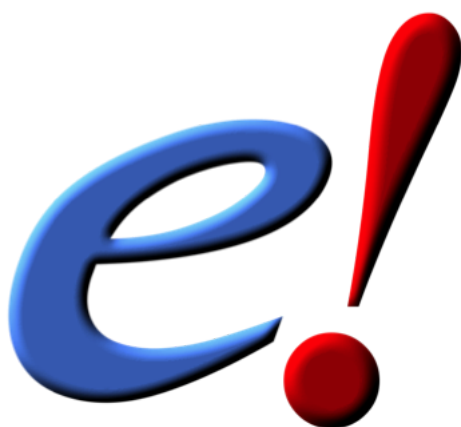


Browsing Genomes with Ensembl



www.ensembl.org

www.ensemblgenomes.org



Contents

- [Finding information about species and genomes in Ensembl, Demo](#)
- [Exploring genomic regions in Ensembl, Demo](#)
- [Human genes and transcripts in Ensembl, Demo](#)
- [Exploring variants in Ensembl, Demo](#)
- [Annotating genetic variants with the VEP, Demo](#)
- [Gene trees and homologues, Demo](#)
- [Whole genome alignments, Demo](#)
- [Features that regulate gene expression, Demo](#)
- [Bulk export of data with BioMart, Demo](#)
- [Custom data, Demo](#)

Finding information about species and genomes in Ensembl, Demo

Demo: Introduction to Ensembl

Ensembl

Homepage

The front page of Ensembl is found at ensembl.org. It contains lots of information and links to help you navigate Ensembl:

The screenshot shows the Ensembl homepage with several red handwritten annotations and arrows pointing to specific features:

- Link back to homepage**: Points to the Ensembl logo in the top left.
- Ensembl tools**: Points to the 'Tools' section in the top navigation bar.
- Blue bar visible on every page**: Points to the top navigation bar.
- Search**: Points to the search bar in the top right.
- Search**: Points to the search bar in the main content area.
- Drop-down list of all species**: Points to the 'All species' dropdown menu in the search bar.
- See the current release and what's new**: Points to the 'Ensembl Release 104 (May 2021)' section.
- See our newest genomes on Rapid Release**: Points to the 'Rapid Release' section.

The page content includes a top navigation bar with links like BLAST/BLAT, VEP, Tools, BioMart, Downloads, Help & Docs, and Blog. The main content area features a search bar, a list of tools, a section for 'All genomes' with a dropdown menu, and a 'Favourite genomes' section. The right-hand panel displays the current release information and news.

On the right-hand panel you can see the current release number and what has come out in this release. To access old releases, scroll to the bottom of the page and click on **View in archive site** in the right-hand corner.

View in archive site

Search

Help topics

Frequently Asked Questions

Video Tutorials

Glossary

Contact HelpDesk

The following archives are available for this page:

- Ensembl GRCh37: Full Feb 2014 archive with BLAST, VEP and BioMart
- Ensembl 110: Jul 2023 (GRCh38.p14) - patched/updated gene set Mar 2023
- Ensembl 109: Feb 2023 (GRCh38.p13) - patched/updated gene set Nov 2022
- Ensembl 108: Oct 2022 (GRCh38.p13) - patched/updated gene set Jul 2022
- Ensembl 107: Jul 2022 (GRCh38.p13) - patched/updated gene set Apr 2022
- Ensembl 106: Apr 2022 (GRCh38.p13) - patched/updated gene set Nov 2021
- Ensembl 105: Dec 2021 (GRCh38.p13) - patched/updated gene set Aug 2021
- Ensembl 104: May 2021 (GRCh38.p13) - patched/updated gene set Mar 2021
- Ensembl 103: Feb 2021 (GRCh38.p13) - patched/updated gene set Aug 2020
- Ensembl 102: Nov 2020 (GRCh38.p13) - patched/updated gene set Sep 2020
- Ensembl 101: Aug 2020 (GRCh38.p13) - patched/updated gene set Mar 2020
- Ensembl 100: Apr 2020 (GRCh38.p13) - patched/updated gene set Jun 2019
- Ensembl 99: Jan 2020 (GRCh38.p13) - patched/updated gene set Aug 2019
- Ensembl 98: Sep 2019 (GRCh38.p13) - patched/updated gene set Jun 2019
- Ensembl 97: Jul 2019 (GRCh38.p12) - patched/updated gene set Mar 2019
- Ensembl 96: Apr 2019 (GRCh38.p12) - patched/updated gene set Nov 2018
- Ensembl 95: Jan 2019 (GRCh38.p12) - patched/updated gene set Jul 2018
- Ensembl 80: May 2015 (GRCh38.p2) - patched/updated gene set Jan 2015
- Ensembl 77: Oct 2014 (GRCh38) - patched/updated gene set Aug 2014
- Ensembl 75: Feb 2014 (GRCh37.p13) - patched/updated gene set Sep 2013
- Ensembl 54: May 2009 (NCBI 36) - patched/updated gene set Oct 2008

More information about the Ensembl archives

Click on the links to go to the archives. Alternatively, you can jump quickly to the correct release by adding **e** plus the **release number** in the URL. For example e98.ensembl.org jumps to Ensembl release 98.

Available species

Scroll back up to the top of the homepage. You can view all available species by clicking the **View full list of all species** link underneath the coloured search block.

You can search for your species of interest (either the common or scientific name) using the search bar at the top right-hand corner of the table. Click on the common name of your species of interest to go to the species information page. We'll click on **Human**.

Human (GRCh38.p13) ▼

Search Human (Homo sapiens)

Search all categories

Search...

Go

e.g. BRCA2 or 17:83952802-84038237 or rs699 or osteoarthritis

Genome assembly: GRCh38.p13 (GCA_000001405.28)

More information and statistics

Download DNA sequence (FASTA)

Convert your data to GRCh38 coordinates

Display your data in Ensembl

Other assemblies

GRCh37 Full Feb 2014 archive with BLAST, VEP and BioMart

Gene annotation

What can I find? Protein-coding and non-coding genes, splice variants, cDNA and protein sequences, non-coding RNAs.

More about this genebuild

Download FASTA files for genes, cDNAs, ncRNA, proteins

Download GTF or GFF3 files for genes, cDNAs, ncRNA, proteins

Update your old Ensembl IDs

Example gene

Example transcript

Comparative genomics

What can I find? Homologues, gene trees, and whole genome alignments across multiple species.

More about comparative analysis

Download alignments (EMF)

Example gene tree

Example region

Regulation

What can I find? DNA methylation, transcription factor binding sites, histone modifications, and regulatory features such as enhancers and repressors, and microarray annotations.

More about the Ensembl regulatory build and microarray annotation

Experimental data sources

Download all regulatory features (GFF)

Example regulatory feature

ENCODE

Variation

What can I find? Short sequence variants and longer structural variants; disease and other phenotypes

More about variation in Ensembl

Download all variants (GVF)

Example variant

Example phenotype

Example structural variant

Species information

Here you can see links to example features and to download flatfiles. To find out more about the genome assembly and genebuild, click on **More information and statistics** under the **Genome assembly** section.

Human (GRCh38.p13)

Human assembly and gene annotation

Assembly

This site provides a data set based on the December 2013 Homo sapiens high coverage assembly GRCh38 from the Genome Reference Consortium. This assembly is derived from the hg38 database. The data set consists of gene models built from the genome alignments of the human proteome as well as from alignments of human cDNAs using the cDNA2genome model of exonerate.

This release of the assembly has the following properties:

- contig length total 3.4 Gb
- chromosome length total 3.1 Gb (excluding haplotypes).

It also includes 261 alt loci scaffolds, mainly in the LINC-ROR complex on chromosome 19 (35 alternate sequence representations) and the MHC region on chromosome 6 (10 alternate sequence representations).

[Watch a video on YouTube](#) about patches and haplotypes in the Human genome.

Patches

As the GRC maintains and improves the assembly, patches are being introduced. Currently, assembly patches are of two types:

Statistics

Summary	
Assembly	GRCh38.p13 (Genome Reference Consortium Human Build 38), INSDC Assembly GCA_000001405.28.6; UCSC hg38
Base Pairs	3,096,649,726
Base Path Length	3,096,649,726
Assembly provider	Genome Reference Consortium
Annotation provider	Ensembl
Annotation method	Full genebuild
Genebuild released	Jan 2014
Genebuild last updated/patched	Mar 2021
Database version	104.38
Gencode version	GENCODE v41

Here you'll find a detailed description of how the genome was produced and links to the original source. You will also see details of how the genes were annotated.

The current genome assembly for human is GRCh38. If you want to see the previous assembly, GRCh37, visit our dedicated site grch37.ensembl.org.

GRCh37 BLAST/BLAT | VEP | Tools | BioMart | Downloads | Help & Docs

Search all species...

Search

All species for

Go

e.g. BRCA2 or human 5:52797363-63627669 or rs999 or coronary heart disease

Browse a Genome

The Ensembl project produces genome databases for vertebrates and other eukaryotic species, and makes this information freely available online.

Available genomes

Human GRCh37.p13

Want to use GRCh38?

Our main site features the GRCh38 Homo sapiens assembly, with the latest gene models, variants, regulatory build and more!

[Learn more](#) about how to migrate your data to GRCh38

About this archive

This archive is based on Ensembl Release 75 data, and gives continuing access to human assembly GRCh37, Human variation and regulation data has since been updated in April 2021.

MySQL dumps of human databases on the most recent schema version are available on our [FTP site](#).

Ensembl GRCh37 Release 104 (May 2021)

- Updated regulatory build.
- Updated variation data including the latest data from dbSNP build 154, ClinVar and COSMIC.

[More release news](#) on our blog

Ensembl Genomes

Homepage

Let's take a look at the Ensembl Genomes homepage at ensemblgenomes.org.

EnsemblGenomes Providing genome data for non-vertebrate species, with tools for the manipulation, analysis and visualisation of that data

Search all genomes

Go

Ensembl COVID-19

SARS-CoV-2 Genome sequence & annotation

Go

Ensembl Rapid Release

2-weekly releases of new assemblies with gene & protein feature annotation

Go

EnsemblMetazoa

Caenorhabditis elegans WBcel235

Drosophila melanogaster BDGP6.28

Bombus morio ASM15162v1

Go to Ensembl Metazoa

Ensembl Plants

Triticum aestivum IWGSC

Oryza sativa Japonica Group IRGSP-1.0

Arabidopsis thaliana

Go to Ensembl Plants

EnsemblProtists

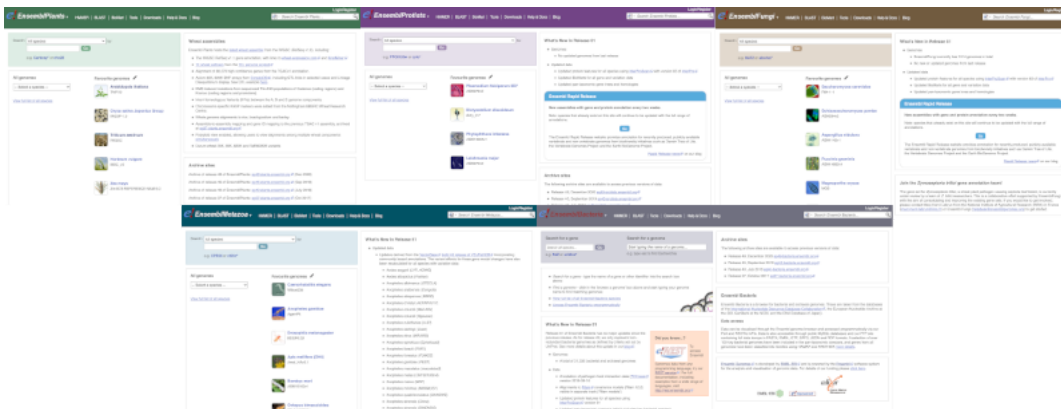
Plasmodium falciparum 3D7 ASM276v2

Dicostelium discoideum dicty_2.7

Phytophthora infestans ASM14294v1

Go to Ensembl Protists

Click on the different taxa to see their homepages. Each one has a different colour-coding, but they are all structured in a similar format to the Ensembl main site.



You can navigate most of the taxa in the same way as you would with Ensembl.

Ensembl Bacteria

Ensembl Bacteria has a large number of genomes and has a slightly different method to the other Ensembl sites. Let's look at it in more detail.

Search for a gene

Search all species... Go

e.g. *ftsZ* or *uridine*

Search for a genome

Start typing the name of a genome...

e.g. type *esc* to find *Escherichia*

Archive sites

The following archive sites are available to access previous versions of data:

- Release 49, December 2020 ftp49.bacteria.ensembl.org/
- Release 45, September 2019 ftp45.bacteria.ensembl.org/
- Release 40, July 2019 ftp40.bacteria.ensembl.org/
- Release 37, October 2018 ftp37.bacteria.ensembl.org/

Ensembl Bacteria

Ensembl Bacteria is a browser for bacterial and archaeal genomes. These are taken from the databases of the [International Nucleotide Sequence Database Collaboration](#), the European Nucleotide Archive at the EBI, GenBank at the NCBI and the DNA Database of Japan.

Data access

Data can be visualised through the Ensembl genome browser and accessed programmatically via our Perl and RESTful APIs. Data is also accessible through public MySQL databases and our FTP site containing full data (GenBank, EMBL, RefSeq, UniProt and Ensembl annotations). A selection of over 100 key bacterial genomes have been classified into families using HAMAP and PANTHER [more details](#).

Ensembl Genomes

Ensembl Genomes is a browser for eukaryotic genomes and is powered by the Ensembl software system for the analysis and visualisation of genomic data. For details of our funding please [click here](#).

Did you know...?

eREST To access Ensembl

Genomes data from any programming language, try our REST service. For full documentation, including examples from a wide range of languages, visit <http://rest.ensembl.org/>

What's New in Release 51

Release 51 of Ensembl Bacteria has no major updates since the previous release. As for release 49, we only represent non-redundant bacterial genomes as defined by criteria set out by UniProt. See more details about this update in our [blog](#).

- Genomes
 - A total of 31,332 bacterial and archaeal genomes
- Data
 - Annotation of pathogen-host interaction data (PHI-base) version 2019-09-16
 - Alignments to Rfam covariance models (Rfam 12.2) visible in separate track (Rfam models)
 - Updated protein features for all species using InterProScan version 5.1
 - Updated pan-taxonomic compara (which includes key bacterial species)

There's no drop-down species list for bacteria as it would be hard to navigate with the number of species. You can click the **View full list of all Ensembl Bacteria species** link underneath the coloured search block. Search for your species of interest using the filter in the top right-hand corner of the table.

Find a Species

Show 25 entries

Filter:

Species	Assembly	Taxonomy ID	Serotype	Publications	Present in pan-taxonomic compara
Clostridioides difficile 630 (GCA_000009205)	ASM920v2	272563			Y
Clostridioides difficile ATCC 9689 = DSM 1296 (GCA_001077535)	ASM107753v2	1121308			N
Clostridioides difficile CD160 (GCA_000449425)	ASM44942v2	1151292			N
Clostridioides difficile CD196 (GCA_000085225)	ASM8522v1	649462			N
Clostridioides difficile F501 (GCA_000450805)	ASM45080v2	1151372			N
Clostridioides difficile str. BA09_70 (GCA_001299495)	ASM129949v1	1496			N
Clostridioides difficile (GCA_002301955)	ASM230195v1	1496			N
Clostridioides difficile str. CD-B18-123 (GCA_005502205)	ASM550220v1	1496			N
Clostridioides difficile str. WC1065050 (GCA_005786645)	ASM578664v1	1496			N

Alternatively, you can find a species by typing the species name into the **Search for a genome** search box at the top of the page. A drop-down list will appear with any species matching the name you entered.

For example, to find a sub-strain of *Clostridioides difficile* start typing in the species name. Due to the auto-complete, you'll see useful results as soon as you get to *Clostridio*.

The screenshot shows the Ensembl search interface. On the left, there's a 'Search for a gene' section with a text input 'Search all species...' and a 'Go' button. Below it, an example 'ftsZ or uridine*' is shown. On the right, the 'Search for a genome' section has a text input with 'clostridio' entered. A dropdown menu is open, showing a list of *Clostridioides difficile* strains with their GCA IDs and TaxIDs. To the right of the search boxes is an 'Archive sites' section with the text 'The following archive sites are av...'. Below the search boxes, there are four links: 'Search for a gene - type the name of a gene or above', 'Find a genome - click in the 'browse a genome' name to find matching genomes', 'View full list of all Ensembl Bacteria species', and 'Access Ensembl Bacteria programmatically'. At the bottom right, there's a logo for the International Nucleotide Sequence Database Consortium (INSDC) and text mentioning the EBI GenBank at the NCBI.

The drop down contains various strains of *C. difficile*. Let's choose *C. difficile* 630. This will take us to another species information page, where we can explore various features.

The screenshot shows the Ensembl species information page for *Clostridioides difficile* 630 (GCA_000009205). The page has a dark header with the species name and ID. Below the header, there's a search bar with 'Search ...' and a 'Go' button. Below the search bar, there's a link to 'Provider EMBL Nucleotide Sequence Database | Taxonomy ID 272563'. Below that, there's a section 'About Clostridioides difficile 630 (GCA_000009205)' with an 'Information and statistics' link. The main content area is divided into four panels: 'Genome assembly: ASM920v2' with links for 'More information and statistics', 'Download DNA sequences (FASTA)', and 'Display your data in Ensembl Bacteria'; 'Gene annotation' with a 'What can I find?' section listing protein-coding genes, splice variants, cDNA, and protein sequences, and links for 'More about this genebuild', 'Download genes, cDNAs, ncRNA, proteins - FASTA - GFF3', and 'Update your old Ensembl IDs'; 'Comparative genomics' with a 'What can I find?' section listing gene families based on HAMAP and PANTHER classification, and links for 'More about comparative analyses' and 'Phylogenetic overview of gene families'; and 'Variation' with a 'This species currently has no variation database. However you can process your own variants using the Variant Effect Predictor' section and a 'Variant Effect Predictor' link. The page also features several icons and diagrams, including a karyotype, an example region, a gene family diagram, and a pan-taxonomic tree.

Unlike the *Homo sapiens* species information page, there is no prose description of the genome or gene annotation, as these pages were generated automatically.

Ensembl Rapid Release

Our newest genomes, such as those coming from the [Darwin Tree of Life](#), are available rapid.ensembl.org with limited annotation.

Tools

BLAST >

[All tools](#)

Search our genomes for your DNA or protein sequence

Search

Camarhynchus parvulus (Small tree finch) for

Go

e.g. Camarhynchus parvulus 2:361680-384534 or Clytia hemisphaerica IPR001650

Ensembl Rapid Release is a new site designed to make our data available more quickly. Release of data occurs on a two-week cycle, meaning we can make our gene sets available with minimal delay once the annotation is complete. For each species we provide a gene set along with additional features such as protein feature annotation and BLAST functionality.

It is important to note that Ensembl Rapid Release is by nature not as fully featured as a typical data release on www.ensembl.org. Currently we do not provide homology data, gene symbol assignment (which uses homology data), data archiving or programmatic access. We are however working on adding more functionality over the coming months to further improve usability.

[More details](#) about Ensembl Rapid Release and the current state.

Latest Genomes

We have 4 new genomes this release:

- [Cimex lectularius \(Bed bug\) - GCA_000648675.3 - Harlan](#)
- [Olive baboon](#)
- [Sumatran orangutan - GCA_90280775.3](#)
- [Salmo salar \(Atlantic salmon\) - GCA_905237065.2](#)

[View all species and download data](#)

Newest genomes

About Rapid Release

All genomes

-- Select a species --

[View and download available data for all species](#)

Highlighted genomes



Camarhynchus parvulus (Small tree finch) - GCA_902806625.1
Camarhynchus_parvulus_V1.1



Clytia hemisphaerica (Z4C2) - GCA_902728285.1
GCA902728285v1

Exploring genomic regions in Ensembl, Demo

Start at the Ensembl front page, ensembl.org. You can search for a region by typing it into a search box, but you have to specify the species.

To bypass the text search, you need to input your region coordinates in the correct format, which is **chromosome, colon, start coordinate, dash, end coordinate**, with no spaces for example: **human 4:122868000-122946000**. Type (or copy and paste) these coordinates into either search box.

Search

All species

▼

for

chicken 4:53544500-53598000

Go

e.g. [BRCA2](#) or [rat 5:62797383-63627669](#) or [rs699](#) or [coronary heart disease](#)

or



Chicken (bGalGal1.mat.broiler.GRCg7b) ▼

Search Chicken (Gallus gallus)

Search all categories

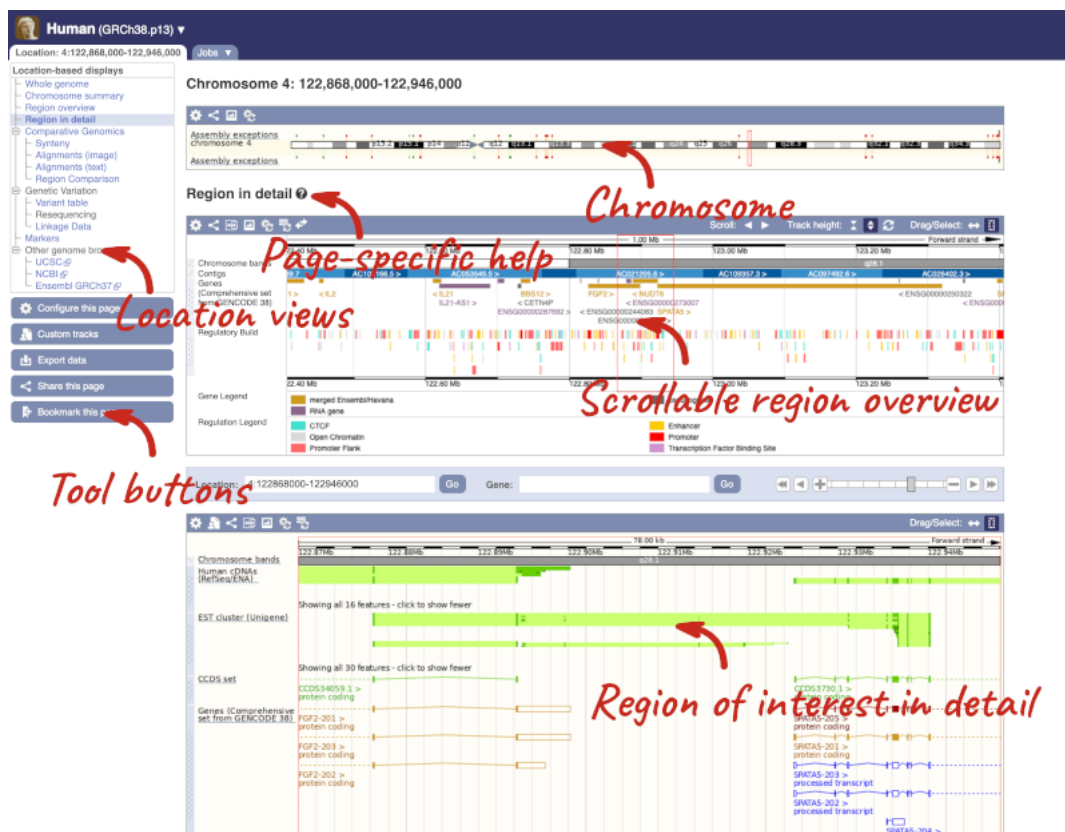
▼

4:53544500-53598000

Go

e.g. [ESRRB](#) or [5:38111022-38265293](#) or [rs13582785](#) or [sphingolipid](#)

Press *Enter* or click *Go* to jump directly to the Region in detail Page.

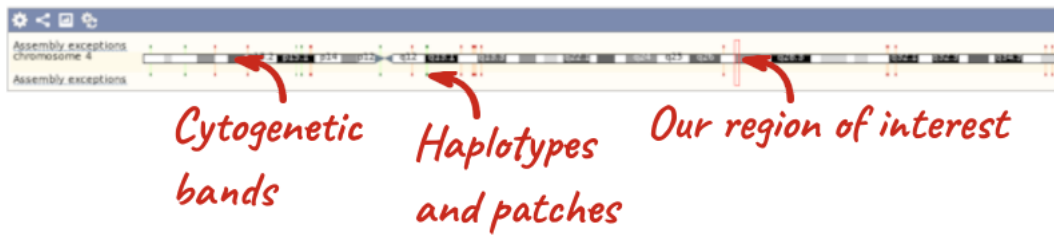


Click on the button to view page-specific help. The help pages provide text, labelled images and, in some cases, help videos to describe what you can see on the page and how to interact with it.

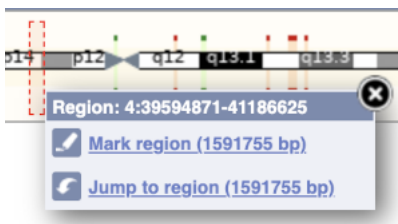
The Region in detail page is made up of three images, let's look at each one in detail.

The first image shows the chromosome:

Chromosome 4: 122,868,000-122,946,000



The region we're looking at is highlighted on the chromosome. You can jump to a different region by dragging out a box in this image. Drag out a box on the chromosome, a pop-up menu will appear.



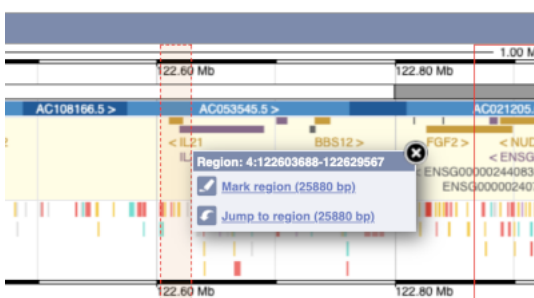
If you wanted to move to the region, you could click on *Jump to region (### bp)*. If you wanted to highlight it, click on *Mark region (###bp)*. For now, we'll close the pop-up by clicking on the X on the corner.

The second image shows a 1Mb region around our selected region. This is always 1Mb in human, but the fixed size of this view varies between species. This view allows you to scroll back and forth along the chromosome.

Region in detail



You can also drag out and jump to or mark a region.

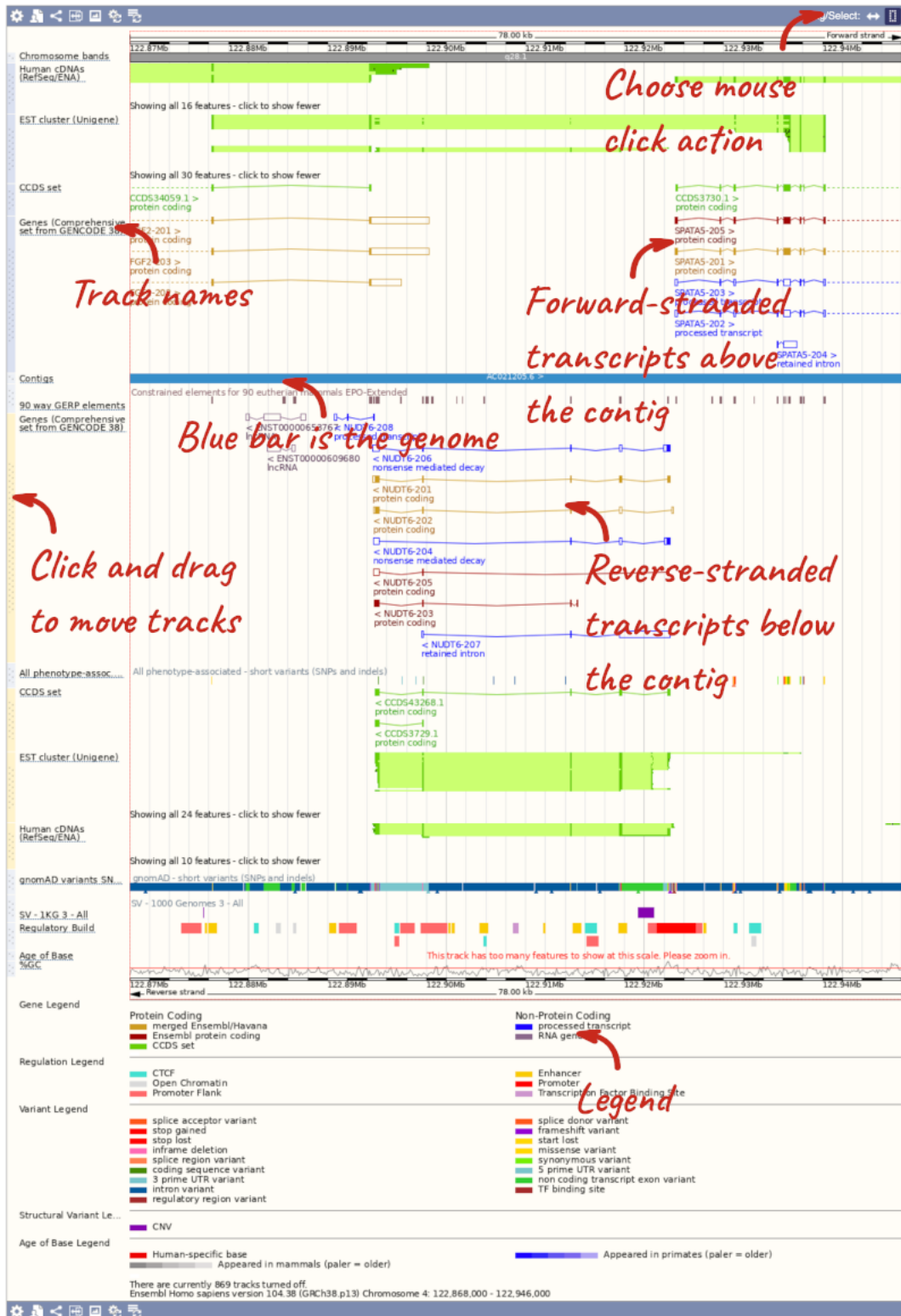


Click on the X to close the pop-up menu.

Click on the *Drag/Select button*  to change the action of your mouse click. Now you can scroll along the chromosome by clicking and dragging within the image. As you do this you'll see the image below grey out and two blue buttons appear. Clicking on *Update this image* would jump the lower image to the region central to the scrollable image. We want to go back to where we started, so we'll click on *Reset scrollable image*.



The third image is a detailed, configurable view of the region.

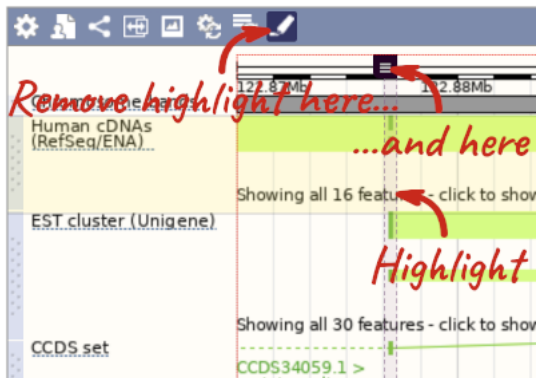


Here you can see various *tracks*, which is what we call a data type that you can plot against the genome. Some tracks, such as the transcripts, can be on the forward or reverse strand. Forward stranded features are shown above the blue contig track that runs across the middle of the image, with reverse stranded features

below the contig. Other tracks, such as variants, regulatory features or conserved regions, refer to both strands of the genome, and these are shown by default at the very top or very bottom of the view.

You can use click and drag to either navigate around the region or highlight regions of interest. Click on the *Drag/Select* option at the top or bottom right to switch mouse action. On *Drag*, you can click and drag left or right to move along the genome, the page will reload when you drop the mouse button. On *Select* you can drag out a box to highlight or zoom in on a region of interest.

With the tool set to *Select*, drag out a box around an exon and choose *Mark region*.

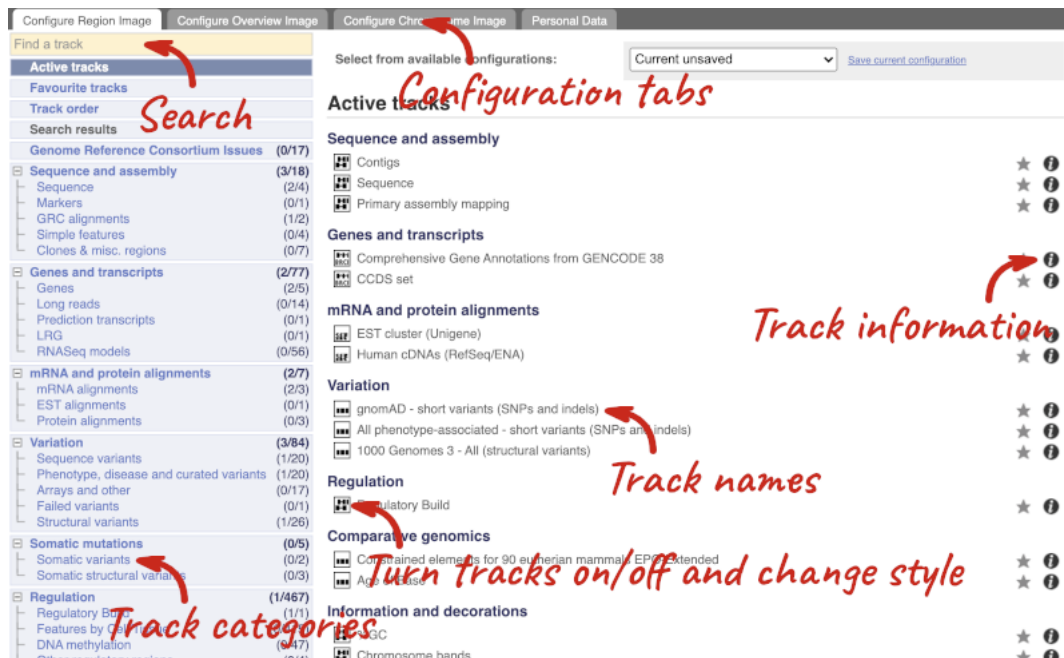


The highlight will remain in place if you zoom in and out or move around the region. This allows you to keep track of regions or features of interest.

We can edit what we see on this page by clicking on the blue *Configure this page* menu at the left.



This will open a menu that allows you to change the image.



There are thousands of possible tracks that you can add. When you launch the view, you will see all the tracks that are currently turned on with their names on the left and an info icon on the right, which you can click on to expand the description of the track. Turn them on or off, or change the track style by clicking on the box next to the name. More details about the different track styles are in this FAQ:

<http://www.ensembl.org/Help/Faq?id=335>.

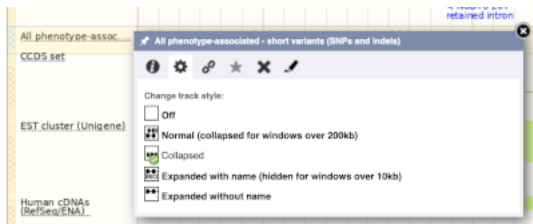
You can find more tracks to add by either exploring the categories on the left, or using the *Find a track* option at the top left. Type in a word or phrase to find tracks with it in the track name or description.

Let's add some tracks to this image. Add:

- Proteins (mammal) from UniProt – Labels
- 1000 Genomes - All - short variants (SNPs and indels) – Normal

Now click on the tick in the top left hand to save and close the menu. Alternatively, click anywhere outside of the menu. We can now see the tracks in the image. The proteins track is stranded, so you will see two tracks, one above and one below the contig, representing the proteins mapped to the forward and reverse strands respectively. The variants track is not stranded, so is found near the bottom of the image.

If the track is not giving you can information you need, you can easily change the way the tracks appear by hovering over the track name then the cog wheel to open a menu. To make it easier to compare information between tracks, such as spotting overlaps, you can move tracks around by clicking and dragging on the bar to the left of the track name.



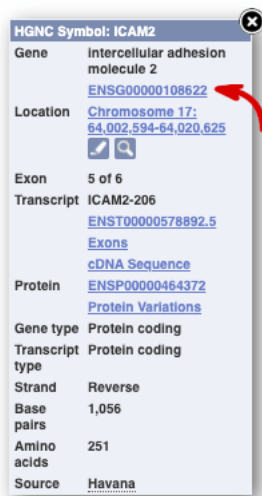
Now that you've got the view how you want it, you might like to show something you've found to a colleague or collaborator. Click on the *Share this page* button to generate a link. Email the link to someone else, so that they can see the same view as you, including all the tracks you've added. These links contain the Ensembl release number, so if a new release or even assembly comes out, your link will just take you to the archive site for the release it was made on.



To return this to the default view, go to *Configure this page* and select *Reset configuration* at the bottom of the menu.

Human genes and transcripts in Ensembl, Demo

You can find out lots of information about Ensembl genes and transcripts using the browser. If you're already looking at a region view, you can click on any transcript and a pop-up menu will appear, allowing you to jump directly to that gene or transcript.



HGNC Symbol: ICAM2

Gene: Intercellular adhesion molecule 2
[ENSG00000108622](#)

Location: Chromosome 17: 64,002,594-64,020,625

Exon: 5 of 6

Transcript: ICAM2-206
[ENST00000578892.5](#)
[Exons](#)
[cDNA Sequence](#)

Protein: [ENSP00000464372](#)
[Protein Variations](#)

Gene type: Protein coding
Transcript type: Protein coding

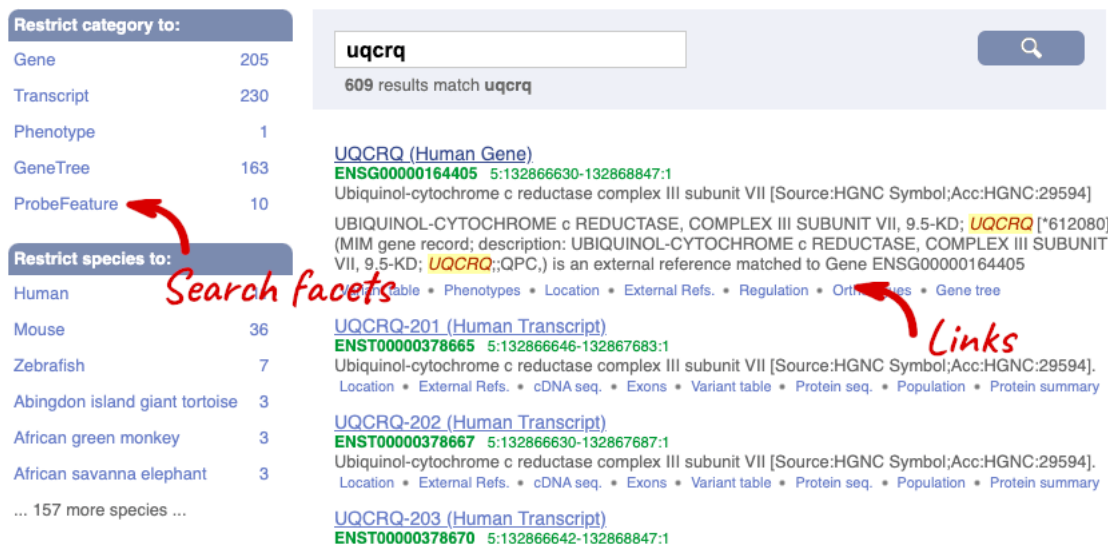
Strand: Reverse
Base pairs: 1,056
Amino acids: 251
Source: Havana

Link to gene

Alternatively, you can find a gene by searching for it. You can search for gene names or identifiers, and also phenotypes or functions that might be associated with the genes.

We're going to look at the human *UQCRCQ* gene. From ensembl.org, type *UQCRCQ* into the search bar and click the Go button. You will get a list of hits with the human gene at the top.

Where you search for something without specifying the species, or where the ID is not restricted to a single species, the most popular species will appear first, in this case, human, mouse and zebrafish appear first. You can restrict your query to species or features of interest using the options on the left.



Restrict category to:

Gene	205
Transcript	230
Phenotype	1
GeneTree	163
ProbeFeature	10

Restrict species to:

Human	609
Mouse	36
Zebrafish	7
Abingdon island giant tortoise	3
African green monkey	3
African savanna elephant	3
... 157 more species ...	

Search facets

uqcrcq 609 results match **uqcrcq**

UQCRCQ (Human Gene)
[ENSG00000164405](#) 5:132866630-132868847:1
Ubiquinol-cytochrome c reductase complex III subunit VII [Source:HGNC Symbol;Acc:HGNC:29594]
UBIQUINOL-CYTOCHROME c REDUCTASE, COMPLEX III SUBUNIT VII, 9.5-KD; **UQCRCQ** [*612080]
(MIM gene record; description: UBIQUINOL-CYTOCHROME c REDUCTASE, COMPLEX III SUBUNIT VII, 9.5-KD; **UQCRCQ**;QPC₁) is an external reference matched to Gene ENSG00000164405

UQCRCQ-201 (Human Transcript)
[ENST00000378665](#) 5:132866646-132867683:1
Ubiquinol-cytochrome c reductase complex III subunit VII [Source:HGNC Symbol;Acc:HGNC:29594].
Location * External Refs. * cDNA seq. * Exons * Variant table * Protein seq. * Population * Protein summary

UQCRCQ-202 (Human Transcript)
[ENST00000378667](#) 5:132866630-132867687:1
Ubiquinol-cytochrome c reductase complex III subunit VII [Source:HGNC Symbol;Acc:HGNC:29594].
Location * External Refs. * cDNA seq. * Exons * Variant table * Protein seq. * Population * Protein summary

UQCRCQ-203 (Human Transcript)
[ENST00000378670](#) 5:132866642-132868847:1

Links

The gene tab

Click on the gene name or Ensembl ID. The *Gene* tab should open:

Human (GRCh38.p13) ▼

Location: 5:132,866,630-132,868,847 **Gene: UQCRCQ** *Gene tab* Jobs ▼

Gene-based displays

- Summary
 - Splice variants
 - Transcript comparison
 - Gene alleles
- Sequence
 - Secondary Structure
- Comparative Genomics
 - Genomic alignments
 - Gene tree
 - Gene gain/loss tree
 - Orthologues
 - Paralogues
 - Ensembl protein families
- Ontologies
 - GO: Molecular function
 - GO: Cellular component
 - GO: Biological process
- Phenotypes
- Genetic Variation *Gene views*
 - Variant table
 - Variant image
 - Structural variants
- Gene expression
- Pathway
- Regulation
- External references
- Supporting evidence
- ID History
 - Gene history

Gene: UQCRCQ ENSG00000164405

Description ubiquinol-cytochrome c reductase complex III subunit VII [Source:HGNC Symbol;Acc:QCR8, QP-C, UQCR7]

Gene Synonyms QCR8, QP-C, UQCR7

Location [Chromosome 5: 132,866,630-132,868,847](#) forward strand.
GRCh38:CM000667.2

About this gene This gene has 6 transcripts ([splice variants](#)), [209 orthologues](#) and is associated with [3](#)

Transcripts [Show transcript table](#) *Option: open table of tra*

Summary ⓘ

Name [UQCRCQ](#) (HGNC Symbol)

CCDS This gene is a member of the Human CCDS set: [CCDS34237.1](#)

UniProtKB This gene has proteins that correspond to the following UniProtKB identifiers: [Q14949](#)

RefSeq This Ensembl/Gencode gene contains transcript(s) for which we have [selected identic](#) they will be in the [External references](#) table

Ensembl version ENSG00000164405.11

Other assemblies This gene maps to [132,202,322-132,204,539](#) in GRCh37 coordinates.
View this locus in the GRCh37 archive: [ENSG00000164405](#)

Gene type Protein coding

Annotation method Annotation for this gene includes both automatic annotation from Ensembl and Havan

Go to Region in Detail for more tracks and navigation options (e.g. zooming)

Genes (Comprehensive set...)

Contigs

22.22 kb

132,857,500 132,860,000 132,862,500 132,865,000 132,867,500 132,870

Transcripts: click for more info

- UQCRCQ-202 > protein coding
- UQCRCQ-206 > retained intron
- UQCRCQ-205 > processed transcript
- UQCRCQ-203 > protein coding
- UQCRCQ-201 > protein coding
- UQCRCQ-204 > retained intron

< ACD04500.1

This page summarises the gene, including its location, name and equivalents in other databases. At the bottom of the page, a graphic shows a region view with the transcripts. We can see exons shown as blocks with introns as lines linking them together. Coding exons are filled, whereas non-coding exons are empty. We can also see the overlapping and neighbouring genes and other genomic features.

There are different tabs for different types of features, such as genes, transcripts or variants. These appear side-by-side across the blue bar, allowing you to jump back and forth between features of interest. Each tab has its own navigation column down the left hand side of the page, listing all the things you can see for this feature.

Let's walk through this menu for the gene tab. How can we view the genomic sequence? Click *Sequence* at the left of the page.

Human (GRCh38.p13) ▼

Location: 5:132,866,600-132,868,847 Gene: UQCRC

Gene-based displays

- Summary
 - Splice variants
 - Transcript comparison
 - Gene alleles
- Sequence
 - Secondary Structure
- Comparative Genomics
 - Genomic alignment
 - Gene tree
 - Gene gain/loss tree

Gene: UQCRC

Description

Gene Synonym

Location

About this gene

Transcripts

Genome assembly

Click sequence

The sequence is shown in FASTA format. The FASTA header contains the genome assembly, chromosome, coordinates and strand (1 or -1) – this gene is on the positive strand.

Marked-up sequence ?

[Download sequence](#) [BLAST this sequence](#)

BLAST or download this sequence

Exons UQCRC exons All exons in this region

Markup loaded

Upstream sequence

Exon of an overlapping gene

```
>chromosome:GRCh38:5:132866030:132869447:1
TGTATTATCTCTGGGAACGTTAACAGCCTGCCAGGGGACTCTAAACAGAGTCCAAGAG
AAGCCTTCGATGTGTGTCCTGATTTTGGTGCTGGGGTCGCTCATGGCCTGAAAATAACG
AAGCCCCCTTGAGCTACTGGCCAAGGCCAAATGAAACCTCGTGCTCCTCTGCTGTGGGG
CGGCCCCGGCTCGGGGCTCAGAGGCGCTCTGTCTACAGCTGGGAAGACTCGCAGGCCCA
AGCTCCTGGCCAGCGCTGCCCGGAGGTCGGCAGGCCCTTCTCGTCACCTTTTGTTC
CTCCCCCGCTCCCGCATTCCGCCGCTTCCTGACTGGGATTCACAGAAAAGCCGAGGG
CTGAGGAGAAGTGTGAGCGCTCCGCTGTCCACTGTCCCCAAAGTCAGTTCAATCCCC
GACGTCCTCCGCTAGGCTCCACCCACCGGCCCGGGCAGGGCTCCAAGGCACCTCCCAC
CTACGGGTCACCCAGTCAGCCCACTTCTTTCTGGGACAAAGGCGTCATCCCTTAGAGACA
GTAGGAAAATGGTATCTCCCGAAGTTACCTCAGACCTCCAAGAGCGGCTTCCAACCTT
GCCGGAATGACGAACGAGTCAACCGGATCGGTGACTGTGGAGGCGAGCTGAGCCCTGT
GCGTGAGTGGGGTCTGGTTGTGCACTGTTCTGAGCCCTGGGAGGCTAGGGGCGCCCCG
TGGGCTGGGAAGGATAAGGAGTGCAGGGGCGAGGCTCTGGGGTGGGGATGGACCCCG
```

UQCRC exon

Exons are highlighted within the genomic sequence, both exons of our gene of interest and any neighbouring or overlapping gene. By default, 600 bases are shown up and downstream of the gene. We can make changes to how this sequence appears with the blue *Configure this page* button found at the left. This allows us to change the flanking regions, add variants, add line numbering and more. Click on it now.

Configure Page

Personal Data

Display options

Manage configurations

Reset configuration

Select from available configurations:

Current unsaved

Save current configuration

Display options

5' Flanking sequence (upstream):

600

*(Maximum of 1000000)

3' Flanking sequence (downstream):

600

*(Maximum of 1000000)

Number of base pairs per row:

60 bps

Additional exons to display:

Core exons

Orientation of additional exons:

Display exons in both orientation

Show variants:

No

Hide variants longer than 10bp:

☒

Hide variants by frequency (MAF):

Don't hide

Filter variants by consequence type:

No filter

3 prime UTR variant

5 prime UTR variant

NMD transcript variant

coding sequence variant

Filter variants by evidence status:

No filter

1000Genomes

Cited

ESP

ExAC

Hide individual variant sources:

No filter

Hide AMDGC

Hide Affy GenomeWideSNP_6 C

Hide Archive dbSNP

Line numbering:

None

Show variants

Filter to only 10

Show line number

Once you have selected changes (in this example, *Show variants*, *1000 Genomes variants* and *Line numbering*) click at the top right.

421

GACGTCCTCCCTAGGCTCCACCCACCGCCCGGGCAGGGCCTCCAAGGCACCTCCCAC

480

431: rs563868690 450: rs153867588

481

CTACGGGTCACCCAGTCAGCCCACTTCTTCTGGGACAAAGGCGTCATCCCTTAGAGACN

540

533: rs114932352 540: rs60321536

541

TTAGGAAAAATGTATCTCCCGAAGTTACCTCAGACCTCCAAAGCNGCTTCCAACCTT

600

541: rs192032018 551: rs548641026

601

GCCGGAAATGACGAACGAGTCAACCGGATCGGTGACTGTGGAGGGCGAGCTGANCCTGT

660

654: rs527907338 656: rs369119794

Links to variant tab

You can download this sequence by clicking in the *Download sequence* button above the sequence:

Download sequence

This will open a dialogue box that allows you to pick between plain FASTA sequence, or sequence in RTF, which includes all the coloured annotations and can be opened in a word processor. If you want run a sequence analysis tool, download as FASTA sequence, whereas if you want to analyse the sequence visually, RTF is best for this. This button is available for all sequence views.

Download sequence

File name:

File format:

[Preview](#) [Download](#) [Download Compressed](#)

Settings

Sequences to export:

- ☐ Select/deselect all
- ☐ cDNA (transcripts)
- ☐ Coding sequences (CDS)
- ☐ Amino acid sequences
- ☐ 5' UTRs
- ☐ 3' UTRs
- ☐ Exons
- ☐ Introns
- ☒ Genomic sequence

5' Flanking sequence (upstream): * (Maximum of 1000000)

3' Flanking sequence (downstream): * (Maximum of 1000000)

Fields marked * are required

Guide to file formats

FASTA

Text sequence(s):
DNA and/or amino acids

```
>11 chr1:chromosome: chromosome: GRCh38:11:19:
CAGCGGGAAGCCACAGGCGCATCCCTAGTAGGCTACTTGG
TCGGCCCTCAGACAGAATCTCCCTACATTTTGCAGTTGGC
CCAGATATGGAGCAGGCTCAGGCGTGACGCGCGGTTGTAGT
TCTAAATCCCTGTAGACTTACCTCCCGCCGCGCTGGAC
```

RTF

Marked-up sequence,
with or without variants

```
ATTAGCAACAAAAAGCAAAACACGGG
GAGTCTCTTCCACAAACATGGGCAI
TCTTAGGGAGTAGAATATTGATGG
TTTTTGGTAATGTGGCTTTCGGT
AGGCCCTCACAAATTCGTCCAAGTG
TTTTCGTTTCCGCACCTGGGACCTG
GCTGGGCATGTGGAGCTGATGCTT
```

To find out what the protein does, have a look at GO terms from the [Gene Ontology consortium](#). There are three pages of GO terms, representing the three divisions in GO: Biological process (what the protein does), Cellular component (where the protein is) and Molecular function (how it does it). Click on *GO: Biological process* to see an example of the GO pages.

GO: Biological process

Find other genes linked to this term

Accession	Term	Evidence	Annotation source	Transcript IDs	
GO:0006122	mitochondrial electron transport, ubiquinol-cytochrome c	IBA	GO_Central	ENST00000378667 ENST00000378670 ENST00000378665	Search BioMart View on karyoty
GO:0021539	cardiac muscle tissue morphogenesis	IEA	Ensembl	ENST00000378670	Search BioMart View on karyoty
GO:0021548	muscle development	IEA	Ensembl	ENST00000378670	Search BioMart View on karyoty
GO:0021680	cerebellar Purkinje cell layer development	IEA	Ensembl	ENST00000378670	Search BioMart View on karyoty

Hover over the codes for definitions

GO terms are linked to transcripts, not genes.

See which ones

Here you can see the functions that have been associated with the gene. There are three-letter codes that indicate how the association was made, as well as links to the specific transcript they are linked to.

We also have links out to other databases which have information about our genes and may focus on other topics that we don't cover, like Gene Expression Atlas or OMIM. Go up the left-hand menu to *External references*:

External references ?

Download all tables as CSV

This gene corresponds to the following database identifiers:

Filter	
External database	Database identifier
Expression Atlas	ENSG00000164405 [view all locations]
HGNC Symbol	UQCQRQ ubiquinol-cytochrome c reductase complex III subunit VII [view all locations]
MIM gene	UBIQUINOL-CYTOCHROME c REDUCTASE, COMPLEX III SUBUNIT VII, 9.5-KD ; UQCQRQ [*612080] UBIQUINOL-CYTOCHROME c REDUCTASE, COMPLEX III SUBUNIT VII, 9.5-KD ; UQCQRQ ; QPC [view all locations]
MIM morbid	MITOCHONDRIAL COMPLEX III DEFICIENCY, NUCLEAR TYPE 4 ; MC3DN4 [#615159] MITOCHONDRIAL COMPLEX III DEFICIENCY, NUCLEAR TYPE 4 ; MC3DN4 [view all locations]
NCBI gene (formerly Entrezgene)	UQCQRQ ubiquinol-cytochrome c reductase complex III subunit VII [view all locations]
Reactome gene	R-HSA-1428517 The citric acid (TCA) cycle and respiratory electron transport [view all locations] R-HSA-1430728 Metabolism [view all locations] R-HSA-163200 Respiratory electron transport, ATP synthesis by chemiosmotic coupling, and heat production by uncoupling proteins. [view all locations] R-HSA-611105 Respiratory electron transport [view all locations]
WikiGene	UQCQRQ ubiquinol-cytochrome c reductase complex III subunit VII [view all locations]

Demo: The transcript tab

We're now going to explore the different transcripts of *UQCQRQ*. Click on *Show transcript table* at the top.

Show transcript table

Show/hide columns (1 hidden)								Filter			
Transcript ID	Name	bp	Protein	Biotype	CCDS	UniProt Match	RefSeq Match	Flags			
ENST00000378670.8	UQCQRQ-203	1573	82aa	Protein coding	CCDS34237	O14949	NM_014402.5	MANE Select v0.95	Ensembl Canonical	GENCODE basic	APPRIS P1
ENST00000378665.1	UQCQRQ-201	586	82aa	Protein coding	CCDS34237	O14949	-		GENCODE basic	APPRIS P1	TSL:1
ENST00000378667.1	UQCQRQ-202	429	82aa	Protein coding	CCDS34237	O14949	-		GENCODE basic	APPRIS P1	TSL:2
ENST00000496429.1	UQCQRQ-205	258	No protein	Processed transcript	-	-	-			TSL:2	
ENST00000498309.1	UQCQRQ-206	874	No protein	Retained intron	-	-	-			TSL:2	
ENST00000480372.1	UQCQRQ-204	509	No protein	Retained intron	-	-	-			TSL:2	

Here we can see a list of all the transcripts of *UQCQRQ* with their identifiers, lengths, biotypes and flags to help you decide which ones to look at.

If we were to only choose one transcript to analyse, we would choose UQCQRQ-203 because it is the MANE Select and Ensembl Canonical. This means it is both 100% identical to the RefSeq transcript NM_014402.5 and both Ensembl and NCBI agree that it is the most biologically important transcript.

Click on the ID, *ENST00000378670.8*.

You are now in the *Transcript tab* for UQCQRQ-203. We can still see the gene tab so we can easily jump back. The left hand navigation column provides several options for the transcript UQCQRQ-203 - many of these are similar to the options you see in the gene tab, but not all of them. If you can't find the thing you're looking for, often the solution is to switch tabs.

Click on the *Exons* link. This page is useful for designing RT-PCR primers because you can see the sequences of the different exons and their lengths.

Protein summary ⓘ

Protein domains for ENSP00000367939.3



You can also see the structure of the protein from the PDB by clicking on *PDB 3D Protein model*.

3D Protein model (PDB) ⓘ

Ensembl protein: [ENSP00000367939](#) | PDB model: [5XTE](#) (Cytochrome b-c1 complex subunit 8 - chains A,N)



This uses LiteMol to show a 3D protein. You can use all the normal controls that you would use with LiteMol, plus plot Ensembl features like Exons and variants onto the structure using the options on the right. We allow you to see the top ten PDB models for this protein, based on coverage and quality scores, you can choose which at the top of the viewer.

Exploring variants in Ensembl, Demo

In any of the sequence views shown in the Gene and Transcript tabs, you can view variants on the sequence. You can do this by clicking on [Configure](#) this page from any of these views.

Let's take a look at the Gene sequence view for *HBB* in human. Search for *HBB* and go to the *Sequence view*.

If you can't see variants marked on this view, click on *Configure this page* and select *Show variants: Yes and show links*. You may also wish to add a filter to the variants to allow them to load more quickly, we'll add *Filter variants by evidence status: 1000Genomes*.

Marked-up sequence ?

[illegible]

Find out more about a variant by clicking on it.

Variation: rs140196387

Class SNP

Source dbSNP

Location 1:26432488

Alleles G/A/C (Forward strand)

Consequences

- 5 prime UTR variant
- non coding transcript exon variant
- Intron variant
- MMMD transcript variant
- non coding transcript variant
- TF binding site
- regulatory region variant

[Explore this variant](#)

[Gene/Transcript Locations](#)

[Population Allele Frequencies](#)

You can add variants to all other sequence views in the same way.

You can go to the Variation tab by clicking on the *variant ID*. For now, we'll explore more ways of finding variants.

To view all the sequence variants in table form, click the *Variant table* link at the left of the gene tab.

Variant table

Variant table

This table shows known variants for this gene. Use the 'Consequence Type' filter to view a subset of these.

Filter: Global MAF: All SIFT: All PolyPhen: All Consequences: missense variant Filter Other Columns

Variant ID	Chr: bp	Alleles	Global MAF	Class	Source	Evidence	Clin. Sig.	Conseq. Type	AA	AA coord	SIFT	Poly-Phen	CADD	REV EL	Meta LR	Mutation Assessor	Transcript
rs33985739	11:5225601	G/C/T	-	SNP	dbSNP		?	missense variant	H/Q	147	0.03	0.351	22	0.828	0.788	0.879	ENST00000335295.4
rs33985739	11:5225601	G/C/T	-	SNP	dbSNP		?	missense variant	H/Q	147	0.03	0.351	22	0.828	0.788	0.879	ENST00000335295.4
rs33954264	11:5225602	T/A/C/G	-	SNP	dbSNP		?	missense variant	H/L	147	0	0.76	24	0.865	0.874	0.955	ENST00000335295.4
rs33954264	11:5225602	T/A/C/G	-	SNP	dbSNP		?	missense variant	H/R	147	0.01	0.76	24	0.72	0.68	0.403	ENST00000335295.4
rs33954264	11:5225602	T/A/C/G	-	SNP	dbSNP		?	missense variant	H/P	147	0.01	0.76	24	0.97	0.908	0.955	ENST00000335295.4

You can filter the table to only show the variants you're interested in. For example, click on *Consequences: All*, then select the variant consequences you're interested in. For display purposes, the table above has already been filtered to only show missense variants.

Consequences (27/27 on)

Turn All Off PTV PTV & Missense Only Exonic Turn All On

PTV = Protein Truncating Variant

transcript ablation (0) On	inframe deletion (92) On	mature miRNA variant (0) On
splice donor variant (131) On	missense variant (2128) On	5 prime UTR variant (385) On
splice acceptor variant (137) On	protein altering variant (14) On	3 prime UTR variant (1336) On
stop gained (83) On	splice region variant (538) On	non coding transcript exon variant (1074) On
frameshift variant (440) On	incomplete terminal codon variant (10) On	intron variant (3098) On
stop lost (8) On	synonymous variant (441) On	NMD transcript variant (2291) On
start lost (40) On	stop retained variant (1) On	non coding transcript variant (407) On
transcript amplification (0) On	start retained variant (0) On	upstream gene variant (0) On
inframe insertion (9) On	coding sequence variant (2731) On	downstream gene variant (0) On

Apply Cancel

You can also filter by the different pathogenicity scores and MAF, or click on *Filter other columns* for filtering by other columns such as Evidence or Class.

The table contains lots of information about the variants. You can click on the IDs here to go to the Variation tab too.

You can also see the phenotypes associated with a gene. Click on *Phenotype* in the left hand menu.

Phenotypes ?

Phenotypes, diseases and traits associated with this gene ENSG00000244734

Phenotype, disease and trait	Source
ALPHA-THALASSEMIA	MIM morbid
BETA-THALASSEMIA	MIM morbid
BETA-THALASSEMIA INTERMEDIA	Orphanet
Beta-thalassemia major	Orphanet
BETA-THALASSEMIA_DOMINANT INCLUSION BODY TYPE	MIM morbid
Delta-beta thalassemia	Orphanet
Dominant beta-thalassemia	MIM morbid
ERYTHROCYTOSIS_FAMILIAL_6	MIM morbid
FETAL HEMOGLOBIN QUANTITATIVE TRAIT LOCUS 1	MIM morbid
HEINZ BODY ANEMIAS	MIM morbid

Showing 1 to 10 of 26 entries

Phenotypes linked to the gene

Phenotype, disease and trait annotations associated with variants in this gene

Phenotype, disease and trait	Source(s)	Number of variants	Show/hide details
ALL variants with a phenotype annotation (WARNING: details table may not load for this number of variants!)	-	1355	Show
Anemia	ClinVar	2	Show
Annotated by HGMD	HGMD-PUBLIC	715	Show
BETA-PLUS-THALASSEMIA	ClinVar	2	Show
BETA-PLUS-THALASSEMIA_DOMINANT	ClinVar	1	Show
BETA-THALASSEMIA INTERMEDIA_DOMINANT	ClinVar	1	Show
BETA-THALASSEMIA_DOMINANT INCLUSION BODY TYPE	ClinVar	16	Show
Beta thalassemia intermedia	ClinVar	3	Show
Beta thalassemia major	ClinVar	3	Show
Beta-houston-thalassemia	ClinVar	1	Show

Showing 1 to 10 of 466 entries

Phenotypes linked to variants in the gene

Phenotypes linked
to variants in
the gene

Phenotype, disease and trait annotations associated orthologues of this gene in other species

Phenotype, disease and trait	Source	Species	Gene
anatomical system quality	ZFIN	Zebrafish (Danio rerio)	ENSDARG00000088330 hbae1.3
liver quality	ZFIN	Zebrafish (Danio rerio)	ENSDARG00000088330 hbae1.3
intestine quality	ZFIN	Zebrafish (Danio rerio)	ENSDARG00000088330 hbae1.3
whole organism shape	ZFIN	Zebrafish (Danio rerio)	ENSDARG00000088330 hbae1.3

Phenotypes linked to orthologues of the gene

Open the transcript table and go to HBB-201 ENST00000335295, then click on *Haplotypes* in the left hand menu.

Haplotypes ?

Export data as JSON Switch to CDS view

Show All entries Show/hide columns Filter

Protein haplotype	Flags	Protein	Protein	Protein	Protein	Protein	Protein	Protein
REF		0.952 (5727)	0.887 (1172)	0.99 (687)	0.985 (993)	0.998 (1004)	0.989 (967)	0.895 (904)
7E>V	D	0.0382 (230)	0.0998 (132)	0.0072 (5)	0 (0)	0 (0)	0 (0)	0.0921 (93)
7E>K		0.00449 (27)	0.0129 (17)	0 (0)	0 (0)	0 (0)	0 (0)	0.0099 (10)
27E>K		0.00233 (14)	0 (0)	0 (0)	0.00794 (8)	0 (0)	0.00613 (6)	0 (0)
18K>* 19del(13)	S	0.00099 (6)	0 (0)	0 (0)	0.00595 (6)	0 (0)	0 (0)	0 (0)
122E>Q		0.000499 (3)	0 (0)	0 (0)	0 (0)	0.00307 (3)	0 (0)	0 (0)

View by protein on CDS

Likely to affect protein

Frequency in 1000

View by protein or CDS

S Likely to affect protein

Frequency in 1000

• Variants that are found together in individuals

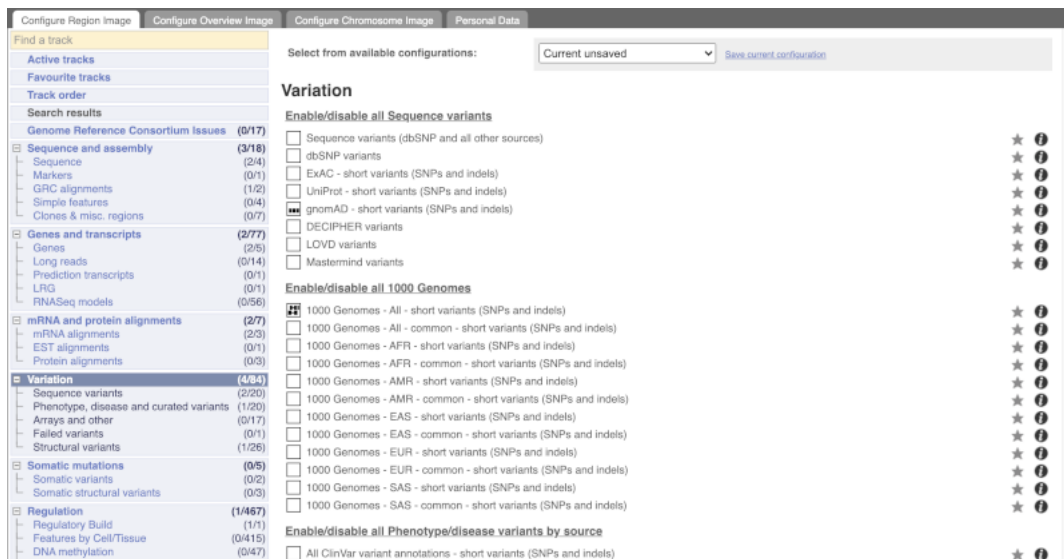
Genomes populations

The Haplotypes view in the transcript tab shows you the actual protein and CDS sequences in 1000 Genomes individuals. This is possible because the 1000 Genomes study has phased genotypes, so we know which alleles occur on which of the chromosome pairs. The table lists all the versions of the protein that occur along with their frequencies, including the reference sequence and sequences with one or more alternative alleles.

Click on one of the haplotypes, we'll go for *18K>*,19del[130]*, to find out more about it. Here you will see the frequency in the 1000 Genomes subpopulations, the sequence and the 1000 Genomes individuals where this protein is found.

Let's have a look at variants in the Location tab. Click on the *Location* tab in the top bar.

Configure this page and open *Variation* from the left-hand menu.



There are various options for turning on variants. You can turn on variants by source, by frequency, presence of a phenotype or by individual genome they were isolated from. You can also turn on genotyping chips.

Let's have a look at a specific variant. If we zoomed in we could see the variant rs334 in this region, however it's easier to find if we put *rs334* into the search box. Click through to open the Variation tab.

The icons show you what information is available for this variant. Click on *Genes and regulation*, or follow the link on the left.

Genes and regulation

Gene and Transcript consequences

Gene	Transcript (strand)	Allele (Tr.)	Consequence Type	Position in transcript	Position in CDS	Position in protein	AA	Codons	SIFT	PolyPhen	CADD	REVEL	MetaLR	Mutation Ass.
ENSG00000244734 HGNC: HBB	ENST00000335295.4 (-) biotype: protein_coding	A (T)	missense variant	70 (out of 628)	20 (out of 444)	7 (out of 147)	E/V	GAG/GTG	0.01	0.007	13	0.535	0.071	0.939
ENSG00000244734 HGNC: HBB	ENST00000335295.4 (-) biotype: protein_coding	C (G)	missense variant	70 (out of 628)	20 (out of 444)	7 (out of 147)	E/G	GAG/GGG	0.05	0.009	14	0.472	0.529	0.766
ENSG00000244734 HGNC: HBB	ENST00000335295.4 (-) biotype: protein_coding	G (C)	missense variant	70 (out of 628)	20 (out of 444)	7 (out of 147)	E/A	GAG/CGC	0.29	0.006	5	0.503	0.459	0.326
ENSG00000244734 HGNC: HBB	ENST00000335295.4 (-) biotype: protein_coding	T (A)	missense variant	70 (out of 628)	20 (out of 444)	7 (out of 147)	E/V	GAG/GTG	0.01	0.007	13	0.535	0.071	0.939
ENSG00000244734 HGNC: HBB	ENST00000380315.2 (-) biotype: protein_coding	C (G)	missense variant	250 (out of 502)	20 (out of 272)	7 (out of 90)	E/G	GAG/GGG	0.07	0.009	13	0.472	0.529	0.766
ENSG00000244734 HGNC: HBB	ENST00000380315.2 (-) biotype: protein_coding	G (C)	missense variant	250 (out of 502)	20 (out of 272)	7 (out of 90)	E/A	GAG/CGC	0.34	0.006	5	0.503	0.459	0.326
ENSG00000244734 HGNC: HBB	ENST00000485743.1 (-) biotype: protein_coding	A (T)	missense variant	71 (out of 680)	20 (out of 336)	7 (out of 111)	E/V	GAG/GTG	0.01	0.013	13	0.535	0.071	0.939
ENSG00000244734 HGNC: HBB	ENST00000485743.1 (-) biotype: protein_coding	C (G)	missense variant	71 (out of 680)	20 (out of 336)	7 (out of 111)	E/G	GAG/GGG	0.04	0.449	14	0.472	0.529	0.766
ENSG00000244734 HGNC: HBB	ENST00000485743.1 (-) biotype: protein_coding	G (C)	missense variant	71 (out of 680)	20 (out of 336)	7 (out of 111)	E/A	GAG/CGC	0.31	0.369	5	0.503	0.459	0.326
ENSG00000244734 HGNC: HBB	ENST00000633227.1 (-) biotype: nonsense_mediated_decay	A (T)	missense variant NMD transcript variant	70 (out of 609)	20 (out of 168)	7 (out of 55)	E/V	GAG/GTG	0	0.054	13	0.535	0.071	0.939

Gene expression correlations

Gene	P-value (-log ₁₀)	Effect size	Tissue
ENSG00000132109	3.4404823802790916	0.664553	Th2_memory
ENSG00000132109	3.2285404012949654	0.550356	Treg_memory
ENSG00000132109	3.021075290019598	0.552243	CD8_T-cell_naive
ENSG00000132109	2.808660057318154	0.417533	CD4_T-cell_naive
ENSG00000132109	2.62442365507921	0.530484	Treg_naive
ENSG00000132109	2.5061573191302773	0.468006	Th17_memory
ENSG00000181074	2.1760728019324116	0.510024	LCL
ENSG00000132109	2.1492582182910924	0.527156	B-cell_naive
ENSG00000181023	1.974640875470262	-3.32732	NK-cell_naive
ENSG00000180878	1.7731527333799302	2.51023	monocyte_CD16_naive

Regulatory feature consequences

Regulatory feature	Active in cell lines	Feature type	Allele	Consequence type	Variant position
ENSR000000952768	IHECRE00003705, IHECRE000000989	promoter_flanking_region	A	regulatory region variant	2601 (out of 4598)
ENSR000000952768	IHECRE00003705, IHECRE000000989	promoter_flanking_region	C	regulatory region variant	2601 (out of 4598)
ENSR000000952768	IHECRE00003705, IHECRE000000989	promoter_flanking_region	A	regulatory region variant	2601 (out of 4598)

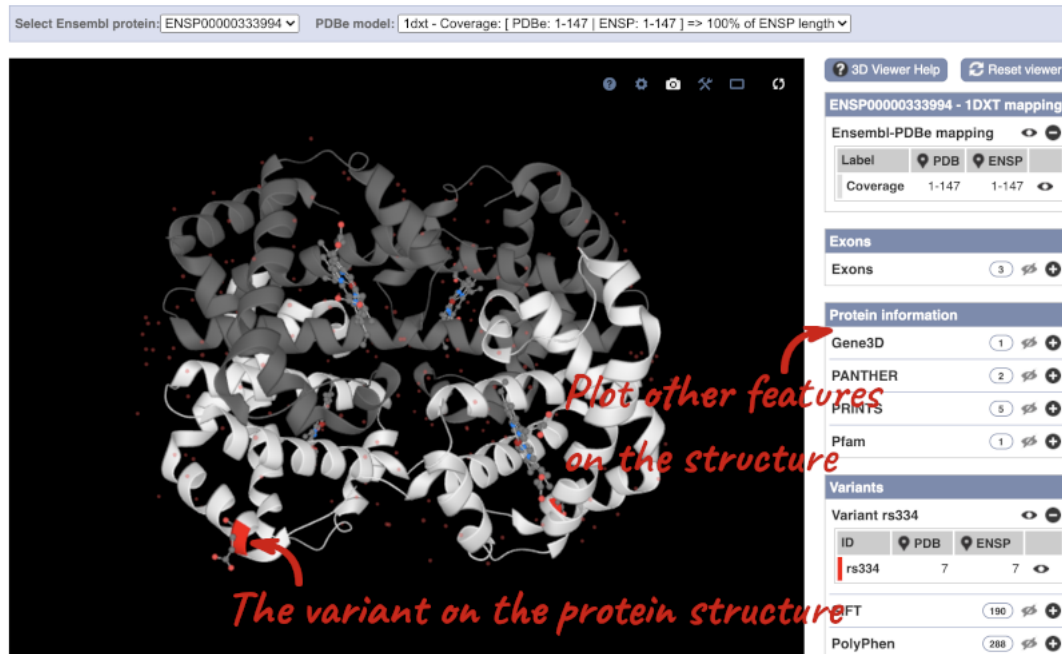
No overlap with Ensembl Motif features

This page illustrates the genes the variant falls within and the consequences on those genes, including pathogenicity predictors. It also shows data from [GTEx](#) on genes that have increased/decreased expression in individuals with this variant, in different tissues. Finally, regulatory features and motifs that the variant falls within are shown.

We can also see the variant in the protein structure by clicking on *3D Protein model*.

3D Protein model (PDBe) ?

Variant rs334 (missense_variant) | Ensembl protein: [ENSP00000333994](#) | PDBe model: [1DXT](#) (Hemoglobin subunit beta - chains B,D)

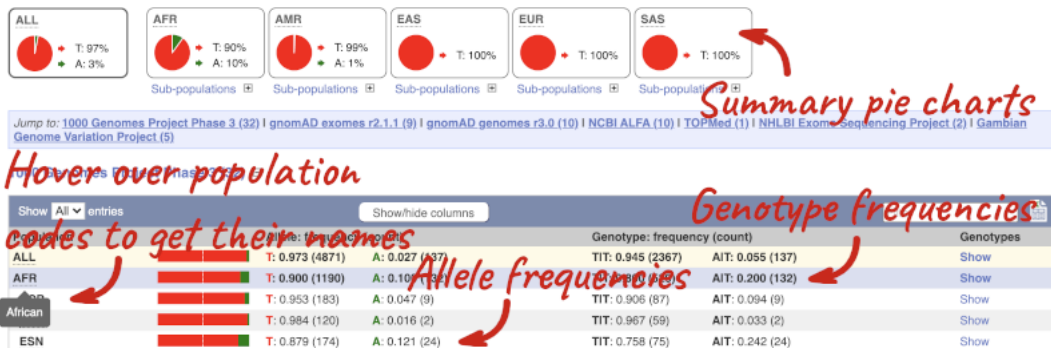


This is a LiteMol viewer, where you can rotate and zoom in on the structure. The variant location is highlighted, so you can see where it lands within the structure.

Let's look at population genetics. Click on *Population genetics* in the left-hand menu.

Population genetics ?

1000 Genomes Project Phase 3 allele frequencies



The population allele frequencies are shown by study, including 1000 Genomes and gnomAD. Where genotype frequencies are available, these are shown in the tables.

There are big differences in allele frequencies between populations. Let's have a look at the phenotypes associated with this variant to see if they are known to be specific to certain human populations. Click on *Phenotype Data* in the left-hand menu.

Phenotype Data

Significant association(s)

Show All entries Show/hide columns Filter								
Phenotype, disease and trait	Source(s)	Mapped Terms	Ontology Accessions	Supporting evidence	External reference	Clinical significance	Reported gene(s)	Statistics
Anemia	ClinVar [New York Genome Cent...]	Anemia, anemia (phenotype)	EFQ:0004272, HP:0001903	-	-	★ ★ ★ ★ ★	LOC106099062, HBB, LOC107133510	-
beta Thalassemia	ClinVar [Mendelics/Myriad Wom...]	-	Orphanet:848	-	-	? ★ ★ ★ ★ ★	LOC106099062, HBB, LOC107133510	-
BETA-THALASSEMIA, DOMINANT INCLUSION BODY TYPE	ClinVar [Division of Human Ge...]	-	Orphanet:231226	-	-	★ ★ ★ ★ ★	LOC106099062, HBB, LOC107133510	-
Blood cell traits multivariate analysis	NHGRI-EBI GWAS catalog	11 terms	11 accessions	-	PMID:31080455	-	HBB	p-value: 5.00e-20
Blood cell traits multivariate analysis	NHGRI-EBI GWAS catalog	11 terms	11 accessions	-	PMID:31080455	-	HBB	p-value: 1.00e-12

This variant is associated with various phenotypes, including sickle cell and malaria resistance. These phenotype associations come from sources including the GWAS catalog, ClinVar, Orphanet and OMIM. Where available, there are links to the original paper that made the association, the allele that is associated with the phenotype and p-values and other statistics.

Annotating genetic variants with the VEP, Demo

We have identified five variants on human chromosome nine, C-> A at 128203516, an A deletion at 128328461, C->A at 128322349, C->G at 128323079 and G->A at 128322917.

We will use the Ensembl VEP to determine:

- Have my variants already been annotated in Ensembl?
- What genes are affected by my variants?
- Do any of my variants affect gene regulation?

Go to the front page of Ensembl and click on the Variant Effect Predictor.

Variant Effect Predictor >

Analyse your own variants and predict the functional consequences of known and unknown variants

This page contains information about the VEP, including links to download the script version of the tool. Click on *Launch VEP* to open the input form:

Variant Effect Predictor ⓘ

The screenshot shows the Ensembl VEP input form. At the top, there's a 'New job' button and 'Clear form' and 'Close' links. The 'Species' dropdown is set to 'Homo_sapiens'. Below it, the 'Assembly' is 'GRCh38.p13'. The 'Name for this job (optional):' field is empty. The 'Input data:' section has a 'Paste data:' box with a red arrow pointing to it and the text 'Name your job' and 'Paste in your data...'. Below the paste box, there are options to 'Or upload file:' (with a 'Choose file' button) and 'Or provide file URL:'. The 'Transcript database to use:' section has four radio buttons: 'Ensembl/GENCODE transcripts' (selected), 'Ensembl/GENCODE basic transcripts', 'RefSeq transcripts', and 'Ensembl/GENCODE gene models'. A red arrow points to the selected option with the text '...or upload a file' and 'Choose the transcript database'.

The data is in VCF format:
chromosome coordinate id reference alternative

Put the following into the Paste data box:

```
9 128328460 var1 TA T
9 128322349 var2 C A
9 128323079 var3 C G
9 128322917 var4 G A
9 128203516 var5 C A
```

The VEP will automatically detect that the data is in VCF.

There are further options that you can choose for your output. These are categorised as *Identifiers*, *Variants and frequency data*, *Additional annotations*, *Predictions*, *Filtering options* and *Advanced options*. Let's open all the menus and take a look.

Additional configurations:

The screenshot shows the 'Identifiers' configuration menu. It has a title 'Identifiers' and a subtitle 'Additional identifiers for genes, transcripts and variants'. There are three checkboxes: 'Gene symbol:' (checked), 'Transcript version:' (checked), and 'Gene build:' (unchecked).

BLOSUM62: ☐

Ancestral allele: ☐

Filtering options Pre-filter results by frequency or consequence type

Filters

Filter by frequency:

- ☒ No filtering
- ☐ Exclude common variants
- ☐ Advanced filtering

Return results for variants in coding regions only: ☐

Restrict results:

NB: Restricting results may exclude biologically important data!

Advanced options Additional enhancements

[Run](#)

Hover over the options to see definitions.

We're going to select some options:

- *HGVS*, annotation of variants in terms of the transcripts and proteins they affect, commonly-used by the clinical community
- *Phenotypes*
- *Protein domains*

When you've selected everything you need, scroll right to the bottom and click *Run*.

Recent jobs

[Refresh](#)

Analysis	Jobs	Submitted at
Variant Effect Predictor	VEP analysis of pasted data in Homo_sapiens Queued	17/09/2021, 14:39 (BST)

The display will show you the status of your job. It will say *Queued*, then automatically switch to *Done* when the job is done, you do not need to refresh the page. You can edit or discard your job at this time. If you have submitted multiple jobs, they will all appear here.

Click *View results* once your job is done.

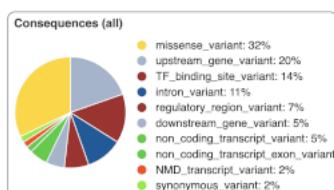
In your results you will see a graphical summary of your data, as well as a table of your results.

Variant Effect Predictor results

[Job details](#)

[Summary statistics](#)

Category	Count
Variants processed	5
Variants filtered out	0
Novel / existing variants	1 (20.0) / 4 (80.0)
Overlapped genes	4
Overlapped transcripts	24
Overlapped regulatory features	2



Results preview

[Navigation \(per variant\)](#)

Show: 1/5 [All](#) variants

[Filters](#)

Uploaded variant

is

defined

[Add](#)

[Download](#)

[New job](#)

All: [VCF VEP](#)

BioMart: [Variants](#) [Genes](#)

Uploaded variant	Location	Consequence	Symbol	Gene	Feature type	Feature	Biotype	Exon	HGVS
var5	9-128203516-128203516	missense_variant	DNM1	ENSG00000106976	Transcript	ENST0000034117.11	protein_coding	1/23	ENST0000034117.11
var5	9-128203516-128203516	missense_variant	DNM1	ENSG00000106976	Transcript	ENST0000034117.11	protein_coding	1/23	ENST0000034117.11
var5	9-128203516-128203516	intron_variant	CI21	ENSG00000148337	Transcript	ENST00000372948.7	protein_coding	-	ENST00000372948.7

The results table is enormous and detailed, so we're going to go through the it by section. The first column is *Uploaded variant*. If your input data contains IDs, like ours does, the ID is listed here. If your input data is

only loci, this column will contain the locus and alleles of the variant. You'll notice that the variants are not necessarily in the order they were in in your input. You'll also see that there are multiple lines in the table for each variant, with each line representing one transcript or other feature the variant affects.

You can mouse over any column name to get a definition of what is shown.

The next few columns give the information about the feature the variant affects, including the consequence. Where the feature is a transcript, you will see the gene symbol and stable ID and the transcript stable ID and biotype. Where the feature is a regulatory feature, you will get the stable ID and type. For a transcription factor binding motif (labelled as a MotifFeature) you will see just the ID. Most of the IDs are links to take you to the gene, transcript or regulatory feature homepage.

Uploaded variant	Location	Consequence	Symbol	Gene	Feature type	Feature	Biotype
var4	9:128322917-128322917	missense_variant	COQ4	ENSG00000167113	Transcript	ENST00000372875.3	protein_coding
var4	9:128322917-128322917	upstream_gene_variant	TRUB2	ENSG00000167112	Transcript	ENST00000372890.6	protein_coding
var4	9:128322917-128322917	upstream_gene_variant	TRUB2	ENSG00000167112	Transcript	ENST00000460320.1	processed_transcript
var4	9:128322917-128322917	missense_variant	COQ4	ENSG00000167113	Transcript	ENST00000608951.5	protein_coding
var4	9:128322917-128322917	missense_variant	COQ4	ENSG00000167113	Transcript	ENST00000609948.1	protein_coding
var4	9:128322917-128322917	regulatory_region_variant	-	-	RegulatoryFeature	ENSR00000241858	promoter
var4	9:128322917-128322917	TF_binding_site_variant	-	-	MotifFeature	ENSM00000348409	-
var4	9:128322917-128322917	TF_binding_site_variant	-	-	MotifFeature	ENSM00150425749	-

This is followed by details on the effects on transcripts, including the position of the variant in terms of the exon number, cDNA, CDS and protein, the amino acid and codon change, transcript flags, such as MANE, on the transcript, which can be used to choose a single transcript for variant reporting, and pathogenicity scores. The pathogenicity scores are shown as numbers with coloured highlights to indicate the prediction, and you can mouse-over the scores to get the prediction in words. Two options that we selected in the input form are the HGVS notation, which is shown to the left of the image below and can be used for reporting, and the Domains to the right. The Domains list the proteins domains found, and where there is available, provide a link to the 3D protein model which will launch a LiteMol viewer, highlighting the variant position.

Exon	HGVS	cDNA position	CDS position	Protein position	Amino acids	Codons	Feature strand	MANE SELECT	Transcript support level	APPRIS	SIFT	PolyPhen	Domains
1/23	ENST00000341179.11:c.46C>A	138	46	16	L/M	CTG/ATG	1	-	1	-	0.02	0.01	Gene3D:3.40.50.300 PANTHER:PTHR11566 PANTHER:PTHR11566-SF32 SMART:SM00053 Superfamily:SSF52540
1/22	ENST00000372948.7:c.-6+670G>T	-	-	-	-	-	-1	-	2	A2	0.02	0.015	Protein structure 48 Protein domains
1/23	ENST00000393594.7:c.46C>A	129	-	-	-	-	-	-	-	-	0.02	0.03	Gene3D:3.40.50.300 PANTHER:PTHR11566 PANTHER:PTHR11566-SF32 SMART:SM00053 Superfamily:SSF52540
1/23	ENST00000475805.5:c.46C>A	138	46	16	L/M	CTG/ATG	1	-	2	A1	0.02	0.034	Gene3D:3.40.50.300 PANTHER:PTHR11566 PANTHER:PTHR11566-SF32 SMART:SM00053 Superfamily:SSF52540
1/22	ENST00000486160.3:c.46C>A	102	46	16	L/M	CTG/ATG	1	-	1	A1	0.02	0.014	Gene3D:3.40.50.300 SMART:SM00053 PANTHER:PTHR11566 PANTHER:PTHR11566-SF32

Where the variant is known, the ID of the existing variant is listed, with a link out to the variant homepage. In this example, only rsIDs from dbSNP are shown, but sometimes you will see IDs from other sources such as COSMIC. The VEP also looks up the variant in the Ensembl database and pulls back the allele frequency (AF in the table), which will give you the 1000 Genomes Global Allele Frequency. In our query, we have not selected allele frequencies from the continental 1000 Genomes populations or from gnomAD, but these could also be shown here. We can also see ClinVar clinical significance and the phenotypes associated with known variants or with the genes affected by the variants, with the variant ID listed for variant associations and the gene ID listed for gene associations, along with the source of the association.

Existing variant	AF	Clinical significance	Associated phenotypes
rs9697215	0.0517	benign	ClinVar: phenotype not specified (rs9697215 , ClinVar)
rs9697215	0.0517	benign	ClinVar: phenotype not specified (rs9697215 , ClinVar)
rs377735694	0.0002	-	Neonatal encephalomyopathy-cardiomyopathy-respiratory distress syndrome (ENSG00000167113 , Orphanet) COENZYME Q10 DEFICIENCY PRIMARY 7 (ENSG00000167113 , MIM morbid & DDG2P) COENZYME Q10 DEFICIENCY PRIMARY 7 (rs377735694 , ClinVar)

For variants that affect transcription factor binding motifs, there are columns that show the effect on motifs (you may need to click on *Show/hide columns* at the top left of the table to display these). Here you can see the position of the variant in the motif, if the change increases or decreases the binding affinity of the motif and the transcription factors that bind the motif.

Motif name	Motif position	High info position	Motif score change	TRANSCRIPTION FACTORS
ENSPFM0597	12	N	-0.005	TFAP2C-MAX
ENSPFM0177	1	N	-0.010	ETV6
ENSPFM0066	4	N	-0.055	E2F1-ELK1
ENSPFM0161	15	N	-0.019	ETV2-RFX5

Above the table is the *Filter* option, which allows you to filter by any column in the table. You can select a column from the drop-down, then a logic option from the next drop-down, then type in your filter to the following box. We'll try a filter of *Consequence*, followed by *is* then *missense_variant*, which will give us only variants that change the amino acid sequence of the protein. You'll notice that as you type *missense_variant*, the VEP will make suggestions for an autocomplete.

Filters

Uploaded variant

Location

Allele

Consequence

Impact

Symbol

is

defined

Add

Amino acids

Codons

Existing variant

AF

Clinic signi

You can export your VEP results in various formats, including VCF. When you export as VCF, you'll get all the VEP annotation listed under **CSQ** in the **INFO** column. After filtering your data, you'll see that you have the option to export only the filtered data. You can also drop all the genes you've found into the Gene BioMart, or all the known variants into the Variation BioMart to export further information about them.

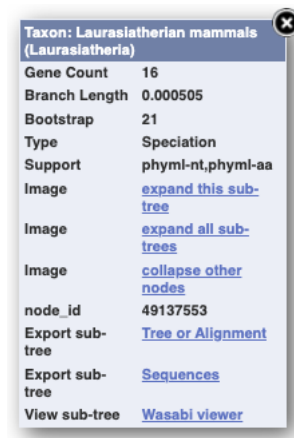
Gene trees and homologues, Demo

Let's look at the homologues of human *BRCA2*. Search for the *gene* and go to the *Gene* tab.

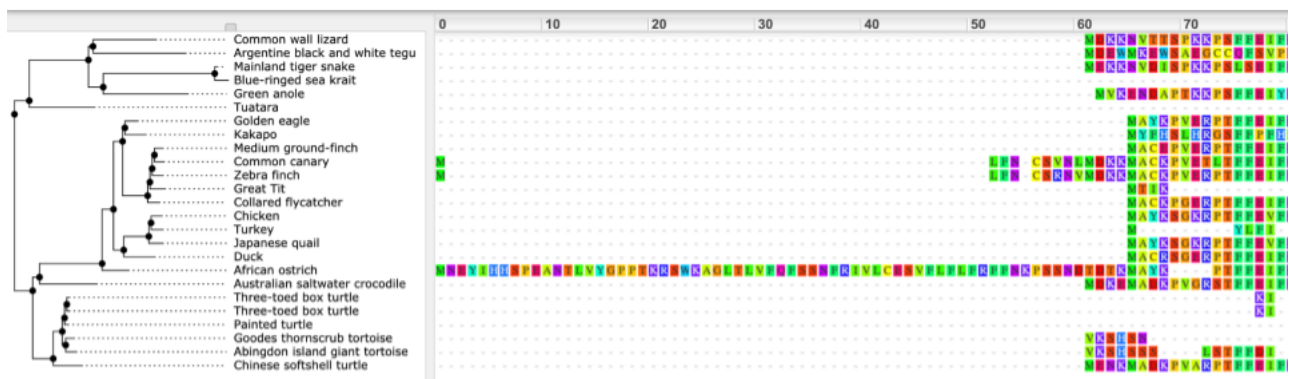
Click on *Gene tree*, which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.



Funnels indicate collapsed nodes. We can expand them by clicking on the node and selecting *Expand this sub-tree* from the pop-up menu.



We can also see the alignment of the sub-tree by clicking on Wasabi viewer, which will open a pop-up:



You can download the tree in a variety of formats. Click on the download icon in the bar at the top of the image to get a pop-up where you can choose your format.

Download

File name:

File format:

Guide to file formats

CLUSTALW	FASTA	Mega	MSF
<pre> CLUSTAL W(1.83) multiple sequence homo_sapiens/1-465588 CCTCAGGAC pan_troglodytes/1-465588 CCCCAGGAC ** ***** homo_sapiens/1-465588 CCGAGTGCC pan_troglodytes/1-465588 CCGAGTGCC ***** </pre>	<pre> >homo_sapiens/1-464308 CCTCAGGACCGACGGCAACCAACACAGAT CCAGTGCCCTTCGACTGCCCTCTGGGCC TGGGACAGAGAGAGAACACAGCTGGCT AGGGCCCTGTTGGGGGTTAGATCAAA CCGAGTGATCTGATATGGGCACCT CCAGGCTCTGTGCAMAAAGTTGCTGTG AGGAGACCTGGTGCTCTGCTGTGT AAGATGGGGTGGGTGGATTCCTCT GGGAGAGGAAGAAGAGGCCCTGGG </pre>	<pre> #mega !Title: ProjectedMultiAlign; !Format datatype=dna Identical=. #homo_sap CCTCAGGACC GACGGCAAC #pan_trog ..C..... #homo_sap CCGAGTGCCCT TCGACTGCCT #pan_trog #homo_sap TGGGACAGAG AGAGAACAC </pre>	<pre> ProjectedMultiAlign MSF: 2 Type: Name: homo_sapiens/1-465588 Len Name: pan_troglodytes/1-465588 Len // homo_sapiens/1-465588 CCTCAGGAC G pan_troglodytes/1-465588 CCCCAGGAC </pre>

We can look at homologues in the *Orthologues* and *Paralogues* pages, which can be accessed from the left-hand menu. If there are no orthologues or paralogues, then the name will be greyed out. *Paralogues* is greyed out for *BRCA2* indicating that there are no paralogues available. Click on *Orthologues* to see the 175 orthologues available.

Show All entries		Show/hide columns	Filter				
Species	Type	Orthologue	Target %id	Query %id	GOC Score	WGA Coverage	High Confidence
Algerian mouse (<i>Mus spretus</i>)	1-to-1 View Gene Tree	Brca2 (MGP_SPRETEU_G0029058) Compare Regions (5:151,625,243-151,672,752:1) View Sequence Alignments	58.86 %	57.31 %	100	99.35	Yes
Alpine marmot (<i>Marmota marmota marmota</i>)	1-to-1 View Gene Tree	BRCA2 (ENSMMSG000000022103) Compare Regions (CZRN01000014.1:27,378,250-27,449,785:1)	69.97 %	69.87 %	100	100.00	Yes

Choose to see only *Rodents and related species* orthologues by selecting the box. The table below will now only show details of rodent orthologues. Let's look at mouse.

Mouse (<i>Mus musculus</i>)	1-to-1	Brca2 (ENSMUSG00000041147)	58.58 %	57.05 %	100	0.00	Yes
	View Gene Tree	Compare Regions (5:150,446,095-150,493,794:1)					
		View Sequence Alignments					

Links from the orthologue allow you to go to alignments of the orthologous proteins and cDNAs. Click on *View Sequence Alignments* then *View Protein Alignment* for the mouse orthologue.

Type: 1-to-1 orthologues

Species	Gene ID	Peptide ID	Peptide length	% identity (Protein)	% coverage	Genomic location
Human (<i>Homo sapiens</i>)	ENSG00000139618	ENSP00000369497	3418 aa	57 %	95 %	13:32315086-32400268
Mouse (<i>Mus musculus</i>)	ENSMUSG00000041147	ENSMUSP00000038576	3329 aa	58 %	98 %	5:150446095-150493794

CLUSTAL W (1.81) multiple sequence alignment

ENSP00000369497/1-3418 MPIGSKERPTTFEIFKTRCNKADLGPISLWNFELSSEAPPYNSEPAEESKHNNYEPN
ENSMUSP00000038576/1-3329 MPVEYKRRPTFWEIFKARCSADLGPISLWNFELSSEAPPYNSEPPSEYKPHGYEPQ
*: *:****:****:*.*****:****:*.***

```
ENSP00000369497/1-3418 LFKTPQRKPSYNQLASTPIIFKEQGLTLPLYQSPVKELDKFKLDLGRNVPNSRHKSRLTV
ENSMU00000038576/1-3329 LFKTPQRNPYPHQFASTPIMFKERSQTLPLDQSPFREL-----GKVVASSKHHTSKK
*****:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:*:
```

Protein IDs

ENSP00000369497/1-3418	KTQMDQADDVSCPLLNSCLSESPVVLQCTHVTQQRDKSVVCGSLFHTPKFVKGRQTPKHI
ENSMUSP00000038576/1-3329	KTQKDPVVDVASPPLKSLSESPLTLRCTQAVLQREKPVVSGSLFYTPKPKLE-GOTPKPI

ENSP00000369497/1-3418 SESLGAEVDPDMSWSSSLATPPTLSSTVLIVRNEEASETVFFHDTTANVKSYSFSNHDESL
 ENSMUSP00000038576/1-3329 SESLGEVDPDMSWSSSLATPPTLSSTVLIVRNEEASETVFFHDTTANVKSYSFSNHDESL

Information on
orthologue pair

Protein IDs

Alignment in Clustal W format

Whole genome alignments, Demo

Let's look at some of the comparative genomics views in the Location tab. Go to the region 2:176087000-176202000 in human, which contains the *HoxD* cluster which is involved in limb development and is highly conserved between species.

You can turn on conservation scores and constrained elements. Click on *Configure this page*, then *Comparative genomics* and turn on the tracks for *Constrained elements for 91 eutherian mammals EPO-Extended* and *Conservation score for 91 eutherian mammals EPO-Extended*. Save and close the menu.



You can now see the conservation scores in pale pink. These were used to determine the peaks indicated in the constrained elements track in dark pink. This track indicates regions of high conservation between species, considered to be “constrained” by evolution.

We can also look at individual species comparative genomics tracks in this view by clicking on *Configure this page*.

Select *BLASTz/LASTz* alignments from the left-hand menu to choose alignments between closely related species. Turn on the alignments for *Mouse*, *Chicken* and *Chimpanzee* in *Normal*. Save and close the menu.



The alignment is greatest between closely related species.

We can also look at the alignment between species or groups of species as text. Click on *Alignments (text)* in the left hand menu.

Select *Select an alignment* to open the alignment menu.

Select another alignment

Alignments Selector

Start typing the name of a species...

All Alignments Pairwise Rodents & Lagomorphs

Rats & Mice Lagomorphs Other Rodents

Rats & Mice

- ☒ Mouse reference (CL57BL6)
- ☐ Algerian mouse
- ☐ Rat
- ☐ Ryukyu mouse
- ☐ Shrew mouse
- ☐ Steppe mouse

Selected species

Mouse reference (CL57BL6) X

Reset All Cancel Apply

Click through the links, *Pairwise*, *Rodents & Lagomorphs*, *Rats and Mice* to select *Mouse reference (CL57BL6)*.

In this case there are two blocks aligned, Block 1 a large (115001 bp) alignment against mouse chr2 and one smaller block against mouse chr7. Click on *Block 1*.

You will see a list of the regions aligned, followed by the sequence alignment. Click on *Display full alignment*. Exons are shown in red.

To compare with both contigs visually, go to *Region comparison*.

To add species to this view, click on the blue *Select species or regions* button. Choose *Mouse Reference* again then close the menu.



You can configure this view for both species. Click on *Configure this page* and look in the top left of the menu.

Species to configure:

✓ Human
Mouse

The drop down allows you to configure each species separately.

We can view large scale syntenic regions from our chromosome of interest. Click on *Synteny* in the left hand menu.

Synteny

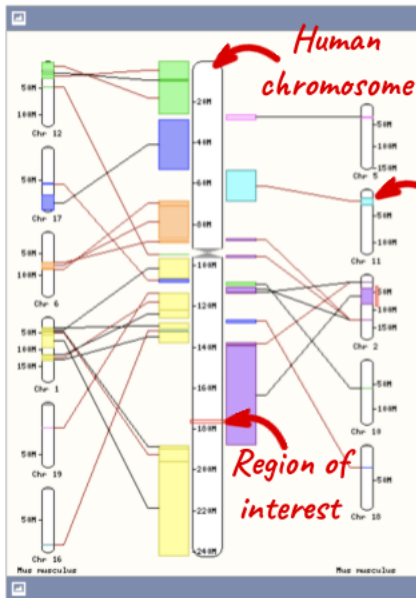
Synteny between Human chromosome 2 and Mouse

Change Species:

Mouse Go

Change chromosome:

2 Go

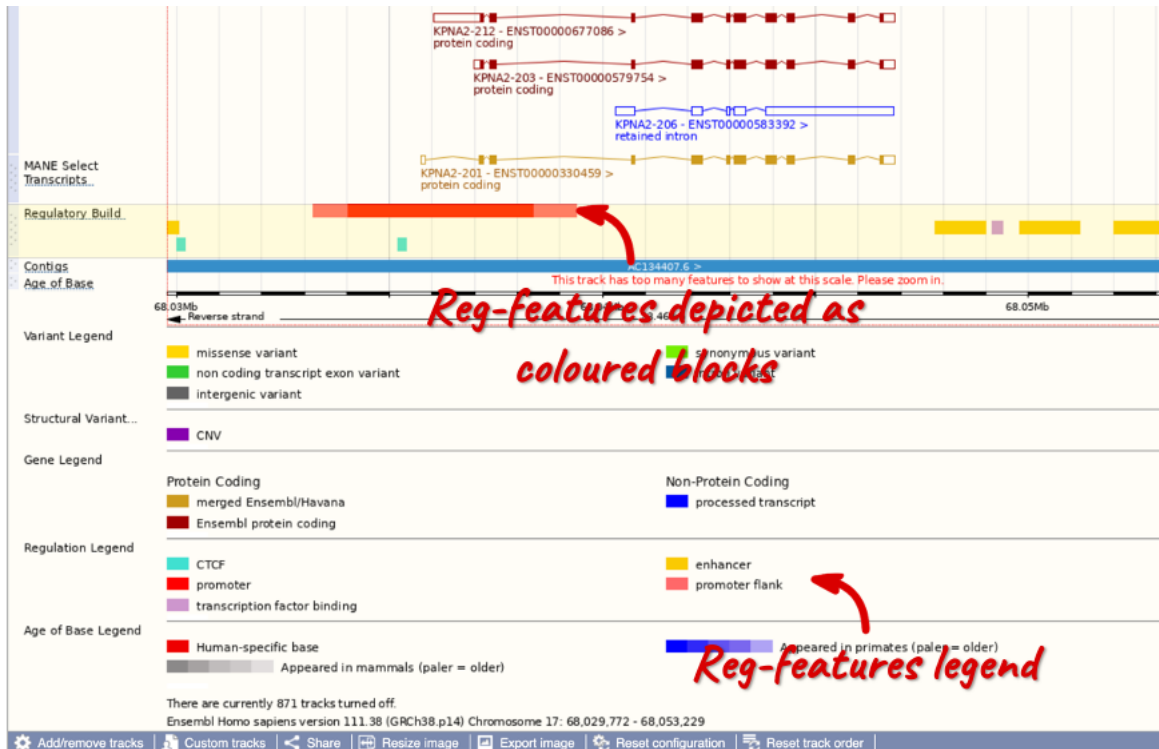


Choose another species or chromosome

Features that regulate gene expression, Demo

We're going to look for regulatory features in the region of a gene and investigate their activity in different cell types. We'll start by searching for the gene *KPNA2* and jumping to the **Location** tab. Scroll down to the **Region in detail** view and zoom out a little to see the gene as well as its flanking regions.

The **Regulatory Build** track is shown by default.



In this region we can see a number of regulatory features, including a red promoter with light red promoter flanks, cyan CTCF binding sites, yellow enhancers and lilac transcription factor (TF) binding sites (don't worry if you have zoomed out further or not as far and can see more/less). Refer to the legend at the bottom of the view to see what each of the colours mean.

You can also click on the individual regulatory features to learn more. Click on the red promoter to open a pop-up menu.

Regulatory Feature
ID: [ENSR00000097453](#)
Type: Promoter
Core bp: [Chromosome 17: 68,034,000-68,038,401](#)
Bounds bp: [Chromosome 17: 68,033,202-68,039,399](#)

Click on the stable ID, *ENSR00000097453*, to jump to the **Regulation** tab.

Human (GRCh38.p14) ▼

Location: 17:68,029,772-68,053,229 Regulation: ENSR0000097453 Jobs ▼

Regulation-based displays

- Activity
- Source Data

Configure this page

Custom tracks

Export data

Share this page

Bookmark this page

Regulatory Feature: ENSR0000097453

Activity

Source Data

Available views

List of regulatory activity in different cell types

Classification

To view the configured tracks visit:

Location [Chromosome 17: 68,034,000-68,038,401](#)

Bound region [Chromosome 17: 68,033,202-68,039,399](#)

Activity ⓘ

Summary of Regulatory Activity [Hide](#) ⓘ

Active (1/118)	Inactive (64/118)	Poised (17/118)	Repressed (36/118)
astrocyte	A549 aorta B B (PB) cardiac muscle CD4+ ab T (PB) CD4+ ab T (PB)_2 CD4+ ab T (Th) CD4+ ab T (VB) CD8+ ab T (PB) CM CD4+ ab T (VB) dermal fibroblast	CD4+ ab T (PB)_1 CMP CD4+_2 CMP CD4+_3 foreskin fibroblast_2 germinal matrix H1-hESC_2 H1-hESC_3 HUES48 HUES6 HUES64 IPS DF 19.11 IPS-15b	A673 adrenal gland bipolar neuron brain_1 CD14+ monocyte (PB) CD14+ monocyte_1 CD4+ CD25+ ab Treg (PB) CD8+ ab T (CB) CMP CD4+_1 EB (CB) EM CD8+ ab T (VB) endodermal

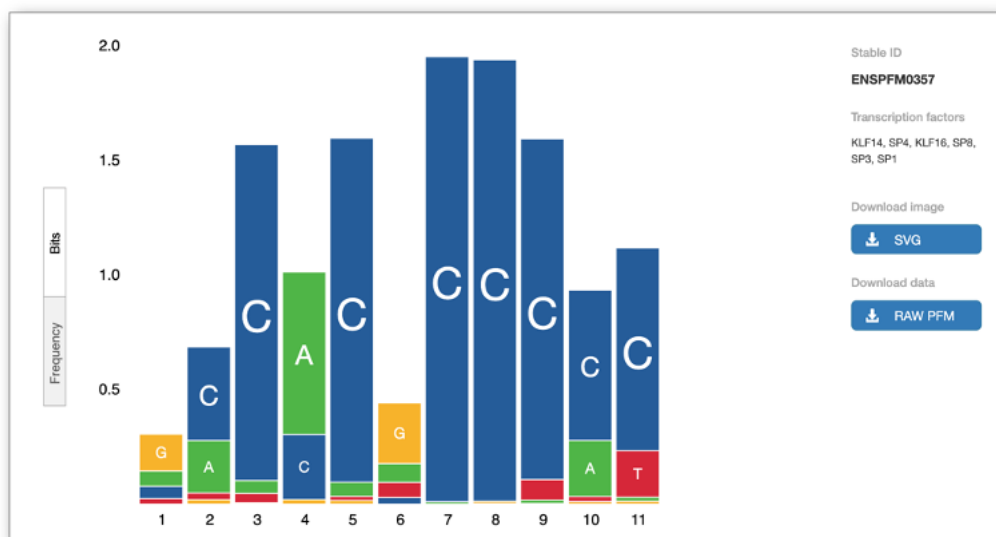
Here, you can find a summary of the activity of the promoter in the different cell types. Scroll down to **Summary of Regulatory Aactivity** to find out in which cells the promoter is active (the feature displays an active epigenetic signature, which can include evidence of open chromatin), inactive (the region bears no epigenetic modifications from the ones included in the Regulatory Build), poised (the feature displays a epigenetic signature with the potential to be activated) or repressed (the feature is epigenetically suppressed). We can see that this promoter is active in one out of the 118 cell types currently in Ensembl.

Let's switch back to the **Location** tab to explore the different regulation tracks that are available. Click on **Configure this page** and in the pop-up window under the **Regulation** section, click on **Other regulatory regions** and enable the **Fantom 5**, **TarBase** and **Motif features** tracks. Close the pop-up window.

The **Fantom 5** track displays transcription start site (TSS) and enhancer predictions from the [FANTOM5 project](#).

The **TarBase** track displays experimentally verified miRNA targets from TarBase.

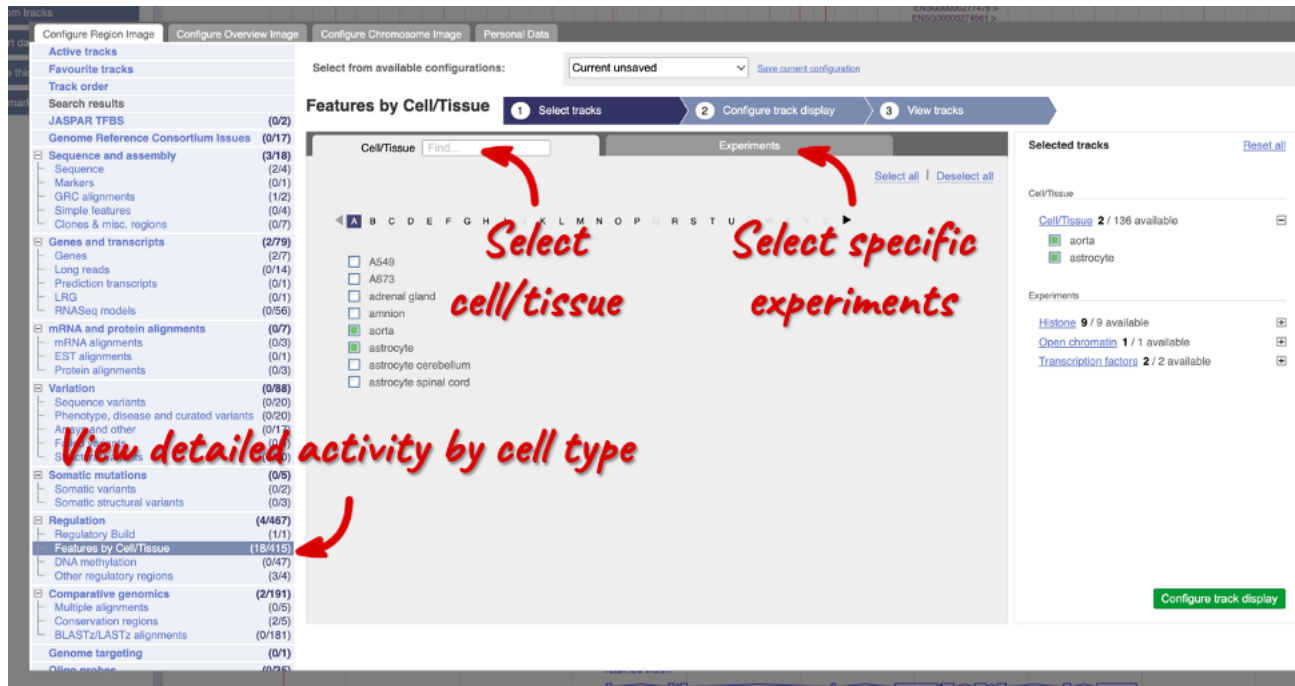
The **Motif features** track indicates the positions of transcription factor binding motifs (TFBMs) in black lines/blocks. You can click on individual features to find out more information about the TFBM, including a list of TFs binding at this site and, if available, in which cells the TFBM was experimentally verified in. You can also view the **Binding matrix**** by clicking on the matrix ID. This opens a pop-up window which displays the binding matrix used and a binding score representing how well a particular site matches the binding matrix.



We can explore more detailed data by adding further **Regulation** tracks. Click on the **Configure this page** button on the left-hand side.

 Configure this page

In the pop-up window, go to **Regulation** and click on **Features by Cell/Tissue** to view the detailed activity of the regulatory feature by cell type.

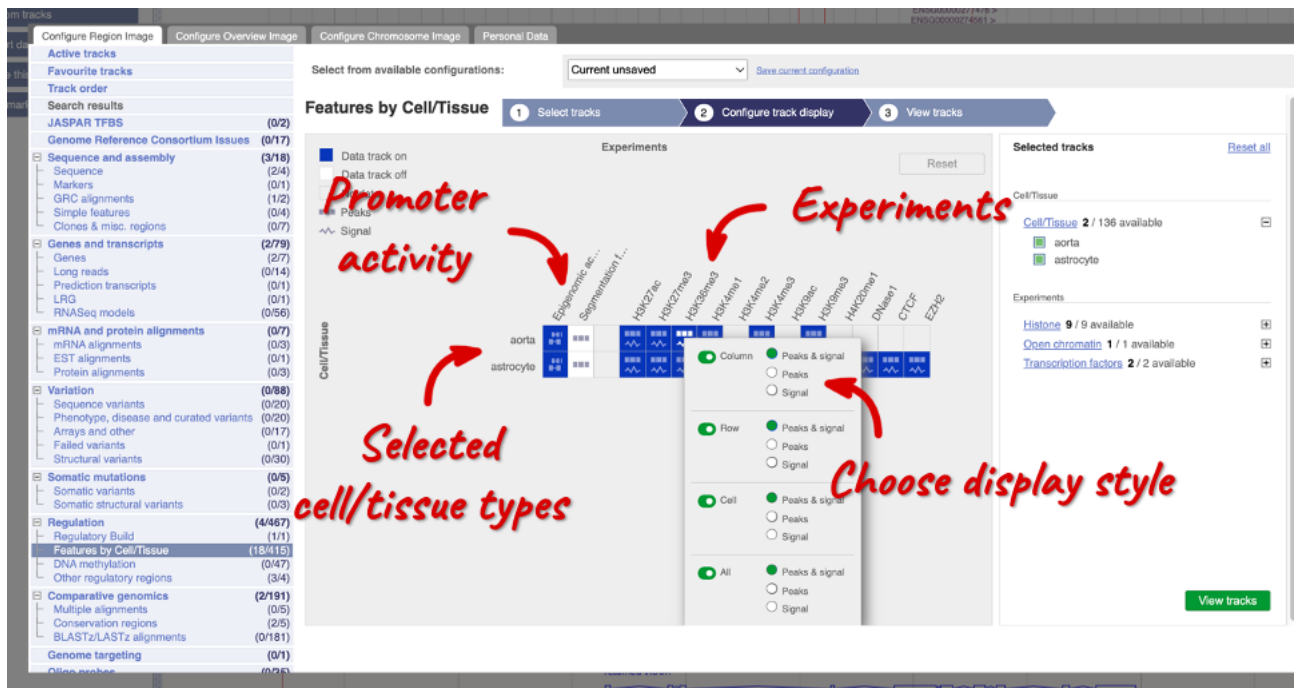


The screenshot shows the 'Features by Cell/Tissue' configuration window. On the left is a sidebar with a tree view of genomic features. The 'Regulation' section is expanded, and 'Features by Cell/Tissue' is selected. The main panel has three tabs: 'Cell/Tissue', 'Experiments', and 'View tracks'. The 'Cell/Tissue' tab is active, showing a search bar and a list of cell types with checkboxes. Red arrows point to the search bar with the text 'Select cell/tissue', to the 'Experiments' tab with 'Select specific experiments', and to the 'Features by Cell/Tissue' item in the sidebar with 'View detailed activity by cell type'. On the right, a 'Selected tracks' panel shows the chosen cell types 'aorta' and 'astrocyte'. A green 'Configure track display' button is at the bottom right.

We can add cells by clicking on them. Find them using the search or the alphabet ribbon. Let's add a cell type where the promoter is inactive, **aorta**, and one where it's active **astrocytes**. Once you've selected the cells, they will appear in the menu on the right, where you can easily view the list by clicking on the + icon and de-select them.

To choose the experiments to see data on, click on the **Experiments** tab at the top of the menu. You can navigate this the same as the **Cell/Tissue** tab, except that you have to choose between **Histone**, **Open Chromatin** and **Transcription factors**. Let's **Select all** in all categories.

When you've chosen your experiments and cells, you can click on the green **Configure track display** button in the bottom right-hand corner.



Now we can see the active feature in astrocytes compared to the inactive feature in aorta.



Bulk export of data with BioMart, Demo

Follow these instructions to guide you through BioMart to answer the following query:

You have three questions about a set of human genes:

ESPN, MYH9, USH1C, CISD2, THRB, WHRN

(these are HGNC gene symbols. More details on the HUGO Gene Nomenclature Committee can be found on <http://www.genenames.org>)

1. What are the NCBI Gene IDs for these genes?
2. Are there associated functions from the GO (gene ontology) project that might help describe their function?
3. What are their cDNA sequences?

Click on BioMart in the top header of a www.ensembl.org page to go to: www.ensembl.org/biomart/martview

You cannot choose any filters or attributes until you've chosen your dataset. Your dataset is the data type you're working with. In this case we're going to choose human genes, so pick *Ensembl Genes* then *Human genes* from the drop-downs.

The first screenshot shows the 'Dataset' dropdown menu with the following options: '- CHOOSE DATABASE -', 'Ensembl Genes 104' (highlighted), 'Mouse strains 104', 'Ensembl Variation 104', and 'Ensembl Regulation 104'. The second screenshot shows the 'Dataset' dropdown menu with the following options: 'Ensembl Genes 104' (selected), '- CHOOSE DATASET -', 'Chicken genes (GRCg6a)', 'Human genes (GRCh38.p13)' (highlighted), and 'Mouse genes (GRCm39)'.

Now that you've chosen your dataset, the filters and attributes will appear in the column on the left. You can pick these in any order and the options you pick will appear.

Click on *Filters* on the left to see the available filters appear on the main page. You'll see that there are loads of categories of Filters to choose from. You can expand these by clicking on them. For our query, we're going to expand *GENE*.

The screenshot shows the 'Filters' section on the left with 'Expand Filters' written in red. The main area shows the 'GENE' filter expanded, with 'Expand GENE' written in red. The 'GENE' filter has two options: 'Limit to genes with external references...' (selected) and 'Input external references list (e.g. NCBI Gene IDs)'. The 'Limit to genes with external references...' option has a dropdown menu with 'Only' selected and 'Excluded' as an option. The 'Input external references list' option has a text input field with 'Gene stable ID(s) (e.g. ENSG00000000003)' as an example.

Our input data is a list of identifiers, so we're going to use the *Input external references ID list* filter. This allows us to input a list of identifiers from different databases. We need to choose what kind of identifier we're using, so that BioMart can look up the right column in a data table. You can pick these from a drop-down list, which lists the type of identifier with an example of how it looks. For our query, we have a list of gene names, so we need to pick *Gene Name(s)*.

Choose ID type

To check if the filters have worked, you can use the *Count* button at the top left, which will show you how many genes have passed the filter. If you get 0 or another number you don't expect, this can help you to see if your query was effective.

Dataset 6 / 67128 Genes
Human genes (GRCh38.p13)

To choose the attributes, expand this in the menu. There are six categories for human gene attributes. These categories are mutually exclusive, you cannot pick attributes from multiple categories. This means that we need to do two separate queries to get our GO terms and NCBI IDs, and to get our cDNA sequences.

Features
Structures
Homologues (Max select 6 orthologues)
Variant (Germline)
Variant (Somatic)
Sequences

The Ensembl gene and transcript IDs, with and without version numbers are selected by default. The selected attributes are also listed on the left.

Selected attributes are selected by default

We can choose the attributes we want by clicking on them. For our query, we're going to select:

- GENE
 - Gene Name
- EXTERNAL
 - NCBI gene ID
 - GO term accession
 - GO term name
 - GO term definition

We need to select the *Gene Name* in order to get back our original input, as this is not returned by default in BioMart. The order that you select the attributes in will define the order that the columns appear in in your output table.

You can get your results by clicking on *Results* at the top left.

Dataset 6 / 67128 Genes
Human genes (GRCh38.p13)

The results table just gives you a preview of the first ten lines of your query. This allows the results to load quickly, so that if you need to make any changes to your query, you don't waste any time. To see the full table you can click on *View ## rows*. You can also export the data to an xls, tsv, csv or html file. For large queries, it is recommended that you export your data as *Compressed web file (notify by email)*, to ensure your download is not disrupted by connection issues.

Export all results to ☐ Unique results only

Email notification to

View rows as

Gene stable ID	Transcript stable ID	Gene name	NCBI gene (formerly Entrezgene) ID	GO term accession	GO term name	GO term definition
ENSG00000151090	ENST00000216181	MYH9	4627	GO:0005524	ATP binding	Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme and enzyme regulator.
ENSG00000100345	ENST00000216181	MYH9	4627	GO:0005515	protein binding	Interacting selectively and non-covalently with any protein or protein complex (a complex of two or more proteins that may include other nonprotein molecules).
ENSG00000100345	ENST00000216181	MYH9	4627	GO:0051015	actin filament binding	Interacting selectively and non-covalently with an actin filament, also known as F-actin, a helical filamentous polymer of globular G-actin subunits.
ENSG00000100345	ENST00000216181	MYH9	4627	GO:0003774	motor activity	Catalysis of the generation of force resulting either in movement along a microfilament or microtubule, or in torque resulting in membrane scission, coupled to the hydrolysis of a nucleoside triphosphate.

You can see multiple rows per gene in your input list, because there are multiple transcripts per gene and multiple GO terms per transcript.

To get the cDNA sequences, go back to the *Attributes* then select the category *Sequences* and expand *SEQUENCES*.

When you select the sequence type, the part of the transcript model you've chosen will be highlighted in the graphic.

SEQUENCES:

Sequences (max 1)

☐ Unspliced (Transcript)
☐ Unspliced (Gene)
☐ Flank (Transcript)
☐ Flank (Gene)
☐ Flank-coding region (Transcript)
☐ Flank-coding region (Gene)

☐ 5' UTR
☐ 3' UTR
☐ Exon sequences
☐ cDNA sequences
☐ Coding sequence
☒ Peptide

Choose *cDNA sequences*, then expand *HEADER INFORMATION* to add *Gene Name* to the header. Then hit *Results* again.

View rows as ☐ Unique results only

```
>ENSG00000151090|ENSG00000151090.20|ENST00000416420|ENST00000416420.5|THRB
AACTTTGCATGAATAATGTGAGTGCGCTTGGAAAAGAGACCTCCTGCTCCGCGGGCTCGG
GGCAAGAGCCCGCAGGCTACCTTCCCGGGCAGGGCGCTCAACCAACCGGCTCCAGGG
CACTGGTAATTTGGCTAGAGGACCGCGCGGAGGCGAGCGGATCTGCGATTCTCTTCTGGT
TGGCTGCTCGCTGGGTGCCAAGTTCACACATGATTAAATGAATAAGAGGGGCTTG
GAGGCTCATAGCAAGGTCTTGAAGAGAGCAAGAAAAGTCTTCCAGGAAAAGTGGAC
ACCTGAGTGAGTGTTACAACGAGAAGGGGGCATCCAGGCAAGGCAGCTACCAAGCACA
GGTGTGAAGTGACTTCACTGTAGGAAGAGGAAATGCAGAAAAACAACTCTTGAAGACT
GAAAGCCTGTATTCAAAGAGGTAGCAGTAGACTGTTGGAAGTAAGGAGATGTCAGTGA
AAAAAGGATCCAGAATGATTACTAACCCTATGACTCCCAACAGTATGACAGAAAATGGCC
TTACAGCCTGGGCAAAACCGAAGCACTGTCCAGACCGAGAACACGACTGGAAGCTAGTAG
GAATGCTGAAGCCTGCCCTACATAGGAAGAGCCATTGAGAGAGGCGCAGCACGTTGAAAA
ATGAACAGTCGTCGCCACATCTCATCCAGACCACTTGGACTAGCTCAATATTCATCTGG
```

For more details on BioMart, have a look at this publication:

Kinsella, R.J. et al

Ensembl BioMart: a hub for data retrieval across taxonomic space.

<http://europepmc.org/articles/PMC3170168>

Custom data, Demo

Demo: Upload small files

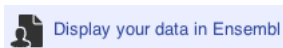
We have some patients that present with microcephaly and developmental delay. They all have large scale deletions on chromosome five:

Patient	Chromosome	Start	End
P1	5	36821632	37091234
P2	5	36731476	36978306
P3	5	36908552	37108671

We can turn them into a BED file and view them in the genome browser:

```
chr5 36821632 37091234 P1
chr5 36731476 36978306 P2
chr5 36908552 37108671 P3
```

You can add data from a Region in Detail page by clicking on the *Custom tracks* button at the left. Alternatively, go to a species homepage and click on *Display your data in Ensembl*.



A menu will appear:

Add a custom track

Please note that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services, including **Google Drive** and **Dropbox**. Please see our [support page](#) for more information.

Name for this data (optional):

Species:

The species is human

Human (*Homo sapiens*)

Assembly: GRCh38

Data:

Paste in data or provide a file URL

Paste in your data

Or upload file (max 20MB)

Choose File

no file selected

Data format:

[Help on supported formats, display types, etc](#)

Add data

The interface detects file types if you upload or attach a file. When you paste in your data, it can't do this so we have to tell it what our file type is. It will give you an option where you can select *BED*.

Click *Add data*.

You should get to a dialogue box telling you your upload has been successful.

Thank you. Your file uploaded successfully

File uploaded: BED demo (Bed file, *Homo sapiens*)

Total features found: 4

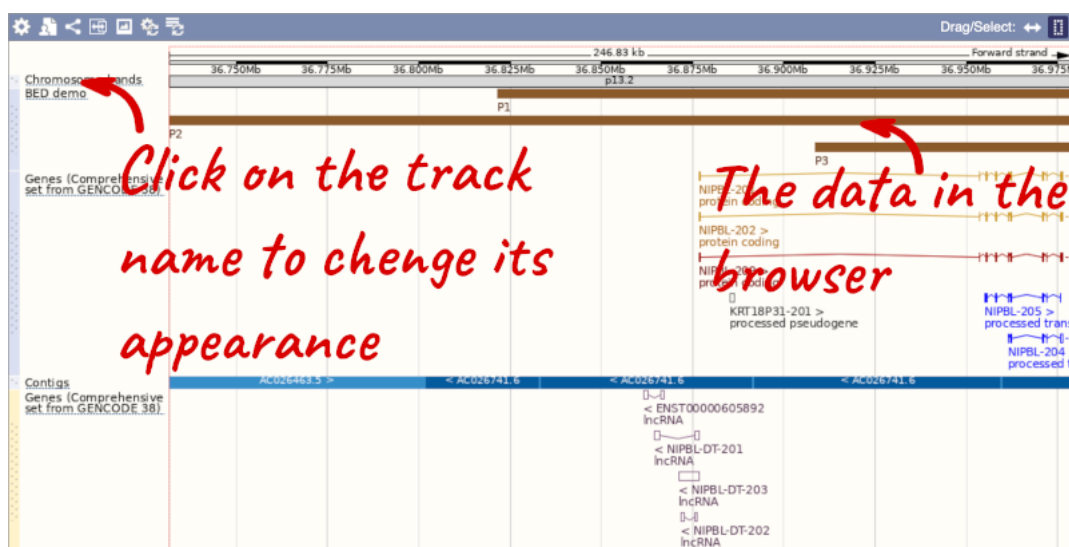
Go to Location view:

- Nearest region with data: [5:36731477-36978306](#)

or

Close this window to return to current page

Click on the genomic coordinates link to go to the nearest region with data.



To have a look at the file, click on *Custom tracks*.

Your data ⓘ

Update selected: [Connect](#) [Disconnect](#) [Delete](#)

Filter						
Select	Type	Source	Species	Assembly	Last updated	Actions
☐	Upload	BED demo View sample location Bed file	<i>Homo sapiens</i>	GRCh38	17/11/21 at 14:40	   

Disconnect, save, share

 Add more data

 Search for public track hub

or delete this data

If you've got an Ensembl account, you can save this data to your account. Accounts are free to set up and allow you to save configurations and data, and share with groups.

Demo: Attach URLs of large files

Larger files, such as BAM files generated by NGS, need to be attached by URL. I've put a BAM file of human chromosome 20 RNASeq data online at: http://ftp.ebi.ac.uk/pub/databases/ensembl/training/emily_BAM/

Let's take a look at the folder.

Index of /pub/databases/ensembl/training/emily_BAM

Name	Last modified	Size	Description
 Parent Directory			-
 GRCh37_Illumina_reads_test.bam	2019-01-16 10:01	4.3M	
 GRCh37_Illumina_reads_test.bam.bai	2019-01-17 14:10	176K	
 GRCh38.20.illumina.merged.1.bam	2019-01-17 14:09	2.8G	
 GRCh38.20.illumina.merged.1.bam.bai	2019-01-17 13:55	169K	
 GRCh38.21.illumina.merged.1.bam	2019-01-17 14:22	2.9G	
 GRCh38.21.illumina.merged.1.bam.bai	2019-01-17 14:11	121K	
 cat_e2.bam	2019-01-16 10:01	5.8M	
 cat_e2.bam.bai	2019-01-16 09:17	131K	

Apache Server at ftp.ebi.ac.uk Port 80

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.

To attach the file, click on *Custom tracks*, then click on *Add more data* to add a new track.

We get to the same dialogue box as before. This time we'll name our data *Illumina reads*.

Paste in the URL of the BAM file itself

(http://ftp.ebi.ac.uk/pub/databases/ensembl/training/emily_BAM/GRCh38.20.illumina.merged.1.bam).

Add a custom track

Please note that track hubs and indexed files (BAM, BigBed, etc) do not work with certain cloud services, including **Google Drive** and **Dropbox**. Please see our [support page](#) for more information.

Name for this data (optional):

BAM demo

Species:

Human (*Homo sapiens*)

Assembly: GRCh38

Data:

http://ftp.ebi.ac.uk/pub/databases/ensembl/training/emily_BAM/GRCh38.20.illumina.merged.1.bam

Or upload file (max 20MB)

Choose file

No file chosen

Data format:

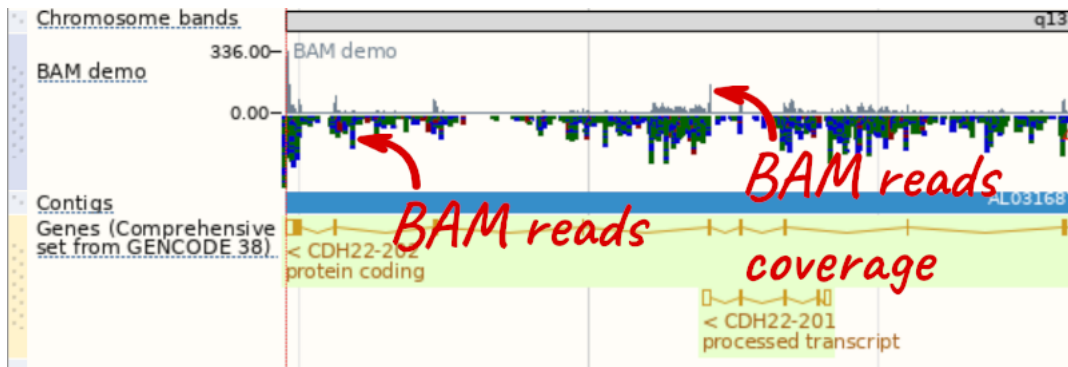
BAM

[Help on supported formats, display types, etc](#)

Add data

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.

To see this data, jump to a region on chromosome 20. Let's go to the region of the CDH22 gene. Search for the gene and click on the location.



We can zoom in to see the sequence itself. Drag out boxes in the view to zoom in, until you see a view like this. Alternatively, jump to location 20:46241000-46241030.

Demo: Track hub registry

Our regulatory data incorporates data from sources such as ENCODE, Blueprint, and Roadmap Epigenomics. To see the data directly from these sources, you can add track hubs.

You can search for track hubs to add in different ways:

- Search for track hubs in the [Track Hub Registry](#) and choose to add them to your genome browser of choice.
- Search the track hub registry using the Track Hub Registry interface in Ensembl (there is a link from the [homepage](#)).

We will now add the track hub containing data from the Blueprint project.

You can add track hubs to view in Ensembl directly via the Track Hub Registry. Go to the [Track Hub Registry homepage](#) and search for *blueprint*.

There are two results for the Blueprint Hub, one for adding the track hub to GRCh37 and one for adding it to GRCh38, plus one RNA-seq alignment hub.

Search Results

Track Collections 1 to 3 of 3

Blueprint Hub
Blueprint Epigenomics Data Hub

Species: 9606 - *Homo sapiens*
Assembly: GCA_000001405.15 - GRCh38

View in Genome Browser View Info

Blueprint Hub
Blueprint Epigenomics Data Hub

Species: 9606 - *Homo sapiens*
Assembly: GCA_000001405.1 - GRCh37

View in Genome Browser View Info

Remote Data Available

RNASeq-er alignment hub for ENA runs in ERP010718

A blueprint for C4 photosynthesis in the rice paddy: insights from the noxious weed *Echinochloa glabrescens*; [ERP010718](#)

Species: 39947 - *Oryza sativa Japonica Group*
Assembly: GCA_001433935.1 - IRGSP-1.0

View in Genome Browser View Info

Unchecked

Choose the genome browser to view the data in

Alternatively, you can add track hubs by searching the Track Hub Registry through Ensembl. Click the *Custom tracks* -> *Track Hub Registry Search* in any region view within Ensembl.

Configure Region Image Configure Overview Image Configure Chromosome Image Personal Data

Custom tracks

Track Hub Registry Search Search the Track Hub Registry

Manage Configurations

Select Track Hub Registry search on the left

Species: Human (Homo sapiens)
Assembly: GRCh38
Track type: -- all --

Text search:

Hint: Leave "text search" empty to show all track hubs for this species

Search

Search box

You can only find track hubs for the selected species and assembly denoted in the search box.

Search for *blueprint*.

Click *Attach this hub* in the search results page.

Configure Region Image Configure Overview Image Configure Chromosome Image Personal Data

Custom tracks

Track Hub Registry Search Search Results

Manage Configurations

Searched Human GRCh38 for "blueprint"

Found 1 track hub - [Search again](#)

Blueprint Hub

Description: Blueprint Epigenomics Data Hub

Data type: genomics

Number of tracks: 5698

Attach this hub

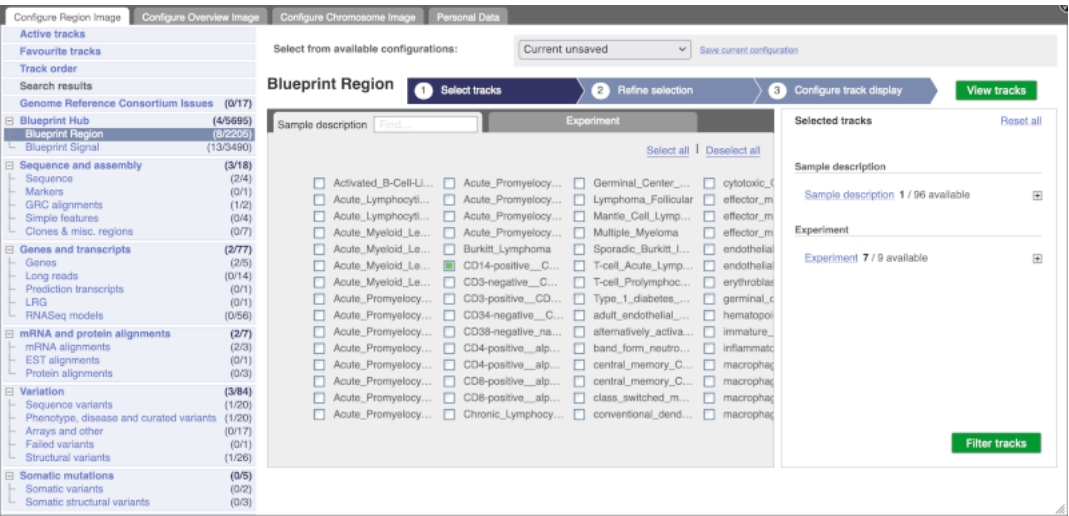
Can't see the track hub you're interested in?

We only search for hubs compatible with assemblies used on this website - please [search the registry directly](#) for data on other assemblies.

Alternatively, you can [manually attach any hub](#) for which you know the URL.

Track Hubs often contain vast amounts of data, which can slow Ensembl down, so only add them if you need them, and trash them when you are finished with them.

Go to *Configure this Page* to see that a new category has been added to your menu.



You can add tracks to the Region in Detail view using the matrix.

