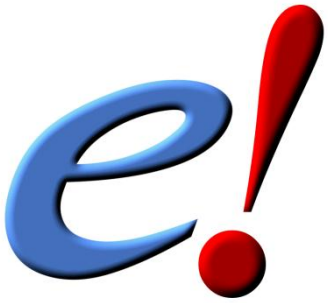
Answers



www.ensembl.org

Exercise Answers v73

Chemical and Biological resources for Toxicology and Toxicogenomics (diXa) at the EBI

http://www.ebi.ac.uk/~gspudich/workshop_presentations/EBI_toxicology

Exploring the Ensembl genome browser

Exercise 1 - Exploring the human *NQO1*gene

(a) Human *NQO1* (ENSG00000181019)

- Chromosome 16, base pair 69,740,899-69,760,854 on the reverse strand.
- Ensembl has seven transcripts annotated for this gene.
- Six transcripts are protein coding.

(b) Yes (click the Orthologues link at the left). Chromosome 5.

(c) These are some of the phenotypes associated to *NQO1:* SUSCEPTIBILITY TO BENZENE TOXICITY and tumours in various tissues.

(d) Click on the Ontology table link to show GO associations to this gene, for example:

- superoxide metabolic process
- response to oxidative stress
- cytochrome-b5 reductase activity, acting on NAD(P)H

Exercise 2 – 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) Click on the gene symbol to go to the Gene tab. Click on Expression in the left hand menu.

The gene is expressed in all tissues listed. This is unsurprising for a metabolic gene.

Hover over the column titles to view definitions.

Intron spanning reads are RNASeq reads that cover exon junctions.

RNASeq alignments are RNASeq reads that align to the genome.

(c) 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.

(d) Click on External references.

The MIM disease ID is 261600.

Orphanet lists multiple forms of phenylketonuria: Classical phenylketonuria, Maternal hyperphenylalaninemia, Mild hyperphenylalaninemia, Mild phenylketonuria and Tetrahydrobiopterin-responsive hyperphenylalaninemia/phenylketonuria

Exercise 3 – Exploring a genomic region in human

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

Select Search: Human and type 13:32448000-33198000 in the text box (or alternatively leave the Search drop-down list like it is and type human 13:32448000-33198000 in the 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 coloured bars in the Contigs track.

(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 clones in the Find a track text box.
Select Tilepath.
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).

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

Select the link and copy.
Compose an email to yourself, paste the link in and send the message.
Open the email and click on your link. You should be able to view the page with the new configuration and data tracks you had added to in the Location tab.

(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 (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

Gene trees and homologues

Exercise 4 – Orthologues, paralogues and gene trees for the human *BRAF* gene.

(a) Go to 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 88 orthologous genes.

There are orthologues in 8 primates (no orthologue has been described in macaque).

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.
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).

Whole genome alignments

Exercise 5 – Zebrafish orthologues

(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, 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 Alignments (images) or Alignments (text) menus in the Location tab. Click on the blue Export data button at the left, and choose Clustal from the list.

(c) Click on Region in detail in the left hand menu. Turn on the multiple alignment and, constrained elements and conservation score for 5 teleost fish EPO tracks, both under the Comparative genomics menu by configuring the page.

The 5 teleost fish EPO track just shows that the whole region for the *dbh* gene can be aligned among those five 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. Note that there are intronic regions that seem to be fairly conserved across the species available.

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

Exercise 6 – 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.

The homologue in dog of human RHO is OPDS_CANFA. Click 15 downstream genes (or upstream) to compare the genes between human and dog in this syntenic block.

Exercise 7 – Whole genome alignments

(a) Go to the Ensembl homepage (http://www.ensembl.org/).
Select Search: Human and type brca2 in the search box.
Click Go.
Click on 13:32889611-32973805:1 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) - BLASTZ_NET – Normal, Chimpanzee (Pan troglodytes) – BLASTZ_NET – Normal, Mouse (Mus musculus) – BLASTZ_NET - Normal and Platypus (Ornithorhynchus anatinus) - BLASTZ_NET - Normal.
Click on Translated blat alignments. Select Anole Lizard (Anolis carolinensis) - TRANSLATED_BLAT_NET - Normal and Zebrafish (Danio rerio) - TRANSLATED_BLAT_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, 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 Conservation regions under the Comparative genomics menu.

Select Conservation score for 36 eutherian mammals EPO_LOW_COVERAGE, Conservation score for 20 amniota vertebrates Pecan and Constrained elements for 20 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 Conservation regions: All conserved regions.
SAVE and close.

The conserved regions will be shown in light blue.

(e) Click on the Gene: BRCA2 tab.
Click on Genomic alignments under Comparative Genomics 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.
SAVE and close.

The conserved regions will be shown in light blue.

BioMart

Exercise 8 – 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) 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 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 to an Excel sheet by using the Export all results to XLS option.

Exercise 9 – Convert IDs

Click New.
Choose the ENSEMBL Genes 73 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.

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. Tick the box Unique results only.

Exercise 10 – Export homologues

Click New.
Choose the ENSEMBL Genes 73 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).

Exercise 11 – Export structural variants

(a) Choose Ensembl Variation 73 and *Homo sapiens* Structural Variation.
Filters: Region: Chromosome 1, Base pair start: 130408, Base pair end: 210597
Count shows 6 out of 3561682 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 73 and Homo sapiens Short Variation (SNPs and indels).
Filters: Filter by Variation ID 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.

Exercise 12 – Find genes associated with array probes

(a) Click New.
Choose the ENSEMBL Genes 73 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 probeset ID(s) and enter the list of probeset IDs in the text box (either comma separated or as a list).

Count shows 25 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 25 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.
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.

Variation

Finding variants in Ensembl

Exercise 13 – 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 Show variants in 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 (CSHL-HAPMAP:HapMap-YRI population), the least frequent genotype is CC at the frequency of 9.7%. This is also the least frequent genotype in in other populations (to find out what the three letter population are, have a look at our FAQ (http://www.ensembl.org/Help/Faq?id=328)

(c) Click on phylogenetic context.

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

Select the 36 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 three of the 36 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 diabetes, 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 PubMed article.

You can view External Data sources that mirror data from SNPedia and LOVD. We share information about the effects of variations in DNA, citing peer-reviewed scientific publications. Click on SNPedia and LOVD in the left hand menu to explore further. No LOVD data was found for this variant so far.

Exercise 14 – 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 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 study where the association was originally described is given in the Phenotype Data table. Click on pubmed/20031578 for more details.

The association between rs1801133 and homocysteine levels is described in the paper 'Novel associations of *CPS1*, *MUT*, *NOX4* and *DPEP1* with plasma homocysteine in a healthy population:

a genome-wide evaluation of 13,974 participants in the Women's Genome Health Study' (Pare *et al*, Cir Cardiovasc Genet. 2009 Apr;2(2):142-50).

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

Select Alignment: 6 primates EPO and click Go.

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

(g) Go to http://neandertal.ensemblgenomes.org/ and type rs1801133 in the Search Neandertal text box.
Click Go.
Click on rs1801133 on the results page.
Click on Jump to region in detail.
Click on Configure this page in the side menu.
Click on Variation features.
Select All variations – Normal.
SAVE and close.
Draw a box of about 50 bp around rs1801133 (shown in yellow in the centre of the display).
Click on Jump to region on the pop-up menu.

The Sequences track shows that there are four reads for Neanderthal at the position of rs1801133, all with a G, so based on these (very limited) data there is no evidence that both alleles were already present in Neanderthal.

Exercise 15 – Structural variation in human

(a) Go to the Ensembl homepage (http://www.ensembl.org/).
Select Search: Human and type *ccl3l1* in the search box.
Click Go.
Click on *CCL3L1* (Human Gene) at the top.

(b) Click on Structural Variation in the side menu.

Yes, CNVs have been annotated for this gene by multiple studies, as indicated by the many bars in the larger and smaller structural variants tracks in the display. Details are given in the table below the display.

Note: Can you do this with BioMart?

Exercise 16 – 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 Additional information Show to reveal information about HGVS nomenclature.

This SNP has got three HGVS names, one at the genomic DNA level (g.73924993C>T), one at the transcript level (721G>A) and one at the protein level (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 Individual 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 17 – VEP (Variant Effect Predictor tool)

(a) Go to 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 117171039 117171039 G/A
7 117171092 117171092 T/C
7 117171122 117171122 T/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

Under the non-synonymous SNP predictions option, select prediction only for SIFT and PolyPhen, then click Next.

The output format is either in HTML or text. You will get a table with the consequence terms from the Sequence Ontology project (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 117171092) and benign/tolerated (coordinate 117171122). 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).

(b) In order to see your uploaded SNPs as a track in Region in detail, you will need to choose a name for this upload (e.g. VEP) when entering the data into the VEP tool. So you may need to enter the data again. Once you have done that and given a name to the upload, click on any link under the location column (in the VEP results table) to see your newly added VEP track with the three variations in the Location tab (or Region in detail view) in Ensembl.