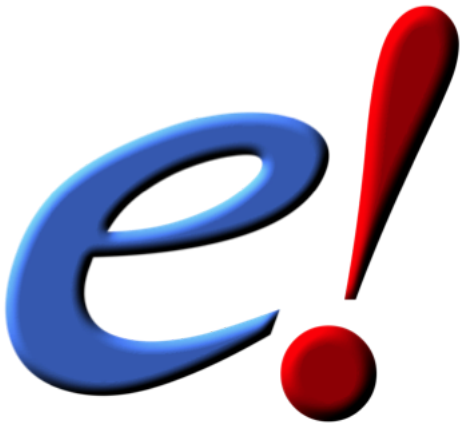
Browsing Genomes with Ensembl



www.ensembl.org

www.ensemblgenomes.org

CC BY

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

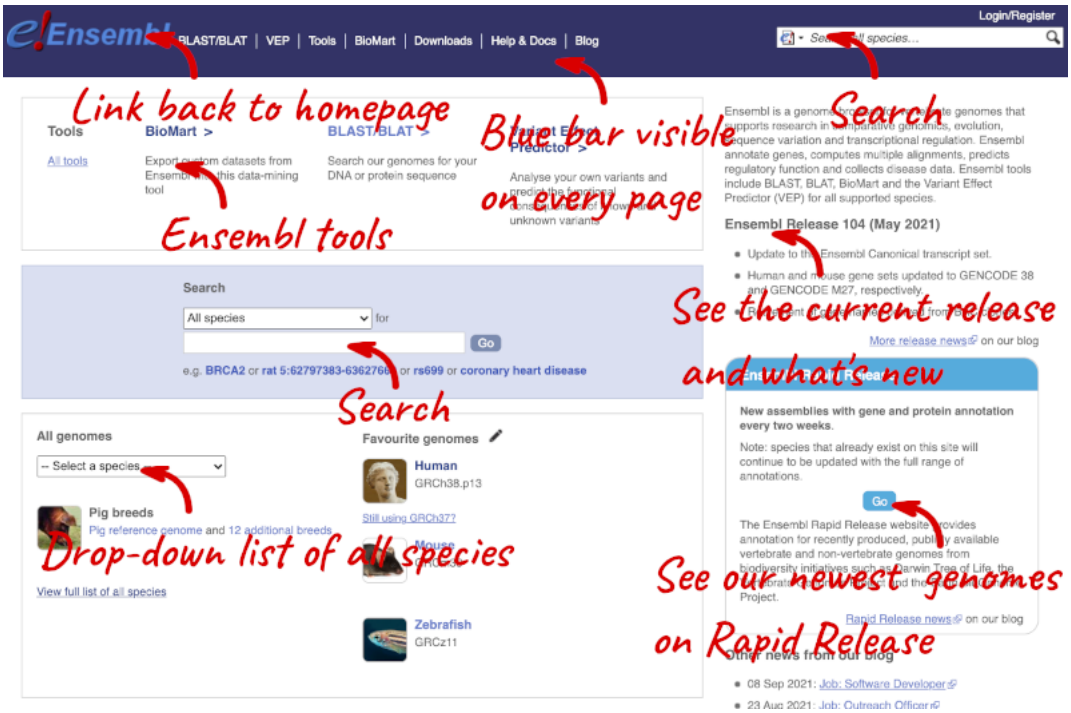
Finding information about species and genomes in Ensembl, Demo

Demo: Introduction to Ensembl

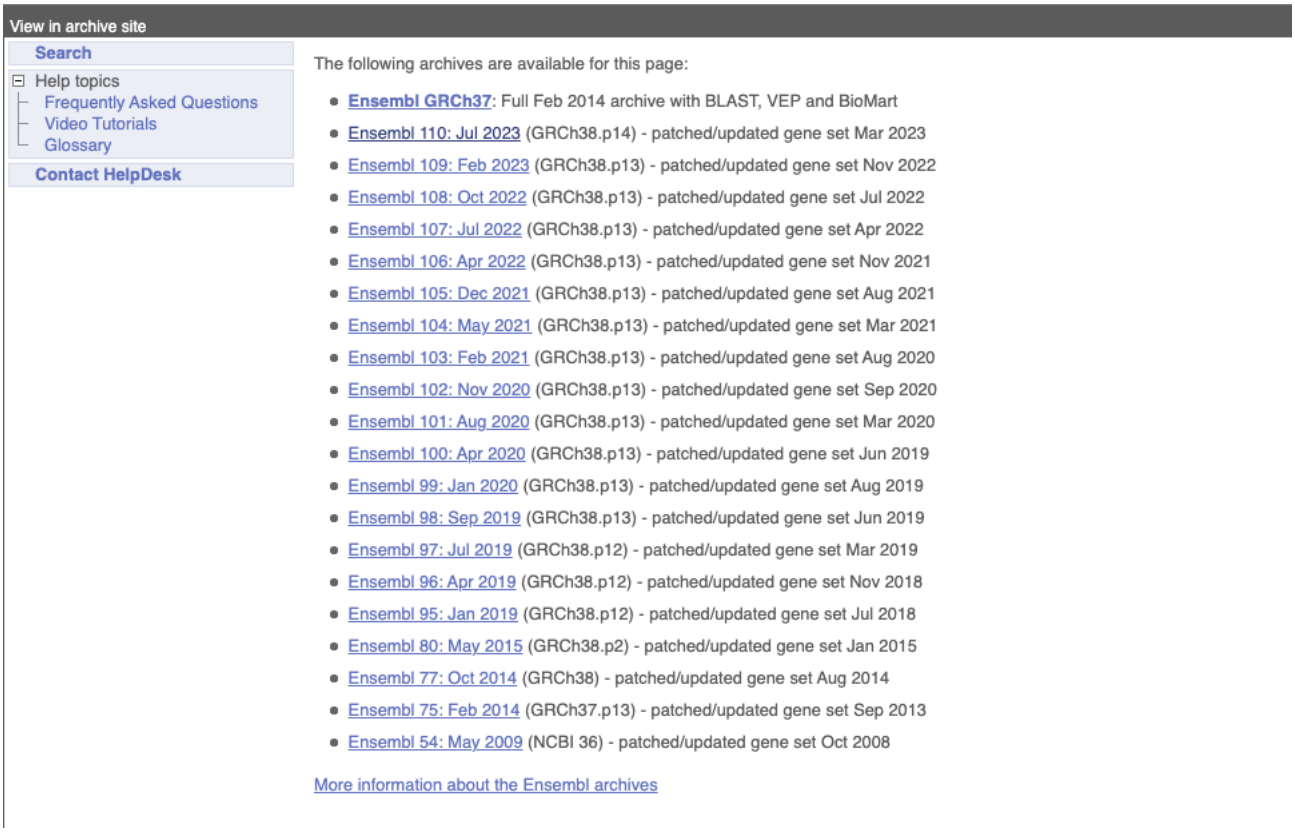
Ensembl

Homepage

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



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.

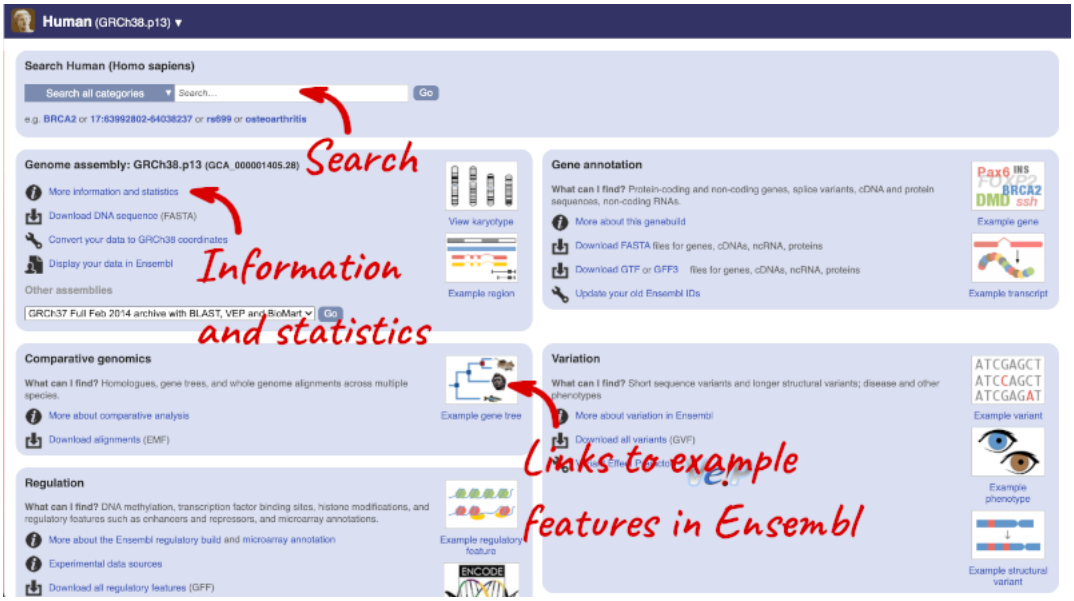


Click on the links to go to the archives. Alternatively, you can jump quickly to the correct release by adding **e** plus the **release number** in the URL. For example e98.ensembl.org jumps to Ensembl release 98.

Available species

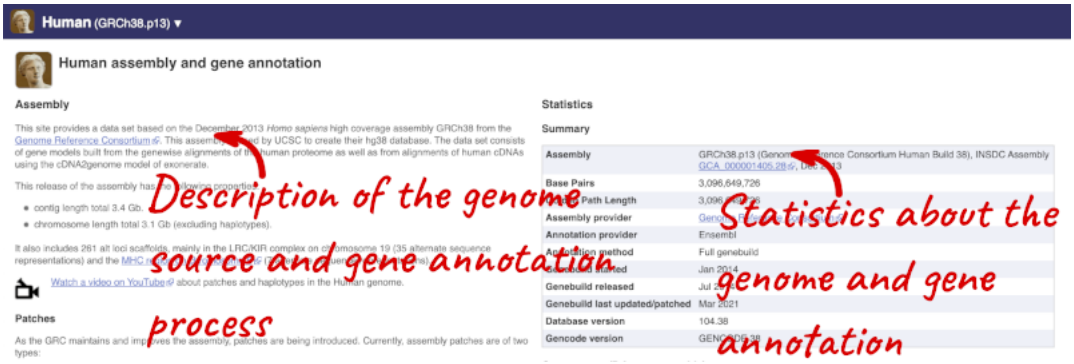
Scroll back up to the top of the homepage. You can view all available species by clicking the **View full list of all species** link underneath the coloured search block.

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**.



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.



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

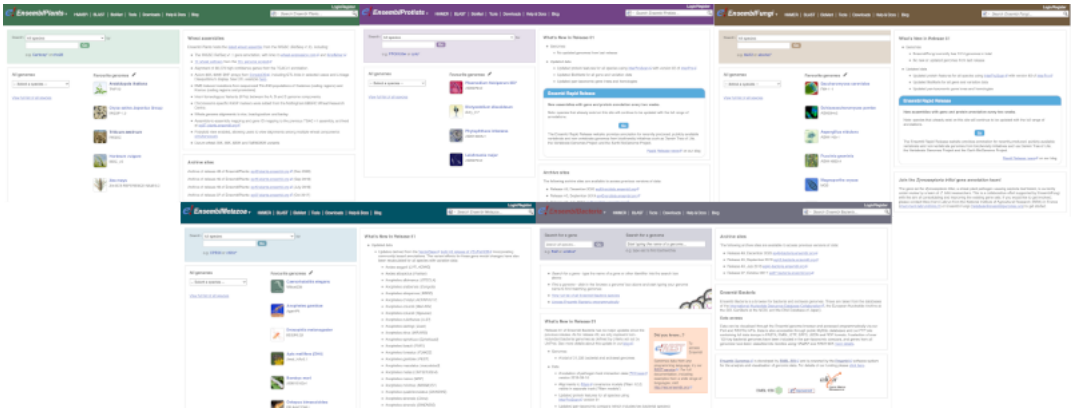
The current genome assembly for human is GRCh38. If you want to see the previous assembly, GRCh37, visit our dedicated site grch37.ensembl.org.

Ensembl Genomes

Homepage

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

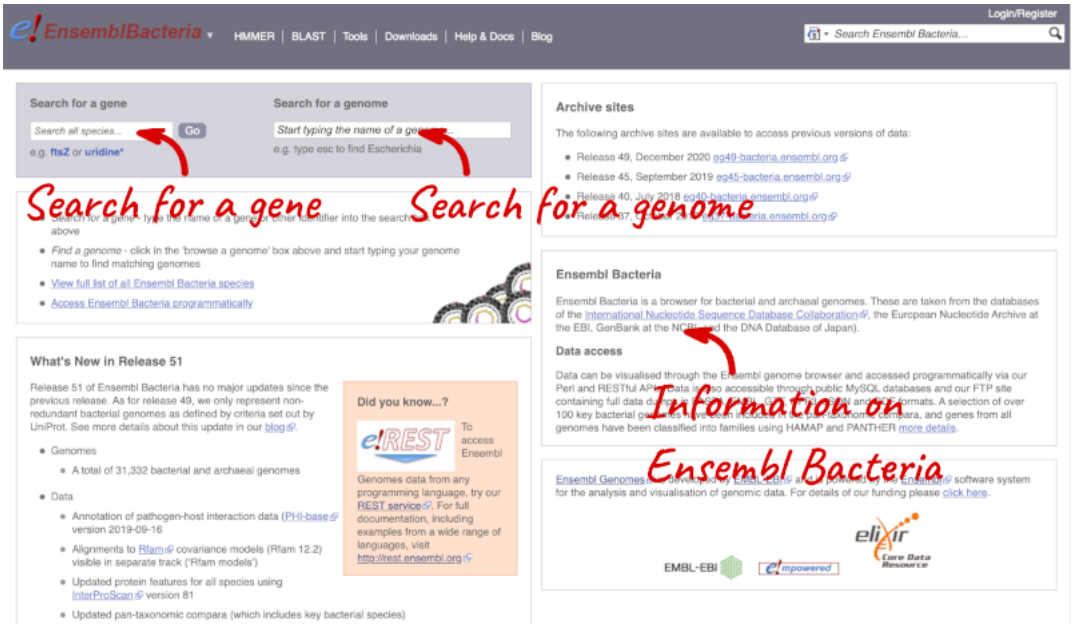
Click on the different taxa to see their homepages. Each one has a different colour-coding, but they are all structured in a similar format to the Ensembl main site.



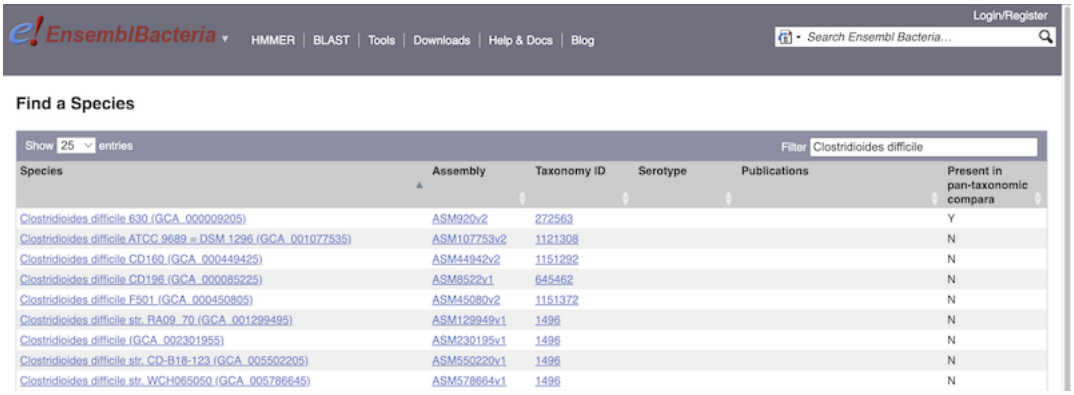
You can navigate most of the taxa in the same way as you would with Ensembl.

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.

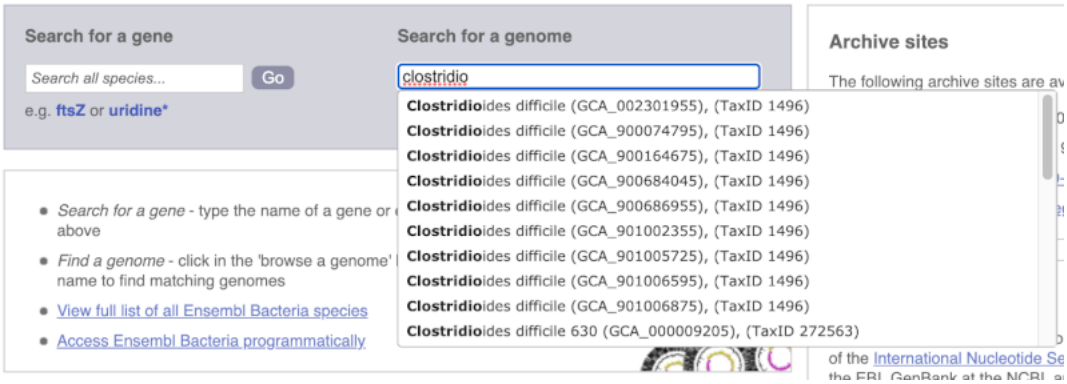


There's no drop-down species list for bacteria as it would be hard to navigate with the number of species. You can click the **View full list of all Ensembl Bacteria species** link underneath the coloured search block. Search for your species of interest using the filter in the top right-hand corner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



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*.

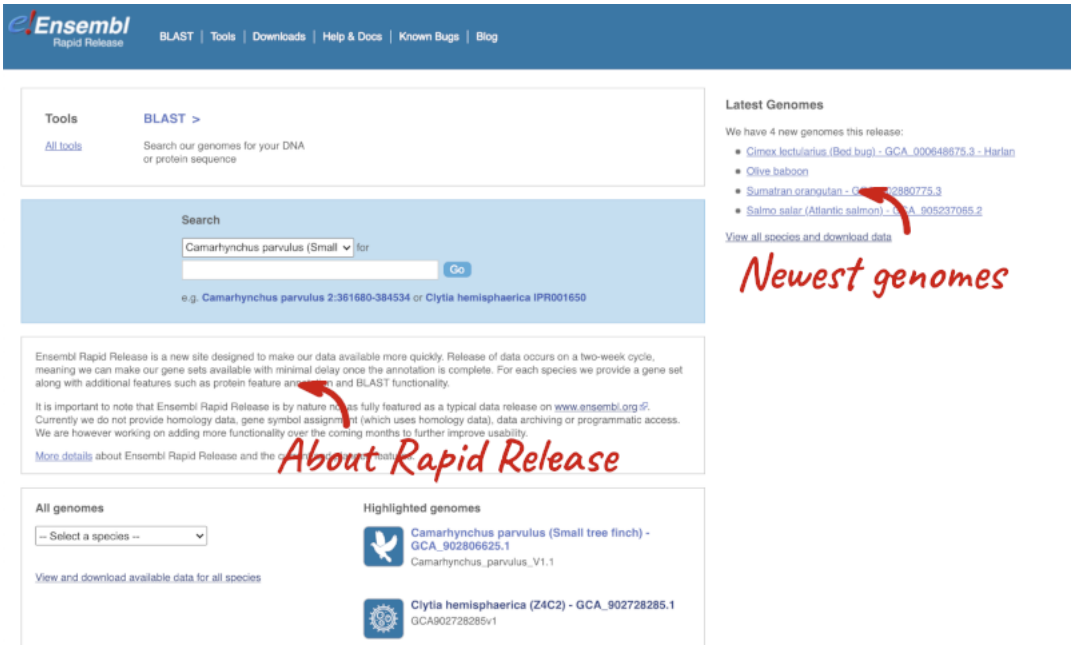


The drop down contains various strains of *C. difficile*. Let's choose *C. difficile 630*. This will take us to another species information page, where we can explore various features.

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

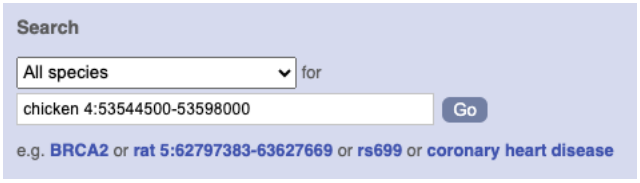
Our newest genomes, such as those coming from the Darwin Tree of Life, are available rapid.ensembl.org with limited annotation.


Ensembl
Rapid Release
BLAST | Tools | Downloads | Help & Docs | Known Bugs | Blog
Tools
BLAST >
All tools
Search our genomes for your DNA or protein sequence
Search
Camarhynchus parvulus (Small
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
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
Latest Genomes
We have 4 new genomes this release:
Cimex lectularius (Bed bug) - GCA_000648675.3 - Harlan
Olive baboon
Sumatran orangutan - GCA_002880775.3
Salmo salar (Atlantic salmon) - GCA_905237065.2
View all species and download data
Newest genomes

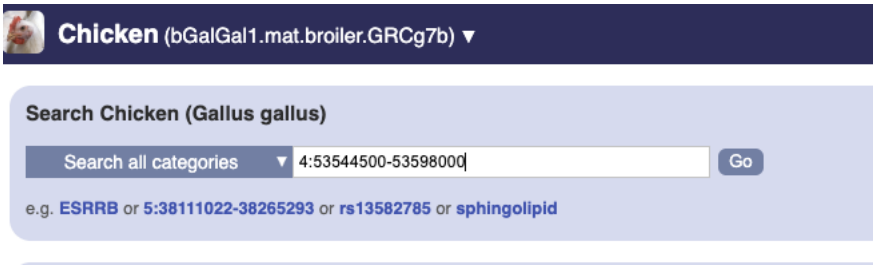
Exploring genomic regions in Ensembl, Demo

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.

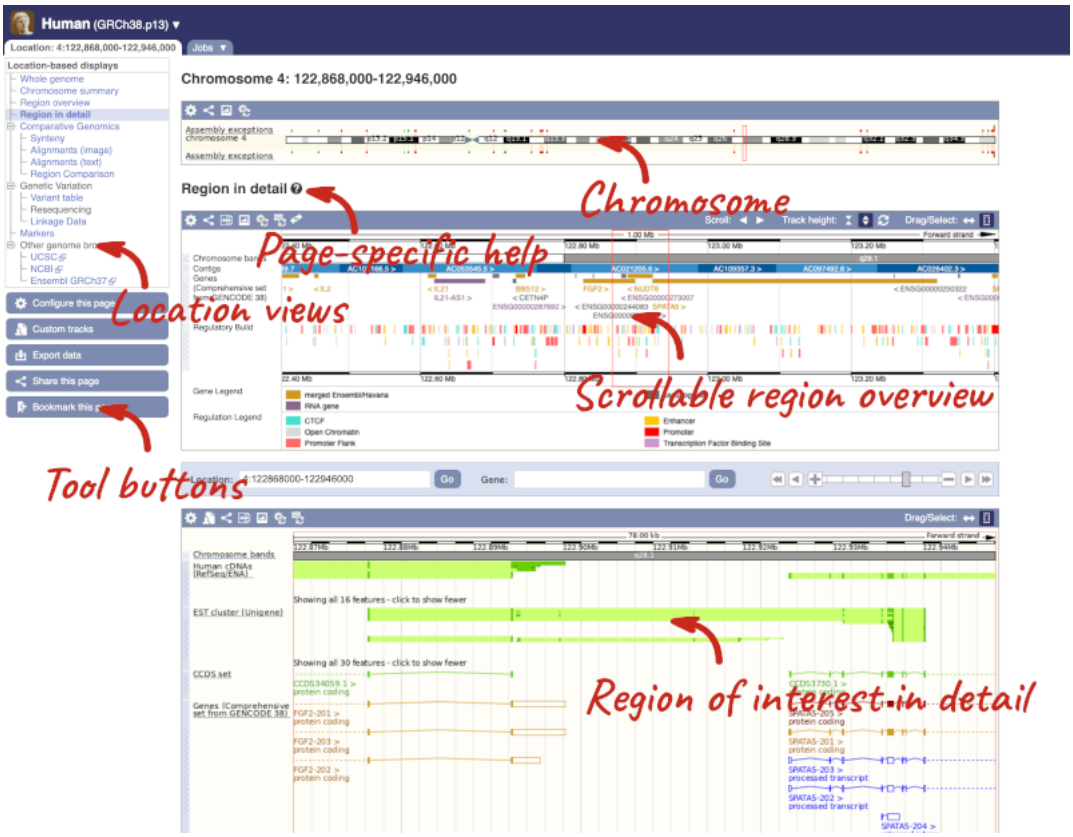
To bypass the text search, you need to input your region coordinates in the correct format, which is **`chromosome, colon, start coordinate, dash, end coordinate`**, with no spaces for example: **`human 4:122868000-122946000`**. Type (or copy and paste) these coordinates into either search box.



or



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

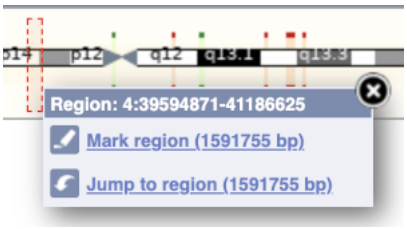


Click on the ❓ button to view page-specific help. The help pages provide text, labelled images and, in some cases, help videos to describe what you can see on the page and how to interact with it.

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:

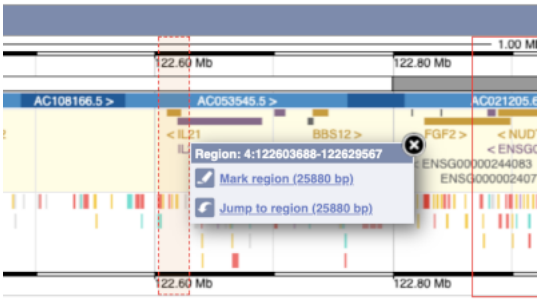
The region we're looking at is highlighted on the chromosome. You can jump to a different region by dragging out a box in this image. Drag out a box on the chromosome, a pop-up menu will appear.



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.

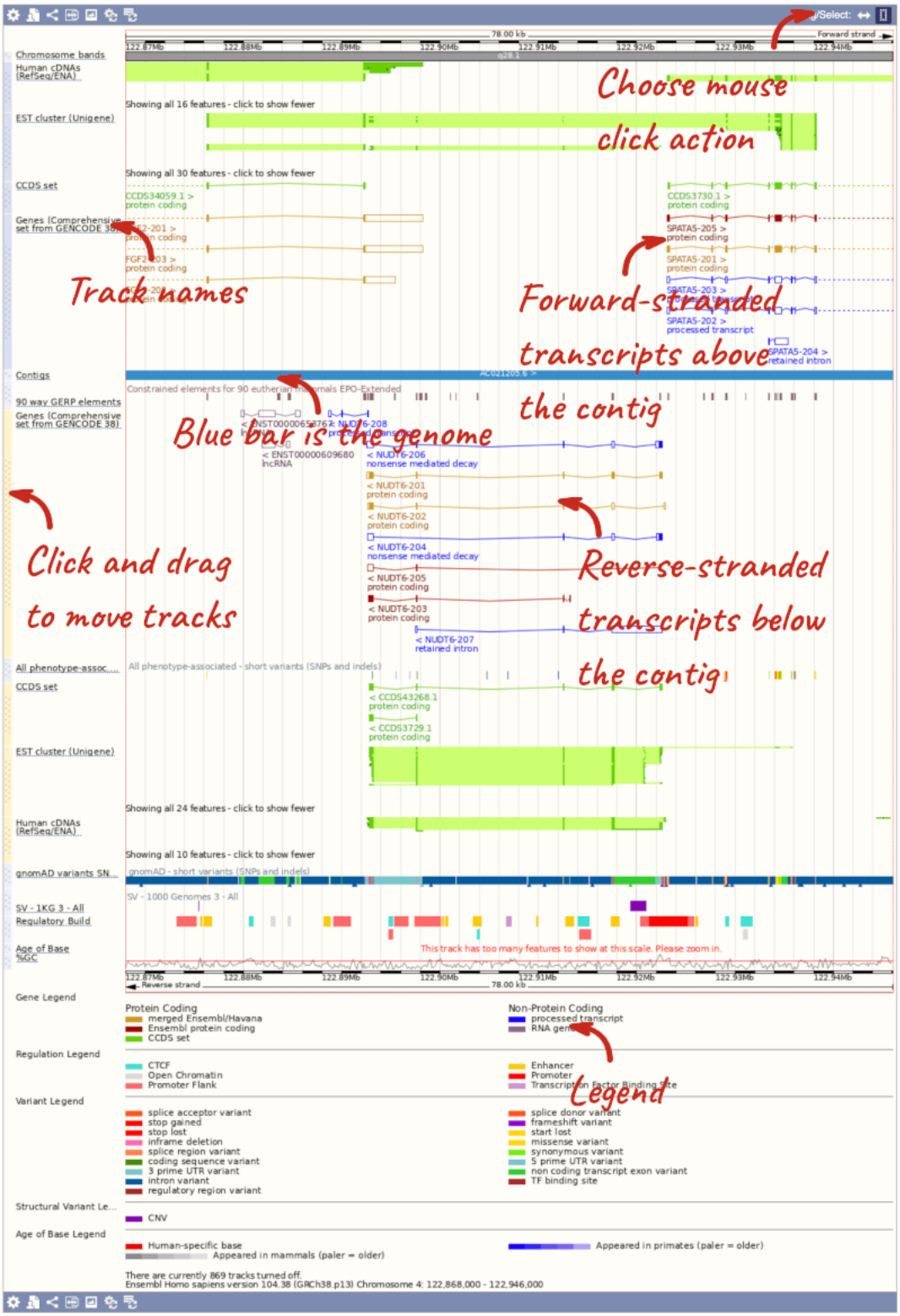
You can also drag out and jump to or mark a region.



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.

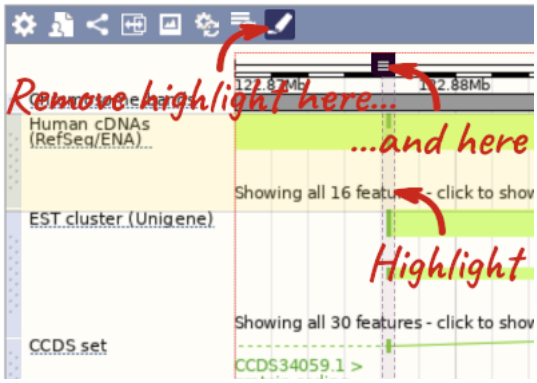


Here you can see various *tracks*, which is what we call a data type that you can plot against the genome. Some tracks, such as the transcripts, can be on the forward or reverse strand. Forward stranded features are shown above the blue contig track that runs across the middle of the image, with reverse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

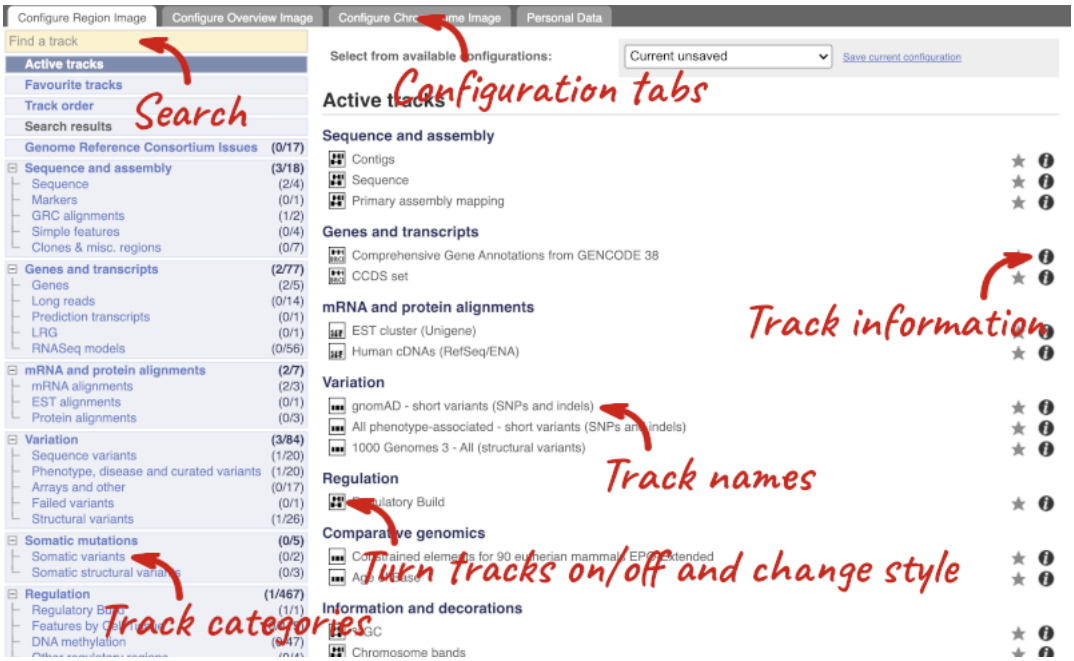
With the tool set to *Select*, drag out a box around an exon and choose *Mark region*.



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.

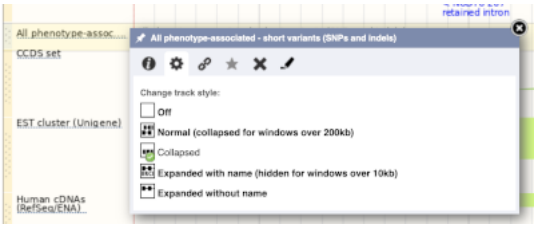
You can find more tracks to add by either exploring the categories on the left, or using the *Find a track* option at the top left. Type in a word or phrase to find tracks with it in the track name or description.

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.

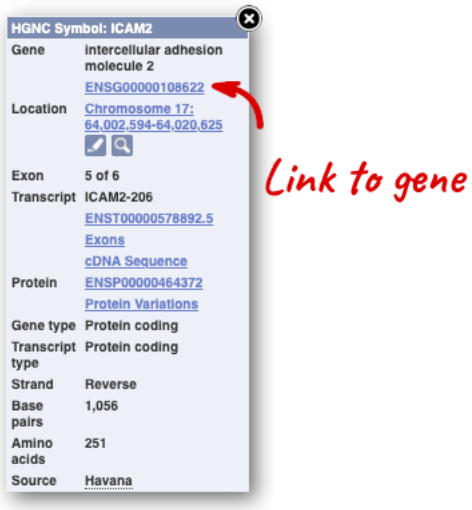


Now that you’ve got the view how you want it, you might like to show something you’ve found to a colleague or collaborator. Click on the *Share this page* button to generate a link. Email the link to someone else, so that they can see the same view as you, including all the tracks you’ve added. These 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

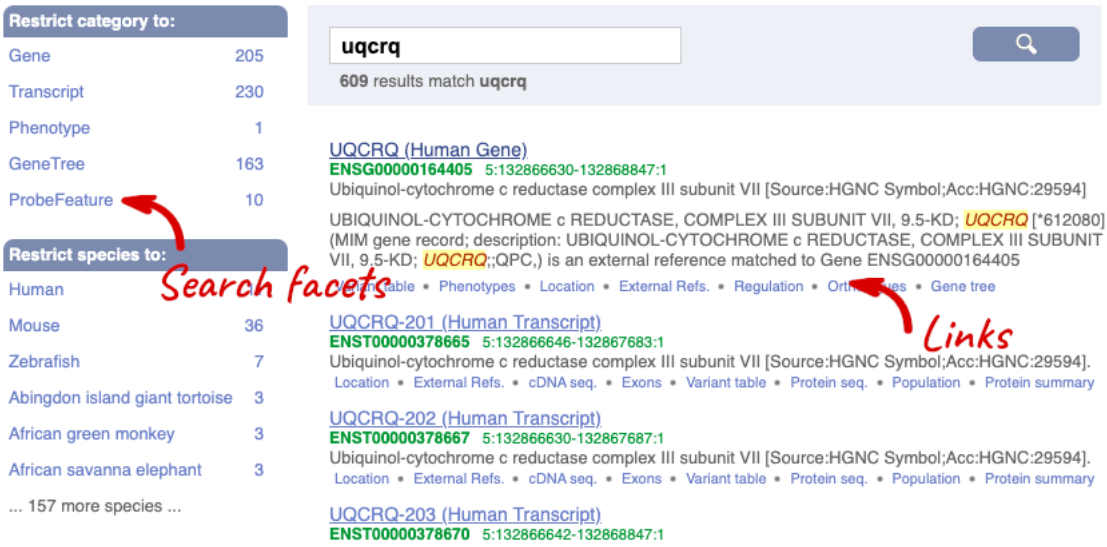
You can find out lots of information about Ensembl genes and transcripts using the browser. If you're already looking at a region view, you can click on any transcript and a pop-up menu will appear, allowing you to jump directly to that gene or transcript.



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.

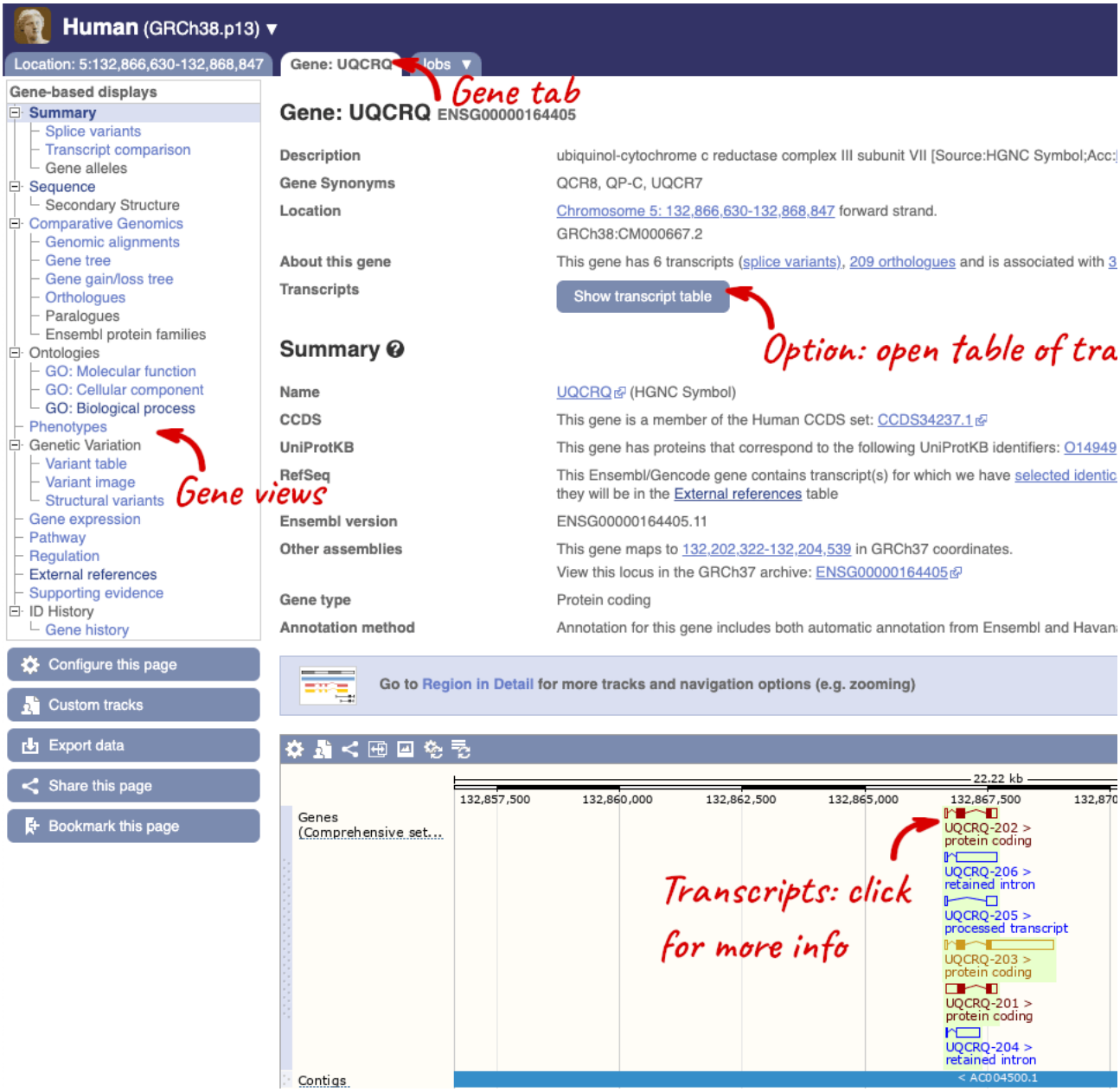
We're going to look at the human *UQCRQ* gene. From ensembl.org, type *UQCRQ* into the search bar and click the *Go* button. You will get a list of hits with the human gene at the top.

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.



The gene tab

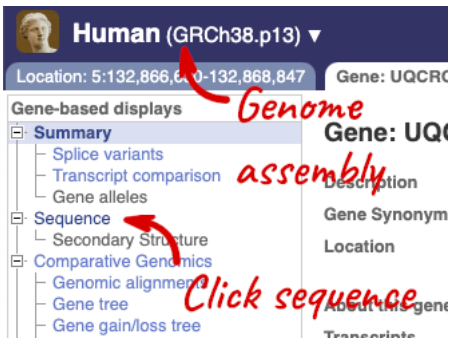
Click on the gene name or Ensembl ID. The *Gene tab* should open:



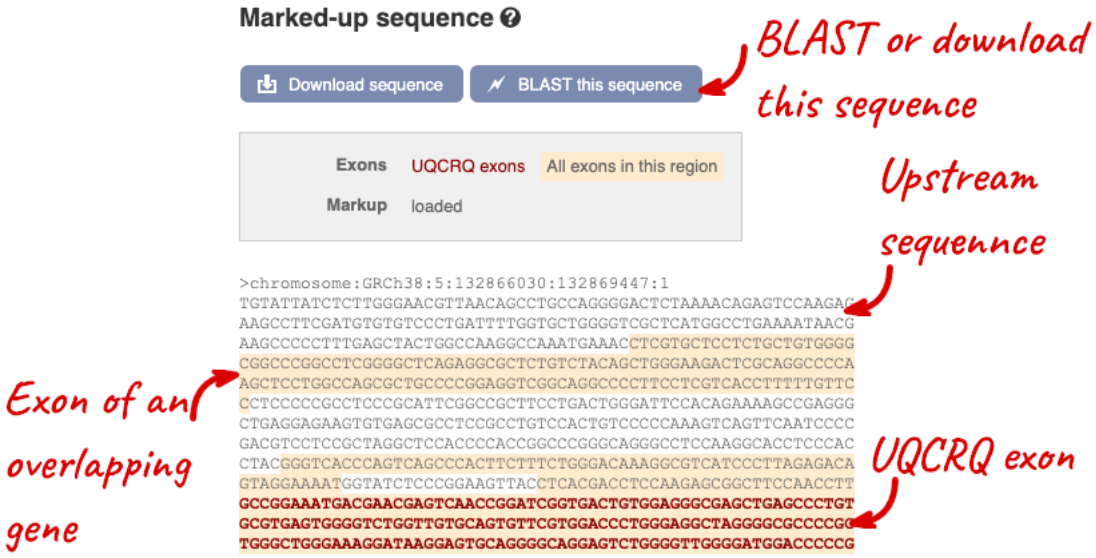
This page summarises the gene, including its location, name and equivalents in other databases. At the bottom of the page, a graphic shows a region view with the transcripts. We can see exons shown as blocks with introns as lines linking them together. Coding exons are filled, whereas non-coding exons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

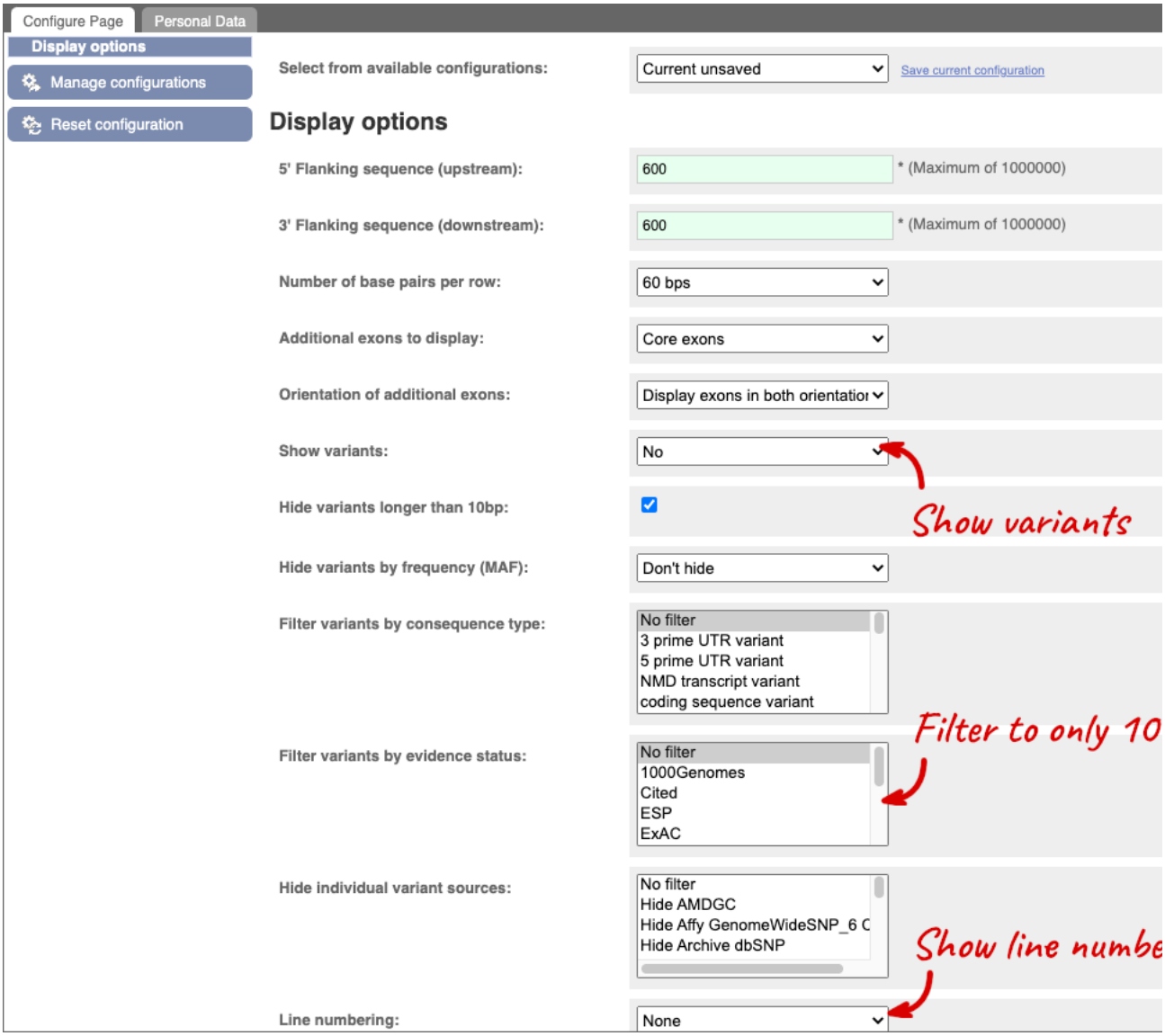
Let's walk through this menu for the gene tab. How can we view the genomic sequence? Click *Sequence* at the left of the page.



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.



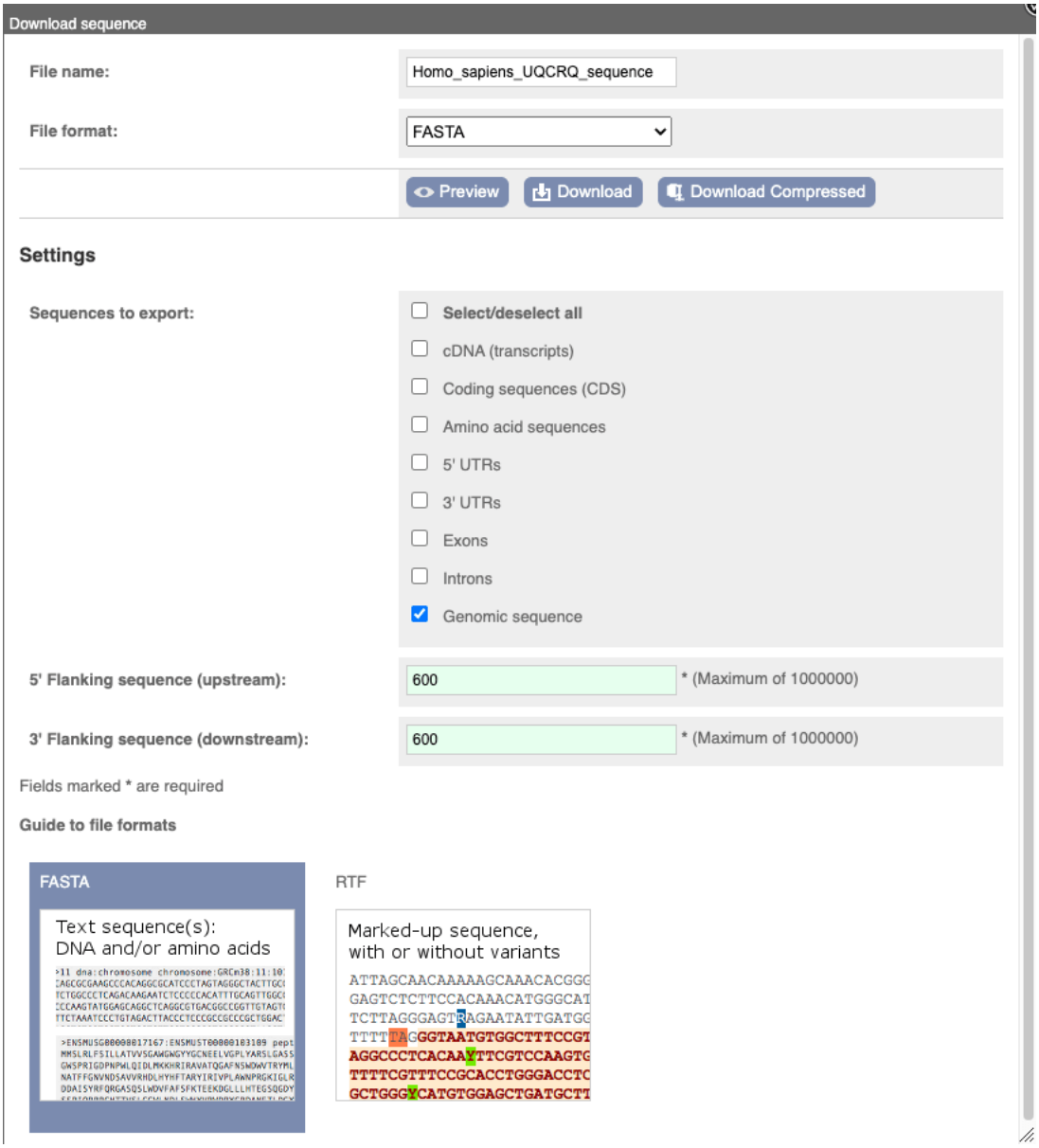
Exons are highlighted within the genomic sequence, both exons of our gene of interest and any neighbouring or overlapping gene. By default, 600 bases are shown up and downstream of the gene. We can make changes to how this sequence appears with the blue *Configure this page* button found at the left.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



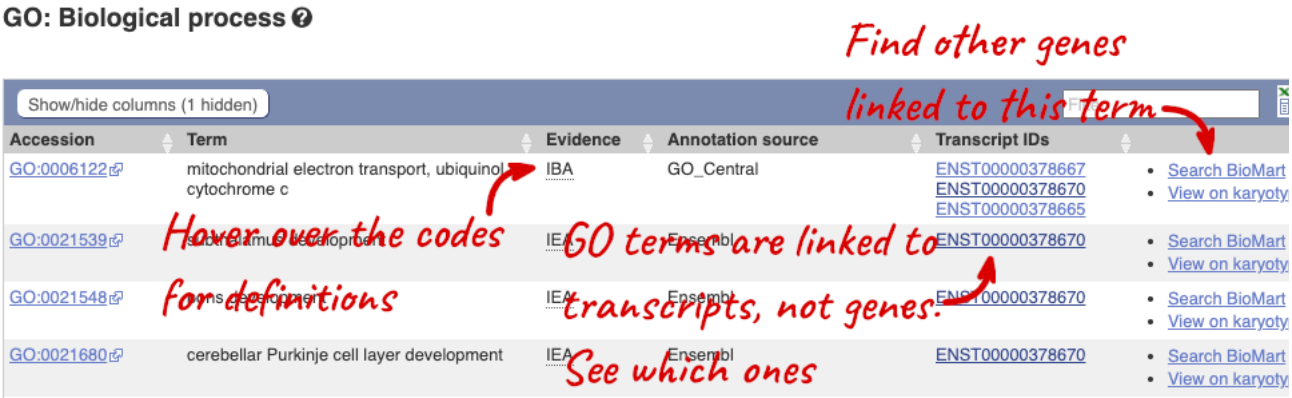
Once you have selected changes (in this example, *Show variants*, *1000 Genomes variants* and *Line numbering*) click at the top right.

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

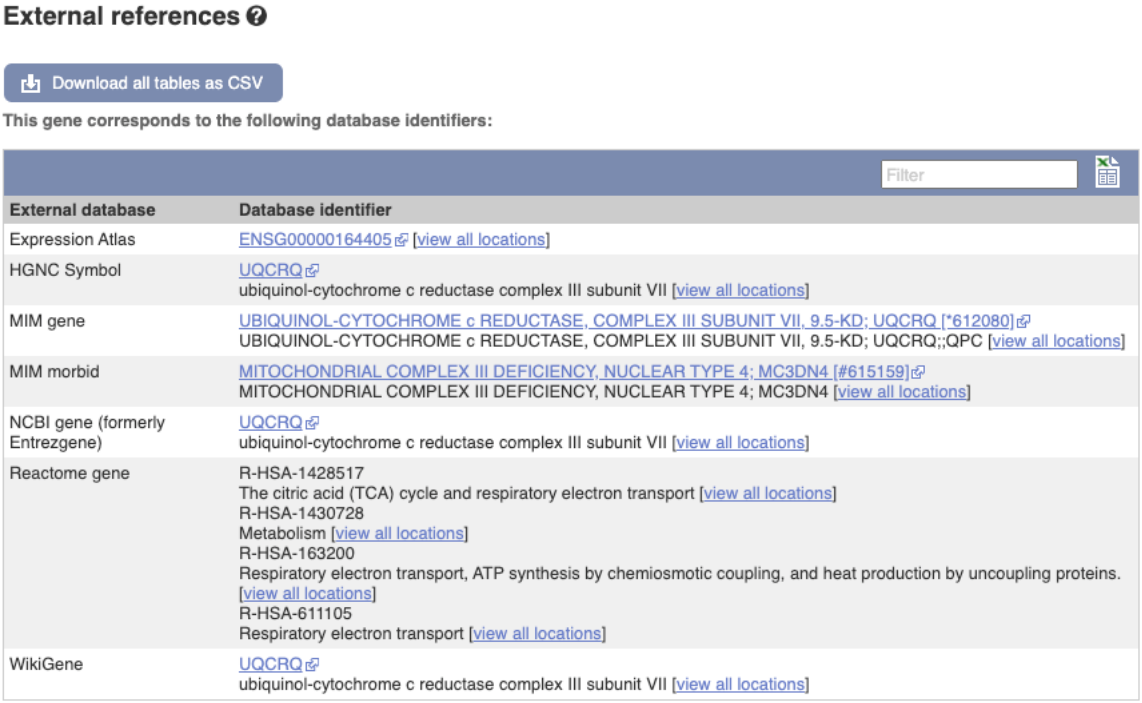


To find out what the protein does, have a look at GO terms from the Gene Ontology consortium. There are three pages of GO terms, representing the three divisions in GO: Biological process (what the protein does), Cellular component (where the protein is) and Molecular function (how it does it). Click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



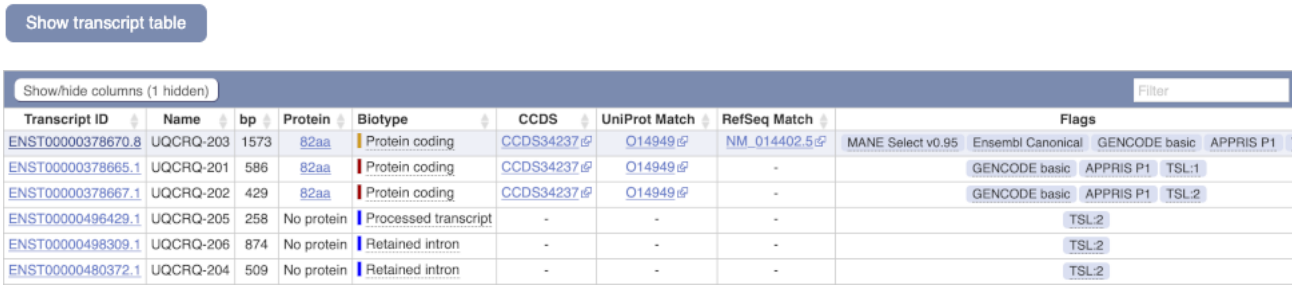
Here you can see the functions that have been associated with the gene. There are three-letter codes that indicate how the association was made, as well as links to the specific transcript they are linked to.

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*:



Demo: The transcript tab

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



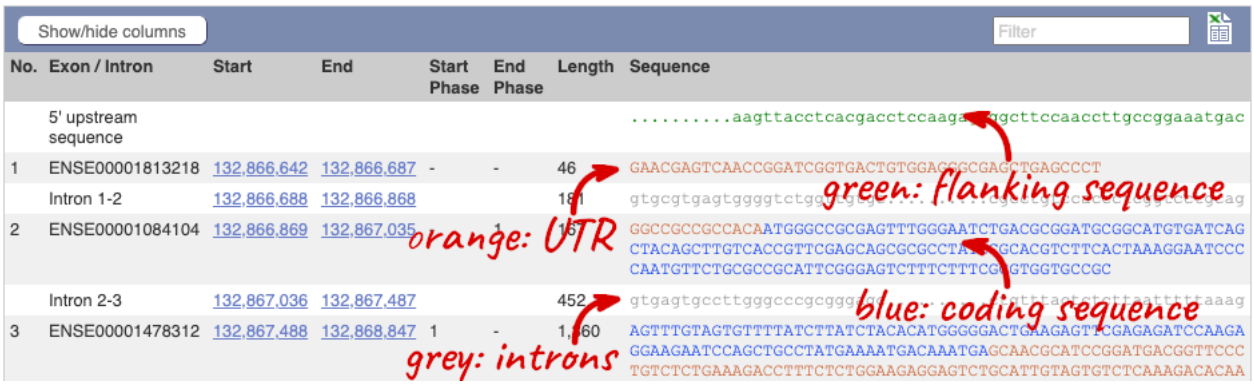
Here we can see a list of all the transcripts of *UQCRQ* with their identifiers, lengths, biotypes and flags to help you decide which ones to look at.

If we were to only choose one transcript to analyse, we would choose UQCRQ-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 UQCRQ-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 UQCRQ-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.



You may want to change the display (for example, to show more flanking sequence, or to show full introns). In order to do so click on *Configure this page* and change the display options accordingly.

Now click on the *cDNA link* to see the spliced transcript sequence with the amino acid sequence. This page is useful for mapping between the RNA and protein sequences, particularly genetic variants.

UnTranslated Regions (UTRs) are highlighted in dark yellow, codons are highlighted in light yellow, and exon sequence is shown in black or blue letters to show exon divides. Sequence variants are represented by highlighted nucleotides and clickable IUPAC codes are above the sequence.

Next, follow the *General identifiers* link at the left. Just like the *External References* page in the gene tab, this page shows links out to other databases such as RefSeq, UniProtKB, PDBe and others, this time linked to the transcript or protein product, rather than the gene.

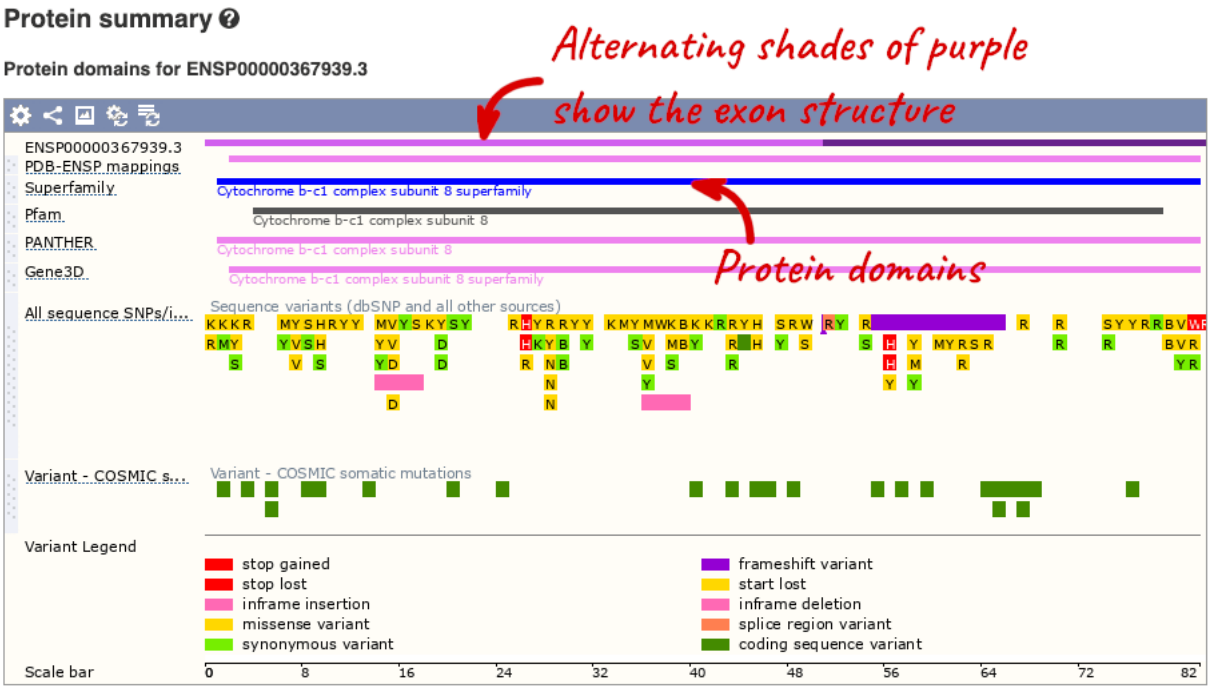
General identifiers

This transcript corresponds to the following database identifiers:

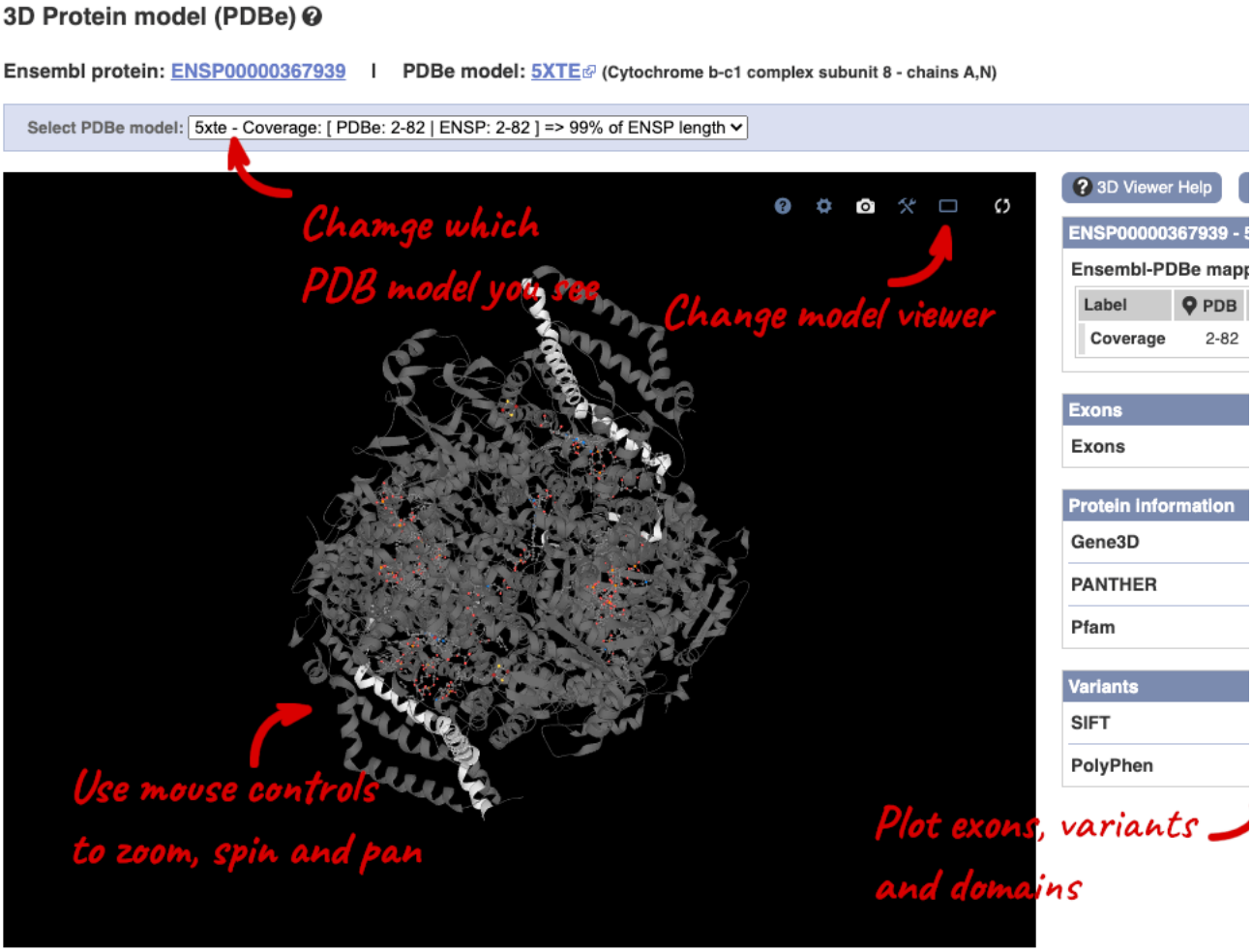
Show All entries Filter

External database	Database identifier
BioGRID Interaction data, The General Repository for Interaction Datasets	117991 [view all locations]
CCDS	CCDS34237.1 [view all locations]
European Nucleotide Archive	BC001390 [align] [view all locations] BC090048 [align] [view all locations] D50369 [align] [view all locations]
Human Protein Atlas	ENSG00000164405 [view all locations] HPA046693 [view all locations] HPA053323 [view all locations]
INSDC protein ID	AAH01390.1 [align] [view all locations] AAH90048.1 [align] [view all locations] BAA23321.1 [align] [view all locations]

If you're interested in protein domains, you could click on *Protein summary* to view domains from Pfam, PROSITE, Superfamily, InterPro, and more. These are all plotted against the transcript sequence, with the exons shown in alternating shades of purple at the top of the page. Alternatively, you can go to *Domains & features* to see a table of the same information.



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



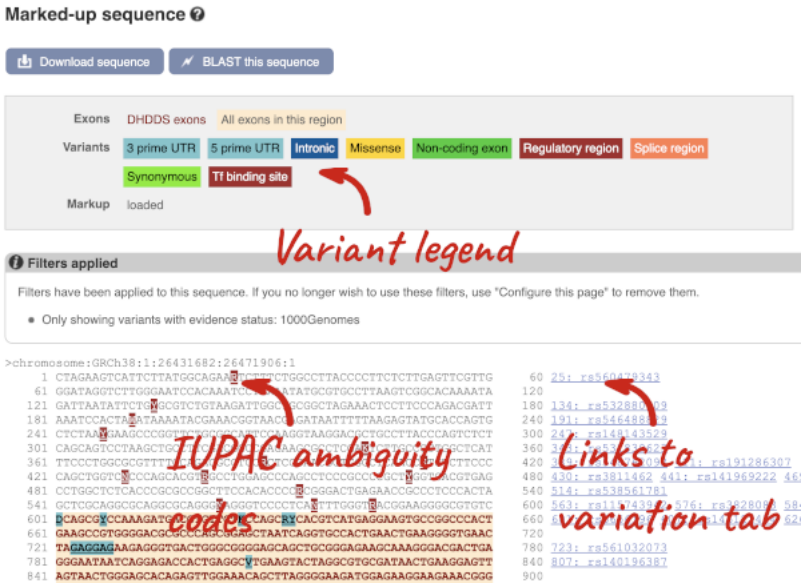
This uses LiteMol to show a 3D protein. You can use all the normal controls that you would use with LiteMol, plus plot Ensembl features like Exons and variants onto the structure using the options on the right. We allow you to see the top ten PDB models for this protein, based on coverage and quality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

Exploring variants in Ensembl, Demo

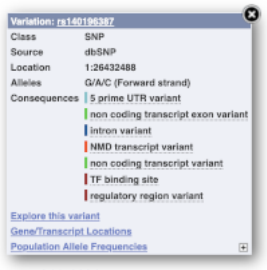
In any of the sequence views shown in the Gene and Transcript tabs, you can view variants on the sequence. You can do this by clicking on Configure this page from any of these views.

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*.



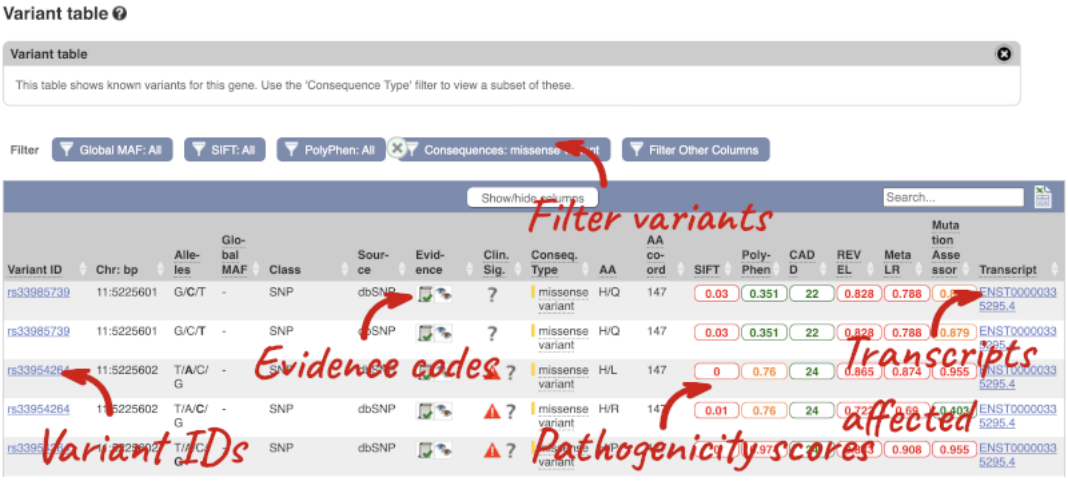
Find out more about a variant by clicking on it.



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.

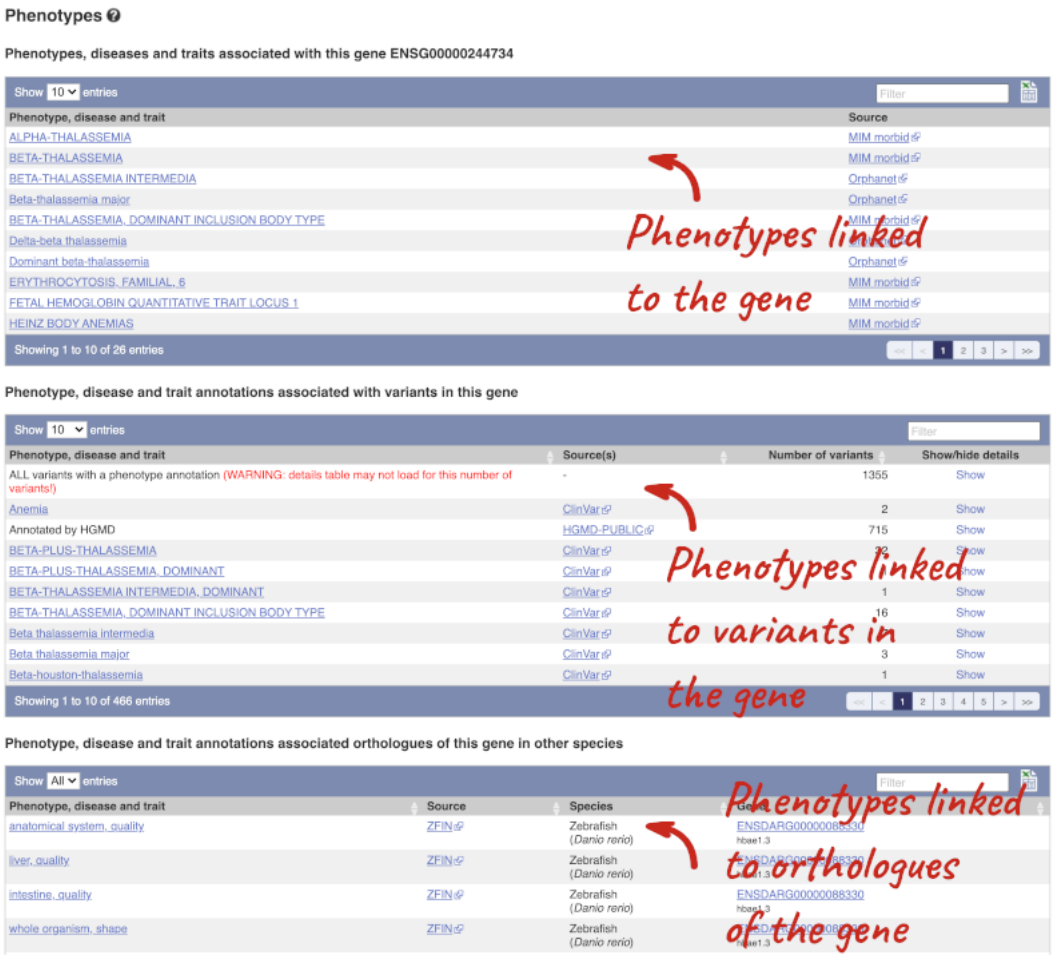


You can filter the table to only show the variants you're interested in. For example, click on *Consequences: All*, then select the variant consequences you're interested in. For display purposes, the table above has already been filtered to only show missense variants.

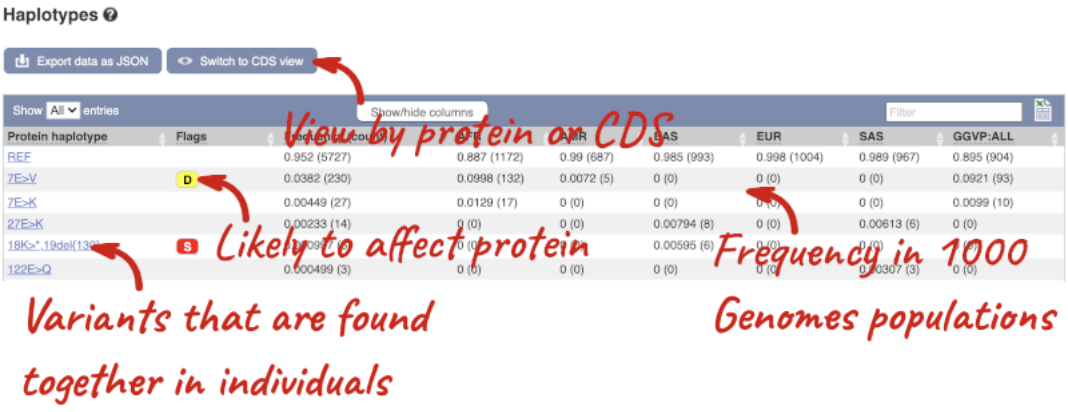
You can also filter by the different pathogenicity scores and MAF, or click on *Filter other columns* for filtering by other columns such as Evidence or Class.

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.



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

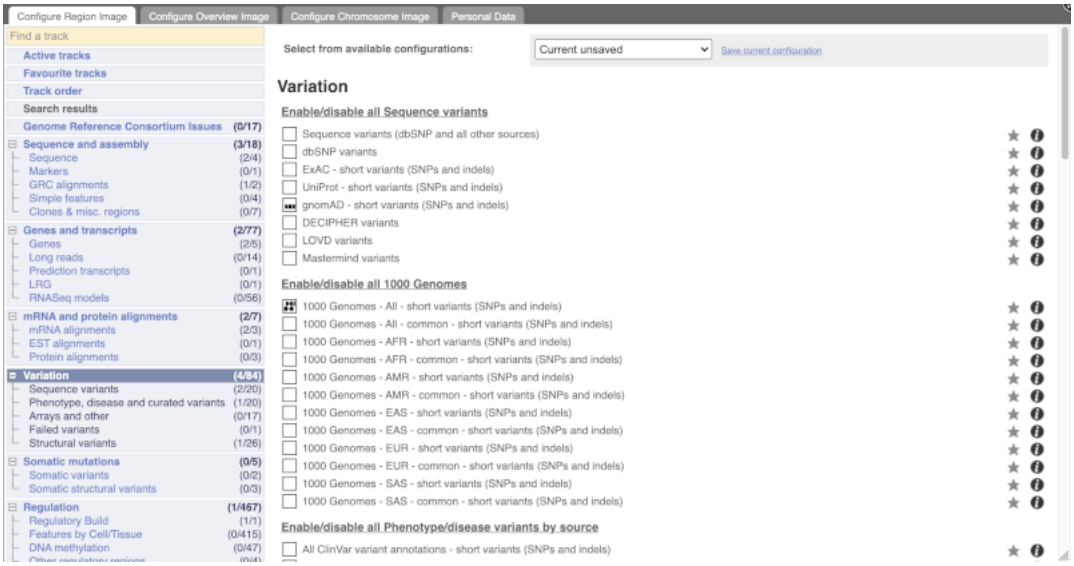


The Haplotypes view in the transcript tab shows you the actual protein and CDS sequences in 1000 Genomes individuals. This is possible because the 1000 Genomes study has phased genotypes, so we know which alleles occur on which of the chromosome pairs. The table lists all the versions of the protein 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.

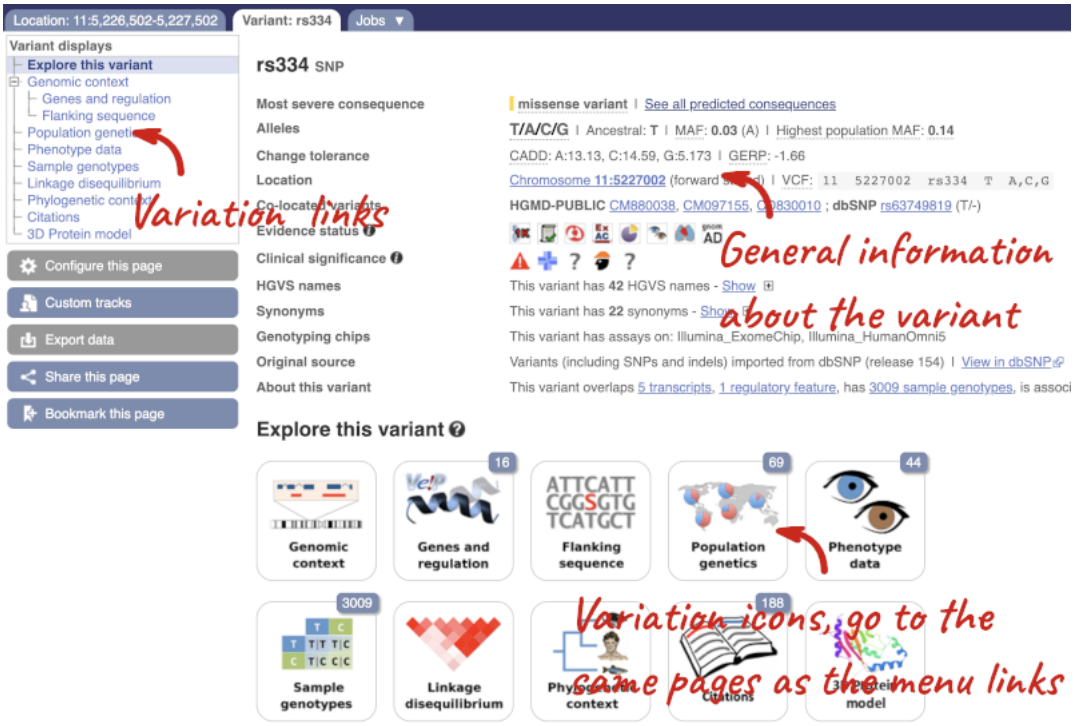
Let's have a look at variants in the Location tab. Click on the *Location* tab in the top bar.

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

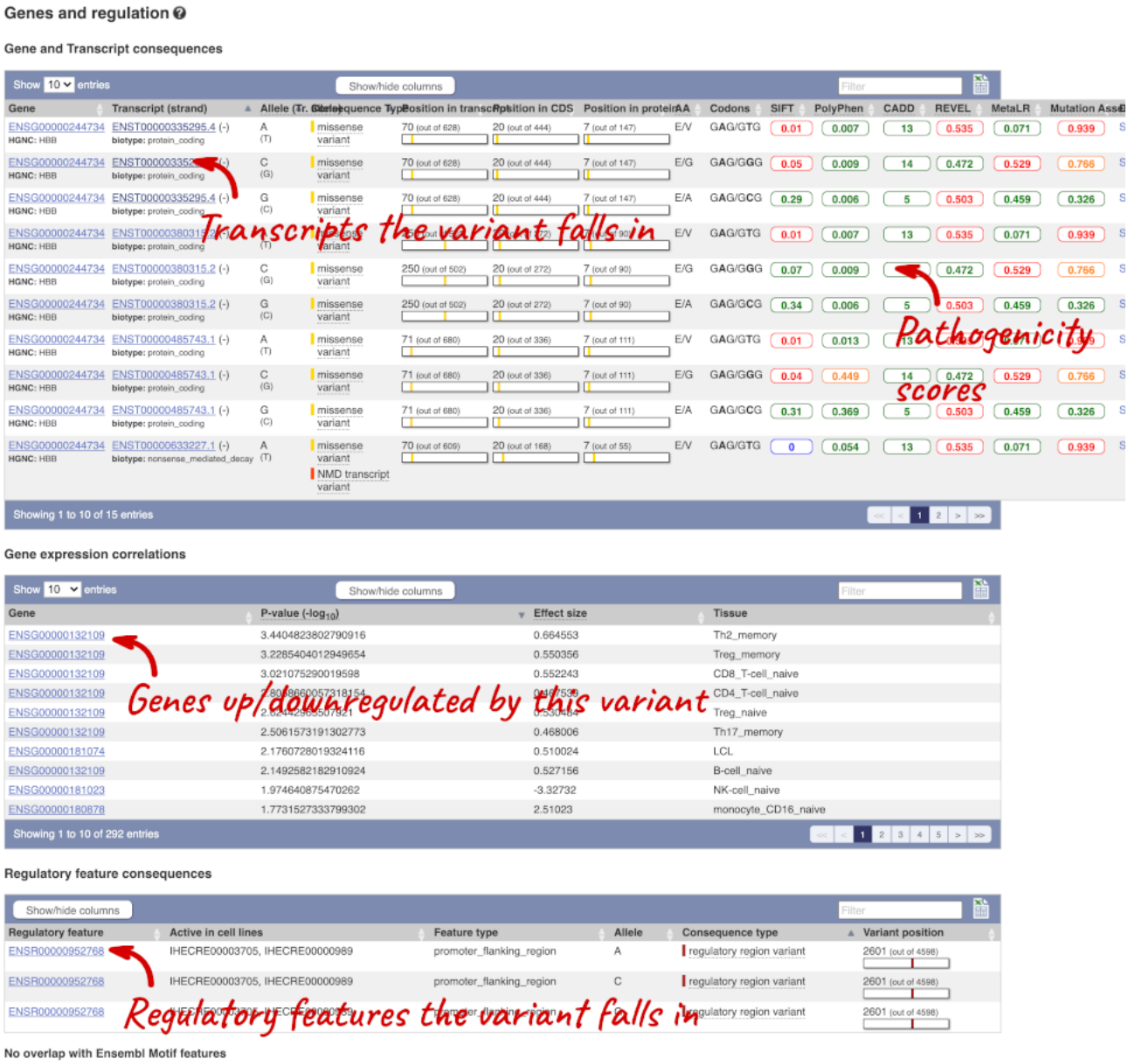


There are various options for turning on variants. You can turn on variants by source, by frequency, presence of a phenotype or by individual genome they were isolated from. You can also turn on genotyping chips.

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.

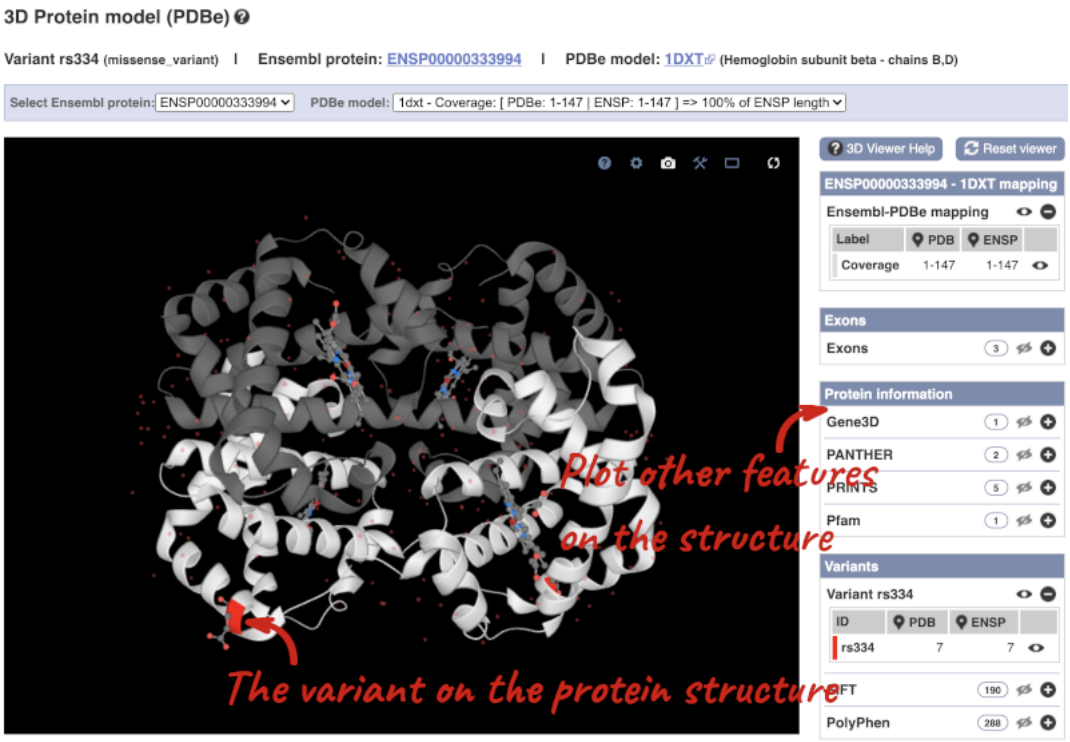


The icons show you what information is available for this variant. Click on *Genes and regulation*, or follow the link on the left.



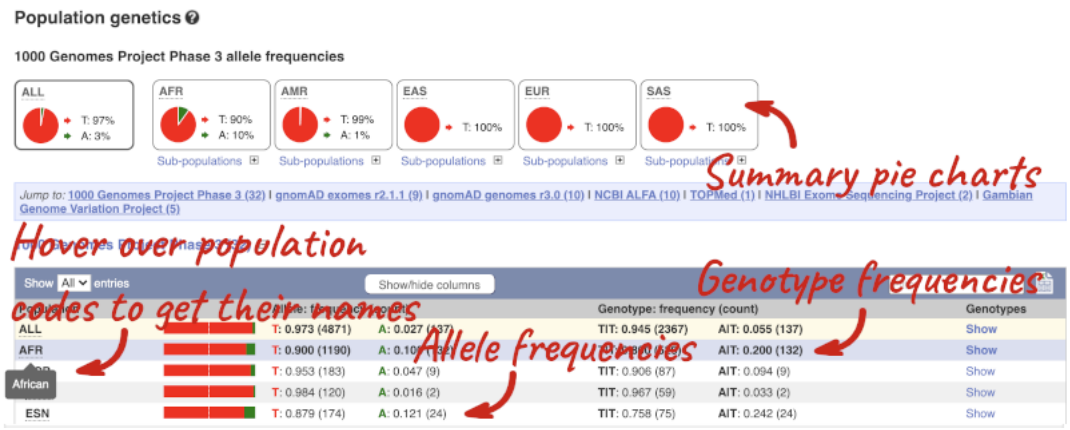
This page illustrates the genes the variant falls within and the consequences on those genes, including pathogenicity predictors. It also shows data from GTEx on genes that have increased/decreased expression in individuals with this variant, in different tissues. Finally, regulatory features and motifs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

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



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.



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)	Associated allele	Statistics
Anemia	ClinVar [New York Genome Cent...]	Anemia, anemia (phenotype)	EFO:0004272, HP:0001903	-	-	⚠ ☆☆☆☆	LOC106099062, HBB, LOC107133510	A	-
beta Thalassemia	ClinVar [MendelicsIMyriad Wom...]	-	Orphanet:848	-	-	? ☆☆☆☆	LOC106099062, HBB, LOC107133510	A	-
BETA-THALASSEMIA, DOMINANT INCLUSION BODY TYPE	ClinVar [Division of Human Ge...]	-	Orphanet:231226	-	-	⚠ ☆☆☆☆	LOC106099062, HBB, LOC107133510	A	-
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

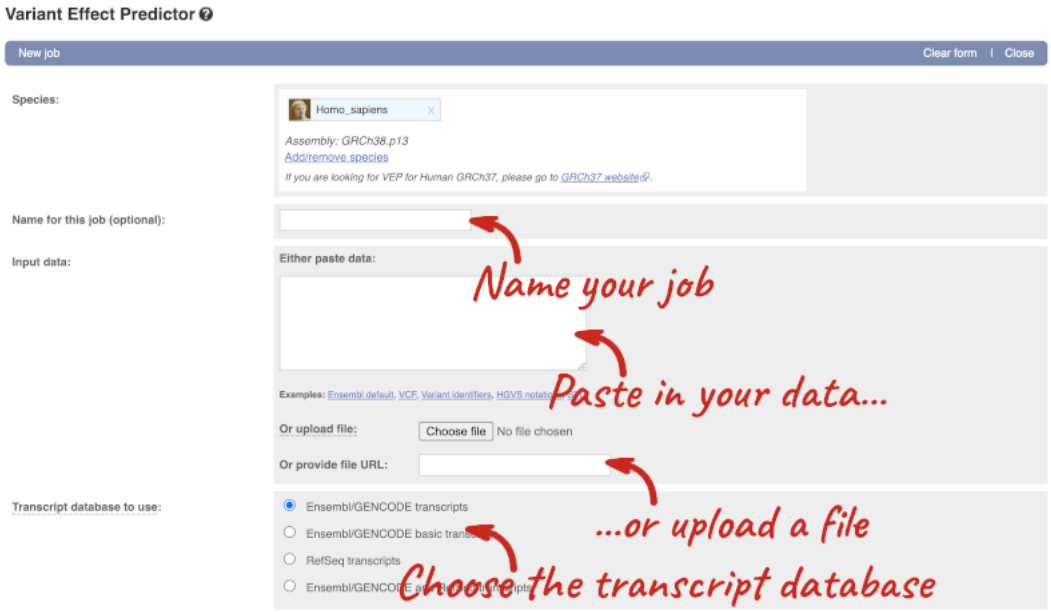
We have identified five variants on human chromosome nine, C-> A at 128203516, an A deletion at 128328461, C->A at 128322349, C->G at 128323079 and G->A at 128322917.

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.

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:



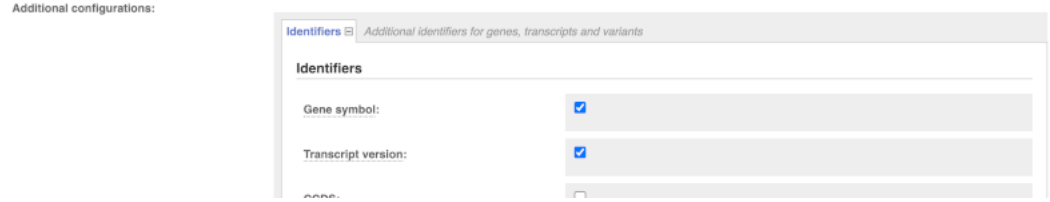
The data is in VCF format:
`chromosome coordinate id reference alternative`

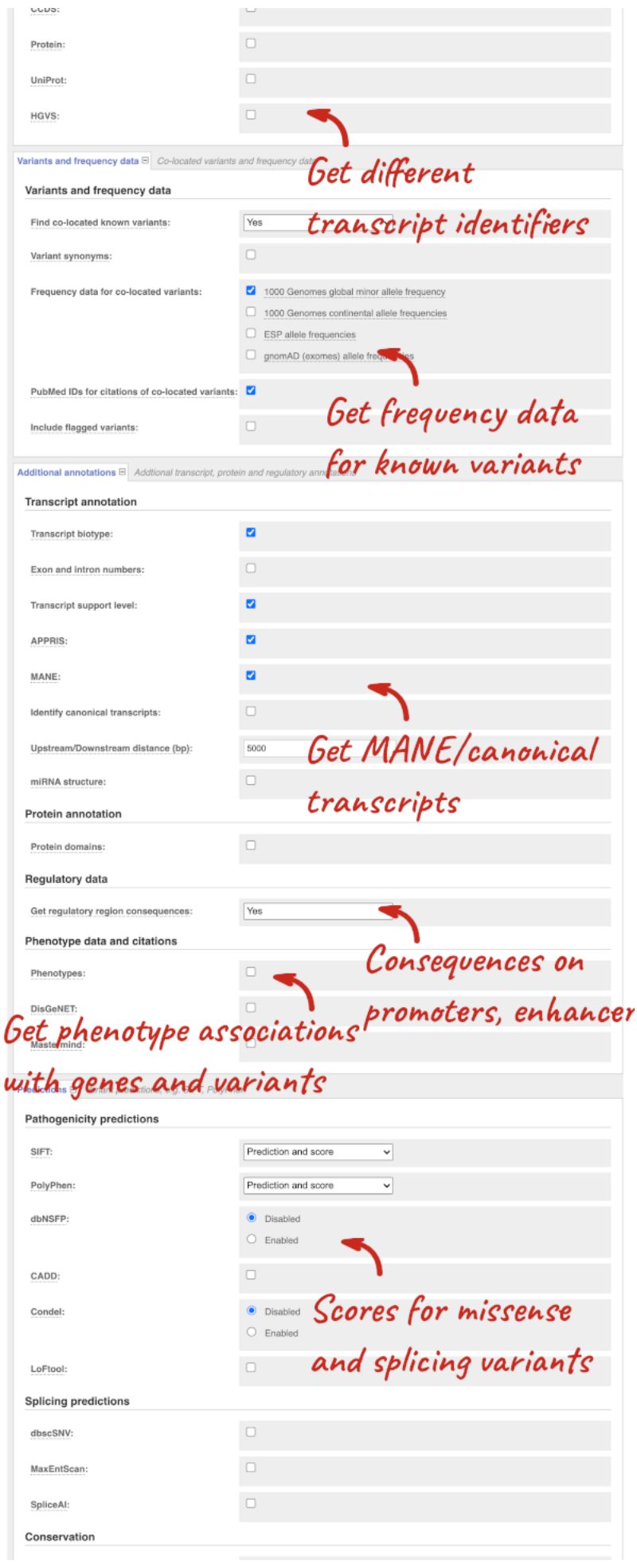
Put the following into the Paste data box:

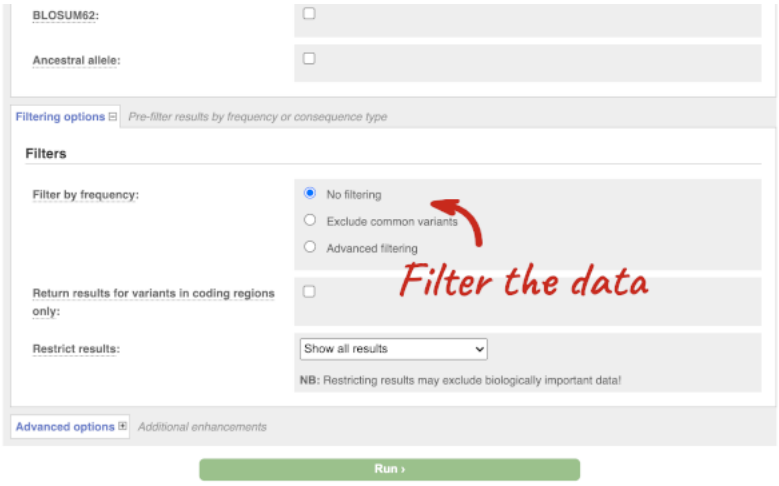
```
9 128328460 var1 TA T
9 128322349 var2 C A
9 128323079 var3 C G
9 128322917 var4 G A
9 128203516 var5 C A
```

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.




CCDS:
Protein:
UniProt:
HGVS:
Get different transcript identifiers
Variants and frequency data
Co-located variants and frequency data
Variants and frequency data
Find co-located known variants:
Yes
Variant synonyms:
Frequency data for co-located variants:
1000 Genomes global minor allele frequency
1000 Genomes continental allele frequencies
ESP allele frequencies
gnomAD (exomes) allele frequencies
Get frequency data for known variants
PubMed IDs for citations of co-located variants:
Include flagged variants:
Additional annotations
Additional transcript, protein and regulatory annotations
Transcript annotation
Transcript biotype:
Exon and intron numbers:
Transcript support level:
APPRIS:
MANE:
Identify canonical transcripts:
Upstream/Downstream distance (bp):
5000
miRNA structure:
Get MANE/canonical transcripts
Protein annotation
Protein domains:
Regulatory data
Get regulatory region consequences:
Yes
Consequences on promoters, enhancer
Phenotype data and citations
Phenotypes:
DisGeNET:
Mastermind:
Get phenotype associations with genes and variants
Pathogenicity predictions
SIFT:
Prediction and score
PolyPhen:
Prediction and score
dbNSFP:
Disabled
Enabled
CADD:
Condel:
Disabled
Enabled
LoFtool:
Scores for missense and splicing variants
Splicing predictions
dbscSNV:
MaxEntScan:
SpliceAI:
Conservation



Hover over the options to see definitions.

We're going to select some options:

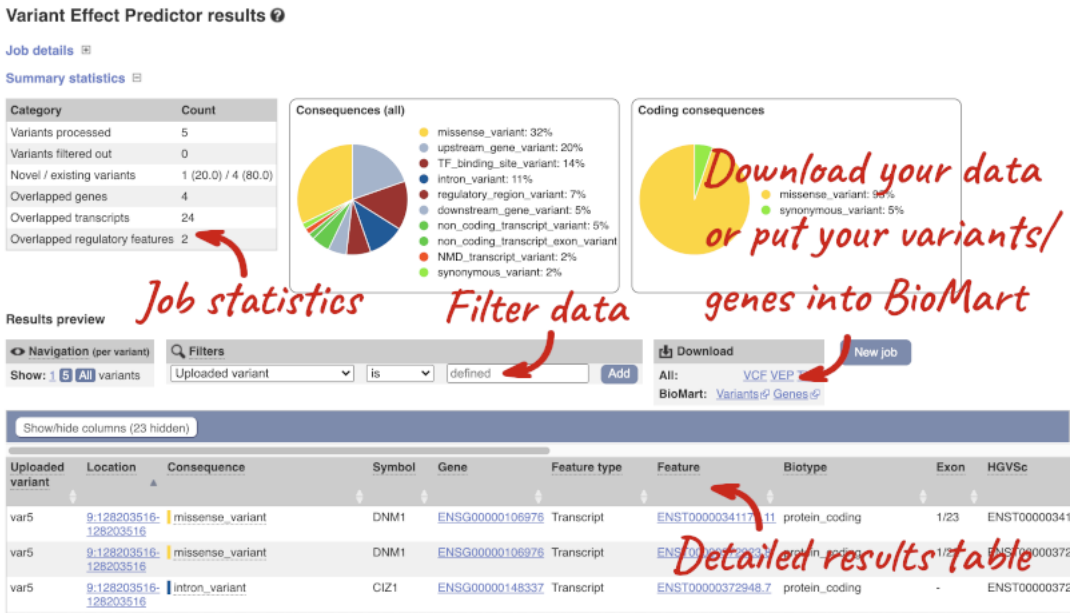
- *HGVS*, annotation of variants in terms of the transcripts and proteins they affect, commonly-used by the clinical community
- *Phenotypes*
- *Protein domains*

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

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.



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

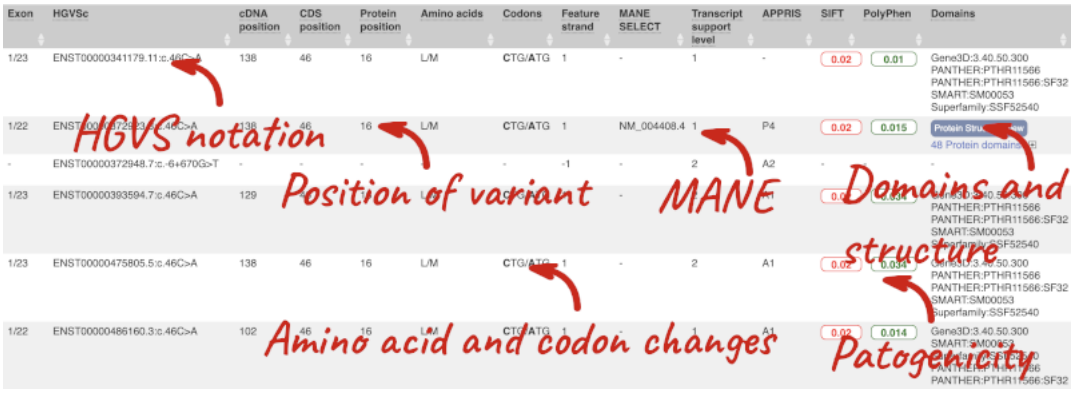
only loci, this column will contain the locus and alleles of the variant. You'll notice that the variants are not neccessarily in the order they were in in your input. You'll also see that there are multiple lines in the table for each variant, with each line representing one transcript or other feature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



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

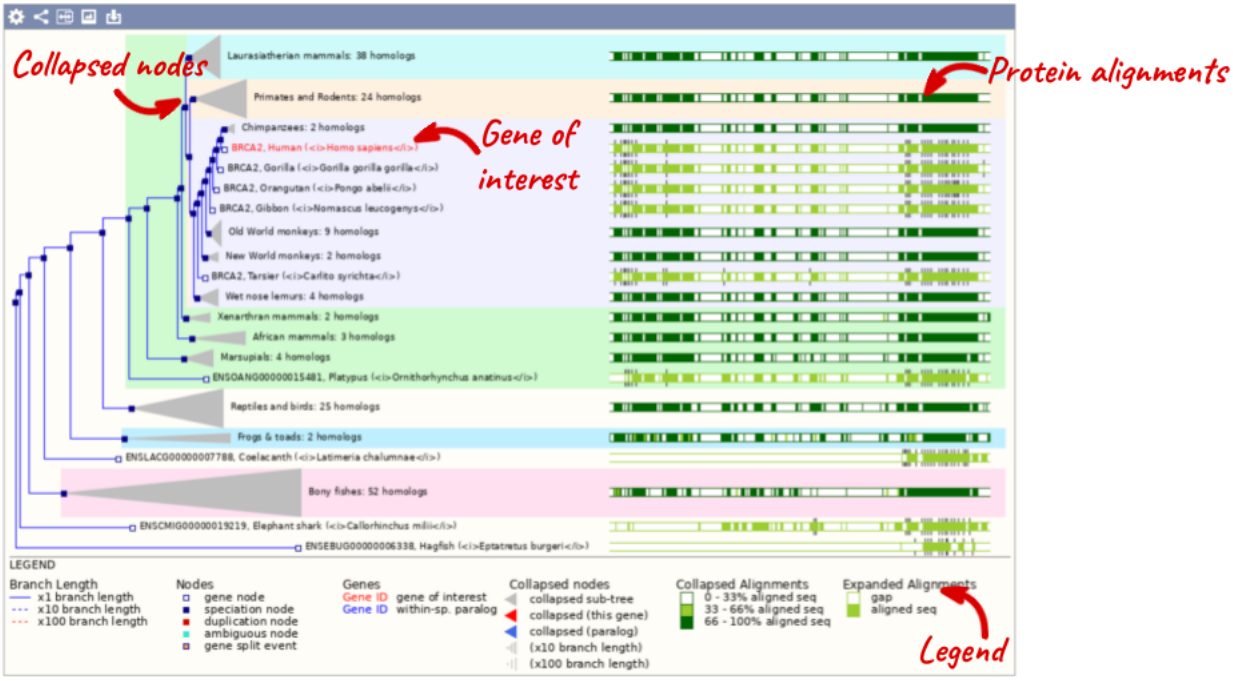
Above the table is the *Filter* option, which allows you to filter by any column in the table. You can select a column from the drop-down, then a logic option from the next drop-down, then type in your filter to the following box. We'll try a filter of *Consequence*, followed by *is* then *missense_variant*, 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.

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.

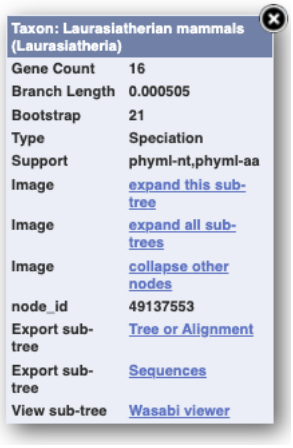
Gene trees and homologues, Demo

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

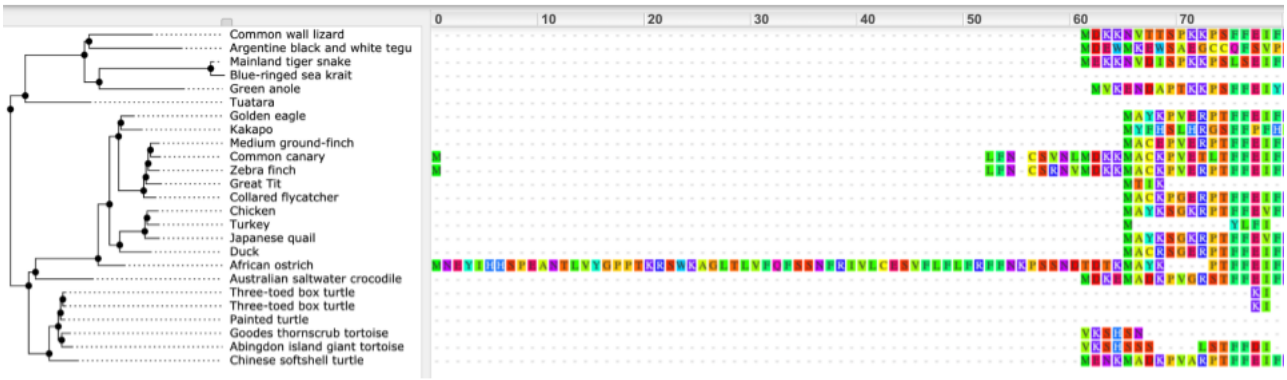
Click on *Gene tree*, which will display the current gene in the context of a phylogenetic tree used to determine orthologues and paralogues.



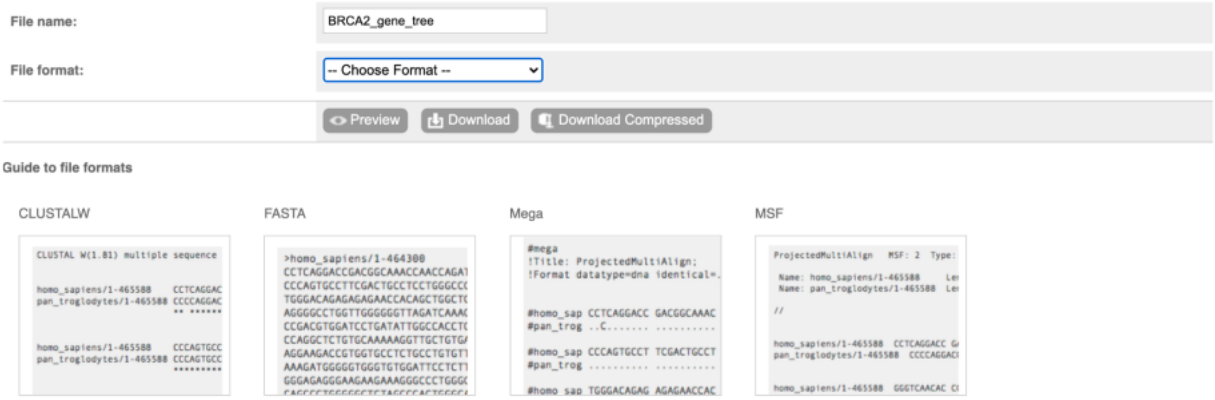
Funnels indicate collapsed nodes. We can expand them by clicking on the node and selecting *Expand this sub-tree* from the pop-up menu.



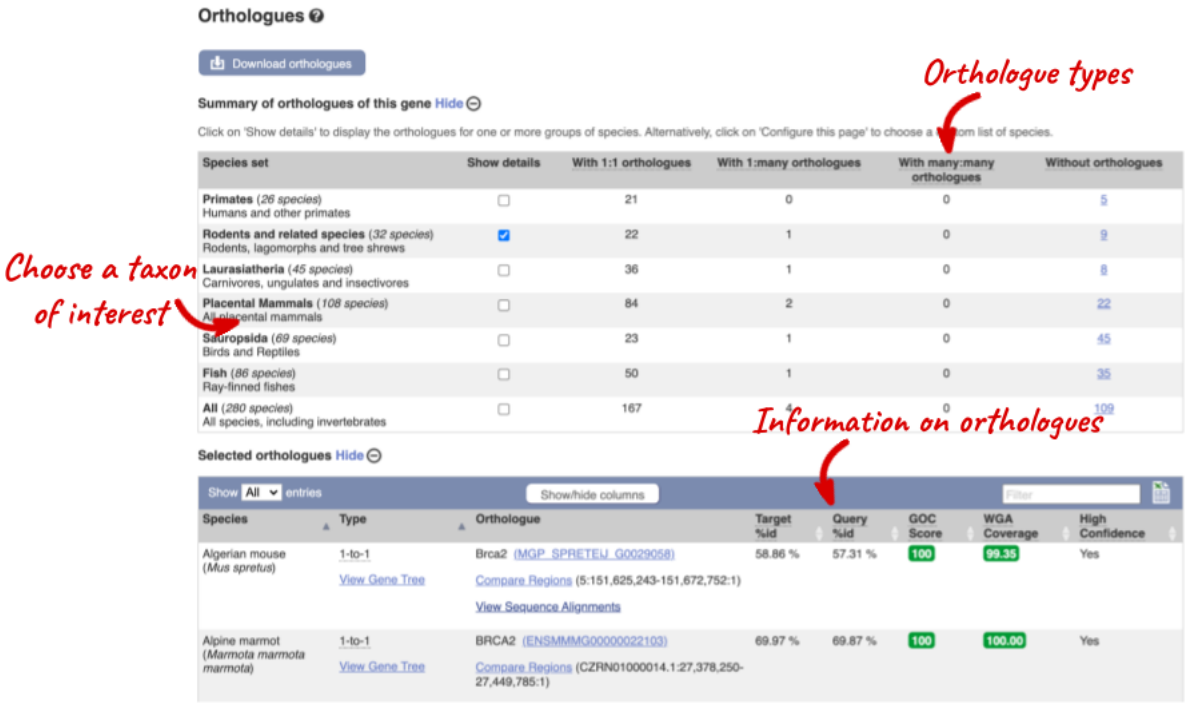
We can also see the alignment of the sub-tree by clicking on Wasabi viewer, which will open a pop-up:



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.



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.



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.

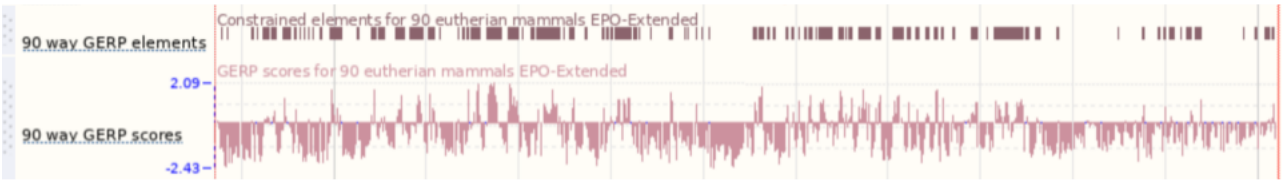
Mouse (*Mus musculus*)	1-to-1 View Gene Tree	Brca2 (ENSMUSG00000041147) Compare Regions (5:150,446,095-150,493,794:1) View Sequence Alignments	58.58 %	57.05 %	100	0.00	Yes

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.

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.

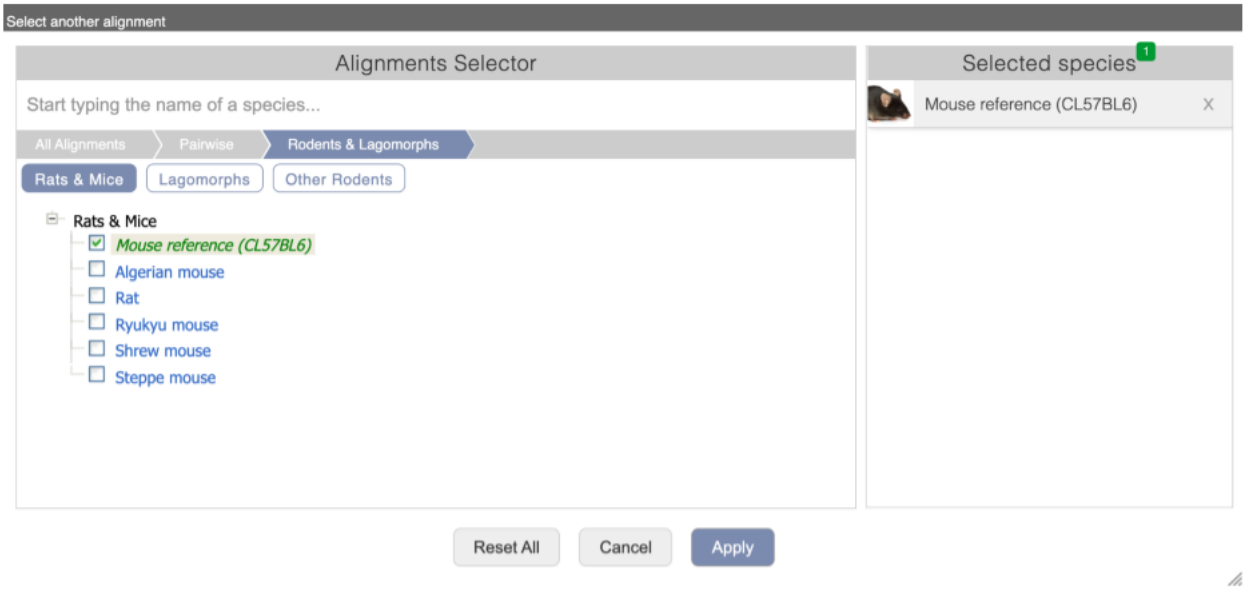
We can also look at individual species comparative genomics tracks in this view by clicking on *Configure this page*.

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.



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.

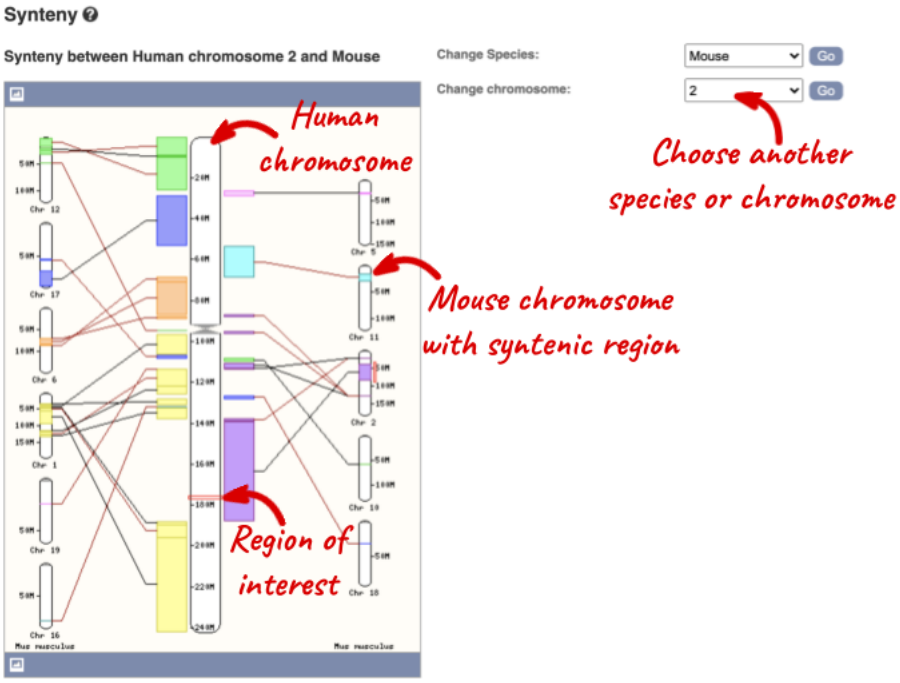
To compare with both contigs visually, go to *Region comparison*.

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.

The drop down allows you to configure each species separately.

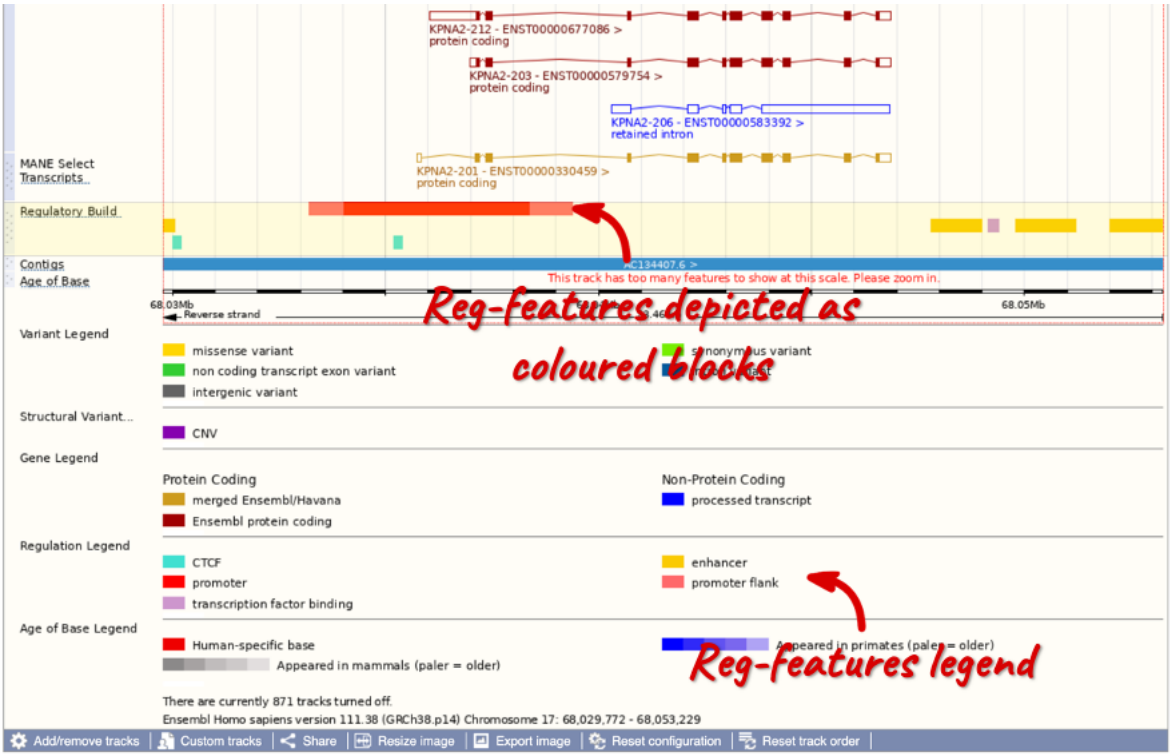
We can view large scale syntenic regions from our chromosome of interest. Click on *Synteny* in the left hand menu.



Features that regulate gene expression, Demo

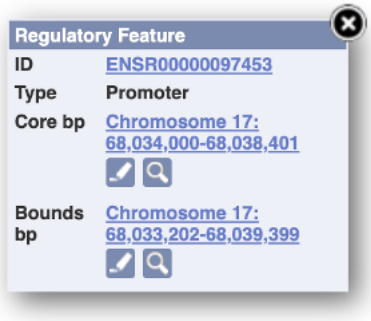
We're going to look for regulatory features in the region of a gene and investigate their activity in different cell types. We'll start by searching for the gene *KPNA2* and jumping to the **Location** tab. Scroll down to the **Region in detail** view and zoom out a little to see the gene as well as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

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

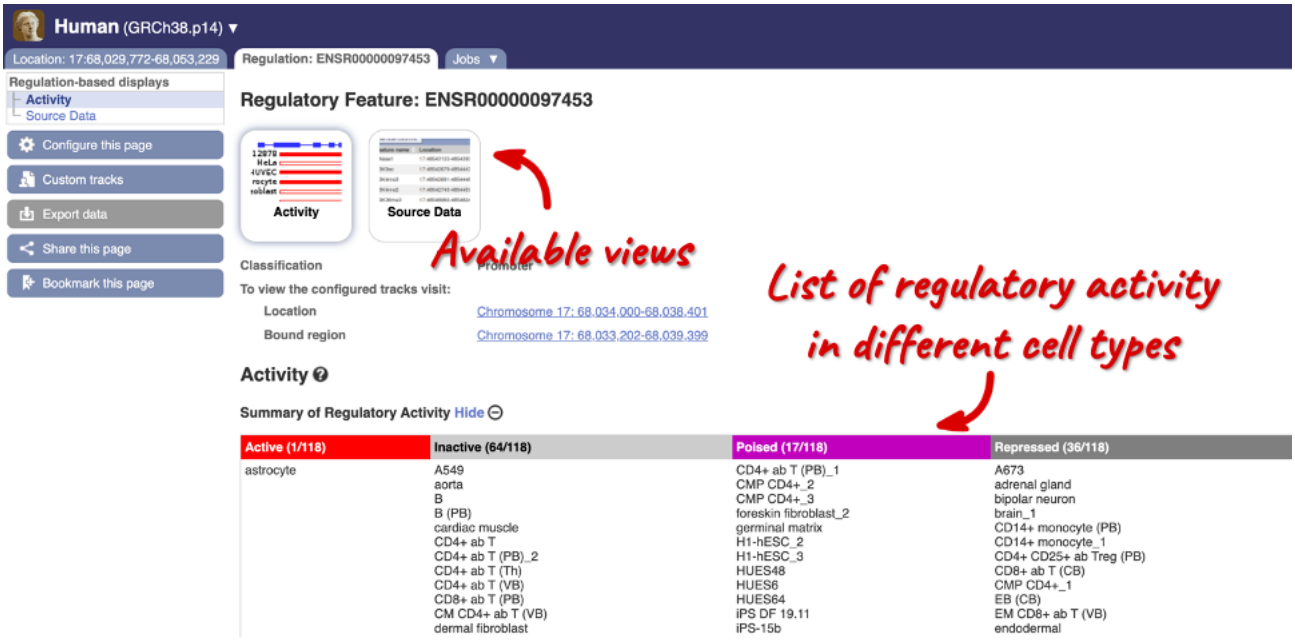


In this region we can see a number of regulatory features, including a red promoter with light red promoter flanks, cyan CTCF binding sites, yellow enhancers and lilac transcription factor (TF) binding sites (don't worry if you have zoomed out further or not as far and can see more/less). Refer to the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



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



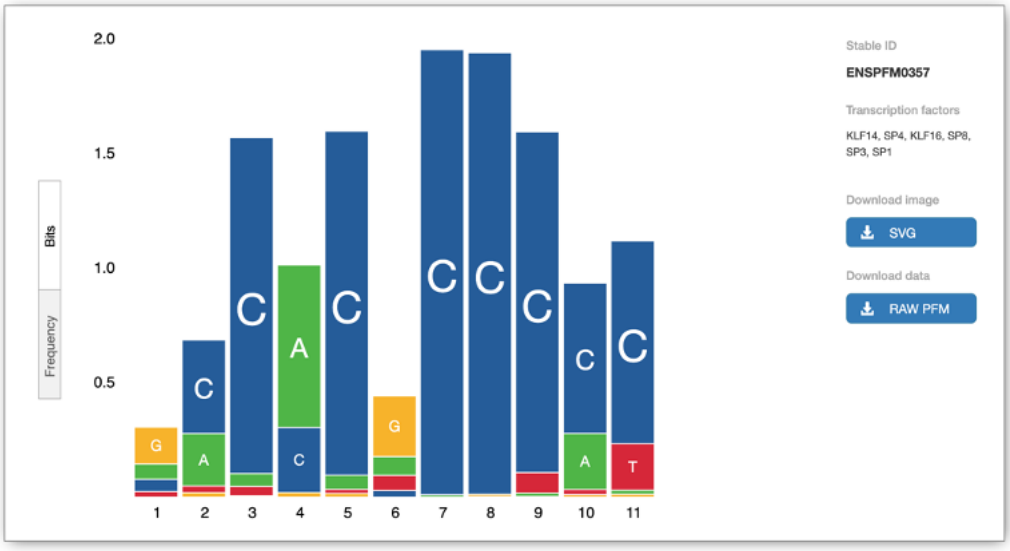
Here, you can find a summary of the activity of the promoter in the different cell types. Scroll down to **Summary of Regulatory Aactivity** to find out in which cells the promoter is active (the feature displays an active epigenetic signature, which can include evidence of open chromatin), inactive (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.

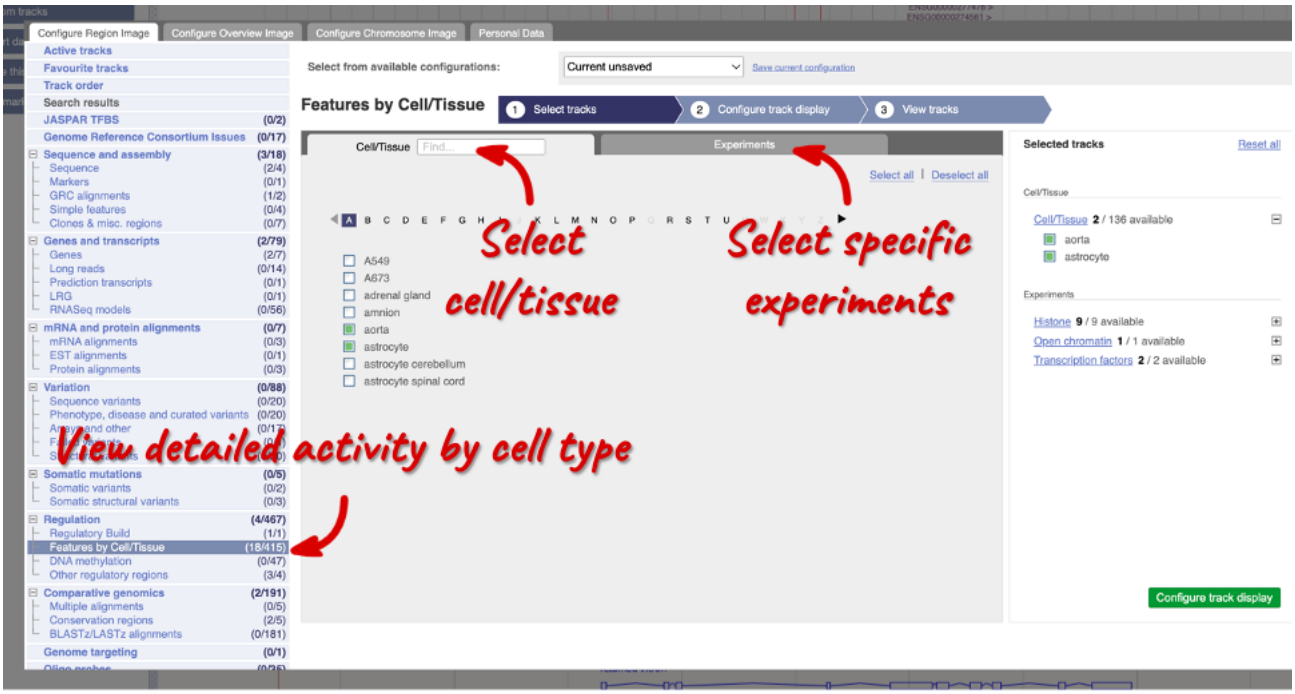
The **TarBase** track displays experimentally verified miRNA targets from TarBase.

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.

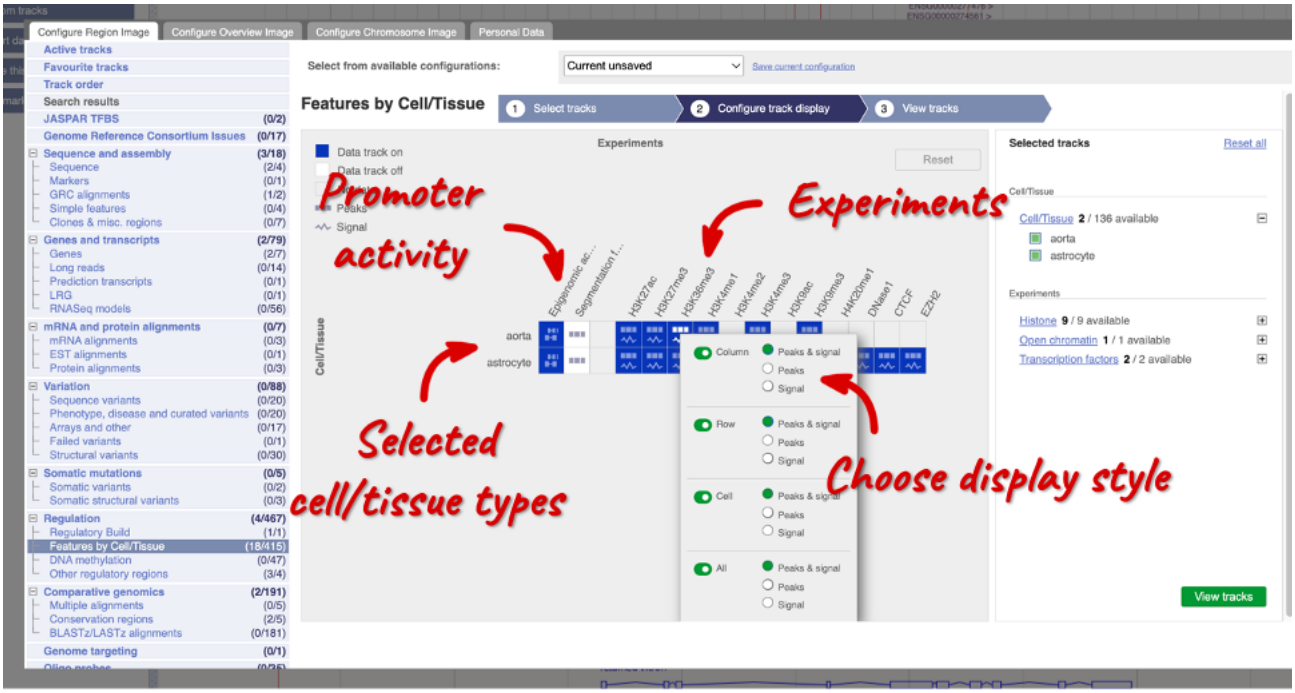
In the pop-up window, go to **Regulation** and click on **Features by Cell/Tissue** to view the detailed activity of the regulatory feature by cell type.



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

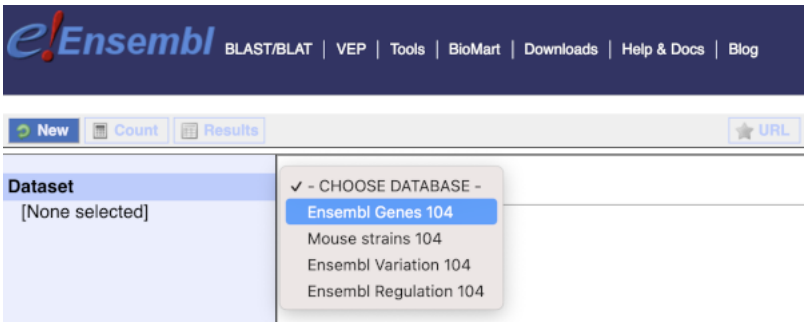
Follow these instructions to guide you through BioMart to answer the following query:

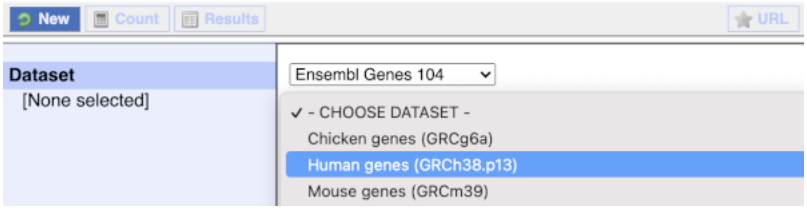
You have three questions about a set of human genes:
ESPN, MYH9, USH1C, CISD2, THRB, WHRN
(these are HGNC gene symbols. More details on the HUGO Gene Nomenclature Committee can be found on http://www.genenames.org)

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

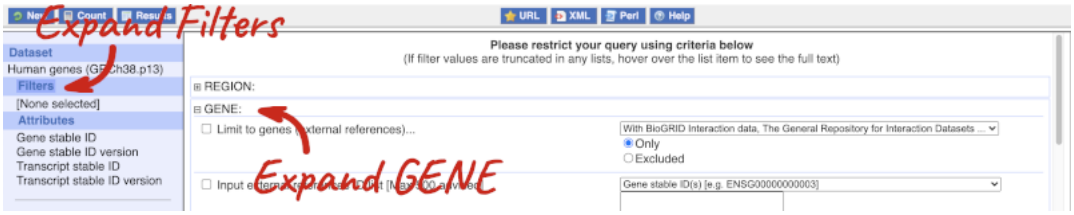
You cannot choose any filters or attributes until you've chosen your dataset. Your dataset is the data type you're working with. In this case we're going to choose human genes, so pick *Ensembl Genes* then *Human genes* from the drop-downs.



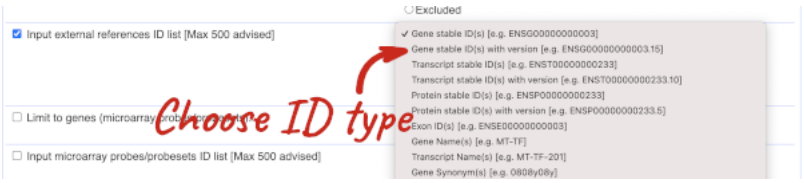


Now that you've chosen your dataset, the filters and attributes will appear in the column on the left. You can pick these in any order and the options you pick will appear.

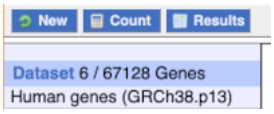
Click on *Filters* on the left to see the available filters appear on the main page. You'll see that there are loads of categories of Filters to choose from. You can expand these by clicking on them. For our query, we're going to expand *GENE*.



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)*.

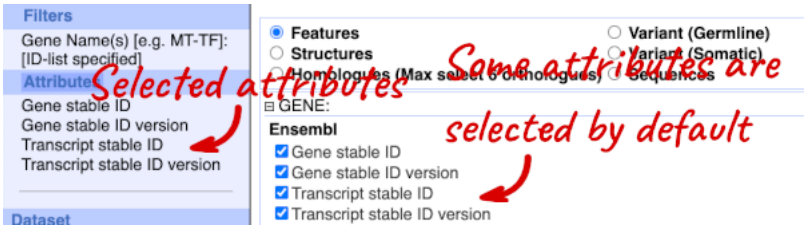


To check if the filters have worked, you can use the *Count* button at the top left, which will show you how many genes have passed the filter. If you get 0 or another number you don't expect, this can help you to see if your query was effective.



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.

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

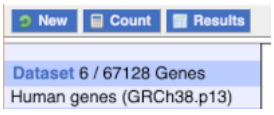


We can choose the attributes we want by clicking on them. For our query, we're going to select:

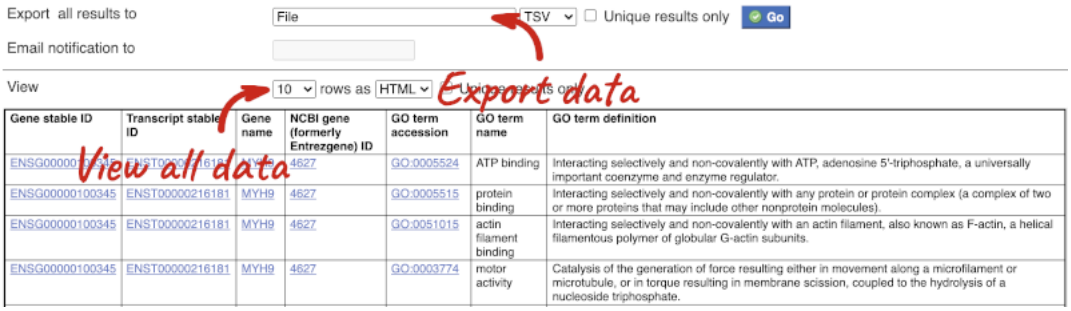
- GENE
  - Gene Name
- EXTERNAL
  - NCBI gene ID
  - GO term accession
  - GO term name
  - GO term definition

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.



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.



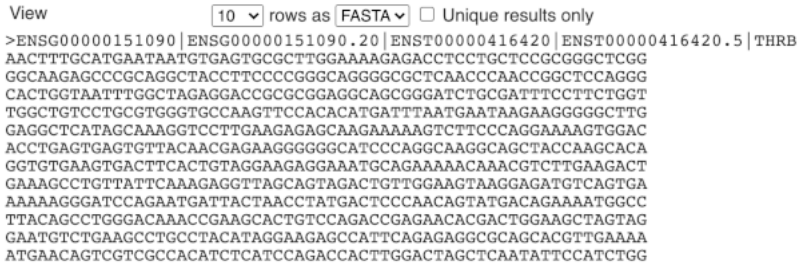
Gene stable ID	Transcript stable ID	Gene name	NCBI gene (formerly Entrezgene) ID	GO term accession	GO term name	GO term definition
ENSG00000100345	ENST00000216181	MYH9	4627	GO:0005524	ATP binding	Interacting selectively and non-covalently with ATP, adenosine 5'-triphosphate, a universally important coenzyme 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 grpahic.

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



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

Kinsella, R.J. et al
Ensembl BioMarts: a hub for data retrieval across taxonomic space.
http://europepmc.org/articles/PMC3170168

Custom data, Demo

Demo: Upload small files

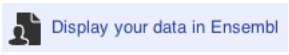
We have some patients that present with microcephaly and developmental delay. They all have large scale deletions on chromosome five:

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

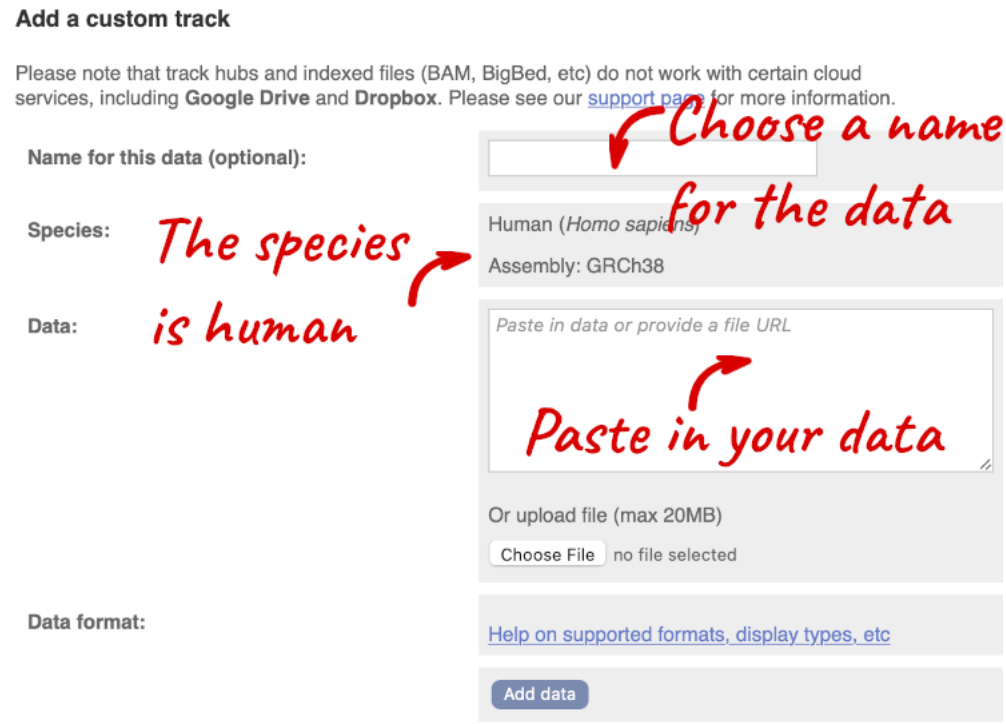
We can turn them into a BED file and view them in the genome browser:

```
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:



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

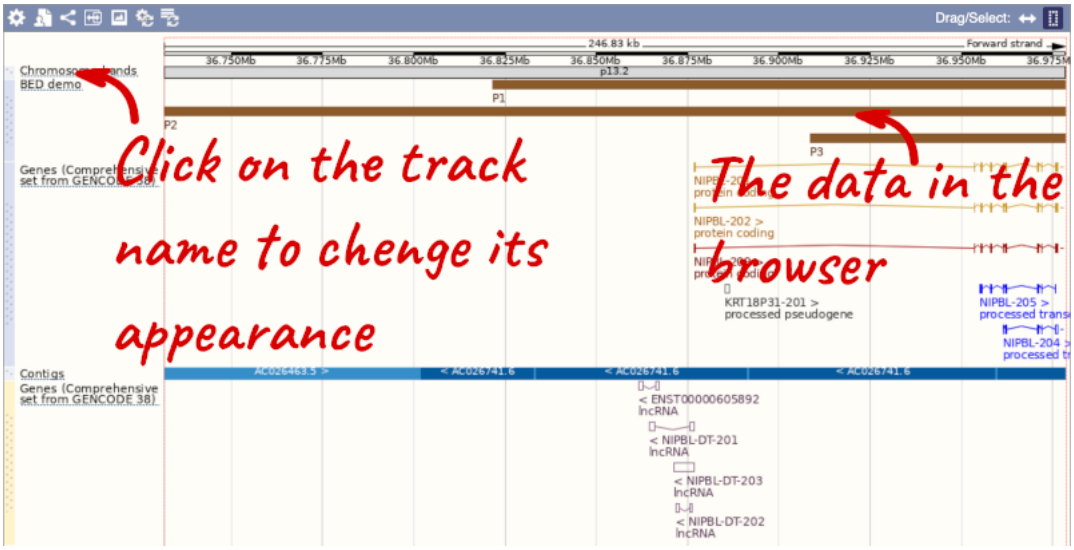
Go to Location view:

- Nearest region with data: 5:36731477-36978306

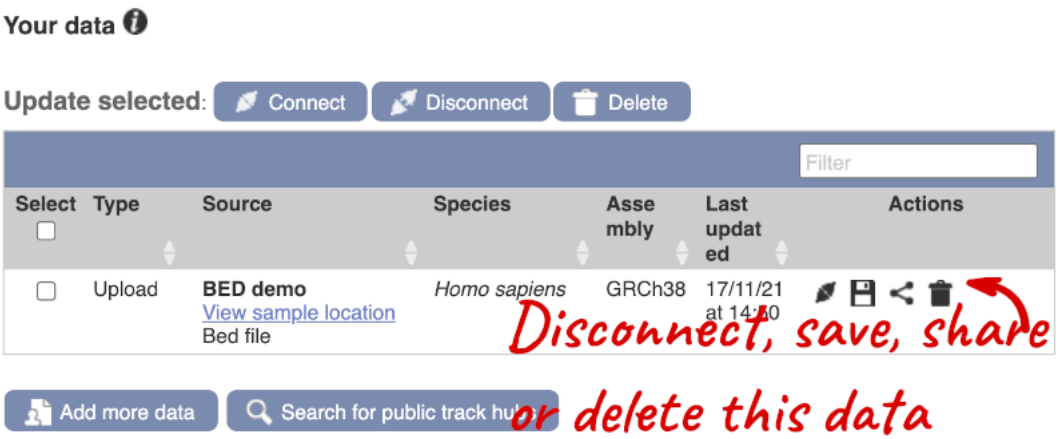
or

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*.

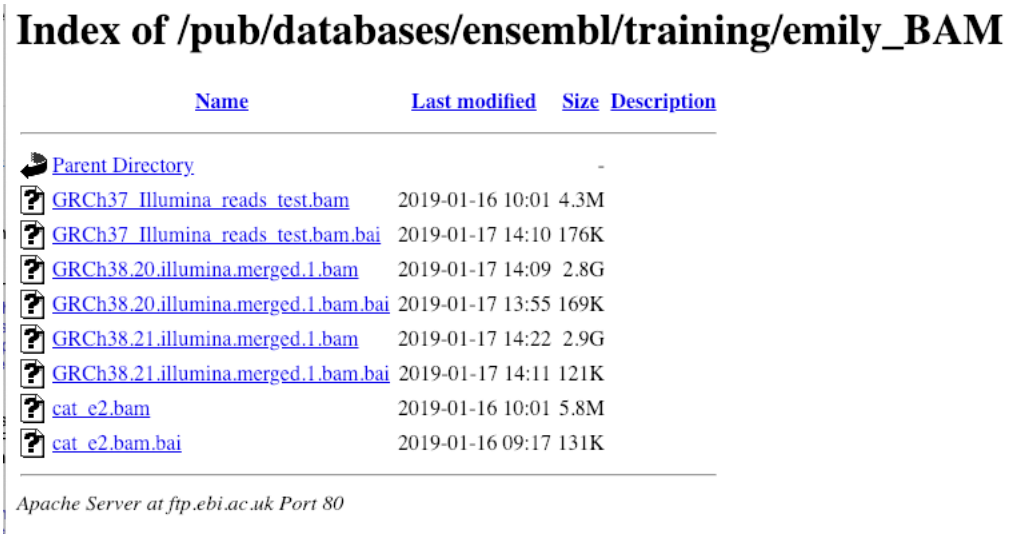


If you've got an Ensembl account, you can save this data to your account. Accounts are free to set up and allow you to save configurations and data, and share with groups.

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	

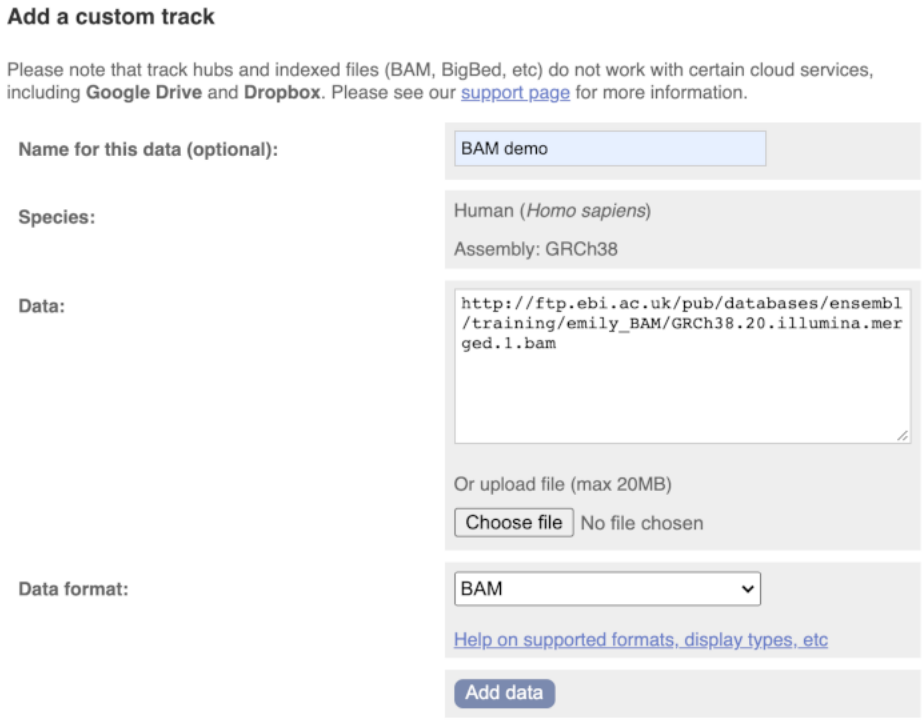
Apache Server at ftp.ebi.ac.uk Port 80

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.

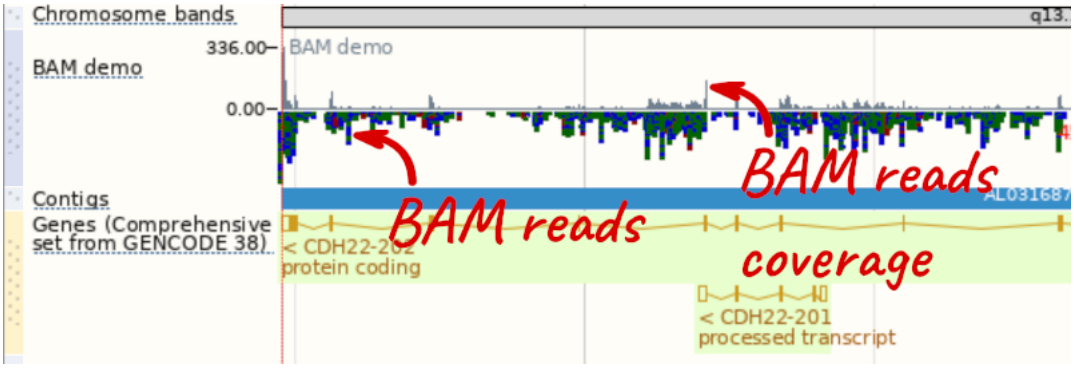
We get to the same dialogue box as before. This time we'll name our data *Illumina reads*.

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).



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

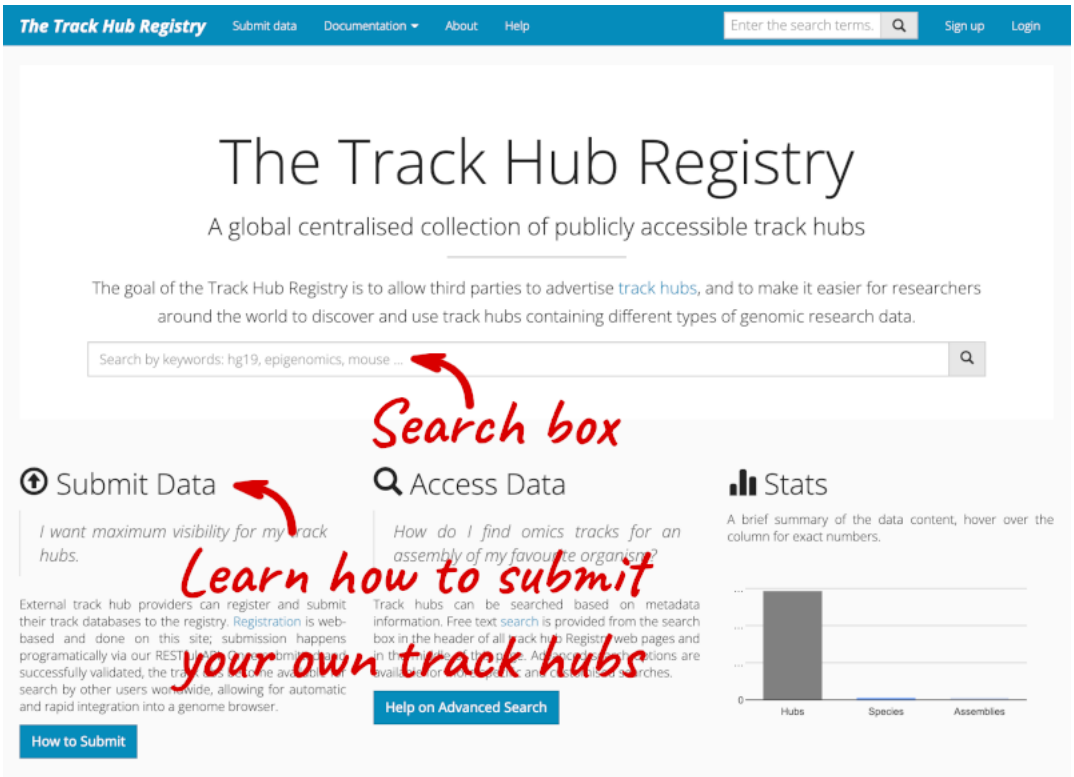
Our regulatory data incorporates data from sources such as ENCODE, Blueprint, and Roadmap Epigenomics. To see the data directly from these sources, you can add track hubs.

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

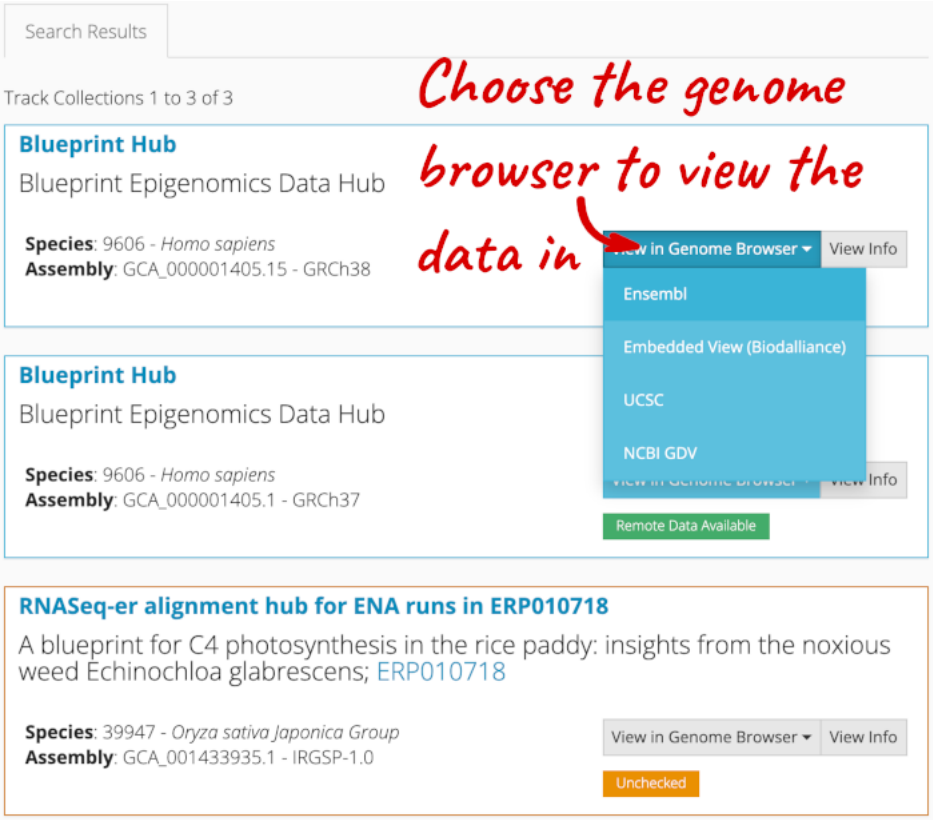
- Search for track hubs in the Track Hub Registry and choose to add them to your genome browser of choice.
- Search the track hub registry using the Track Hub Registry interface in Ensembl (there is a link from the homepage).

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

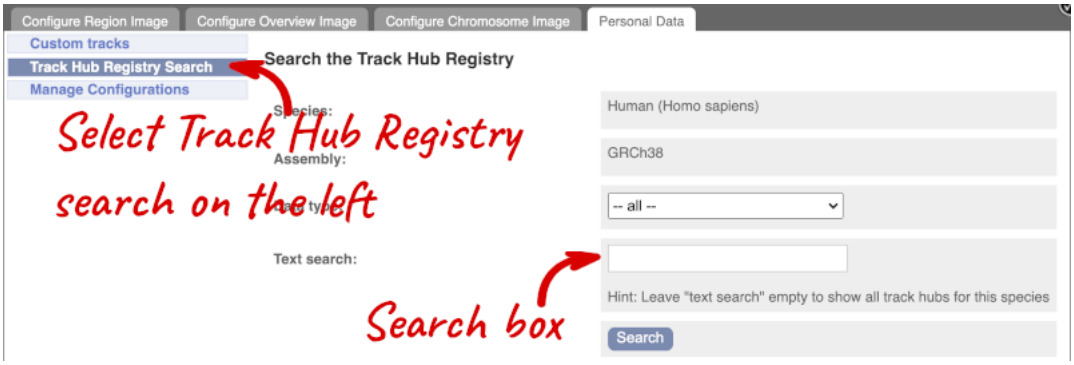
You can add track hubs to view in Ensembl directly via the Track Hub Registry. Go to the Track Hub Registry homepage and search for *blueprint*.



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.



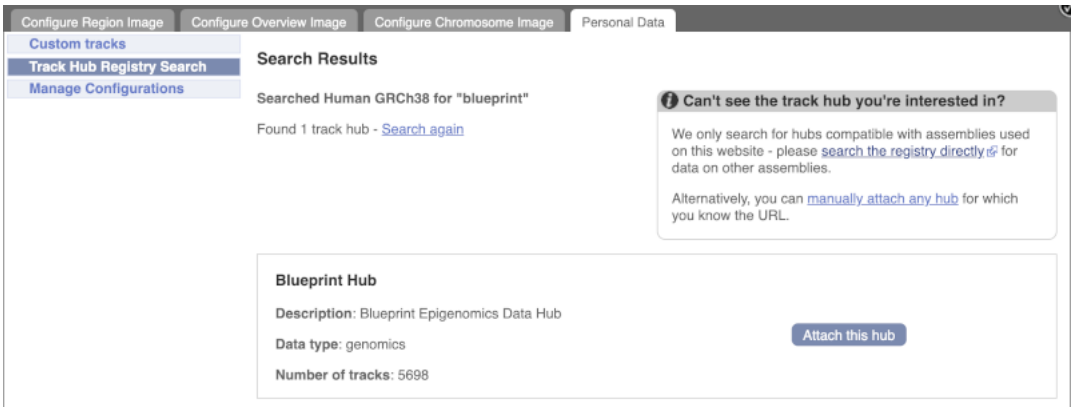
Alternatively, you can add track hubs by searching the Track Hub Registry through Ensembl. Click the *Custom tracks -> Track Hub Registry Search* in any region view within Ensembl.



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

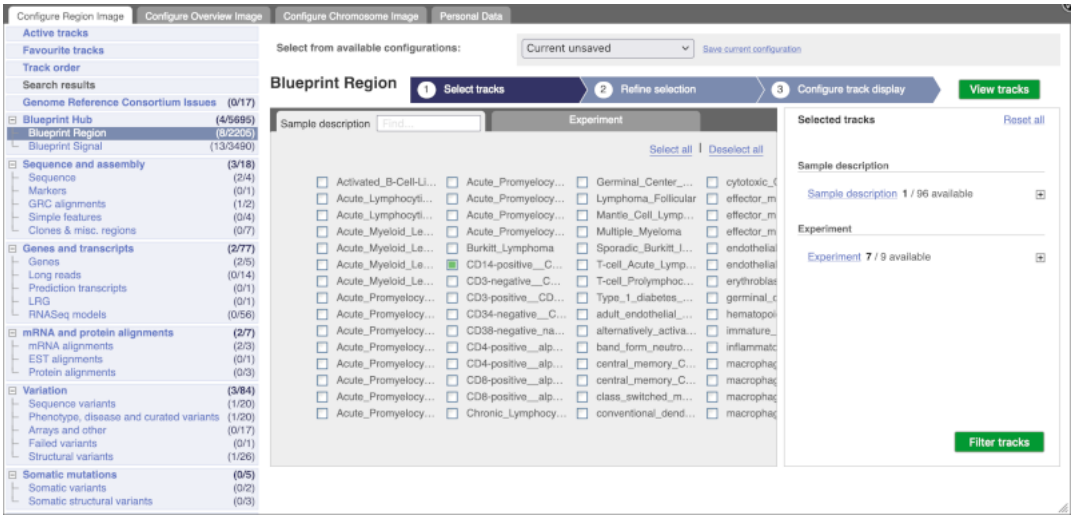
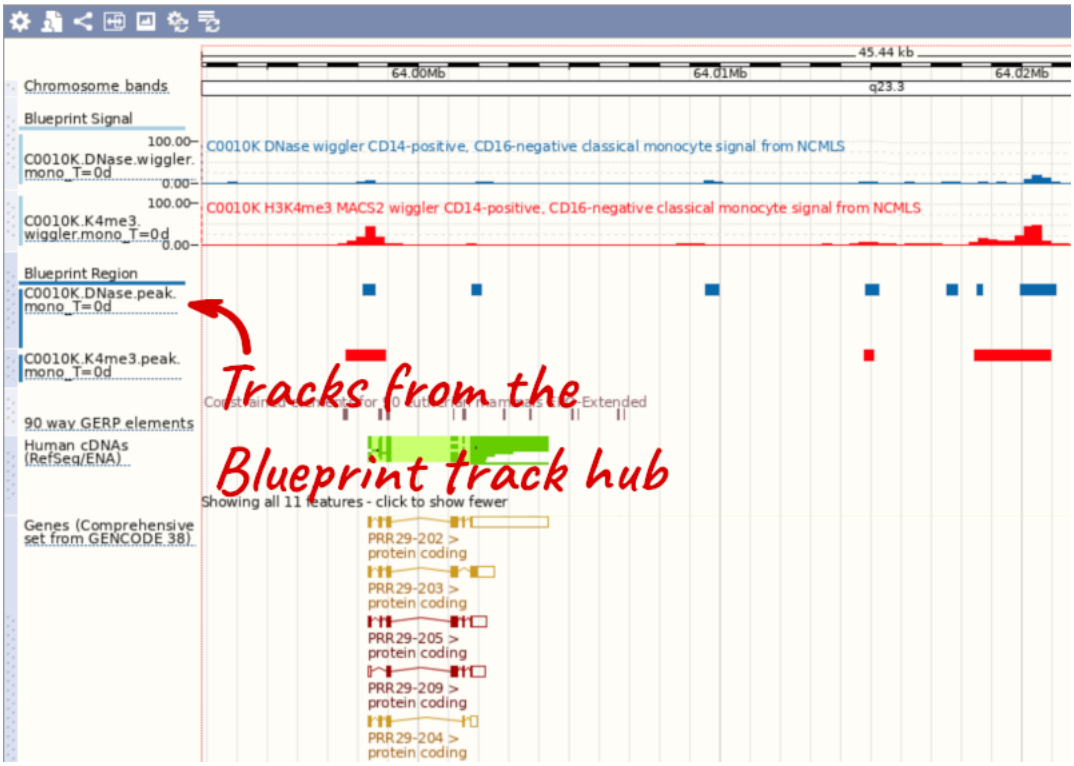
Search for *blueprint*.

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



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.