Module 1: Browsing genomes with Ensembl

Aims

- Explain why it can be useful to look at the whole genome.
- Demonstrate some of the features and applications of Ensembl.

Introduction

Ensembl is a joint project between the EBI (European Bioinformatics Institute) and the Wellcome Trust Sanger Institute that annotates **chordate** genomes (i.e. vertebrates and closely related invertebrates with a notochord such as sea squirt). Gene sets from model organisms such as yeast and worm are also imported for comparative analysis by the Ensembl 'compara' team. Most annotation is updated every two months, leading to increasing Ensembl versions (such as version 79), however the gene sets are determined less frequently. A sister browser at www.ensemblgenomes.org is set up to access non-chordates, namely bacteria, plants, fungi, metazoa, and protists.

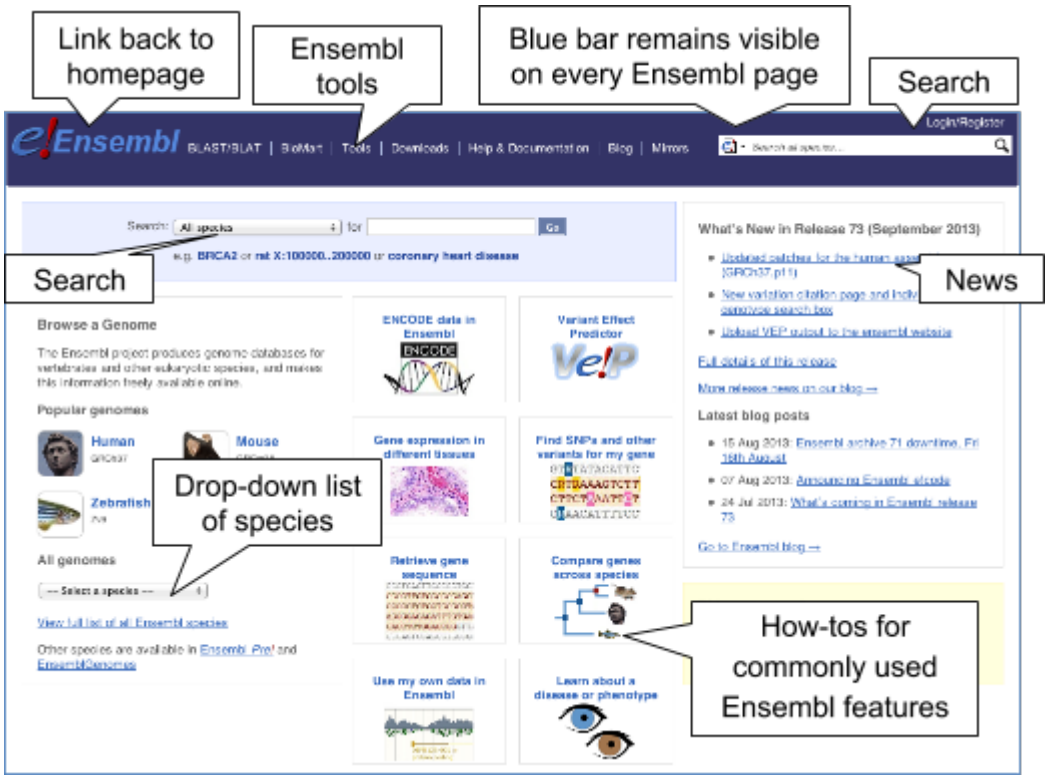
Ensembl provides genes and other **annotation** such as regulatory regions, conserved base pairs across species, and sequence variations. The Ensembl gene set is based on protein and mRNA evidence in **UniProtKB** and **NCBI RefSeq** databases, along with manual annotation from the **VEGA/Havana** group. All the data are freely available and can be accessed via the web browser at www.ensembl.org. Perl programmers can directly access Ensembl databases through an Application Programming Interfaces (**Perl APIs**). Gene sequences can be downloaded from the Ensembl browser itself, or through the use of the **BioMart** web interface, which can extract information from the Ensembl databases without the need for programming knowledge by the user.

While browsers can be very useful tools, they do not provide the definitive answer to every question!

Also, new data and updates make genome browsing a fluid, changing, and improving, process.

Demo: Exploring the Ensembl genome browser

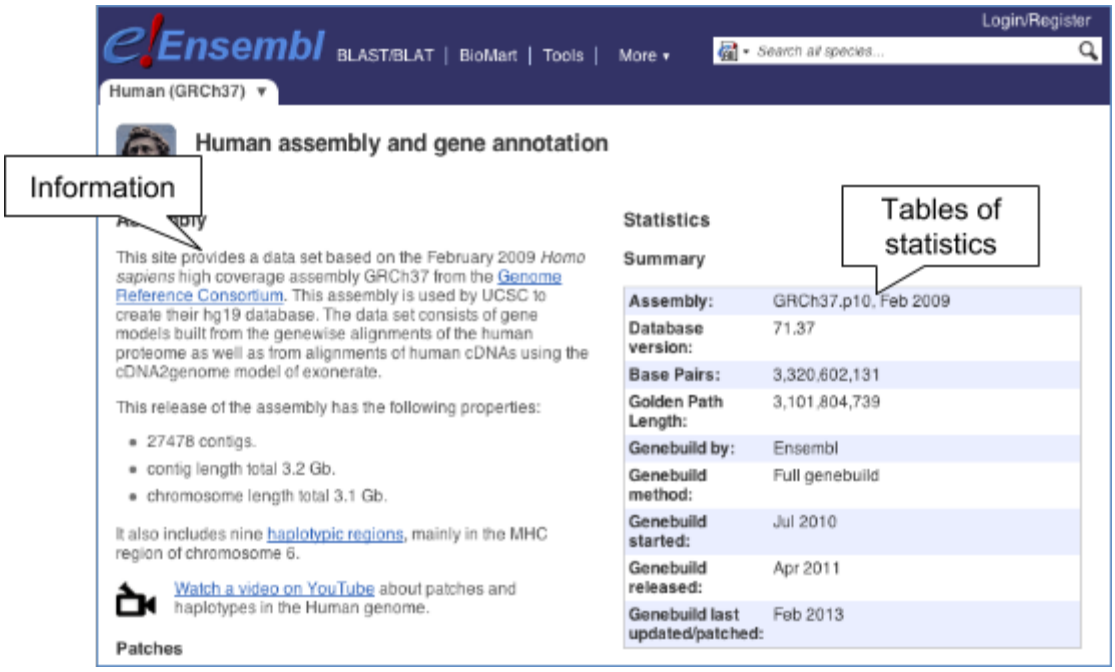
The front page of Ensembl is found at ensembl.org. It contains lots of information and links to help you navigate Ensembl:



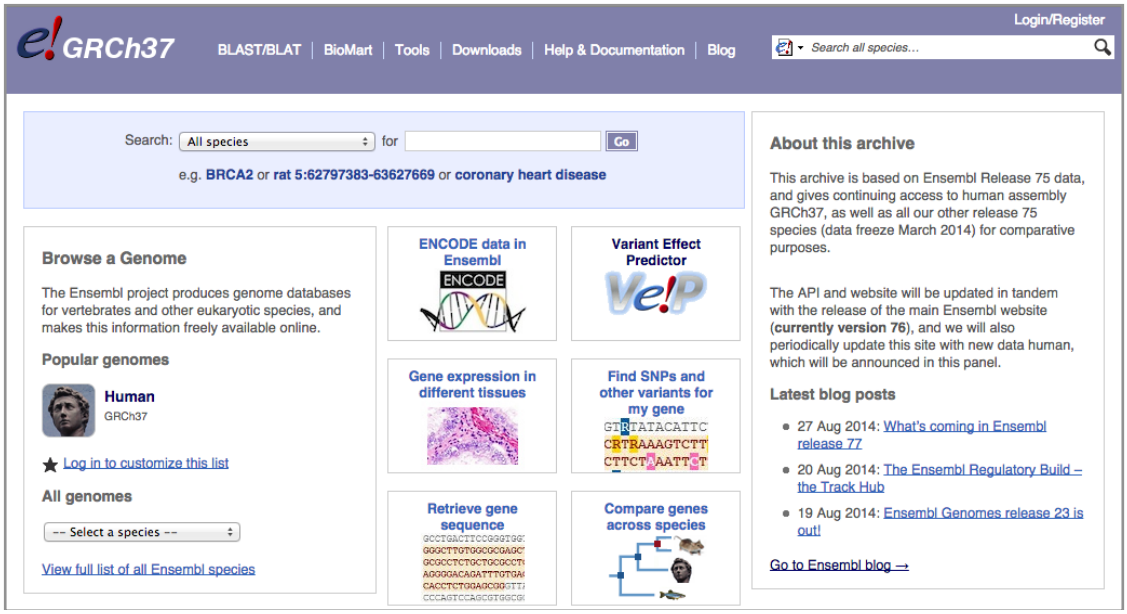
Click on View full list of all Ensembl species.

Click on the common name of your species of interest to go to the species homepage. We'll click on Human.

To find out more about the genome assembly and genebuild, click on More information and statistics.



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



Demo: The Region in detail view

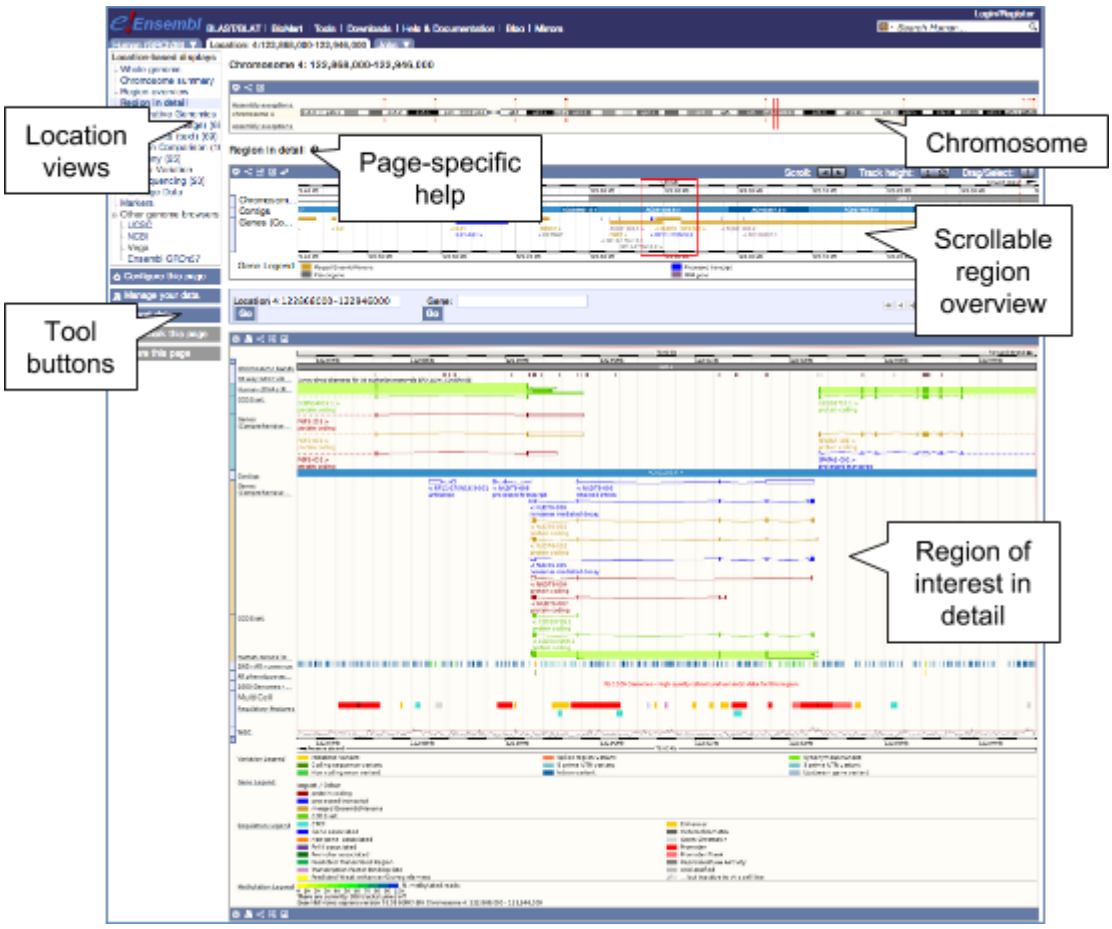
Start at the Ensembl front page, ensembl.org. You can search for a region by typing it into a search box, but you have to specify the species.

Type (or copy and paste) human 4:122868000-122946000 into either search box.

or

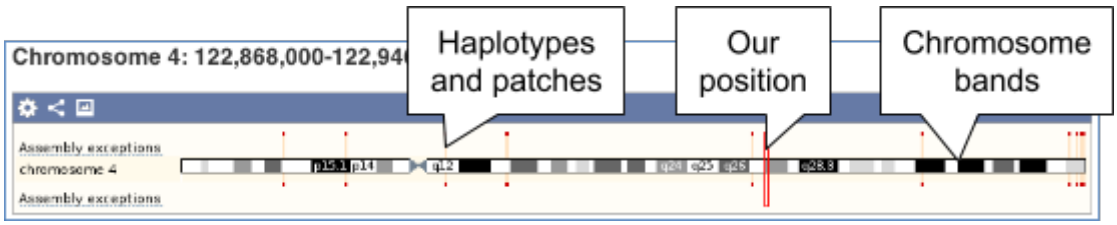
Press Enter or click Go to jump directly to the **Region in detail** Page.

Click on the button to view page-specific help. The help pages provide links to Frequently Asked Questions, a Glossary, Video Tutorials, and a form to Contact HelpDesk.There is a help video on this page at http://youtu.be/tTKEvgPUq94.

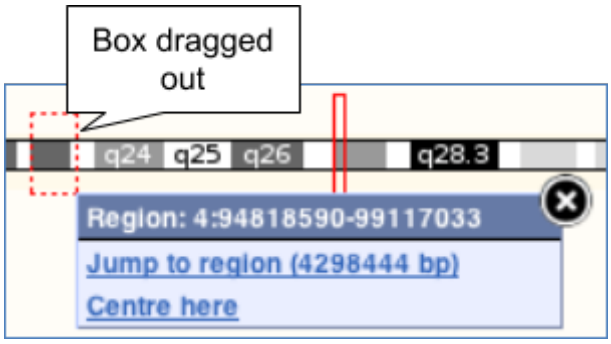


The Region in detail page is made up of three images, let's look at each one on detail.

The first image shows the chromosome:

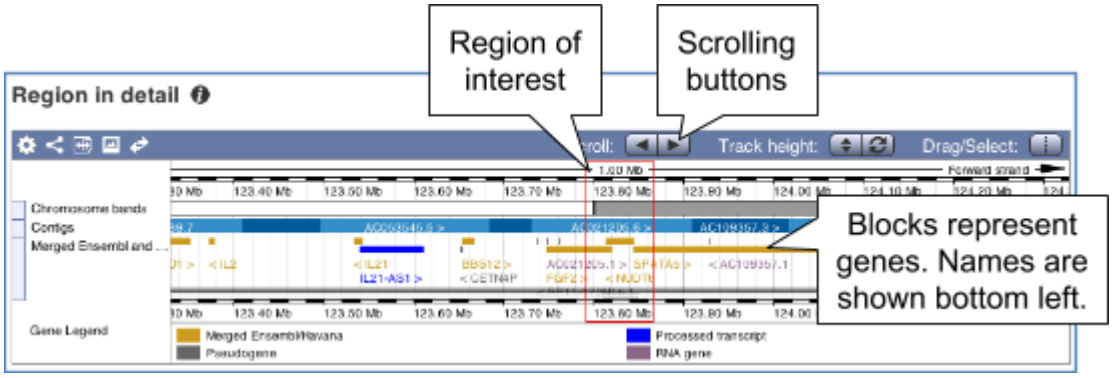


You can jump to a different region by dragging out a box in this image. Drag out a box on the chromosome, a pop-up menu will appear.



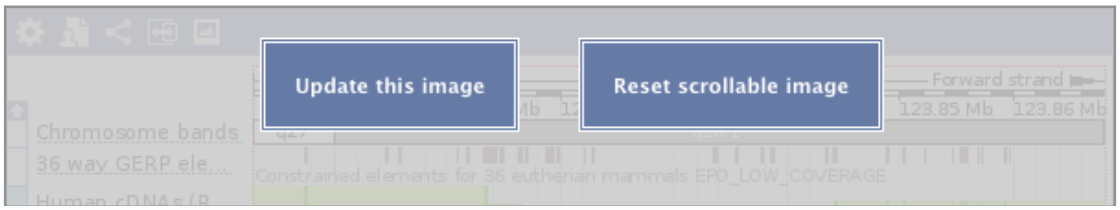
If you wanted to move to the region, you could click on Jump to region (### bp). For now, we'll close the pop-up by clicking on the X on the corner.

The second image shows a 1Mb region around our selected region. This view allows you to scroll back and forth along the chromosome.

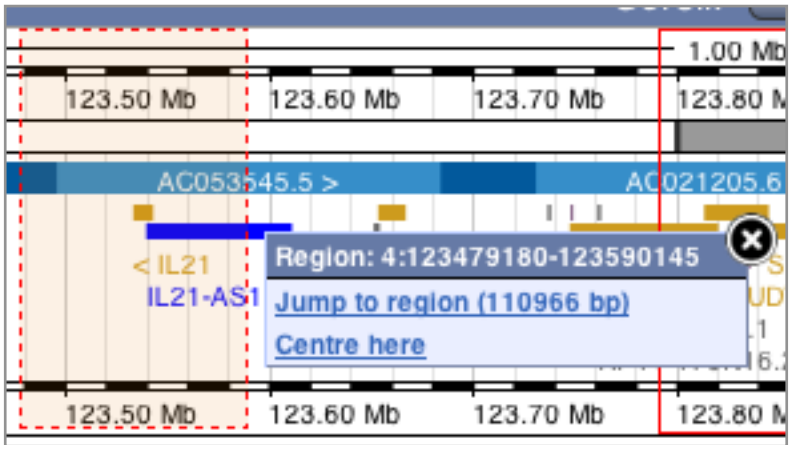


At the moment the gene track is set to a fixed height. Click on the Automatic track height button  to expand the image to include all possible data in the track.

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.

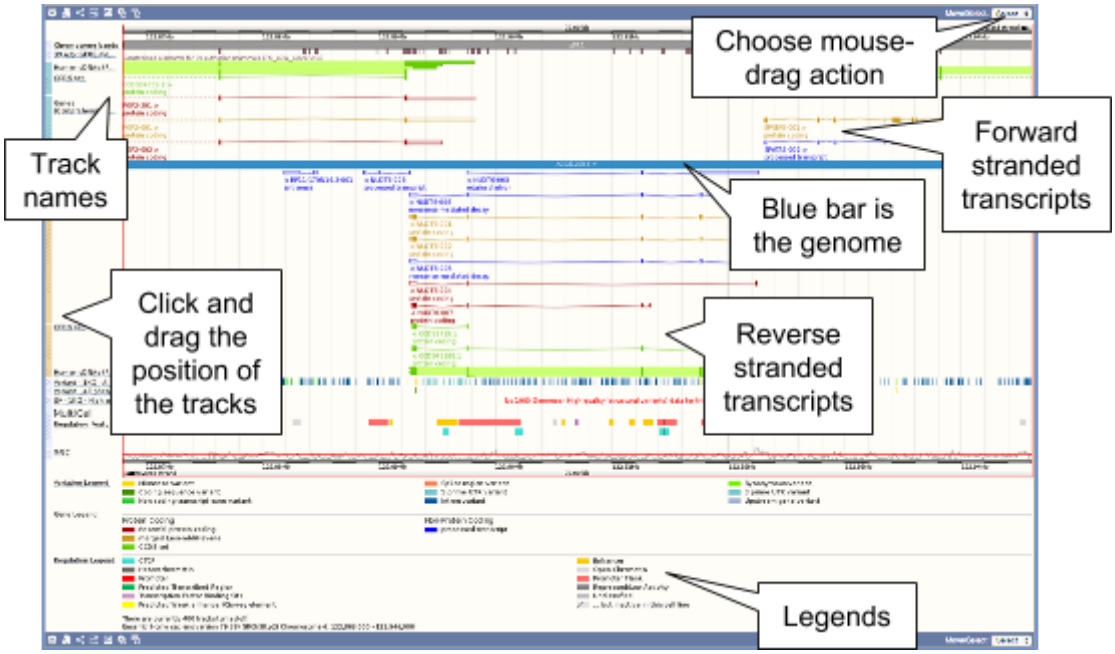


You can also drag out and jump to a region. Either hold down shift and drag in the image, or click on the Drag/Select button  to change the action of your mouse click, and drag out a box.



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

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

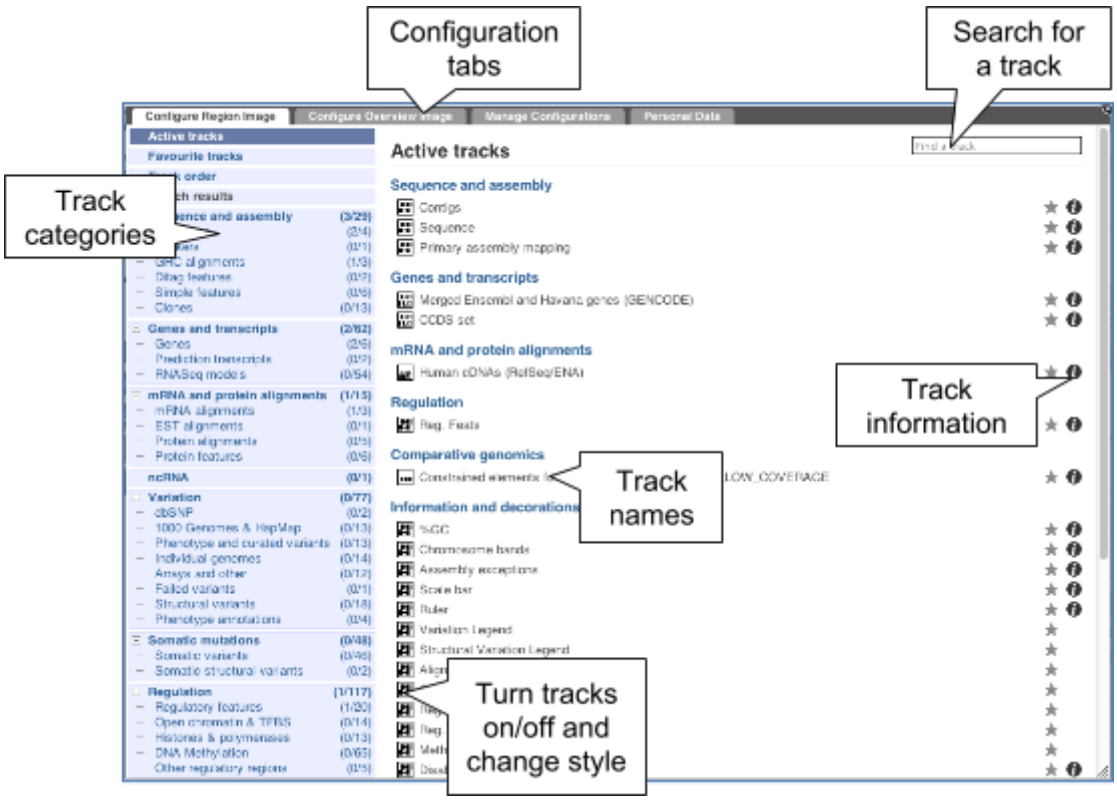


Click on the Move/Select option at the top or bottom right to switch mouse action. On Move, you can click and drag left or right to move along the genome, the page will reload when you drop the mouse button. On Select you can drag out a box to highlight or zoom in on a region of interest.

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.

You can put some tracks on in different styles; more details are in this FAQ: http://www.ensembl.org/Help/Faq?id=335.

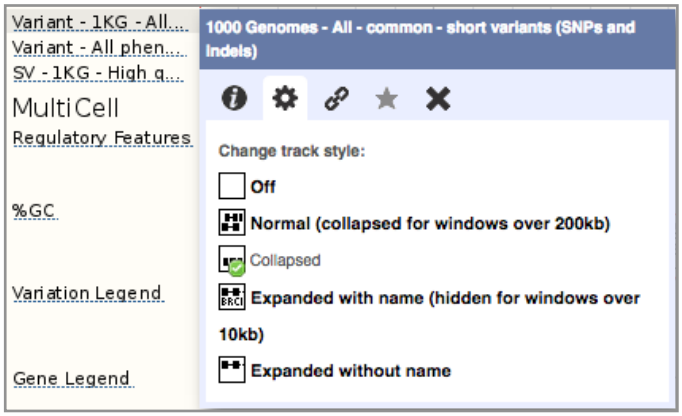


Let's add some tracks to this image. Add:

- Proteins (mammal) from UniProt – Labels
- dbSNP variants – 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.

We can also change the way the tracks appear by hovering over the track name then the cog wheel to open a menu. We can move tracks around by clicking and dragging on the bar to the left of the track name.



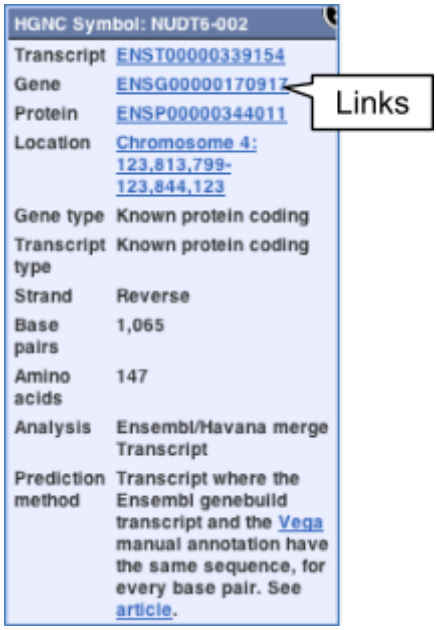
Now that you've got the view how you want it, you might like to show something you've found to a colleague or collaborator. Click on the Share this page button to generate a link. Email the link to someone else, so that they can see the same view as you, including all the tracks you've added. These links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

Demo: The gene tab

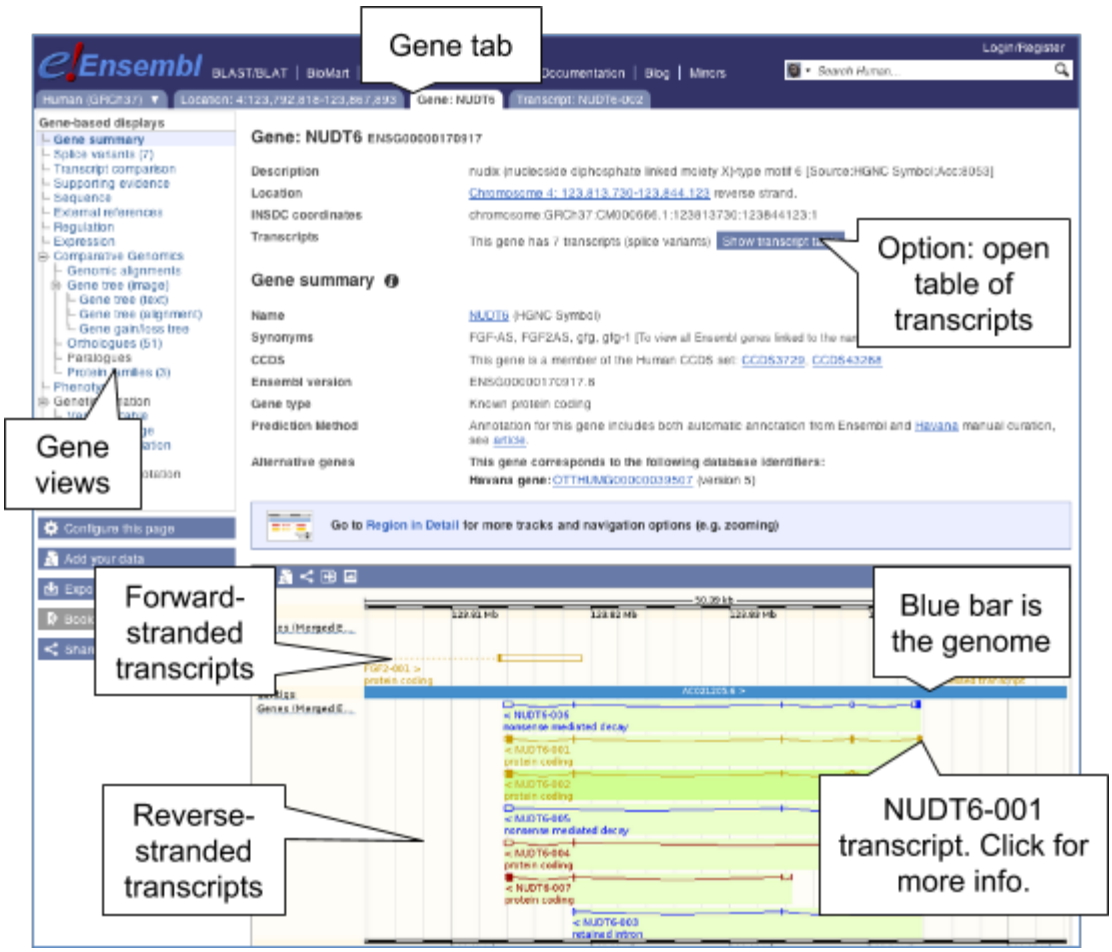
Now let's look at a gene.

If you click on any one of the transcripts in the Region in detail image, a pop-up menu will appear, allowing you to jump directly to that gene or transcript.

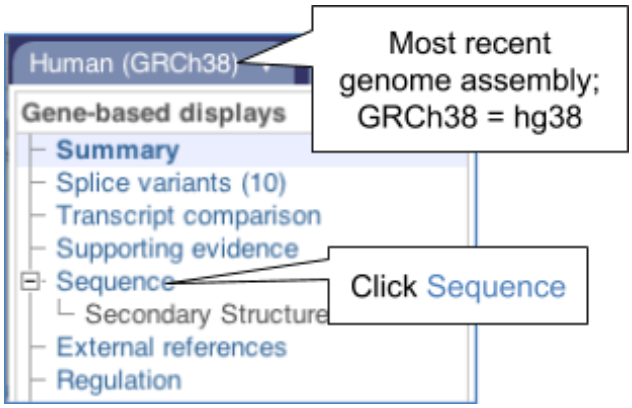


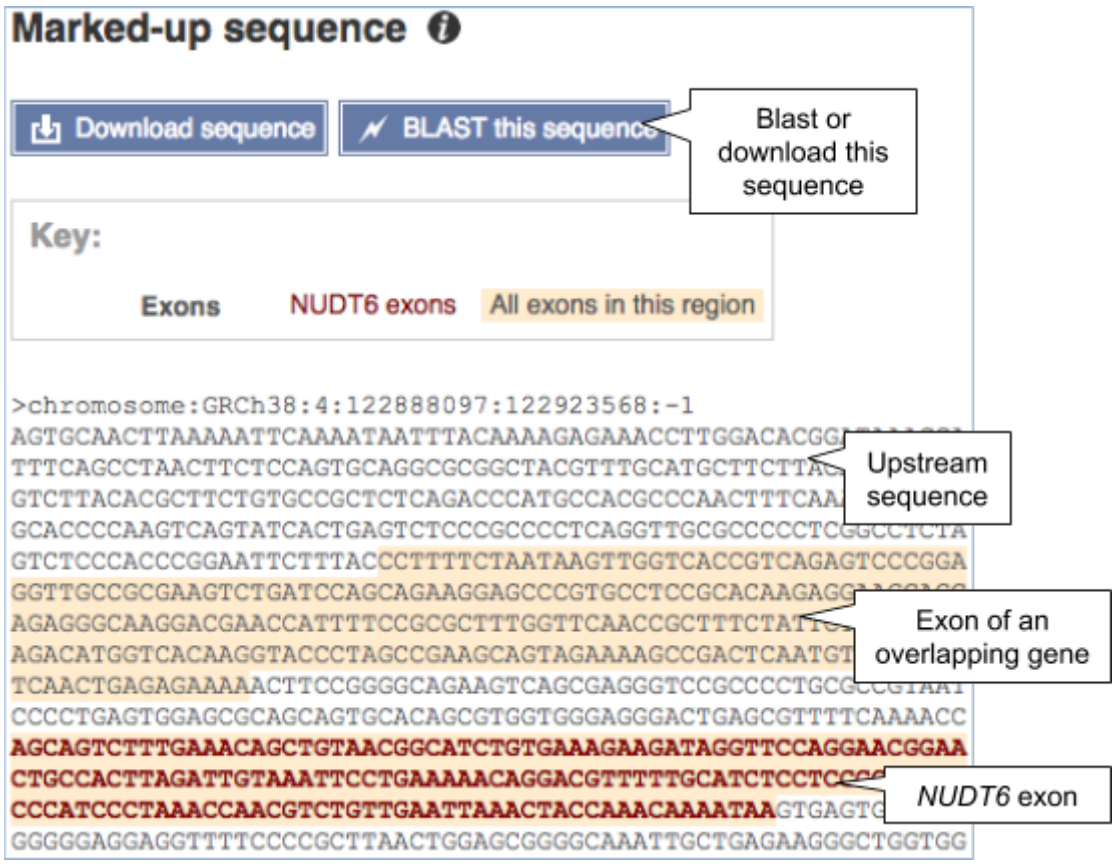
Click on any of the transcripts for NUDT6, then click on the Ensembl gene ID ENSG00000170917.

The **Gene tab** should open:

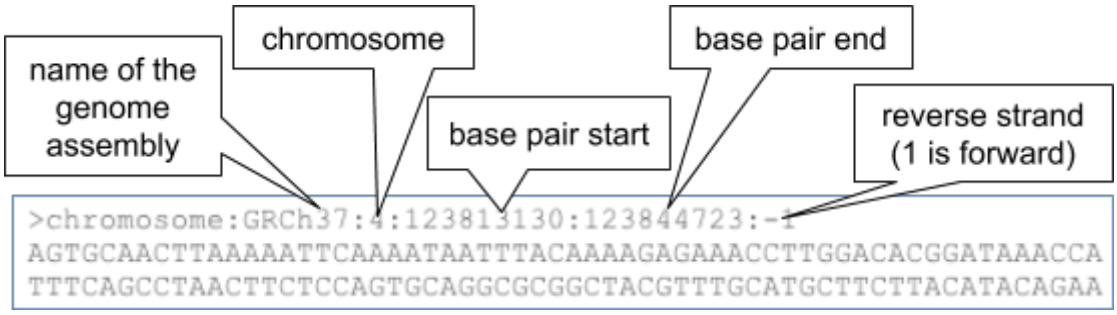


Let's walk through some of the links in the left hand navigation column. How can we view the genomic sequence? Click Sequence at the left of the page.

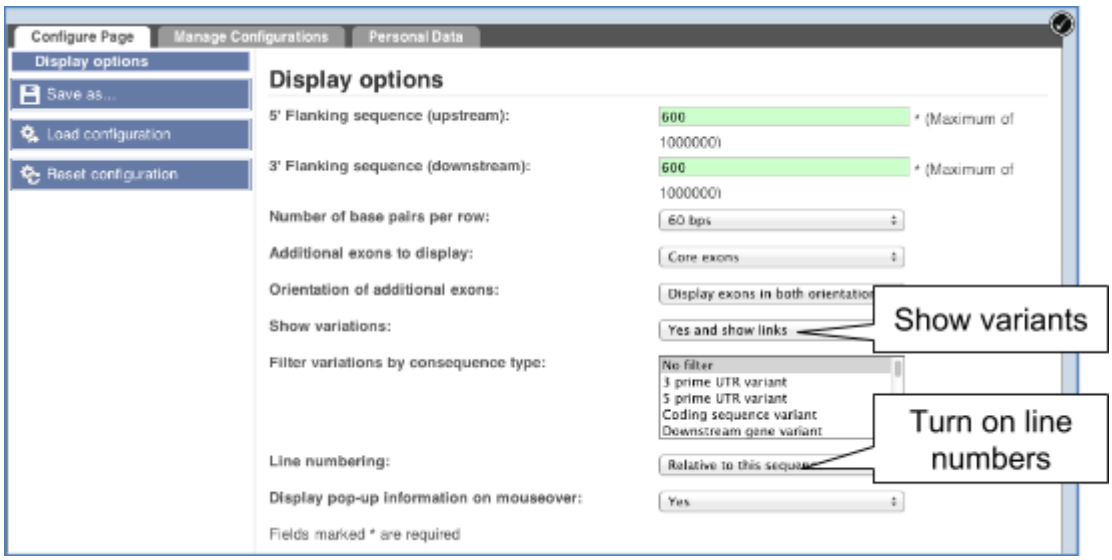




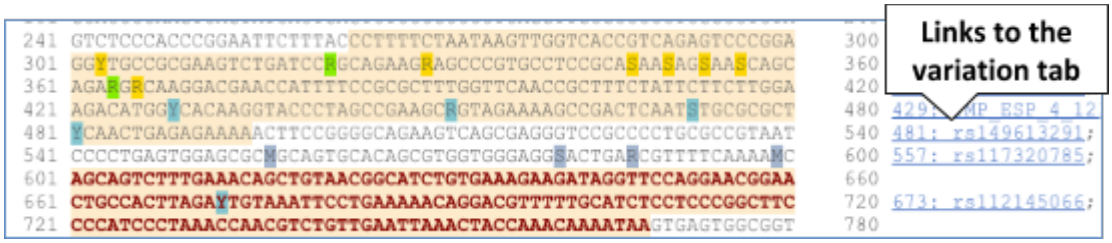
The sequence is shown in FASTA format. Take a look at the FASTA header.



Exons are highlighted within the genomic sequence. Variations can be added with the Configure this page link found at the left. Click on it now.

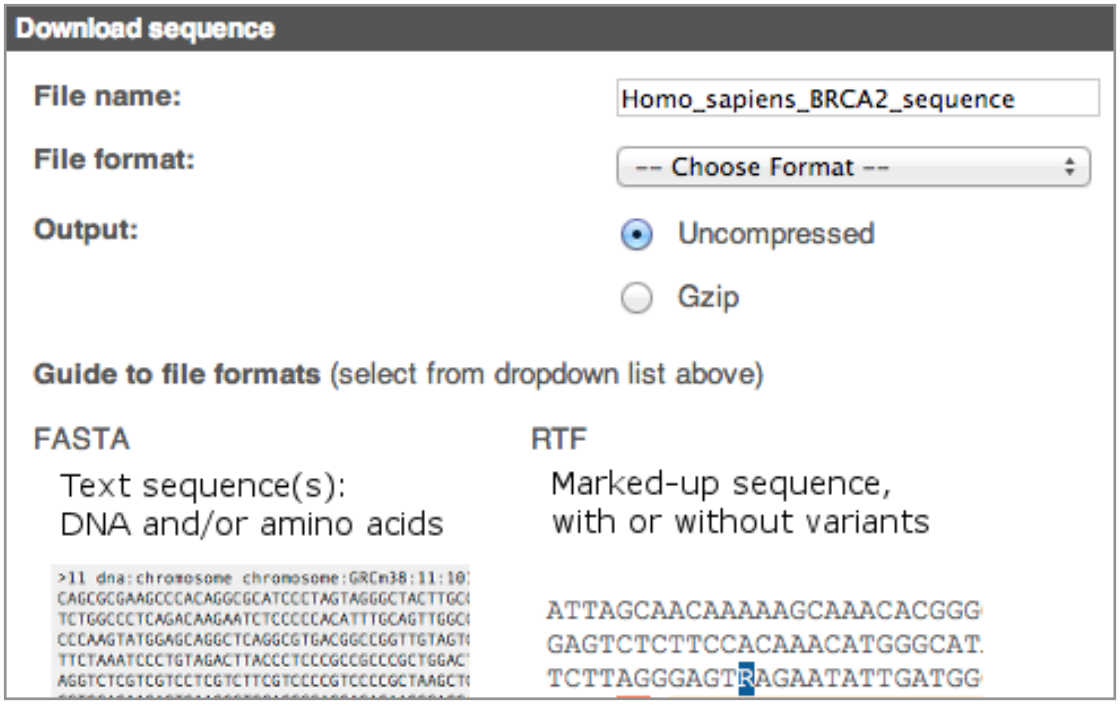


Once you have selected changes (in this example, Show variations and Line numbering) click at the top right.

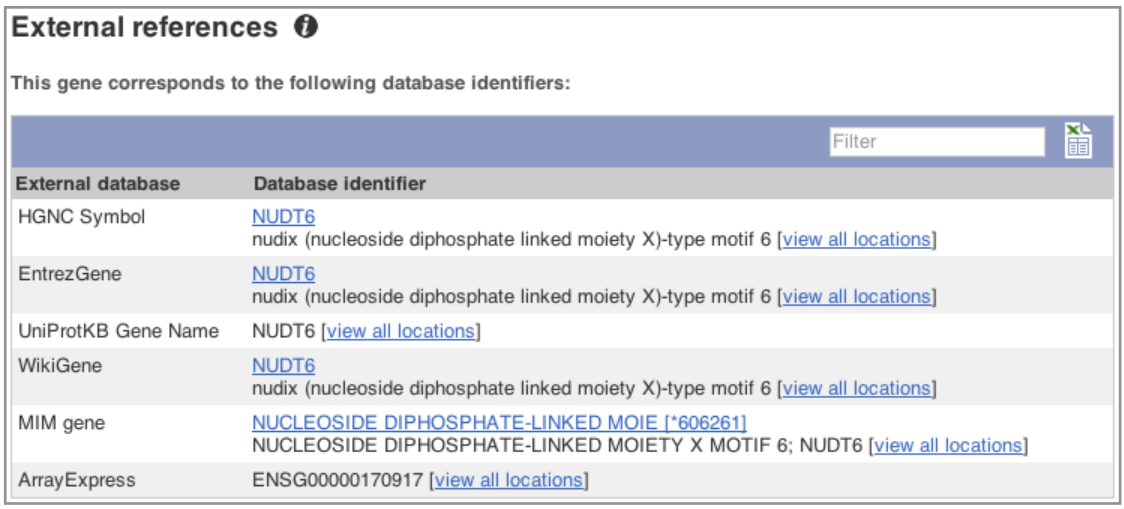


You can download this sequence by clicking in the Download sequence button above the sequence:

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. This button is available for all sequence views.



Can our gene be found in other databases? Go up the left-hand menu to External references:



External references

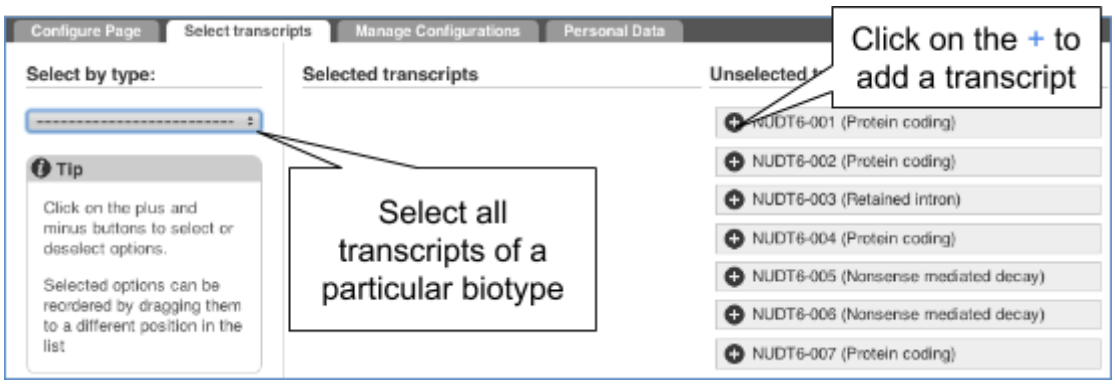
This gene corresponds to the following database identifiers:

External database	Database identifier
HGNC Symbol	NUDT6 nudix (nucleoside diphosphate linked moiety X)-type motif 6 [view all locations]
EntrezGene	NUDT6 nudix (nucleoside diphosphate linked moiety X)-type motif 6 [view all locations]
UniProtKB Gene Name	NUDT6 [view all locations]
WikiGene	NUDT6 nudix (nucleoside diphosphate linked moiety X)-type motif 6 [view all locations]
MIM gene	NUCLEOSIDE DIPHOSPHATE-LINKED MOIE [*606261] NUCLEOSIDE DIPHOSPHATE-LINKED MOIETY X MOTIF 6; NUDT6 [view all locations]
ArrayExpress	ENSG00000170917 [view all locations]

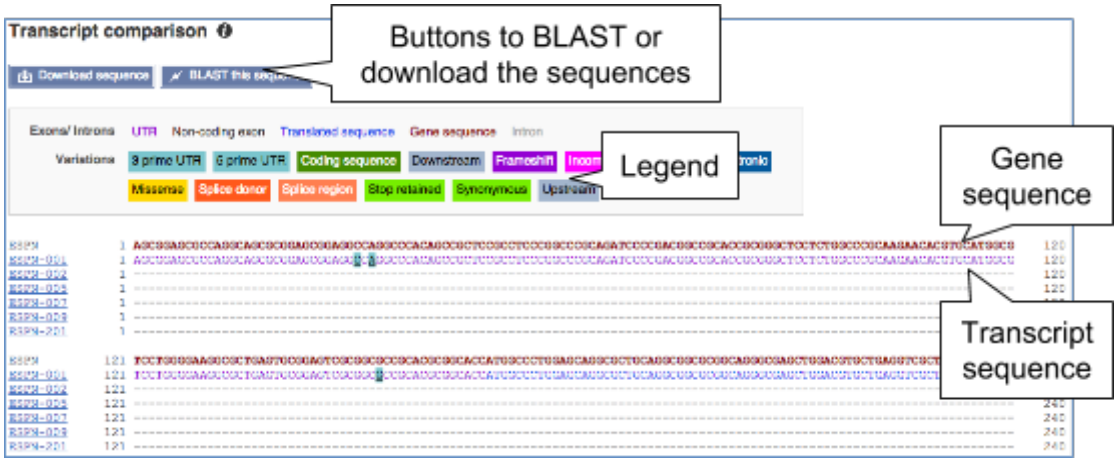
This contains links to the gene in other projects, such as EntrezGene.

To find out more about the individual transcripts of this gene, click on Transcript comparison in the left-hand menu.

You must now choose the transcripts you'd like to see, click on the blue Select transcripts button.



Let's select all the protein-coding transcripts, then close the menu.



Demo: The transcript tab

Let's now explore one splice isoform. Click on Show transcript table at the top.

Have a look at the largest one, NUDT6-001.

Show/hide columns (1 hidden) Filter

Name	Transcript ID	bp	Protein	Biotype	CCDS	RefSeq	Flags
NUDT6-001	ENST00000304430	1169	316 aa	Protein coding	CCDS43268	NM_007083 NP_009014	TSL:1 GENCODE basic APPRIS PI

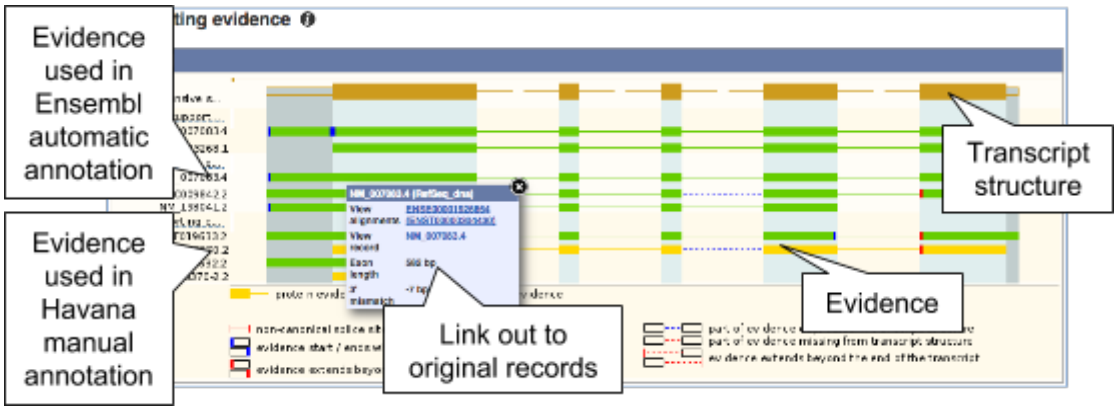
If we were to only choose one transcript to analyse, we would choose this one because it has:

- Matching annotation between automatic and manual methods (Gold in Biotype column).
- Matching annotation with RefSeq giving it a CCDS.
- High transcript support (TSL1).
- A complete structure, making it a member of GENCODE Basic.
- The highest protein expression, making it an APPRIS Principal Isoform.

Click on the ID, ENST00000304430.

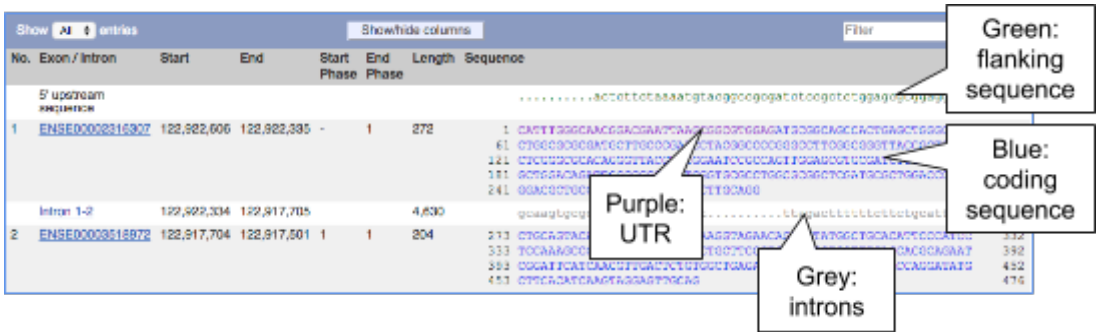
You are now in the Transcript tab for NUDT6-001. The left hand navigation column provides several options for the transcript NUDT6-001.

For detailed information on the support for this transcript, click on Supporting evidence.

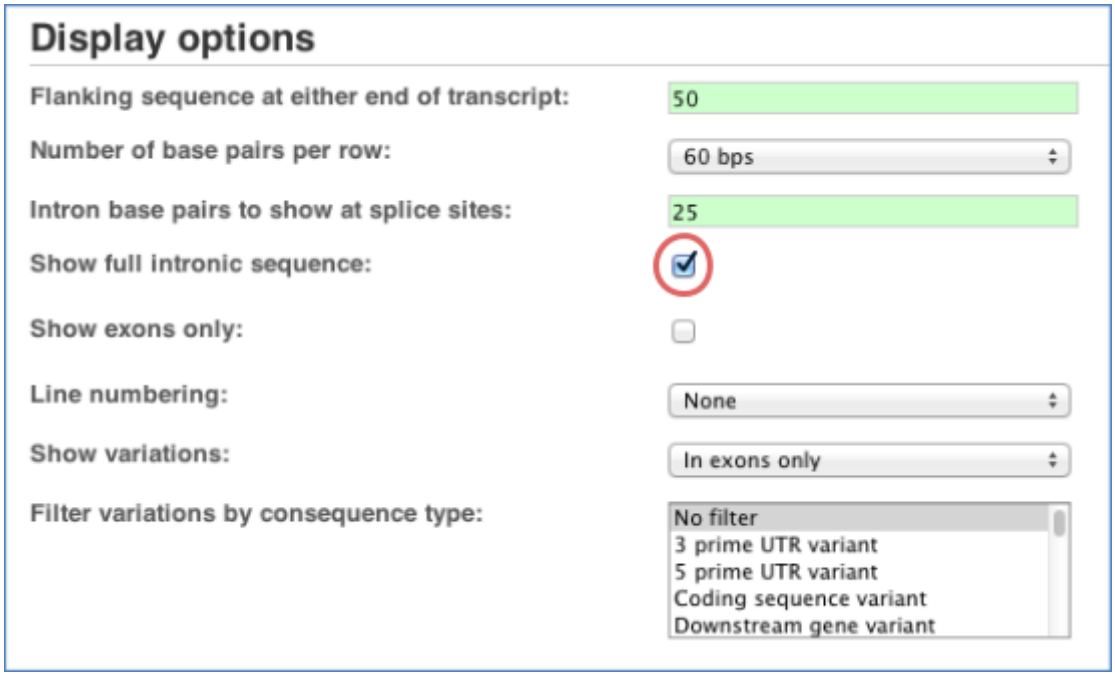


Click on the identifiers of the evidence to get a pop-up. This links out to the original records of these data in, for example, RefSeq, Uniprot or ENA.

Click on the Exons link.



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.

```
                M  RR     ***                          R           *
 1  CATTTGGGCAACGGACGAATTAAGCGGCGTGGAGATGCGGCAGCCACTGAGCTGGGGCCG
    ...................................ATGCGGCAGCCACTGAGCTGGGGCCG
    ...................................-M--R--Q--P--L--S--W--G--R

                                  S   Y
61  CTGGCGCGCGATGCTTGCCCGAACCTACGGCCCCGGGCCTTCGGCGGGTTACCGCTGGGC
27  CTGGCGCGCGATGCTTGCCCGAACCTACGGCCCCGGGCCTTCGGCGGGTTACCGCTGGGC
 9  --W--R--A--M--L--A--R--T--Y--G--P--G--P--S--A--G--Y--R--W--A
```

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.

This page shows information from other databases such as RefSeq, UniProtKB, CCDS and others that match to the Ensembl transcript and protein.

Show All entries Filter

External database	Database identifier
HGNC Symbol	NUDT6 nudix (nucleoside diphosphate linked moiety X)-type motif 6 [view all locations]
UniParc	UPI00001308E2 [view all locations]
CCDS	CCDS43268.1 [view all locations]
UniProtKB/Swiss-Prot	NUDT6_HUMAN [align] Nucleoside diphosphate-linked moiety X motif 6 [view all locations]
RefSeq peptide	NP_009014.2 [Target %id: 100; Query %id: 100] [align] nucleoside diphosphate-linked moiety X motif 6 isoform a [view all locations]
RefSeq mRNA	NM_007083.4 [align] [view all locations]
UCSC Stable ID	uc003iew.3 [view all locations]
Human Protein Atlas	HPA039202 [view all locations] HPA039202 [view all locations]
European Nucleotide Archive	AF019632 [align] [view all locations] AF019633 [align] [view all locations] AK291871 [align] [view all locations] BC009842 [align] [view all locations] L31408 [align] [view all locations]
HGNC transcript name	NUDT6-001 nudix (nucleoside diphosphate linked moiety X)-type motif 6 [view all locations]
INSDC protein ID	AAA67062.1 [align] [view all locations] AAD01635.1 [align] [view all locations] AAD01636.2 [align] [view all locations] AAH09842.1 [align] [view all locations] BAF84560.1 [align] [view all locations]
PDB	3FXT [view all locations] 3H95 [view all locations]

Click on GO table to see GO terms from the Gene Ontology consortium. www.geneontology.org

Ontology table

- GO: Biological process
- GO: Cellular component
- GO: Molecular function

Descendants of GO: Biological process

Show/hide columns Filter

Accession	Term	Evidence	Annotation Source	GOSlim Accessions	GOSlim Terms
GO:0008150	biological_process	ND			

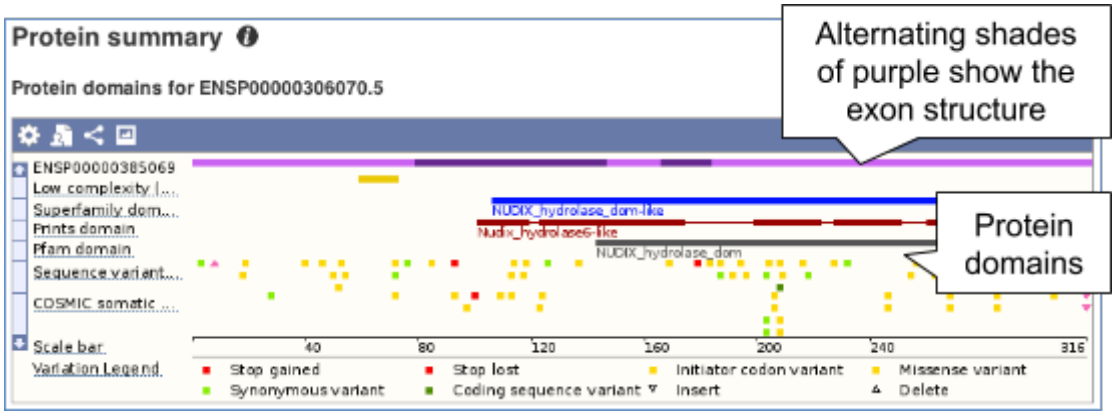
Descendants of GO: Cellular component

Show/hide columns Filter

Accession	Term	Evidence	Annotation Source	GOSlim Accessions	GOSlim Terms
GO:0005575	cellular_component	ND			
GO:0005634	nucleus	IEA		GO:0005575 GO:0043226 GO:0005623 GO:0005622	cellular_component organelle cell intracellular

Hover over the three-letter Evidence codes to see what they mean.

Now click on Protein summary to view domains from Pfam, PROSITE, Superfamily, InterPro, and more.



Clicking on Domains & features shows a table of this information.

Domains & features

Domains

Show/hide columns Filter

Domain type	Start	End	Description	Accession	InterPro
Pfam	143	265	NUDIX_hydrolase_dom	PF00293	IPR000086 [Display all genes with this domain]
Superfamily	106	306	NUDIX_hydrolase_dom-like	SSF55811	IPR015797 [Display all genes with this domain]
Prints	101	119	Nudix_hydrolase6-like	PR01356	IPR003293 [Display all genes with this domain]
Prints	123	139	Nudix_hydrolase6-like	PR01356	IPR003293 [Display all genes with this domain]
Prints	139	157	Nudix_hydrolase6-like	PR01356	IPR003293 [Display all genes with this domain]

Exercises - Browser

Exercise 1 – Exploring a genomic region in human

(a) Go to the region from 31,937,000 to 32,633,000 bp on human chromosome 13. On which cytogenetic band is this region located? How many contigs make up this portion of the assembly (contigs are contiguous stretches of DNA sequence that have been assembled solely based on direct sequencing information)?

(b) Zoom in on the *BRCA2* gene.

(c) Turn on the Tilepath track in this view. What is this track? Are there any Tilepath clones that contain the complete *BRCA2* gene?

(d) Create a Share link for this display. Email it to your neighbour. Open the link they sent you and compare. If there are differences, can you work out why.

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

(f) Turn off all tracks you added to the Region in detail page.

Exercise 2 – Exploring assembly exceptions in human

(a) Go to the region 21:32630000-32870000 in human. What is the red highlighted region?

(b) Can you see the assembly exceptions in the chromosome view? How many are there on chromosome 21? Drag out a box to jump to a region containing the leftmost assembly exception in 21q11.2 (note: you must drag out a region smaller than 1Mb). What is the name of this assembly exception?

(c) Can you compare this assembly exception with the reference? What is different between this assembly exception and the version on the primary assembly?

Exercise 3 – Exploring the human *MYH9* gene

(a) Find the human *MYH9* (myosin, heavy chain 9, non-muscle) gene, and go to the Gene tab.

- On which chromosome and which strand of the genome is this gene located?
- How many transcripts (splice variants) are there and how many are protein coding?
- What is the longest transcript, and how long is the protein it encodes?
- Which transcript is the best quality?

(b) Click on Phenotype at the left side of the page. Are there any diseases associated with this gene, according to OMIM (Online Mendelian Inheritance in Man)?

(c) In the transcript table, click on the transcript ID for MYH9-001, and go to the Transcript tab.

- How many exons does it have?

- Are any of the exons completely or partially untranslated?
- Is there an associated sequence in UniProtKB/Swiss-Prot? Have a look at the General identifiers for this transcript.
- What are some functions of MYH9-001 according to the Gene Ontology consortium? Have a look at the GO table for this transcript.

(d) Are there microarray (oligo) probes that can be used to monitor ENST00000216181 expression?

Exercise 4 – Finding a gene associated with a phenotype

Phenylketonuria is a genetic disorder caused by an inability to metabolise phenylalanine in any body tissue. This results in an accumulation of phenylalanine causing seizures and mental retardation.

(a) Search for phenylketonuria from the Ensembl homepage. What gene is associated with this disorder?

(b) How many protein coding transcripts does this gene have? View all of these in the transcript comparison view.

(c) What is the MIM disease identifier for this gene?

Exercise Answers:

Exercise 1 – Exploring a genomic region in human

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

Select Search: Human and type 13:31937000-32633000 in the text box (or alternatively leave the Search drop-down list like it is and type human 13:31937000-32633000 in the text box).
Click Go.

This genomic region is located on cytogenetic band q13.1. It is made up of eight contigs, indicated by the alternating light and dark blue coloured bars in the Contigs track. Note that KF455761.1 is a tiny contig that splits AL137143.8 in two.

(b) Draw with your mouse a box encompassing the *BRCA2* transcripts. Click on Jump to region in the pop-up menu.

(c) Click Configure this page in the side menu (or on the cog wheel icon in the top left hand side of the bottom image).

Type tilepath in the Find a track text box.
Select Tilepath.
Click on the (i) button to find out more

The tilepath track shows the BAC clones that the assembly was based upon.

Save and close the new configuration by clicking on □ (or anywhere outside the pop-up window).

There is not just one clone that contains the complete *BRCA2* gene. The BAC clone RP11-37E23 contains most of the gene, but not its very 3' end (contained in RP11-298P3). This was reflected on the two contigs that make up the entire *BRCA2* gene (the Contigs track is on by default).

(d) Click Share this page in the side menu.

Select the link and copy.
Get your neighbour's email address and compose an email to them, paste the link in and send the message.

When you receive the link from them, open the email and click on your link. You should be able to view the page with the new configuration and data tracks they have added to in the Location tab. You might see differences where they specified a slightly different region to you, or where they have added different tracks.

(e) Click Export data in the side menu. Leave the default parameters as they are.
Click Next>.
Click on Text.

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

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

(f) Click Configure this page in the side menu.
Click Reset configuration.
Click .

Exercise 2 – Exploring assembly exceptions in human

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

Select Search: Human and type 21:32630000-32870000 in the text box (or alternatively leave the Search drop-down list like it is and type human 21:32630000-32870000 in the text box).
Click Go.

You will see a red highlighted region in the middle of this region. Click on the thin dark red bar in any of the three views to see the label **CHR_HSCHR21_3_CTG1_1:32769079-32843731**. Click on What are assembly exceptions? to open a new window which explains assembly exceptions.

(b) Assembly exceptions are marked in the chromosome view at the top.
There are seven haplotypes on chromosome 21.

Drag a box around the assembly exception in 21q11.2 (less than 1Mb) then click on Jump to region.

Scroll down to the Region in detail view and click on the thin dark red bar at the top of the assembly exception. A drop-down containing the name of the assembly exception will appear.
CHR_HSCHR21_1_CTG1_1

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

Scroll down the page to see the comparison between the haplotype and primary assembly. Aligned sequences are highlighted in pink and linked together in green.
The assembly exception CHR_HSCHR21_1_CTG1_1 contains an extra region compared to the primary assembly.

Exercise 3 – Exploring the human *MYH9* gene

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

Select Search: Human and type *MYH9*. Click Go.

Click on either the Ensembl ID ENSG00000100345 or the HGNC official gene name *MYH9*.

- Chromosome 22 on the reverse strand.
- Ensembl has 11 transcripts annotated for this gene, of which three are protein coding.

- The longest transcript is MYH9-001 and it codes for a protein of 1,960 amino acids
- MYH9-001 is the best quality transcript, as it has a CCDS associated with it, is TSL:1 and is Golden.

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

(c) Click on ENST00000216181

- It has 41 exons. This is shown in the Transcript summary or in the left hand side menu Exons.
- Click on the Exons link in this side menu. Exon 1 is completely untranslated, and exons 2 and 41 are partially untranslated (UTR sequence is shown in purple). You can also see this in the cDNA view if you click on the cDNA link in the left side menu.
- P35579 from UniProt/Swiss-Prot matches the translation of the Ensembl transcript. Click on P35579 to go to UniProtKB, or click align for the alignment.
- The Gene Ontology project (http://www.geneontology.org/) maps terms to a protein in three classes: biological process, cellular component, and molecular function. Meiotic spindle organisation, cell morphogenesis, and cytokinesis are some of the roles associated with MYH9-001.

(d) Click on Oligo probes in the side menu.

Probesets from Affymetrix, Agilent, Codelink, Illumina, and Phalanx match to this transcript sequence. Expression analysis with any of these probesets would reveal information about the transcript. Hint: this information can sometimes be found in the ArrayExpress Atlas: www.ebi.ac.uk/arrayexpress/

Exercise 4 – Finding a gene associated with a phenotype

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

Type phenylketonuria into the search box then click Go. Choose Gene from the left hand menu.

The gene associated with this disorder is *PAH*, phenylalanine hydroxylase, ENSG00000171759.

(b) If the transcript table is hidden, click on Show transcript table to see it.

There are four protein coding transcripts.

Click on Transcript comparison in the left hand menu. Click on Select transcripts. Either select all the transcripts labelled protein coding one-by-one, or click on the drop down and select Protein coding. Close the menu.

(c) Click on External references.

The MIM disease ID is 261600.