Ensembl browser webinar series

Between the 6th April and 18th May 2017, we will be holding a series of live Ensembl webinars which come together to provide a complete browser course.

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 1 hour chunks.

Videos of all the webinars will be available here, along with exercises so that you can practice what you have learned, and there will be opportunities for you to discuss these exercises online.

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

A previous course was held in 2016 and the lectures and exercises will remain available here until they have been replaced.

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 every Thursday between 6th April - 18th May 2017

This course will consist of a series of 1 hour live webinars, held between 6th April and 18th May 2017. Each webinar will take place on Thursdays at 09:00 BST (GMT+1).
How do I sign up for the live webinars?

To attend the live webinars, you need to sign up to our webinar series here [2] (or here if you're in China [3]).

Once you have registered, you will get a weekly email reminder of when the webinars are taking place, and a link to the live session. We will use GoToTraining to run the webinars, which requires you to install a small piece of software, but this is free and you will receive instructions via email. We recommend that you install the software in advance of the first webinar as it may take a couple of minutes to set up. If you wish to view these webinars on a Linux machine you will be able to access a browser based version of the webinar (GoToTraining should automatically recognise that you are using a Linux based system).

The videos of the webinars will be posted in this course and are open to everyone. You do not need to be signed up in order to view them, however if you sign up you can get reminders when the new videos are available.

The programme will be:

<table>
<thead>
<tr>
<th>Date</th>
<th>Webinar topic</th>
<th>Instructor</th>
</tr>
</thead>
<tbody>
<tr>
<td>6th April</td>
<td>Introduction to Ensembl [4]</td>
<td>Helen Sparrow</td>
</tr>
<tr>
<td>13th April</td>
<td>Ensembl genes [5]</td>
<td>Emily Perry</td>
</tr>
<tr>
<td>4th May</td>
<td>Comparing genes and genomes with Ensembl Compara [8]</td>
<td>Ben Moore</td>
</tr>
<tr>
<td>11th May</td>
<td>Finding features that regulate genes – the Ensembl Regulatory Build [9]</td>
<td>Ben Moore</td>
</tr>
<tr>
<td>18th May</td>
<td>Uploading your data to Ensembl and advanced ways to access Ensembl data [10]</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 [11] and embed the video in the corresponding page, so if you cannot make a particular webinar, you can always catch up later.

Weekly exercises to test what you have learnt

After each webinar, we will post new 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 over the week.

Getting help with the exercises

We have set up a Facebook group [12] where you can discuss these exercises. We will be monitoring the Facebook group to offer help, but you may also find it helpful to talk to your fellow course attendees. If you prefer not to use Facebook, you can always helpdesk [at] ensembl.org (email us) your questions.
Unfortunately, as we have no way of determining who has completed all parts of the course, we are unable to provide certificates.

**Use the old course**

The current version of the course was created March-May 2016. The new webinar course will include updates to the old course, however most of the old course is still valid. We will keep the old course up here until the new one starts.

**Introduction to Ensembl**

This webinar was held on the 6th April 2017 at 9am BST, presented by Helen Sparrow.

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 [13].

You can also download the slides [14].

**Exploring Ensembl exercises**

**Finding out about Ensembl 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 genes have been annotated?

**Exercise 2 — Mouse**

(a) What's new in release 88 for mouse?

(b) What previous assemblies are available for mouse?
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 view

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) Export the genomic sequence of the region you are looking at in FASTA format.

(e) 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 there 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 solutions
Finding out about Ensembl species - solutions

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 allMel1 or GCA 000004335.1

(b) Click on More information and statistics. Statistics are shown in the tables on the left.

The length of the genome is 2,245,312,831 bp.

There are 19,343 coding genes.

Exercise 2 — Mouse
(a) Click on Mouse on the front page of Ensembl to go to the species homepage. News is in the top right.

What's new in Mouse release 88:

- Update to Ensembl-Havana GENCODE gene set
- External database reference update
- Updated cDNA alignments

(b) Under Other reference assemblies there is one previous assembly and the release you can find it in is listed.
NCBIM37 is available in release 67.

Exercise 3 — Mosquitos
(a) Go to metazoa.ensembl.org. Open the drop down list or click on View full list of all Ensembl Metazoa species. Type Anopheles into the filter box in the top left.

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 search species 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.
Region in Detail view

Exercise 5 — Exploring a genomic region in human

(a) Go to the Ensembl homepage (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 bars in the Contigs track. Note that KF455761.1 is a tiny contig that splits AL137143.8 in two.

(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 the tick (or anywhere outside the pop-up window).

There is not just one clone that contains the complete BRCA2 gene. The BAC clone RP11-37E23 contains most of the gene, but not its very 3’ end (contained in RP11-298P3). This was reflected on the two contigs that make up the entire BRCA2 gene (the Contigs track is on by default). You may find this easier to see if you highlight the 3’ exon on BRCA2.

(d) Click Export data in the side menu. Leave the default parameters as they are.

Click Next.

Click on Text.

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

13 dna:chromosome chromosome:GRCh38:13:32311910:32405865:1

(e) Click Configure this page in the side menu.

Click Reset configuration.

Click on the tick.
Exercise 6 — Exploring assembly exceptions in human

(a) Go to the Ensembl homepage (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 any of the three views to see the label CHR_HSCHR21_3_CTG1_1:32769079-32843731. Click on What are assembly exceptions? to open a new window which explains assembly exceptions.

(b) Assembly exceptions are marked in the chromosome view at the top.

There are seven haplotypes on chromosome 21.

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

Scroll down the page to see the comparison between the haplotype 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.

Ensembl genes

This webinar was held on the 13th April 2017 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 [15].

You can also download the slides [16].

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 MYH9-001, 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 morbid identifier for this gene?

Exercise 3 – Exploring a bacterial gene (Clostridium sporogenes)

Start in http://bacteria.ensembl.org/index.html [17] and select the Clostridium sporogenes (GCA_001020205) genome.

(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 [18])

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 MYH9-001 and it codes for a protein of 1,960 amino acids
- MYH9-001 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/ [19]) 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/ [20]

Exercise 2 - Finding a gene associated with a phenotype

(a) Start at the Ensembl homepage (http://www.ensembl.org [18]).

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 88.

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 morbid ID is 261600.

Exercise 3 - Exploring a bacterial gene (Clostridium sporogenes)

(a) Go to http://bacteria.ensembl.org/index.html [17]

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.

Data export with BioMart

This webinar was held on the 20th April 2017 at 9am BST and was presented by Victoria Newman.

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 [21].

You can also download the slides [22].
BioMart exercises

Using BioMart to Export Data from Ensembl

Exercise 1 — Finding Genes by Protein Domain
Download the sequences of all mouse proteins with transmembrane domains located on chromosome 9 between bp 100,000 and 10,000,000.

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 structural 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 rs1801500 and rs1801368. 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, 226652_at, 221641_s_at, 202055_at, 226743_at, 228393_s_at, 225120_at, 218515_at, 202224_at, 223461_at, 225135_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: Transmembrane proteins on chromosome 9

Attributes: Ensembl gene and transcript IDs and Associated gene names

Go to the Ensembl homepage (http://www.ensembl.org) and click on BioMart at the top of the page.

Select Ensembl genes as your database and Mus musculus genes as the dataset.

Click on Filters on the left of the screen and expand REGION. Change the chromosome to 9, and Base Pair put in 100000 as start and 1000000 as end (no commas or spaces).

Now expand PROTEIN DOMAINS, also under filters, and select Limit to genes, choosing with Transmembrane helices from the drop-down and then Only. Clicking on Count should reveal that you have filtered the dataset down to 15 genes (of 51158 genes).

Click on Attributes and select Sequences. Expand Sequences and select Protein. Now click on Results. The first 10 results are displayed by default; to download your results, use the top panel to choose the file type you want by clicking file and then click GO.

The output will be a file containing table the Ensembl gene ID, Ensembl Transcript ID and Associated
gene names of all proteins with a transmembrane domain on mouse chromosome 9-100000:10000000.

Exercise 2 — Export Homologues

Click New.

Dataset: Choose the Ensembl Genes database and then the Ciona savignyi genes (CSAV2.0) dataset.

Filters: Expand the GENE section by clicking on the + box and enter the gene list in the Input external references ID list box.

Attributes: Select the Homologues attributes page, also expand the Orthologues section by clicking on the + box to select Human Ensembl Gene ID.

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

Exercise 3 — Convert IDs

Click New.

Dataset: Choose the ENSEMBL genes database and then the Homo sapiens genes (GRCh38.p10) dataset.

Filters: Expand the GENE section by clicking on the + box, select Input external references ID list - RefSeq protein ID(s) and enter the list of IDs in the text box (either comma separated or as a list).

HINT: You may have to scroll down the menu to see these.

Count: Shows 11 genes (remember one gene may have multiple splice variants coding for different proteins, that is the reason why these 29 proteins do not correspond to 29 genes).

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.

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

Exercise 4 — Export Structural Variants

(a) Dataset: Choose Ensembl Variation and Homo sapiens Structural Variation (GRCh38.p10).

Filters: Expand Region and select Chromosome 1, Base pair start: 130408, Base pair end: 210597. Also expand General Structural Variant features and click on Limit to Variants from source: DGVa

Count: Shows 85 out of 5,892,964 structural variants.

Attributes: Expand Structural Variation (SV) Information and click DGVa Study Accession and Source Name. Next, expand Structural Variant (SV) Location and choose Chromosome name, also expand Supporting Structural Variant (SSV) Location and select Sequence region start (bp) and Sequence region end (bp).
Results: Click Unique Results only and expand the preview table to All

(b) Dataset: Choose Ensembl Variation and Homo sapiens Short Variation (SNPs and indels) (GRCh38.p10).

Filters: Filter by Variation name enter: rs1801500, rs1801368

Attributes: Variant Name, Variant Alleles, Phenotype description and Associated gene with phenotype.

Click the Results button on the toolbar.

You can view this same information in the Ensembl browser. Click on one of the variation IDs (names) in the result table. The variation tab should open in the Ensembl browser. Click Phenotype Data.

Exercise 5 — Find Genes Associated with Array Probes

(a) Click New.

Dataset: Choose the ENSEMBL Genes database, then the Homo sapiens genes (GRCh38.p10) dataset.

Filters: Expand the GENE section by clicking on the + box. Then select Input microarray probes/probesets ID list - Affy hg u133 plus 2 probeset ID(s) and enter the list of probeset IDs in the text box (either comma separated or as a list).

Count: Shows 24 genes match this list of probesets.

Attributes: Select the Features attributes page. Next, expand the GENE section by clicking on the + box and in addition to the default selected attributes, select Description. Now, expand the External section by clicking on the + box. Select HGNC symbol from the External References section and AFFY HG U133-PLUS-2 from the Microarray probes/probesets section.

Click the Results button on the toolbar. Select View All rows as HTML or export all results to a file. Tick the box Unique results only.

Your results should show that the 25 probes map to 24 Ensembl genes.

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

Attributes: Select the Sequences attributes page. Expand the Sequences section by clicking on the + box. Select Flank (Transcript) and enter 2000 in the Upstream flank text box. Expand the Header information section by clicking on the + box. Select, in addition to the default selected attributes, Description and Associated Gene Name.

Note: Flank (Transcript) will give the flanks for all transcripts of a gene with multiple transcripts. Flank (Gene) will give the flanks for one possible transcript in a gene (the most 5’ coordinates for upstream flanking).

Click the Results button on the toolbar.
(c) You can leave the Dataset and Filters the same, and go directly to the Attributes section:

**Attributes**: Select the Homologues attributes page. Expand the Gene section by clicking on the + box, select Associated Gene Name and deselect Ensembl Transcript ID. Expand the Orthologues section by clicking on the + box. Select Mouse Ensembl Gene ID, Mouse Chromosome Name, Mouse Chr Start (bp) and Mouse Chr End (bp).

**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.

---

**Variation data in Ensembl and the Ensembl VEP**

This webinar was held on the 27th April 2017 at 9am BST and was presented by Victoria Newman.

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 watch the video on YouKu [23].

You can also download the slides [24].

---

**Variation exercises**

**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 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 HapMap set?

(c) What is the ancestral allele? Is it conserved in the 40 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?

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

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
Use the VEP tool in Ensembl and choose the options to see SIFT and PolyPhen predictions. Do these variants result in a change in the proteins encoded by any of the Ensembl genes? Which gene? Have the variants already been found?

Exercise solutions

Exploring variants in Ensembl - solutions

Exercise 1 - Human population genetics and phenotype data

(a) Please note there is more than one way to get this answer. Either go to the Variation Table for the human TAGAP gene, and Filter variants to the 5'UTR, or search Ensembl for rs1738074 directly.

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

This SNP is found in three transcripts (ENST00000326965, ENST00000338313, and ENST00000367066).

(b) Click on Population genetics at the left of the variation tab. (Or, click on Explore this variation at the left and click the Population genetics icon.)

In Yoruba (HapMap-YRI population), the least frequent genotype is CC at the frequency of 9.7%. This is also the least frequent genotype in other populations (to find out what the three letter population are, hover over the names).

(c) Click on Phylogenetic context.

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

Select the 40 eutherian mammals EPO LOW COVERAGE alignment and click on Go.

A region containing the SNP (highlighted in red and placed in the centre) and its flanking sequence are displayed. The T allele is conserved in all but four of the 40 eutherian mammals displayed. Note that two species have no alignment in that region and many other species have no variation database.

(d) Click Phenotype Data at the left of the Variation page.

This variant 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) [25]).

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 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: 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  rs29522348  in the search box. Click on  rs29522348 (Mouse Variation).

SNP rs29522348 is located on 17:73924993. In Ensembl, its alleles are provided as in the forward strand.

(b) Click on  HGVS names  to reveal information about HGVS nomenclature.

This SNP has three  HGVS names, one at the genomic DNA level (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).

(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 - 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](http://www.ensembl.org/info/docs/variation/vep/vepFormats.html) 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.

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. You can also upload the VEP results as a track and view them on Location pages in Ensembl. SIFT and PolyPhen are available for missense SNPs only. For two of the entered positions, the variations have been predicted to be probably damaging/deleterious (coordinate 117531038) and benign/tolerated (coordinate 117531068). All the three variations have been already described and are known as in rs1800078, rs1800077 and rs35516286 in dbSNP and other sources (databases, literature, etc).

Comparing genes and genomes with Ensembl Compara

This webinar was held on the 4th May 2017 at 9am BST and was presented by Ben Moore.

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](http://www.ensembl.org/info/docs/variation/vep/vepFormats.html).

You can also [download the slides](http://www.ensembl.org/info/docs/variation/vep/vepFormats.html).
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 Tarsius syrichta protein have to the human one? Click on the Alignment link next to the Ensembl identifier column 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 [18] to find the sardh gene on the zebrafish genome.

(a) Go to the Location page 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 the 11 fish EPO by configuring the page.

What is the difference between the 11 fish 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 [18]

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 Ensembl 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 39 mammals and the set of 23 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.

Bonus Exercise 5 - Exporting paralogues with BioMart

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

Exercise solutions

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

(a) Go to www.ensembl.org [18], choose human and search for BRAF. Click through to the Gene tab view.

On the gene tab, click on Orthologues at the left side of the page to see all the orthologous genes.

There are orthologues in 10 of the 11 primates represented in Ensembl. Only Vervet-AGM has no predicted BRAF orthologue.

The percentage of identical amino acids in the Tarsier protein (the orthologue) compared with the gene of interest. i.e. human BRAF (the target species/gene) is 69%. This is known as the Target %ID. The identity of the gene of interest (human BRAF) when compared with the orthologue (Tarsier BRAF, the query species/gene) is 62% (the query %ID).

Note the difference in the values of the Target and Query %ID reflects the different protein lengths for the human and tarsier BRAF genes.

(b) There is more than one way to get to the answer.
Option 1: Go to the Orthologues page and click on the marmoset orthologue to open the gene tab.

Click Genomic alignments at the left. Then select Select an Alignment: Human (Homo_sapiens) and click Apply. Choose Block 1 to get the largest block of aligned sequence, then click Display full alignment.

The red sequence is present in exons, so there is a gene in both species in this region. You can find where the start and stop codons are located if you configure this page and select START/STOP codons.

Option 2: Go to location tab of the marmoset BRAF gene and then click on Region Comparison view at the left. Click on Select species or regions at the left and select Human (Homo_sapiens) 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 select join genes. Zoom out on the location view to see blue lines connecting all the homologous genes between marmoset and human genes in that region).

Exercise 2 – Zebrafish orthologues

(a) Start in the Location tab (region in detail) for sardh. Click on Alignments (Image) at the left, and click Select an alignment to select the Multiple: 11 fish EPO_LOW_COVERAGE alignment in the pop-up view.

The zebrafish, cave fish, cod, tilapia, Amazon molly, platyfish, spotted gar, stickleback, medaka, fugu, and tetradon are shown in this region. All the species show a gene in the aligned region. This can also be seen in the Alignments (text) page (the exons are highlighted in red).

(b) You can export the alignments from either the Alignments (text) or Alignments (image) pages in the Location tab. Click on the blue Download alignments button at the top of the text page, or the icon at the top of the image, and choose ClustalW from the list.

(c) Click on Region in detail in the left hand menu. Turn on the multiple alignment, constrained elements and conservation score for 11 fish EPO tracks, all under the Comparative genomics menu by configuring the page.

The 11 fish EPO track just shows that the whole region for the sardh gene can be aligned among those eleven species of fish. The Constrained elements and Conservation score tracks show the conserved sequence is located where in the alignment.

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

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

Exercise 3 – Synteny

(a) Change the species to dog next to the image.
Yes, there are multiple syntenic regions in dog to human chromosome 3, which is in the centre of this view. Dog chromosomes 6, 20, 23, 31, 33, and 34 have syntenic regions to human chromosome 3.

(b) Scroll down to the bottom of the page.

There is a homologue in dog of human \textit{RHO}. Click Centre on gene RHO to compare the genes between human and dog in this syntenic block.

\textbf{Exercise 4 - Whole genome alignments}

(a) Go to the Ensembl homepage (http://www.ensembl.org/ [25]).

Select \textit{Search: Human} and type \textit{brca2} in the search box.

Click Go.

Click on 13:32315474-32400266:1 below \textit{BRCA2 (Human Gene)}.

You may want to turn off all tracks that you added to the display in the previous exercises as follows:

Click Configure this page in the side menu.

Click Reset configuration.

SAVE and close.

(b) Click Configure this page in the side menu

Click on BLASTZ/LASTZ alignments under the Comparative genomics menu. Select Chicken (\textit{Gallus gallus}) - LASTZ\_NET – Normal, Chimpanzee (\textit{Pan troglodytes}) – LASTZ\_NET – Normal, Mouse (\textit{Mus musculus}) – LASTZ\_NET – Normal and Platypus (\textit{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 \textit{BRCA2} seem to be highly conserved between the various species, which is what is to be expected because these are supposed to be under higher selection pressure than intronic and intergenic sequences.

(c) Click Configure this page in the side menu.

Click on Conservation regions under the Comparative genomics menu.

Select Conservation score for 40 eutherian mammals EPO\_LOW\_COVERAGE, Conservation score for 24 amniota vertebrates Pecan and Constrained elements for 24 amniota vertebrates Pecan.

SAVE and close.

Both the Conservation score and Constrained elements tracks largely correspond with the data seen in the pairwise alignment tracks; all exons of the \textit{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.

SAVE and close.

The conserved regions will be shown in light blue.

**Bonus Exercise 5 – Exporting paralogues with BioMart**

Start at ensembl.org/biomart/martview. Choose the ENSEMBL Genes database. Choose the Homo sapiens 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 Associated Gene Name. Under PARALOGUES: select Human Paralogue Ensembl Gene ID, Human Paralogue associated 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.

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

This webinar was held on the 11th May 2017 at 9am BST and was presented by Ben Moore.

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.

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 [30].
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?

(c) Can you observe the relevant transcription factor binding to the transcription factor binding motifs you identified in part (a) in any cell types? What other transcription factors are also found at this location in this/these cell type(s)?

Exercise 3 - Finding regulatory features with BioMart

An ~30 kb deletion on chromosome 18 (46035000-46065000) has been identified in a patient with blood defects. This region is non-genic and you suspect it may contain an enhancer.

(a) Use BioMart to export a list of predicted enhancers within this locus.

(b) You are particularly interested in enhancers that are active in eosinophil:VenousBlood cells, studied by Blueprint. Which of the enhancers in this region have evidence of activity in these cells?

(c) Click through from the active enhancers to the Regulation tab. Can you find what histone
modifications occur in these loci in this cell type?

### 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 R6:132445867-132513198:-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.

H3K4me1/2/3 modifications have been found in the 5’ region of STX7 in all three cell types, along with a variety of other modifications, including H3K9ac.

(e) Click on Configure this page and choose the DNA Methylation menu. Turn on the track Jurkat
Ensembl browser webinar series
Published on EMBL-EBI Train online (http://www.ebi.ac.uk/training/online)

**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](http://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 red predicted promoter at the 5' end of **APOE**. Click on it to get a pop-up with its ID: ENSR00000110117.

It contains five black lines, which are transcription factor binding motifs (you may need to scroll to see all sites)

(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 (click on ‘Active in’ to view the full list).

(c) 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 click to turn on evidence for the transcription factors, then close.

SP1, TAF7, and USF1 binding is detected in the area of the TF binding sites in H1ESC cells; HEY1, Rad21, TAF1, and USF1 binding is detected in this area in HepG2 cells; ZBTB7A binding is detected in this area in K562 cells.

**Exercise 3 - Finding regulatory features with BioMart**

(a) Start at [ensembl.org/biomart/martview](http://ensembl.org/biomart/martview). Choose the ENSEMBL Regulation database. Choose the Homo sapiens Regulatory Features dataset.

Click on Filters in the left panel. Select Chromosome – 18. Select Base Pair and input the coordinates. Select Feature Type – Enhancer.

Click on Attributes in the left panel. Select Regulatory stable ID along with the default attributes.

Click the Results button on the toolbar. Select View All rows as HTML or export all results to a file.

(b) Click on Filters again. Select Epigenome: eosinophil (VB).

Click on Attributes. Select Activity.

Click the Results button.

There is one active enhancers in this region in eosinophil:VenousBlood cells: **ENSR00000102855**.
(c) Click on the reg-feat IDs ENSR00000102855 to get to the regulation tab for each. Go to Details by cell type.

Go to Select cells then select eosinophil (VB). Save and close the menu. Open the Select evidence menu and choose all the histone modification, then save and close.

The enhancer has the histone modifications H3K36me3 and H3K4me1 in eosinophil:VenousBlood cells, which are known to be associated with activation.

**Viewing your data in Ensembl and advanced ways to access Ensembl data**

This webinar will be held on the 18th May 2017 at 9am BST and will be presented by Emily Perry.

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

The first part of this final webinar will show you how you can view custom data, such as BED or BAM files, in the Ensembl browser.

We then introduce some of the more advanced methods of accessing Ensembl data, such as using REST API, Perl API, FTP site and MySQL queries.

You can catch last year's webinar on our [YouTube channel](http://www.ebi.ac.uk/~emily/Workshops/BAM/) or you can [watch the video on YouKu](http://www.ebi.ac.uk/~emily/Workshops/BAM/).

You can also [download the slides](http://www.ebi.ac.uk/~emily/Workshops/BAM/).

**Advanced exercises**

**Advanced Access**

**Exercise 1 — Attaching URLs of large files**

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

Here you can see a number of BAM files (.bam) with corresponding index files (.bam.bai). We’re 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.

(a) Attach and view the BAM file of human chromosome 20 RNASeq data.
(b) Go to the region on chromosome 20 that contains gene CDH22. 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 individual sequence of the RNASeq reads.

(d) Remove the track from your region in detail view.

**Exercise 2 — REST API endpoint queries**

Complete the following exercises using single REST API endpoint queries.

(a) Get the sequence for the region from basepair 32889000 to 32891000 of human chromosome 13 in FASTA format. Hard-mask and soft-mask the sequence. How many repeat regions are there in this sequence?

(b) Get the Ensembl Gene ID for the human **CCR5** (chemokine (C-C motif) receptor 5) gene.

(c) Has an orthologue for this gene been identified in chimpanzee?

(d) A famous variant in the human **CCR5** gene is the delta 32 allele, a 32-basepair deletion at position 46373456-46373487 (rs333). Individuals carrying one copy of the delta 32 allele seem to be resistant to infection by HIV, the virus that causes AIDS, and individuals with two copies (delta 32 homozygotes, ~1% of Caucasians) are almost completely immune to infection by HIV. The delta 32 allele may have been selected for in European populations because it confers resistance to plague (Black Death) or smallpox.

The HGVS notation for this variant is 3:g.46373456_46373487

What is the effect of this variant on the CCR5 protein?

**Exercise 3 — Methylation data in human (synoptic exercise)**

This exercise requires you to combine the knowledge you have gained about different aspects of Ensembl. It is designed to be challenging and force you to come up with solutions yourself.

The human **PDHA2** gene, that encodes for a subunit of the pyruvate dehydrogenase complex, is exclusively expressed in spermatogenic cells. In the paper ‘Human testis-specific **PDHA2** gene: Methylation status of a CpG island in the open reading frame correlates with transcriptional Activity’ (Pinheiro et al Mol Genet Metab. 2010 Apr;99(4):425-30), two CpG islands in the **PDHA2** gene are reported, one encompassing the core promoter region and extending into the open reading frame, the other exclusively located in the coding region. The latter CpG island was shown to be methylated in somatic tissues but demethylated in testicular germ cells and has therefore been proposed to play an important role in the tissue-specific expression of the **PDHA2** gene.

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

(b) Turn on the CpG islands track. Two CpG islands are reported in the **PDHA2** gene by Pinheiro et al (2010). Do they appear in this track? If not, why not?

(c) Confirm the existence of the two CpG islands using the EMBASS program CpGPlot (http://www.ebi.ac.uk/Tools/seqstats/emboss_cpgplot) on the sequence around the **PDHA2** gene.
(d) Upload the CpG islands found by CpGPlot using Custom tracks. Use BED format, which in its simplest form just consists of the chromosome and the start and end coordinates, separated by spaces (as an optional fourth field, you can add a name/description). The genomic start and end coordinates of the CpG islands can be calculated from the genomic start coordinate of the sequence on which the CpGPlot program was run and the relative location of the CpG islands on this sequence as given by the CpGPlot output.

(e) Create a link to allow you to show your new BED track to colleagues, compared to the %GC track.

(f) Is there a regulatory feature at the 5’ end of the PDHA2 gene? What type? Which cell type(s) is it active in? What evidence supports the presence of this feature?

(g) Turn on the RNA-seq tracks for different tissues. Is there evidence that PDHA2 is expressed in one tissue more than others?

(h) How well conserved is the region of the PDHA2 gene amongst the 40 eutherian mammals? Are the CpG islands conserved?

(i) How many biological processes are associated with PDHA2 by Gene Ontology (GO) terms? Can you export the sequences of all human genes that are also associated with the first of these terms?

(j) Can you fetch the gene sequence for PDHA2 in FASTA using the Ensembl REST API?

**Exercise solutions**

**Advanced Access - solutions**

**Exercise 1 — Attaching URLs of large files**

(a) 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 in the URL of the BAM file itself (http://www.ebi.ac.uk/~emily/Workshops/BAM/GRCh38.20.illumina.merged.1.bam)

Since this is a file, the interface is able to detect the “.BAM” file extension, so 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 add the custom track from the ‘Personal Data’ menu. Select Unlimited track style. You can see that there are more RNASeq reads that map to the exons 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 the Page. and turn off this track by selecting Off in the track style of the personal data track.

You can also remove the custom data by clicking on Manage your Data and then clicking on the trash can icon associated with this data.

**Exercise 2 — REST API endpoint queries**

Complete the following exercises using single REST API endpoint queries.
(a) Go to the REST API documentation page at http://rest.ensembl.org/documentation

Click on GET sequence/region/:species/:region to get the documentation for this command.

Use the documentation to construct a URL in the correct form and add the genomic co-ordinates from chromosome 13 to create the URL
i.e: http://rest.ensembl.org/sequence/region/human/13:32889000..32891000:1?content-type=fasta

This URL will give you the sequence.

To return a hard-masked or soft-masked version of the region sequence, use the optional URL additions using the provided format: http://rest.ensembl.org/sequence/region/human/13:32889000..32891000:1?content-type=fasta;mask=hard

http://rest.ensembl.org/sequence/region/human/13:32889000..32891000:1?content-type=fasta;mask=soft

Hard mask will mask all repeats as N's and soft mask will mask repeats as lower-case characters. There are four separate repeat features in this region.

(b) Click on GET xrefs/symbol/:species/:symbol to get the documentation for this command.

Use the documentation to construct a URL in the correct form and add the gene name to create the URL i.e: http://rest.ensembl.org/xrefs/symbol/homo_sapiens/CCR5?content-type=application/json

This URL will give you the Ensembl stable ID: ENSG00000160791.

(c) Click on GET homology/id/:id or GET homology/symbol/:species/:symbol to get the documentation for this command (there are two separate queries that allow you the search by either ID or symbol).

Use the documentation to construct a URL in the correct form and add the Ensembl stable gene ID (from (b)) to create the URL. You can filter by the Latin name of chimpanzee (Pan troglodytes) i.e: http://rest.ensembl.org/homology/id/ENSG00000160791?content-type=application/json;target_species=pan_troglodytes

Or

Use the documentation to construct a URL in the correct form and add the associated name (CCR5) to create the URL. You can filter by the Latin name of chimpanzee (Pan troglodytes) i.e: http://rest.ensembl.org/homology/symbol/human/CCR5?content-type=application/json;target_species=pan_troglodytes

Either of these URLs will give you the human and chimpanzee orthologous pair.

(d) Click on GET vep/:species/hgvs/:hgvs_notation or GET vep/:species/id/:id to get the documentation for this command (there are two separate queries that allow you the search by either variant ID or HGVS notation).

Use the documentation to construct a URL in the correct form and add the HGVS notation to create the URL i.e: http://rest.ensembl.org/vep/human/hgvs/3:g.46373456_46373487del?content-type=application/json

Or

Use the documentation to construct a URL in the correct form and add the variant ID to create the URL i.e: http://rest.ensembl.org/vep/human/id/rs333?content-type=application/json
This variant is a frameshift variant on the \textit{CCR5} gene (listed as “most severe consequence”).

\textbf{Exercise 3 — Methylation data in humans (synoptic exercise)}

(a) Go to the Ensembl homepage (\url{www.ensembl.org} [18]).

Select \textit{Search: Human} and type \textit{PDHA2} in the for text box. Click \textit{Go}.

Click on \texttt{4:95840019-95841474:1}.

Zoom out, so that approximately 5kb region around the PDHA2 gene is shown.

You may want to turn off all tracks that you added to the display in the previous exercises as follows:

Click \textit{Configure this page} in the side menu.

Click \textit{Reset configuration}. SAVE and close.

(b) Click \textit{Configure this page} in the side menu.

Type \texttt{cpg} in the Find a track box.

Select \textit{CpG islands}. SAVE and close.

No CpG islands are shown. As for the inclusion of CpG islands into the Ensembl database for human a minimum length of 400 bp is required, the reason for this could be that the CpG islands in the \textit{PDHA2} gene are shorter than 400 bp. However, there is a %GC track, which shows that the region that comprises the 5’ part of the \textit{PDHA2} gene and the region directly upstream of the gene has a high %GC (the red line in the %GC track indicates 50% GC). It is difficult / impossible to distinguish individual CpG islands in this track, though.

(c) Click \textit{Export data} in the side menu.

Click \texttt{Next}.

Click on \texttt{Text}.

Select and copy the sequence.

Go to \url{http://www.ebi.ac.uk/Tools/emboss/cpgplot/index.html} [36].

Paste the sequence into the text box. Click \textit{Run}.

CpGPlot does confirm the existence of two CpG islands in the \textit{PDHA2} gene region of lengths 200 and 263 bp, respectively. So, it is indeed because of their length being less than 400 bp that these CpG islands are not present in the Ensembl database.

(d) The genomic coordinates of your CpG islands are the start coordinates of your region of interest (found at the top of your exported FASTA) plus the coordinates of the islands within that region (from EMBOSS). In my case this is:

First island: start = 95839291 + 734 = 95840025

end = 95839291 + 933 = 95840224

Second island: start = 95839291 + 1058 = 95840349
This gives coordinates for my CpG islands in BED format as:

```plaintext
chr4 95840025 95840224 cpg_island_1
chr4 95840349 95840611 cpg_island_2
```

Click Custom tracks in the side menu.

Type CpG islands in the Name for this upload (optional) box.

Select Data format: BED.

Copy the BED formatted data into the Paste file box. Click Add data.

Click on Go to nearest region with data: 4:95790125-95890125.

The two CpG islands should now be shown on the Region in detail page. They should coincide with the regions of high %GC.

Zoom in on the two CpG islands.

To display the names of the CpG islands:

Hover over the CpG islands track name.

Hover over the icon of the cog-wheel.

Select Labels.

(e) Drag your CpG islands track so that it is next to the %GC track.

Click Share this page in the side menu.

Select the link and copy.

Paste into your internet browser to view.

(f) There is a cyan CTCF binding site 5’ of PDHA2.

Click on the feature then the ID ENSR00002011748 to get to the regulatory tab.

The CTCF binding site is active in A549, DND-41, GM12878, H1ESC, HMEC, HSMM, HUVEC, HeLa-S3, K562 and NHEK cells.

Click on Details by cell type, then Select cells and choose A549, DND-41, GM12878, H1ESC, HMEC, HSMM, HUVEC, HeLa-S3, K562 and NHEK and close, then Select evidence and choose ALL ON and close.

CTCF-binding is found in many of the cell types. Rad21 binding is also seen in H1ESC cells.

(g) Click on Configure this page, then select RNASeq models. Turn on the BAM files for all the tissues in Coverage only.

You will see histograms of RNA-seq coverage for each of the tissues. The largest number is for the merged read. For the tissue-specific read, Testes have a higher peak than all the other tissues. There
are also wider peaks in the Testes track that cover the whole gene, whereas other tissues only have a peak at the 3’ end of PDHA2.

(h) Click on Configure this page, then select Comparative genomics. Turn on the tracks for the Constrained elements for 40 eutherian mammals and Conservation score for 40 eutherian mammals.

The region of the gene itself has high GERP scores, indicated by constrained elements over most of the gene. There is no apparent difference in the conservation score between the CpG islands and their flanking regions.

(i) Click on the Gene Tab, Gene: PDAH2 and select GO: Biological process.

There are seven terms in the table, the first being GO:0005975, carbohydrate metabolic process.

To export the list use BioMart.

Click on Search BioMart in the table. This will take you to a BioMart results table with the gene and transcript IDs, GO terms and gene position.

Click on Attributes.

Choose Sequences.

Expand SEQUENCES and select Unspliced (Gene).

Expand Header information and deselect Ensembl Transcript ID.

Click Results.

You can export these results if you wish.

(j) Go to the REST API documentation page at http://beta.rest.ensembl.org/documentation.

Click on GET sequence/id/:id to get the documentation for this command.

You will need the stable ID of PDHA2, go to the browser page to find that it is ENSG00000163114.

Use the documentation to construct a URL in the correct form, ie: http://beta.rest.ensembl.org/sequence/id/:id?format=fasta

Add the ID to the URL to create: http://beta.rest.ensembl.org/sequence/id/ENSG00000163114?format=fasta

This URL will give you the sequence.

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 [37] or YouKu [38] channel.
- Try out an Ensembl tutorial [39]. There are videos and coursebooks on a variety of subjects.
- See if your question has already been answered in the Ensembl FAQs [40].
- Have a look at the Ensembl Glossary [41].
- View technical documentation on Ensembl [42].
- Visit our further courses on Train online: Ensembl: Browsing chordate genomes [43] and Ensembl: Filmed API workshop [44].

Support

- If you cannot find the answer to your question, contact the Ensembl helpdesk [45].
- If you are using Ensembl programmatically, our dev list [46] 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 [47] 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.

Denise Carvalho-Silva [55]
EMBL-EBI
Ensembl Outreach Officer

Denise joined the Ensembl Outreach team in September 2011 and has been training and supporting worldwide communities on the usage of both Ensembl and Ensembl Genomes databases through workshop, social media and helpdesk. She got a PhD degree in Biochemistry (Universidade Federal de Minas Gerais, Belo Horizonte, Brazil), a post-doctoral training in Genetics (Australian National University, Canberra, Australia) and was a Research Associate in Human Evolution at the Wellcome Trust Sanger Institute (Hinxton, UK). She has also worked as a Senior Computer Biologist in the HAVANA team on the GENCODE, EUCOMM and Pig Genome projects before joining the EMBL-EBI.
Ensembl browser webinar series
Published on EMBL-EBI Train online (http://www.ebi.ac.uk/training/online)

Ben Moore [56]
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 [57]
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.

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

Links
[1] http://www.ebi.ac.uk/training/online/trainers/emily
[2] https://docs.google.com/forms/d/e/1FAIpQLSe8Zw7SjQfr7WtKh0PCNstwSHVaQU2ZaNUhR0WdKgSjAvUQ-g/viewform
[3] https://www.surveymonkey.co.uk/r/NT5JZWZ
[7] http://www.ebi.ac.uk/training/online/course/ensembl-online-training-series/variation-data-ensembl-
[53] http://europepmc.org/articles/PMC4407537
[54] http://europepmc.org/articles/PMC4271150
[55] http://www.ebi.ac.uk/training/online/trainers/denise
[56] http://www.ebi.ac.uk/training/online/trainers/bmoore
[57] http://www.ebi.ac.uk/training/online/trainers/hsparrow