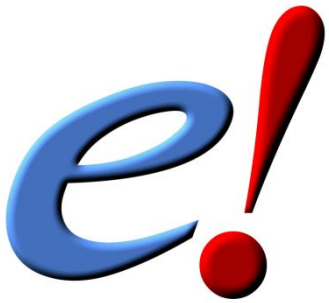
Browsing Genomes with Ensembl



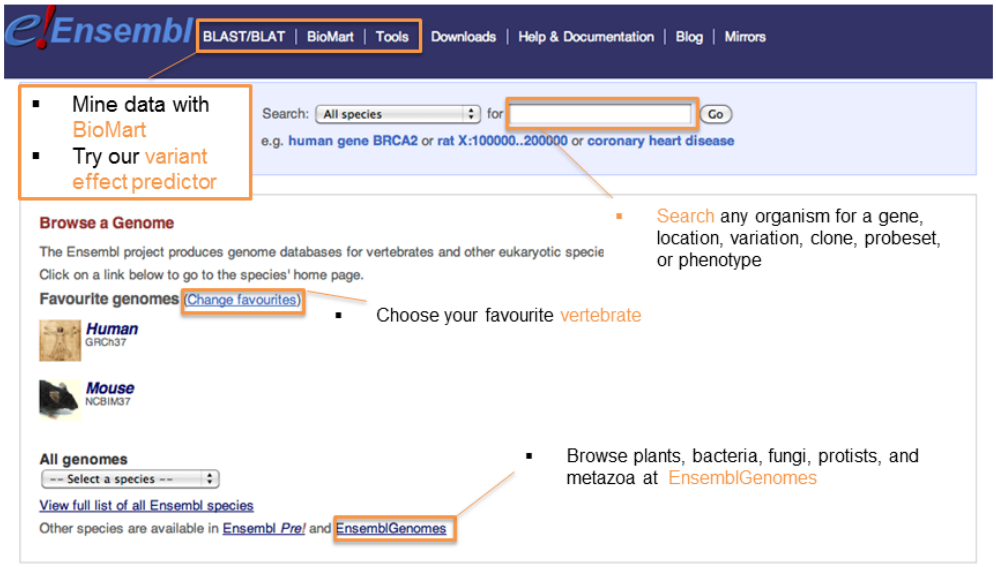
www.ensembl.org
www.ensemblgenomes.org

Course Manual
V64 Oct/Nov, 2011

TABLE OF CONTENTS

Getting started with Ensembl
www.ensembl.org

Ensembl provides genes and other **annotation** such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in **UniProtKB** and **NCBI RefSeq** databases, along with manual annotation from the **VEGA/Havana** group. All the data are freely available and can be accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interface (**Perl API**). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the **BioMart** web interface, which can extract information from the Ensembl databases without the need for programming knowledge by the user!



You will learn about

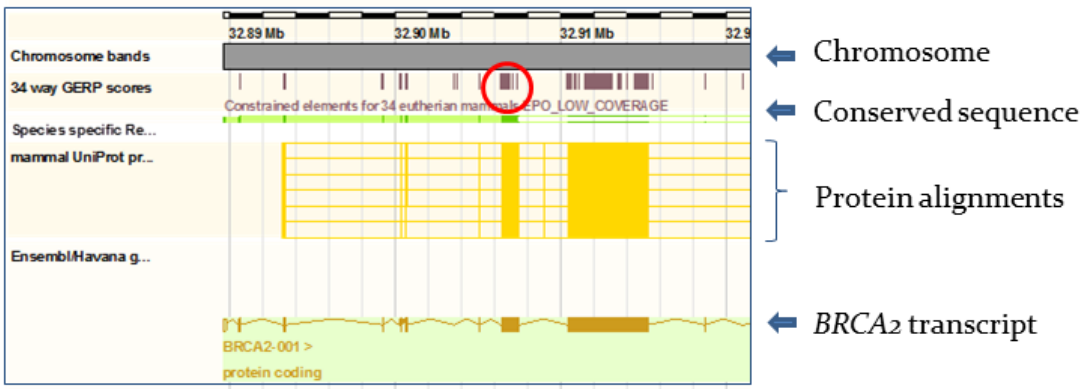
- Why do we need genome browsers?
- An introduction to Ensembl
- How information can be obtained from the site
- An overview of Ensembl tools

Tired of reading?
Check our video tutorial instead!
http://www.youtube.com/user/EnsemblHelpdesk

The Ensembl Genome browser
Introduction to BioMart

Introduction to Ensembl

Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute that annotates **chordate** genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'compara' team. Most annotation is updated every two months, leading to increasing Ensembl versions (such as version 64), however the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non-chordates, namely bacteria, plants, fungi, metazoa, and protists.



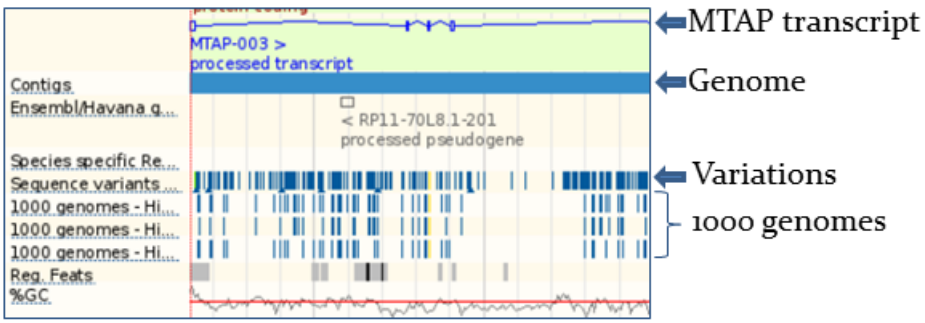
The Region in Detail view

The vast amount of information associated with the genomic sequence demands a way to organise and access that information. This is where genome browsers come in. Ensembl strives to display many layers of genome annotation into a simplified view for the ease of the user. The picture above shows the '**Region in Detail**' page for the BRCA2 gene in human. The example shows blocks of conserved sequence reflecting conservation scores of sequence identity on a base pair level across 35 species. Conserved regions are displayed as dark blocks that represent local regions of alignment. One of the blocks is circled in red. You would only have to click on this block to see more details.

Also in this figure are proteins from the UniProtKB aligned to the same genomic region. Filled yellow blocks show where these UniProtKB proteins align to the genome, and gaps in the alignment are shown as empty yellow blocks. Note, in this case, the UniProtKB proteins support most of the exons shown in the Ensembl BRCA2-001 transcript (in gold).

Use the *Configure this page* tool button to add more data tracks to Ensembl views. 

Both **Ensembl** and **Vega (Havana)** transcripts are portrayed as exons (boxes) and introns (connecting lines). In fact, filled boxes show coding sequence, and empty boxes reflect UnTranslated Regions (**UTRs**). This 'Region in Detail' view is useful for comparing Ensembl gene models with current proteins and mRNAs in other databases like **NCBI RefSeq**, **EMBL-Bank**, and, in the example above, UniProtKB. Everything in this view is aligned to the genome.

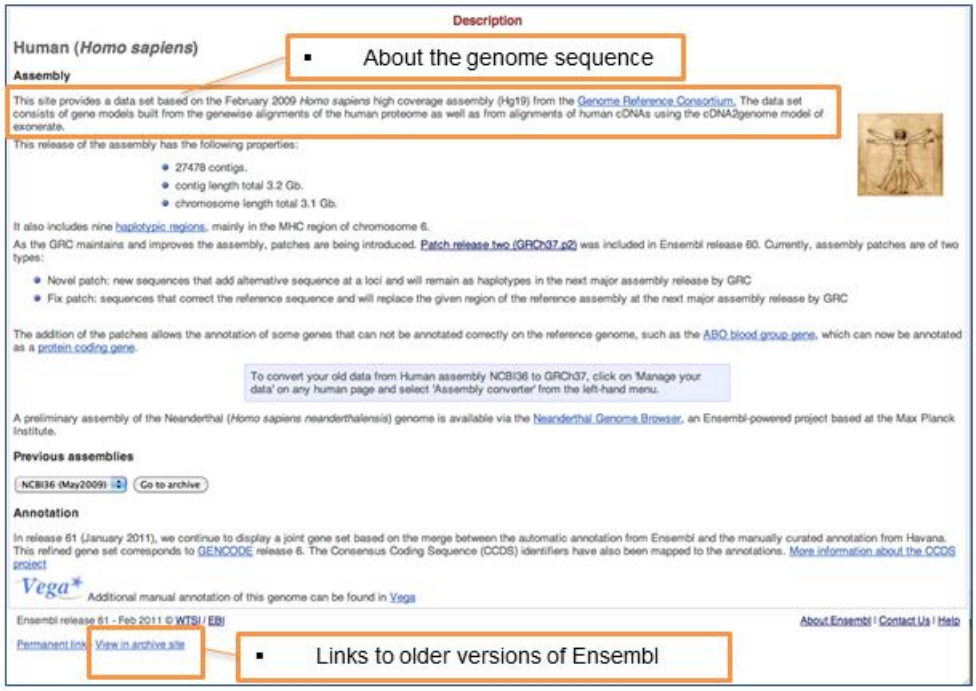


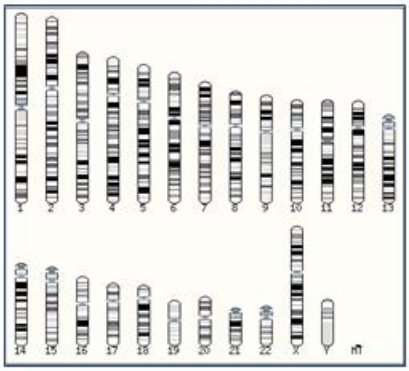
The Region in Detail view: 1000 genomes tracks

The region in detail view can be configured (using the *Configure this page* tool button) to show regulatory features, sequence variation, and more! Click on any vertical line in the variation track for a menu about the SNP (single nucleotide polymorphism) or InDel (insertion deletion mutation). For example, click on "Sequence variants" under the "Germline variation" track and turn on the sequence variants (all sources) at the right side of this page. Save and close. Back to the region in detail view, click on the sequence variation of interest. A pop-up box will show you a few variation properties such as rs number, alleles, type, and others. Click on the

rs properties link to take you to an information page for the genetic variation, including links to population frequencies, if known. You can do the same for regulatory features as well.

An index page is provided for each species with information about the source of the genomic sequence assembly, a karyotype (if available), and a link to past or archive sites. The picture below shows the Ensembl homepage for human. Links to the human karyotype, a summary of gene and genome information, and the most common InterPro domains in the genome are found at the left of this index page.



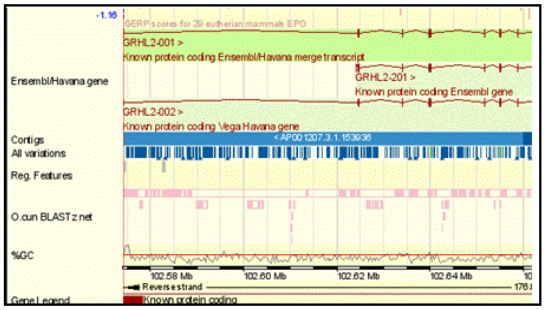


Ensembl devotes separate pages and views in the browser to display a variety of information types, using a tabbed structure

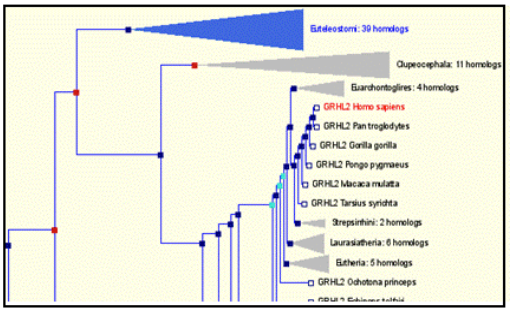
View genotype information in the variation tab, gene trees in the gene tab, a chromosomal region in the location tab, and cDNA sequence alongside the protein translation in the **transcript** pages. Compare conserved regions with the position of genes and population variation in the **Region in Detail** view. See homology relationships in the **gene** page, or perform a **BLAST** or **BLAT** search against any species in Ensembl.

Transcript Sequence w/Variations

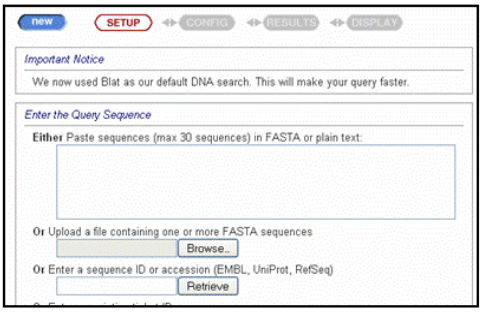
Genes, SNPs, and Conserved Regions



Homologues in Gene Trees



BLAST and BLAT aligners



Retrieving Data from Ensembl

BioMart is a very popular web-interface that can extract information from the Ensembl databases and present the user with a table of information without the need for programming. It can be used to output sequences or tables of genes along with gene positions (chromosome and base pair locations), single nucleotide polymorphisms (SNPs), homologues, and other annotation in HTML, text, or Microsoft Excel format. BioMart can also translate

one type of ID to another, identify genes associated with an **InterPro** domain or gene ontology (**GO**) term, export gene expression data and lots more.

Ensembl uses MySQL relational databases to store its information. A comprehensive set of Application Programme Interfaces (APIs) serve as a middle-layer between underlying database schemes and more specific application programmes. The API aims to encapsulate the database layout by providing efficient high-level access to data tables and isolate applications from data layout changes.

Synopsis- What can I do with Ensembl?

- View genes along with other annotation along the chromosome
- View alternative transcripts (including splice variants) for a gene
- Explore homologues and phylogenetic trees across more than 50 species for any gene
- Compare whole genome alignments and conserved regions across species
- View microarray sequences that match to Ensembl genes
- View ESTs, clones, mRNA and proteins for any chromosomal region
- Examine single nucleotide polymorphisms (SNPs) for a gene or chromosomal region
- View SNPs across strains (rat, mouse), populations (human), or even breeds (dog)
- View positions and sequence of mRNA and protein that align with an Ensembl gene
- Upload your own data
- Use BLAST, or BLAT, a similar sequence alignment search tool, against any Ensembl genome
- Export sequence, or create a table of gene information with BioMart
- Use the Variant Effect Predictor

Need more help?

- Check Ensembl documentation
- Watch video tutorials on YouTube
- View the FAQs
- Try some exercises
- Read some publications
- Go to our online course

Stay in touch!

- Email the team with comments or questions at helpdesk@ensembl.org
- Follow the Ensembl blog
- Sign up to a mailing list

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)

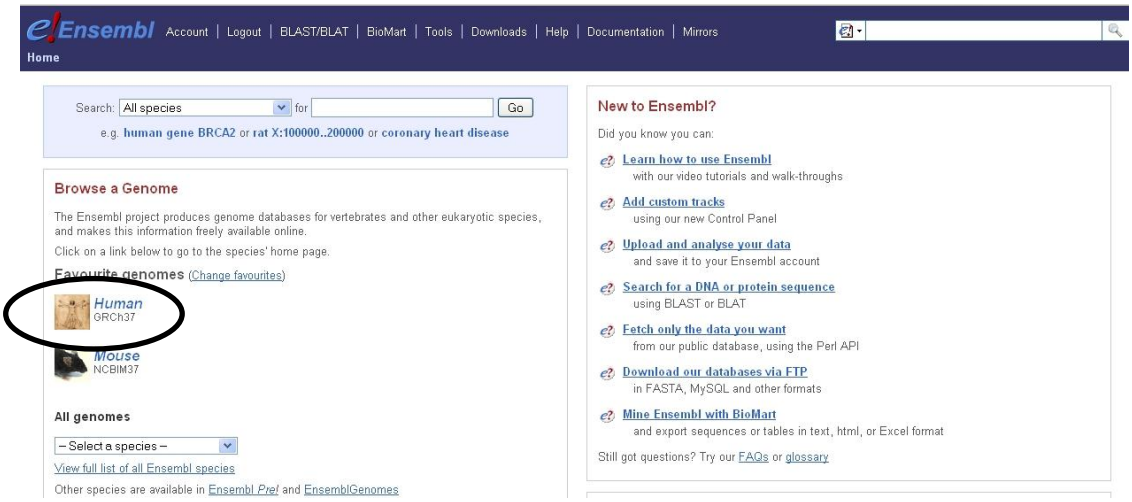
Vilella, A.J. *et. al*
EnsemblCompara GeneTrees: Analysis of complete, duplication aware phylogenetic trees in vertebrates.
Genome Res. 2009 Feb 19(2):327-35

II) WALKING THROUGH THE WEBSITE

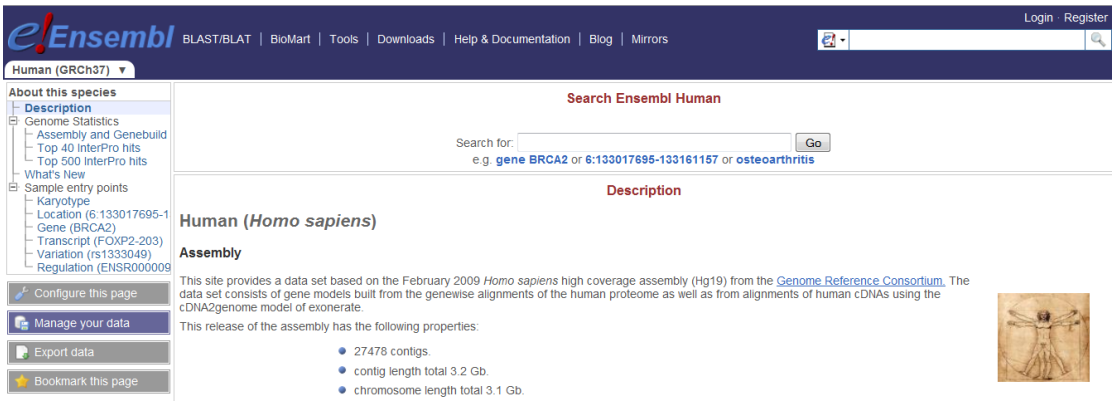
The instructor will guide you through the website using the human and mouse ***ESPN*** genes. The following points will be addressed:

- **The Gene Summary tab and gene-related links:**
  - 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?
- **The Location tab and related links:**
  - How do I zoom in and change the gene focus.
  - Adding a track (e.g. KO alleles)
- **Exporting a sequence and running BLAT/BLAST**

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

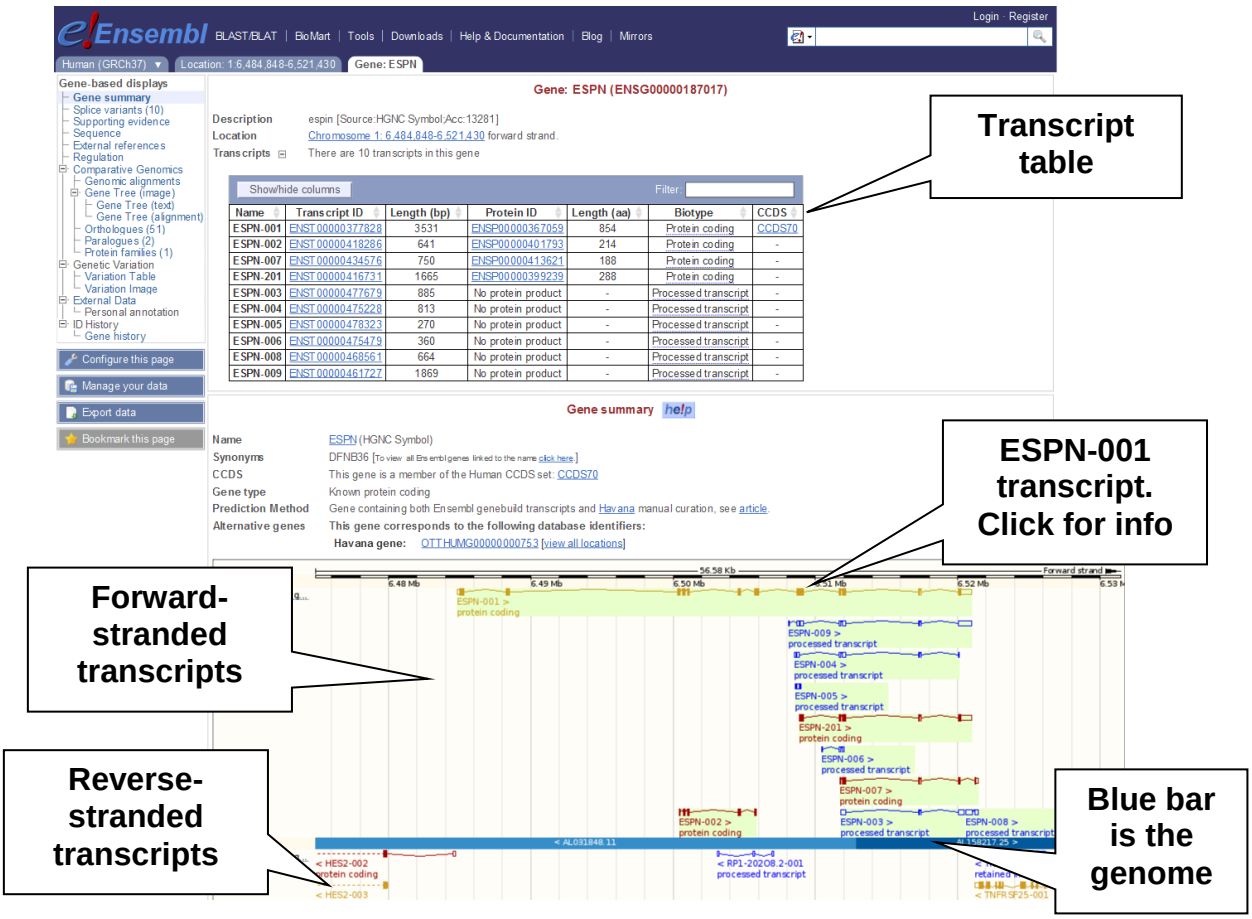


Click on the human icon to open the human home page.

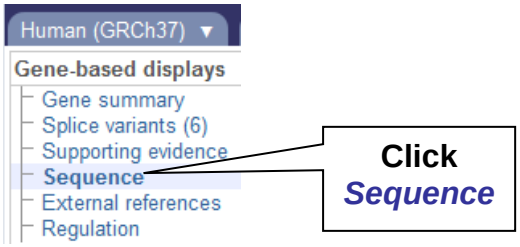


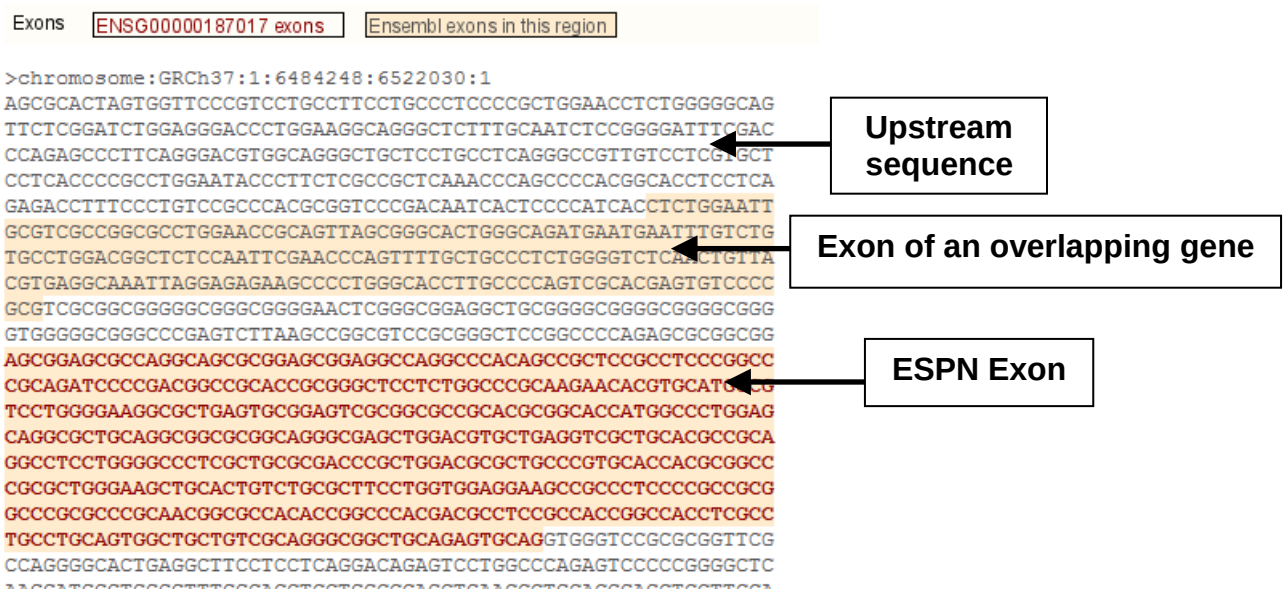
Type 'gene ESPN' into the search bar and click the 'Go' button. Click on 'Gene' and '*Homo sapiens*' to find the hits.

Click on it. The following 'Gene' tab should open:

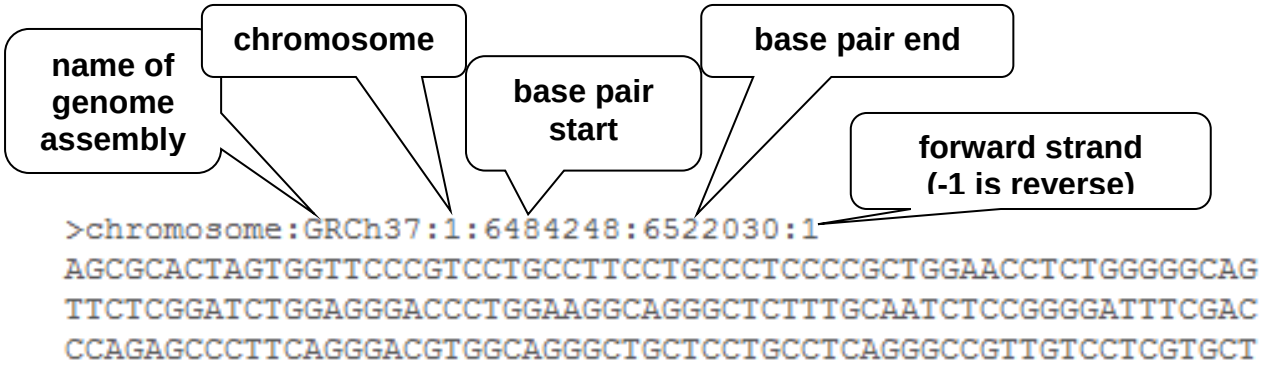


Let's walk through some of the links in the left hand navigation column. How can we view the genomic sequence? Click *Sequence* at the left of the page.



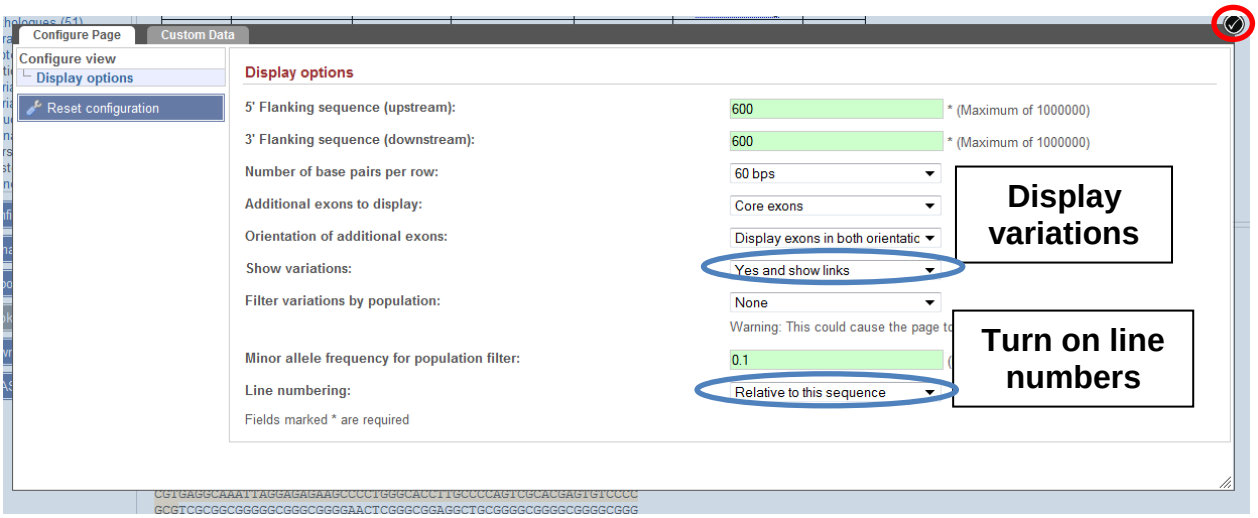


The sequence is shown in FASTA format. Take a look at the FASTA header:



Exons are highlighted within the genomic sequence. Variations can be added with the 

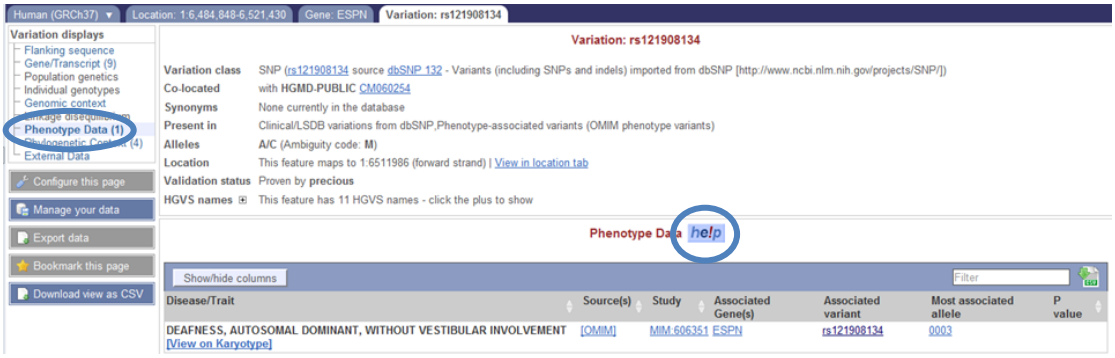
link found at the left. Click on it now.



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

Click the variation (highlighted 'A' in the sequence) for a popup box of information. Follow the link to rs121908134 or click on the link 27739: rs121908134 at the right of the view to open the variation tab.

Click on 'Phenotype Data' at the left of the variation tab.

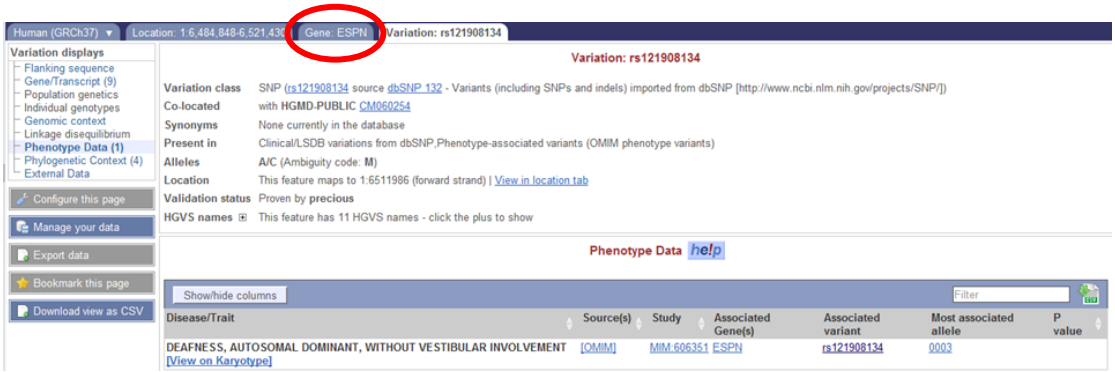


This variation has been associated with Autosomal Dominant deafness by OMIM (Online Mendelian Inheritance in Man).

Click on the blue 'Help' button to view page-specific help.

The help pages provide links to Frequently Asked Questions (FAQs), a glossary, video tutorials, and a form to contact our helpdesk.

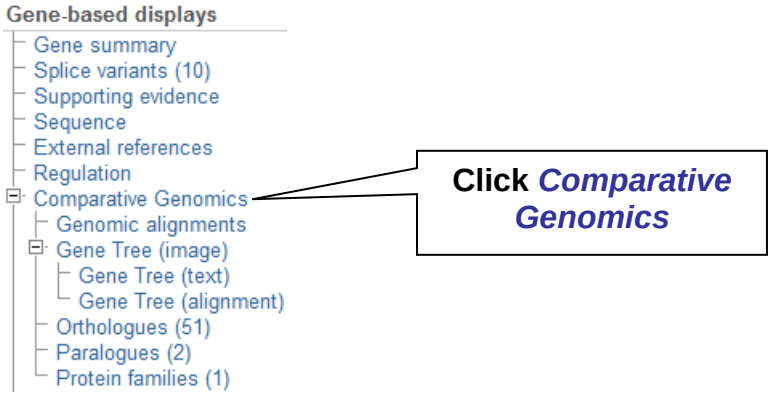
Now let's go back to the Gene tab. Click on the gene tab circled in red in the screenshot below.



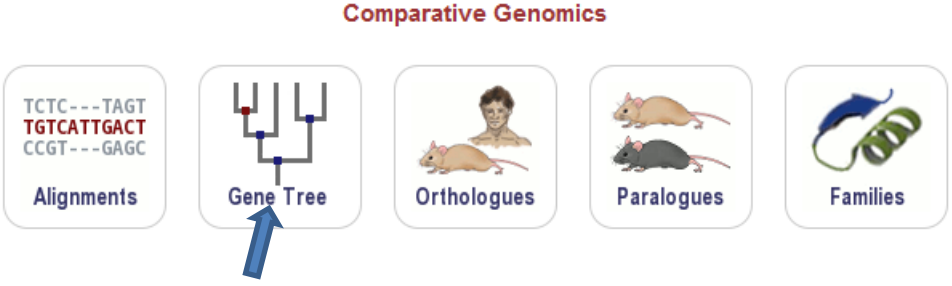
To view all the sequence variations in this locus, click the *Variation Table* link at the left of the gene tab.

The table is divided into consequence types. Click on 'Show' to expand a detailed table for any of the consequence types available.

Comparative genomics offers a new type of browsing menu! Let's try it. Click on Comparative Genomics in the left hand menu of the gene tab.



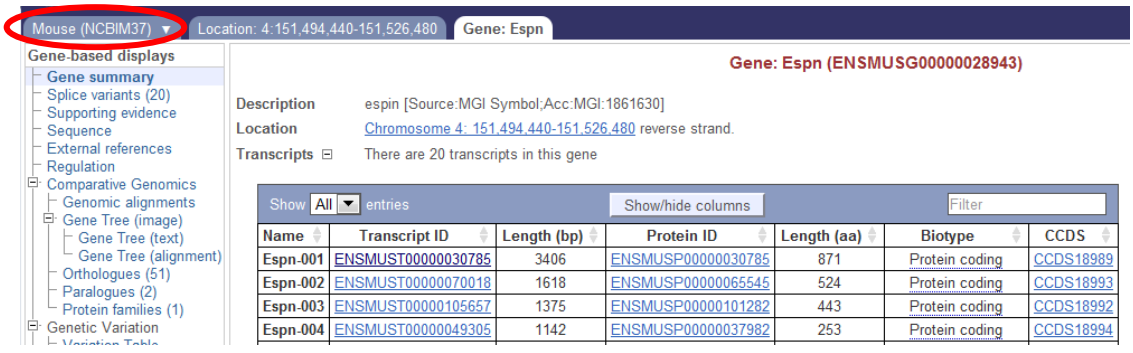
A menu of icons should appear:



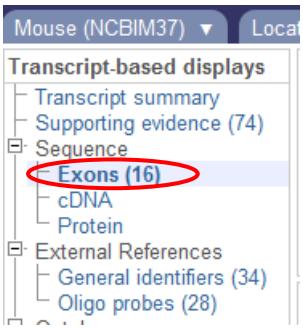
Click on *Gene tree,* which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.

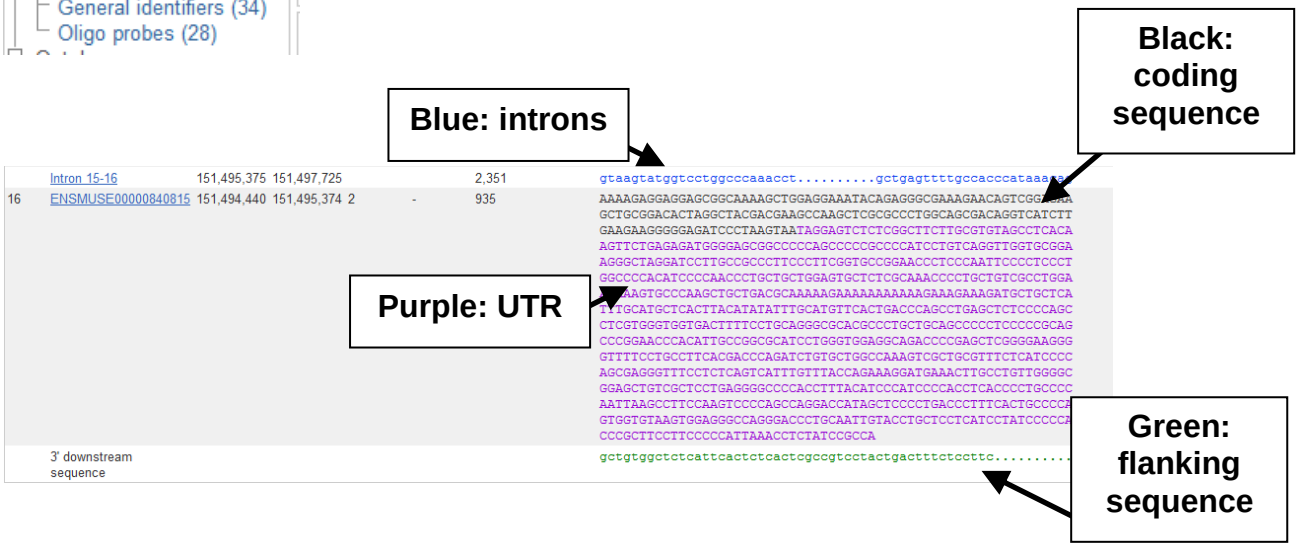
Click the *Orthologues* link at the left of this page (or from the Comparative genomics icons) to view homologues detected in this tree. Find the Mouse orthologue, and click the ENSMUSG ID.

We are now in the mouse pages, as indicated at the left. Espn in mouse has many splice isoforms! Let's click on the Transcript ID for the first one, ENSMUST00000030785.



The left hand navigation column provides several options for the transcript Espn-001. Click on the *Exons* link, circled in the image below.



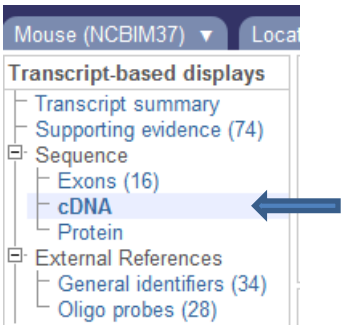


You may use the 

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 

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.

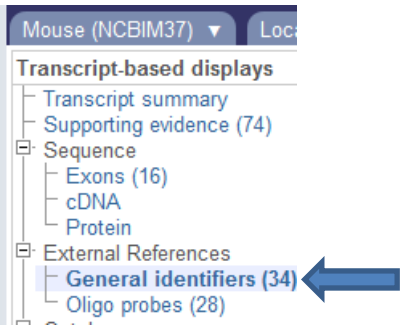


A non-synonymous SNP changes the amino acid sequence.

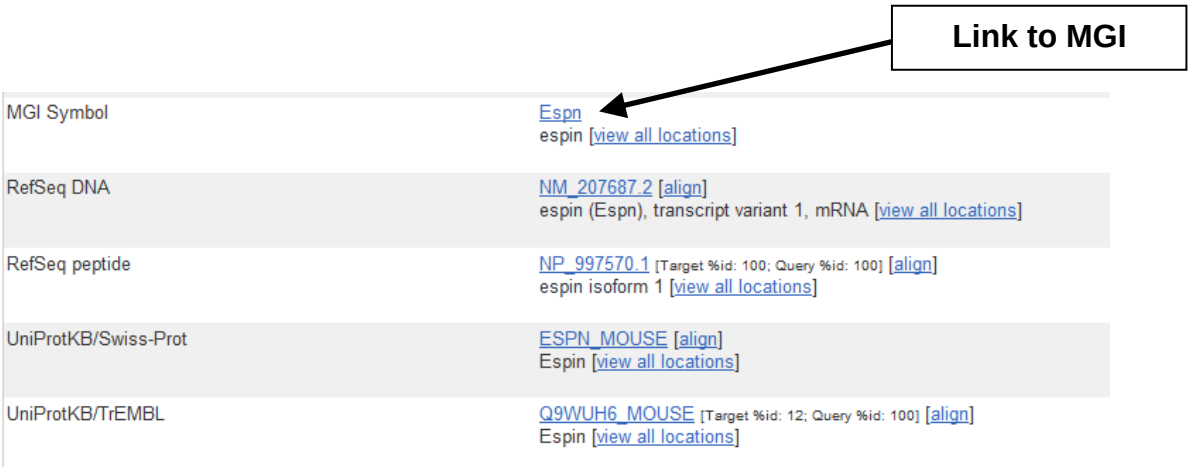
UnTranslated Region (UTR) is highlighted in dark yellow, codons are highlighted in light yellow, and exon sequence is shown in black or blue letters to show exon divides.

Sequence variants are represented by highlighted nucleotides, and clickable IUPAC codes above the sequence.

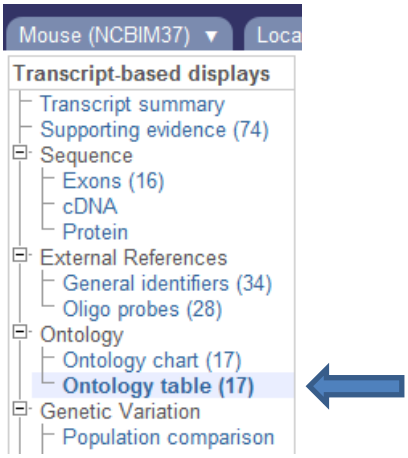
Next, follow the *General identifiers* link at the left.



This page shows information that matches to the Ensembl transcript and protein from RefSeq, EntrezGene, OMIM, UniProtKB, and others.

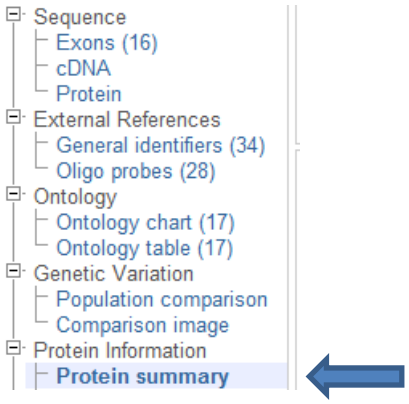


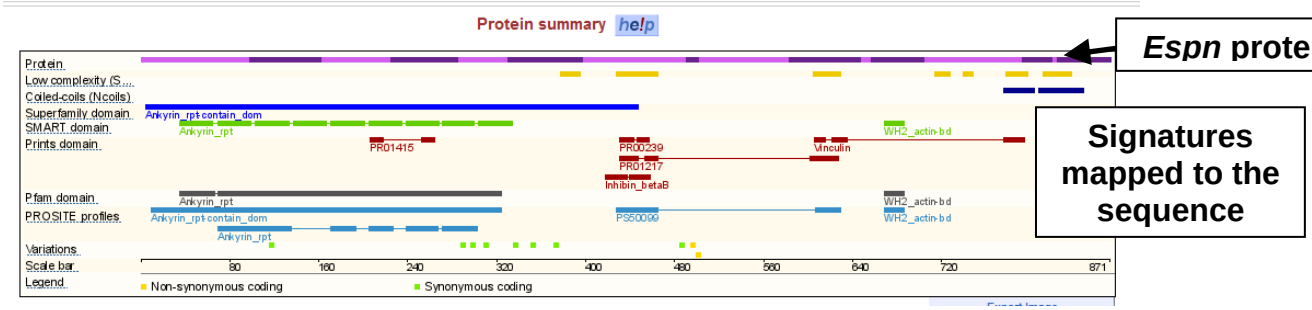
Click on *Ontology table* to see GO terms from the Gene Ontology consortium. *www.geneontology.org*



Click on the Help page to see a guide to the three letter Evidence codes.

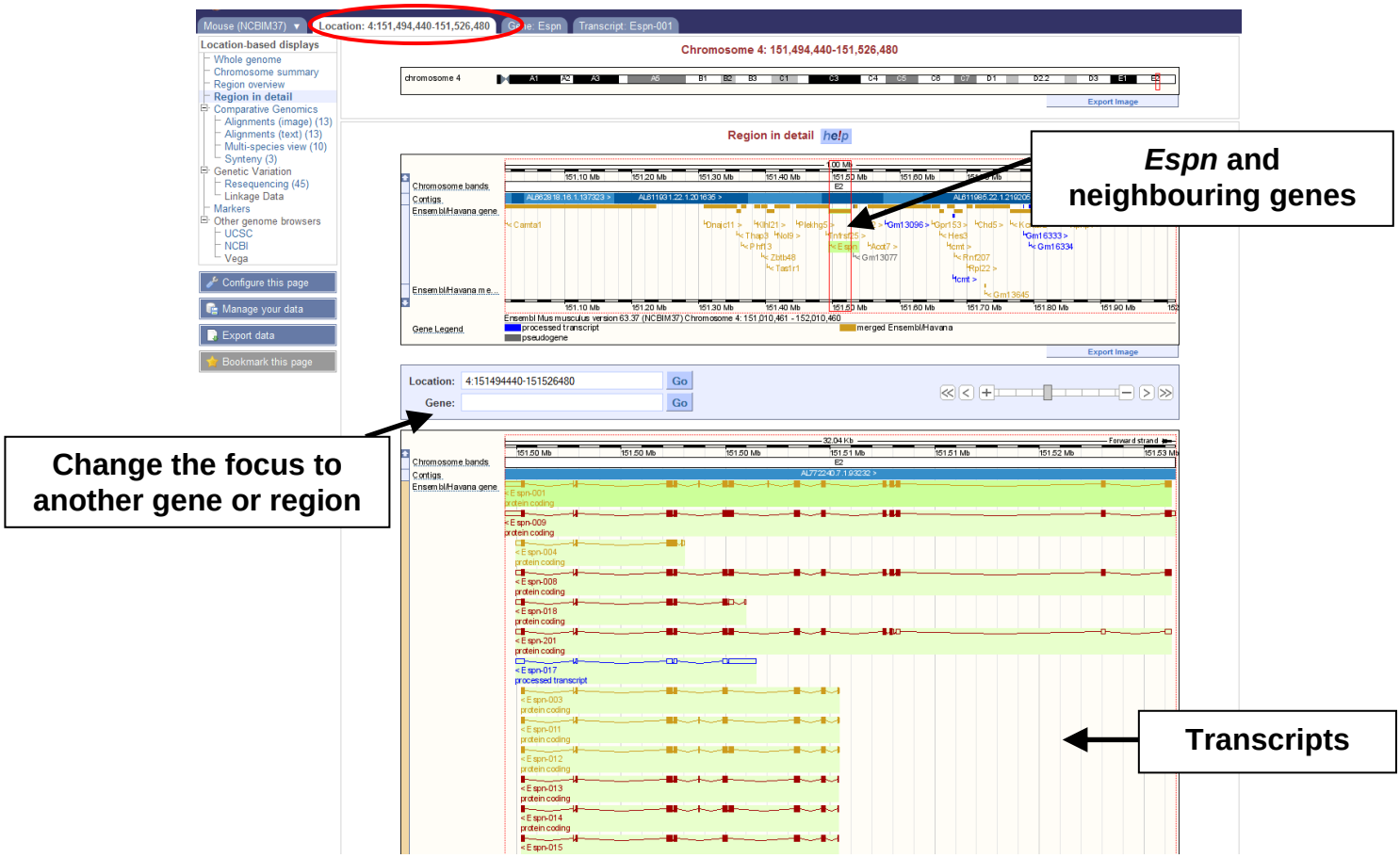
Now click on *Protein summary* to view mapped domains and signatures.





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 the *Location* tab.



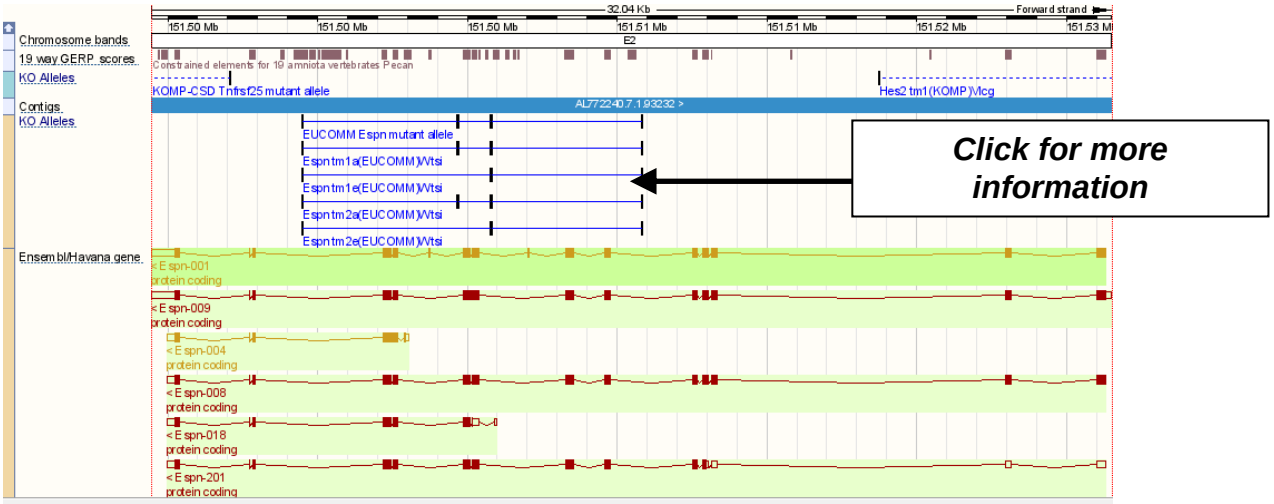
Ensembl *Location* displays are highly configurable. You can switch on additional tracks to view a wealth of data along the genome.

Use the Configure this page **link to add**

- ***Constrained elements for 19 amniota vertebrates Pecan:***
  Under the "Comparative genomics" menu click on "conservation regions" and then select Constrained elements;
- ***KO Alleles:***
  Under the "Genes and transcripts" menu click on "Targetted analysis" and then select DAS KO.

Click on information.

Have a look at the changes in the region in detail view. Click and drag tracks to reorder them, if it helps with comparing the data.

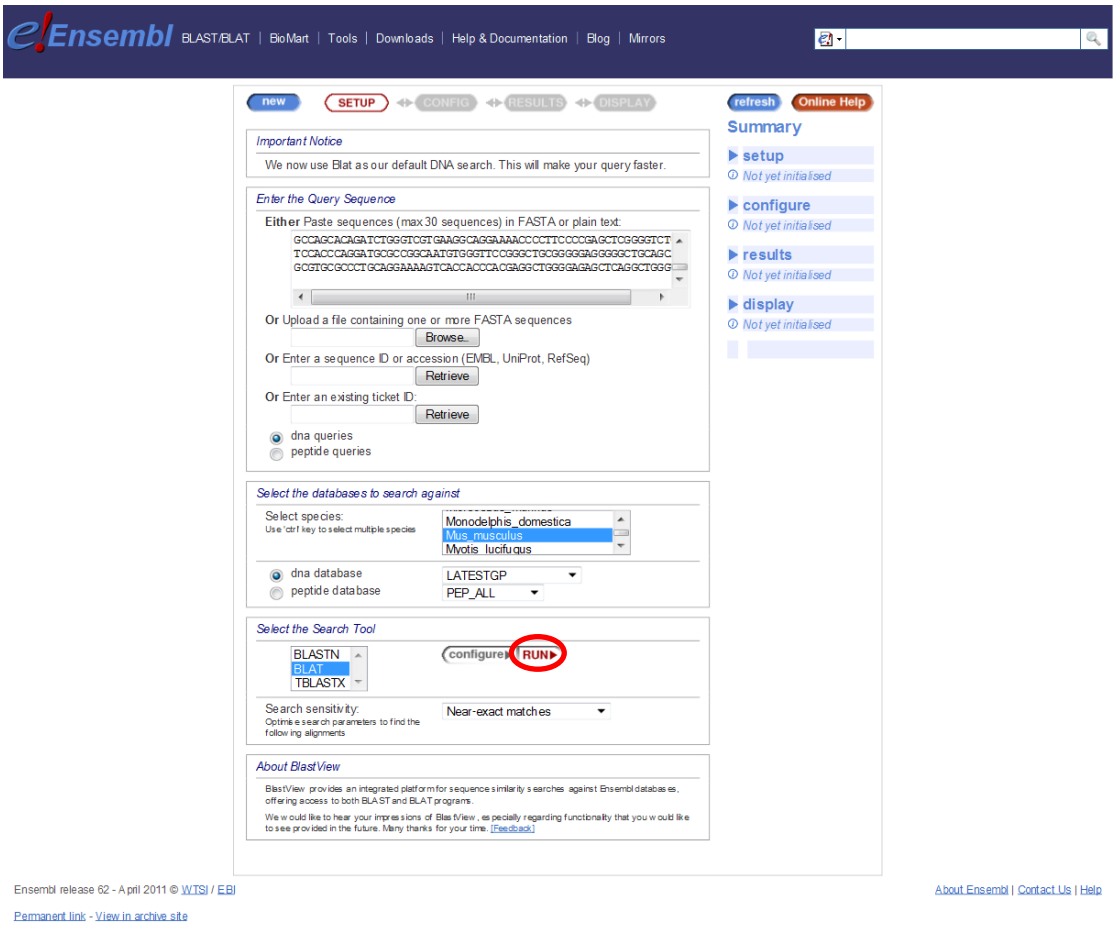


Note: You can search for ENSMUSG00000028943 at *www.knockoutmouse.org*

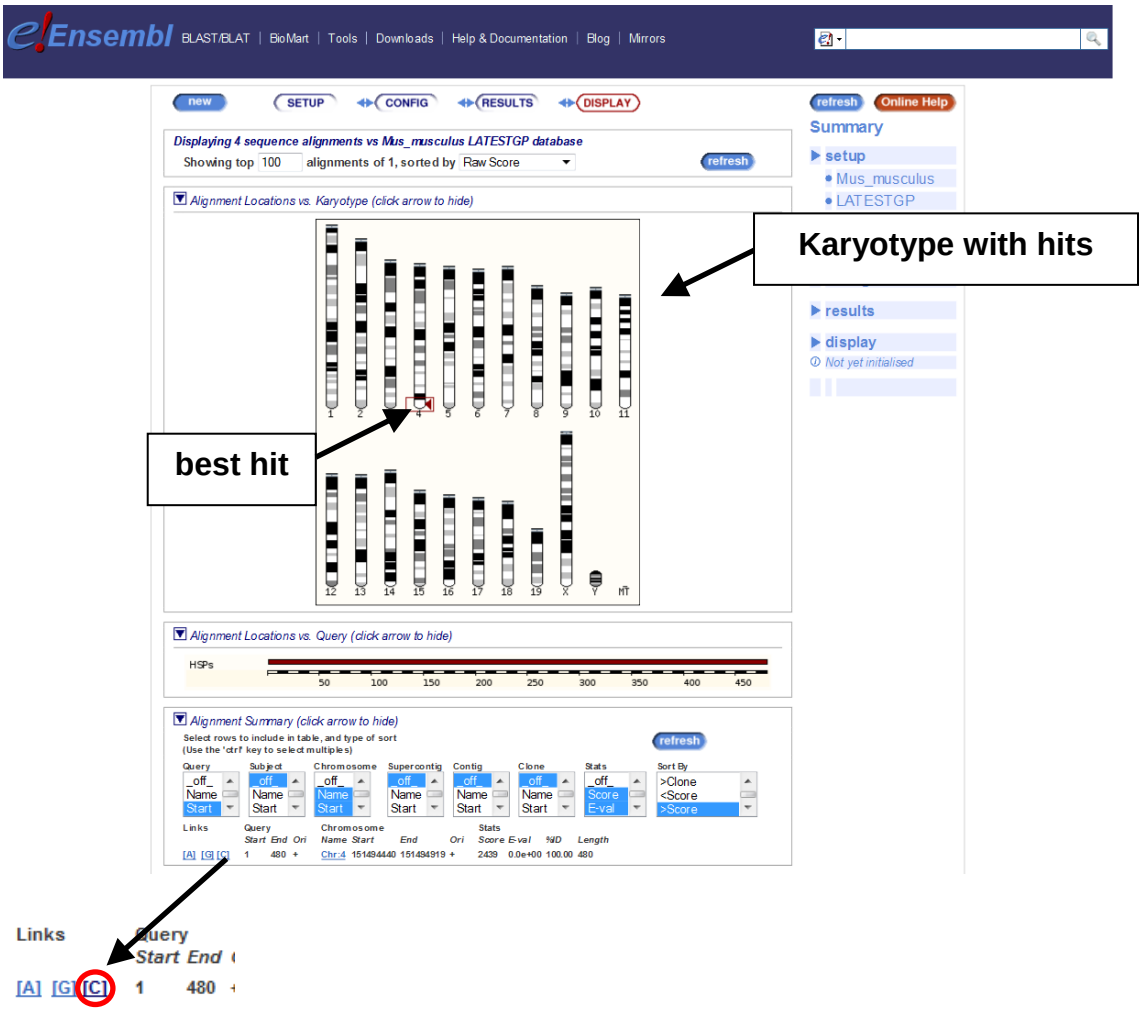
Our last task is to export genomic sequence. Click the Export data option and click *Next*. Now click *HTML*.

```
>4 dna:chromosome chromosome:NCBIM37:4:151494440:151526480:1
TGGCGGATAGAGGTTTAATGGGGGAAGGAAGCGGGTGGGGGATAGGATGAGGAGCAGGTA
CAATTGCAGGGTCCCTGGCCCTCCACTTACACCACTGGGGCAGTGAAAGGGTCAGGGGAG
CTATGGTCCTGGCTGGGGACTTGGAAGGCTTAATTGGGGCAGGGGTGAGGTGGGGATGGG
ATGTAAAGGTGGGGCCCCTCAGGAGCGACAGCTCCGCCCCAACAGGCAAGTTTCATCCTT
TCTGGTAAACAAATGACTGAGAGGAAACCCTCGCTGGGGATGAGAAACGCAGCGACTTTG
GCCAGCACAGATCTGGGTCGTGAAGGCAGGAAAACCCCTTCCCCGAGCTCGGGGTCTGCC
TCCACCCAGGATGCGCCGGCAATGTGGGTTCCGGGCTGCGGGGGAGGGGGCTGCAGCAGG
GCGTGCGCCCTGCAGGAAAAGTCACCACCCACGAGGCTGGGGAGAGCTCAGGCTGGGTCA
```

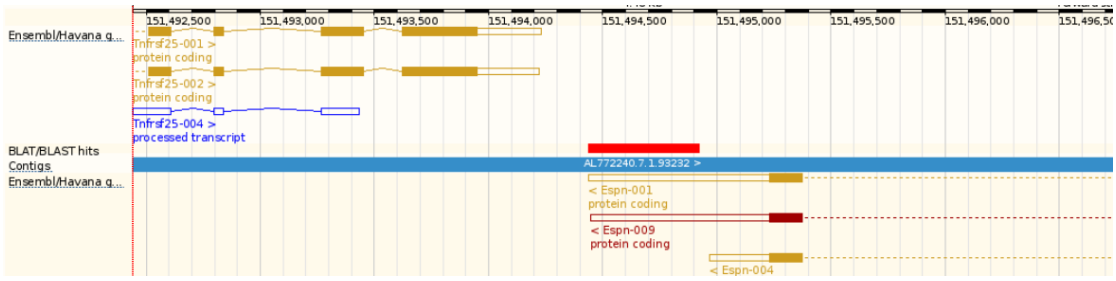
Select the header and a few lines of the nucleotide sequence using Edit/Copy in your browser. Click on the BLAST/BLAT link in the bar at the top of the page. Paste the sequence into the appropriate box and select *BLAT* as the search algorithm.



Finally, click *Run*.



Follow links to an alignment [A], genomic sequence [G] and the corresponding Location View [C] (to C (see) the hit!). Clicking on [C] should reveal the BLAT hit in "Region in detail":



Click on the red bar for the score, %ID, and other BLAST/BLAT values.

END OF THE WORKED EXAMPLE

III) EXERCISES and ANSWERS

Note: the answers to these exercises correspond to current version (64) of www.ensembl.org. If you use these exercises in the future, after Ensembl is updated, please use the archive site for version 64.

BROWSING ENSEMBL

Exercise 1 – Exploring the human *MYH9* gene

Use www.ensembl.org

(a) Find the human *MYH9* (myosin, heavy chain 9, non-muscle) gene, and go to the Gene tab.

- 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?
- What is the longest transcript, and how long is the protein it encodes?
- Which transcript has a CCDS record associated with it?

? *Why is the CCDS important- what does it tell us?*

(b) Click on ***Phenotypes*** at the left of the page. Are there any diseases associated with this gene, according to MIM (*Mendelian Inheritance in Man*?)

(c) In the transcript table, click on the transcript ID for the longest protein, and go to the Transcript tab.

- How many exons does it have?
- Are any of the exons completely or partially untranslated?
- Have a look at the *General identifiers* for the transcript. Is there an associated sequence in UniProtKB/SwissProt?
- ? *What is the difference between these matches and the associations you saw in question 1(b)?*
- Have a look at the *Ontology table views* for the transcript. What are some functions of MYH9-001?

(d) Are there probesets on microarray platforms that can be used to monitor ENST00000216181 expression?

(e) What protein signatures can be found in the amino acid sequence?

(f) Go back to the Gene tab. View the cDNA alignment with the mouse orthologue.

Exercise 2 – Exploring a mouse region

Start in http://www.informatics.jax.org/

(a) Search the MGI browser for *Ush1c*, and click on the hit. Where is it located (which chromosome and basepairs?)

(b) Jump to the Ensembl Genome Browser under the Sequence Map section. What are the neighbouring genes to mouse *Ush1c*?

(c) Is there a knockout mouse available for this gene?

? *How would you confirm that the KO Allele targets all protein coding transcripts determined by Ensembl?*

(d) Zoom in to the region spanned by the KO Alleles (you can click and drag a box with your mouse to do this, and then 'jump to region'.). How well are the Ensembl exons supported by CCDS and the UniProtKB tracks?

(e) Are there any coding sequence variants in this region?

(f) Go to the synteny view. Which human chromosomes have synteny to mouse chromosome 7? What is the gene ID of the human homologue of mouse *Ush1c* (you can find this below the figure).

Exercise 3 – Exploring a bacteria gene

Use www.ensemblgenomes.org, Ensembl Bacteria

(a) Who sequenced the *Bacillus subtilis* genome? Where are the genes from?

(b) Find the *Bacillus subtilis* dhaS gene. Does Ensembl predict orthologues for this gene? What is the difference between 'Bacterial Compara' and 'Pan-taxonomic Compara'?

(c) What is the protein sequence for this gene? Are there any motifs or signatures present in the amino acid sequence?

Answers (Browsing Ensembl)

Answer 1 – Exploring the human MYH9 gene

(a) *Go to the Ensembl homepage (http://www.ensembl.org).*
*Select '**Search: Human**' and type '**MYH9 gene**'*
Click [Go].
*Click on '**Homo sapiens**' on the page with search results.*
*Click on '**Gene**'.*
*Click on '**Ensembl protein_coding Gene: ENSG00000100345 (HGNC Symbol: MYH9)**'.*

- *Chromosome 22 on the reverse strand.*
- *Ensembl has 14 transcripts annotated for this gene.*
- *Five transcripts are protein coding.*
- *The longest transcript is MYH9-202 and it codes for a protein of 562 amino acids*
- *MYH9-001 has a CCDS record. CCDS is the consensus coding sequence set. These coding sequences (CDS) have been agreed upon by Ensembl, NCBI, UCSC and VEGA/Havana.*

? *The CCDS set is a collection of reviewed, agreed-upon coding sequences (for human mouse). These sequences are high-confidence, and unlikely to change in the future.*

(b) *Autosomal dominant deafness, Epstein syndrome, and Fechtner syndrome are among the MIM diseases associated with MYH9. Click on any of these for more information in the MIM record itself.*

(c) *Click on **ENST00000216181***

- *It has 41 exons. This is shown in the Transcript summary.*
- *Click on '**Exons**' in the side menu. Exon 1 is completely untranslated, and exons 2 and 41 are partially untranslated (UTR sequence is shown in purple). You can also see this in the cDNA view.*
- *Click on '**General identifiers**' in the side menu. MYH9-HUMAN from Swiss-Prot matches the Ensembl transcript. Click on it to go to UniProtKB, or click 'align' for the alignment.*

? *These are associated with a specific transcript. In **1(b)** we looked at gene associations with* diseases, according to MIM (*Mendelian Inheritance in Man*).

- *Have a look at '**Ontology table**'. The Gene Ontology project maps terms to a protein in three classes: biological process, cellular component, and molecular function. Roles in meiotic spindle organization, cell morphogenesis, and cytokinesis are associated with MYH9-001.*

***(d)** Click on '**Oligo probes**' in the side menu. Probesets from Affymetrix, Agilent, Codelink, Illumina, and Phalanx match to this transcript sequence.*

***(e)** Click on '**Protein summary**' in the side menu. Click on the Help button if you're not sure how to read the image.*

*Click on '**Protein Information - Domains & features**' in the side menu to see a table of these same peptidase signatures.*

(f) *Click on the Gene tab, then '**Orthologues**' in the side menu. There is one mouse orthologue predicted for human MYH9; scroll down to see it. Click on '**Alignment (cDNA)**' in the mouse row.*

Answer 2 – Exploring a mouse region

(a) *Search http://www.informatics.jax.org/ for Ush1c, and click on the first hit. In the mouse genome NCBI 37, the gene is on chromosome 7 from base pair 53450720 to 53493873.*

(b) *In the Ensembl Region in Detail page, you can see the genes on either side of Ush1c are Abcc8 and Otog.*

(c) *Turn on the KO Alleles track by searching for it in the **Configure this page** menu under Targetted analysis. Back in the Region in Detail page, click on the blue horizontal bar in the KO allele track and then on the clickable KOMP* link (KOMP-CSD Ush1c mutant allele) *to see the knockout mouse.*

- *Look at the transcript structures, and make sure all of the protein coding transcripts have exons in the knock-out region. In this case, they do.*

(d) *Click and drag a box around the KO Alleles, and 'jump to region'. The view should be zoomed in. Reorder the tracks if it helps to see that the CCDS and UniProtKB entries line up nicely with the Ensembl exons in this region. Hint: To expand the UniProtKB track, click the track name and use the 'tool' icon to change it from 'stacked' to 'normal'.*

(e) *The Sequence variants (all sources) track should be on- if not, turn it on with **Configure this page** (in the Germline*

variation menu). There are yellow (non-synoymous coding) variations in this region. Click on a yellow line for more information about the variant. You may need to zoom in further.

***(f)** Click **Synteny** at the left. In the centre of the view, mouse chromosome 7 is shown. Human chromosomes with synteny are shown as little chromosomes circling the mouse chromosome. They are human chromosomes 10, 15, 19, 11 and 16.*
Looking below the image, human gene ID ENSG00000006611 is homologous to the mouse Ush1c gene.

Answer 3 – Exploring a bacterial gene

***(a)** Go to the Ensembl Bacteria homepage: (http://bacteria.ensembl.org). Select the genome for Bacillus subtilis. The papers show the scientists involved in the sequencing project (Kunst, F. et. al. and Barbe, CV. et. al.) Genes come from EMBL/Genbank/DDBJ sequences.*

***(b)** A search for dhaS should lead you to the gene tab. Click '**Orthologues**' at the left. 79 Orthologues are predicted across the bacteria. Note there is another Orthologues link in the Pan-taxonomic Compara heading. This shows 188 orthologues across all Ensembl species (including plants, vertebrates, etc).*

***(c)** Click on the Protein ID **EBBACP00000002246** in the Transcript table at the top of the view. The view shows Aldehyde_DH signatures from various projects (for example Superfamily and Pfam). These motifs can also be found in the '**Domains & features**' link at the top left of the menu. Click on '**Protein sequence**' at the top left for the amino acid sequence.*

IV) BioMart

Mining data- worked example

You have three questions about a set of human genes.
These gene symbols come from 'HGNC' www.genenames.org

Espn, Myh9, Ush1c, Chd7, Cisd2, Thrb, Whrn

- What are the EntrezGene IDs for these genes?
- Are there associated functions from the GO (gene ontology) project that might help describe their function?
- What are their cDNA sequences?

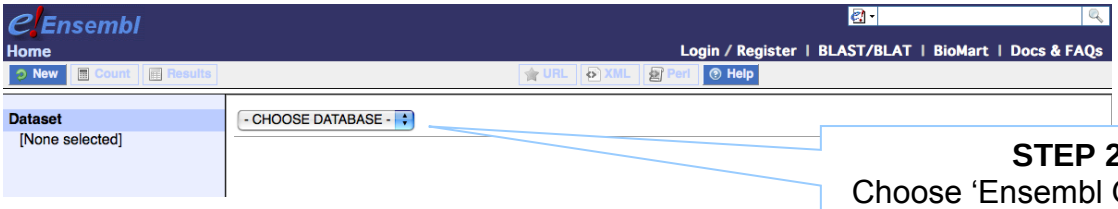
For more about BioMart, have a look at these publications!

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

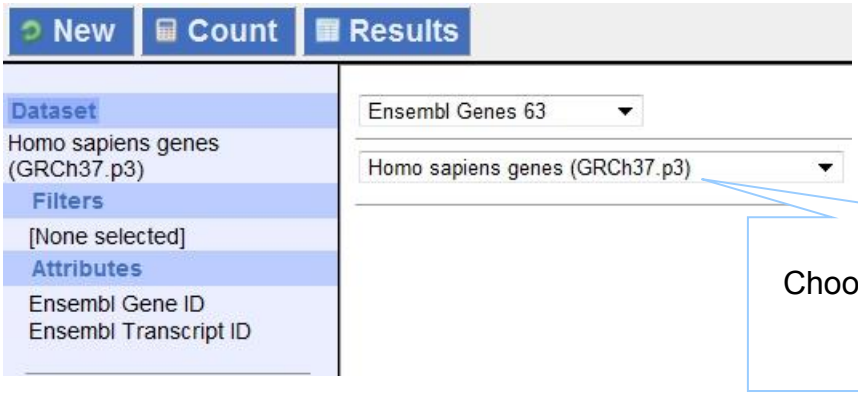
Kinsella, R.J. *et al*
Ensembl BioMarts: a hub for data retrieval across taxonomic space.
Database (Oxford) 2011:bar030

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

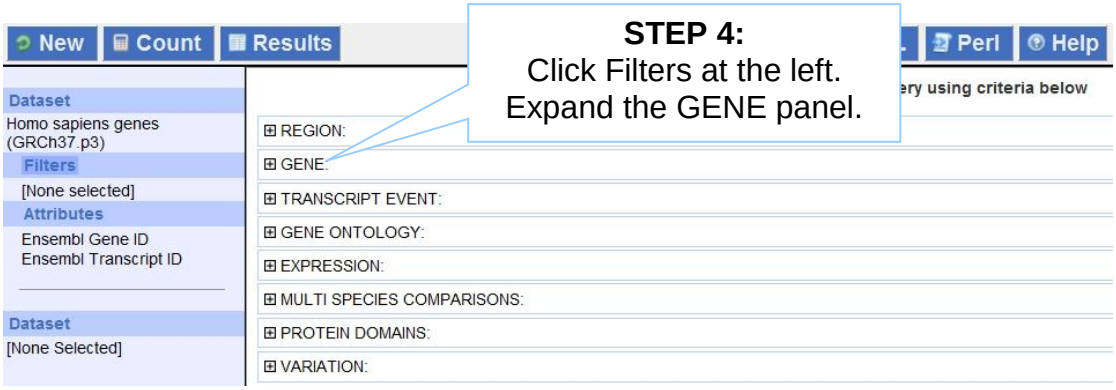
NOTE: These answers were determined using Ensembl Mart 64



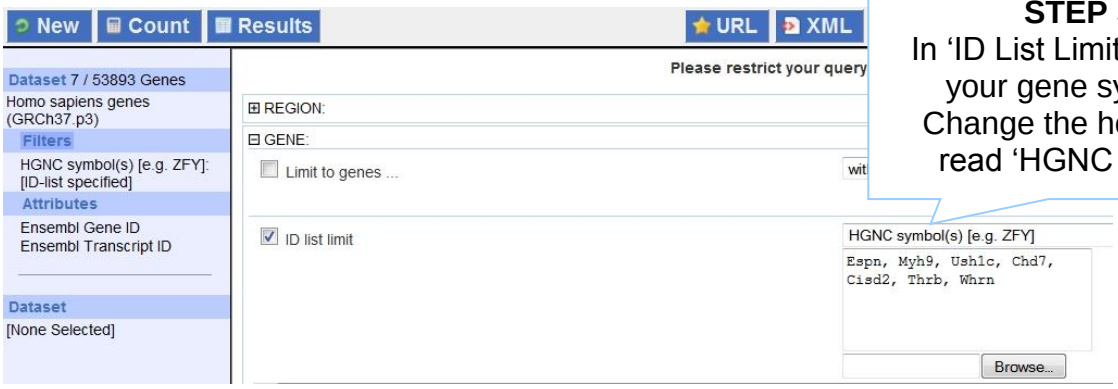
STEP 2:
Choose 'Ensembl Genes 64 as the primary database.



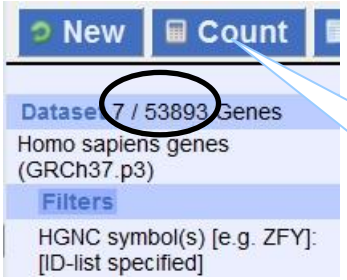
STEP 3:
Choose '*Homo sapiens* genes' as the dataset.



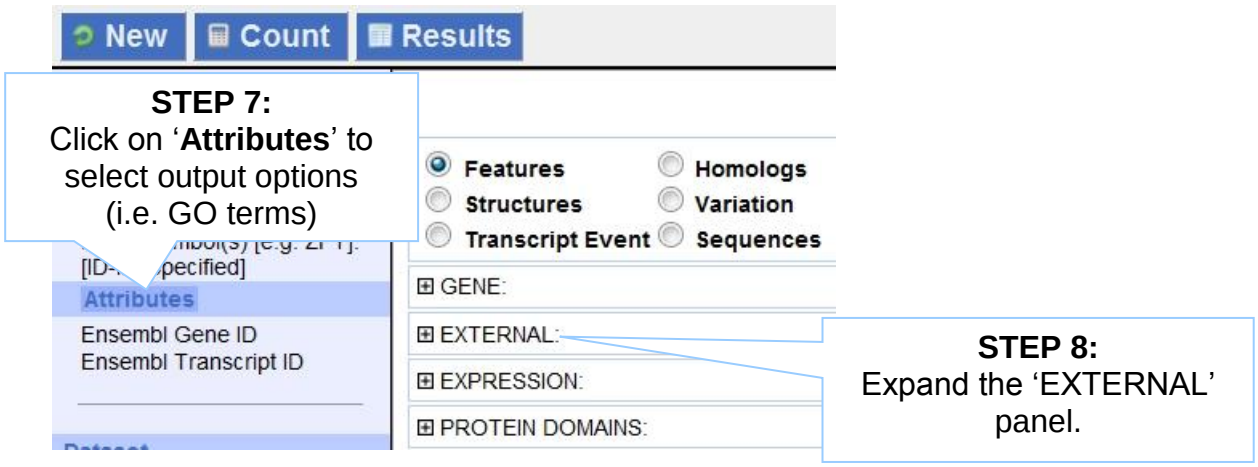
STEP 4:
Click Filters at the left.
Expand the GENE panel.



STEP 5:
In 'ID List Limit', paste in your gene symbols. Change the heading to read 'HGNC symbol'

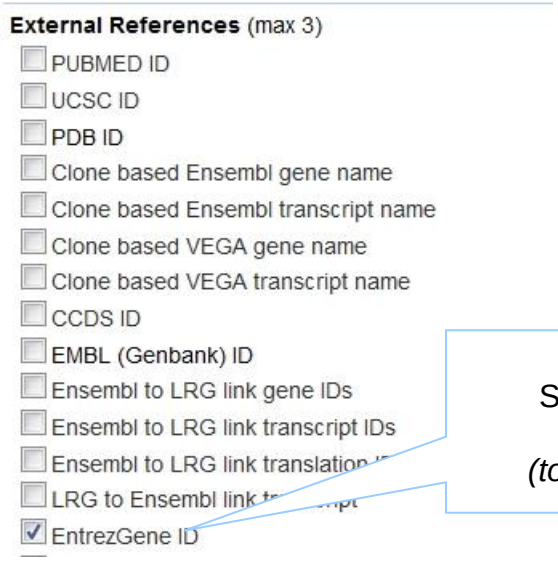


STEP 6:
Click '**Count**' to see BioMart is reading 7 genes out of 53,893 possible *H. sapiens* genes. (Note: in v64, the total gene count is 54,345)

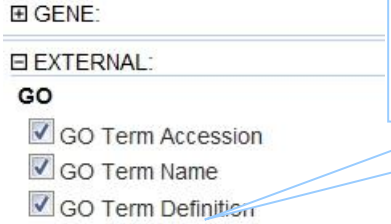


STEP 7:
Click on '**Attributes**' to select output options (i.e. GO terms)

STEP 8:
Expand the 'EXTERNAL' panel.

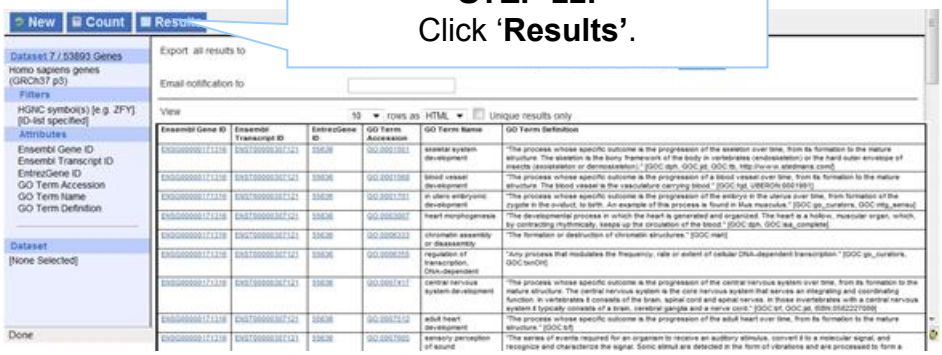


STEP 9:
Scroll down to select '**EntrezGene ID**'
(to answer question 1)



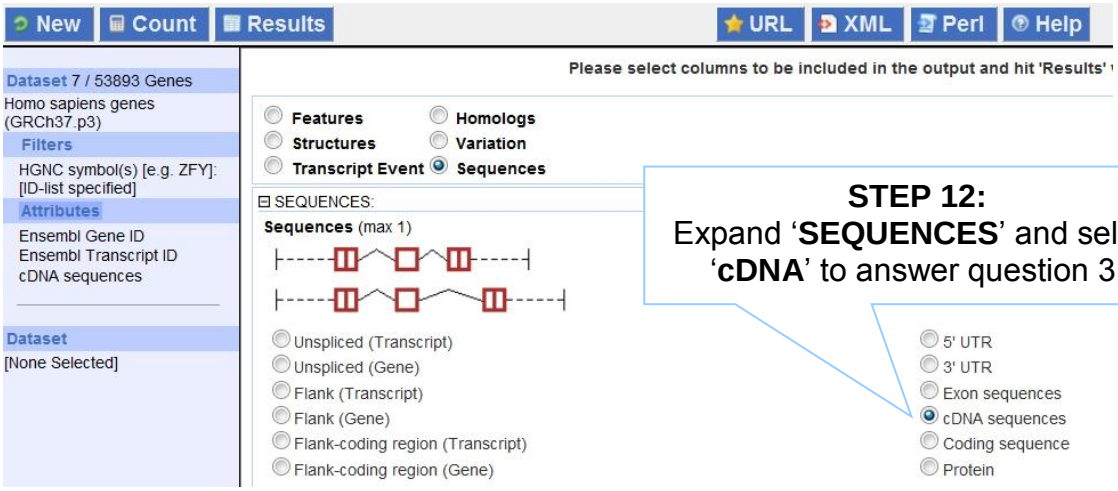
STEP 10:
Scroll back up to select '**GO term'** fields
(to answer question 2)

STEP 11:
Click '**Results**'.

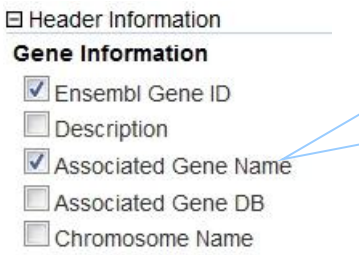


Why are there multiple rows for one gene ID? For example, look at the first few rows?

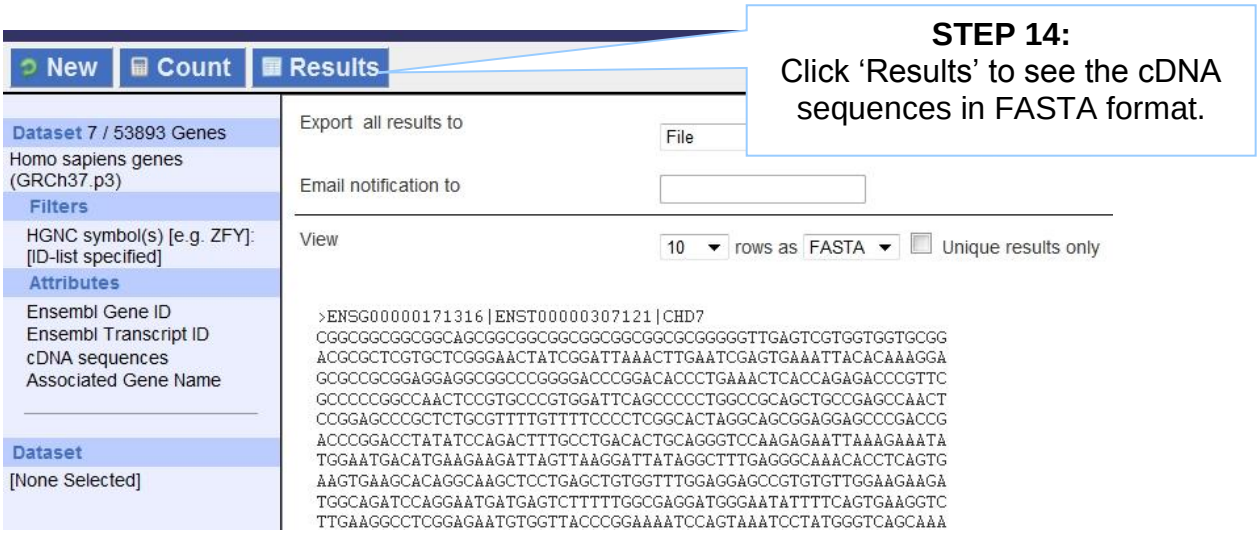
Ensembl Gene ID	Ensembl Transcript ID	EntrezGene ID	GO Term Accession	GO Term Name
ENSG00000171316	ENST00000307121	55636	GO:0001501	skeletal system development
ENSG00000171316	ENST00000307121	55636	GO:0001568	blood vessel development
ENSG00000171316	ENST00000307121	55636	GO:0001701	in utero embryonic development



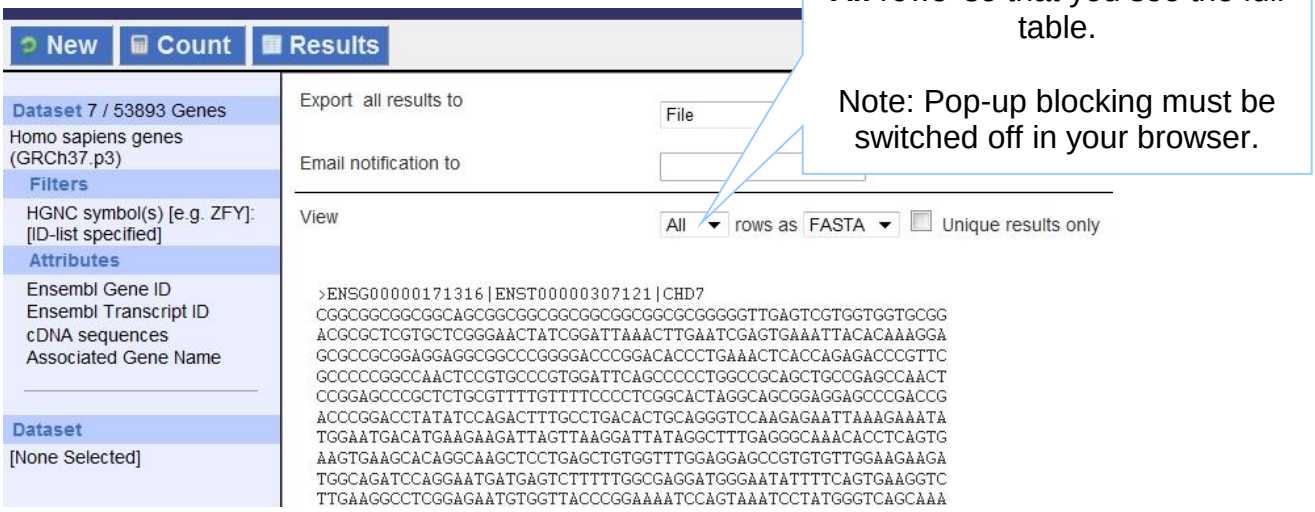
STEP 12:
Expand '**SEQUENCES**' and select '**cDNA**' to answer question 3.



STEP 13:
Expand **'Header Information'** to select the **'Associated Gene Name'**



STEP 14:
Click 'Results' to see the cDNA sequences in FASTA format.



STEP 15:
Change 'View **10** rows' to 'View **All** rows' so that you see the full table.

Note: Pop-up blocking must be switched off in your browser.

Note: you can use the 'Go' button to export a file.

- *What did you learn about the human genes in this exercise?*
- *Could you learn these things from the Ensembl browser? Would it take more time?*

V) BioMart Exercises and Answers

These exercises demonstrate the benefit of using BioMart as a data-mining tool. The aim is to enable you to become familiar with the BioMart interface and to encourage you to explore the different data queries that are possible.

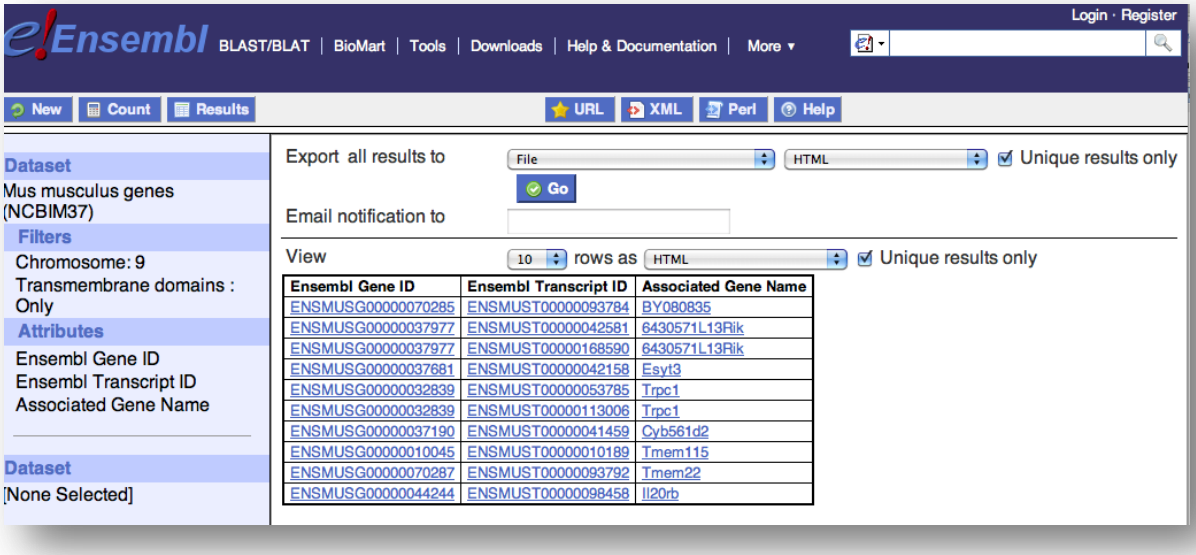
1. We guide you through using BioMart to find mouse proteins with a trans-membrane domain that fall on chromosome 9 of the mouse genome.

As with all BioMart queries you must select the **dataset**, set your **filters** and define your **attributes**. For this exercise:
Dataset: Ensembl genes in mouse
Filters: Transmembrane proteins on chromosome 9
Attributes: Ensembl gene and transcript IDs and Associated gene names

- Go to the Ensembl homepage (http://www.ensembl.org) and click on BioMart at the top of the page.
- Select 'Ensembl genes' as your database and 'Mus musculus genes' as the dataset.
- Click on 'Filters' on the left of the screen and expand 'REGION'. Change the chromosome to '9'.
- Now expand 'PROTEIN DOMAINS', also under filters, and select 'Transmembrane domains' and then ' Only'. Clicking on 'Count' should reveal that you have filtered the dataset down to 426 genes (this result was obtained using the Ensembl Mart 64).
- Click on 'Attributes' and expand 'GENE'. Select 'Associated gene name'.

Now click on 'Results'. The first 10 results are displayed by default; display all results by selecting 'ALL' from the drop down menu.

You should see a similar output to the screenshot below, which displays the Ensembl gene ID, Ensembl Transcript ID and Associated gene names of all proteins with a transmembrane domain on chromosome 9. If you prefer, you can also export to an Excel sheet by using the 'Export all results to' XLS option.



Exercise 2

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

http://www.ebi.ac.uk/~gspudich/workshop_presentations/coursebook_64m.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 3

For a list of Ciona savignyi Ensembl genes, export the human orthologues.

ENSCSAVG00000000002, ENSCSAVG00000000003, ENSCSAVG00000000006, ENSCSAVG00000000007, ENSCSAVG00000000009, ENSCSAVG00000000011

Exercise 4

You can use BioMart to query variations, not just genes.

Step 1: Export the locations, names, and descriptions of human structural variations (CNVs) on chromosome 1.

Step 2: New query. Start with two human dbSNP IDs: rs1801500 and rs1801368. Find their alleles, phenotype descriptions, and associated genes using BioMart. Can you view this same information in the Ensembl browser?

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 human 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.

Answers: BIOMART

Answer 2.
Click [New].
*Choose the '**ENSEMBL Genes 64**' 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 **11** genes (you can also see this by clicking 'Count').*

Answer 3. *Click* ***[New].***
Choose the ***'ENSEMBL Genes 64'*** *database.*
*Choose the '****Ciona savignyi genes*** *(CSAV2.0)' dataset.*

Click on ***'Filters'*** *in the left panel.*
Expand the ***'GENE'*** *section by clicking on the + box.*
Enter the gene list in the ***ID List Limit*** *box.*

Click on ***'Attributes'*** *in the left panel.*
Select the ***'Homologs'*** *attributes page.*
Expand the ***'Orthologs'*** *section by clicking on the + box.*
Select ***'Human Ensembl Gene ID.***
Click [Results] (remember to tick the 'unique results only' box)

Answer 4. Step 1: *Choose 'Ensembl Variation 64' and 'Homo sapiens Structural Variation'.*
Filters: *Region: Chromosome 1*
Attributes: *Chromosome name, Sequence region start (bp), Sequence region end (bp), Structural Variation Accession, Structural Variation Description*

Step 2: *Choose 'Ensembl Variation 64' and 'Homo sapiens Variation (dbSNP 132, ENSEMBL).*
Filters: *'Filter by Variation ID' enter: rs1801500, rs1801368*
Attributes: *'Variation ID', 'Variant Alleles', 'Phenotype description', 'Associated gene'.*

You can view this same information in the Ensembl browser. Click on one of the variation IDs (names) in the result table. The variation tab should open in the Ensembl browser. Click Phenotype Data at the left for more details on the SNP.

Answer 5. (a) *Click* ***[New].***
Choose the ***'ENSEMBL Genes 64'*** *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.*
In addition to the default selected attributes, select ***'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 most of the human genes at least one mouse orthologue has been identified.

TASK – A DELETION ON CHROMOSOME 6 ASSOCIATED WITH MENTAL RETARDATION

GOAL

In this tutorial we will explore a deletion on the short arm of human chromosome 6 associated with mental retardation using the Ensembl browser and BioMart.

BACKGROUND

Krepischi *et al.* detected in a patient with mental retardation a heterozygous deletion on 6p21. Using array CGH the deletion was mapped to a 812.77 kb segment on 6p21.31–21.32 (chr6:33,273,955–34,086,729); the distal breakpoint was mapped to a 14.3 kb interval (chr6:33,259,651–33,273,955); the proximal breakpoint was located on a 123 kb segment (chr6:34,086,729–34,209,880).

Note: the genomic coordinates mentioned above are relative to the GRCh37 (hg19) genome assembly.

See also: Krepischi ACV, Rosenberg C, Costa SS, Crolla JA, Huang S, Vianna-Morgante AM. 2010. A novel de novo microdeletion spanning the *SYNGAP1* gene on the short arm of chromosome 6 associated with mental retardation. Am J Med Genet Part A 152A:2376–2378.

TOPICS COVERED

(1) Viewing the deletion and its breakpoints using Ensembl
(2) Retrieving a list of the genes encompassed by the deletion using BioMart

---

(1) Viewing the deletion and its breakpoints using Ensembl

To get an overview of the whole region of the deletion, including the breakpoints, search from the homepage for the region '6:33259651-34209880'.

- Go to the Ensembl v64 homepage (http://www.ensembl.org).
- Select 'Search: Human' and type '6:33259651-34209880' in the 'for' text box.
- Click [Go].

This directly leads to the 'Region in detail' page. Note that our region of interest is indicated by the red box in the 1Mb overview panel and shown in more detail in the bottom panel.

Before adding new tracks, first remove any tracks added in the previous task.

- Click [Configure this page] in the side menu.
- Click [Reset configuration].
- Click (✓).

Note that individual tracks can also be deleted by hovering over the track title and clicking the 'Turn track off' icon in the pop-up menu.

Add the DECIPHER (DatabasE of Chromosomal Imbalance and Phenotype in Humans using Ensembl Resources) track.

- Click [Configure this page] in the side menu.
- Type 'decipher' in the 'Find a track' box.
- Select 'decipher – Normal'.
- Click (✓).

To have a closer look at the distal breakpoint, simply change the coordinates at the top of the bottom panel.

- Change the coordinates in the 'Location' box to '6:33259651-33273955'.
- Click [Go].

Similar for the proximal breakpoint.

- Change the coordinates in the 'Location' box to '6:34086729-34209880'.
- Click [Go].

? Question

Do the breakpoints encompass any genes?

---

(2) Retrieving a list of the genes encompassed by the deletion using BioMart

Using BioMart it is very easy to generate an Excel spreadsheet that contains the following information for all genes that are encompassed by the deletion (including the breakpoint regions): Ensembl Gene ID, genomic coordinates, name, description. To see whether any of the genes has been reported before to be

associated with mental retardation according to the Online Mendelian Inheritance in Man (OMIM) database (http://www.omim.org), also include the MIM Morbid Description.

Step 1: Dataset

- Go to the Ensembl homepage (http://www.ensembl.org).
- Click the 'BioMart' link on the toolbar.

Start with all human Ensembl genes.

- Choose the 'Ensembl Genes 64' database.
- Choose the 'Homo sapiens genes (GRCh37.5)' dataset.

Note that everything you do is confirmed in the side menu.

Step 2: Filters

Now filter for the genes on chromosome 6.

- Click on 'Filters' in the left panel.
- Expand the 'REGION' section by clicking on the + box.
- Select 'Chromosome - 6'.

Make sure that the box in front of the filter is checked (this should happen automatically), otherwise the filter is not turned on.

- Click on [Count].

This should give you 2781 / 54345 Genes.

Now filter further for genes that are in the region from bp 33,259,651–34,209,880.

Select 'Gene Start (bp) – 33259651'.
Select 'Gene End (bp) – 34209880'.
Click on [Count].

This should give you 31 / 54345 Genes.

Step 3: Attributes

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.
Expand the 'GENE' section by clicking on the + box.
Deselect 'Ensembl Transcript ID'.
Select, in addition to the default selected 'Ensembl Gene ID', 'Chromosome Name', 'Gene Start (bp)', 'Gene End (bp)', 'Associated Gene Name' and 'Description'.
Expand the 'EXTERNAL' section by clicking on the + box.
Select 'MIM Morbid Description'.

Step 4: Results

Have a look at a preview of the results (only 10 rows of the results will be shown).

Click the [Results] button on the toolbar.

If you are happy with how the results look in the preview, output all the results to an Excel spreadsheet.

Select 'Export all results to File – XLS'.
Tick the box 'Unique results only'.

Note: BioMart is 'transcript-centric', which means that it will often 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. To get rid of these redundant rows, use the 'Unique results only' option.
Click [Go].

This should give you an Excel spreadsheet with 31 rows of results.

Question

Is any of the genes encompassed by the deletion associated with mental retardation according to the OMIM database?

ANSWERS - A DELETION ON CHROMOSOME 6 ASSOCIATED WITH MENTAL RETARDATION

---

(1) Yes, the distal breakpoint encompasses (parts of) three genes, i.e. *PFDN6*, *RGL2* and *TAPBP*, while the proximal breakpoint also encompasses (parts of) three genes, i.e. *GRM4*, *CYCSP55* and *HMGA1*.

(2) Yes, according to the OMIM database the *SYNGAP1* gene has been associated with autosomal dominant mental retardation (http://www.omim.org/entry/612621). Note that Krepischi *et al.* end their paper saying "SYNGAP1 haploinsufficiency might be the main cause of mental retardation and severe speech impairment associated with the novel 6p deletion described here.".

VII) EXERCISES VARIATIONS AND REGULATION

Exercise 1 – A Human SNP

The SNP rs1738074 in the 5' UTR of the human TAGAP gene has been identified as a genetic risk factor for a few diseases.

(a) In which transcripts is this SNP found?

(b) What is the least frequent genotype for this SNP in Caucasians (CEU populations)?

(c) With which diseases is this SNP associated? Are there any known risk alleles?

Exercise 2 – Rice variation

(a) Find the **LOC_Os01g42030** gene in *Oryza sativa*. Are there any sequence variants from dbSNP? Are any non-synonymous coding SNPs?

(b) Draw variations on the genomic sequence using the 'Sequence' link in the gene tab. Can you find the non-synonymous variation?

Exercise 3 – Human: Exploring a region

Use www.ensembl.org to learn more about a region of human chromosome 13 that you find aberrant in a patient.

(a) Search for human chromosome 13, base pairs 32,591,538-32,976,859. How many gold genes are in this region?

(b) Turn on the track from the Decipher project. Is there data from the project in this region?

(c) Click 'Region overview' at the left. This view allows you to zoom out over 1Mb (The 'Region in detail' view is limited to 1Mb maximum). Is there a structural variant that covers regions in band q12.3 and q13.1 (zoom out to see this).

Exercise 4 - Gene regulation: Human 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) Use 'Configure this page' to turn on tracks for one cell type. Are there sites enriched for marks of open chromatin (e.g. DNase1 and FAIRE) and for polymerase binding (both PolII and PolIII) in HeLa cells at the 5' end of STX7?

ANSWERS VARIATIONS AND REGULATION

Answer 1

(a) There is more than way to get this answer. Either go to the '**Variation Table**' for the human TAGAP gene, and 'Show' variants in the 5'UTR, or search Ensembl for rs1738074 directly.

Once you're in the Variation tab, click on the Gene/Transcript link. This SNP is found in three transcripts (*ENST00000326965, ENST00000338313, and ENST00000367066).*

(b) *Click on '**Population genetics**" at the left of the variation tab.*
In Caucasians (CSHL-HAPMAP:HapMap-CEU population) the least frequent genotype is T/T. This is also the least frequent genotype in CID, GIH, TSI, and other populations (to decode the three letter population codes, see HapMap or dbSNP: http://www.ncbi.nlm.nih.gov/SNP/snp_viewTable.cgi?pop=1409)

(c) *Click '**Phenotype Data**' at the left of the Variation page to see the variant is associated with diabetes and celiac. The allele A is associated with celiac disease. Note that the alleles reported by Ensembl are T/C, and Ensembl reports alleles on the forward strand. This suggests that A was reported on the reverse strand.*

Answer Exercise 2 - Rice

(a) *Go to plants.ensembl.org and search for LOC_Os01g42030 in the Oryza sativa genome. You can find this information in the Gene tab, Variation Table. Three variations map to the gene, and one is non-synonymous (though it is listed three times, there is one rsID for this SNP (rs18388594).*

(b) *Click on **Sequence** at the left. Use **Configure this page** to choose 'Show variations: Yes and show links'. Either search for **rs18388594** (using Cntrl F) or scroll down the page to find the yellow highlighted variation.*

Answer Exercise 3 – Exploring a region

(a) *Search www.ensembl.org for human 13:32591538-32976859. You can view this region in either 'Region in detail' or 'Region Overview' (click on 'Region overview' in the left hand menu for this view.)*

There are 4 gold genes in the region: FRY-AS1, FRY, ZAR1L, and BRCA2.

***(b)** Turn on the decipher track by clicking on 'Configure this page'. Either search the configuration panel for 'decipher', or click on 'Germline variation' and then select the 'decipher' track. Close the menu.*

There are 7 patient records in this region. Click on the red or blue bars to see a pop-up box of information about the record. Click on the Link (for example 00001604) to jump to Decipher. Red and blue tracks refer to 'loss' and 'gain' in sequence, respectively.

***(c)** Zoom out using the slide, or click and drag your mouse around a region of interest in the chromosome at the top. Yes, structural variation: nsv436168 covers band q13.1 and part of band q12.3.*

Answer- Exercise 4: Gene Regulation

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

In the location tab, configure the page and turn on the "Regulatory features" track under the Regulation menu.

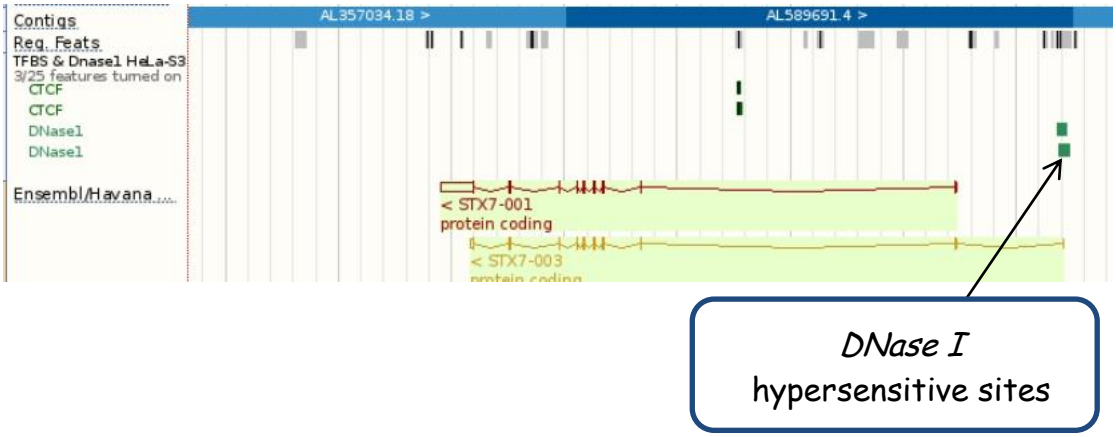
You can click on either Reg. Feats for a summary of all regulatory features (shaded grey boxes) or you can choose cell specific regulatory features tracks (e.g CH4, HeLa-S3).

There are quite a few of regulatory features mapping to the STX7 transcripts, including the 5' end.

(b) *In the location tab, 'Configure this page' and click on Open chromatin & TFBS (under Regulation). Filter by 'Open Chromatin' and click on FAIRE for HeLa-S3 cells.*

Now choose on 'Histones & polymerases' at the left hand menu and filter by 'Polymerase'. PolII is already turned on for HeLa cells, and PolIII is not available. Close the menu.

As in the picture below, there are two DNase I hypersensitive sites found at the 5' end of STX7 (a reverse-stranded gene). Remember you can drag tracks to reorder them on the page for easier comparison.



VIII) EXERCISES COMPARATIVE GENOMICS

Exercise 1 - Orthologues, paralogues and genetrees

Find the human *BRAF* gene.

(a) How many orthologues are predicted for this gene in primates? Look in the orthologues page.

Note the Target %id and Query %id.
How much sequence identity does the *Tarsius syrichta* protein have to the human one? Click on the **Alignment** link next to the **Ensembl identifier** column to view a protein alignment in Clustal format.

(b) Go to the orthologue in marmoset. Is there a genomic alignment between marmoset and human? Are there genes for both species in this region?

(c) Go to the location tab for the marmoset BRAF gene, and click on Multi-species view at the left. Can you turn on the alignment between human and marmoset? Are the BRAF genes viewable for both species?

Exercise 2 – Zebrafish orthologues

Go to www.ensembl.org to find the *dbh* gene on the zebrafish genome.

(a) Go to the Location page for this gene. View the Alignments (image) and Alignments (text) for the 5 teleost fish. Which fish genomes are represented in the alignment? Do all the fish show a gene in these alignments?

(b) Export the alignments (as Clustal).

(c) Go to the 'Region in Detail' view and turn on the tracks for:

- 5 teleost fish EPO
- Conservation score for 5 teleost fish EPO
- Constrained elements for 5 teleost fish EPO

What is the difference between the '5 teleost fish EPO' track and the 'Constrained elements'? What do most of the constrained element blocks match up to?

Can you find more information on how the constrained elements track was generated?

Exercise 3 – Synteny

Go to www.ensembl.org

Find the **Rhodopsin (RHO) gene** for Human. Go to the **Location tab**.

(a) Click 'Synteny' at the left. Are there any syntenic regions Dog? If so, which chromosomes are shown in this view?

(b) Stay in the Synteny view. Is there a homologue in dog for human RHO? Are there more genes in this syntenic block with homologues?

ANSWERS COMPARATIVE GENOMICS

1 (a) *Go to www.ensembl.org , choose human and search for BRAF.* Click on gene and then on '7:140424943-140624564:-1' below 'BRAF [Ensembl/Havana merge: ENSG00000157764]'

On the gene tab, click on "Orthologues" at the left side of the page to see all the 55 *orthologous genes. There are 9 orthologues in the primates.*

The BRAF protein in Tarsier has a % identity of 69%, (the human gene has a % id of 62%). Note the difference in Target and Query % ID reflects the different protein lengths.

(b) *Go to the orthologues page and click on the marmoset orthologue to open the gene tab.*

Click 'Genomic alignments' at the left to see the pairwwse alignment with human. *You will have to choose the alignment in the menu above the sequence.*

The red sequence is present in exons, so there is a gene in both species in this region.

(c) *Go to location and then 'Multi-species view'. 'Select species' at the left and select 'Marmoset'. You should see an alignment between the human BRAF gene region and the BRAF gene region for the marmoset.*

2 (a) *Start in the Location tab (region in detail) for dbh (ENSDARG00000069446). Click on 'Alignments (Image)' at the left, and select the '5 teleost fish EPO' alignment in the pull-down menu in the view. The zebrafish, stickleback, medaka,*

takifugu, and tetradon are shown in this region. All the species show a gene in the aligned region. This can also be seen in the Alignments (text) page.

(b) *You can export the alignments from either 'Alignments (images)' or 'Alignments (text)' menus in the Location tab. Click on the blue 'export data' tab at the left, and choose 'Clustal' from the list.*

(c) *Click on 'Region in detail' in the left hand menu. Turn on the comparative tracks with 'Configure this page'; the three tracks are in the 'Multiple alignments' menu.*

The '5 teleost fish EPO' track just shows the whole region for the dbh gene could be aligned. The 'Constrained elements' track shows where in the alignment the conserved sequence is (and it matches up to exons, which tend to be highly conserved).

Click on the Track name and the 'i' (information button) to read more.

3 (a) *Change the species to Dog below the image. Yes, there are multiple syntenic regions to human chromosome 3 on Dog. Human chromosome 3 is in the center of this view. Dog chromosomes 6, 20, 23, 31, 33, and 34 have syntenic regions to human chromosome 3.*

(b) *Scroll down to the bottom of the page. The homologue of human RHO is OPDS_CANFA. Click '15 downstream genes' to compare the genes between human and dog in this syntenic block.*

IX) 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.

e! Data Upload: If you want to try data upload to Ensembl, our exercises are here:

http://www.ensembl.org/info/website/tutorials/

Other help:

The Ensembl Glossary:
http://www.ensembl.org/Help/Glossary

Ensembl FAQs:
http://www.ensembl.org/Help/Faq

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).*