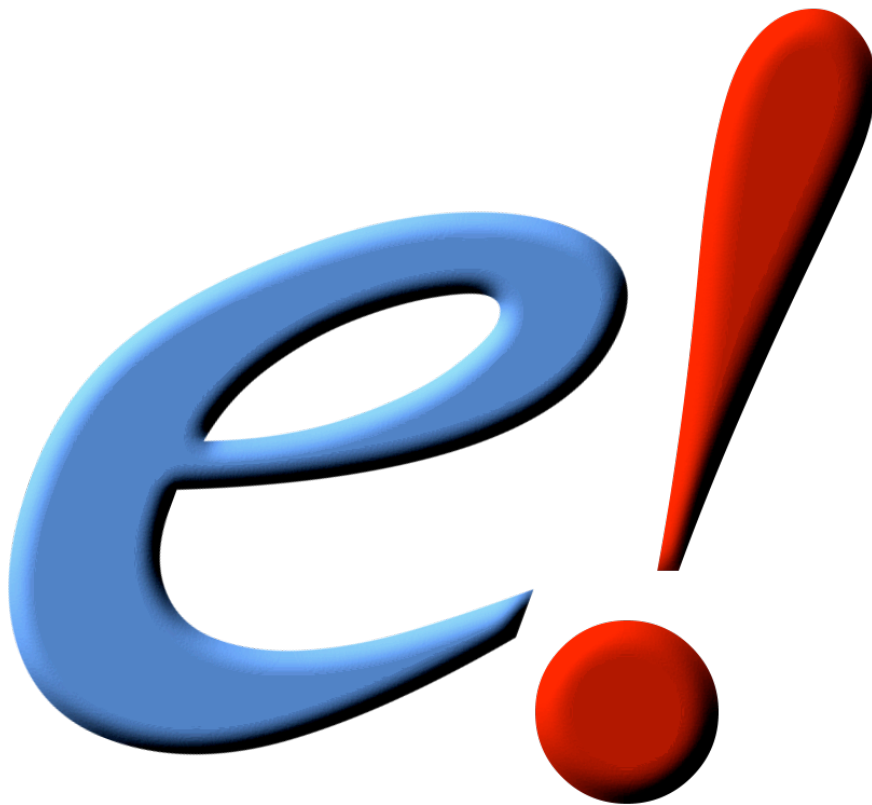


---

**Computational & Comparative Genomics  
Cold Spring Harbor, New York, United States  
9-15 November 2011**

---

**BROWSING GENES AND GENOMES  
WITH ENSEMBL**



---

**EXERCISES AND ANSWERS**

---

Note: These exercises are based on Ensembl version 64 (September 2011). After in future a new version has gone live, version 64 will still be available at <http://e64.ensembl.org>. If your answer doesn't correspond with the given answer, please consult the instructor.

---

## BROWSER

---

### Exercise 1 – Exploring a gene

(a) Find the human *F9* (Coagulation factor IX) gene. On which chromosome and which strand of the genome is this gene located? How many transcripts (splice variants) have been annotated for it?

(b) What is the longest transcript? How long is the protein it encodes? Has this transcript been annotated automatically (by Ensembl) or manually (by Havana)? How many exons does it have? Are any of the exons completely or partially untranslated?

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

(d) Is it possible to monitor expression of ENST00000218099 with the CodeLink microarray? If so, can it also be used to monitor expression of the other two transcripts?

(e) In which part (i.e. the N-terminal or C-terminal half) of the protein encoded by ENST00000218099 does its peptidase activity reside?

(f) How many non-synonymous coding variants have been discovered for the protein encoded by ENST00000218099?

(g) Is there a mouse ortholog predicted for the human *F9* gene?

(h) If you have yourself a gene of interest, explore what information Ensembl displays about it!

---

### Answer

(a)

☞ Go to the Ensembl homepage (<http://www.ensembl.org>).

☞ Select 'Search: Human' and type 'f9' or 'factor IX' in the 'for' text box.

☞ Click [Go].

☞ Click on 'Gene' on the page with search results.

☞ Click on 'Human'.

☞ Click on 'F9 [ Ensembl/Havana merge: ENSG00000101981 ]'.

The human *F9* gene is located on the X chromosome on the forward strand. There have been three transcripts annotated for this gene, ENST00000218099 (F9-001), ENST00000394090 (F9-201) and ENST00000479617 (F9-002).

(b)

The longest transcript is ENST00000218099 (F9-001). The length of this transcript is 2780 base pairs and the length of the encoded protein is 461 amino acids.

☞ Click on the transcript in the 'Gene Summary' display.

It is an Ensembl/Havana merge transcript that has been annotated both automatically and manually. This can also be seen from the fact that the transcript is golden colored.

☞ Click on the Ensembl Transcript ID 'ENST00000218099' in the list of transcripts.

It has eight exons.

☞ Click on 'Sequence - Exons' in the side menu.

The first and last exon are partially untranslated (sequence shown in purple).

(c)

☞ Click on 'External References - General identifiers' in the side menu.

☞ Explore some of the links (a good place to start is usually 'UniProtKB/Swiss-Prot').

☞ Do the same for 'Ontology - Ontology table'.

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

(d)

☞ Click on 'External References - Oligo probes' in the side menu.

The CodeLink microarray contains one probe, GE80895, that maps to ENST00000218099, so it is possible to monitor expression of this transcript using this array.

☞ Click on 'ENST00000394090' and 'ENST00000479617' in the list of transcripts.

The probe GE80895 also maps to ENST00000394090. There is no CodeLink probe that maps to ENST00000479617 though, so expression of this transcript cannot be monitored using this array.

(e)

☞ Click on 'ENST00000218099' in case you are not already on the 'Transcript: F9-001' tab.

☞ Click on 'Protein Information - Protein summary' in the side menu.

☞ Click on 'Protein Information - Domains & features' in the side menu.

The peptidase activity of the protein resides in the peptidase domain that is located in the C-terminal half of the protein.

(f)

☞ Click on 'Protein Information - Variations' in the side menu.

☞ Type 'non-synonymous' in the 'Filter' text box at the top of the 'Variations' table.

For the protein encoded by ENST00000218099 seven non-synonymous coding variants are shown, four from dbSNP (i.e. rs104894807, rs6048, rs1801202 and rs4149751) and three from the COSMIC (Catalogue of Somatic Mutations in Cancer) database (COSM70673, COSM27239 and COSM24388).

(g)

☞ Click on the 'Gene: F9' tab.

☞ Click on 'Comparative Genomics - Orthologues' in the side menu.

☞ Type 'mouse' in the 'Filter' text box at the top of the 'Selected orthologues' table.

There is one mouse ortholog predicted for human *F9*, i.e. ENSMUSG00000031138.

---

## Exercise 2 – Exploring a region

(a) Go to the region from bp 32,448,000 to 33,198,000 on human chromosome 13. On which cytogenetic band is this region located? How many contigs make up this portion of the assembly (contigs are contiguous stretches of DNA sequence that have been assembled solely based on direct sequencing information)?

- (b) Are there any BAC clones that contain the complete *BRCA2* gene?
- (c) Zoom in on the *BRCA2* gene.
- (d) Add the track with RefSeq gene models. Has RefSeq annotated the *BRCA2* gene? If so, how many transcripts have been annotated? Do they differ from the Ensembl transcripts?
- (e) Export the genomic sequence of the region you are looking at in FASTA format.
- (f) Turn off all tracks you added to the 'Region in detail' page.
- (g) If you have yourself a genomic region of interest, explore what information Ensembl displays about it!
- 

### **Answer**

(a)

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Select 'Search: Human' and type '13:32448000-33198000' in the 'for' text box (or alternatively leave the 'Search' drop-down list like it is and type 'human 13:32448000-33198000' in the 'for' text box).
- ☞ Click [Go].

This genomic region is located on cytogenetic band q13.1. It is made up of seven contigs, indicated by the alternating light and dark blue colored bars in the 'Contigs' track.

(b)

- ☞ Click [Configure this page] in the side menu.
- ☞ Type 'clones' in the 'Find a track' text box.
- ☞ Select 'Tilepath', '1Mb clone set' and '32k clone set'.
- ☞ Click (✓).

It doesn't look like there is a clone that contains the complete *BRCA2* gene. For example clone RP11-37E23 contains most of the gene, but not its 5' end.

Note that the tilepath clones do correspond to the contigs and it is easy to see from which BAC clone which contig sequence in the assembly is derived, e.g. AC002525.1 is derived from RP1-257C22, AL137143.8 is derived from RP11-207N4 etc.

(c)

- ☞ Draw with your mouse a box around the *BRCA2* transcripts.
- ☞ Click on 'Jump to region' in the pop-up menu.

(d)

- ☞ Click [Configure this page] in the side menu.
- ☞ Type 'refseq' in the 'Find a track' text box.
- ☞ Select 'Human RefSeq import - Expanded with labels'.
- ☞ Click (✓).
- ☞ Click on transcript models to retrieve more information about them.

There has been one transcript annotated by RefSeq for the *BRCA2* gene, i.e. NM\_000059.3. This transcript is almost identical to Ensembl transcript BRCA2-001 (ENST00000380152). Both encode a 3418 aa protein. The RefSeq transcript is 6 bp shorter at the 5' end and 462 bp longer at the 3' end.

(e)

- ☞ Click [Export data] in the side menu.
- ☞ Click [Next>].
- ☞ Click on 'Text'.

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

```
>13 dna:chromosome chromosome:GRCh37:13:32883613:32978196:1
```

(f)

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

---

## COMPARATIVE GENOMICS

---

### Exercise 1 – Orthologs, paralogs and genetrees

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

- (a) Find the gene encoding the long-wave-sensitive (red) opsin.
- (b) How many within-species paralogs have been identified for this gene? Note the 'Target %id' and 'Query %id'. Which paralog has the most sequence similarity with the long-wave-sensitive opsin?
- (c) Have a look at the genomic location of the long-wave-sensitive (red), medium-wave-sensitive (green) and short-wave-sensitive (blue) opsin genes. Does this explain why red-green color blindness is much more prevalent in males than in females (e.g. in the US population 7% vs 0.4%)?
- (d) Have a look at the gene tree for the long-wave-sensitive opsin gene. Which of its paralogs is due to the most recent duplication event? Is this reflected in the sequence similarity between the long-wave-sensitive opsin and this paralog when compared with the other paralogs (see question b)? On which taxonomic level did this duplication take place?
- (e) Retrieve an alignment between the long-wave-sensitive opsin and all its paralogs in Jalview. To this end, select all aligned protein sequences using 'Select > Select all', then order them by using 'Calculate > Order > by ID', subsequently select all human protein sequences and finally use 'Select > Invert Sequence Selection' and 'Edit > Delete' to delete all non-human protein sequences.

---

### Answer

(a)

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Select 'Search: Human' and type 'long wave sensitive opsin' in the 'for' text box.
- ☞ Click [Go].
- ☞ Click on 'Gene' on the page with search results.
- ☞ Click on 'Human'.

☞ Click on 'OPN1LW [ Ensembl/Havana merge: ENSG00000102076 ]'.

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

(b)

☞ Click on 'Comparative Genomics - Paralogues' in the side menu.

There have six within-species paralogs been identified for the human long-wave-sensitive (red) opsin gene. ENSG00000147380 (*OPN1MW*) and ENSG00000166160 (*OPN1MW2*), the genes encoding the medium-wave-sensitive (green) opsins, show the highest Target %id and Query %id. The medium-wave-sensitive opsins thus have the highest sequence similarity to long-wave-sensitive opsin (Target %id indicates the percentage of the sequence of long-wave-sensitive opsin matching the sequence of the paralog protein. Query %id indicates the percentage of the sequence of the paralog protein matching the sequence of long-wave-sensitive opsin).

(c)

☞ Click on the 'Location: X:153,409,698-153,424,507' tab.

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

(d)

☞ Click on the 'Gene: OPN1LW' tab.

☞ Click on 'Comparative Genomics - Gene Tree (image)' in the side menu.

☞ Click on 'View options: View paralogs of current gene' below the gene tree image.

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

A duplication event on the level of the Hominoidea (Apes and man) has given rise to the long-wave-sensitive and medium-wave-sensitive opsins. The other paralogs are due to earlier duplication events. This agrees with the fact that the medium-wave-sensitive opsins show the highest sequence similarity with long-wave-sensitive opsin (see question b) and the fact that the genes for the

long-wave-sensitive and medium-wave-sensitive opsins are located close to each other on the genome (see question c).

(e)

- ☞ Click on the speciation node (blue square) that is at the base of the complete gene tree.
- ☞ Click on 'Expand for Jalview' in the pop-up menu (that should say 'Taxon: Coelomata').
- ☞ Click [Start Jalview].
- ☞ Close the pop-up window with the gene tree.
- ☞ Click on 'Select > Select all' on the menu bar of the pop-up window with the protein sequence alignment.
- ☞ Click on 'Calculate > Sort > by ID' on the menu bar.
- ☞ Select the protein sequences of the human paralogs.
- ☞ Click on 'Select > Invert Sequence Selection' on the menu bar.
- ☞ Click on 'Edit > Delete' on the menu bar.

As the alignment is based on the complete set of protein sequences in the gene tree, the alignment of this subset of seven proteins will contain empty columns. These can be removed using the option 'Edit > Remove Empty Columns' on the menu bar.

- ☞ Click on 'Edit > Remove Empty Columns' on the menu bar.

---

## Exercise 2 – Whole genome alignments

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

(b) Turn on the 'BLASTZ alignment' tracks for chicken, chimp, mouse and platypus and the 'Translated BLAT alignment' tracks for anole lizard and zebrafish. Does the degree of conservation between human and the various other species reflect their evolutionary relationship? Which parts of the *BRCA2* gene seem to be the most conserved? Did you expect this?

(c) Have a look at the 'Conservation score' and 'Constrained elements' tracks for the set of 35 mammals and the set of 19 vertebrates. Do these tracks confirm what you already saw in the tracks with pairwise alignment data?

(d) Retrieve the genomic alignment for a constrained element. Highlight the bases that match in >50% of the species in the alignment.

(e) Retrieve the genomic alignment for the *BRCA2* gene for the primates. Highlight the bases that match in >50% of the species in the alignment.

---

## Answer

(a)

- 🔗 Go to the Ensembl homepage (<http://www.ensembl.org>).
- 🔗 Select 'Search: Human' and type 'brca2 ' in the 'for' text box.
- 🔗 Click [Go].
- 🔗 Click on 'Gene' on the page with search results.
- 🔗 Click on 'Human'.
- 🔗 Click on '13:32889611-32973805:1' below 'BRCA2 [ Ensembl/Havana merge: ENSG00000139618 ]'.

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].
- 🔗 Click (✓).

(b)

- 🔗 Click [Configure this page] in the side menu
- 🔗 Click on 'Comparative genomics - BLASTZ/LASTz alignments'.
- 🔗 Select 'Chicken (*Gallus gallus*) - BLASTZ\_NET - Normal', 'Chimpanzee (*Pan troglodytes*) - BLASTZ\_NET - Normal', 'Mouse (*Mus musculus*) - BLASTZ\_NET - Normal' and 'Platypus (*Ornithorhynchus anatinus*) - BLASTZ\_NET - Normal'.
- 🔗 Click on 'Comparative genomics - Translated blat alignments'.
- 🔗 Select 'Anole Lizard (*Anolis carolinensis*) - TRANSLATED\_BLAT\_NET - Normal' and 'Zebrafish (*Danio rerio*) - TRANSLATED\_BLAT\_NET - Normal'.
- 🔗 Click (✓).

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, chicken, lizard and zebrafish, 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 'Comparative genomics - Conservation regions'.
- ☞ Select 'Conservation score for 35 eutherian mammals EPO\_LOW\_COVERAGE', 'Constrained elements for 35 eutherian mammals EPO\_LOW\_COVERAGE', 'Conservation score for 19 amniota vertebrates Pecan' and 'Constrained elements for 19 amniota vertebrates Pecan'.
- ☞ Click (✓).

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 that the UTRs don't, though.

(d)

- ☞ Click on a constrained element.
- ☞ Click on 'View alignments (text)' in the pop-up menu.
- ☞ Click [Configure this page] in the side menu.
- ☞ Select 'Conservation regions: All conserved regions'.
- ☞ Click (✓).

(e)

- ☞ Click on the 'Gene: BRCA2' tab.
- ☞ Click on 'Comparative Genomics - Genomic alignments' in the side menu.
- ☞ Select 'Alignment: 6 primates EPO'.
- ☞ Click [Go].
- ☞ Click [Configure this page] in the side menu.
- ☞ Select 'Conservation regions: All conserved regions'.
- ☞ Click (✓).

### Exercise 3 – Synteny

- (a) Find the Ensembl *CYCS* (Cytochrome c) gene for human.
- (b) Is there an ortholog predicted for the human *CYCS* gene in mouse?
- (c) Have for all orthologs a look at the 'Multi-species view' page. Based on synteny information, which seems the best candidate to be the mouse cytochrome c gene?

### Answer

(a)

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Select 'Search: Human' and type 'cycs' in the 'for' text box.
- ☞ Click [Go].
- ☞ Click on 'Gene' on the page with search results.
- ☞ Click on 'Human'.
- ☞ Click on 'CYCS [ Ensembl/Havana merge gene: ENSG00000172115 ]'.

(b)

- ☞ Click on 'Comparative Genomics - Orthologues' in the side menu.
- ☞ Type 'mouse' in the 'Filter' text box at the top of the 'Selected orthologues' table.

There are four 1-to-many orthologs predicted for the human *CYCS* gene in mouse (ENSMUSG00000058927, ENSMUSG00000062038, ENSMUSG00000062813 and ENSMUSG00000063694), as well as one possible ortholog (ENSMUSG00000056436).

(c)

- ☞ Click on 'Multi-species view' for each of the orthologs.
- ☞ Click each time on the back button of the browser to get back to the 'Orthologues' page.

In human the *CYCS* gene is located downstream of the *MPP6*, *DFNA5* and *OSBPL3* genes and upstream of the *NPVF* gene. In mouse, ENSMUSG00000063694 is flanked by the same genes, while the others aren't. Based on this information, ENSMUSG00000063694 seems the best candidate to be the mouse cytochrome c gene.

Note: That ENSMUSG00000063694 is only recognized as one of four 1-to-many orthologs of the human *CYCS* gene in the Ensembl ortholog/paralog prediction pipeline is due to the fact that the cytochrome c protein is very conserved between different species. Due to this high conservation, the amount of phylogenetic signal present in this gene family is low, which makes the tree topology determination less accurate. This makes certain orthologous/paralogous relationships difficult to infer.

#### Exercise 4 – Pan-taxonomic compara

(a) Use the Ensembl Genomes browser (<http://www.ensemblgenomes.org>) to find the fumarase gene of *Saccharomyces cerevisiae*.

(b) Is this gene conserved across all the kingdoms, i.e. animals, plants, fungi, protists and prokaryotes? Is that what you would expect?

## Answer

(a)

- ~ Go to the Ensembl Genomes homepage (<http://www.ensemblgenomes.org>).
- ~ Click on 'Fungi' in the menu bar.
- ~ Select 'Search: *Saccharomyces cerevisiae*' and type 'fumarase' in the 'for' text box.
- ~ Click [Go].
- ~ Click on 'FUM1 (SGD: YPL262W)' on the page with search results.

(b)

- ~ Click on 'Pan-taxonomic Compara - Gene Tree (image)' in the side menu.
- ~ Click on 'View fully expanded tree' below the gene tree image.

There are 53 orthologs predicted for the fumarase gene, covering all kingdoms, from bacteria to human. This is what you would expect as fumarase catalyzes a step in the citric acid cycle (TCA cycle), a process that is of central importance in all aerobic organisms (see also <http://en.wikipedia.org/wiki/Fumarase>).

---

---

## REGULATION

---

### Exercise 1 – Regulatory build

The *HLA-DRB1* and *HLA-DQA1* genes are part of the human major histocompatibility complex class II (MHC-II) region and are located about 44 kb from each other on chromosome 6. In the paper 'The human major histocompatibility complex class II HLA-DRB1 and HLA-DQA1 genes are separated by a CTCF-binding enhancer-blocking element' (Majumder *et al.* J Biol Chem. 2006 Jul 7;281(27):18435-43) a region of high acetylation located in the intergenic sequences between *HLA-DRB1* and *HLA-DQA1* is described. This region, termed XL9, coincided with sequences that bound the insulator protein CCCTC-binding factor (CTCF). Majumder *et al.* hypothesize that the XL9 region may have evolved to separate the transcriptional units of the *HLA-DR* and *HLA-DQ* genes.

- (a) Go to the region from bp 32,540,000 to 32,620,000 on human chromosome 6
- (b) Is there a regulatory feature annotated in the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes that has CTCF binding data as (part of) its core evidence?
- (c) Has CTCF binding been detected at this position in all cell/tissue types analyzed?
- (d) Have a look at the 'Regulatory evidence - Histones & Polymerases' configuration matrix. For which cell/tissue type are the most histone acetylation data sets available? Is in this cell/tissue type the region that shows CTCF binding also a region of high acetylation, as found by Majumder *et al.*?

---

### Answer

(a)

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Select 'Search: Human' and type '6:32540000-32620000' in the 'for' text box.
- ☞ Click [Go].

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].

☞ Click (✓).

(b)

☞ Click on the regulatory features shown in the 'Reg. Feats' track that are located in the intergenic region between the *HLA-DRB1* and *HLA-DQA1* genes.

Yes, there is one regulatory feature, ENSR00000488025, that has CTCF binding data as part of its core evidence.

(c)

- ☞ Click [Configure this page] in the side menu.
- ☞ Click on 'Regulation - Open chromatin & TFBS'.
- ☞ Select 'MultiCell - Track style: Peaks'.
- ☞ Click (✓).

CTCF binding has been detected at this position in eight of the cell/tissue types analyzed.

(d)

- ☞ Click [Configure this page] in the side menu.
- ☞ Click on 'Regulation - Histones & polymerases'.

According to the 'Histones & Polymerases' configuration matrix the most information on histone acetylation is available for CD4 cells.

- ☞ Hover over 'CD4' in the 'Histones & Polymerases' configuration matrix.
- ☞ Select 'Select features for CD4 - All'.
- ☞ Click (✓).

Yes, the region that shows CTCF binding is in CD4 cells also a region of high acetylation, of histone 2, 3 and 4.

---

## Exercise 2 – Methylation

The human *PDHA2* gene, that encodes for a subunit of the pyruvate dehydrogenase complex, is exclusively expressed in spermatogenic cells. In the paper 'Human testis-specific PDHA2 gene: Methylation status of a CpG island in the open reading frame correlates with transcriptional Activity' (Pinheiro *et al.* Mol Genet Metab. 2010 Apr;99(4):425-30), two CpG islands in the *PDHA2* gene are reported, one encompassing the core promoter region and extending into the open reading frame, the other exclusively located in the coding region. The latter CpG island was shown to

be methylated in somatic tissues but demethylated in testicular germ cells and has therefore been proposed to play an important role in the tissue-specific expression of the *PDHA2* gene.

(a) Find the *PDHA2* gene for human and go to the 'Region in detail' page. Zoom out one step, so that 5 kb around the *PDHA2* gene is shown.

(b) Can you identify the two CpG islands in the *PDHA2* gene?

(c) Confirm the existence of the two CpG islands using the EMBOSS program CpGPlot (<http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html>) on the sequence around the *PDHA2* gene.

(d) Upload the CpG islands found by CpGPlot using [Manage your data]. Use BED format, which in its simplest form just consists of the chromosome and the start and end coordinates, separated by spaces (as an optional fourth field, you can add a name/description). The genomic start and end coordinates of the CpG islands can be calculated from the genomic start coordinate of the sequence on which the CpGPlot program was run and the relative location of the CpG islands on this sequence as given by the CpGPlot output.

(e) Are the two CpG islands methylated in somatic tissues but demethylated in sperm?

---

### **Answer**

(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 'Gene' on the page with search results.
- ☞ Click on 'Human'.
- ☞ Click on '4:96761239-96762625:1'.
- ☞ 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].
- ☞ Click (✓).

(b)

- ☞ Click [Configure this page] in the side menu.
- ☞ Type 'cpg' in the 'Find a track' text box.
- ☞ Select 'CpG islands'.
- ☞ Click (✓).

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)

- ☞ Click [Manage your data] in the side menu.
- ☞ Click on 'Upload Data'.
- ☞ Type 'CpG islands' in the 'Name for this upload (optional)' text box.
- ☞ Select 'Data format: BED'.
- ☞ Type the following in the 'Paste file' text box:

```
chr4 96761176 96761375 cpg_island_1
chr4 96761500 96761762 cpg_island_2
```

- ☞ Click [Upload].
- ☞ Click on 'Go to nearest region with data: 4:96711276-96811276'.

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 wrench.

☞ Select 'Labels'.

(e)

☞ Click [Configure this page] in the side menu.

☞ Click on 'Regulation - DNA Methylation'.

☞ Select all 'MeDIP' tracks in 'Normal' mode.

☞ Click (✓).

Yellow, green and blue represent unmethylated, intermediately methylated and methylated regions, respectively. It can be seen that the region around the 5' part of the *PDHA2* gene is methylated in all assayed tissues and cell lines, except in sperm. The 'MeDIP-seq' track for sperm shows that the unmethylated regions coincide with the CpG islands found by CpGPlot.

---

---

## VARIANT EFFECT PREDICTOR

---

Resequencing of the genomic region of the human *CFTR* (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)) gene (ENSG0000001626) has amongst others revealed the following variants (alleles defined in the forward strand):

G/A at 7:117,171,039

T/C at 7:117,171,092

T/C at 7:117,171,122

(a) Determine if these variants result in a change in the proteins encoded by any of the Ensembl transcripts of the *CFTR* gene. Which of these variants is the most damaging / deleterious according to the SIFT and PolyPhen tools? Have the variants already been annotated by Ensembl?

(b) Have a look at the uploaded variants in 'Region in detail'.

---

### Answer

(a)

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Click on the 'Tools' link on the toolbar.
- ☞ Click on 'Variant Effect Predictor - Upload your data'.
- ☞ Enter 'My variants' in the 'Name for this upload (optional)' text box.
- ☞ Enter the following list in the 'Paste file' text box (be sure to separate the five values on each row by spaces):

```
7 117171039 117171039 G/A +
7 117171092 117171092 T/C +
7 117171122 117171122 T/C +
```

- ☞ Select 'SIFT predictions: Prediction only'.
- ☞ Select 'PolyPhen predictions: Prediction only'.
- ☞ Click [Next>].
- ☞ Click on 'HTML'.

Two of the variants are non-synonymous coding in three of the encoded proteins and thus cause an amino acid change. One variant is synonymous coding and thus doesn't change the amino acid sequence. Of the two non-synonymous coding variants, the one at position 117171092 seems to be more deleterious / damaging according to SIFT and PolyPhen than the one at position 117171122. All variants have already been annotated by Ensembl as

can be seen from the dbSNP and COSMIC accession numbers that are shown in the 'Co-located Variation' column.

(b)

☞ Click for one of the variants on the coordinates in the 'Location' column on the 'Variant Effect Predictor Results' page.

The uploaded variants should be shown on 'Region in detail' in a new track, named 'My variants'.

☞ Draw with your mouse a box around the uploaded variants.

☞ Click on 'Jump to region' in the pop-up menu.

It can be clearly seen that all three uploaded variants correspond to an already annotated variant shown in the 'Sequence variants (all sources)' track, one synonymous coding (green) and two non-synonymous coding (yellow).

---

---

## CUSTOM ANNOTATION

---

### Exercise 1 – Attaching a BAM file

The following file contains alignments to the GRCh37 assembly of low coverage Illumina sequencing reads of chromosome 20 of individual HG00096 from the 'British from England and Scotland, UK' cohort ([http://ccr.coriell.org/Sections/Search/Sample\\_Detail.aspx?Ref=HG00096&PgId=166](http://ccr.coriell.org/Sections/Search/Sample_Detail.aspx?Ref=HG00096&PgId=166)):

[http://www.ebi.ac.uk/~bert/HG00096.chrom20.ILLUMINA.bwa.GBR.low\\_coverage.20100901.bam](http://www.ebi.ac.uk/~bert/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20100901.bam)

The file is in BAM format. BAM is the compressed binary version of the SAM (Sequence Alignment/Map) format, a compact and indexable representation of nucleotide sequence alignments:

<http://samtools.sourceforge.net/SAM1.pdf>

To display these data in Ensembl also the .bam.bai index file is needed:

[http://www.ebi.ac.uk/~bert/HG00096.chrom20.ILLUMINA.bwa.GBR.low\\_coverage.20100901.bam.bai](http://www.ebi.ac.uk/~bert/HG00096.chrom20.ILLUMINA.bwa.GBR.low_coverage.20100901.bam.bai)

The .bam and .bam.bai files should be placed in the same directory.

Attach the file to Ensembl and have a look at the result. Can you find any individual reads containing a nucleotide that differs from the sequence of the reference genome? And a position where individual HG00096 differs from the reference genome or where individual HG00096 is heterozygous?

---

### Answer

- ☞ Go to the Ensembl homepage (<http://www.ensembl.org>).
- ☞ Click on the picture of Leonardo da Vinci's 'Vitruvian Man' or the word 'Human' next to it.
- ☞ Click [Manage your data] in the side menu.
- ☞ Click on 'Attach Remote File'.
- ☞ Enter the URL of the file in the 'File URL' text box.
- ☞ Select 'Data format: BAM'.
- ☞ Enter 'HG00096' in the 'Name for this track' text box.
- ☞ Click [Next>].
- ☞ Click (✓).
- ☞ Go to any 'Region in detail' page for chromosome 20.

A new track named 'HG00096' should have been added to the 'Region in detail' page.

🔗 Zoom in to see the actual reads.

Individual reads are shown in grey, with the consensus sequence shown above the reads in color.

If you want to compare the reads to the reference genome sequence:

🔗 Click [Configure this page] in the side menu.

🔗 Click on 'Sequence and assembly - Sequence'.

🔗 Select 'Sequence'.

🔗 Click (✓).

Nucleotides that differ from the sequence of the reference genome are shown in red:

[http://www.ensembl.org/Homo\\_sapiens/Location/View?db=core&r=20:44861207-44861246](http://www.ensembl.org/Homo_sapiens/Location/View?db=core&r=20:44861207-44861246)

An example of a position where individual HG00096 is heterozygous:

[http://www.ensembl.org/Homo\\_sapiens/Location/View?db=core&r=20:44854706-44854746](http://www.ensembl.org/Homo_sapiens/Location/View?db=core&r=20:44854706-44854746)

---

## Exercise 2 – Attaching a VCF file

The following file contains variants from the Sanger Mouse Genomes Project (<http://www.sanger.ac.uk/resources/mouse/genomes/>):

[ftp://ftp-mouse.sanger.ac.uk/current\\_snps/20110602-final-snps.vcf.gz](ftp://ftp-mouse.sanger.ac.uk/current_snps/20110602-final-snps.vcf.gz)

The file is in VCF (Variant Call Format) format. VCF is a tab delimited format for storing variant calls and individual genotypes:

<http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-40>

To display these data in Ensembl also the .vcf.gz.tbi index file is needed:

[ftp://ftp-mouse.sanger.ac.uk/current\\_snps/20110602-final-snps.vcf.gz.tbi](ftp://ftp-mouse.sanger.ac.uk/current_snps/20110602-final-snps.vcf.gz.tbi)

The .vcf.gz and .vcf.gz.tbi files should be placed in the same directory.

Attach the file to Ensembl and have a look at the result. Is the set of variants from the Sanger Mouse Genomes Project less or more comprehensive than the set of mouse variants present in Ensembl?

---

### **Answer**

- 🔗 Go to the Ensembl homepage (<http://www.ensembl.org>).
  - 🔗 Click on the picture of the mouse or the word 'Mouse' next to it.
  - 🔗 Click [Manage your data] in the side menu.
  - 🔗 Click on 'Attach Remote File'.
  - 🔗 Enter the URL of the file in the 'File URL' text box.
  - 🔗 Select 'Data format: VCF'.
  - 🔗 Enter 'Sanger variants' in the 'Name for this track' text box.
  - 🔗 Click [Next>].
  - 🔗 Click (✓).
- 
- 🔗 Go to any 'Region in detail' page.
  - 🔗 Click [Configure this page] in the side menu.
  - 🔗 Type 'variants' in the 'Find a track' text box.
  - 🔗 Select 'Sequence variants (all sources)'.
  - 🔗 Click (✓).

A new track named 'Sanger variants' should have been added to the 'Region in detail' page. This track contains many more variants than the 'Sequence variants (all sources)' track. Once the Sanger Mouse Genomes Project has submitted these variants to dbSNP they will subsequently be imported into Ensembl and shown in the 'Sequence variants (all sources)' track.

- 🔗 Zoom in to see the individual variants.
- 

### **Exercise 3 – Removing custom annotation**

Remove your attached and uploaded annotations.

---

### **Answer**

- 🔗 Go to the Ensembl homepage (<http://www.ensembl.org>).
- 🔗 Click on the picture of Leonardo da Vinci's 'Vitruvian Man' or the word 'Human' next to it.
- 🔗 Click [Manage your data] in the side menu.
- 🔗 Click for each dataset on 'Delete'.
- 🔗 Click (✓).

Your annotations should be removed now.

---