Ensembl browser webinar series

Emily Perry [1], Ben Moore [2], Helen Sparrow [3], Denise Carvalho-Silva [4], Victoria Newman [5], Astrid Gall [6], Erin Haskell [7]

- DNA & RNA
- Beginner
- > 3 hours

This course includes an introduction to Ensembl, what kind of data Ensembl has, and how to browse and export the data - essentially providing one of our day-long training courses in short chunks. It consists of video introductions with the background the data and demos using Ensembl, along with exercises so that you can practice what you have learned.

You do not need to complete the full course; you can dip in and out to complete only the modules you are interested in.

Go to 'how to take this course [8]' to learn more about the course content

Learning objectives:
- Know the data types available in Ensembl and how to access them
- Be able to view data in the Ensembl browser
- Be able to mine Ensembl data using BioMart

How to take this course

Live webinars 4th, 6th, 11th and 13th September 2018

This course consisted of a series of 1.5 hour live webinars, held between 4th and 13th September 2018.

The programme was be:

<table>
<thead>
<tr>
<th>Date</th>
<th>Webinar topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>4th September</td>
<td>Introduction to Ensembl [9]</td>
<td>Astrid Gall</td>
</tr>
<tr>
<td></td>
<td>Ensembl genes [10]</td>
<td>Emily Perry</td>
</tr>
<tr>
<td>6th September</td>
<td>Variation data in Ensembl and the</td>
<td>Erin Haskell</td>
</tr>
<tr>
<td>Date</td>
<td>Webinar topic</td>
<td>Instructor</td>
</tr>
<tr>
<td>--------------</td>
<td>----------------------------------------------------</td>
<td>-----------------------------</td>
</tr>
<tr>
<td>Comparing genes and genomes with Ensembl Compara</td>
<td>[12]</td>
<td></td>
</tr>
<tr>
<td>11th September</td>
<td>Finding features that regulate genes – the Ensembl Regulatory Build [13]</td>
<td>Emily Perry, Erin Haskell</td>
</tr>
<tr>
<td></td>
<td>Data export with BioMart [14]</td>
<td></td>
</tr>
<tr>
<td>13th September</td>
<td>Uploading your data to Ensembl [15]</td>
<td>Astrid Gall</td>
</tr>
<tr>
<td></td>
<td>Introduction to the Ensembl REST APIs</td>
<td>Emily Perry</td>
</tr>
</tbody>
</table>

You will notice that the topic titles correspond to the pages of the course listed on the left. After each webinar, we will upload the recording to our YouTube channel [16] and embed the video in the corresponding page, so if you cannot make a particular webinar, you can always catch up later.

**Exercises to test what you have learnt**

After each webinar video, there are pages in each section of this course with exercises (and their solutions), which will enable you to practice using what you have learnt. Each set of exercises should take you no more than an hour, and you can space them out and do them at your own pace.

**Getting help with the exercises**

You can helpdesk [at] ensembl.org (email us) with any questions.

Unfortunately, as we have no way of determining who has completed all parts of the course, we are unable to provide certificates.

**Introduction to Ensembl**

This webinar is presented by Emily Perry.

This first webinar introduces you to the Ensembl project and provides an overview of the resources it provides.

Using the browser you will find out how to explore your species of interest, look at the region in detail view and how to browse the genome.

If you have problems accessing YouTube, you can download the video [17].

You can also download the slides [18] and the demo [19].
Introduction exercises

Homepage, assemblies and species

Exercise 1 – Panda
(a) Go to the species homepage for Panda. What is the name of the genome assembly for Panda?

(b) Click on More information and statistics. How long is the Panda genome (in bp)? How many coding genes have been annotated?

Exercise 2 – Zebrafish
What previous assemblies are available for zebrafish?

Exercise 3 – Mosquitoes
(a) Go to Ensembl Metazoa. How many species of the genus Anopheles are represented in Ensembl Metazoa?

(b) When was the current Anopheles gambiae genome assembly last revised?

Exercise 4 – Bacteria
Go to Ensembl Bacteria and find the species Belliella baltica. How many coding and non-coding genes does it have?

Region in detail

Exercise 5 – Exploring a genomic region in human
(a) Go to the region from 31,937,000 to 32,633,000 bp 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) Zoom in on the BRCA2 gene.

(c) Configure this page to turn on the Tilepath track in this view. What is this track? Are there any Tilepath clones that contain the complete BRCA2 gene?

(d) Create a link for this display. Email it to your neighbour. Open the link they sent you and compare your display. If there are differences, can you work out why?

(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.
Exercise 6 – Exploring CRISPR sites

You want to do some CRISPR manipulation of the human SMC3 gene. You’re looking for a CRISPR editing site within the locus 10:110578600-110578700.

(a) Go to the locus above and turn on the CRISPR track. How many CRISPR sites can you see in this locus?

(b) Do any of the CRISPR sites overlap any phenotype causing variants? What are the identifiers of these sites and variants?

(c) Mark the region of the negative strand CRISPR site that overlaps these variants, then zoom out to see the whole SMC3 gene. What exon number is your CRISPR site found in the SMC3-201 transcript?

Exercise 7 – Exploring assembly exceptions in human

(a) Go to the region 21:32630000-32870000 in human. What is the red highlighted region? What is its name?

(b) Can you see the assembly exceptions in the Chromosome overview at the top? How many regions with assembly exceptions are on chromosome 21?

(c) Can you compare this assembly exception with the reference? What is different between this assembly exception and the version on the primary assembly?

Exercise 7 – Exploring a genomic region in Anopheles gambiae

(a) Go to the region from 7,300,000 to 7,450,000 bp on Anopheles gambiae chromosome 2L. On which cytogenetic band is this region located?

(b) How many genes are found in this region? Zoom in on the second exon of AGAP004970-RA. Turn on the track Start/Stop codons. Can you see the start codon of AGAP004970-RA?

(c) Highlight the start codon of AGAP004970-RA. Zoom out to view the whole gene. Can you see where you highlighted?

Exercise 8 – Exploring a genomic region in rice

(a) Go to the region 1:405000-453000 in Oryza sativa Japonica.

(b) Turn on the AGILENT:G2519F-015241 microarray track. Are there any oligo probes that map to this region?

(c) Highlight the region around any reverse strand probes you can see. Do they map to any transcripts?

Exercise 9 – Exploring a genomic region in bacteria

(a) Search for the Salmonella enterica subsp. enterica serovar Typhi str. CT18 (Hint: type CT18 into the Search for a genome box).

(b) Go to the region Chromosome:2000605-2009742.

(c) How many genes are annotated in this region? How many are on the forward strand? How many are on the reverse strand?
Exercise solutions

Homepage, assemblies and species

Exercise 1 – 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 right.

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

Exercise 2 – Zebrafish

Click on Zebrafish on the front page of Ensembl to go to the species homepage.

Under Other assemblies three previous assembly names and the releases you can find them in are listed.

Assembly GRCz10 is available in the archived release 91, assembly Zv9 is available in the archived release 79 and assembly Zv8 is available in the archived release 54.

Exercise 3 – Mosquitoes

(a) Go to metazoa.ensembl.org. Click on View full list of all Ensembl Metazoa species. Type Anopheles into the filter box at the right above the table.

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.

Exercise 4 – Bacteria

Go to bacteria.ensembl.org and start to type the name Belliella baltica into the genome search 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 genes.

Region in detail

Exercise 5 – Exploring a genomic region in human
(a) Go to the Ensembl homepage 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. You may need to zoom in to find it.

(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 no single clone that contains the complete BRCA2 gene. The BAC clone RP11-37E23 contains most of the gene, but its very 3’ end is contained in RP11-298P3. This is reflected by the two contigs that make up the complete 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 it.

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

When you receive the link from your neighbour, open the email and click on the 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 have specified a slightly different region to you, or where they have added different tracks.

(e) Click the **Export data** button in the side menu. Leave the default parameters as they are (**FASTA sequence** should already be selected). Click **Next>**.
Click **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:

```plaintext
<>13 dna:chromosome chromosome:GRCh38:13:32311910:32405865:1
```

(f) Click **Configure this page** in the side menu.
Click **Reset configuration**.
Click **?** or click anywhere outside the or anywhere outside the pop-up window.

Exercise 6 – Exploring CRISPR sites

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

Select **Search: Human** and type **10:110578600-110578700** in the text box (or alternatively leave the Search drop-
down list like it is and type **human 10:110578600-110578700** in the text box). Click Go.

Click Configure this page. Type **crispr** in the Find a track text box. Select the CRISPR Cas9 track in Labels track style.

There are five positive strand and three negative strand CRISPR sites.

(b) Click on the variants in the track All phenotype-associated - short variants (SNPs and indels) which is shown by default, and the CRISPR sites to get their identifiers.

CRISPR sites 1074131234, 1074131235 and 1074131236 overlap variants rs113411202 and rs1064797151. 074131234 and 1074131235 also overlap rs779773957. 1074131233 overlaps rs78663177.

(c) In the Region image at the bottom on the page, click and drag a box around the negative strand CRISPR sites, then select Mark region in the pop-up window. In the middle overview image on the page, click and drag a box around the SMC3 gene, then select Jump to region in the pop-up window. Count the exons of the SMC3-201 transcript or click on them to get the number of the exon with the marked region.

The negative strand CRISPR sites are found in exon 7 of the SMC3-201 transcript.

**Exercise 7 – Exploring assembly exceptions in human**

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

Select Search: Human and type **21:32630000-32870000** in the text box (or alternatively leave the Search dropdown 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 the Chromosome or Region view to see the label **CHR_HSCHR21_3_CTG1_1:32769079-32843731**. Click on What are assembly exceptions? to open a new window that explains assembly exceptions (Alternative sequence).

(b) Assembly exceptions are marked in the Chromosome view at the top. There are seven assembly exceptions on chromosome 21.

(c) Another option in the pop-up is Compare with reference. Click on this.

Scroll down the page to see the comparison between the assembly exception 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.

**Exercise 8 – Exploring a genomic region in rice**

(a) Go to the Ensembl Plants homepage.

Select Search: Oryza sativa Japonica and type **1:405000-453000**.

Click Go.

(b) Click on Configure this page to open the menu. You can find the **AGILENT:G2519F-015241** track under Oligo
probes in the left hand menu, or by using the Find a track box at the top right. Turn on the track then save and close the menu.

As the AGILENT:G2519F-015241 track is stranded, it appears at the top and bottom of the view, in green.

There are five probes mapped to this region on the positive strand and one probe on the reverse strand.

(c) Drag a box around the reverse strand probe then click on Mark region to highlight.

The highlighted region maps to three transcripts: LOC_Os01g01790, Os01t0107900-02 and Os01t0107900-01

Exercise 9 – Exploring a genomic region in bacteria

(a) Go to the Ensembl Bacteria homepage.

Type CT18 into the Search for a Genome box. Click on the auto-completed genome name to navigate to the species homepage.

(b) Type Chromosome:2000605-2009742 into the search box.

Click Go.

(c) There are eleven genes annotated in this region. Two are on the forward strand, nine are on the reverse strand.

Ensembl genes

This webinar is presented by Emily Perry.

Ensembl uses an automated pipeline to annotate genomic assemblies and this annotation is presented alongside Havana manual annotation in the browser.

Focusing on the gene and transcript tabs within the Ensembl browser, we will explain the process of genome annotation and explore the range of data available for genes and transcripts.

If you have trouble accessing YouTube you can download the video [20].

You can also download the slides [21] and demo [22].

Ensembl genes and transcripts exercises

Finding out about Ensembl genes and transcripts
Exercise 1 – Exploring the human *MYH9* gene

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

- On which chromosome and which strand of the genome is this gene located?
- How many transcripts (splice variants) are there and how many are protein coding?
- What is the longest transcript, and how long is the protein it encodes?
- Since this gene does not have a MANE Select, which transcript would you take forward for further study?

(b) Click on Phenotype at the left side of the page. Are there any diseases associated with this gene, according to OMIM (Online Mendelian Inheritance in Man)?

(c) What are some functions of *MYH9* according to the Gene Ontology consortium? Have a look at the GO pages for this gene.

(d) In the transcript table, click on the transcript ID for MYH9-201, and go to the Transcript tab.

- How many exons does it have?
- Are any of the exons completely or partially untranslated?
- Is there an associated sequence in UniProtKB/Swiss-Prot? Have a look at the General identifiers for this transcript.

(e) Are there microarray (oligo) probes that can be used to monitor ENST00000216181 expression?

Exercise 2 – Finding a gene associated with a phenotype

Phenylketonuria is a genetic disorder caused by an inability to metabolise phenylalanine in any body tissue. This results in an accumulation of phenylalanine causing seizures and mental retardation.

(a) Search for phenylketonuria from the Ensembl homepage and narrow down your search to only genes. What gene is associated with this disorder?

(b) How many protein coding transcripts does this gene have? View all of these in the transcript comparison view.

(c) What is the MIM gene identifier for this gene?

(d) Go to the MANE Select transcript and look at its 3D structure. In the model 2pah, how many protein molecules can you see?

Exercise GT3 – Exploring a plant gene (*Vitis vinifera*, grape)

Start in http://plants.ensembl.org/index.html and select the *Vitis vinifera* genome.

(a) What GO: biological process terms are associated with the *MADS4* gene?

(b) Go to the transcript tab for the only transcript, Vv01s0010g03900.t01. How many exons does it have? Which one is the longest? How much of that is coding?

(c) What domains can be found in the protein product of this transcript? How many different domain prediction methods agree with each of these domains?
**Exercise 4 - Exploring the mouse Dpp6 gene**

Genetic variation in the dipeptidylpeptidase 6 Gene (DPP6) in humans has previously been strongly associated with amyotrophic lateral sclerosis (ALS), a lethal disorder caused by progressive degeneration of motor neurons in the brain.

(a) Search for the Dpp6 gene in mouse and click on the ENSMUST00000071500 transcript to open the transcript tab. How many exons make up this transcript?

(b) Click on Exons to display the exon sequences of the transcript. Which exon contains the translation start? What is the exon ID of the largest exon? What is the start and end phase of exon 2?

(c) Click on Protein summary. How many protein domains or features fall within the second exon? What is the PFAM protein domain at the C-terminus of the protein and how many exons does it fall into? Which amino acid positions does the domain above cover?

(d) Click on Domains and features. Which domains are associated with Pfam? How many genes in the mouse genome have the IPR002469 domain? What chromosomes are these genes found on?

---

**Exercise 5 – Exploring pathway interactions of a human gene**

Navigate to the Gene tab for the human STAR gene.

(a) What pathway(s) has STAR been implicated in?

(b) Scroll to zoom in on STAR in the diagram for Pregnenolone biosynthesis. Where is STAR located? Does this correlate with information from the information in GO:Cellular Component?

---

**Exercises solutions**

### Finding out about Ensembl genes and transcripts - solutions

**Exercise 1 – 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 MYH9-201 and it codes for a protein of 1,960 amino acids
- MYH9-201 is the transcript I would take forward for further study, as it is an APPRIS P1, has a CCDS associated with it, is TSL:1 and is Golden.

(b) Click on Phenotypes at the left to see the associated phenotypes. There is a large table of phenotypes. To see only the ones from MIM, type mim into the filter box at the top right of the table.
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.


(d) Click on **ENST00000216181**

- It has 41 exons. This is shown in the Transcript summary or in the left hand side menu **Exons** page.
- 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.

(e) 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/](http://www.ebi.ac.uk/arrayexpress/) [25]

---

**Exercise 2 – Finding a gene associated with a phenotype**

(a) Start at the Ensembl homepage ([http://www.ensembl.org](http://www.ensembl.org)) [23]).

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 six protein coding transcripts in release 90.

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 gene ID is 612349.

(d) Open the transcript table and click on the ID for the MANE Select: **ENST00000553106.6**. Go to **3D protein model** in the left-hand menu.

The model 2pah is shown by default. It has two protein molecules in it. You may need to rotate the model to see this clearly.

---

**Exercise 3 – Exploring a plant gene (**Vitis vinifera**, grape)**

(a) Go to [http://plants.ensembl.org/index.html](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 **Vitis vinifera**.
Type MADS4 and click on the gene link VIT_01s0010g03900.

Click on GO: Biological process in the side menu.

There are six 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. 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 MADS-box domain near the N-terminus is identified by eight domain prediction methods. A K-box domain near the C-terminus is identified by two. Two coiled-coils are identified by one.

Exercise 4 - Exploring the mouse Dpp gene

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

Select Search: Mouse and type Dpp6. Click Go.

Click on either the Ensembl ID ENSMUSG00000061576 or the MGI official gene name Dpp6. From the transcript table, click on the link for transcript ENSMUST00000071500 to open the transcript tab.

ENSMUST00000071500 consists of 26 exons.

(b) Click on Exons, which can be found on the left of the page. The translation start is found in the first exon (ENSMUSE00000725552), shown in dark blue text.

The largest exon is the final exon (856 bp), which has the exon ID ENSMUSE00000773588.

Exon 2 has a start and end phase of 0 and 1 respectively, which means that the codon at the start of the exon starts at the first nucleotide and the codon at the end of the exon ends at nucleotide 2. Notice that the end phase of each exon is the same as the start phase of the next exon.

(c) Click on Protein Summary in the menu on the left hand side of the page. Alternating exons are shown on the protein as different shades of purple.

There are two predicted protein domains that fall within the second exon: low complexity (seg) and a transmembrane helix.

Click on a domain or feature to view further information.

The C-terminal PFAM domain is Peptidase_S9B (PF00326), which spans or partially spans seven exons, covering amino acid positions 582-787.

(d) Click on Domains & features.

Looking at the Domains table you should notice that there are two domains associated with Pfam: PF00326 and PF00930.

Click on Display all genes with this domain next to IPR002469. This should now display the genes that have the IPR002469 domain located on the karyotype and as a table.
Six genes have this domain and they are found on chromosomes 1, 2, 5, 9 and 17.

Exercise 5 – Exploring pathway interactions of a human gene

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

Select Search: Human and type STAR. Click Go. Click on either the Ensembl ID ENSG00000147465 or the gene name STAR to navigate to the gene tab.

(a) Click on Pathway in the menu on the left hand side of the page.

STAR is involved in the Pregnenolone biosynthesis pathway, as well as metabolism, metabolism of lipids, metabolism of steroids and metabolism of steroid hormones.

(b) Scroll to zoom in on STAR in the diagram for Pregnenolone biosynthesis.

STAR is located in the mitochondrial intermembrane space.

Click on GO:Cellular Component in the menu on the left hand side of the page.

GO terms locate the STAR protein in the mitochondrial intermembrane space and the mitochondrial cristae (as well as the cytosol and neuronal cellbody and projections).

Variation data in Ensembl and the Ensembl VEP

This webinar is presented by Emily Perry.

Ensembl imports variation and phenotype data from a number of sources. In this webinar, we will learn how to find variants in genes and regions, and access additional information, including population frequencies.

We will then introduce the Variant Effect Predictor (VEP), a tool which allows you to analyse your own variation data for potential effects on genes.

If you have trouble accessing YouTube you can download the video [26].

You can also download the slides [27] and demo [28].

Variation exercises

Known variants in Ensembl

Exercise 1 — Human population genetics and phenotype data

The SNP rs1738074 in the 5’ UTR of the human TAGAP gene has been identified as a genetic risk factor for a few
Ensembl browser webinar series
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

diseases.

(a) In which transcripts is this SNP found?

(b) What is the least frequent genotype for this SNP in the Yoruba (YRI) population from the 1000 Genomes phase 3?

(c) What is the ancestral allele? Is it conserved in the 91 eutherian mammals?

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

Exercise 2 — Exploring a SNP in human

The missense variation rs1801133 in the human \textit{MTHFR} gene has been linked to elevated levels of homocysteine, an amino acid whose plasma concentration seems to be associated with the risk of cardiovascular diseases, neural tube defects, and loss of cognitive function. This SNP is also referred to as ‘A222V’, ‘Ala222Val’, as well as other HGVS names.

(a) Find the page with information for rs1801133.

(b) Is rs1801133 a Missense variation in all transcripts of the \textit{MTHFR} gene? What is the amino acid change?

(c) Why are the alleles for this variation in Ensembl given as G/A and not as C/T, as in the literature?

(d) What is the major allele in rs1801133?

(e) In which paper(s) is the association between rs1801133 and homocysteine levels described?

(f) According to the data imported from dbSNP, the ancestral allele for rs1801133 is G. Ancestral alleles in dbSNP are based on a comparison between human and chimp. Does the sequence at this same position in other primates confirm that the ancestral allele is G?

Exercise 3 — Exploring a SNP in mouse

Madsen \textit{et al} in the paper ‘Altered metabolic signature in pre-diabetic NOD mice’ (PloS One. 2012; 7(4): e35445) have described several regulatory and coding SNPs, some of them in genes residing within the previously defined insulin dependent diabetes (IDD) regions. The authors describe that one of the identified SNPs in the murine \textit{Xdh} gene (rs29522348) would lead to an amino acid substitution and could be damaging as predicted as by SIFT (http://sift.jcvi.org/).

(a) Where is the SNP located (chromosome and coordinates)?

(b) What is the HGVS recommended nomenclature for this SNP?

(c) Why does Ensembl put the C allele first (C/T)?

(d) Are there differences between the genotypes reported in NOD/LTJ and BALB/cByJ, according to the PERLGEN panel?

Exercise 4 – Variation data in the tomato (\textit{S. lycopersicum}) genome

(a) Find the Solyc02g084570.3 gene in tomato and go to its Location tab. Can you see the variation track?
(b) Zoom in around the last exon of this gene. What are the different types of variants seen in that region? What are the locations of any splice region variants mapped in the region?

**Exercise 5 – Variation data in fungi**

(a) How many species in Ensembl Fungi have variation data?

(b) Select *Fusarium oxysporum* and search for FOXG_13574T0 gene. One of its upstream variants is SNP tmp_10_6610. What are the possible alleles for this polymorphic position? Which one is on the reference genome?

(c) What is the most frequent allele at this position?

(d) Which samples have the genotypes C|T and T|T?

**Exercise 6 – Variation structure viewer**

You are interested in the missense variant rs998717588.

(a) Navigate to the Variant tab and click on 3D Protein model.

(b) Identify the location of the variant rs998717588. Is it located in a beta strand, helix or linking region?

(c) Is it in a pfam domain?

**VEP**

**Exercise 7 — The VEP**

Resequencing of the genomic region of the human *CFTR* (cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7) gene (ENSG00000001626) has revealed the following variants (alleles defined in the forward strand):

- G/A at 7: 117,530,985
- T/C at 7: 117,531,038
- T/C at 7: 117,531,068

Use the VEP tool in Ensembl and choose the options to see SIFT and PolyPhen predictions.

(a) Do these variants result in a change in the proteins encoded by any of the Ensembl genes? Which gene?

(b) Do these variants already exist in the Ensembl database?

(c) Which of these variants is predicted to be the most damaging, based on the SIFT and PolyPhen scores?
Exercise 8 – viewing structural variants with the VEP

We have details of a genomic deletion in a breast cancer sample in VCF format:

```
13 32307062 sv1 . <DEL> . . SVTYPE=DEL;END=32332466
```

(a) How many genes have been affected?
(b) Does the SV cause deletion of any complete transcripts?
(c) Display your variant in the Ensembl browser.

Exercise solutions

Known variants in Ensembl

Exercise 1 – Human population genetics and phenotype data

(a) First, find the variant tab for the variant rs1738074. Search Ensembl for rs1738074 directly from the homepage.

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

This SNP is found in four transcripts of TAGAP. It is also intronic to five non-coding transcripts.

(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 (YRI), the least frequent genotype is CC at the frequency of 5.6%.

(c) Click on Phylogenetic context.

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

Select the 91 eutherian mammals EPO LOW COVERAGE alignment and click on Apply.

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 seven of the 75 eutherian mammals displayed.

(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 2 – Exploring a SNP in human

(a) Go to the Ensembl homepage (http://www.ensembl.org/ [29]).

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.

The amino acid change is A/V.

(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 Medellín; 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: 13 primates EPO** and click **Go**.

Gorilla, bonobo, orangutan, chimp, macaque, gibbon, vervet, crab-eating macaque, mouse lemur, olive baboon and marmoset all have a G in this position.

**Exercise 3 – 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 chromosome 17:73924993. In Ensembl, variant alleles are always provided as on the forward strand.

(b) Find the ‘HGVS names’ header in the summary information. Click on **Show** to reveal information about HGVS nomenclature.

This SNP has got five HGVS names, one at the genomic DNA level (NC_000083.6:g.73924993C>T), three at the transcript level (ENSMUST0000024866.4:c.721G>A, ENSMUST0000023162.1:n.738G>A and ENSMUST0000023621.1: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 **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.

**Exercise 4 – Variation data in the tomato (*S. lycopersicum*) genome**

(a) Search for **Solyc02g084570.3** and click on the Location link in the results page. The variation track is shown at
(b) Zoom in around the last exon of this gene by drawing a box in the respective region. Please note the gene is on the reverse strand, so the last exon will be on the left hand side of that image.

The variation legend is shown at the bottom of the page, telling you what the colours mean.

The types of variants seen in that region are 3’ UTR variants, missense variants, synonymous variants and splice region variants.

Splice region variants are shown in orange. Click on the variants to get additional information on that variant including location.

The variants are found at 2:48285642 2:48285640-48285641.

**Exercise 5 – Variation data in fungi**

(a) Go to fungi.ensembl.org, click on View full list of all Ensembl Fungi species.

Click on the upward triangle next to the Variation database column to sort the table by species with variation data.

The table shows that we have eight fungi species currently with variation databases.

(b) Click on Fusarium oxysporum in the table and on the species page search for FOXG_13574T0.

From the Gene tab, click on Variant table and then scroll down to find tmp_10_6610 or use the table search box to find it.

The alleles are C/T, where C is the reference allele.

(c) Click on tmp_10_6610 in the table to open the Variant tab. Then click on Genotype Frequency from the menu on the left hand side of the page.

The most frequent allele at this position is C with a frequency of 0.850.

(d) Click on Sample genotypes in the left menu.

The table shows that sample 909454 has the C|T genotype and 909455 has the T|T genotype.

**Exercise 6 – 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, to see an information box in the top left hand corner of the image.

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).

VEP

Exercise 7 – VEP

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

7 117530985 117530985 G/A
7 117531038 117531038 T/C
7 117531068 117531068 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](http://www.ensembl.org/info/docs/variation/vep/vep_formats.html)

Click Run.

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

(a) 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.

The variants with ‘missense’ as their consequence annotation are those which cause a change in the protein sequence. All these variants affect the CFTR gene (ENSG00000001626). Note that you can also upload the VEP results as a track and view them on Location pages in Ensembl, just click on the link in the ‘Location’ column.

(b) Scroll across the table and find the ‘Existing variant’ column.

You can see that all these variants match an existing record for dbSNP (rs IDs) and some match COSMIC (COSM IDs) records.

(c) SIFT and PolyPhen are available for missense SNPs only.

The missense variant at coordinate 7 117531038 T/C been predicted to be probably damaging/deleterious. The other missense variant at coordinate 117531068 is predicted to be benign/tolerated.

Exercise 8 – viewing structural variants with the VEP

(a) Give your data a name, such as Patient deletion.

Paste 13 32307062 sv1 . <DEL> . . SVTYPE=DEL;END=32908738

into the Paste data field then hit Run.
13 genes have been affected.

(b) Use the Filters, selecting Consequence is transcript_ablation.

Yes, there is deletion of complete transcripts of PDS5B, N4BP2L1, BRCA2, RNY1P4, RF00190, IFIT1P1, ATP8A2P2, N4BP2L2, AL137247.1 and AL138820.1.

(c) To view your variant in the browser click on the location link in the results table [13: 32307062-32908738].

The link will open the Region in detail view in a new tab. If you have given your data a name it will appear automatically in red. If not, you may need to Configure this page and add it under Your data.

Comparing genes and genomes with Ensembl Compara

This webinar is presented by Emily Perry.

Ensembl Compara allows you to perform detailed analysis of gene models between species.

During this webinar we take a look at the gene trees and homologues of a set of genes, and at whole genome alignments between pairs and groups of species.

If you have trouble accessing YouTube you can download the video [31].

You can also download the slides [32] and the demo [33].

Compara exercises

Exercise 1 – Orthologues, paralogues and gene trees for the human BRAF gene

(a) How many orthologues are predicted for this gene in primates? How much sequence identity does the Tarsier (Carlito syrichta) protein have to the human one? Click on the Alignment link next to the Ensembl identifier to view a protein alignment in Clustal format.

(b) Go to the gene tree for this gene. View the Wasabi alignment of all the proteins in primates.

Exercise 2 –Cow orthologues

Go to www.ensembl.org to find the sardh gene in the cow genome.

(a) Go to the Location tab for this gene. View the Alignments (image) and Alignments (text) for the 38 eutherian mammals. Which mammalian genomes are represented in the alignment? Do all the mammals show a gene in these alignments?
(b) Export the alignments (as Clustal).

(c) Click on the Region in detail link at the left and turn on the tracks for multiple alignments, constrained elements and conservation score for 91 eutherian mammals by configuring the page.

What is the difference between the 91 eutherian mammals EPO multiple alignment track and the constrained elements track? Which regions of the gene do most of the constrained element blocks match up to?

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

Exercise 3 – Synteny

Go to www.ensembl.org. Find the Rhodopsin (RHO) gene for Human. Go to the Location tab.

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

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

Exercise 4 – Whole genome alignments

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

(b) Turn on the LASTZ-net alignment tracks for chicken, chimp, mouse and platypus. 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 91 mammals and the set of 54 amniota 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.

Exercise solutions

Exercise 1 – 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 in the menu on the left to see all orthologues.

Of the 23 primates in Ensembl, there are 1-to-1 orthologues in 20 primates and 1-to-many orthologues in one (chimpanzee).

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 95.26%. 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 94.52% (the query %id).

Note that the difference in the values of the Target and Query %id reflects the different protein lengths for the human and tarsier BRAF genes.
(b) Click on Gene tree in the left hand menu. All of the primates are enclosed in a blue box. Click on the furthest left node in the box to get a pop-up labelled Primates. Click on Wasabi viewer to see the alignment.

Exercise 2 – Cow orthologues

(a) Start in the Location tab (region in detail) for sardh. Click on Alignments (Image) at the left, and select the 38 eutherian mammals EPO_LOW_COVERAGE alignment in the view and click Apply.

The Human, Bonobo, Gorilla, Chimpanzee, Orangutan, Macaque, Olive baboon, Mouse lemur, Mouse, Rat, Shrew mouse, Prairie vole, Algerian mouse, Ryukyu mouse, Horse, Cat, Dog, Pig and Cow are shown in this region. All the species (with the exception of Crab-eating macaque, Marmoset, Gibbon, Goat, Rabbit, Sheep and Vervet-AGM) have an alignment to this genomic region in cow. This can also be seen in the Alignments (text) page (the exons are highlighted in red).

(b) You can export the alignments from the Alignments (text) pages in the Location tab. Click on the blue Download alignment button at the top of the image, and choose CLUSTALW.

(c) Click on Region in detail in the left hand menu. Click on Configure this page and turn on the Multiple alignments, Constrained elements and Conservation score for 91 eutherian mammals EPO tracks, all under Comparative genomics.

The 91 eutherian mammals EPO tracks just shows that the region for the sardh gene can be aligned. The Constrained elements and Conservation score tracks show where the conserved sequence is located in the alignment.

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

Click on the track name Track name, then the (information button) and the link GERP conservations scores to read more about constrained elements (or any other data track).

Exercise 3 – Synteny

(a) Change the species to Dog using the drop-down menu below the image and click Go.

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.

Exercise 4 – Whole genome alignments

(a) Go to the Ensembl homepage www.ensembl.org.
Select Search: Human and type brca2 in the search box.
Click Go and then on 13:32315474-32400266:1 below BRCA2 (Human Gene). This link takes you to the Region in detail view of the gene location.

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.
Close the menu.

(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 on the Location tab to go back to the main view.
Click Configure this page in the side menu.
Click on Conservation regions under the Comparative genomics menu.
Select Conservation score and Constrained elements for 91 eutherian mammals EPO-Low-Coverage, Conservation score and Constrained elements for 54 amniota vertebrates Mercator-Pecan.
Close the menu.

Both the Conservation score and Constrained elements tracks largely correspond with the data seen in the pairwise alignment; 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.
Close the menu.

The conserved regions in the alignment at the bottom of the page will be highlighted in light blue.

**Finding features that regulate genes – the Ensembl Regulatory Build**

This webinar is presented by Astrid Gall.

The Ensembl Regulatory Build incorporates data from sources including ENCODE, Roadmap Epigenomics and Blueprint to predict the positions of features involved in regulating gene expression, such as promoters and enhancers. Learn about how the build works and how to find regulatory features on the genome.

If you have trouble accessing YouTube you can download the video [34].

You can also download the slides [35] and the demo [36].
Regulation exercises

Exercise 1 – Gene regulation: Human PPP1R37

(a) Find the Location tab (Region in detail page) for the region between the genes ASCL2I and TSPAN32. Are there any predicted enhancers in this region?

(b) Go to the regulation tab for the enhancer ENSR00000952450. How many cell types is this enhancer active in? Are there any cell types where its activity is repressed?

(c) Go to the Details by cell type page. Take a look at the histone modifications across this promoter in NK (PB) cells, where this promoter is active, compared to NPC_1 cells, where it is poised. What differences can you observe?

(d) Go back to the Summary page. Are there any verified transcription factor binding motifs in this enhancer? In what cells? Can you see the binding peaks of these transcription factors?

Exercise 2 – Regulatory features in human

(a) Search for the regulatory feature ENSR00000262400. What type of feature is this? What is its genomic location?

(b) Which cell types is this feature inactive or repressed in? Look at the details by cell type, why do you think it has been called differently in these cell types?

(c) Why do so many cells have this feature listed as NA?

Exercise solutions

Exercise 1 – Gene regulation: Human PPP1R37

(a) Search for human ASCL2I from the home page. Click on Location in the search results. In the region overview, drag out a box to encompass the neighbouring gene TSPAN32I.

There is one yellow enhancer in the region between these genes.

(b) Click on the enhancer to get a pop-up with its ID and other information. Click on the ID ENSR00000952450.

ENSR00000952450 is active in five cell types. It is repressed in 32.

(c) Click on Details by cell type then choose cells by clicking on clicking on Configure cell/tissue, selecting NK (PB) and NPC_1. Add evidence tracks by clicking on Evidence and Select all in all categories. Click View tracks to load the page.

In NK (PB) cells there is H3K4me1 modification, whereas in NPC_1 cells there is H3K27me3 modification.

(d) In the motifs track there are two black markers indicating verified TF motifs. Click on them to cell which motifs and which cells.

There is an NFATC1 motif verified in GM127878 cells and a MNT, MAX, CLOCK, HEY2, HES5 motif verified in HepG2 cells.
Go to the Details by cell type page and use the matrix to turn on the TF binding tracks for these cells and TFs. You will find that not all the TFs are available for HepG2.

You can see binding of NFATC1 in GM12878 and binding of MNT in HepG2.

Exercise 2 – Regulatory features in human
(a) Search for ENSR00000262400 from the home page and click through to the search results,

ENSR00000262400 is a CTCF binding site found at chromosome 11: 1,998,201-2,003,400.

(b) The CTCF binding site is inactive in A673 and HepG2 cells.

Click on Details by cell type then choose cells by clicking on Configure cell/tissue, selecting the cell types listed above. Add evidence tracks by clicking on Evidence and Select all. Click View tracks to load the page.

There is no evidence of any histone modifications or transcription factor binding in the cell types where this CTCF binding site is inactive, but there is H3K9me3 modification in HepG2, an inactivating modification.

(c) Cells which do not have CTCF ChIP-seq data cannot have an activity listed for this feature.

Data export with BioMart

This webinar is presented by Ben Moore

As well as browsing genome information in Ensembl you can export data directly from the database.

BioMart is a powerful tool that allows you to export customised data, using a simple point-and-click interface.

This webinar covers the principles of the tool and how to perform advanced searching to export tables of Ensembl data.

If you have trouble accessing YouTube you can download the video [37].

You can also download the slides [38] and the demo [39].

BioMart exercises
Using BioMart to Export Data from Ensembl

Exercise 1 — Finding Genes by Protein Domain
Download the peptide sequences of all mouse proteins with Signalp cleavage sites located on chromosome 9.

Exercise 2 – Export homologues
Export the human orthologues of these Ciona savignyi Ensembl genes:
ENSCSAVG00000000002
ENSCSAVG00000000003
ENSCSAVG00000000006
ENSCSAVG00000000007
ENSCSAVG00000000009
ENSCSAVG00000000011

Exercise 3 – Convert IDs
BioMart is a very handy tool when you want to convert IDs from different databases. Below is a list of 29 IDs of human proteins from the NCBI RefSeq database:
NP_001218, NP_203125, NP_203124, NP_203126, NP_001007233, NP_150635, NP_001214, NP_150637, NP_150634, NP_150649, NP_001216, NP_116787, NP_001217, NP_127463, NP_001220, NP_004338, NP_004337, NP_116786, NP_036246, NP_116756, NP_116759, NP_001221, NP_203519, NP_001073594, NP_001219, NP_001073593, NP_203520, NP_203522

Generate a list that shows to which Ensembl Gene IDs and to which HGNC symbols these RefSeq protein IDs correspond. Do these 29 proteins correspond to 29 genes?

Exercise 4 – Export variants
You can use BioMart to query variants, not just genes. (Make sure you use the right Datasets.)

(a) Export the study accession, source name, chromosome, sequence region start and end (in bp) of human structural variations (SV) on chromosome 1, starting at 130,408 and ending at 210,597.

(b) In a new BioMart query, find the alleles, phenotype descriptions, and associated genes for the human SNPs rs566014072 and rs754099015. Can you view this same information in the Ensembl browser?

Exercise 5 – Find genes associated with array probes
Forrest et al. performed a microarray analysis of peripheral blood mononuclear cell gene expression in benzene-exposed workers (Environ Health Perspect. 2005 June; 113(6): 801–807). The microarray used was the human Affymetrix U133A/B (also called U133 plus 2) GeneChip. The top 25 up-regulated probe-sets were:

207630_s_at, 221840_at, 219228_at, 204924_at, 227613_at, 223454_at, 228962_at, 214696_at, 210732_s_at, 212370_at, 225390_s_at, 227645_at, 226652_at,
221641_s_at, 202055_at, 226743_at, 228393_s_at, 225120_at, 218515_at, 202224_at, 200614_at, 212014_x_at, 223461_at, 209835_x_at, 213315_x_at

(a) Retrieve for the genes corresponding to these probe-sets the Ensembl Gene and Transcript IDs as well as their HGNC symbols and descriptions.

(b) In order to analyse these genes for possible promoter/enhancer elements, retrieve the 2000 bp upstream of the transcripts of these genes.

(c) In order to be able to study these human genes in mouse, identify their mouse orthologues. Also retrieve the genomic coordinates of these orthologues

Exercise B6 – Exporting paralogues with BioMart

Export a list of all human genes on chromosome 14 which have a parologue, including the gene names, the last common ancestor and the identity between the genes.

Exercise B7 – Exporting histone modification sites

Using the Human Regulatory Evidence dataset, export a list of all H3K9me3 modified loci in Aorta.

Exercise solutions

Exercise 1 — 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: Signalp cleavage sites 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.

Step 1: Dataset: Select Ensembl Genes as your database, and then select Mouse genes as the dataset.

Step 2: Filters: click on Filters on the left of the screen

Expand REGION. Change the chromosome to 9.

Scroll down and expand the PROTEIN DOMAINS section, and select Limit to genes, choosing with With Cleavage site (Signalp) from the drop-down and then Only. Clicking on Count should reveal that you have filtered the dataset down to 218 genes.

Step 3: Attributes: click on Attributes on the left of the screen

Select Sequences. Expand Sequences and select Peptide, it may already be selected.

Step 4: Results: Click the Results button at the top left of the page.
The first 10 results are displayed by default; to download your results click GO. Note that we only have the option to download as FASTA format because we are downloading sequences, other format options are available for exporting tables.

Exercise 2 — Export Homologues

Click the New button at the top left of the page.

**Step 1: Dataset** Choose the Ensembl Genes database and then the Ciona savignyi genes dataset.

**Step 2: Filters**

Expand the GENE section and enter the gene list in the Input external references ID list box.

Note that you have to ensure the format you are inputting must match the format in the drop down menu above the box. You can check using the Count button that your IDs have been accepted.

**Step 3: Attributes**

Select the Homologues option at the top of the Attributes page, expand the ORTHOLOGUES section, scroll down to find the Human Orthologues section and choose Human Ensembl Gene ID.

**Step 4: Results**

Exercise 3 — Convert IDs

Click New.

**Step 1: Dataset** Choose the Ensembl Genes database and then the Human genes dataset.

**Step 2: Filters**

Expand the GENE section, select Input external references ID list. From the drop down list choose RefSeq peptide ID(s) [e.g. NP_001001130] and enter the list of IDs in the text box (either comma separated or as a carriage-returned list).

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

**Step 3: Attributes**

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.

**Step 4: Results**

Select View All rows as HTML or export all results to a file.

Exercise 4 — Export Variants

(a) Click New.
Step 1: Dataset
Choose Ensembl Variation and Human Structural Variants.

Step 2: Filters
Expand REGION and select Chromosome 1, Base pair start: 130408, Base pair end: 210597. Also expand GENERAL STRUCTURAL VARIANT FILTERS and click on Limit to Variants from source: DGVa if this is not already selected.

Click on count, this shows 87 structural variants.

Step 3: Attributes
Click Study accession and Source Name. Ensure that Chromosome/scaffold name, position start and end are selected.

Step 4: Results
Click Unique Results only and expand the preview table to All

(b) Click New.

Exercise 5 — Find Genes Associated with Array Probes
(a) Click New.

Step 1: Dataset
Choose the Ensembl Genes database, then the Human genes dataset.

Step 2: Filters
Expand the GENE section and select Input microarray probes/probesets ID list. Choose AFFY HG U133 Plus 2 probe ID(s) [e.g. 1553551_s_at] from the drop down list above and enter the list of probeset IDs in the text box (either comma separated or as a list). Count shows that 27 genes match this list of probesets.

Step 3: Attributes
Expand GENE, select Description (Gene and Transcript IDs are already selected). Scroll down and expand the EXTERNAL. Find the External References section and choose HGNC symbol, scroll down to find the Microarray
Step 4: Results

Select View All rows as HTML or export all results to a file. Tick the box Unique results only.

(b) Don’t change Dataset and Filters – simply click on Attributes.

Step 3: Attributes

Select the Sequences option at the top of the attributes page.

Expand the SEQUENCES section. Select Flank (Transcript) and enter 2000 in the Upstream flank text box. Expand the HEADER INFORMATION section. Select, in addition to the default selected attributes, Gene description and 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).

Step 4: Results

Download the FASTA file.

(c) Don’t change Dataset and Filters – simply click on Attributes.

Step 3: Attributes

Select the Homologues option at the top of the attributes page.

Expand the GENE section, select Gene name and deselect Transcript stable ID. Expand the ORTHOLOGUES.

Scroll down to find the Mouse Orthologues section. Select Mouse gene stable ID, Mouse chromosome/scaffold name, Mouse chromosome/scaffold start (bp) and Mouse chromosome/scaffold end (bp).

Step 4: Results:

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.

Exercise 6 – Exporting paralogues with BioMart

Start at ensembl.org/biomart/martview. Choose the Ensembl Genes database. Choose the Human Genes dataset.


Click on Attributes in the left panel. Select Homologues from the six options at the top. Under GENE: deselect Ensembl Transcript ID and select gene name. Under PARALOGUES: select Human Parologue Ensembel Gene ID, Human Parologue gene name, Last common ancestor with Human, %id. target Human gene identical to query gene and %id. query gene identical to target Human gene.

Click the Results button on the toolbar. Select View All rows as HTML or export all results to a file.
Exercise 7 – Exporting histone modification sites

Chose the Ensembl Regulation – Human Regulatory Evidence dataset.

In Filters, expand REGULATORY EVIDENCE. Select Feature Type – H3K9me3. Select Epigenome – Aorta.

Click Results to see the table.

### Uploading your data to Ensembl

This webinar was held on the 13th September 2018, presented by Astrid Gall.

As well as exploring genomic data through the web interface, you are also able to upload data to view within the browser.

This webinar will show you how you can view your own data, such as BED or BAM files, in the Ensembl browser. You will also learn about Track Hubs and how to find and visualise them in Ensembl.

If you have trouble accessing YouTube you can watch the video on YouKu. [40]

You can also download the slides [41].

### Data upload exercises

**Exercise 1 - Attach URLs of large files**

Larger files, such as BAM files generated by NGS, need to be attached by a URL. There is a BAM file of human chromosome 20 RNAseq data online at: https://www.ebi.ac.uk/~emily/Workshops/BAM/ [42]

Here you can see a number of BAM files (.bam) with corresponding index files (.bam.bai). We are interested in the files GRCh38.20.illumina.merged.1.bam and GRCh38.20.illumina.merged.1.bam.bai. These files are the BAM file and the index file, respectively. When attaching a BAM file to Ensembl, there must be an index file in the same folder, but only the BAM file itself needs to be uploaded to Ensembl.

(a) Attach and view the BAM file of human chromosome 20 RNAseq data.

(b) Go to the region on chromosome 20 that contains the CDH22 gene. Configure the page to show your added track in the ‘Unlimited’ style. What is the relationship between the number of RNAseq reads and the exons of CDH22?

(c) Zoom onto exon 1 of CDH22 so that you can see the the sequence of the individual RNAseq reads.

(d) Remove the track from your Region in detail view.
Exercise 2 - Track Hubs

(a) Add the ENCODE Analysis Hub to the Region in detail view for the genomic region surrounding the BRCA2 gene.

Hint: You will need to add and view this Track Hub to the human GRCh37 genome assembly.

(b) Turn on all the available tracks relating to Histone Modification Peaks and Transcription Factor Peaks in HeLa-S3 cells.

(c) Which Transcription Factors and Histone Modifications are annotated in this region?

(d) Add the tracks showing Signals for the ENCODE Histone Modifications and Transcription Factors that have peaks in this region. Compare the signal intensity to the location of annotated peaks.

(e) Remove the ENCODE Analysis Hub from your list of custom tracks.

Exercise solutions

Exercise 1 - Attach URLs of large files

(a) There are two main ways to attach a file containing your own data to view in Ensembl. Either:

From the Ensembl homepage, click on Use my own data in Ensembl, then click on Custom tracks to add a new track.

Or:

Click on the Custom tracks button in any Region in detail view in Ensembl.

A dialogue box labelled ‘Add a custom track’ will appear. We can name our data, for this exercise we will label our data ‘Illumina reads’.

Paste the URL of the BAM file itself ([http://www.ebi.ac.uk/~emily/Workshops/BAM/GRCh38.20.illumina.merged.1.bam](http://www.ebi.ac.uk/~emily/Workshops/BAM/GRCh38.20.illumina.merged.1.bam)) into the data box.

Since this is a file, the interface is able to detect the “.BAM” file extension, thus it automatically labels the format as BAM. Click on Add data and close the menu.

(b) Search for the CDH22 gene and click on the Location tab. Click on Configure this page, and then on Your data in the menu. Select the Unlimited track style for your ‘Illumina reads’ track. Close the menu.

You can see that there are more RNASeq reads that map to the exons than reads that map to the introns of the gene.

(c) Zoom in to see the sequence itself by dragging out boxes in the view to zoom in or use the scale bar in the top right of the Region in detail image.

(d) Click on Configure this page and turn off this ‘Illumina reads’ track by selecting Off as the track style of the Your data track.

You can also remove the custom data by clicking on the tab Personal Data and then clicking on the ‘Trash’ icon associated with this data.

Exercise 2 - Track Hubs

(a) There are two ways to add the ENCODE Analysis Hub to view in Ensembl. Either:
Search for Encode from the Track Hub Registry homepage (https://trackhubregistry.org [44]). Find the ENCODE Analysis Hub in the search results, click on View in Genome Browser and select Ensembl.

Once the Track Hub is added, search for the BRCA2 gene in the GRCh37 human genome assembly. Switch to the Location tab.

Or:
Search for the BRCA2 gene in the GRCh37 human genome assembly. Switch to the Location Tab.

Click on the Custom tracks button. Click Track Hub Registry Search and Search for Encode. Click Attach this hub for the ENCODE Integrative Analysis Data Hub.

(b) Click on Configure this page and click on ENCODE Histone Modifications Peaks. Turn on all available tracks for HeLa-S3 cells by hovering over the cell line name and clicking Select all HeLa-S3 in the pop up window (the boxes for HeLa-S3 will turn blue).

Click on ENCODE Transcription Factor Peaks. Turn on all available tracks for HeLa-S3 cells by hovering over the cell line name and clicking Select all HeLa-S3 in the pop up window (the boxes for HeLa-S3 will turn blue).

Close the menu to add the selected tracks.

(c) A number of different Transcription Factors and Histone Modifications, mainly surrounding the BRCA2 5’ region, are annotated.

Transcription Factors:
USF2, STAT1, Pol2S2, Mxl1, MAX, InI1, E2F6, E2F4, MYC, C-MYC, CEBPB, POLR2A, POL2, POL2B

Histone Modifications:
H3K4me3, H3K9ac, H3K79me2, H3K4me3, H3k4me2, H3K36me3, H3K27ac

(d) Click on Configure this page and click on ENCODE Histone Modifications Signal. Turn on all available tracks for HeLa-S3 cells by hovering over the cell line name and clicking Select all HeLa-S3 in the pop up window (the boxes for HeLa-S3 will turn blue).

Click on Configure this page and click on ENCODE Transcription Factor Signals. Turn on all available tracks for HeLa-S3 cells by hovering over the cell line name and clicking Select all HeLa-S3 in the pop up window (the boxes for HeLa-S3 will turn blue).

By comparing the signal intensity and annotated peaks for each of the Histone Modifications and Transcription Factors, you can see that the increased signal intensity corresponds to the regions where a peak has been annotated.

(e) Click on Custom tracks and click the ‘Trash’ icon from the Actions section of the ENCODE Analysis Hub.

Ensembl REST API

The webinar is new for the 2018 course and will be presented by Emily Perry on the 13th September 2018.

This final webinar will introduce the Ensembl REST API, which can be used for language agnostic programmatic access to the Ensembl database.
Your feedback

Please tell us what you thought about this course. Your feedback is invaluable and helps us to improve our courses and thus enhance your learning experience.

Get help and support on Ensembl

Questions or any comments about Ensembl or this course? helpdesk [at] ensembl.org (Contact the Ensembl helpdesk).

Tutorials

- Check out our YouTube [48] or YouKu [49] channel.
- Try out an Ensembl tutorial [50]. There are videos and coursebooks on a variety of subjects.
- See if your question has already been answered in the Ensembl FAQs [51].
- Have a look at the Ensembl Glossary [52].
- View technical documentation on Ensembl [53].
- Visit our further courses on Train online: Ensembl: Browsing chordate genomes [54] and Ensembl: Filmed API workshop [55].

Support

- If you cannot find the answer to your question, contact the Ensembl helpdesk [56].
- If you are using Ensembl programmatically, our dev list [57] is a community of Ensembl developers where you can ask and answer questions.
- If you thought this course was useful, and you and your colleagues would like to experience our face-to-face training, consider hosting an Ensembl workshop [58] at your institution.

References

- McLaren W., et al. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP


Contributors

Emily Perry [1]
EMBL-EBI
Ensembl Outreach Project Leader

Emily is the Outreach Project Leader for Ensembl: she is responsible for the team that teaches workshops, creates training materials and help pages, manages social media, answers helpdesk queries and aids development of new tools for the resource. Emily started at EMBL-EBI as an Ensembl Outreach Officer in September 2012 and became the Project Leader in March 2015. Before working at EMBL-EBI, Emily did her PhD in molecular biology at the MRC Human Genetics Unit in Edinburgh, then worked for the University of Edinburgh's SCI-FUN group, touring Scottish secondary schools with an interactive science roadshow.

Erin Haskell [7]
EMBL-EBI
Ensembl Outreach Officer

Erin was a part of the Ensembl [66] Outreach Team between August 2017 and July 2019. As an Outreach officer she responded to user helpdesk queries, taught workshops on how to use Ensembl and related resources, and contributed to website help pages and social media. Before joining EMBL-EBI she completed a PhD in fungal-plant interactions at the University of York, where she also studied to become an Associate fellow of the Higher
Education Academy (AFHEA), qualified in adult education.

Astrid Gall [6]

EMBL-EBI
Ensembl Outreach Officer

Astrid joined Ensembl [66] as an Outreach Officer in August 2017. As a member of the Outreach Team she delivers workshops, creates training materials and help pages, manages social media, answers helpdesk queries and aids development of new Ensembl tools and displays. Before starting at EMBL-EBI, Astrid obtained her PhD in Molecular Virology from the University of Veterinary Medicine Hannover, Germany, then worked in Virology, Genomics and Bioinformatics at the Federal Research Institute for Animal Health, Germany, the Wellcome Sanger Institute and the University of Cambridge, UK. Astrid is a Fellow of Lucy Cavendish College at the University of Cambridge, where she held a Teaching Fellowship providing small-group teaching to undergraduate students, and acted as a Tutor for graduate students.

Ben Moore [2]

EMBL-EBI
Ensembl Outreach Officer - Flicek team: Vertebrate genomics

Ben joined Ensembl as an Outreach Officer in September 2015. As a member of the Outreach team he delivers workshops, answers helpdesk queries, creates training materials and help pages, and manages social media. Before starting at EMBL-EBI, Ben did his PhD in Biological Sciences at the MRC Laboratory of Molecular Biology at the University of Cambridge. Ben has previously hosted the BlueSci science radio show and now organises the annual Cambridge Pint of Science festival.
Helen Sparrow [3]
EMBL-EBI
Ensembl Outreach Officer - Flicek team: Vertebrate genomics

Helen joined Ensembl as an Outreach Officer in September 2015. As a member of the Outreach Team she delivers workshops, answers helpdesk queries, creates training materials and help pages, and manages social media. Before starting at EMBL-EBI, Helen did her PhD in environmental biotechnology at the University of York, then worked as the community manager at academia.edu and as a museum educator at the California Academy of Sciences.

Victoria Newman [5]
EMBL-EBI
Ensembl Outreach Officer

Victoria is an Outreach Officer for Ensembl, delivering workshops, answering help-desk queries, creating training and support materials and engaging with users on social media. Before joining EMBL-EBI in 2016, Victoria obtained her PhD in Biochemistry and Molecular Biology from the University of California, San Francisco, then worked in academic publishing at Macmillan and Frontiers.

Denise Carvalho-Silva [4]
EMBL-EBI
Scientific Outreach Lead - Open Targets

Denise joined EMBL-EBI’s Open Targets team in June 2016, where she delivers workshops, provides user support and creates online training materials for the Open Targets projects. Denise holds a PhD in Biochemistry from the Universidade Federal de Minas Gerais, Belo Horizonte, Brazil, and completed her postdoctoral training in Genetics at the Australian National University, Canberra, Australia. Before coming to Open Targets, she worked in Ensembl at EMBL-EBI, at the Wellcome Trust Sanger Institute as a Research Associate in Human Evolution, and as a Senior Computer Biologist in the HAVANA team on the GENCODE, EUCOMM and Pig genome projects.

The 2018 course was taught by Emily Perry, Erin Haskell and Astrid Gall.

The 2017 course was taught by Emily Perry, Ben Moore, Helen Sparrow and Victoria Newman.

The 2016 course was taught by Emily Perry, Denise Carvalho-Silva, Ben Moore and Helen Sparrow.

Source URL: https://www.ebi.ac.uk/training/online/course/ensembl-browser-webinar-series-2016

Links
[1] https://www.ebi.ac.uk/training/online/trainers/emily
[2] https://www.ebi.ac.uk/training/online/trainers/bmoore
[3] https://www.ebi.ac.uk/training/online/trainers/hsparrow
[4] https://www.ebi.ac.uk/training/online/trainers/denise
[5] https://www.ebi.ac.uk/training/online/trainers/vnewman_9726
[6] https://www.ebi.ac.uk/training/online/trainers/agalli_13476
[7] https://www.ebi.ac.uk/training/online/trainers/ehaskell_13392
[8] https://www.ebi.ac.uk/training/online/course/ensembl-browser-webinar-series-2016/how-take-course
[9] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/introduction-ensembl
[10] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/ensembl-genes
[12] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/comparing-genes-and-genomes-ensembl-compara
[14] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/data-export-biomart
[16] https://www.youtube.com/user/EnsemblHelpdesk/featured?&ab_channel=EnsemblHelpdesk
[17] http://ftp.ebi.ac.uk/pub/databases/ensembl/online_course/1_exploring_emily.mp4
[18] https://www.ebi.ac.uk/training/online/ftp.ebi.ac.uk/pub/databases/ensembl/online_course/1_Exploring_Emily.pdf
[19] https://www.ebi.ac.uk/training/online/ftp.ebi.ac.uk/pub/databases/ensembl/online_course/1_intro_demo.pdf
[26] http://ftp.ebi.ac.uk/pub/databases/ensembl/training/online_course/3_variation_emily.mp4
[27] http://ftp.ebi.ac.uk/pub/databases/ensembl/training/online_course/3_Variation_Emily.pdf
[31] http://ftp.ebi.ac.uk/pub/databases/ensembl/training/online_course/4_compara_emily.mp4
Ensembl browser webinar series
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[34] http://ftp.ebi.ac.uk/pub/databases/ensembl/training/online_course/5_Regulation_Astrid.mp4
[37] http://ftp.ebi.ac.uk/pub/databases/ensembl/training/online_course/6_BioMart_Ben.mp4
[38] http://ftp.ebi.ac.uk/pub/databases/ensembl/training/online_course/6_BioMart_Ben.pdf
[40] https://v.youku.com/v_show/id_XMzgyMzgwMzEzNg==.html?spm=a2h3j.8428770.3416059.1
[41] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u13476/7_UploadData_UniAndes_Astrid.pdf
[42] https://www.ebi.ac.uk/~emily/Workshops/BAM/
[43] http://www.ebi.ac.uk/~emily/Workshops/BAM/GRCh38.20.illumina.merged.1.bam
[44] https://trackhubregistry.org
[45] http://v.youku.com/v_show/id_XMzgyMzQ5NzUzMg==.html?spm=a2h3j.8428770.3416059.1
[46] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u1218/8_REST_API_Emily.pdf
[48] http://www.youtube.com/user/EnsemblHelpdesk
[49] http://i.youku.com/u/id_UMzM1NjkzMTI0
[54] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes
[55] https://www.ebi.ac.uk/training/online/course/ensembl-filmed-api-workshop
[56] http://www.ensembl.org/Help/Contact
[58] http://training.ensembl.org/
[59] http://europepmc.org/articles/PMC4702834
[60] http://europepmc.org/articles/PMC4702859
[61] http://europepmc.org/articles/PMC3170168
[63] http://europepmc.org/articles/PMC2652215
[64] http://europepmc.org/articles/PMC4407537
[65] http://europepmc.org/articles/PMC4271150
[66] https://www.ebi.ac.uk/training/online/glossary/ensembl