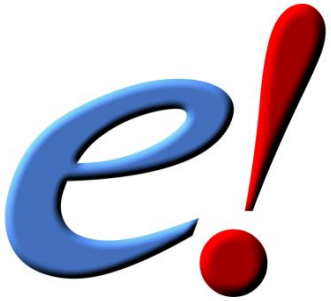
Browsing Variation Data with Ensembl



www.ensembl.org
www.ensemblgenomes.org

Exercise Answers v95

http://training.ensembl.org/events/2019/2019-04-02-VEP_Breda

Variant Effect Prediction Course, Breda
2nd-5th April 2019

Exercise answers

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 all four transcripts of TAGAP (ENST00000326965, ENST00000338313, ENST00000367066 and ENST00000642909).

(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 populations are, hover over the names).

(c) 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/).

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 seven *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 Medellin; 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: 12 primates EPO and click Go.
All twelve primates, including human, have a G in this position.

Exercise V3 – Exploring a SNP in mouse

(a) Go to 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 five HGVS names, one at the genomic DNA level (NC_000083.6:g.73924993C>T), three at the transcript level (ENSMUST00000024866.4:c.721G>A, ENSMUST00000233162.1:n.738G>A,

ENSMUST00000233621.1:c.*284G>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 Mouse Genomes Project, or search for the two strain names.
There are indeed differences between the genotypes reported in those two different strains. The genotype reported in ARK/J is T|T whereas in C57BL/6NJ the genotype is C|C.

Exercise V4 – Variation structure viewer

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

Type rs998717588 into the search box. Click Go.

Click on the variant ID rs998717588 from the search results. Click on 3D Protein model from the menu on the left hand side of the page.

(b) You can find the amino acid residue affected by this variant in the Variants panel. This variant affects residue number 121 of the PDB structure and of the ENSP protein.

Use the LiteMol interactive viewer to identify the location of the variant (red). The variant is located in a helix.

Hover your mouse over the red highlighted residue in the interactive viewer. An information box in the top left hand corner of the image shows you that this residue is a Valine.

(c) Click on the grey eye icon in the Pfam section of the Protein Information panel to highlight the Pfam domains in the image. Click on the + icon in the Pfam section of the Protein Information panel to show information about the Pfam domains. The variant falls in Pfam domain PF00337 (purple).