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.

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 [8]</td>
<td>Astrid Gall</td>
</tr>
<tr>
<td></td>
<td>Ensembl genes [9]</td>
<td>Emily Perry</td>
</tr>
<tr>
<td>6th September</td>
<td>Variation data in Ensembl and the Ensembl VEP [10]</td>
<td>Erin Haskell</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Astrid Gall</td>
</tr>
<tr>
<td>Date</td>
<td>Webinar topic</td>
<td>Instructor</td>
</tr>
<tr>
<td>----------------</td>
<td>---------------------------------------------------</td>
<td>---------------------</td>
</tr>
<tr>
<td>11th September</td>
<td>Comparing genes and genomes with Ensembl Compara</td>
<td>Emily Perry</td>
</tr>
<tr>
<td></td>
<td>Finding features that regulate genes – the Ensembl Regulatory Build</td>
<td>Erin Haskell</td>
</tr>
<tr>
<td></td>
<td>Data export with BioMart</td>
<td></td>
</tr>
<tr>
<td>13th September</td>
<td>Uploading your data to Ensembl</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 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.

![Help Icon] 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 was held on the 4th September 2018 at 3pm BST, presented by Astrid Gall.

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 watch the video on YouKu.

You can also download the slides.
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 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?

Extra Exercise 8 – Exploring CRISPR sites

You want to do some CRISPR manipulation of the human SMC3 gene. You are looking for a CRISPR 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 sites that overlap these variants, then zoom out to see the whole SMC3 gene. In which exon of the SMC3-201 transcript are the CRISPR sites found?

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.

Click Next>.
Click on Text.

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

```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 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 drop-down list like it is and type `human 21:32630000-32870000` in the text box).

Click Go.

You will see a red highlighted region in the middle of this region. Click on the thin dark red bar in 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 7 – Exploring a genomic region in Anopheles gambiae**

(a) Go to [metazoa.ensembl.org](http://metazoa.ensembl.org).

Select Search: Anopheles gambiae and type `2L:7300000-7450000` in the text box. Click Go.

The region is located on the cytogenetic band 21B.

(b) There are seven genes in this region and two that overlap the ends.

Drag out a box around the second exon of the gene, the second red box from the left, and click on Jump to region. 
You may wish to drag out more boxes to zoom in further.

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 Start/stop codons in the Find a track text box. 
Select Start/stop codons. 
Close the menu by clicking on ? or anywhere outside the window.

Start codons are shown in green. You should see one that coincides with the start of the filled red box of the second exon.

(c) Drag out a box around the green start codon and select Mark region.

Scroll up to the Overview image to drag out a box around the gene and select Jump to region (or use the Zoom out button).
The highlighted region is now visible as a grey dotted line.

**Extra Exercise 8 – Exploring CRISPR sites**

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

---

**Ensembl genes**

This webinar was held on the 4th September 2018 and was 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 [watch the video on YouKu](https://www.youtube.com/watch?v=video_id) [18].

You can also [download the slides](https://www.ensembl.org) [19].
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?
- Which transcript is the best quality?

(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?

Exercise 3 – Exploring a bacterial gene (Clostridium sporogenes)

(a) What GO: biological process terms are associated with the polC CLSPOx_12590 gene?
(b) Go to the transcript tab for the only transcript, PolC-1. How long is the transcript?

(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?

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) [21])

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 best quality transcript, as it has a CCDS associated with it, is TSL:1 and is Golden.

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

(c) The Gene Ontology project ([http://www.geneontology.org/](http://www.geneontology.org/) [22]) maps terms to a protein in three classes: biological process, cellular component, and molecular function. Meiotic spindle organisation, cell morphogenesis, and cytokinesis are some of the roles associated with *MYH9*.

(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/) [23]

Exercise 2 – Finding a gene associated with a phenotype

(a) Start at the Ensembl homepage ([http://www.ensembl.org](http://www.ensembl.org) [21]).

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.

Exercise 3 – Exploring a bacterial gene (*Clostridium sporogenes*)

(a) Go to http://bacteria.ensembl.org/index.html [20]

Select *Clostridium sporogenes* by beginning to write the species name, and selecting the species option.

Type PolC and click on the gene name link PolC [CLSPOx_12590].

Click on GO: biological process in the side menu.

There is one term listed: GO:0006260, DNA replication.

(b) Click on the transcript named PolC-1 (or on the Transcript tab).

PolC-1 is 4299 bp in length.

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

A Ribonuclease H-like domain is identified by two domain prediction methods. A DNA polymerase, alpha subunit is identified by three. An exonuclease domain is identified by two, a nucleic acid-binding domain is identified by two and a DNA Polymerase III epsilon subunit is identified by one.

Variation data in Ensembl and the Ensembl VEP

This webinar was held on the 6th September 2018 at 15:00 BST and was presented by Erin Haskell.

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.

The demonstrations in this video were carried out using Ensembl version 93.

If you have trouble accessing YouTube you can watch the video on YouKu [24].
Variation exercises

Exercise 1 — Human population genetics and phenotype data
The SNP rs11725853 in an intron of the human \textit{GLRA3} gene has been identified as a genetic risk factor for kidney disease in individuals with type 1 diabetes.

(a) How many alternate alleles does this SNP have? Which allele is the risk (or associated) allele for urinary albumin excretion rate in type 1 diabetes?

(b) Which super-population (AFR, AMR, EAS, EUR or SAS) in the 1000 Genomes Project has the highest frequency of the risk allele?

(c) In which paper(s) is the association between rs11725853 and urinary albumin excretion rate described?

(d) What is the ancestral allele? Is it conserved in the 70* eutherian mammals?

*This number is correct for Ensembl release 93 (http://jul2018.archive.ensembl.org/). In later releases this group may contain more than 70 species.

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?

(c) Why are the alleles for this variation in Ensembl given as G/A and not as C/T, as in dbSNP and literature? (http://www.ncbi.nlm.nih.gov/projects/SNP/snp_ref.cgi?rs=1801133)

(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 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?

\textbf{Exercise 4 — The VEP}

Resequencing of the genomic region of the human \textit{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?

\textbf{Exercise solutions}

\textbf{Exploring variants in Ensembl - solutions}

\textbf{Exercise 1 – Human population genetics and phenotype data}

First, find the variant tab for the variant rs11725853. Search Ensembl for \textit{rs11725853} directly from the homepage.

(a) Find the ‘Alleles’ header on the variant summary page shows.

\textbf{G/A/C} is shown as the possible nucleotides at this position, so there are two alternate alleles (A and C), and one reference allele (G).

Click on \textit{Phenotype data} in the left-hand navigation panel.
The first table in this page has a column called ‘Associated allele’, we can see that A is listed as the risk allele for Urinary albumin excretion rate in type 1 diabetes.

(b) Click on Population genetics in the left-hand navigation panel.

The 1000 Genomes Project Phase 3 is the first resource listed. The variant frequencies are summarised by super-population in the pie-charts, and also in the table below. The East Asian (EAS) population has the highest frequency of the risk allele A, at 35%. The table shows us that the most common genotype in this population is A|G, with ~45% of the population having this genotype.

(c) Click on Citations in the left-hand navigation panel.

This variant is described in the paper PMID:24595857, ‘Genome-wide association study of urinary albumin excretion rate in patients with type 1 diabetes’. Note that you can also find out the PubMed ID from the Phenotype Data page, this may be easier to find relevant papers if the variant has many attributed phenotypes.

(d) The ancestral allele is G, as shown next to the allele information in the summary at the top of the page. To find out more information, click on Phylogenetic context in the left-hand navigation panel.

Click on the Select another alignment button in the blue box and select the 70 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 G allele is conserved in the majority of this group.

Exercise 2 – Exploring a SNP in human

(a) Go to the Ensembl homepage (http://www.ensembl.org/ [26]).

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 Medellin; 1000 Genomes), HCB (Han Chinese in Beijing, China; HapMap).

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

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

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

Select Alignment: 8 primates EPO and click Go.
Gorilla, vervet, chimp, macaque, olive baboon and marmoset all have a G in this position. Please note that there is no variation database for gorilla, olive baboon, vervet or marmoset though.

Exercise 3 – Exploring a SNP in mouse

(a) Go to www.ensembl.org, type \textbf{rs29522348} in the search box. Click on \textbf{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 \textbf{Show} to reveal information about HGVS nomenclature.

This SNP has three \textbf{HGVS names}, one at the genomic DNA level (NC_000083.6:g.73924993C>T), one at the transcript level (ENSMUST00000024866.4:c.721G>A) and one at the protein level (ENSMUSP00000024866.4:p.Val241Ile).

(c) In Ensembl, the allele that is present in the reference genome assembly is always put first (C is the allele for the reference mouse genome, strain C57BL/6J). This is referred to as the ‘Reference’ or ‘Major’ allele.

(d) Click on \textbf{Sample genotypes} is the left hand side menu. In the summary of genotypes by population, click on \textbf{Show} for \textbf{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 – VEP

Go to \textbf{www.ensembl.org} \cite{21} and click on the link tools at the top of the page. Click on \textbf{Variant Effect Predictor} and enter the three variants as below:

\begin{verbatim}
7 117530985 117530985  G/A \\
7 117531038 117531038  T/C \\
7 117531068 117531068  T/C
\end{verbatim}

Note: Variation data input can be done in a variety of formats. See more details here \textbf{http://www.ensembl.org/info/docs/variation/vep/vep_formats.html} \cite{27}.

Click \textbf{Run}.

When your job is listed as \textbf{Done}, click \textbf{View Results}.

(a) 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.

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}
(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.

Comparing genes and genomes with Ensembl Compara

This webinar was held on the 6th September 2018 and was 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 watch the video on YouKu [28].

You can also download the slides [29].

Compara exercises

Exercise 1 – Orthologues, paralogs 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 View Sequence Alignments link next to the Ensembl identifier to view a protein alignment in Clustal format.

(b) Go to the orthologue in marmoset. Is there a genomic alignment between marmoset and human? Is there a gene for both species in this region?

Exercise 2 – Zebrafish orthologues

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

(a) Go to the Location tab for this gene. View the Alignments (image) and Alignments (text) for the 11 fish. Which fish genomes are represented in the alignment? Do all the fish 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 11 fish EPO_LOW_COVERAGE by configuring the page. What is the difference between the 11 fish EPO_LOW_COVERAGE track and the 11 way GERP elements and scores tracks? 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 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 32 amniota vertebrates Mercator-Pecan track. 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 70 mammals and the set of 32 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.

Extra Exercise 5 – Pan-taxonomic Compara

Find the NMA2179 gene in the genome of the bacterial strain Neisseria meningitidis Z2491, (TaxID 122587).

(a) What is the function of this gene?

(b) How many orthologues are predicted for this gene in bacteria? What is the maximum identity of a bacterial protein with the one of Neisseria meningitidis Z2491?

(c) How many vertebrate species have predicted orthologues for NMA2179? How many orthologues are predicted in human and what is their type of relationship with NMA2179?

(d) Export the Gene Tree in Newick format.

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) There is more than one way to get to the answer.
Option 1: Go to the Orthologues page and click on the marmoset orthologue (ENSCJAG00000013536) to open the Gene tab.
Click Genomic alignments at the left, then the Select an alignment button.
Select All Alignments -> Pairwise -> Primates -> Human and click Apply. Choose Block 1 to get the largest block of aligned sequence.
Click on Display full alignment at the bottom of the page.
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 Codons: 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 the Select species or regions button and select All Divisions -> Primates -> Human, then click Apply. 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, tick the box Join genes and close the menu. Zoom out on the location view to see blue lines connecting all homologous genes between marmoset and human genes in that region).

Exercise 2 – Zebrafish orthologues
(a) Start in the Location tab (region in detail) for sardh. Click on Alignments (Image) at the left, and select the 11 fish EPO_LOW_COVERAGE alignment in the view and click Apply.

The zebrafish, cave fish, cod, stickleback, medaka, amazon molly, platyfish, tilapia, fugu, tetradon and spotted gar are shown in this region. All species except the medaka 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 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 11 fish EPO_LOW_COVERAGE, all under Comparative genomics.

The 11 fish EPO_LOW_COVERAGE just shows that the region for the sardh gene can be aligned for these eleven species of fish. The 11 way GERP elements and scores 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 11 way GERP elements, 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 **Alignments (text)** under the **Comparative Genomics** menu. Select **All alignments -> Multiple -> 32 amniota vertebrates Mercator-Pecan** and click **Apply**.

Yes, the degree of conservation does reflect the evolutionary relationship between human and the other species. E.g. the highest degree of conservation is found within the primates cluster, while the degree of conservation between human and birds is low.

Especially the exonic sequences of BRCA2 seem to be highly conserved between the various species. This is to be expected because these are likely 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 70 eutherian mammals EPO-Low-Coverage, Conservation score and Constrained elements for 32 amniota vertebrates Mercator-Pecan**.
Close the menu.

Both the 32 way GERP elements and scores track 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.

**Extra Exercise 5 – Pan-taxonomic Compara**

(a) Go to **bacteria.ensembl.org** and start typing **Neisseria meningitidis Z2491** in the search box to find the genome. Select **Neisseria meningitidis Z2491, (TaxID 122587)** from the drop-down menu. Then type **NMA2179** in the search box on the homepage of Neisseria meningitidis Z2491 to find the gene. Click on the Gene ID **NMA2179** in the search results to open the Gene tab.

This gene encodes for the dihydrofolate reductase. You can find this information in the description of the gene at the top of the page and under **GO: Molecular function** in the menu on the left.

(b) Click on **Orthologues** under **Pan-taxonomic Compara** in the side menu to see all orthologues.

There are 68 orthologues in bacteria. Tick the box in the table at the top to show details for bacteria only.

Now click on the header **Target %id** of the table at the bottom to sort the table. The **Target %id** is defined as the percentage of the orthologous sequence matching the Neisseria meningitidis Z2491 sequence.

The maximum identity is 48.47% and found between the protein of Aeromonas hydrophila subsp. hydrophila ATCC 7966 and Neisseria meningitidis Z2491.

Likewise, **Query %id** is defined as the percentage of the Neisseria meningitidis Z2491 sequence matching the...
sequence of the orthologue. The maximum identity is 50.00% for Azotobacter vinelandii DJ.

(c) Thirteen vertebrate species have predicted orthologues for NMA2179. Tick the box in the table at the top to show details for vertebrates only.

Now type human in the box above the table at the bottom to find the orthologues in human.

There are two orthologues predicted in human, both with a 1-to-many relationship with NMA2179.

You can click on ENSG00000228716 next to DHFR to open the Gene tab for this human gene. Note that DHFR indeed encodes for the human dihydrofolate reductase.

(d) Click on Gene Tree in the side menu to view the image of the tree.

Click on the rightmost button above the image (it will display Download data from this image when you mouse over it) to open a pop-up window.

Select Newick format and click the Download button.

Finding features that regulate genes – the Ensembl Regulatory Build

This webinar was held on the 11th September 2018 at 15:00 BST and was presented by Emily Perry.

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 watch the video on YouKu [30].

You can also download the slides [31].

Regulation exercises

Finding features that regulate genes – the Ensembl Regulatory Build

Exercise 1 – Gene regulation: Human

(a) Find the Location tab (Region in detail page) for the STX7 gene. Are there any predicted promoters in this gene region?

(b) Click Configure this page and on the Regulatory features menu in the left hand side. Turn on Regulatory features for HUVEC, HeLa-S3 and HepG2 cell types. Is the predicted promoter active in any of these cell lines?

(c) Use Configure this page to add supporting data indicating open chromatin for HeLa-S3, HUVEC and HepG2
cells. Are there sites enriched for marks of open chromatin (DNase1) in these cells at the 5’ end of STX7?

(d) Configure this page once again to add histone modification supporting data for the same cell type as above. Which ones are present at the 5’ end of STX7?

(e) Do any data support methylated CpG sites in this region (5’ end) of STX7 in Jurkat cells?

Exercise 2 – Regulatory features in human

(a) Go to the Location tab (Region in detail page) for human APOE and zoom out a little to see the flanking region. Is there a regulatory feature annotated at the 5’ end of the gene? What kind of feature is it and what is its stable ID? Does it contain any transcription factor binding motifs?

(b) In which cell types is this feature active? And poised?

(c) Are there any transcription factor binding motifs in this feature? For what transcription factor?

(d) Can you observe binding of this transcription factor in any cell types? What other transcription factors are also found at this location in this/these cell type(s)?

Exercise solutions

Finding features that regulate genes – the Ensembl Regulatory Build

Exercise 1 – Gene regulation: Human STX7

(a) Search for human STX7 from the home page. Click on 6:132445867-132513198:-1 below STX7 (Human Gene).

Regulatory features from the Ensembl 'regulatory build' are based on indicators of open chromatin such as CTCF binding sites, DNase I hypersensitive sites, and Transcription Factor binding sites. The Regulatory features are turned on by default in the Region in detail view.

There is one promoter predicted in the region of STX7.

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

(b) See the legend below the Region in detail view to find the predicted promoter is coloured red or purple in the different cell lines chosen. The purple colouring of the predicted promoter in the HUVEC cell type (out of the three cells chosen) indicates that this promoter is ‘poised’. The red colouring of the predicted promoter in the HeLa-S3 and HepG2 cell lines indicates that these promoters are ‘active’.

(c) Configure this page and click on Open chromatin & TFBS. Turn on both peaks and signal for DNase 1 in HeLa-S3, HUVEC and HepG2 cells (the boxes in this configure this page window will turn blue. For more information on how to select and view the supporting data, click on Show tutorial in the pop up window). Close the menu.

There’s a DNase1 hypersensitive site in the first exon of STX7 that is detected in HUVEC and HepG2 cells; a second DNase1-hypersensitive site in exon 2 is detected in HUVEC cells. Click on the coloured block to find out that the DNase1 enriched sites come from the ENCODE project.
(d) Configure this page and click on Histones & polymerases. Change the Filter by menu from All classes to Histone. Select all the histone modifications available for HeLa-S3, HepG2 and HUVEC cells (some of them might be on by default). Save and close the menu.

H2AZ, H3K27me3, H3K4me1, H3K4me2, H3K4me3, H3K79me2, H3K9ac and H4K20me1 sites have been found in the 5’ region of STX7 in HepG2 cells.

(e) Click on Configure this page and choose the DNA Methylation menu. Turn on the track Jurkat RRBS ENCODE. Save and close the menu.

Some CpG sites at the 5’ end of STX7 are not highly methylated (note the yellow bars) whilst others are (blue bars). See the Methylation Legend, at bottom, for more details. Additional information on human DNA methylation tracks can be found at

www.ensembl.org/info/docs/funcgen/index.html [32]

Exercise 2 – Regulatory features in human

(a) Search for human APOE from the home page. Click on Location in the search results and zoom out to view the genomic region surrounding the APOE gene. The gene is positive stranded so look for features at the left hand side. Click on the features to get their IDs.

There is a pink promoter flank at the 5’ end of APOE. Click on it to get a pop-up with its ID: ENSR00000110117.

(b) Click on the stable ID ENSR00000110117. Click on the ID to go to the regulation tab.

ENSR00000110117 is active in 18/68 cell types studied and poised in 32.

(c) Click on the promoter to see a list of its TF motifs.

There are SP1 motifs in this promoter.

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

SP1 binding is only observed in H1ESC cells.

Open Select evidence again and choose all the transcription factors, then close. You may find it easier to see if you also go into Select cells and turn the cell types ALL OFF, then turn on H1ESC only.

TAF7 and USF1 also bind to this locus.

Data export with BioMart

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 watch the video on YouKu [33].
You can also download the slides [34].

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:

(a) 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?

NP_001218, NP_203125, NP_203124, NP_203126, NP_001007233, NP_150636, 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
**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 are below.

(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

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, 26652_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

**Exercise solutions**

**Using BioMart to Export Ensembl Data**

**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 Cleavage site (Signalp) from the drop-down and then Only. Clicking on Count should reveal that you have filtered the dataset down to 221 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 GENES 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**

Click Unique Results only and expand the preview table to All

---

**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 GENES 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 out of 6,007,985 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.

**Step 1: Dataset** Choose Ensembl Variation and Human Short Variation (SNPs and indels excluding flagged variants)

**Step 2: Filters**

Expand the GENERAL VARIANT FILTERS, choose Filtter by Variation name and enter: rs566014072, rs754099015

**Step 3: Attributes**

Expand the VARIANT ASSOCIATED INFORMATION, choose Variant name, Variant alleles, scroll down to the Phenotype annotation section and choose Phenotype description and Associated gene with phenotype. Uncheck the information about Chromosomes and start/end positions.

**Step 4: Results**

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.
Ensembl browser webinar series
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

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 probes/probesets section and choose AFFY HG U133 Plus 2 probe.

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.
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. [35]

You can also download the slides [36].

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/ [37]

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)[38]) 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](https://trackhubregistry.org)[39]). 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, Inl1, 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.

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

You can also download the slides [41].

You can also access the online Jupyter notebooks [42] in Python, R and Perl to learn more about the REST API.
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 [43] or YouKu [44] channel.
- Try out an Ensembl tutorial [45]. There are videos and coursebooks on a variety of subjects.
- See if your question has already been answered in the Ensembl FAQs [46].
- Have a look at the Ensembl Glossary [47].
- View technical documentation on Ensembl [48].
- Visit our further courses on Train online: Ensembl: Browsing chordate genomes [49] and Ensembl: Filmed API workshop [50].

Support

- If you cannot find the answer to your question, contact the Ensembl helpdesk [51].
- If you are using Ensembl programmatically, our dev list [52] 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 [53] at your institution.

References

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 joined the Ensembl [61] Outreach Team in August 2017. As an Outreach officer she responds to user helpdesk queries, teaches workshops on how to use Ensembl and related resources, and contributes 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 [61] 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/agall_13476
[7] https://www.ebi.ac.uk/training/online/trainers/ehaskell_13392
[8] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/introduction-ensembl
[9] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/ensembl-genomes
[12] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/finding-features-regulate-genes-%E2%80%93-ensembl-regulato
[13] https://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/data-export-biomart
[15] https://www.youtube.com/user/EnsemblHelpdesk/featured?&ab_channel=EnsemblHelpdesk
[16] https://v.youku.com/v_show/id_XMzgxMjA2MTQ2MA==.html?spm=a2hdp.8244740.0.0
[17] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u1218/1_Introduction_UniAndes_Astrid.pdf
[18] https://v.youku.com/v_show/id_XMzgxMjEyNjE1Ng==.html?spm=a2hdp.8244740.0.0
[19] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u1218/2_GenesandTranscripts_Emily.pdf
[23] http://www.ebi.ac.uk/arrayexpress/
[24] https://v.youku.com/v_show/id_XMzgxNDk3Njc5Ng==.html?spm=a2hdp.8244740.0.0
[25] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u13392/3_EnsemblVariation.pdf
[28] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u1218/4_compara_ensemblwebinar.pdf
[30] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u13392/5_Regulation_Emily.pdf
[31] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u13392/6_BioMart_Erin.pdf
[32] https://www.ebi.ac.uk/~emily/Workshops/BAM/
[33] https://www.ebi.ac.uk/~emily/Workshops/BAM/GRCh38.20.illumina.merged.1.bam
[34] https://www.trackhubregistry.org
[36] https://www.ebi.ac.uk/training/online/sites/ebi.ac.uk.training.online/files/u13476/7UploadData_UniAndes_Astrid.pdf
[37] https://v.youku.com/v_show/id_XMzgyMDk1MjE2NA==.html?spm=a2hdp.8428770.3416059.1
[38] https://notebooks.azure.com/ensembl-training
Ensembl browser webinar series
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[43] http://www.youtube.com/user/EnsemblHelpdesk
[44] http://i.youku.com/u/id_UMzM1NjkzMTI0
[49] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes
[50] https://www.ebi.ac.uk/training/online/course/ensembl-filmed-api-workshop
[54] http://europepmc.org/articles/PMC4702834
[55] http://europepmc.org/articles/PMC4702859
[56] http://europepmc.org/articles/PMC3170168
[57] http://europepmc.org/articles/PMC2916720
[58] http://europepmc.org/articles/PMC2652215
[59] http://europepmc.org/articles/PMC4407537
[60] http://europepmc.org/articles/PMC4271150
[61] https://www.ebi.ac.uk/training/online/glossary/ensembl