Ensembl: Browsing genomes

Emily Perry [1], Giulietta M. Spudich [2], Ben Moore [3], Helen Sparrow [4]

- DNA & RNA
- Beginner
- 3 hours

This course focuses on genomes on the Ensembl website at www.ensembl.org. It provides a quick beginner's guide to the overall structure of the Ensembl genome browser.

Learning objectives:

- Understand what Ensembl is and where the data come from
- Know how to access and navigate the Ensembl homepage
- Be able to search the Ensembl browser with a gene, location (a region of a genome), or polymorphism of interest
- Know how to explore a region on a genome, a gene and a transcript
- Understand where to view gene trees, sequence variation, and possible regions involved in gene regulation in Ensembl
- Be able to view a sequence for a gene, protein, or a genome of interest
- Understand where the genes come from
- Understand the different ways the results are presented in Ensembl (including the gene, location, transcript and variation tabs)
- Be able to export Ensembl data quickly and easily with the browser and BioMart

What you will learn

After completing this course, you should:

- know how to access and navigate [7] the Ensembl homepage;
• be able to search the Ensembl browser with a gene, location (a region of a genome), or polymorphism of interest; [8]

• know how to explore a region [9] on a genome, a gene and a transcript;

• understand where to view gene trees, sequence variation [10], and possible regions involved in gene regulation [11] in Ensembl;

• be able to view a sequence [12] for a gene, protein, or a genome of interest;

• understand where the genes [13] come from;

• understand the different ways the results are presented in Ensembl (including the gene, location, transcript and variation tabs [14]);

• be able to export Ensembl data quickly and easily with the browser and BioMart [15];

Why Ensembl?

Figure 1. The Ensembl genome browser provides access to organised information from the analysis of biological data.

The vast amount of information that comes with annotating a genomic sequence demands a way of organising and accessing that information (Figure 1). This need is met by Ensembl – a genome browser providing free access to the complete sequences of higher and model organisms.

Biological databases are an important resource for the life sciences community. Keeping up-to-date with the hundreds of databases supporting molecular biology and related fields is a daunting and time-consuming task. Integrating this information into one access point is a necessity.

Genome browsers and their underlying databases act as single entry points to data from multiple projects and genomic analyses, such as genes and proteins, sequence variation, comparative genomics and motifs involved in gene regulation [11].

Ensembl and Ensembl Genomes [16] are major projects integrating and displaying genome annotation [17] for multiple species.
What is Ensembl?

Ensembl [5] provides a genome browser that acts as a single point of access to annotated genomes for mainly vertebrate species (Figure 2).

Information such as gene sequence, splice variants and further annotation [17] can be retrieved at the genome, gene and protein level. This includes information on protein domains, genetic variation, homology, syntenic regions and regulatory elements. Coupled with analyses such as whole genome alignments and effects of sequence variation on protein, this powerful tool aims to describe a gene or genomic region in detail.

Ensembl imports genome sequences from consortia which keeps us consistent with many other bioinformatics projects. Each species in Ensembl has its own home page, where you can find out who provided the genome sequence and which version of the genome assembly is represented. For an example, see the human [18] home page.

Figure 2. Ensembl contains information on chordate genomes, including human, mouse, rat, zebrafish, panda, and takifugu.

The Ensembl project
Ensembl project seeks to enable genomic science by providing high quality, integrated annotation on chordate and selected eukaryotic genomes within a consistent and accessible infrastructure.

All data are open access and all software is open source, i.e. freely available to the scientific community.

The project began in 1999 as a joint project between the EMBL European Bioinformatics Institute and the Wellcome Trust Sanger Institute (then named the Sanger Centre). This changed between 2014 and 2017, with all staff and eventually all hardware moving over to the EMBL-EBI.

Supported species include comprehensive, evidence-based gene annotations and a selected set of genomes includes additional data focused on variation, comparative, evolutionary, functional and regulatory annotation. The most advanced resources are provided for key species including human, mouse, rat and zebrafish reflecting the popularity and importance of these species in biomedical research.

Frequent updates!

The website and underlying databases (along with the application programming interfaces (APIs)) are updated every 2-3 months. Previous releases of Ensembl can be accessed from the Ensembl Archive.

As of Ensembl release 89 (May 2017), 69 species are supported, along with 16 mouse strains. Associations between sequence variants and diseases, and potential gene regulation sequences across different cell types are among the cutting edge data integrated and displayed in the Ensembl browser and underlying databases.

Where are we?

Ensembl headquarters are in Hinxton, in Cambridgeshire, England.

When to use Ensembl
You can use Ensembl [5] if you:

- have a gene of interest, and you would like to know if there are homologues in other species, or any sequence variations in the gene;
- would like to know what the sequence is for your gene of interest, and what the sequences of the splice variants (transcripts) are;
- want to explore the region around a gene of interest, and find neighbouring genes;
- want to find sequences that may be involved in gene regulation [11] (open chromatin signatures, transcription factor binding sites, etc.);
- are interested in how conserved a gene or region is across species;
- want to know a selection of sequence variants that have been associated with a disease, for example, diabetes;
- have questions about a gene, variant [27], or chromosomal region;

Ensembl cannot help you if:

- want to submit sequence files (see the course on ENA [28]);
- you are looking for metabolic pathways (learn more about Reactome [29]);
- your species of interest is not a chordate [20] (see a sister project, Ensembl Genomes) [30];

How to access Ensembl

Access Ensembl [5]:

- Search [8] our website at www.ensembl.org [31]
- Use BioMart [15] to quickly obtain tables of gene information
- Access data using the Perl APIs [32]

Our website

The menu bar. The menu bar at the top provides links to the homepage, to tools, [33], to help and documentation [34] and a search box. It is visible on every Ensembl page.

Searching [8] Ensembl. Look for a gene, location, variant [27] and more using the search box on the homepage or the box that is provided in the top right corner of any Ensembl page (Figure 3).

Browse a Genome. Choose your species of interest in this section. The drop down menu under ‘All genomes’ allows you to select from the full list.

What’s New in Ensembl Release [26] 89? To find out what release you are working with, have a look at the news section of the homepage, at the top right of the page. Below this you’ll see links to social media, including blog posts and Twitter. Archive sites [25] allow access to previous versions, or releases, of Ensembl.

How to search Ensembl

Figure 4. The Ensembl [5] search box
Search [www.ensembl.org](https://www.ensembl.org) using:

- a **gene name** (for example, *BRCA2*);
- an **identifier** ([35](#)) from an external database, such as [UniProt](https://www.uniprot.org) **accession** ([37](#)) number or a [PDBe](https://www.ebi.ac.uk/pdbe) **ID**;
- a **disease** name (for example, coronary heart disease);
- a **variation** (for example, rs1223);
- a **location** - a genomic region (for example, rat X:100000..200000);
- a [Gene Ontology](https://www.geneontology.org) (**GO**) term

Most search results will take you to the appropriate Ensembl view through a results page. If you search using a location you will be directed straight to the **location tab** ([9](#)) (this tab provides a view of a region of a genome).

The different tabs in Ensembl will be explained in the later section 'Navigating Ensembl' ([7](#)). For now we will concentrate on how to search Ensembl.

Want to find genes or genomic regions for a sequence? See the section 'Searching with a sequence using BLAT or BLAST' ([40](#)).

### Searching for the BRCA2 gene

Let's search for a gene. In our example, we want to know more information about the *BRCA2* gene, such as the genomic location, neighbouring genes and the sequence.
Figure 5. The search box.

You can use the search box on the home page, as shown in Figure 5. Or, click on a species first, and search from, for example, the Human home page.

Try it!

1. Open the Ensembl homepage at www.ensembl.org [31]
2. Choose your species of interest (Human) using the pull-down menu to the left of the search box.
3. Type in your search term of interest into the search box. In our example we are using the gene name 'BRCA2'.
4. Click ‘Go’ to obtain the search results
5. You should see the BRCA2 gene at the top of the list.

Searching with a sequence using BLAT or BLAST

If you have a sequence, but you are not sure what the gene name or Ensembl [5] ID is, you can align it to the genome with BLAST [41] or BLAT [42].
BLAT with the *MTAP4* gene sequence

Try it! The beginning of the *MTAP4* gene sequence is shown. Copy it and move on to the steps below.

CTCCGCACGTCTCACTCCCCGAGTGAGGTTGGCAGACAGCCACC
CTCTCGCTGTGCCTGTTGCTCCCTTTAGCCTCCGAGGGTCGTCGCCAC
TGCAGATTCTTTTTGGTCGACATGGCCT

1. Click on the BLAST/BLAT [43] link at the top of the page.
2. Paste your sequence into the box.
3. Check the options are correct. For example, we have selected *Homo sapiens* as the species to search against and the BLAT search tool because we're looking for an identical match.
4. Click 'Run'.

BLAT vs BLAST ... What's the Difference?
BLAT (The BLAST-Like Alignment Tool) is fast, but it demands more exact matches. BLAST will allow lower-scoring hits, and allows more gaps in alignments. You'll get more hits with BLAST (but it may be slower).

Quiz: Searching Ensembl

Questions: 5
Attempts allowed: Unlimited
Available: Always
Pass rate: 75 %
Backwards navigation: Allowed

Exploring sources of biological data

A wealth of biological data can be viewed, downloaded and compared such as:

- genes
- conserved sequences across species
- sequence variation

*Ensembl* [5] brings together information from multiple resources, using the genome as a base for this annotation [17].

Tired of reading? Check out what we had to say about Ensembl at Erasmus MC!
Ensembl genes

The Ensembl [5] gene set is based on evidence, and includes manual annotation [44] for our most used species (Figure 6).

![Diagram showing the relationship between genome sequence, mRNA, protein, and genes, transcripts, proteins, and positions.]

Figure 6. Sequences in public databases are aligned to the genome in order to determine positions of genes, along with splice variants.

How can I view the genes, and information like sequence? Jump to this section [45], or watch the video below.

The GeneBuild

The initial step is to obtain sequenced genomes from official centres. The sequenced genomes are then annotated in the Ensembl pipeline (also known as the Ensembl genebuild [46]) using both automatic and manual annotation for some species. Human, mouse, zebrafish, pig and rat gene sets include manual annotation from the HAVANA [47] project. The Ensembl gene set for human, including Havana transcripts, is the GENCODE [48] set.

All Ensembl transcripts are based on experimental evidence, and draw on mRNAs and protein sequences deposited into public databases (such as UniProt [36] and NCBI RefSeq [49]) from the scientific community. The Ensembl gene set also includes automatically-annotated pseudogenes, non-coding RNAs [50], and alternative splicing events [51] for model organisms. The resulting analyses of the genomes are stored in the Ensembl databases and can be accessed via the Ensembl website [52], BioMart [53] and programmatically [54].
See the [annotation article](#) for more about the Ensembl genebuild pipeline, gene names and annotation.

**Multiple species**

Comparative genomics analyses performed by Ensembl [5] allow you to directly compare included species across many mostly chordate [20] genomes. Whole genome alignments, [56] homologues arising from gene trees, protein families and synteny [57] are available.

Whole genome alignments include sequence alignments between two species ('pairwise alignments'), and multispecies alignments incorporating genomes of more than two species (Figure 7). Analysis of the pairwise alignments results in syntenic regions (i.e. where the sequence and gene order is conserved between two species). To infer evolutionary relationships between species, phylogenetic gene trees (or protein trees) are constructed, resulting in homologues.

![Alignment, Gene Tree, Orthologues, Paralogues, Families](#)

[58]

**Figure 7.** A sampling of comparative genomics views is shown.

Read more about comparative genomics data in Ensembl [here](#).

**Sequence variation**

The majority of sequence variants in Ensembl [5] are single nucleotide polymorphisms (SNPs), insertions and deletions (indels) imported from NCBI dbSNP [60]. For human SNPs in particular, we aim to keep current with dbSNP, updating these with every Ensembl release [26] (every 2-3 months). Projects submitting their variants to dbSNP include individual labs, the 1000 genomes project [61], ExAC [62] and gnomAD [63]. Small sequence variants are mapped onto the reference genome, and effects on Ensembl transcripts are determined. Larger structural variations (such as copy number variation) are also viewable on the genomic sequence. These include structural variants from dGVA [64] and somatic mutations.
Determine the effect of your variants on the Ensembl transcript set using the Variant Effect Predictor [65] tool.

Genotype [66] information for human variants is imported from dbSNP, and reflects data from individual submissions, HapMap [67], and the 1000 Genomes project. Disease and phenotype [68] associations are imported from projects such as the GWAS [69] catalog, OMIM [70] and EGA [71].

Learn how to navigate variation views in this section [10]. Read more about variation sources [72] in Ensembl.

Gene regulation

The Ensembl [5] regulatory build (available for human and mouse) provides a set of regulatory features [73], sequences that may be implicated in gene regulation [11]. These sequences have indicators suggesting they may be promoters, enhancers, or other elements. The features are based on regions of open chromatin identified using DNase I hypersensitivity assays, transcription factor binding sites, CTCF [74] binding regions and histone modifications. These data come from experiments analysed by mainly ChIP-seq, but also DNase-seq experiments, from the ENCODE [75] (Encyclopaedia of DNA elements), Blueprint [76] and Roadmap Epigenomics [77] projects. Ensembl displays the activity of these features in different cell types, and the evidence used to determine this activity.

Find out how to view sequences implicated in gene regulation in this section [78]. Read more about the regulatory build here [79].

Navigating Ensembl

After you have searched Ensembl [5] with, for example, a gene name, you will be taken to the appropriate main page (or tab) (see ‘How to search Ensembl [8]’ for more information). The navigation of the Ensembl website is organised into tabs, or main pages (Figure 8).

The location, gene, transcript, variant [27] and regulation tabs allow data browsing at that level. For example, the gene tab provides information on one specific gene.

Figure 8. The navigation of the Ensembl website is divided into tabs, or summary pages. Note: to get this figure, we opened all the possible tabs by clicking on a specific transcript, variant and regulatory feature. Normally, after searching for a gene, you see only the location and gene tabs.
The location tab[9] (Location: 7:114,414,997-114,693,768). The location tab allows viewing of a region, including genes in that area, variations, regulatory regions, along with over a hundred different data types.

The gene tab[80] (Gene: FOXP2). This contains all the information associated at the gene level. For example, the genomic sequence for the FOXP2 gene will be found in the gene tab.

The transcript tab[81] (Transcript: FOXP2-210). This tab contains transcript-level information, for example the cDNA sequence and position of untranslated regions (UTRs).

The variation tab[10] (Variant: rs181717058). The variation tab is focused on a SNP[82], an indel, a somatic mutation or a structural variant. Find population frequencies, alleles[83] and associated phenotypes and diseases in this tab.

The regulation tab[78] (Regulation: ENSR00000216893). This tab provides potential genomic regions involved in gene regulation[11], with their activity in different cells types.

The jobs tab (Jobs). Any jobs submitted to Ensembl tools[33] can be accessed here.

A first look at our views

The tabs are split into four main areas (Figure 9).
Figure 9. The four main areas of the Gene tab for human FOXP2 (view it in our browser [85]).

**Top section [blue].** The top section contains general information on the data in the tab selected. For example, in the gene tab it contains the gene name and ID, the gene's location and a button to view a table of the transcripts.

**Bottom section [green].** This section contains more specific information relating to the view. It is represented in either graphic or tabular format. For example, on the gene tab it contains the gene summary which displays the transcripts on the chromosome.

**Displays [pink].** The displays on the left-hand side are specific to the tab you are on; you can access more information on the data of interest. For example, on the gene tab, from the links in the 'Gene-based displays' section you can access information such as: splice variants, the gene sequence, homologues and variation data.

**Configuring and managing your data [yellow].** The options in these boxes allow you to configure the data (i.e. change the data types) displayed in the main sections (top and bottom) of the page, to manage your data (i.e. upload your own information) and export it.

*Note!* When you navigate between views in one tab, only the bottom section [green] changes. The top section [blue] remains the same. Open the gene tab and click on another view like ‘Sequence’ to see this.
Investigating a gene

When you search the browser with a gene name, you can click on the results to go the gene tab. This tab contains all the information associated at the gene level; for example, the genomic sequence for the *BRCA2* gene.

![Figure 10.](image)
The gene tab top section provides an overview of the splice variants for one specific gene (in this example, BRCA2 transcripts).

**Gene IDs.** [Ensembl](https://www.ensembl.org) gene IDs begin with ENS for Ensembl, and then a G for gene. In this case, **ENSG00000139618** is the Ensembl gene ID for BRCA2. The number is unique, and should not change, even if the gene is updated. A three-letter code is inserted into the identifier, in the case of species other than human. For example, **ENSMUSG00000041147** would indicate the BRCA2 gene in mouse (*Mus musculus*).

**Synonyms.** Alternative gene symbols that have been used in the literature to describe this gene.

**Location.** The genomic location and strand where this gene is found.

**Transcript table.** At the top of the gene summary, the number of transcripts, or splice isoforms, are shown in a table. In our example, for [Ensembl release](https://www.ensembl.org) 89, there are seven transcripts for the human BRCA2 gene. You can find more information about the transcripts in the 'Biotype' column. In our example, two transcripts in the table are protein coding (BRCA2-201 and BRAC2-206), two result in proteins that undergo nonsense mediated decay (BRCA2-202 and BRCA2-203), and three not translated (BRCA2-204, BRCA2-205 and BRCA2-207).

It’s a good idea to write down the gene identifier of any gene you work with, and the date. This allows you to go back to that specific gene model in Ensembl, even if an update has occurred.

See the Ensembl documentation for extensive information about [genome annotation](https://www.ensembl.org/info/about/).
Gene summary

The gene summary view in more detail

Let's look more in depth at the figure on the gene summary view (Figure 11).

Figure 11. The lower half of the display gives a summary of the gene source, biotype and annotation [17]. All transcripts in the genomic region are depicted below.

Summary. The official gene name is displayed at the top, along with other names and symbols the gene is known by in other major databases, such as RefSeq [49] and Uniprot [36]. For human genes, the official gene name comes from the HGNC [86] (HUGO [87] Gene Nomenclature Committee).

There is also information about the gene biotype and how the gene was annotated.

BRCA2 Transcripts. All transcripts in the region are displayed. Gold and red transcripts are protein coding whilst blue, pink and grey transcripts are non-coding. The gold transcripts indicate a Havana [47]/Ensembl [5] merge. In this case, BRCA2-201 is gold, indicating the sequence is agreed on by both projects. A help page [88] explains the colours more in depth.

Transcript structure. Boxes are exons, and connecting lines are introns. Filled boxes indicate coding sequence, while unfilled boxes are UTR [89](UnTranslated Regions).

Figure 12. Transcript structure
For example, the transcript above is gold, showing it's a protein coding transcript (Havana/Ensembl merge). It has 27 exons. Some exons have filled and unfilled regions, meaning these exons contain coding sequence and UTR. The introns are lines between the exons.

Transcripts drawn above the blue bar (genome) are on the forward strand of the chromosome (like \textit{BRCA2}), and transcripts drawn underneath (like the neighbouring \textit{N4BP2L} transcripts, [B]) are on the reverse strand of the chromosome.

\textbf{Export image}. To export the image in a variety of formats (including pdf), click the image icon.

### Sequence

To view the gene sequence, click the Sequence link at the left of the gene tab.

\textbf{Figure 13.} The sequence of the \textit{BRCA2} gene

Watch the video to find out how to add sequence variation to the gene sequence.

### Variation

The \textit{variant} [27] \textbf{table} lists short sequence variations like SNPs and indels found within the gene of interest (Figure...
14). The table can be filtered by various facets, such as the consequence on the resultant protein, allele [83] frequency or the evidence for the variant, using the filter buttons at the top.

Figure 14. View the variation table for all SNPs, insertions and deletions mapped to this gene.

For structural variation, such as copy number variation, see the structural variation [90] view.

**Homology**

The orthologues page allows you to view homologues across different species (Figure 15).
Figure 15. divides homologues into different taxonomic groups.

Select a taxonomic group in the table on the orthologues page to reveal genes in the group. A separate paralogues view shows homologues within the same species. Orthologues and paralogues are determined using a representative protein for every gene in Ensembl [5], and comparing them in a phylogenetic tree. See this article [91] for more.

Investigating a transcript

The transcript tab allows exploration of one splice variant [27], such as BRCA2-201 (Figure 17). You can get to the transcript tab by searching for the transcript name, by clicking on a transcript in the transcript table in the gene tab, or by clicking on a transcript in one of the graphical displays.
Figure 17. The transcript tab for BRCA2-001.

We will use the menu on the left to explore:

- **Exon** [92] and cDNA sequence
- Protein domains

Exons and introns

The exons view provides information about specific exons and introns (Figure 18). It may be useful for selecting sequence to design primers.
Figure 18. The exons view, accessed from the transcript tab's left-hand menu, shows UTR [89] (in orange), coding sequence (blue), introns (grey) and flanking sequence (green). Variants are highlighted on the sequence in the colours shown in the legend.

Click on 'configure this page' to add the variant [27] highlighting on the sequence.

**cDNA sequence**

The cDNA view shows the spliced transcript along the protein and CDS [93] sequences (Figure 19).
**Figure 19.** Compare the transcript sequence, coding sequence, and variation in the cDNA view.

The cDNA view shows three sequences aligned on top of each other: the transcript sequence in the first row, the coding sequence alone (no UTR [89]) in the second row, and the protein sequence in the third row. Use the blue 'Configure this page' button on the left to change which sequences you can see.

The UTR is shown highlighted in bright yellow. Codons are shown in alternating pale yellow highlight and no highlight. Alternating exons are shown by switching between black and blue text.

By default, variants are highlighted on the sequence, coloured by consequence according to the legend at the top. Ambiguity codes for the variants are shown above the sequence. Click on one to open the variant [27] tab, explored further on [94] in this course. Amino acids affected by sequence variation are marked in red. Hover over a red amino acid [95] with your mouse to view the alternatives. You can turn off the variants by going to the 'Configure this page' menu.

How do you access the cDNA sequence? Try our guided example [96].
Protein domains and features

The domains of the protein product of the transcript can be viewed graphically (Figure 20) or in a table (Figure 21). The domains were determined by running InterProScan [97] on the sequences, which uses multiple protein domain [98] prediction algorithms. This means you may see multiple predictions of the same domain, with slightly different boundaries.

![Protein domains plotted against the exon structure.](image)

**Figure 20.** Protein domains plotted against the exon [92] structure.

Click on ‘Protein summary’ in the left-hand menu to see a graphical display of the protein domains (Figure 20). The exon structure is shown as alternating shade of purple at the top, while the variants are shown below the domains.

![A table of protein domains](image)

**Figure 21.** A table of protein domains

Click on 'Domains and features' to see a table of the same data (Figure 21). Click on the links in the table to go to the domains in Interpro [99] or view a karyotype highlighting the locations of all transcripts with the domain.
Investigating a genomic region

You can search with a location (see 'How to search Ensembl' [8]) for more information) using the search box on the Ensembl homepage to access the location tab directly. Alternatively, a gene search will also display a link to the location, shown as coordinates (Figure 22).

![Link to the location tab](image)

**Figure 22.** From the results shown for the 'BRCA2' search, you can click on the location link to go directly to base pairs 32,315,474 to 32,400,266 on chromosome 13.

You can also click through the the location tab when in another tab in the browser.

The location tab allows exploration of a genomic region. We will investigate:

- The region in detail view, allowing visualisation of hundreds of data tracks along the genome
- Genes in a region
- How to access a view of a whole chromosome

Finding the location tab

You can search with a gene or location (see 'How to search Ensembl' [100]) for more information) using the search box on the Ensembl homepage to access the location tab directly.

From the results shown for the 'BRCA2 gene' search, you can click on the location link to go directly to base pairs 32,889,611 to 32,973,805 on chromosome 13.

Alternatively, open the location tab (Figure 23) from other views in Ensembl by clicking on a location link, or the location tab itself.
Figure 23. Click on on the location tab to open a view of the genomic region for the *BRCA2* locus.

A region in detail

The location tab's main display is the 'region in detail', which shows a region of the genome up to 1 Mb long in a highly customisable view (Figure 23). The 'Region in detail' display supports data visualisation from numerous sources.

Figure 23. The region in detail display for *BRCA2*.

- **Whole chromosome.** The chromosome is shown at the top, with the position of the region of interest marked by a red box. You can use this image to zoom in and out of the chromosome.
- **Overview.** An overview of all genes, centred on the gene of interest (in this case, *BRCA2*), is shown. Each gene is clickable, and following links to gene or transcript IDs opens the gene or transcript tab, respectively.
The genome is divided into light and dark blue bars, reflecting individual contigs. Information such as markers can be added to the top panel by turning on data types in 'configure this page'.

- **Main panel.** The section in the lower part of the region in detail panel is highly customisable, and can be zoomed in (down to 1 bp) or out (up to 1 Mb). To change the zoom you can use the zoom slide, type in a new location or gene, or click and drag a box around a region with your mouse. A number of data tracks (sets of data that can be plotted against the genome) are shown, and can be turned on and off using 'configure this page'. Find out more information about any track by clicking on the name.

Data tracks can be turned on using the 'Configure this page' tool at the left of the page.

---

### Adding data tracks

The region view can be altered by clicking on the blue 'Configure this page' button (Figure 24).

**Figure 24.** The blue 'Configure this page' button can be used to add or remove tracks to the region view.

Data tracks that can be drawn in the 'region in detail' view range from sequence variation from dbSNP and the 1000 Genomes project, through to regulatory features, to whole genome alignments. In the following example, a whole genome alignment between human and gorilla is selected using the 'configure this page' tool (Figure 25).

**Figure 25.** A whole genome alignment between human and gorilla is indicated by pink bars. Click on any bar to find out which region of the gorilla genome aligns to the BRCA2 locus.

The data tracks in the 'configure this page' menu are separated into various menus on the left (Figure 26). Sections include Genes and transcripts, mRNA and protein alignments, Variation, Regulation, and Comparative genomics.
Figure 26. The configure this page button opens a window of track options. Menus on the left allow you to explore tracks by category and indicate how many tracks are turned on in each category. Tracks that have been turned on can be seen in the active tracks menu above.

Test yourself! Can you configure data tracks in Region in Detail? Try this exercise [105] to find out.

More location information

More information is available in the location tab (Figure 27).
Some popular views are explained below:

- **Whole genome.** A view of the karyotype for the species, if one is available. You can upload your data on this view.
- **Chromosome summary.** Zoom in on one chromosome with the chromosome summary link.
- **Region overview.** The 'Region overview' view displays a region of 1 Mb or larger in a customisable view. View splice variants, syntenic blocks, markers, sequence variations, cDNA alignments, CpG islands, and much more along the genome. You can upload your data on this view.
- **Comparative genomics.** These comparative genomics displays provide visualisation of whole genome alignments, text-based views of sequence alignments, and graphics-based views of whole chromosome alignments and synteny across species.
- **Genetic variation.** Compare sequence across different individuals, strains, or breeds in the 'Resequencing' view. Or, visualise linkage disequilibrium in 'Linkage data'.
- **Other genome browsers.** Ensembl provides links out to the UCSC Genome Browser, NCBI Map Viewer and (in human) to our own GRCh37 site. We strive to use the most recent assembly, and jumping between the browsers should show the same region with the same coordinates, provided that updates are synchronised among the different browsers. There may be a short lag time where some browsers are on the old assembly, however usually there is no problem with directly comparing genomic coordinates.
The variation tab provides a wealth of information about a **SNP** [82], insertion, deletion, copy number **variant** [27], or somatic mutation.

We will explore:

- The genomic sequence in the region of a variant
- Genes and transcripts associated with a SNP of interest
- Population frequencies
- Associated diseases and phenotypes

For more about variation, including sources of information, see ‘Exploring the sources of biological data’ [6]? For more extensive background, have a look at the variation **documentation** [72].

You can get to the variant tab by clicking on variants in the **gene variant table** [112], any of the sequence views for the **gene** [12] or **transcript** [113], or from the **region in detail** [9] view. You can also **search directly** [8] for variants using rsIDs, **COSMIC** [114] or phenotypes.

The **Variant Effect Predictor** [115] is a popular tool that allows you to upload your own variations and calculate any effect on transcripts and proteins in Ensembl. See the **tools** [33] section for more.

**Finding the variation tab**

Finding the variation tab

**Access from the Ensembl** [5] home page:

- Search for a **variant** [27] (for example, rs1333049) using the search box on the Ensembl [116] home page (Figure 29). The search results will take you to the variation tab. (see ‘How to search Ensembl’ [8] for more information).
Expand 'Variation'. Click on Human.

rs1333049
Description
A dbSNP Variation. Phenotype(s): Coron 8; Coronary Artery Disease; Multiple comp. Diabetes Mellitus, Hypertension, Coronar Gene Association(s): CDKN2B, CDKN2A

Variation ID
rs1333049
Location
9:22125503 (view in location tab)
Source
e68

Figure 29. The search results will show a quick summary of information about the variant. To open the variation tab for rs1333049, just click on the variation ID.

Access from other views:

- Click on an ambiguity code [117] from the transcript cDNA sequence view (Figure 30). (See 'More transcript information [118])

Figure 30. Click on an ambiguity or IUPAC [119] code above a highlighted nucleotide in the cDNA view.

- Draw the variations on the sequence [12] view, and click on any of the variation IDs
- Click on a variation ID from the gene variation table [112]
- Turn on the variation track in the location tab [120], and click on a variant of interest.
Variation summary

When you click on a **variant** ID (for example, rs1333049) from **Ensembl** views or a search, the variant tab will open up (Figure 28).

![Figure 28](image)

**Figure 28.** Information about rs1333049 is accessible through links in the left-hand menu, and through icons. Click on any link or icon to go to information such as population genetics, [linkage disequilibrium](#) [109], or phenotypes and diseases associated with this **SNP** [82].

**Click on the icons** to find information about this SNP. Specifically:

- **Genomic context.** View the variant graphically on the genome
- **Genes and regulation.** Find out which genes and [regulatory features](#) [73] might be affected by this variant
- **Flanking sequence.** View the sequence upstream and downstream of the variant
- **Population genetics.** View the [allele](#) [83] frequencies across populations
- **Phenotype** [68] data. See diseases and phenotypes associated with the variant
- **Sample genotypes** [66]. Find out which [alleles](#) [83] are carried by individuals
- **Linkage disequilibrium** [109]. View LD plots and export values
- **Phylogenetic context.** Compare the nucleotide across species
- **Citations.** Find [open access](#) [21] papers that mention this variant
Figure 29. The top panel of the variant tab provides a summary of information about this variant (in this example, a SNP).

The top panel shown in Figure 29 is shown in all variant tab views. It provides basic information about the variant.

- **Variant class.** The type (SNP), variant ID (rs1333049) and link to dbSNP [103] (if a record exists) are provided.
- **Alleles.** The possible alleles are shown. The reference allele (the nucleotide in the reference sequence [121]) is listed first. In this case the reference allele is ‘G’. Note this may not be the major allele.
- **Synonyms.** Any other IDs associated with this variant are shown.
- **Validation status [122].** This suggests the confidence level of the existence of the polymorphism [123].
- **HGVS name.** A specific name from the HGVS project.

Phenotype and disease

The ‘phenotype’ [68] data view shows associations to diseases and phenotypes for variants.

Figure 30. The first three rows of the phenotype data table for rs1333049

In Figure 30, the variant rs1333049 is shown to be associated with Coronary Artery Disease by the NHGRI [124]-EBI GWAS [69] catalog. Click on the Europe PMC [125] link under ‘Study’ in order to see a publication demonstrating the association.
Did you know? You can search Ensembl [5] with a disease or phenotype to find associated genes and sequence variants.
Try it! Can you find a SNP [82] phenotype? Go to this exercise [126] to find out.

Population genetics

Allele [83] and genotype [66] frequencies are shown across different populations in the population genetics view (Figure 31).

Figure 31. Allele frequencies from the 1000 Genomes Project [61] for rs1333049.

Allele frequencies in different populations are shown graphically, and in tabular format. Pie charts show the allele frequency in different populations. This information is repeated in the table below, along with a specific count of the number of alleles [83] (allele count) and the genotype frequency codes. The three-letter population codes are populations analysed by the 1000 Genomes project. Hover over them to find out what they mean.

For more information about the view, see the population genetics page help [127].

Investigating gene regulation

The regulation tab allows an in-depth look at regulatory features [79] based on ENCODE [128] data (for human and mouse). For more about the regulatory features, see 'Exploring sources of biological data' [129].
Regulatory features are visible in the region in detail view. They can also be accessed by

We will explore:

- Find regulatory features in a location.
- Exploring a regulatory feature
  - Activity by cell type
  - Evidence for regulatory activity

**Regulation in location views**

Regulatory features are shown by default in the 'region in detail' view. A legend indicates what the colours mean.

![Regulatory features](image)

**Figure 32.** The regulatory features in the region view

You can click on a regulatory feature to navigate to the regulation tab.

Multiple data tracks relating to gene regulation can be turned on by configuring the page. You can add feature activity in different cell types, as well as individual transcription factor binding sites and histone modifications using a specially designed matrix.

**Investigating a regulatory feature**

To access the regulation tab (Figure 34), click on any ENSR ID on the Ensembl website.
Figure 34. The regulation tab has links to view details by cell type, the context of the feature in the genomic region and a table of evidence.

For detailed information about the data used in determining the regulatory features [73], go to the Ensembl regulation documentation [79].

Activity by cell type

The summary page in the regulation tab indicates which cell types the feature is active, inactive, repressed or poised in. This is summarised in a graphic (where you can select which cells you wish to view) and a table (Figure 35).
Figure 35. The activity of a regulatory feature by cell type.

Evidence

The 'Details by cell type' page in the regulation tab allows you to view the evidence for regulatory feature activity in different cell types (Figure 37). You can choose which cell types you wish to see and what evidence you’re interested in using the blue buttons above the graphic. In Figure 37, we can see all the available evidence for DND-41, GM12878, HepG2 and Placenta cells. The regulatory features [73] are still visible, showing you that the feature is inactive in DND-41, active in GM12878, repressed in HepG2 and poised in Placenta cells.
Figure 37. Evidence for regulatory feature activity shown for DND-41, GM12878, HepG2 and Placenta cells.

**Customise Ensembl**

On each of the tab pages you will find options that will enable you to configure the data presented on the screen, manage your own data, and export sequence and alignments (Figure 37). Here we discuss configuring a page. We'll investigate the other options in following pages.
Figure 37. Buttons to configure the page, add or manage custom tracks, export data and share or bookmark the page.

Configure this page

Available for all tabs (location, gene, transcript, variant [27] and regulation). The configuration panel enables you to control data types displayed in the view (Figure 38).

Figure 38. The configure menu for the gene summary view.

The configuration panel on the gene summary view is shown above. Note, the selection on the gene summary view is a subset of what is available on region in detail [120].

The data is divided into menus, highlighted in blue at the left of the image. Select data tracks from the menu, or use
the 'Find a track' option in the upper right-hand corner. Close the menu (click on the tic in the upper right-hand corner) to visualise new tracks on your display.

See more in the module [120] about adding data tracks. Check out or FAQ [131] on track styles.

**Manage your data**

**Figure 39.** Button to add or manage your own data.

Ensembl [5] allows users to upload or attach data in a variety of formats: BAM [132], BED [133], BedGraph, BigWig, GBrowse, GFF [134], GTF, PSL, VCF and WIG.

The 'Custom tracks' button (Figure 39) is available for most tabs. This opens a menu to attach data (Figures 40 and 41).

**Figure 40:** Data can be attached or uploaded using the options at the left of the Custom Data window.
Figure 41: Results of the attachment of a BAM file. Individual reads (grey) and the consensus sequence (coloured) are shown. Positions that differ from the reference genome are shown in red.

Upload data. You can upload your own data on the gene summary page, or the karyotype or a region on location pages. Various file formats are accepted. More information can be found on this help page [135].

Export data

Figure 42. The Export data button.

The gene, transcript and location tabs allow you to download (export) sequences. Use the Export data button (Figure 42) at the left of Ensembl [5] views.
Export data. The 'Export data' button on the gene, location and transcript tabs opens a view that allows you to export sequence as **FASTA** [136], allowing options for the sequence such as cDNA, coding sequence, and genomic sequence (Figure 43).

If you export protein from the gene tab, you will get all isoforms. If you export protein from the transcript tab, the information will be for one transcript (isoform) only.

The Export data button from genomic alignments pages (such as Gene tab, Genomic alignments view and the Orthologues view [137]) will export sequence alignments viewed in the browser. Formats include **CLUSTAL** [138], FASTA, Mega, MSF, Nexus and more. You can click on the Export data button when viewing alignments to export those alignments.

**Download data with BioMart**

Much of **Ensembl** [5]'s data can be quickly exported in text format, as an Excel table, or as **FASTA** [136] sequences through the BioMart interface [139].

BioMart allows you to:

- ‘Translate’ one ID type into another (for example, an Ensembl gene ID to an NCBI RefSeqID;
- Export data in different formats, including: html [140], csv, tsv [141] and xls (Excel) file types;
- Quickly generate tables of information.

**Figure 44.** Access BioMart from the top of Ensembl pages to export genes, variations and sequences.

You can use the 'Biomart' link at the top of every Ensembl page to access BioMart (Figure 44).

**Database.** You can choose the database you are interested in. In our example, we have selected the Ensembl genes database, version 89.

**Dataset.** You can choose the dataset you want to access. In our example we chose the Human data set.

**Options.** Apply filters on the data set in order to select a set of genes of interest. Attributes allow you to select output options. Additional datasets can be combined in your query as an advanced function.

This [tutorial] [142] is a good guide for getting started.

You can also view the Ensembl [FAQs page] [143].

**Get whole genomes with FTP**

Download the genome sequence for an organism, all the cDNA, genes, proteins, or ncRNAs for a species, and more with the [ftp site] [144] (Figure 45). You can get the whole mouse genome sequence, all the proteins in the human genome, or the genes for zebrafish, for example.
Figure 45. The ftp site allows sequence download for Ensembl species.

You can also download GenBank files, gene sets in GTF formats, or the MySQL [145] tables themselves.

Ensembl tools

Figure 46. Click on Tools in the blue bar at the top of Ensembl

Click the tools [147] button at the top of the browser to reveal some useful programs.

Each tool has an online interface and an API [24] script to be used in conjunction with the Ensembl Perl [148] API.

The Variant Effect Predictor [149] is our most popular tool. Enter in transcript or genomic coordinates to determine the effect of sequence variation on transcripts and proteins. A dbSNP [103] identifier [35] will be given in the output, if there is a matching one.

The Assembly Converter [150] allows coordinates from an older genome sequence to be updated to new coordinates (and vice-versa). As genomes are sequenced, the improved technology allows current genome sequence to be more accurate, containing fewer gaps and fewer mistakes. Using the most recent genome version or assembly is advised. Ensembl, the UCSC genome browser [110], and NCBI Map Viewer [111] strive to show all annotation [17] on the newest assembly possible, once the genome sequence is released to the public.
ID History converter [151] displays IDs that are in the current version of Ensembl. Start with a list of old IDs, and see which ones are still used, and which ones have been ‘retired’, or changed into a different ID. Though Ensembl IDs are stable (a gene or transcript should always have the same ID), the ID can change if one gene is split into two, or two genes that were erroneously split in a previous release are fused together into one.

File Chameleon [152] allows customised download of genome-wide files for use with NGS analysis tools.

Archive sites [25] let you view past versions of Ensembl, including older gene sets.

Summary

- Why Ensembl? [153] Ensembl aims to organise the vast amount of information in biological databases, and provide one point of access.

- What is Ensembl? [154] Ensembl provides a browser and underlying databases for annotation [17] of mostly vertebrate genomes. Ensembl genes are based on mainly cDNA and protein evidence in scientific databases. The Ensembl annotation pipeline is combined with Havana [47] manual annotation [44] for our major species (human, mouse, zebrafish, pig and rat). Human and mouse gene sets also include the CCDS [155] set. Comparative analyses, sequence variation and regulatory features [73] are all included in Ensembl.

- Where do the data come from? [6] The gene sets come from analyses described above. Whole genome alignments and homology are provided by comparative genomics analyses by the Ensembl team. Variants from dbsNP [103], COSMIC [114] and other projects are included. Finally, the regulatory features result from Ensembl analyses of ENCODE [128], Roadmap Epigenomics and Blueprint data.

- How do you access Ensembl? [156] You can access the Ensembl browser through the EBI homepage [22] or by going directly to Ensembl [31]. The database is directly accessible through MySQL [145] queries, or using our Perl API [54].

- How can you search Ensembl? [8] You can search the Ensembl browser with gene names, symbols, IDs, genomic regions, variations, descriptions and even diseases or phenotypes.

- How is the data presented in Ensembl? [7] Ensembl presents data in different ways. The information in Ensembl is split into separate tabs. The gene, transcript, location, variation and regulation tabs centre the information on a particular gene, transcript, location, and so on. Each tab has specific displays that you can access from the left-hand menus.

- How do I explore genomes with Ensembl? [14] You can explore a region on a genome, a gene or a splice variant [27] using the location, gene and transcript tabs. The left-hand menu on each tab leads you to more specific views.

- How do I download data? [157] Ensembl data (sequences and alignments) can be exported using the ‘Export data’ [157] link on the gene, location and transcript tab, BioMart [15], or the ftp site [158].
Guided examples of using Ensembl

The examples allow you to revise your knowledge gained on this course by providing guided examples of how Ensembl [5] can be used. The following scenarios demonstrate some uses of the Ensembl browser:

- Finding the cDNA sequence;
- Finding a gene.

Finding cDNA sequence for a gene

Background

The human OSM (oncostatin M) gene is involved in regulation of cell growth. It regulates cytokine production, and is involved in the maturation of foetal hepatocytes. The OSM protein stimulates proliferation of Kaposi's sarcoma cells in HIV-infected individuals.

Scenario

Imagine that you are working with the OSM gene, studying effects of mutation on the protein sequence. You would like to get the cDNA sequence, to make targeted mutations in the DNA that will code for alternative amino acids. You can use Ensembl [5] to search for the OSM gene and find answers to these questions.

Step 1 - Search

1. To search Ensembl [5], choose 'human' from the roll-down menu, and type 'OSM' into the search
box (Figure 47). (Alternatively use the search box at the top right of every Ensembl page).
2. Click 'Go'.
3. Click through to the gene tab. For a reminder of how to do this see the 'How to search Ensembl' [8] section.

**Step 2 - Choose a transcript**

After searching for the gene, and clicking through the results, you should see the OSM gene tab shown below.

![Gene tab for human OSM](image)

Figure 48. The gene tab for human OSM.

Three protein-coding splice variants are shown in the transcript table, each with a unique ENST ID. They are also shown in the image below the table.

How do you choose a transcript? Here are some criteria:

- **Ensembl [5]/Havana [47] merge**: ENST00000215781.2 is golden, which means it was annotated identically by Havana manual annotation [44] and Ensembl automatic annotation [159].
- **CCDS [155]**: Both ENST00000215781.2 and ENST00000403389.1, which means that Ensembl and RefSeq [49] agree on its coding region.
- **TSL**: ENST00000215781.2 is TSL1, meaning that it is fully supported by non-suspect mRNA data, whereas both ENST00000403389.1 and ENST00000403463.1 are TSL3, meaning that they are only supported by a single EST.
- **APPRIS**: ENST00000215781.2 is annotated as APPRIS P3, a candidate for the principal isoform of this gene, whereas ENST00000403389.1 is APPRIS ALT2, a possible principal isoform that is only conserved in three species.

Following these criteria ENST00000215781.2 is the best annotated of the transcripts and is likely to be the most relevant for our studies. Click on 'ENST00000215781.2' to go to the transcript tab.

**Step 3 - Access the cDNA sequence**
Figure 49. The cDNA view shows the cDNA, **CDS** [93] and protein sequences.

1. Click on the 'cDNA' link in the left hand menu to open the cDNA sequence view (Figure 49)
2. The **UTR** [89] is shown in the image, highlighted in yellow. Following that, the coding sequence is shown along with the protein sequence.
3. To export the sequence, click on 'Download sequence' above the legend. Use the dialogue box that opens to select the sequence you want to export.

**Did you know?** You could also look at the **Exons view** [160] in the transcript to see the **exon** [92] sequences... but not the protein.

**Using a sequence to find a gene (BLAST/BLAT)**

**Background**

If you do not have a gene name, an ID, or an **accession** [37] number for your sequence of interest, **Ensembl** [5] provides an interface that allows you to use **BLAST** [41] or **BLAT** [42] to align your sequence to the genome. You...
can then find out whether there is an Ensembl gene in that area.

Scenario

Imagine that you have sequenced a human gene that is associated with cancer. You know part of the DNA sequence, and that mutations in the sequence are associated with colon cancer, but you do not know which gene it belongs to. You can use Ensembl BLAST/BLAT to see if any genes correspond to your sequence.

CTCCGCACTGCTCACTCCCAGCGACGTAGGTGGTGACACGCCCGCTCTGT
GGCTCGCTTGGTTCCCTTAGTCCCGAGCGCTCGCCCACTGCAGATTCCTTT
CCCGTGCAGACATGGCCT

Step 1 - Use BLAST/BLAT

Use this sequence to enter the browser.

CTCCGCACTGCTCACTCCCAGCGACGTAGGTGGTGACACGCCCGCTCTGT
GGCTCGCTTGGTTCCCTTAGTCCCGAGCGCTCGCCCACTGCAGATTCCTTT
CCCGTGCAGACATGGCCT
Figure 50. Paste the sequence into the BLAST [41] interface.

2. Click on BLAST/BLAT [42] at the top of the page (Figure 50).
3. Enter the sequence at the top of the BLAST/BLAT [43] view.
4. Make sure your species of interest is chosen (Homo sapiens).
5. Click RUN. The default program is BLASTN, against Genomic sequence.

Step 2 - View the results

BLAST [41] shows us a full-length hit to chromosome nine and two shorter hits to chromosome three.
Figure 51. The BLAST results, shown as a table and mapped to the karyotype.

The sequence matches to chromosome 9, base pairs 21,802,636 to 21,802,755. The E-value [161] is near zero, and the %ID is 100, so this is a good hit.

Click on the locus link (9:21802636-21802755) from the result table to view the locus, and explore known features in that area.

Step 3 - Viewing the hit

The BLAST [41]/BLAT [42] track (the dark red block with sequence in Figure 52) shows us where our query sequence matches the genome, and allows us to compare the hit to any known genes in the region.
Figure 52. The BLAST hit on a genomic regions

The link from the BLAST result table has taken us to the Location tab, region in detail view. For more about this view, see the section [A region in detail](#) in this course.

A red box shows the position of the BLAT/BLAST hit. It matches the 5’ exon [92] of the MTAP gene for numerous coding and non-coding transcripts.

The sequence you are working with is part of the 5’ end of the MTAP gene, according to Ensembl!

Exercises

The following exercises allow you to apply your knowledge gained on this course. You will be asked how [Ensembl](#) [5] can be used to solve the given tasks. Start by clicking on one of the exercises, for example: [Finding the sequence and region of a gene](#) [163].

For hints, click on 'Need some Help'? Check your answer by viewing 'How we did it'.
Finding the sequence and region of a gene

Scenario
The human ‘furry homolog (Drosophila)’ gene (or FRY) is involved in cell morphogenesis, regulating the actin cytoskeleton, and patterning sensory neuron dendritic fields

Exercise
You want to design primers to clone the human FRY gene, and understand a little bit about the genomic region. Try the following tasks with Ensembl [5]:

- View and export the sequence, paying attention to exon [92] boundaries.
- Find the neighbouring genes to FRY.

Need some help?
There are multiple ways to do this. Some tips follow ...

1. Search for the human FRY gene, and go to the sequence link from the gene tab.
2. Have a look at the location tab to see which genes are in the neighbourhood.

How we did it
To find the sequence of the human FRY gene, follow the steps below:

1. Go to Ensembl [31].
2. Choose ‘Human’ and search for ‘FRY’ in the search box.
3. Click on ‘FRY (Human Gene)’. You should now be in the gene summary page for FRY (Figure 53).
4. Click on ‘Sequence’ at the left hand side. You should now see the sequence (Figure 54).

![Figure 54: Red highlight shows where the exon (92) is. You can use this information to design the primers.](image)

There are alternative ways to get the sequence! You could have used the Export data button, or BioMart.

Now let's find out which genes neighbour *FRY*.

5. Click on the location tab to see *FRY* and its surrounding region (Figure 55).
Figure 55: The neighbouring genes are a non-coding (blue) gene called FRY-AS1, and a small protein-coding (gold) gene called ZAR1L. If it's difficult to see which name matches which gene, click on any coloured bar for more information.

Finding a phenotype for a SNP

Scenario

You are working with the sequence variation rs2068824. This is a Single Nucleotide Polymorphism (SNP) located in the MMEL1 gene in human. The gene codes for a metalloprotease that cleaves polypeptides preferentially between hydrophobic residues.

Exercise

This sequence variation has turned up in several of your samples in patients with digestive troubles. Use Ensembl to find out if any phenotypes or diseases are known to be associated with this SNP.

Need some help?
1. Try searching Ensembl [5] with the variant [27].
2. Click through the search results until you reach the variation tab.
3. Find information on the phenotype [68].

## How we did it

1. Go to the Ensembl homepage [116]
2. Enter 'rs2068824' in the search box.
3. Click through to the variation tab (see the section 'How to search Ensembl' for more information).
4. Click on the phenotype [68] data link at the left of the variation tab.

In release 89, the only phenotype associated to this variation was Coeliac disease (Figure 56).

![Figure 56. The phenotype table for rs2068824.](image)

The association came from the NHGRI [124] GWAS [69] catalog, and the Pubmed [164] ID for the publication was 23936387.

## Finding markers for a mouse gene

### Scenario

The catalase (Cat) gene is present in almost all aerobically respiring organisms, and protects cells from the toxic effects of hydrogen peroxide. It can also promote cell growth.

### Exercise

You are working with a mouse model, and want to clone the Cat gene in mouse. Are there any markers from UniSTS in this locus (you can use them to help clone the gene)? Do any probesets from the Illumina [165] Mouse WG 6 V1 Microarray [166] platform match up to any of the Catalase exons, so you can study some expression patterns?

### Need some help?
1. Try to find the location tab, region in detail [167] page in Ensembl. If you search for *Cat* in the Ensembl mouse home page, you can click the location from the search hit.
2. Use the 'Configure this page' tool button at the left of the location tab to add markers and the *Illumina* [165] Mouse WG 6 V1 platform to the view.

How we did it

1. Start at the [Ensembl homepage][116]. Choose 'Mouse' as the species from the drop down menu, and search for 'Cat'.
2. Click through to the location tab.
3. From the 'Region in detail' view you can see there are two *Cat* transcripts, and both are protein coding (Figure 57).

![Figure 57. The region in detail view, showing two protein coding transcripts in the *Cat* locus.](image)

4. To turn on markers and probeset tracks, click 'Configure this page' at the left. Turn on 'Marker' and the probeset *Illumina* [165] Mouse WG 6 V1'.
5. Close the menu.

If you can't find the tracks you want in the Configure this page window, try using the search box.

You should now see the new tracks on Region in detail (Figure 58).
Figure 58. A marker and probe [168] are found in the 3’ exon [92] of the Cat-201 transcript.

Both the Marker track and the Illumina track show data at the end of the Cat transcripts. Remember, transcripts under the blue bar (contig track) are on the reverse strand of the genome, so the very left of the Cat transcripts is the 3’ end, in this case.

You can zoom into the tracks further using your mouse to draw a box around the region of interest.

To find out more information about the Marker or probe, click on it. The marker name is L25069. The probe name is ILMN_1254174.

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

Contact us!
Email our helpdesk [169] with any questions

Join our mailing lists, [170] Join 'dev' for discussion about the API [24] and local installations

Keep updated!

- Blog [171]
- Twitter [172]
- Facebook [173]

Confused?

- YouTube Videos. Watch the Ensembl videos on the Ensembl YouTube channel [174]
- Tutorials. Learn more about using Ensembl from the tutorials page [175]
- FAQs. Frequently asked questions [143] (and answers)
- Glossary. Definitions [176] of terms found in Ensembl
- Workshops. We run workshops worldwide! If you're interested in attending or organising one, view our workshops page [177].
- Courses held at the EBI. Ensembl takes part in several in-house courses [178] at the EBI. We go on the road with other EBI resources in external training [179] events.

Learn more! Read more about Ensembl data and analyses in publications [180].

Acknowledgements

Project funding

Current and previous sources of funding can be found on the Ensembl website [181].

Ensembl receives majority funding from the Wellcome Trust (grant number WT108749/Z/15/Z) with additional funding for specific project components from the National Human Genome Research Institute (U41HG007234 and U41HG007823), the Biotechnology and Biological Sciences Research Council (BB/L024225/1 and BB/M011615/1), Open Targets, the Wellcome Trust (WT104947/Z/14/Z, WT200990/Z/16/Z, and WT201535/Z/16/Z) and the European Molecular Biology Laboratory. This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement n° 634143 (MedBioinformatics). This project has
received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement n° 733161 (MultipleMS). We also receive funding from the ‘Save the Tasmanian Devil Program’.

Funding statement correct as of 11th July 2017.

Contributors

The Ensembl project

Course contributors

The original course was compiled by Giulietta Spudich in 2012. The course was updated in 2016 and 2017 by Emily Perry, Helen Sparrow and Ben Moore.
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.

Helen Sparrow [4]

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.

Giulietta M. Spudich [2]

EMBL-EBI
Outreach project leader, Ensembl

Giulietta was the outreach project leader for the Ensembl at EMBL’s European Bioinformatics Institute (EBI). The Ensembl project freely provides high quality annotation such as genes, sequence variation, and whole genome alignments across mainly vertebrate genomes. She leads a small team that organises and delivers training courses worldwide, and supports scientific communication about the project.

Before she started working with the Ensembl project in 2006, she obtained her PhD in Susan Marqusee’s lab at UC Berkeley in 2002. She ‘hopped the pond’ to carry out postdoctoral research in biochemical studies of Myosin VI at...
the MRC-LMB in Cambridge, UK (2002-2006).

Ben Moore [3]
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.

Funding and partners

We thank the contributors [181] that support the Ensembl project.

Source URL: https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes

Links
[1] https://www.ebi.ac.uk/training/online/trainers/emily
[2] https://www.ebi.ac.uk/training/online/trainers/giulietta
[3] https://www.ebi.ac.uk/training/online/trainers/bmoore
[4] https://www.ebi.ac.uk/training/online/trainers/hsparrow
[5] https://www.ebi.ac.uk/training/online/glossary/ensembl
[7] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/navigating-ensembl
[8] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/how-search-ensembl
[10] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/navigating-ensembl/investigating-sequence-variation
[16] https://www.ebi.ac.uk/training/online/glossary/ensembl-genomes
[17] https://www.ebi.ac.uk/training/online/glossary/annotation
Ensembl: Browsing genomes
Published on EMBL-EBI Train online (https://www.ebi.ac.uk/training/online)

[20] https://www.ebi.ac.uk/training/online/glossary/chordate
[21] https://www.ebi.ac.uk/training/online/glossary/open-access
[22] http://www.ebi.ac.uk/
[23] http://www.sanger.ac.uk/about/
[24] https://www.ebi.ac.uk/training/online/glossary/api
[26] https://www.ebi.ac.uk/training/online/glossary/ensembl-release
[27] https://www.ebi.ac.uk/training/online/glossary/variant
[28] https://www.ebi.ac.uk/training/online/course/european-nucleotide-archive-using-primary-nucleotides
[29] https://www.ebi.ac.uk/training/online/course/reactome-exploring-and-analysing-biological-pathways-1
[30] https://www.ebi.ac.uk/training/online/course/ensembl-genomes-non-chordates-quick-tour
[32] https://www.ebi.ac.uk/training/online/course/ensembl-filmed-api-workshop
[33] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/how-export-sequence-and-download-data/ensembl-tools
[34] http://www.ensembl.org/info/
[35] https://www.ebi.ac.uk/training/online/glossary/identifier
[36] https://www.ebi.ac.uk/training/online/glossary/uniprot
[37] https://www.ebi.ac.uk/training/online/glossary/accession
[38] https://www.ebi.ac.uk/training/online/glossary/pdb
[39] https://www.ebi.ac.uk/training/online/glossary/gene-ontology
[40] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/how-search-ensembl/searching-sequence-using-blast-or-blast
[41] https://www.ebi.ac.uk/training/online/glossary/blat
[42] https://www.ebi.ac.uk/training/online/glossary/blat
[44] https://www.ebi.ac.uk/training/online/glossary/manual-annotation
[45] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/navigating-ensembl/investigating-gene
[47] https://www.ebi.ac.uk/training/online/glossary/havana
[49] https://www.ebi.ac.uk/training/online/glossary/refseq
[51] http://www.ensembl.org/info/genome/genebuild/genome_annotation.html#Other
[56] http://www.ensembl.org/info/genome/compara/analyses.html
[57] https://www.ebi.ac.uk/training/online/glossary/synteny
[61] http://www.1000genomes.org/
[64] http://www.ebi.ac.uk/dgva/
[65] https://www.ebi.ac.uk/training/online/course/ensembl-variant-effect-predictor-vep-webinar
[66] https://www.ebi.ac.uk/training/online/glossary/genotype
[67] https://www.ebi.ac.uk/training/online/glossary/hapmap
[68] https://www.ebi.ac.uk/training/online/glossary/phenotype
[69] https://www.ebi.ac.uk/training/online/glossary/gwas
[70] https://www.ebi.ac.uk/training/online/glossary/omim
[71] https://www.ebi.ac.uk/training/online/glossary/ega
[73] https://www.ebi.ac.uk/training/online/glossary/regulatory-features
[122] https://www.ebi.ac.uk/training/online/glossary/validation-status
[123] https://www.ebi.ac.uk/training/online/glossary/polymorphism
[124] https://www.ebi.ac.uk/training/online/glossary/nhgri
[125] https://www.ebi.ac.uk/training/online/glossary/europe-pmc
[126] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/exercises/finding-phenotype-snp
[128] https://www.ebi.ac.uk/training/online/glossary/encode
[129] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/exploring-biological-data/gene-regulation
[132] https://www.ebi.ac.uk/training/online/glossary/bam
[133] https://www.ebi.ac.uk/training/online/glossary/bed
[134] https://www.ebi.ac.uk/training/online/glossary/gff
[136] https://www.ebi.ac.uk/training/online/glossary/fasta
[138] https://www.ebi.ac.uk/training/online/glossary/clustal
[140] https://www.ebi.ac.uk/training/online/glossary/html
[141] https://www.ebi.ac.uk/training/online/glossary/tsv
[143] http://www.ensembl.org/Help/Faq
[145] https://www.ebi.ac.uk/training/online/glossary/mysql
[148] https://www.ebi.ac.uk/training/online/glossary/perl
[153] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/why-ensembl
[154] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/what-ensembl
[155] https://www.ebi.ac.uk/training/online/glossary/ccds
[156] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/how-access-ensembl
[157] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/ensembl-tools/export-sequence
[158] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/get-whole-genomes-ftp
[159] https://www.ebi.ac.uk/training/online/glossary/automatic-annotation
[161] https://www.ebi.ac.uk/training/online/glossary/e-value
[162] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/navigating-ensembl/investigating-genomic-region/region-data
[163] https://www.ebi.ac.uk/training/online/course/ensembl-browsing-chordate-genomes/exercises/finding-sequence-and-region-gene
[164] https://www.ebi.ac.uk/training/online/glossary/pubmed
[165] https://www.ebi.ac.uk/training/online/glossary/illumina
[166] https://www.ebi.ac.uk/training/online/glossary/microarray
[167] http://www.ensembl.org/Mus_musculus/Location/View?db=core;g=ENSMUSG00000027187;r=2:10329406-103325317
[168] https://www.ebi.ac.uk/training/online/glossary/probe
[170] http://www.ensembl.org/info/about/contact/mailing.html
[172] https://twitter.com/ensembl
[174] http://www.youtube.com/user/EnsemblHelpdesk
[178] http://www.ebi.ac.uk/training/handson/
[179] http://www.ebi.ac.uk/training/roadshow?ebi-type=course