Browsing Genomes with Ensembl

www.ensembl.org
www.ensemblgenomes.org

Exercise Answers v82

http://www.ebi.ac.uk/~emily/Workshops/2015/Denver/

UC Denver – 14th October 2015
<table>
<thead>
<tr>
<th>TABLE OF CONTENTS</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Exploring the Ensembl genome browser</td>
<td>3</td>
</tr>
<tr>
<td>Ensembl species</td>
<td>3</td>
</tr>
<tr>
<td>Region in Detail</td>
<td>4</td>
</tr>
<tr>
<td>Genes and Transcripts</td>
<td>7</td>
</tr>
<tr>
<td>BioMart</td>
<td>10</td>
</tr>
<tr>
<td>Variation</td>
<td>14</td>
</tr>
<tr>
<td>Exploring variants in Ensembl</td>
<td>14</td>
</tr>
<tr>
<td>VEP</td>
<td>16</td>
</tr>
<tr>
<td>Regulation</td>
<td>18</td>
</tr>
<tr>
<td>Comparative Genomics</td>
<td>20</td>
</tr>
<tr>
<td>Advanced exercise</td>
<td>24</td>
</tr>
</tbody>
</table>
Exploring the Ensembl genome browser

Ensembl species

Exercise S1 – Panda

(a) Select Panda from the drop down species list, or click on View full list of all Ensembl species, then choose Panda from the list.

The assembly is ailMel1 or GCA 000004335.1

(b) Click on More information and statistics. Statistics are shown in the tables on the left.

The length of the genome is 2,245,312,831 bp.
There are 19,343 coding genes.

Exercise S2 – Zebrafish

(a) Click on Zebrafish on the front page of Ensembl to go to the species homepage. News is in the top right.

What’s new in Zebrafish release 82:
- Zebrafish developmental stage RNASeq data set
- Structural variants
- Phenotype data updates

(b) Under Other assemblies two previous assembly names and the releases you can find them in is listed.

Assembly Zv9 is available in the archived release 79 and assembly Zv8 is available in the archived release 59.

Extra Exercise S3 – Mosquitos

(a) Go to metazoa.ensembl.org. Open the drop down list or click on View full list of all Ensembl Metazoa species. Type Anopheles into the filter box in the top left.

There are two Anopheles species: Anopheles gambiae and Anopheles darlingi.

(b) Click on Anopheles gambiae, then on More information and statistics.

The genome was revised in April 2014.
**Extra Exercise S4 – Bacteria**

Go to [bacteria.ensembl.org](http://bacteria.ensembl.org) and start to type the name *Belliella baltica* into the search species box. It will autocomplete, allowing you to select *Belliella baltica DSM 15883, (TaxID 866536)* from the drop-down list. Click on More information and statistics. *Belliella baltica* has 3,680 coding genes and 53 non-coding.

**Region in Detail**

**Exercise RD1 – Exploring a genomic region in human**

(a) Go to the Ensembl homepage ([http://www.ensembl.org/](http://www.ensembl.org/)).

Select *Search: Human* and type **13:31937000-32633000** in the text box (or alternatively leave the Search drop-down list like it is and type **human 13:31937000-32633000** in the text box). Click **Go**.

This genomic region is located on cytogenetic band q13.1. It is made up of eight contigs, indicated by the alternating light and dark blue coloured bars in the Contigs track. Note that KF455761.1 is a tiny contig that splits AL137143.8 in two.

(b) Draw with your mouse a box encompassing the *BRCA2* transcripts. Click on **Jump to region** in the pop-up menu.

(c) Click **Configure this page** in the side menu (or on the cog wheel icon in the top left hand side of the bottom image).

Type **tilepath** in the **Find a track** text box. Select **Tilepath**.

Click on the (i) button to find out more

The tilepath track shows the BAC clones that the assembly was based upon.

Save and close the new configuration by clicking on ✓ (or anywhere outside the pop-up window).

There is not just one clone that contains the complete *BRCA2* gene. The BAC clone RP11-37E23 contains most of the gene, but not its very 3' end (contained in RP11-298P3). This was
reflected on the two contigs that make up the entire BRCA2 gene (the Contigs track is on by default). You may find this easier to see if you highlight the 3’ exon on BRCA2.

(d) Click Share this page in the side menu.

Select the link and copy.

Get your neighbour’s email address and compose an email to them, paste the link in and send the message.

When you receive the link from them, open the email and click on your link. You should be able to view the page with the new configuration and data tracks they have added to in the Location tab. You might see differences where they specified a slightly different region to you, or where they have added different tracks.

(e) Click Export data in the side menu. Leave the default parameters as they are.
Click Next>.
Click on Text.

Note that the sequence has a header that provides information about the genome assembly (GRCh38), the chromosome, the start and end coordinates and the strand. For example:

>13 dna:chromosome
chromosome:GRCh38:13:32311910:32405865:1

(f) Click Configure this page in the side menu.
Click Reset configuration.
Click ✓.

Extra Exercise RD2 – Exploring assembly exceptions in human

(a) Go to the Ensembl homepage (http://www.ensembl.org/).

Select Search: Human and type 21:32630000-32870000 in the text box (or alternatively leave the Search drop-down list like it is and type human 21:32630000-32870000 in the text box). Click Go.
You will see a red highlighted region in the middle of this region. Click on the thin dark red bar in any of the three views to see the label **CHR_HSCHR21_3_CTG1_1:32769079-32843731**. Click on **What are assembly exceptions?** to open a new window which explains assembly exceptions.

(b) Assembly exceptions are marked in the chromosome view at the top.

There are seven haplotypes on chromosome 21.

(c) Another option in the drop-down is **Compare with reference**. Click on this.

Scroll down the page to see the comparison between the haplotype and primary assembly. Aligned sequences are highlighted in pink and linked together in green.

The assembly exception **CHR_HSCHR21_3_CTG1_1** contains an extra region compared to the primary assembly.
Genes and Transcripts

Exercise GT1 – Exploring the human *MYH9* gene

(a) Go to the Ensembl homepage ([http://www.ensembl.org](http://www.ensembl.org)).

Select Search: Human and type *MYH9*. Click Go.

Click on either the Ensembl ID ENSG00000100345 or the HGNC official gene name *MYH9*.

- Chromosome 22 on the reverse strand.
- Ensembl has 11 transcripts annotated for this gene, of which three are protein coding.
- The longest transcript is MYH9-001 and it codes for a protein of 1,960 amino acids
- MYH9-001 is the best quality transcript, as it has a CCDS associated with it, is TSL:1 and is Golden.

(b) These are some of the phenotypes associated to *MYH9* according to MIM: autosomal dominant deafness, Epstein syndrome, and Fechtner syndrome. Click on the records for more information.

(c) Click on ENST00000216181

- It has 41 exons. This is shown in the Transcript summary or in the left hand side menu Exons.
- Click on the Exons link in this 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 if you click on the cDNA link in the left side menu.
P35579 from UniProt/Swiss-Prot matches the translation of the Ensembl transcript. Click on P35579 to go to UniProtKB, or click align for the alignment.

The Gene Ontology project (http://www.geneontology.org/) maps terms to a protein in three classes: biological process, cellular component, and molecular function. Meiotic spindle organisation, cell morphogenesis, and cytokinesis are some of the roles 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. Expression analysis with any of these probesets would reveal information about the transcript. Hint: this information can sometimes be found in the ArrayExpress Atlas: www.ebi.ac.uk/arrayexpress/

**Exercise GT2 – Finding a gene associated with a phenotype**

(a) Start at the Ensembl homepage (http://www.ensembl.org).

Type phenylketonuria into the search box then click Go. Choose Gene from the left hand menu.
The gene associated with this disorder is PAH, phenylalanine hydroxylase, ENSG00000171759.

(b) If the transcript table is hidden, click on Show transcript table to see it.
There are four protein coding transcripts.
Click on Transcript comparison in the left hand menu. Click on Select transcripts. Either select all the transcripts labelled protein coding one-by-one, or click on the drop down and select Protein coding. Close the menu.

(c) Click on External references.
The MIM disease ID is 261600.
Extra Exercise GT3 – Exploring a plant gene (Vitis vinifera, grape)

(a) Go to http://plants.ensembl.org/index.html

Select *Vitis vinifera* from the drop down menu All genomes – select a species or click on View full list of all Ensembl Plants species and then choose *V. vinifera*.

Type *SEP3* and click on the gene name link *SEP3 [VIT_01s0010g03900]*.

Click on GO: biological process in the side menu.

There are nine terms listed including GO:0006351, transcription, DNA-templated, and GO:0006355, regulation of transcription, DNA-templated.

(b) Click on the transcript named *Vv01s0010g03900.t01* (or on the Transcript tab). Click on Exons in the left hand menu.

There are eight exons, of which exon 8 is longest with 303 bp, of which 13 are coding.

(c) Click on either Protein Summary or Domains & features in the left hand menu to see graphically or as a table respectively.

A TF_MADSbox is identified by six domain prediction methods. A TF_Kbox domain is identified by two. Two coiled-coils are identified by one.
**BioMart**

**Exercise BM1 – Finding genes by protein domain**

As with all BioMart queries you must select the dataset, set your filters (input) and define your attributes (desired output). 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](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 Limit to genes, choosing with Transmembrane domains from the drop-down and then Only. Clicking on Count should reveal that you have filtered the dataset down to 425 genes.

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.

The output will display the Ensembl gene ID, Ensembl Transcript ID and Associated gene names of all proteins with a transmembrane domain on mouse chromosome 9. If you prefer, you can also export as an Excel sheet by using the Export all results to XLS option.

**Exercise BM2 – Export homologues**

Click New.  
Choose the ENSEMBL Genes 82 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 Input external references ID list 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.

**Exercise BM3 – Convert IDs**

Click New.
Choose the ENSEMBL Genes 82 database.
Choose the Homo sapiens genes (GRCh38) dataset.

Click on Filters in the left panel.
Expand the GENE section by clicking on the + box.
Select Input external references ID list - 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.

Count shows 11 genes (remember one gene may have multiple splice variants coding for different proteins, that is the reason why these 29 proteins do not correspond to 29 genes).

Click on Attributes in the left panel.
Select the Features attributes page.
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.

**Extra Exercise BM4 – Export structural variants**

(a) Choose Ensembl Variation 82 and Homo sapiens Structural Variation (GRCh38).
**Filters:** Region: Chromosome 1, Base pair start: 130408, Base pair end: 210597
Count shows 71 out of 4,358,541 structural variants.

Attributes: Structural Variation (SV) Information: DGVa Study Accession and Source Name
Structural Variation (SV) Location: Chromosome name, Sequence region start (bp) and Sequence region end (bp).

(b) Choose Ensembl Variation 82 and Homo sapiens Short Variation (SNPs and indels) (GRCh38).

Filters: Filter by Variation name enter: rs1801500, rs1801368

Attributes: Variation Name, Variant Alleles, Phenotype description and 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.

Extra Exercise BM5 – Find genes associated with array probes

(a) Click New.
Choose the ENSEMBL Genes 82 database.
Choose the Homo sapiens genes (GRCh38) dataset.

Click on Filters in the left panel.
Expand the GENE section by clicking on the + box.
Select Input microarray probes/probesets ID list - Affy hg u133 plus 2 probeset ID(s) and enter the list of probeset IDs in the text box (either comma separated or as a list).

Count shows 24 genes match this list of probesets.

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 24 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 one possible transcript in a gene (the most 5’ coordinates for upstream flanking).

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 ORTHOLOGS section by clicking on the + box.
Select Mouse Ensembl Gene ID, Mouse Chromosome Name, Mouse Chr Start (bp) and Mouse Chr End (bp).

Click the Results button on the toolbar.
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.
**Variation**

**Exploring variants in Ensembl**

**Exercise V1 – Human population genetics and phenotype data**

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

Once you’re in the Variation tab, click on the Genes and regulation link or icon.

This SNP is found in three transcripts (ENST00000326965, ENST00000338313, and ENST00000367066).

(b) Click on Population genetics at the left of the variation tab. (Or, click on Explore this variation at the left and click the Population genetics icon.)

In Yoruba (HapMap-YRI population), the least frequent genotype is CC at the frequency of 9.7%. This is also the least frequent genotype in other populations (to find out what the three letter population are, hover over the names).

(c) Click on Phylogenetic context.

The ancestral allele is T and it’s inferred from the alignment in primates.

Select the 39 eutherian mammals EPO LOW COVERAGE alignment and click on Go.

A region containing the SNP (highlighted in red and placed in the centre) and its flanking sequence are displayed. The T allele is conserved in all but four of the 39 eutherian mammals displayed. Note that two species have no alignment in that region and many other species have no variation database.

(d) Click Phenotype Data at the left of the Variation page.

This variation is associated with multiple sclerosis and coeliac. There are known risk alleles for both multiple sclerosis and coeliac and the corresponding P values are provided. The allele A is associated with coeliac disease. Note that the alleles reported by Ensembl are T/C. Ensembl reports alleles on the
forward strand. This suggests that A was reported on the reverse strand in the original paper. Similarly, one of the alleles reported for Multiple sclerosis is G.

**Exercise V2 – Exploring a SNP in human**

(a) Go to the Ensembl homepage ([http://www.ensembl.org/](http://www.ensembl.org/)).

Type `rs1801133` in the Search box, then click Go. Click on `rs1801133`.

(b) Click on Genes and Regulation in the side menu (or the Genes and Regulation icon).

No, rs1801133 is Missense variant in four `MTHFR` transcripts. It's a downstream gene variant of ENST00000418034.

(c) In Ensembl, the alleles of rs1801133 are given as G/A because these are the alleles in the forward strand of the genome. In the literature and in dbSNP, the alleles are given as C/T because the `MTHFR` gene is located on the reverse strand. The alleles in the actual gene and transcript sequences are C/T.

(d) Click on Population genetics in the side menu.

In all populations but two (from the 1000 genomes and HapMap projects), the allele G is the major one. The two exceptions are: CLM (Colombian in Medelin; 1000 Genomes), HCB (Han Chinese in Beijing, China; HapMap).

(e) Click on Phenotype Data in the left hand side menu.

The specific studies where the association was originally described is given in the Phenotype Data table. Links between rs1801133 and homocysteine levels were described in two papers. Click on the pubmed IDs `pubmed:20031578` and `pubmed:23824729` for more details.

(f) Click on Phylogenetic Context in the side menu.

Select Alignment: 8 primates EPO and click Go.

Gorilla, orangutan, chimp, macaque, olive baboon and marmoset all have a G in this position. Please note that there is
no variation database for gorilla, olive baboon or marmoset though.

**Extra Exercise V3 – Exploring a SNP in mouse**

(a) Go to [www.ensembl.org](http://www.ensembl.org), type **rs29522348** in the search box. Click on **rs29522348 (Mouse Variation)**. SNP rs29522348 is located on 17:73924993. In Ensembl, its alleles are provided as in the forward strand.

(b) Click on **HGVS names** to reveal information about HGVS nomenclature. This SNP has got three HGVS names, one at the genomic DNA level (17:g.73924993C>T), one at the transcript level (ENSMUST00000024866.4:c.721G>A) and one at the protein level (ENSMUSP00000024866.4:p.Val241Ile).

(c) In Ensembl, the allele that is present in the reference genome assembly is always put first (C is the allele for the reference mouse genome, strain C57BL/6J).

(d) Click on **Sample genotypes** is the left hand side menu. In the summary of genotypes by population, click on **Show** for PERLEGEN:MM_PANEL2, or search for the two strain names. There are indeed differences between the genotypes reported in those two different strains. The genotype reported in NOD/LTJ is TT whereas in BALB/cByJ the genotype is CC.

**VEP**

**Exercise V4 – VEP**

Go to [www.ensembl.org](http://www.ensembl.org) and click on the link tools at the top of the page. Currently there are 5 tools listed in that page. Click on **Variant Effect Predictor** and enter the three variants as below:

7 117530985117530985G/A
7 117531038117531038T/C
7 117531068117531068T/C
Note: Variation data input can be done in a variety of formats. See more details here [http://www.ensembl.org/info/docs/variation/vep/vep_formats.html](http://www.ensembl.org/info/docs/variation/vep/vep_formats.html)

Click Run.

When your job is listed as Done, click View Results.

You will get a table with the consequence terms from the Sequence Ontology project [http://www.sequenceontology.org/](http://www.sequenceontology.org/) (i.e. synonymous, missense, downstream, intronic, 5’ UTR, 3’ UTR, etc) provided by VEP for the listed SNPs. You can also upload the VEP results as a track and view them on Location pages in Ensembl. SIFT and PolyPhen are available for missense SNPs only. For two of the entered positions, the variations have been predicted to be probably damaging/deleterious (coordinate 117531038) and benign/tolerated (coordinate 117531068). All the three variations have been already described and are known as in rs1800078, rs1800077 and rs35516286 in dbSNP and other sources (databases, literature, etc).
Regulation

Exercise R1 – Gene regulation: Human STX7

(a) Search for human STX7 from the home page. Click on Location in 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. The Regulatory features are turned on by default in the Region in detail view.

There are five enhancers predicted in the region of STX7, two near the 3’ end, two near the middle and one near the 5’ end.

Click on the Reg. Feats track name to jump to an article explaining the underlying data. Click and drag the Reg. Feats track next to the Genes (Merged Ensembl/Havana) track to better compare where the Regulatory features are in the gene.

(b) See the legend below the Region in detail view to find the predicted enhancers are coloured in yellow. Two appear in the HUVEC cell type only (out of the three cells chosen).

(c) Configure this page and click on Open chromatin & TFBS. Turn on both peaks and signal for DNase 1 in HeLa-S3 cells (the boxes in this configure this page window will turn blue. For more information on how to select and view the supporting data, click on Show tutorial in the pop up window). Close the menu.

There’s a DNase 1 hypersensitive site in the 5’ exon of STX7. Click on the coloured block to find out that the DNase1 enriched sites in HeLa-S3 cells come from the ENCODE project.

(d) Configure this page and click on Histones & polymerases. Change the Filter by menu from All classes to Histone. Select the all the histone modifications available for HeLa cells (some of them might be on by default). Save and close the menu.

H3K4me3, H3K9ac, H3K27ac H3K36me3 and H3K4me2 sites have been found in the 5’ region of STX7 in HeLa-S3 cells.
(e) Click on configure this page and choose the DNA Methylation menu. Turn on the track Jurkat RRBS ENCODE. Save and close the menu.

Some CpG sites at the 5’ end of STX7 are not highly methylated (note the yellow bars) whilst others are (blue bars). Yellow, green, and blue bars represent unmethylated, intermediately methylated, and methylated regions, respectively. For more information on human DNA methylation tracks, see: www.ensembl.org/info/docs/funcgen/index.html

**Extra Exercise R2 – Regulatory features in human**

(a) Search for human gene APOE from the home page. Click on Location in the search results. Hold down shift and drag out a box in the middle display to zoom out. The gene is positive stranded so look for features at the left hand side. Click on the features to get their IDs.

There is a pink promoter flank at the 5’ end of APOE. Click on it to get a pop-up with its ID: ENSR00000347288. It contains five black lines, which are transcription factor binding motifs. The pop-up reveals that they are binding sites for SP1.

(b) Click on the stable ID ENSR00000347288. Click on the ID to go to the regulation tab.

ENSR00001636517 is active in 6/18 cell types studied: A549, H1ESC, HMEC, HepG2, IMR90 and K562

(c) Click on Details by cell type, then open the Select cells menu. Choose ALL ON to select all cell types, then close the menu. Open the Select evidence menu and choose SP1 only, then close.

SP1 binding is only observed in H1ESC cells, at both of the binding motifs identified.

Open Select evidence again and choose all the transcription factors, then close. You may find it easier to see if you also go into Select cells and turn the cell types ALL OFF, then turn on H1ESC only. TAF1 binds covering both SP1 binding sites. TAF7 and USF1 bind only to the left SP1 binding site.
Comparative Genomics

Exercise C1 – Orthologues, paralogues and gene trees for the human *BRAF* gene.

(a) Go to [www.ensembl.org](http://www.ensembl.org), choose human and search for *BRAF*. Click through to the Gene tab view.

On the gene tab, click on Orthologues at the left side of the page to see all the orthologous genes.

There are orthologues in 9 primates.

The percentage of identical amino acids in the Tarsier protein (the orthologue) compared with the gene of interest. i.e. human *BRAF* (the target species/gene) is 69%. This is known as the Target %ID. The identity of the gene of interest (human *BRAF*) when compared with the orthologue (Tarsier *BRAF*, the query species/gene) is 62% (the query %ID).

Note the difference in the values of the Target and Query %ID reflects the different protein lengths for the human and tarsier *BRAF* genes.

(b) There is more than one way to get to the answer.

Option 1: Go to the orthologues page and click on the marmoset orthologue to open the gene tab. Click Genomic alignments at the left. Then select Alignment: Human (Homo sapiens) – lastz and click Go. Choose Block 1 to get the largest block of aligned sequence.

The red sequence is present in exons, so there is a gene in both species in this region. You can find where the start and stop codons are located if you configure this page and select START/STOP codons.

Option 2: Go to location tab of the marmoset *BRAF* gene and then click on Region Comparison view at the left. Click on Select species or regions at the left and click on the + to select Human (Homo sapiens) – lastz then save and close. You should see an alignment between the human *BRAF* gene region and the *BRAF* gene region for the marmoset.
(Note: To see a blue line connecting homologous genes in the Region Comparison view page, click on configure this page and under Comparative features select join genes. Zoom out on the location view to see blue lines connecting all the homologous genes between marmoset and human genes in that region).

Exercise C2 – Zebrafish orthologues

(a) Start in the Location tab (region in detail) for sardh. Click on Alignments (Image) at the left, and select the 11 teleost fish EPO alignment in the pull-down menu in the view.

The zebrafish, cave fish, cod, tilapia, Amazon molly, platyfish, spotted gar, stickleback, medaka, fugu, 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 (the exons are highlighted in red).

(b) You can export the alignments from either the Alignments (text) or Alignments (image) pages in the Location tab. Click on the blue Download alignments button at the top of the text page, or the icon at the top of the image, and choose ClustalW from the list.

(c) Click on Region in detail in the left hand menu. Turn on the multiple alignment, constrained elements and conservation score for 11 teleost fish EPO tracks, all under the Comparative genomics menu by configuring the page.

The 11 teleost fish EPO track just shows that the whole region for the sardh gene can be aligned among those eleven species of fish. The Constrained elements and Conservation score tracks show the conserved sequence is located where in the alignment.

Higher conservation regions match up with exonic regions (exons tend to be highly conserved) of the gene.

Hover over the the Track name and the 🌐 (information button) to read more about constrained elements (or any other data track).


**Extra Exercise C3 – Synteny**

(a) Change the species to **dog** next to the image.  
Yes, there are multiple syntenic regions in dog to human chromosome 3, which is in the centre 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.  
There is a homologue in dog of human RHO. Click **Centre on gene RHO** to compare the genes between human and dog in this syntenic block.

**Extra Exercise C4 – Whole genome alignments**

(a) Go to the Ensembl homepage ([http://www.ensembl.org/](http://www.ensembl.org/)).  
Select **Search: Human** and type **brca2** in the search box.  
Click **Go**.  
Click on 13:32315474-32400266:1 below **BRCA2 (Human Gene)**.

You may want to turn off all tracks that you added to the display in the previous exercises as follows:  
Click **Configure this page** in the side menu.  
Click **Reset configuration**.  
SAVE and close.

(b) Click **Configure this page** in the side menu  
Click on **BLASTZ/LASTZ alignments** under the **Comparative genomics** menu. Select **Chicken (Gallus gallus) - LASTZ_NET – Normal**, **Chimpanzee (Pan troglodytes) – LASTZ_NET – Normal**, **Mouse (Mus musculus) – LASTZ_NET – Normal** and **Platypus (Ornithorhynchus anatinus) - LASTZ_NET – Normal**.  
SAVE and close.

Yes, the degree of conservation does reflect the evolutionary relationship between human and the other species; the highest degree of conservation is found in chimp, followed by mouse, platypus and chicken, respectively. Especially the exonic sequences of **BRCA2** seem to be highly conserved between the various species, which is what is to be expected.
because these are supposed to be under higher selection pressure than intronic and intergenic sequences.

(c) Click Configure this page in the side menu. Click on Conservation regions under the Comparative genomics menu. Select Conservation score for 39 eutherian mammals EPO_LOW_COVERAGE, Conservation score for 23 amniota vertebrates Pecan and Constrained elements for 23 amniota vertebrates Pecan. SAVE and close.

Both the Conservation score and Constrained elements tracks largely correspond with the data seen in the pairwise alignment tracks; all exons of the BRCA2 gene show a high degree of conservation (Note the UTRs which are not conserved).

(d) Click on a constrained element (brown block). Click on View alignments (text) in the pop-up menu. Click Configure this page in the side menu. Select Show conservation regions. SAVE and close.

The conserved regions will be shown in light blue.
**Advanced exercise**

**Methylation data in human**

(a) Go to the Ensembl homepage (http://www.ensembl.org/). Select Search: Human and type PDHA2 in the for text box. Click Go. Click on 4:95840019-95841474:1. Zoom out one step, so that the 5kb region around the PDHA2 gene is shown.

You may want to turn off all tracks that you added to the display in the previous exercises as follows:

Click Configure this page in the side menu.
Click Reset configuration.
SAVE and close.

(b) Click Configure this page in the side menu. Type cpg in the Find a track box. Select CpG islands. SAVE and close.

No CpG islands are shown. As for the inclusion of CpG islands into the Ensembl database for human a minimum length of 400 bp is required, the reason for this could be that the CpG islands in the PDHA2 gene are shorter than 400 bp. However, there is a %GC track, which shows that the region that comprises the 5’ part of the PDHA2 gene and the region directly upstream of the gene has a high %GC (the red line in the %GC track indicates 50% GC). It is difficult / impossible to distinguish individual CpG islands in this track, though.

(c) Click Export data in the side menu. Click Next>. Click on Text. Select and copy the sequence. Go to http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html. Paste the sequence into the text box. Click Run. CpGPlot does confirm the existence of two CpG islands in the PDHA2 gene region of lengths 200 and 263 bp, respectively.
So, it is indeed because of their length being less than 400 bp that these CpG islands are not present in the Ensembl database.

(d) The genomic coordinates of your CpG islands are the start coordinates of your region of interest (found at the top of your exported FASTA) plus the coordinates of the islands within that region (from EMBOSS). In my case this is:

First island: start = 95839291 + 734 = 95840025
end = 95839291 + 933 = 95840224

Second island: start = 95839291 + 1058 = 95840349
end = 95839291 +1320 = 95840611

This gives coordinates for my CpG islands in BED format as:
chr4 95840025 95840224 cpg_island_1
chr4 95840349 95840611 cpg_island_2

Click Add your data in the side menu (Note that if you have previously uploaded data to Ensembl, this box will say Manage your data instead).
Click on Upload Data.
Type CpG islands in the Name for this upload (optional) box.
Select Data format: BED.
Copy the BED formatted data into the Paste file box.
Click Upload.
Click on Go to nearest region with data: 4:95790125-95890125.
   The two CpG islands should now be shown on the Region in detail page. They should coincide with the regions of high %GC.

Zoom in on the two CpG islands.

To display the names of the CpG islands:

Hover over the CpG islands track name.
Hover over the icon of the cog-wheel.
Select Labels.

(e) Drag your CpG islands track so that it is next to the %GC track.
Click Share this page in the side menu.
Select the link and copy.  
Paste into your internet browser to view.

(f) There is a lilac TFBS at the 5′ end of PDHA2.  
Click on the feature then the ID ENSR00001245785 to get to the regulatory tab.  
The TFBS is active in H1ESC cells.  
Click on Details by cell type, then Select cells and choose H1ESC and close, then Select evidence and choose ALL ON and close.  
This region has DNase sensitivity, Rad21 binding and H3K27me3 and H3K36me3 histone modifications.

(g) Click on Configure this page, then select RNaseq models. Turn on the BAM files for all the tissues in Coverage only.  
You will see histograms of RNaseq coverage for each of the tissues. The largest number is for the merged read. For the tissue-specific read, Testes have a higher peak than all the other tissues. There are also wider peaks in the Testes track that cover the whole gene, whereas other tissues only have a peak at the 3′ end of PDHA2.

(h) Click on Configure this page, then select Comparative genomics.  
Turn on the tracks for the Constrained elements for 39 eutherian mammals and Conservation score for 39 eutherian mammals.  
The region of the gene itself has high GERP scores, indicated by constrained elements over most of the gene. There is no apparent difference in the conservation score between the CpG islands and their flanking regions.

(i) Click on the Transcript Tab, Transcript: PDAH2-001 and select GO table.  
There are ten terms in the table, the first being GO:00060606, pyruvate metabolic process.

To export the list use BioMart.  
Click on BioMart in the top bar.  
Choose Ensembl Genes 80 and Homo sapiens genes (GRCh38).

Click on Filters.  
Open the menu for GENE ONTOLOGY.  
Select GO Term Accession and put GO:000606 into the box.
Click on Attributes.
Choose Sequences.
Expand SEQUENCES and select Unspliced (Gene).
Expand Header information and deselect Ensembl Transcript ID.

Click Results.
You can export these results if you wish.

(j) Go to the REST API documentation page at http://beta.rest.ensembl.org/documentation.
Click on GET sequence/id/:id to get the documentation for this command.

You will need the stable ID of PDHA2, go to the browser page to find that it is ENSG00000163114.

Use the documentation to construct a URL in the correct form, ie:
http://beta.rest.ensembl.org/sequence/id/:id?format=fasta

Add the ID to the URL to create:
http://beta.rest.ensembl.org/sequence/id/ENSG00000163114?format=fasta

This URL will give you the sequence.